# DOCTORAL (PHD) THESIS

GERGELY HORVÁTH

University of Pannonia

2026.

**University of Pannonia**

**Department of Process Engineering**

# Model–Based Analysis and Optimization of Industrial MDI Manufacturing

**DOCTORAL (PhD) THESIS**

**Gergely Horváth**

DOI:10.18136/PE.2026.984

**Supervisor:**

**Tamás Varga, PhD, associate professor**

**Alex Kummer, PhD, associate professor**

**Doctoral School of Chemical Engineering and Material Sciences**

**2026**

**Model–Based Analysis and Optimization of Industrial MDI Manufacturing**

The thesis was prepared for the award of a doctoral degree (PhD) within the framework of the Doctoral School of Chemical Engineering and Material Sciences at University of Pannonia

in the discipline of Bio-, Environmental and Chemical Engineering Sciences

written by: Gergely Horváth

Supervisors: Dr. Tamás Varga, Dr. Alex Kummer

I recommend the dissertation for acceptance: yes / no.

…….
supervisor

I recommend the dissertation for acceptance: yes / no.

…….
supervisor

I recommend the dissertation for peer review.

…….
chair of the DDHC

The PhD-candidate has achieved ……. % at the public debate.

The composition of the Final Examination Committee:

chair:…….

reviewers:.……

members:…….

Veszprém, ….....

…….
chair of the committee

Qualification of degree: ….....

Veszprém, ….....

…….
chair of the UDHC

*“To study and not think is a waste. To think and not study is dangerous”*

- Confucius

*“Always and never are two words you should always remember never to use.”*

- Wendell Johnson

# Acknowledgements

First, I would like to thank Réka, my wife, who has always helped me through difficult times with patience, attention, and lots of encouragement. Without her support and patience, and many sacrifices – especially the long weekends spent working or studying – it would not have been possible. I would like to thank my parents, my mother and father, for all the love and support they have given me over the years and for my brother, who never hid his opinion and always challenged my thoughts, pushing me to never say anything unless I was completely sure. I also want to thank my mother-in-law and father-in-law, Erika and András, who always helped me with my mid-semester duties with great kindness, hospitality, and support in Veszprém.

I would like to thank my supervisor, Dr. Tamás Varga, who was the primary force behind my decision to resume my studies after a brief pause, and who was there at the beginning, but unfortunately could not be there at the end. I am very grateful to Dr. Alex Kummer, my supervisor, who had to grow up very quickly to take on a huge task, but was able to rise to the challenge and set an example to everyone in terms of work capacity, support, and never accepting imperfect work.

I would also like to thank my colleagues at BorsodChem, namely József Réti, Dr. Zoltán Kozár, Szilágyiné D. Szilvia and Zhao Nan, who, with their decades of professional experience and knowledge, have always been tireless, quick and eager to help whenever I had a question or something to discuss about the technologies and made it possible for me to start my studies in the first place. A huge thanks to my team, to all my colleagues at Process Engineering, for always being supportive, helpful, honest, and kind to me. I consider myself lucky to be part of this team!

Finally, I would like to thank all my friends who always charged me up with energy during our meetups, all the teachers and staff at Pannon University, as well as those who accompanied me on this long journey and believed in me, I hope I have earned the trust.

# Abstract

## Model–Based Analysis and Optimization of Industrial MDI Manufacturing

The production of methylene diphenyl diisocyanate (MDI) is a cornerstone of polyurethane chemistry, with growing industrial importance due to its broad application range and increasing demand for high-performance polymer materials. MDI synthesis, however, remains a complex, multi-step process, characterized by intricate reaction networks, sensitive operational parameters, and environmental challenges. This dissertation presents a comprehensive, model-based and data-driven investigation of the industrial production of MDI, with particular focus on the synthesis and behavior of its primary precursor, methylenedianiline (MDA). The research encompasses both kinetic modeling and machine learning techniques to understand, predict, and optimize process behaviors. A detailed kinetic model was developed to describe the formation of MDA through the reaction of aniline and formaldehyde under acidic conditions. The model includes an extended reaction network and supports parameter sensitivity analysis, offering insight into the influence of molar ratios, residence times, and temperature profiles on product distribution, ring formation, and isomer ratios The identification and validation of model parameters was done using laboratory-scale experiments and available literature data demonstrating its reliability and adaptability across varying process conditions. In parallel, multiple soft sensor models were created to estimate complex product quality attributes, such as the color of MDI mixtures. These models employ advanced machine learning algorithms, including linear regression, regression trees, neural networks, support vector machines, and Gaussian process regression. Feature selection methods—such as Minimum Redundancy Maximum Relevance (MRMR), F-test method, ReliefF algorithm, Correlation-based and Aggregated techniques — were applied to identify the most influential process variables, while time-delay analysis and Bayesian optimization were used to enhance model accuracy and robustness. Among the models, Gaussian Process Regression showed superior predictive performance and the optimal operating parameters were determined by combining the model with Genetic Algorithm. The integration of mechanistic and data-driven modeling approaches enables the identification of optimal production conditions, reduction of undesirable by-products, and the development of explainable, interpretable models aligned with industrial requirements. These outcomes contribute to the advancement of sustainable, efficient, and intelligent MDI production technologies.

# Kivonat


## Modellalapú elemzés és optimalizálás az ipari MDI-gyártásban

A metiléndifenil-diizocianát (MDI) előállítása a poliuretánkémia meghatározó folyamata, mely egyre nagyobb ipari jelentőséggel bír a széleskörű alkalmazhatósága és a magas minőségű polimerek iránti növekvő igény miatt. Az MDI szintézise azonban összetett, többlépcsős folyamat, amelyet bonyolult reakcióhálózatok, érzékeny működési paraméterek és környezeti kihívások jellemeznek. A disszertáció célja az ipari MDI-gyártás komplex, modellalapú és adatvezérelt elemzése, különös tekintettel a fő intermedier, a metiléndianilin (MDA) szintézisére és viselkedésére. A kutatás kiterjed mind a kinetikai modellezésre, mind a gépi tanulási technikák alkalmazására a folyamatok megértése, előrejelzése és optimalizálása érdekében. Egy részletes kinetikai modellt dolgoztam ki az MDA képződésének leírására, amely az anilin és a formaldehid savas közegben történő reakcióján alapul. A modell kiterjesztett reakcióhálózatot tartalmaz, és lehetővé teszi a folyamat paraméterérzékenység vizsgálatát, betekintést nyújtva a mólarányok, tartózkodási idők és hőmérsékleti profilok termékeloszlásra, gyűrűképződésre és izomerarányokra gyakorolt hatásába. A modell paramétereit laboratóriumi méretű kísérletek és szakirodalmi adatok segítségével identifikáltuk és validáltuk, ami igazolja annak megbízhatóságát és alkalmazhatóságát változó folyamatkörülmények között is. Ezzel párhuzamosan különböző soft szenzor modellek kerültek fejlesztésre az MDI keverékek színének, mint komplex minőségi paraméternek a becslésére. A modellek különböző gépi tanulási algoritmusokat alkalmaznak, többek között lineáris regressziót, döntési fákat, neurális hálókat, Support Vector Machine és Gauss-folyamat alapú regressziót. A legfontosabb változók azonosításához változószelekciós módszereket – például Minimális Redundancia, Maximális Relevancia algoritmust (MRMR), F-teszt módszert, ReliefF algoritmust, Korrelációalapú és Aggregált technikákat – alkalmaztam, holtidő-analízissel és Bayes-féle optimalizációval pedig a modellek pontosságát javítottam. A modellek közül a Gauss-folyamat regresszió kiváló predikciós teljesítményt mutatott, az optimális működési paraméterek pedig a modell Genetikai Algoritmussal való kombinálásával kerültek meghatározásra. A modellezési megközelítések kombinációja lehetővé tette az optimális működési feltételek meghatározását, a melléktermékek csökkentését és ipari szinten is értelmezhető, magyarázható modellek kidolgozását. Az eredmények hozzájárulnak a fenntartható, hatékony és intelligens MDI-gyártási technológiák fejlődéséhez.

# Auszug


## Modellbasierte Analyse und Optimierung der industriellen MDI-Herstellung

Die Herstellung von Methylendiphenyldiisocyanat (MDI) ist ein zentraler Prozess der Polyurethan-Chemie mit wachsender industrieller Bedeutung, jedoch durch komplexe Reaktionsnetzwerke, empfindliche Prozessparameter und ökologische Herausforderungen gekennzeichnet. Diese Dissertation präsentiert eine modellbasierte und datengestützte Untersuchung der industriellen MDI-Produktion mit Fokus auf die Synthese und das Verhalten des Vorläufers Methylendianilin (MDA). Zur Beschreibung der MDA-Bildung aus der Reaktion von Anilin und Formaldehyd unter sauren Bedingungen wurde ein detailliertes kinetisches Modell mit erweitertem Reaktionsnetzwerk entwickelt. Sensitivitätsanalysen erlauben die Untersuchung des Einflusses von molaren Verhältnissen, Verweilzeiten und Temperaturprofilen auf Produktverteilung, Ringbildung und Isomerenverhältnisse. Die Identifikation und Validierung der Modellparameter erfolgte anhand von Technikumsversuchen sowie Literaturdaten und belegt die Übertragbarkeit des Modells auf unterschiedliche Prozessbedingungen. Ergänzend wurden Soft-Sensor-Modelle zur Schätzung komplexer Produkteigenschaften, insbesondere der Farbe von MDI-Mischungen, entwickelt. Zum Einsatz kamen lineare Regression, Regressionsbäume, künstliche neuronale Netze, Support Vector Machines und Gaußsche Prozessregression. Die Auswahl relevanter Prozessvariablen erfolgte mittels Minimum Redundancy Maximum Relevance (MRMR), F-Test-Methode, ReliefF-Algorithmus, korrelationsbasierter sowie aggregierter Merkmalsauswahlverfahren. Zusätzlich wurden Zeitverzögerungsanalysen und Bayessche Optimierung zur Berücksichtigung dynamischer Effekte und zur Hyperparameterabstimmung angewendet. Die Gaußsche Prozessregression zeigte die beste Vorhersageleistung; in Kombination mit einem genetischen Algorithmus wurden optimale Betriebsparameter zur Qualitätsverbesserung und Reduktion von Nebenprodukten bestimmt. Die Integration mechanistischer und datengetriebener Ansätze ermöglicht erklärbare, industriell einsetzbare Modelle und trägt zu einer nachhaltigen, effizienten MDI-Produktion bei.

# Table of Contents

# List of Notations

## Mathematical symbols

| | |
|---|---|
| $R^2$ | Coefficient of determination |
| $y_k$ | k-th observed output value |
| $\hat{y}_k$ | k-th calculated output value |
| $\bar{y}$ | mean of all observed output value |
| $y_m^l$ | output of neuron *m* in layer *l* |
| N | total number of data points |
| f | function, machine learning models or activation function |
| d | time delay |
| $\theta$ | parameter matrix of machine learning models |
| $\varepsilon_k$ | measurement error for the *k*-th observation |
| $\beta_0$ | intercept term |
| P | Genetic Algorith population |
| $F_t$ | fitness value of t-th solution |
| $\alpha$ | crossover weight or blend factor |
| p | selected predictors for stepwise linear regression |
| X | predictor matrix |
| X* | test point matrix |
| $x_S$ | partial dependence subset of variables from X |
| $X_C$ | complementary set of $X^S$ in X |
| $x_k$ | input value for the k-th observation |
| $\beta_i$ | i-th regression coefficient |
| n | number of features, population size or reaction order |

| | |
|---|---|
| $R_i$ | i-th regression tree region |
| i | i-th feature |
| s | split point for regression tree |
| $w_{mj}$ | weight of neurons from neuron *j* to neuron *m* |
| Φ | bias term |
| $a_j$ | activated output |
| ε | margin of tolerance |
| w | weight vector |
| $\xi_i$ | slack parameters for positive deviations beyond ε |
| $\xi_i^*$ | slack parameters for negative deviations beyond ε |
| λ | regularization parameter |
| μ | mean, mean function |
| $N$ | normal distribution function |
| Φ (·) | feature mapping function |
| p (·) | probability function |
| $c_\kappa$ | concenration of component κ |
| ν | stoichiometric coefficient |
| k | reaction rate constant, covariance function |
| $\omega_{catalyst}$ | catalyst weight fraction |
| $k_0$ | preexponential factor |
| $P_{enter}$, $P_{remove}$ | parameter selection criteria parameters for stepwise linear regression |
| c | Pearson correlation |
| ψ | Shapley value |
| υ | characteristic function for Shapley values |

| | |
|---|---|
| $S$ | Shapley subsets |
| $\beta$ | GPR model mean value |
| $\sigma_L$ | characteristic length scale |
| $\sigma_F$ | signal standard deviation |
| $\sigma$ | noise standard deviation |

**Abbreviations**

| | |
|---|---|
| MDI | Methylenediphenyl diisocyanate |
| MDA | Methylenedianiline |
| LR | Linear Regression |
| SLR | Stepwise Linear Regression |
| RT | Regression Tree |
| NN | Neural Network |
| SVM | Support Vector Machine |
| GPR | Gaussian Process Regression |
| ANN | Artificial Neural Networks |
| BIC | Bayesian Information Criterion |
| LCA | Life Cycle Assessment |
| APC | Advanced Process Control |
| MPC | Model Predictive Control |
| AI | Artificial Intelligence |
| ML | Machine Learning |
| DL | Deep Learning |
| k-NN | k-nearest neighbors |

| | |
|---|---|
| MSE | Mean Squared Error |
| RMSE | Root Mean Squared Error |
| MAE | Mean Average Error |
| PCA | Principal Component Analysis |
| t-SNE | t-Distributed Stochastic Neighbor Embedding |
| KDDM | Knowledge Discovery and Data Mining |
| CRISP-DM | Cross-Industry Standard Process for Data Mining |
| CRISP-ML(Q) | Cross-Industry Standard Process for the development of Machine Learning applications with Quality assurance methodology |
| GA | Genetic Algorithm |
| PSO | Particle Swarm Optimization |
| SSE | Sum of Squared Errors |
| SVR | Support Vector Regression |
| RBF | Radial Basis Function |
| GP | Gaussian Process |
| TDI | Toluene Diisocyanate |
| KPI | Key Performance Indicator |
| ABA | Aminobenzylaniline |
| A/F | Aniline / Formaldehyde molar ratio |
| H/A | HCl / Aniline molar ratio |
| TOS | Time On Stream |
| MMM | Mono-methyl-MDA |
| RR | Remaining Reactants |
| IR | Isomer Ratio |

| | |
|---|---|
| OF | Oligomer Fraction |
| NS | N-substituted by-products |
| NMB | N-methylbenzene |
| LB | Lower Bounds |
| UB | Upper Bounds |
| W/A | Water / Aniline molar ratio |
| Tc | Condensation reactor temperature |
| $T_r$ | Rearrangement reactor temperature |
| $t_c$ | Condensation reaction time |
| $t_r$ | Rearrangement reaction time |
| $F_{R1}$ | Formalin feed ratio to the first reactor |
| $F_{R2}$ | Formalin feed ratio to the second reactor |
| $t_F$ | Formalin dosing time |
| 2r | 2-ring MDA content |
| 3r | 3-ring MDA content |
| 4r | 4-ring MDA content |
| 5r | 5-ring MDA content |
| >6r | ≥ 6-ring MDA content |
| >4r | ≥ 4-ring MDA content |
| O-O | Ortho-ortho isomer ratio |
| O-P | Ortho-para isomer ratio |
| P-P | Para-para isomer ratio |
| IQR | Interquartile Range |
| MIMO | Multi-Input and Multi-Output |

| | |
|---|---|
| ODCB | Ortho-dichlorobenzene |
| CCC | chloroformamidine-N-carbonyl chloride |
| PC | Principal Component |
| PDP | Partial Dependence Plot |
| ICE | Individual Conditional Expectation |

# 1 Introduction

In the modern chemical industry, mathematical modeling plays a crucial role in understanding, designing, and optimizing chemical processes. As industries strive for efficiency, safety, and sustainability, mathematical models provide a structured approach to predict system behavior, minimize costs, and improve operational performance. By using functions that describe mass and energy balances, reaction kinetics, and transport phenomena, engineers can simulate real-world processes in order to optimize operation, decrease production costs, prepare the foundations of feasibility studies or can be a tool of investigation before implementing costly experimental trials.

This mathematical model based predictive capability not only accelerates innovation but also enhances process control, ensuring consistent product quality, process safety and adherence to environmental regulations. In addition, with increasing complexities in chemical production—such as multi-phase reactions, dynamic process control, and integration of renewable resources—advanced mathematical models help industries overcome challenges. These models facilitate enhanced decision-making in domains such as process optimization, risk assessment, and equipment design, thereby contributing to the development of safer and more efficient chemical facilities.

In this context, the application of mathematical modeling extends beyond traditional engineering calculations; it now integrates computational tools, artificial intelligence, and real-time simulations. These advancements empower chemical engineers to refine processes, reduce environmental impact, and drive the industry toward a more sustainable future. Thus, mathematical modeling is not just a theoretical exercise but a fundamental pillar of modern chemical engineering, ensuring innovation, safety, and efficiency in an increasingly complex industrial landscape.

A prime example of the impact of mathematical modeling in chemical engineering is its role in optimizing complex industrial processes, such as the synthesis of isocyanates. Through the utilization of computational simulations and machine learning algorithms, engineers have the capacity to construct predictive models that precisely represent reaction kinetics, heat and mass transfer, as well as system dynamics. Such models facilitate real-time adjustments to processes, thereby enhancing efficiency and sustainability while minimizing dependence on expensive and labor-intensive experimental trials. Furthermore, the integration of artificial

intelligence into process modeling augments decision-making capabilities by discerning optimal operating conditions and revealing patterns that may be disregarded by conventional methods.

The synthesis of various isocyanates represents a process that is both highly energy-demanding and complex, thus offering significant opportunities and potential for the advancement of technologies and formulations aimed at sustainable development. Nevertheless, the optimization of this production process necessitates the employment of a model that is both validated with empirical measurement data and accurately represents the actual system. This is crucial for achieving notable reductions in time and cost, while simultaneously enhancing product quality. Both kinetic models and machine learning models can facilitate the manufacturing process with greater speed and precision, thereby increasing product selectivity to align with market demands, minimizing the formation of by-products and energy consumption, and enhancing key performance indicators to desired levels.

In my research, I have developed methods for model-based and data-driven analyses of the formation process and reactions of Methylenediphenyl diisocyanate (MDI) and Methylenedianiline (MDA) intermediates, the kinetics of the reactions and other difficult-to-define parameters such as the prediction of the colour of MDI mixtures. The soft-sensors and models developed were used to support industrial processes by defining different flexible and challenging objectives. We also aimed to ensure that the results of each of the developed models could be interpreted in an explainable way based on industrial experience. I have devised methodologies for both model-based and data-driven analyses concerning the formation processes and reactions of MDI and MDA intermediates. This includes the kinetics of these reactions and other complex parameters, such as predicting the color of MDI mixtures. The developed soft-sensors and models were utilized to aid industrial processes by defining various flexible and challenging objectives. Furthermore, efforts were made to ensure that the outcomes of each developed model could be interpreted in an explainable way, supported by industrial experience.

The aim of my research was to examine the individual components and processes of MDI production as an integrated system. Accordingly, the objectives and main research questions were formulated to reflect this systemic approach, while the answers and findings were summarized in the form of thesis points:

- Is it possible to define a kinetic model capable of accurately describing the formation of MDA and other secondary components generated during the synthesis reactions, while remaining applicable under relevant industrial production parameters (e.g. aniline/formaline molar ratio)?
- Is it possible to use machine learning models to describe the synthesis of MDA in order to achieve a deeper understanding of the effects of various production parameters on product quality, and potentially predict these effects?
- Is it possible to identify the reaction pathways responsible for the formation of by-products that cause color degradation during MDI production, and if identified, what strategies can be applied in industrial systems to reduce their formation?
- Is it possible to create and implement a regression-based machine learning model as a soft sensor in industrial systems to predict MDI color using operating parameters, thus aiding in the management and enhancement of industrial processes?

Section 2 initially highlights the significance of mathematical modeling and the methodologies for its utilization. It discusses the types of models that can be developed based on the basis of reasoning and the availability of data over time. In addition to traditional chemical engineering models, a brief introduction to data-driven modeling and machine learning is anticipated.

Subsequent to the modeling techniques, Section 3 delivers a comprehensive analysis of the raw materials employed in the production of isocyanates, as well as the design, structure, and operation of industrial systems. It further elucidates the challenges and complexities inherent in the MDI synthesis process.

In Section 4, a comprehensive kinetic model was devised to describe the MDA synthesis process. MDA is synthesized industrially through the reaction of aniline and formaldehyde, catalyzed by inorganic acids, with hydrochloric acid (HCl) being the most prevalent. The parameters employed in the MDA synthesis critically influence the quality attributes of the resulting MDA product mixture, which in turn affect the characteristics of the final MDI product mixture. While MDA serves as a crucial industrial intermediate, its synthesis at an industrial scale is predominantly guided by empirical methodologies, experiential knowledge, and heuristic techniques. However, the impact of production parameters and their extents on product properties remain undocumented in the relevant literatures. The findings of our study indicate that increasing the complexity of existing kinetic models by defining new components

and reactions to MDA synthesis can yield a more robust and adaptable model for the description of the reaction system.

Subsequently in Section 5, a machine learning model was developed leveraging experimental laboratory synthesis data, wherein the correlations between independent operating parameters and dependent product quality parameters characterizing MDA mixtures were determined and explored through the development of various regression models acting as soft-sensors. Upon examining the laboratory experimental data from a set of several individual experiments, it was discovered that machine learning models could describe independent parameters with satisfactory accuracy. The machine learning models discussed in this thesis illustrate the potential to develop models capable of adjusting ring distribution, isomer ratios, and selectivity of the reaction while concurrently minimizing by-product quantities in accordance with market demands or different objective functions by fine-tuning production parameters, thereby achieving an optimal product portfolio and operational cost for the industrial process.

Estimating the color of MDI mixtures as a quality parameter presents significant challenges when utilizing conventional kinetic models. Consequently, a soft-sensor was developed and presented in Section 6 to facilitate the estimation of this parameter. Throughout our research, we employed five distinct feature selection methodologies – namely, MRMR, F-test, ReliefF, Correlation-based and a so-called Aggregated approach, which fused the results of the first four methods in order to derive an optimal set of features, from which MRMR results proved to be the best for selecting important features. Correlation-based methods were specifically employed to ascertain and incorporate time delay for each operational parameter, which significantly influenced the models' accuracy. Subsequently, the performance of five different machine learning models, including Linear Regression (LR), Regression Tree (RT), Neural Network (NN), Support Vector Machine Learning (SVM), and Gaussian Process Regression (GPR), was assessed with Bayesian hyperparameter optimization applied where it was feasible. Among these models, the GPR models exhibited superior performance. To elucidate the outcomes of the GPR model, partial dependence plots were computed and compared with empirical industrial experiences and observations. Furthermore, genetic algorithms were implemented based on the developed GPR model to acquire an optimal parameter set, and the robustness of the optimal solution was also verified by sensitivity studies.

# 2 Importance of model-based analysis and optimization

The preparation of mathematical models is essential for advancing the understanding of isocyanate production. These models provide a systematic framework to analyze the complex chemical reactions involved in the synthesis of isocyanates, which are key intermediates in the production of polyurethane and other polymers. Mathematical modeling has long played a crucial role in both science and engineering, transforming qualitative questions about observed phenomena into mathematical problems. From the earliest stages of mathematical development, formulating real-world phenomena as mathematical problems has been both a driving force and an essential component of the discipline.

In chemical engineering, computational power is particularly crucial, as the main physical and chemical principles are inherently complex. These processes often involve not only heat, mass, and momentum transfer but also chemical reactions, reaction enthalpy, adsorption, desorption, phase transitions, and multiphase flows. The complex interaction of these factors makes mathematical modeling both challenging and inevitable to understand complex interactions. Despite being abstractions of real systems, mathematical models are invaluable tools for scientists and engineers. They provide a structured approach to understanding and predicting the behavior of chemical and industrial processes, enabling optimization, innovation, and more efficient decision-making [1].

The rise of artificial intelligence (AI) and machine learning (ML) has significantly altered mathematical modeling in the chemical industrial sector as well [2]. Traditional kinetic models are increasingly integrated with ML frameworks that identify complex connections among process variables without requiring direct physical equations. Advanced techniques like deep learning, support vector machines, and Bayesian optimization are used to improve process predictability and automate decision-making processes. Data-driven surrogate modeling lowers the computational expenses tied to high-fidelity simulations by developing simplified models that maintain predictive accuracy. These surrogate models allow for real-time optimization, making process alterations more rapid and effective. Furthermore, there is ongoing research into using reinforcement learning algorithms for autonomous process control, where intelligent agents develop optimal control strategies by continuously interacting with simulated environments.

The general aims of developing mathematical models to represent industrial processes include:

1. **Process Optimization and Efficiency Increasing:** One of the primary uses of mathematical modeling in the chemical industry involves optimizing industrial processes. By developing kinetic or data-driven models, engineers can forecast system behavior under various operating conditions. This helps in selecting optimal parameters to increase yield, reduce energy consumption, and minimize waste. Classical optimization methods, such as linear and nonlinear programming, are widely applied to improve process conditions. Meanwhile, kinetic models, modern machine learning techniques, and hybrid models offer more dynamic solutions through the examination of both historical and real-time process data [3].
2. **Understanding Reaction Kinetics and Mechanisms:** In the realm of chemical reaction engineering, mathematical models are essential for grasping reaction kinetics and reactor performance. Reaction rate equations, drawn from the foundational principles of chemical kinetics, are employed to delineate concentration profiles, reaction pathways, and states of equilibrium [4]. Kinetic models are categorized as empirical, semi-empirical, or fundamental, each varying in accuracy and computational demand. Traditional kinetic modeling uses differential equations to describe how reactant and product concentrations change over time. Recent advancements in computational techniques have led to the development of hybrid models that integrate kinetic theory with data-driven approaches, such as artificial neural networks (ANNs) and genetic algorithms. These new methods improve the ability to forecast unique reaction behaviors, including nonlinear dynamics, catalyst deactivation, and enhancement of selectivity even within multi-phase reactors.
3. **Sustainable Process Development:** With increasing environmental concerns and stricter regulatory requirements, mathematical modeling plays a crucial role in designing sustainable and environmentally friendly chemical processes. Life Cycle Assessment (LCA) models help evaluate the environmental impact of raw material extraction, production, transportation, and waste disposal, enabling industries to implement greener alternatives [5]. Process models are also extensively used in waste minimization strategies, and energy efficiency improvement. For instance, mathematical models of chemical absorption and adsorption processes aid in designing more efficient $CO_2$ capture systems, reducing greenhouse gas emissions from industrial sources. Similarly, energy integration models using pinch analysis assist in optimizing heat exchanger networks, leading to significant energy savings in large-scale industrial chemical plants.

4. **Real-Time Process Control:** Advanced process control (APC) strategies rely heavily on mathematical modeling to regulate complex chemical processes in real-time. Model predictive control (MPC), on the other hand, uses dynamic models to anticipate future process behavior and adjust control variables proactively, minimizing disturbances and improving process stability [6]. The integration of digital twin technology—virtual replicas of physical systems based on mathematical models — further enhancing real-time decision-making by continuously updating models based on live plant data [7].

## 2.1 Modeling strategies

Within the discipline of the aforementioned chemical engineering, modeling strategies can be categorized as either a priori and a posteriori, or as deductive and inductive approaches. Consequently, a thorough analysis of the differences, similarities, advantages, and disadvantages associated with various modeling strategies is imperative for the development of an optimal model that accurately describes and facilitates the investigation of isocyanate production through modeling techniques.

This section presents the four previously mentioned modeling strategies, each offering distinct benefits and tailored for different uses. The decision of when, how, or in what combination to use them is the responsibility of the modeler:

- **Basis of reasoning:** Deductive vs. Inductive modeling
- **Timing of data utilization:** A priori vs. A posteriori modeling (before or after data)

The four primary scientific modeling approaches – deductive modeling, inductive modeling, a priori modeling, and a posteriori modeling – can be categorized based on the reasoning method and the timing of data incorporation. Deductive modeling is driven by theory, where models are constructed using established principles and laws to derive specific predictions. It aligns primarily with a priori modeling, as it is developed before data collection. Inductive modeling is data-driven, wherein models are built from observed patterns or relationships within empirical data. This approach is most commonly associated with a posteriori modeling, as it involves refining models after data collection. A priori modeling builds models based on theoretical assumptions or principles, without the need for prior data. It is often used when systems are well understood and theoretical predictions are needed before experimentation. A posteriori modeling refines or develops models based on collected data, validating theoretical models or uncovering new patterns. It includes both inductive (data-driven) and deductive (data-informed) approaches.

The relationship among the different modelling strategies is presented in Table 1 below:

**Table 1: Framework of modeling approaches**

| | **A priori** (before data) | **A posteriori** (after data) |
|---|---|---|
| **Deductive** | Theoretical model preparation *Example: Deriving fluid flow equations from Navier-Stokes equations* | Model validation or tuning *Example: Validating a kinetic model for laboratory data* |
| **Inductive** | Knowledge-based generalization *Example: Preparation of a rule-based system with historical datasets* | Data-driven modeling *Example: Training a machine learning model for operational data* |

Each approach serves specific objectives based on the system under consideration, data availability, and modeling goals. During model execution, various strategies may be employed depending on the available data, existing knowledge, or relevant theoretical frameworks. For instance, the clarification of the isocyanate reaction mechanism, which serves as the validation objective, illustrates such an application. These approaches are complementary and can be integrated in hybrid models to improve both predictive accuracy and theoretical understanding.

Table 2 provides a comparative analysis of two modeling approaches with respect to the timing of data utilization:

**Table 2: Comparison of A priori and A posteriori modeling features**

| Feature | A Priori Modeling | A Posteriori Modeling |
|---|---|---|
| **Model basis** | First principles (chemistry, physics) | Experimental data and correlations |
| **Prediction Accuracy** | High for well known systems | High inside experimental range, limited outside |
| **Data Requirement** | Low to moderate | High |
| **Flexibility** | Applicable to new processes | Limited to known conditions |
| **Example** | Thermodynamic equations | Regression models |

A comparison can also be made between two distinct modeling methodologies based on the foundational principles of reasoning: inductive and deductive approaches. The differences between these methodologies are illustrated in Figure 1:

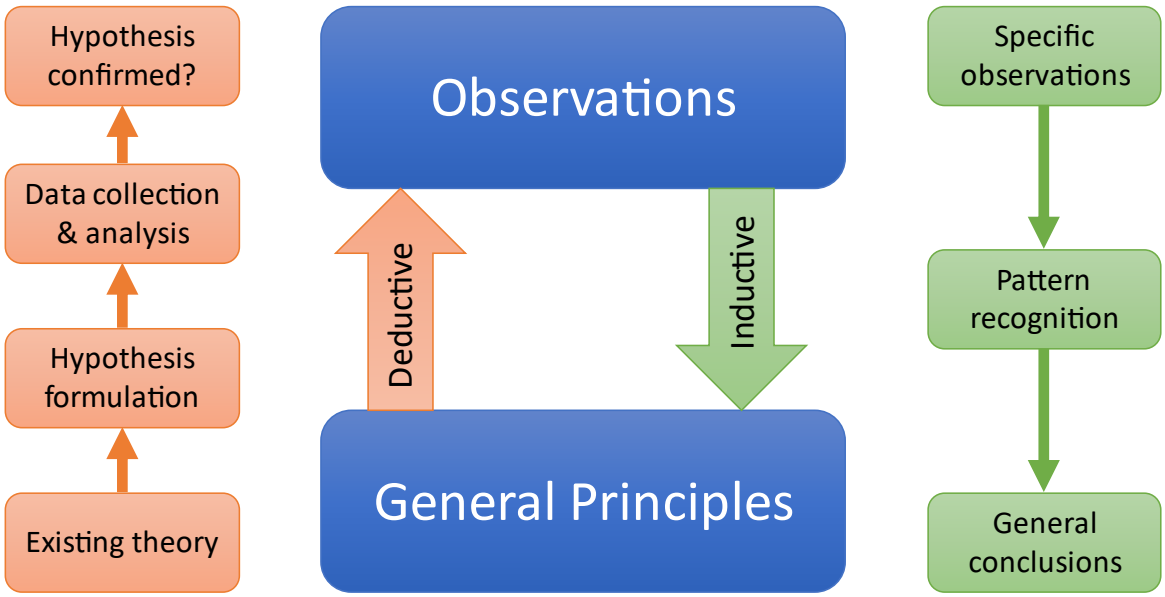


**Figure 1: Comparison of Inductive and Deductive modeling features**

In the subsequent sections, a comprehensive examination of the four modelling approaches previously introduced will be undertaken, focusing on clarifying their distinct strengths, general frameworks for application and fundamental concepts.

### 2.1.1 A priori approach

A priori modeling is a fundamental approach in chemical engineering that relies on first principles rather than empirical data to predict and analyze chemical processes. These models are constructed using well-established scientific laws such as mass and energy conservation, reaction kinetics, and thermodynamics. The predictive nature of a priori models makes them indispensable in process design, optimization, and scale-up, offering reliable insights even in the absence of extensive experimental data [8].

A key advantage of a priori modeling lies in its ability to develop theoretical or computational models based on established principles, prior knowledge, or assumptions made before any direct empirical observation or experimental data collection. By utilizing theoretical foundations, such models can simulate system responses under different scenarios, examine parameter sensitivities, and pinpoint critical thresholds or emerging phenomena [9]. Unlike empirical or data-driven models, a priori models are not fundamentally dependent on observational data for their construction but can be validated or adjusted with empirical data afterward. This makes them exceptionally valuable when data is limited, costly, or challenging to obtain for various reasons. Additionally, they provide a mechanistic insight into system dynamics, thereby improving interpretability and applicability.

However, the effectiveness of a priori modeling depends heavily on the validity of its underlying assumptions and the completeness of theoretical understanding. Simplifying assumptions made for tractability may limit model accuracy, especially in highly nonlinear or stochastic systems. Consequently, hybrid approaches that integrate a priori modeling with empirical calibration are increasingly being adopted to harness the strengths of both paradigms.

### 2.1.2 A posteriori approach

A posteriori modeling refers to the development or calibration of models based on empirical data or observations obtained after experimentation or measurement. Unlike a priori modeling, which relies on theoretical principles and assumptions established prior to data collection, a posteriori approaches are inherently data-driven. They are often used to uncover patterns, refine theoretical models, or validate predictions by fitting model parameters to real-world observations [10]. In this framework, models are constructed or adjusted to minimize discrepancies between simulated outcomes and observed phenomena. This includes statistical models, machine learning algorithms, and empirical curve-fitting techniques, as well as the parameter tuning of mechanistic models using optimization methods. A posteriori modeling

plays a central role in fields such as climate science, epidemiology, economics, and engineering, where complex systems are difficult to represent solely through first-principles equations.

The strength of a posteriori modeling lies in its ability to closely reflect observed behavior, allowing high predictive accuracy within the domain of the available data. It is particularly advantageous when the underlying mechanisms are poorly understood or too complex to model analytically. Moreover, a posteriori analysis is essential for model validation, uncertainty quantification, and performance assessment. However, such models may suffer from overfitting, limited generalizability, or dependence on the quality and quantity of the training data. Without adequate theoretical grounding, they may lack interpretability and provide limited insight into causal relationships. As such, a posteriori modeling is often most powerful when used in conjunction with a priori frameworks, enabling a balanced approach that leverages both theoretical understanding and empirical evidence.

### 2.1.3 Inductive approach

Inductive modeling is employed to characterize the process of deriving a general conclusion based on specific observations, such as identified patterns. This method is occasionally referred to as induction. The process of induction commences with a set of premises, mainly based on empirical observations or experimental data, and utilizes these premises to extrapolate a general conclusion [11].

Inductive modeling is a data-driven methodology that involves constructing models based on observed patterns, correlations, or regularities within empirical data. Unlike deductive or a priori approaches, which begin with theoretical assumptions or governing equations, inductive modeling relies on the extraction of knowledge directly from data, often without prior assumptions about the underlying system dynamics [12]. This approach is particularly useful when the system under study is too complex, nonlinear, or poorly understood to be accurately described by first-principles models. By leveraging statistical learning, machine learning algorithms, and pattern recognition techniques, inductive models aim to generalize from specific observations to predict unseen or future behavior. Common examples include regression analysis, decision trees, neural networks, and support vector machines. Inductive modeling is widely used across scientific disciplines, including bioinformatics, environmental science, chemical engineering, and economics. In chemical process optimization, inductive models can predict product yield or energy consumption based on historical operational data.

One of the key advantages of inductive modeling is its flexibility and adaptability to large, high-dimensional datasets. It can uncover complex relationships that may not be obvious through traditional analysis, making it suitable for exploratory research and predictive analytics. Furthermore, it facilitates rapid model development, especially in scenarios where theoretical models are infeasible or unavailable. However, inductive models also present several limitations. They are inherently dependent on the quality and representativeness of the input data, and their predictive power may degrade outside the domain of the training set. Additionally, many inductive models function as "black boxes," offering limited interpretability or insight into causal mechanisms. This can pose challenges in scientific fields where understanding system behavior is as important as prediction. To address these challenges, hybrid modeling strategies are increasingly adopted, combining the strengths of both inductive and deductive a priori approaches. Such integration enables the development of robust, interpretable, and accurate models that are grounded in both data and theory.

### 2.1.4 Deductive approach

Deductive modeling refers to the approach of obtaining particular conclusions or forecasts based on general principles or established theoretical foundations. This approach, commonly linked to formal logic and the scientific method, is recognized as deduction. Within a deductive modeling process, one commences with a set of axioms, physical laws, or mathematical formulations, employing logical reasoning to infer system behavior under specified conditions. These premises are typically grounded in first-principles knowledge, such as thermodynamics, conservation laws, or reaction kinetics, with the model being constructed independently of observational data [13].

Deductive modeling is a methodology guided by theory, highlighting causality, comprehension, and internal coherence. In contrast to inductive or data-driven methods that derive models from observed patterns, deductive modeling uses a priori knowledge to describe system dynamics through structured equations and logical rules. This makes it especially suitable for domains where the governing principles of the system are well understood and quantitatively expressed. Common examples of deductive modeling include the use of differential equations to describe mass and energy balances in chemical reactors, the application of Newtonian mechanics in physical simulations, or the use of rate laws in kinetic modeling. In practice, deductive models are used extensively in engineering, physics, and environmental sciences, where compliance with the fundamental laws of physics is essential. For instance, in chemical engineering, the design of a distillation column can be modeled using thermodynamic

equilibrium relationships and mass transfer equations derived deductively. Such models enable robust simulation, optimization, and control of complex processes, even before empirical data are collected [14].

One of the primary strengths of deductive modeling is its explanatory power. Since models are built on transparent and interpretable mechanisms, they provide insights into how and why systems behave as they do. This interpretability also enhances their generalizability, allowing for extrapolation beyond observed data domains when assumptions remain valid. Moreover, deductive models can support hypothesis testing, theoretical exploration, and scenario analysis in fields where data collection may be limited or hazardous. However, deductive modeling also has inherent limitations. The effectiveness of deductive modeling depends on the existence and accuracy of underlying theoretical understanding. This knowledge might be lacking for some systems, particularly those marked by significant nonlinearity, randomness, or incomplete understanding, as encountered in isocyanate reaction systems. Additionally, deductive models often require simplifying assumptions (e.g., ideal behavior, steady-state operation), which may not hold in real-world applications. The complexity of solving analytical or numerical formulations can also be a barrier to their practical implementation.

## 2.2 Introduction to Machine Learning models

In recent years and decades, the design, optimisation, fault detection and other uses of chemical processes supported by different models and model types has accelerated, especially with the rise of various data-driven modelling technologies (e.g. artificial neural networks). There are several available articles where the authors could successfully implement models mainly for applications where both the objective functions and the constrains are non-linear, a good example is the study of the polymerization of nylon-6,6 using neural networks and the optimization of the process parameters, which models are promising for industrial applications [15]. Several applications and proposed improvements have been published on the topic of fault-detection in order to increase the availability of industrial systems, to detect and diagnose real faults, avoid unnecessary emergency shutdowns and protect process equipment. Successful application of these deep learning models can significantly reduce the number of accidents and shutdowns in industrial facilities, reduce unnecessary environmental impact from malfunctions and reduce the likelihood of damage to assets and equipment [16,17]. In many cases, the use of these data models can also significantly reduce the required computation times. An excellent example of this is the Steam Methane Reformer model published by N.D. Vo *et al.*, where

conventional kinetic model and its performance were compared with the performance of a neural network model and correlations between the different parameters were also mapped [18]. From the results, it was seen that by applying the neural network in parallel with the kinetic model, the required computational time could be reduced by 3 orders of magnitude without significantly affecting the prediction accuracy. Another excellent way to use the data and knowledge available in our existing chemical systems or to minimize the possibility of manufacturing defects is to use this information during the design of additional equipment. In a related article, Rojek *et al.* show how different neural network models can support a process engineer in performing design tasks [19]. Machine learning models can be used in many other applications and have been used with impressive results in many cases, for example in the petrochemical, oil and gas industries to estimate oil properties from pressure-volume-temperature data [20], or for example to optimize industrial hydrocracking units [21]. There are also applications to the description and comparison of fixed-bed reactors for ethylene oxide production with conventional models [22], and Balabin *et al.* have published results on feature selection methods for estimating biodiesel properties [23]. By the available literature it can be seen that these machine learning methods can be used in a very wide range of applications with good accuracy and lower computational requirements compared to traditional models, and with their use industrial processes can be improved and optimised to a higher extent.

Machine Learning regression models and soft-sensor approaches have become increasingly indispensable in modern chemical engineering. In many industrial processes, key quality or state variables are either expensive, slow, or technically difficult to measure directly; soft-sensors – data-driven regression models built from readily accessible sensor signals – bridge this gap by estimating those hard-to-measure variables in real-time [24]. By capturing nonlinearities, multicollinearity and dynamic process changes better than purely first-principles models, these ML-based tools enable enhanced monitoring, faster fault detection, tighter control and improved yield or quality in chemical-engineering applications [25]. Incorporating soft-sensors into the workflow allows process engineers to transform large streams of online sensor data into actionable estimates of internal states or product attributes, thus supporting real-time optimization, predictive maintenance and increased robustness in complex process environments [26].

The objective of Machine Learning models is to utilize existing data to facilitate a more profound understanding of chemical processes [27], address challenges associated with the inability to identify certain chemical steps and transformations, and discover novel materials

[28,29], as well as ascertain optimal reaction conditions in a continuous online manner [30,31]. The implementation of these machine learning approaches in the synthesis of MDA and, by extension, MDI, holds significant potential due to the intricate nature of the material system and the versatility of various analytical methods, particularly in the context of industrial-scale manufacturing processes where the simplicity and precision of application are paramount.

Understanding the technologies that drive innovation has become essential, not only an optional approach. Machine Learning stands as a key advancement in this transformative age. This section aims to clarify the idea of machine learning, providing a comprehensive guide. It is important to understand what machine learning is, become familiar with its various types, applications, and the tools utilized in the industry [26].

To achieve a comprehensive differentiation, it is essential to define Artificial Intelligence, Machine Learning, and Deep Learning (DL), as well as to highlight the interrelations among these models and modeling methodologies.

Artificial Intelligence includes the creation of software systems that exhibit intelligent behavior and simulate human cognitive processes via diverse algorithms. These algorithms emphasizing three core capabilities: learning, reasoning, and self-correction, to achieve optimal performance. Artificial Intelligence can relates to systems derived from machine learning techniques as well as those engineered through explicit programming methodologies.

Machine learning constitutes a subfield of artificial intelligence, employing algorithms that acquire knowledge from data to formulate predictions. These predictions may be derived through supervised learning, wherein algorithms learn patterns from existing data, or through unsupervised learning, wherein they identify general patterns within data. Machine learning models possess the capability to forecast numerical values utilizing historical data, classify events as either true or false, and group data points according to shared characteristics.

Deep Learning represents a specialized part within machine learning, focusing on algorithms fundamentally grounded in multi-layered artificial neural networks, which draw inspiration from the architecture of the human brain. In contrast to traditional machine learning algorithms, deep learning algorithms are generally have greater complexity and non-linearity. These algorithms have the capability to learn from huge volumes of data, therefore yielding highly precise outcomes. Exemplary applications of deep learning include language translation, image recognition, and the development of personalized medicines.

On the other hand, Data Science is an interdisciplinary field that combines statistics, computer science, and domain expertise to collect, process, analyze, and interpret large volumes of data to generate actionable insights and guide decision-making. Figure 2 presents a structural overview of the previously mentioned data models.

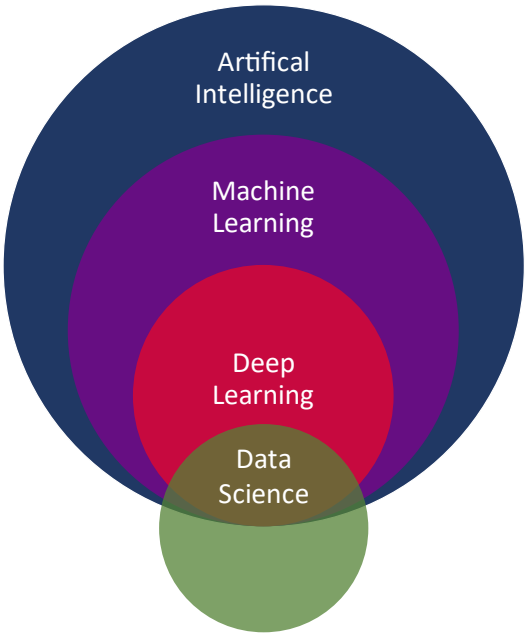


**Figure 2: Framework of AI, Machine Learning, Deep Learning, and Data Science**

Machine Learning, commonly referred to as ML, constitutes a specialized branch within artificial intelligence (AI), dedicated to the creation of computer algorithms that undergo enhancement autonomously through experiential learning and the utilization of data. Essentially, machine learning empowers computers to derive insights from data and to formulate decisions or forecasts independently, without explicit programming. Presently, across different disciplines, including industry, laboratories, research and development institutions or any other, a vast amount of data is being produced and is readily available. We are actively enhancing our capabilities in processing, comprehending, and utilizing this data for predictive and other scholarly objectives.

Fundamentally, machine learning applies to the development and implementation of algorithms that enable these decisions and predictions. These algorithms are engineered to enhance their performance over time, achieving greater accuracy and effectiveness as they assimilate more data. In conventional programming, a computer follows a set of predefined set of instructions to execute a task. In contrast, within the realm of machine learning, the computer is provided with a collection of examples, i.e. data and a specific task to achieve; it is then the

role of the computer or algorithm to determine the methodology to accomplish the task based on the provided examples.

Machine learning models demonstrate proficiency in estimating quantities of diverse contaminants or by-products within industrial systems, as well as parameters that present challenges for traditional chemical engineering calculations, such as the color of isocyanate products [32]. The implementation of such models enables the modeler to either support, reduce, or, when the model exhibits reliability and robustness, completely substitute laboratory measurements related to the parameter being estimated. The capability to evolve and enhance based on data makes machine learning extremely potent and adaptable. It fuels numerous technological innovations we observe today, such as voice assistants, recommendation engines, autonomous vehicles, and predictive analytics.

### 2.2.1 Types of Machine Learning

In this chapter, I summarize various Machine Learning applications in order to present different models, their importance and applicability, as well as their advantages and disadvantages. In the following paragraphs, the classification of these models, the steps of model development, and the mathematical formalizations of different models and related methods (e.g. extreme value search, feature selection, etc.) are introduced.

Machine learning can be systematically categorized into three different types based on the characteristics of the learning system and the nature of the data accessible: supervised learning, unsupervised learning. Figure 3 illustrates the classification of various machine learning types.

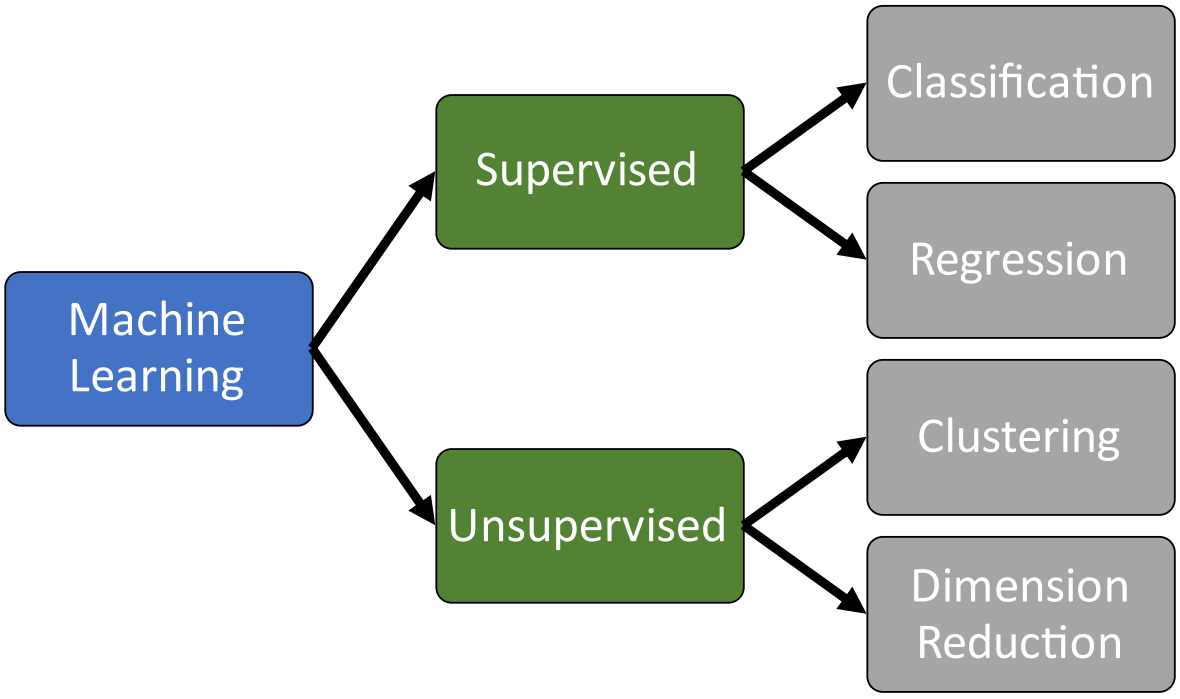


**Figure 3: Classification of different Machine Learning applications**

Supervised machine learning is a subfield of Artificial Intelligence that involves training a predictive model using labeled data. In this paradigm, the learning algorithm receives a dataset composed of input-output pairs, where each input instance is associated with a corresponding output label. The objective is to derive a function that accurately maps inputs to their corresponding outputs, thereby enhancing the model's ability to generalize to new, unseen datasets.

Supervised learning tasks are primarily categorized into classification and regression. In classification tasks, the output space generally consists of distinct categories or classes. In contrast, regression is concerned with continuous output variables, where the model predicts values that are real numbers. Conventional algorithms employed in supervised learning comprise decision or regression trees, support vector machines (SVMs), k-nearest neighbors (k-NN), and neural networks [33].

Model performance is typically evaluated using metrics such as accuracy, precision, recall, F1-score for classification, or mean squared error (MSE) and R-squared for regression. To avoid overfitting, which occurs when a model excels on training data but underperforms on new data, methods like cross-validation, regularization, and pruning are used. The success of supervised learning models is largely reliant on the quality and volume of labeled data, alongside the suitability of the selected algorithm and feature representation.

Unsupervised machine learning refers to a set of approaches within machine learning aimed at identifying hidden structures or patterns in data without any explicit labels. Unlike supervised learning, the data in unsupervised learning consists solely of feature vectors, without accompanying output labels. The primary goal is to discover the natural organization of the data and to describe its underlying distribution or intrinsic characteristics [34].

The two primary categories of unsupervised learning tasks comprise clustering and dimensionality reduction. Clustering entails the aggregation of analogue data points predicated on a concept of similarity or distance, with widely recognized algorithms encompassing k-means and hierarchical clustering. These techniques are designed to partition the data into discrete groups or clusters with the objective of maximizing similarity within clusters while minimizing similarity between clusters.

Dimensionality reduction techniques, including Principal Component Analysis (PCA) and t-Distributed Stochastic Neighbor Embedding (t-SNE), are employed to decrease the number of input features whilst retaining crucial structural information. These techniques are notably advantageous for visualization, noise reduction, and enhancing the efficiency of subsequent analytical tasks [35].

Because unsupervised learning does not utilize output labels, assessing the performance of models becomes more difficult. To evaluate their performance, it is recommended to use metrics like the Silhouette score, Davies–Bouldin index, or reconstruction error, chosen specifically for the task at hand. Interpreting results and validating the importance of discovered patterns heavily relies on domain expertise. Unsupervised learning is especially beneficial for exploratory data analysis, detecting anomalies, segmenting customers, and learning features. This approach is frequently used when labeled data is scarce or expensive to obtain, rendering it an efficient technique for discovering insights and forming data-driven hypotheses.

### 2.2.2 General steps of Machine Learning Model Development

Various process models exist for industrial purposes within Knowledge Discovery and Data Mining (KDDM), yet the Cross-Industry Standard Process for Data Mining (CRISP-DM) is generally utilised as the most widely adopted model. This model has been developed based on experiences and notable cases encountered in the realm of industrial data mining [36]. Despite its popularity, Studer *et al.* identified two significant limitations [37]. First, CRISP-DM is primarily centered on data mining without considering the evolving nature of the environment, which necessitates that ML models undergo continuous monitoring and updates

post-deployment to avert a decline in performance. Second, CRISP-DM lacks a comprehensive quality assurance approach, which is crucial for early error detection to reduce costs at later stages. Studer *et al.* proposed a process model, namely CRISP-ML(Q) that adheres to the core principles of CRISP-DM—maintaining neutrality across industries and applications— while addressing the specific requirements of ML applications, and presents an approach to monitoring, maintenance and quality assurance that is recognized as industry leading practice.

Understanding the operations of machine learning in order to support industrial processes necessitates participation in a systematic process that transforms raw data into insightful and explainable information. Each phase in this continous process is illustrated in Figure 4 below:

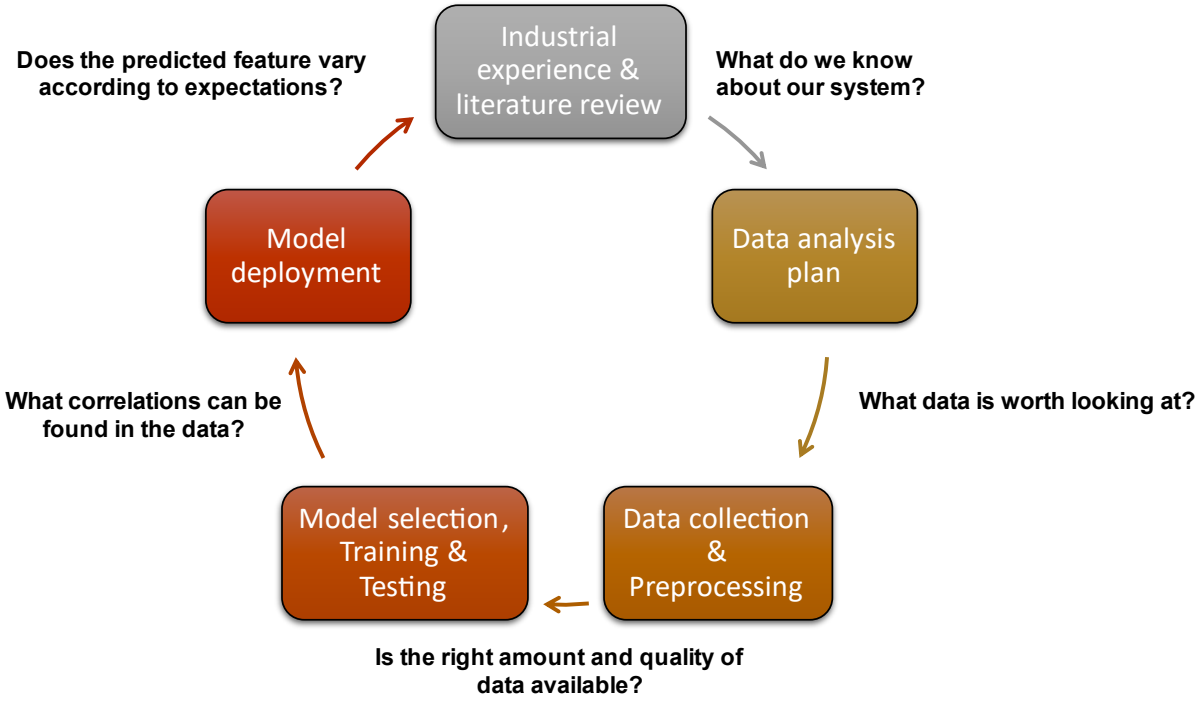


**Figure 4: Steps of continous model-based analysis**

The beginning stage of the machine learning workflow involves grasping the essence of the process, be it a laboratory experiment, a pilot test, or a functioning industrial plant. It is crucial to comprehend the problem that needs addressing via machine learning. Leveraging reliable literature and real-world industrial experience, it is possible to establish preliminary assumptions and insights into the process to enhance problem comprehension.

The next phase of the machine learning process involves the collection of data. Data serves as a fundamental component in machine learning, wherein both its quality and quantity have a direct impact on the performance of the model. Data can be acquired from a diverse range of sources, including industrial operational data, databases, text files, images, audio files,

or even extracted from the web. Subsequent to its collection, the data must be appropriately prepared for machine learning. This preparation entails organizing the data into a suitable format and ensuring its relevance to the specific problem being addressed.

Data preprocessing is a key step within the machine learning process. The cleaning of data is the first step, which includes the removal of duplicates and the removal of errors, the management of missing data, whether through elimination or filling, and the normalization of data, which involves scaling it to a standard format. Preprocessing enhances the quality of the dataset and facilitates the accurate interpretation by the machine learning model.

Following the completion of data preparation, the next step is to choose a suitable machine learning model. A variety of models are available, such as linear regression, decision trees, neural networks and others [38]. The selection of a model is contingent upon the data characteristics and the specific problem to be addressed. Considerations in model selection include the scale and category of the data, the complexity of the task and the computational resources that are accessible.

After choosing a suitable machine learning model, the next step involves training it using the pre-processed data. Training necessitates inputting the data into the model, thereby enabling it to optimize its internal parameters to enhance output prediction accuracy. During the training process, it is imperative to mitigate the risks of overfitting—where the model exhibits high performance on the training dataset but fails to generalize to new datasets—and underfitting, whereby the model underperforms on the training dataset as well as new datasets.

To reduce the risk of overfitting, K-fold cross-validation is commonly used. This technique involves the resampling of data and dividing it into distinct subsets for testing and training the model over multiple cycles. It is often used in predictive scenarios to assess how well a model is likely to perform in real-world applications. Additionally, it can be utilized to analyze the quality of a model fit and to check the consistency of its parameters [39].

Following the training phase of a model, it is crucial to assess its performance on unseen data prior to deployment. The result evaluation and monitoring of machine learning models includes continuous assessment to identify changes known as model drift, which occurs when a model's performance deteriorates due to shifts in data patterns, thereby a sustainable model prediction quality can be achieved over time. Ongoing monitoring and retraining protocols assist organizations in ensuring that their models remain efficient and dependable within production environments. Metrics commonly employed for evaluating a model's performance

include accuracy (for classification tasks), precision and recall (relevant for binary classification problems), and metrics like mean squared error (applicable to regression problems).

In addition to optimizing for accuracy, hyperparameter optimization for machine learning incorporates tools for automated hyperparameter searches, thereby ensuring both efficiency and reproducibility. A variety of algorithms, such as Bayesian optimization [40], Grid Search [41], Genetic Algorithms (GA) [42], and Particle Swarm Optimization (PSO) [43], can be employed for hyperparameter tuning wherein diverse combinations of hyperparameters are evaluated. These methods guarantee that experiments are reproducible and well-documented, allowing for consistent optimization throughout time.

The deployment of machine learning models necessitates their integration into production environments, thereby enabling the delivery of real-time predictions or insights. In industrial environments, machine learning models are employed to forecast various qualitative or quantitative data derived from operational parameters. The objective of implementing and deploying these models is to assist operators and engineers in an explainable manner, thereby ensuring that the models are reliable, robust and reproducible [44].

### 2.2.3 Formalization of Machine Learning models

For machine learning models, the measured target variables can generally be expressed as follows in Equation 1, accounting for measurement errors $\varepsilon_k$.

$$y_k = f(\boldsymbol{x}_{k-\boldsymbol{d}}, \boldsymbol{\theta}) + \varepsilon_k, k = 1, \dots, N \quad (1)$$

where f denotes the machine learning models, $y_k$ is the k-th observed value of the target variable $\boldsymbol{y}$, and $\boldsymbol{x}_k$ is the k-th sample vector drawn from a matrix $\boldsymbol{X}$. The matrix $\boldsymbol{X}$ contains the independent variables and has dimension $N \times N_x$, where $N_x$ is the number of independent variables. The vector $\boldsymbol{d}$, with dimensions $1 \times N_x$, represents the time delays associated with the input variables, and $\theta$ is the paramter matrix of the machine learning models. The $\varepsilon_k$ term corresponds to the measurement error for the $k$-th observation.

Model predictions, $\hat{y}_k$ can be written in general as in Equation 2:

$$\hat{y}_k = f(\boldsymbol{x}_{k-\boldsymbol{d}}, \boldsymbol{\theta}) \quad (2)$$

Where $\hat{y}_k$ is the model prediction for the output value.

Equations 3 to 6 provide the formal definitions of the metrics earlier referenced for evaluating the performance of regression models:

$$R^2 = 1 - \frac{\sum_{k=1}^{N}(y_k - \hat{y}_k)^2}{\sum_{k=1}^{N}(y_k - \bar{y}_k)^2} \quad (3)$$

$$MSE = \frac{1}{N}\sum_{k=1}^{N}(y_k - \hat{y}_k)^2 \quad (4)$$

$$RMSE = \sqrt{\frac{1}{N}\sum_{k=1}^{N}(y_k - \hat{y}_k)^2} \quad (5)$$

$$MAE = \frac{1}{N}\sum_{k=1}^{N}|y_k - \hat{y}_k| \quad (6)$$

Where $y_k$ is the k-th observed output value, $\hat{y}_k$ is the k-th calculated output value, $\bar{y}$ is the mean of all observed output values and N is the total number of data points.

Throughout our research, we created models designed to predict target variables, particularly concerning the formation of MDA and MDI. We utilized and assessed the effectiveness of five separate model types for various tasks:

- linear regression (LR)
- regression tree (RT)
- neural networks (NN)
- support vector machine (SVM)
- gaussian process regression (GPR)

Figure 5 demonstrates the balance between interpretability and predictive accuracy across different machine learning models. Basic models, including linear regression and decision trees, provide good interpretability but might fall short in predictive strength for complex tasks. Conversely, models such as random forests, gradient boosting machines, and deep neural networks typically deliver better accuracy but with diminished interpretability. Recognizing this trade-off is essential when choosing models for situations where transparency, explainability, or regulatory compliance is important.

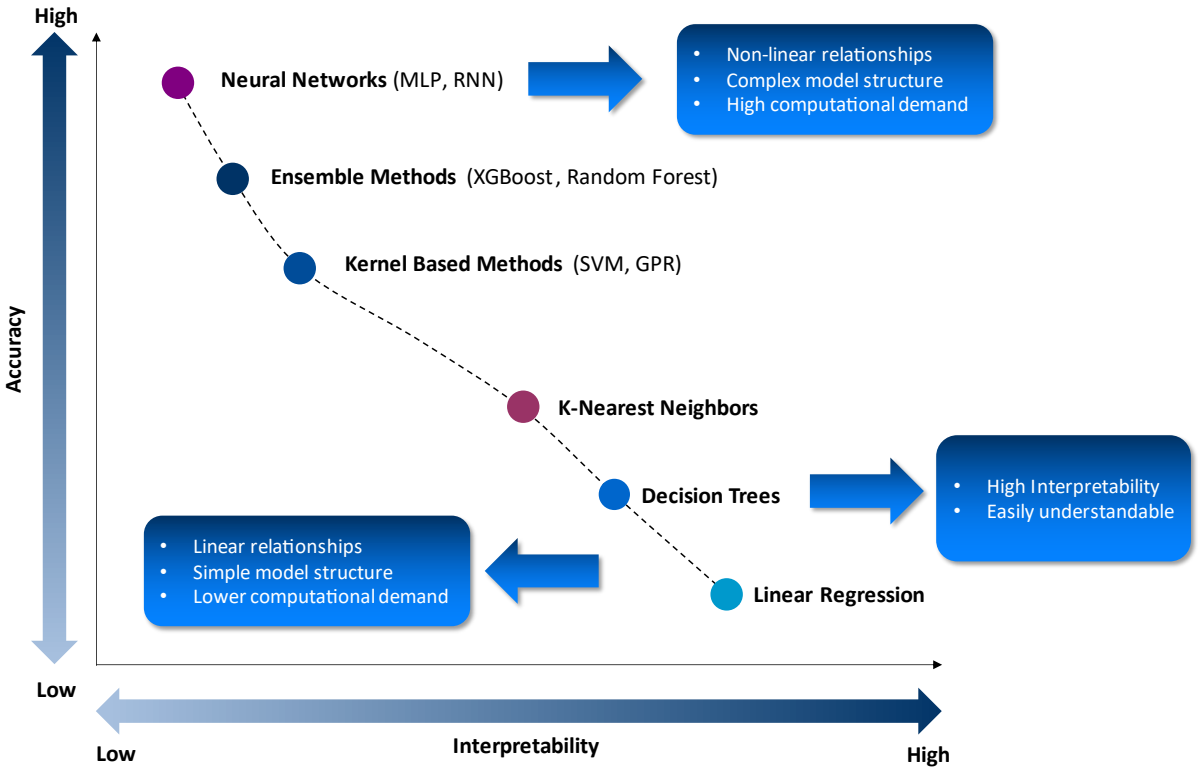


**Figure 5: Comparison of accuracy and interpretability of different Machine Learning models [45]**

### 2.2.3.1 Linear Regression

Linear regression is a foundational statistical modeling technique used to describe the linear relationship between a dependent variable and one or more independent variables [46]. Linear regression operates on the assumption that the dependent variable can be expressed as a linear combination of the independent variables, accompanied by a random error term. Due to its simplicity, interpretability, and computational efficiency, linear regression is widely applied to model and characterize systems in which the intrinsic relationships are assumed to be approximately linear. In case of linear regression, the model prediction can be calculated as the following Equation 7:

$$\hat{y}_k = \beta_0 + \sum_{i=1}^{n} x_{k,i}\,\beta_i + \varepsilon_k \qquad (7)$$

Where $\beta_0$ is the intercept term, $x_{k,i}$ is the value of the i-th feature for the k-th observation $\beta_i$ is the i-th regression coefficient and n is the number of features (excluding the intercept term).

The model calculates coefficients through the minimization of the sum of squared residuals, whereby residuals denote the discrepancies between observed and predicted values. These residuals function as indices of the model's accuracy – larger residuals indicate a greater

deviation from the observed data, thereby yielding an elevated value of the objective (loss) function. In contrast, a model that accurately aligns with the data will result in smaller residuals and, therefore, a reduced value for the objective function. Despite its utility, linear regression is inherently limited in its capacity to model non-linear relationships. When the true relationship between variables is non-linear, more advanced techniques such as polynomial regression, kernel methods, or non-linear machine learning models may be required to achieve adequate predictive performance. The general illustration of Linear Regression can be seen in Figure 6.

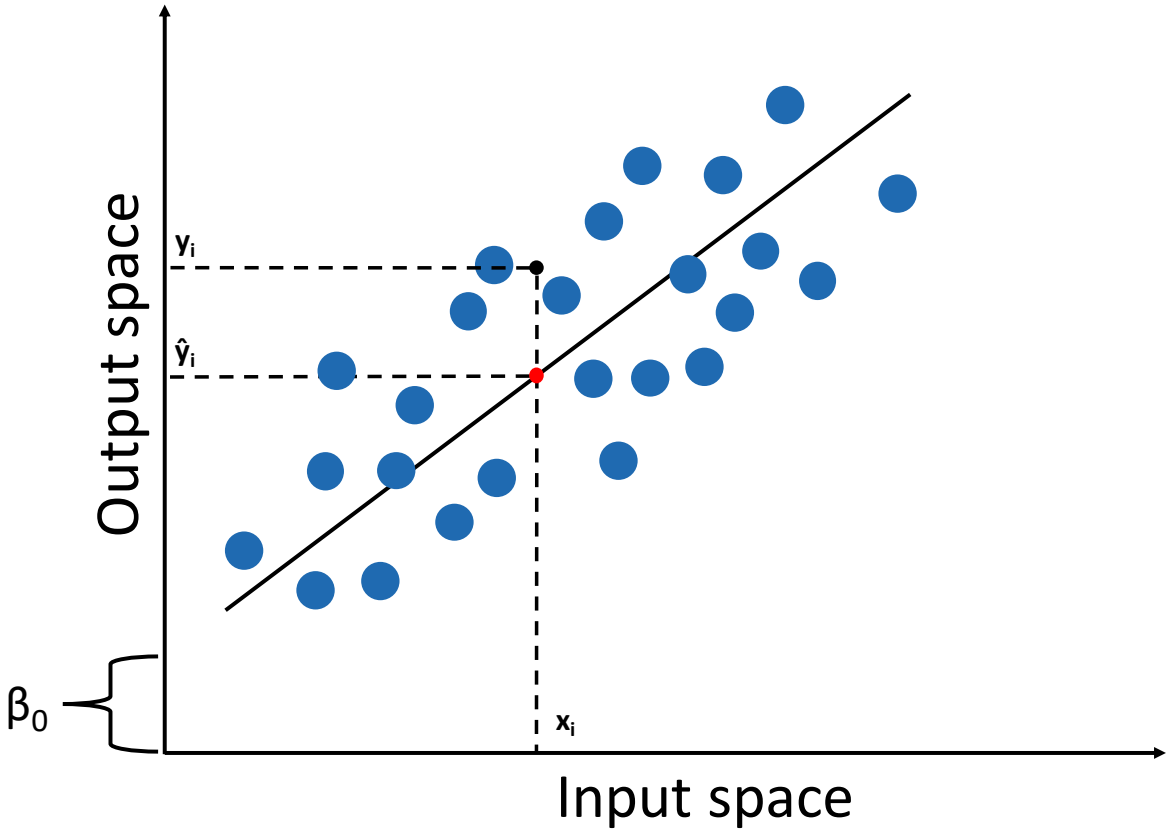


**Figure 6: General illustration of Linear Regression**

### 2.2.3.2 Stepwise Linear Regression

Stepwise linear regression (SLR) is an iterative predictor-selection method for linear regression, which constructs a model by repeatedly adding and/or removing variables from a pool of candidate explanatory variables according to a predefined criterion. Instead of incorporating all predictors simultaneously, this method progressively adds or removes predictor variables or features based on statistical criteria, with the goal of identifying a smaller set of variables that delivers effective predictive performance.

In Machine Learning, SLR is frequently employed for feature selection tasks as well: by choosing the most important input features, it is possible to decrease model complexity,

enhance interpretability, and decrease the risk of overfitting [47]. For instance, the model can start the training with no predictors and use forward selection, incorporating the features that can increase a specific criterion (e.g., AIC, BIC, $R^2$ or adjusted $R^2$) at every step; alternatively, the training of SLR can start with all predictors and employ backward elimination, discarding the least significant variable at each step [48].

Stepwise Linear Regression, though simple and easy to use, has considerable limitations. Firstly, it can produce suboptimal models, as it follows a greedy approach and lacks assurance of reaching global optimality, and it may become unstable when predictor variables are highly correlated [49]. Secondly, it is prone for overfitting if not used with caution, which means a model trained through SLR might perform well on training data but could possibly fail to predict effectively based on newer set of test data.

Considering a multiple linear regression model according to Equation 7, stepwise linear regression evaluates possible model candidates using different fit criteria, for example Bayesian Information Criterion (BIC) according to Equation 8:

$$\mathrm{BIC} = \mathrm{N} \cdot \ln\left(\frac{RSS}{N}\right) + p \cdot \ln(N) \tag{8}$$

Where $N$ is the number of observations, $p$ is the number of estimated parameters in the model and $RSS$ is the residual sum of squares, which can be calculated according to Equation 9:

$$RSS = \sum (y_k - \hat{y}_k)^2 \tag{9}$$

If we analyze the two terms of the Equation 8, the term $\mathrm{N} \cdot \ln\left(\frac{RSS}{N}\right)$ can be interpreted as a 'fit term', where smaller RSS leads to a lower BIC, while term $p \cdot \ln(N)$ represents a penalty term, which aims to penalize having too many predictors. Stepwise linear regression chooses the model with the smallest BIC, adding a predictor only when the improvement in fit is large enough to offset the added complexity penalty. BIC is a Bayesian-based approximation that is asymptotically consistent – which means when the true model is included among the candidates, BIC tends to select it as $\mathrm{N} \to \infty$. In contrast to AIC, BIC applies a heavier penalty for model complexity, and therefore more often prefers simpler models.

#### 2.2.3.3 Regression tree

A regression tree functions as a non-linear machine learning model designed to predict continuous outcomes based on the values of input variables. It represents a type of decision tree algorithm, where the objective is to predict a continuous variable rather than a categorical one.

This model partitions the input space into regions that achieve maximal uniformity with respect to the target variable. Decisions at each node depend on a splitting criterion, which is generally intended to minimize variance or mean squared error within each data segment. The terminal nodes, known as leaf nodes, offer the predicted values, commonly denoted by the mean value of the target variable within the respective region.

Regression tree performs a recursive binary splitting of data, which can be formulated as Equation 10 for each split:

$$\begin{aligned} \mathrm{R}_1(i,s) &= \{x \mid x_i \leq s\} \\ \mathrm{R}_2(i,s) &= \{x \mid x_i > s\} \end{aligned} \tag{10}$$

Where $R_1$ and $R_2$ are the two regions after splitting, feature i and split point s is selected to minmimze the sum of squared errors (SSE) according to Equation 11:

$$\min_{i,s} \left[ \sum_{x_i \in R_1(i,s)} \left(y_i - \bar{y}_{R_1}\right)^2 + \sum_{x_i \in R_2(i,s)} \left(y_i - \bar{y}_{R_2}\right)^2 \right] \tag{11}$$

Where $\bar{y}_{R_1}$ and $\bar{y}_{R_2}$ are the average y values in the respective regions. The partitioning of regions will proceed until one of the termination criteria is met, which may include the minimum quantity of samples per leaf, the maximum tree depth, or a minimum enhancement in model performance or in SSE. The computation of model output is facilitated by Equation 12, which necessitates navigating through the tree according to the partitioning rules until a terminal leaf is encountered. Subsequently, the computed mean for the specified region serves as the model prediction:

$$\hat{\mathrm{y}} = \frac{1}{\mathrm{n}} \sum_{i=1}^{n} y_i \tag{12}$$

Where $\hat{y}$ is the predicted output for the input x that falls into this leaf node, n is the number of training samples in the region and $y_i$ is the actual target value of the i-th training sample in the node.

One of the key benefits of regression trees is their capability to model non-linear connections between predictors and the response variable. In contrast to linear regression, regression trees do not presume any specific functional relationship between the independent and dependent variables, enabling them to identify intricate interactions and non-linear trends. Furthermore, regression trees are understandable and can be illustrated graphically, providing clarity into the decision-making process.

However, single regression trees are prone to overfitting and may exhibit high variance. To address this, ensemble methods such as random forests [50] and gradient boosting machines [51] are frequently employed, wherein multiple trees are combined to improve predictive accuracy and generalization.

Regression trees are frequently employed within industrial applications due to their inherent simplicity and robust predictive capabilities, primarily in the development of soft sensors and system control related tasks [52,53]. Additionally, they are particularly effective for estimating a wide range of properties, such as mechanical features [54]. The general illustration of Regression Trees can be seen in Figure 7.

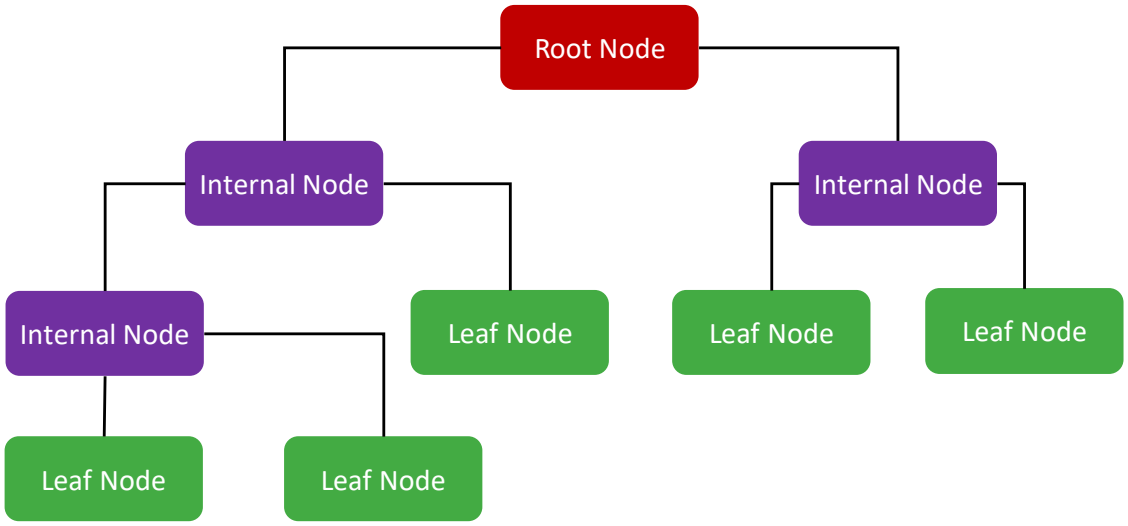


**Figure 7: General illustration of Regression Tree structure**

### 2.2.3.4 Neural networks

Artificial Neural Networks or ANNs are computational frameworks that draw inspiration from the architecture and functionality of biological neural systems. In regression contexts, these networks serve to model intricate and frequently nonlinear relationships between input variables and their corresponding continuous output values. In contrast to linear regression approaches, neural networks excel at detecting complex patterns within high-dimensional datasets by employing multilayered, nonlinear transformations. The general structure of a multi-layer feed forward neural network is shown in Figure 8.

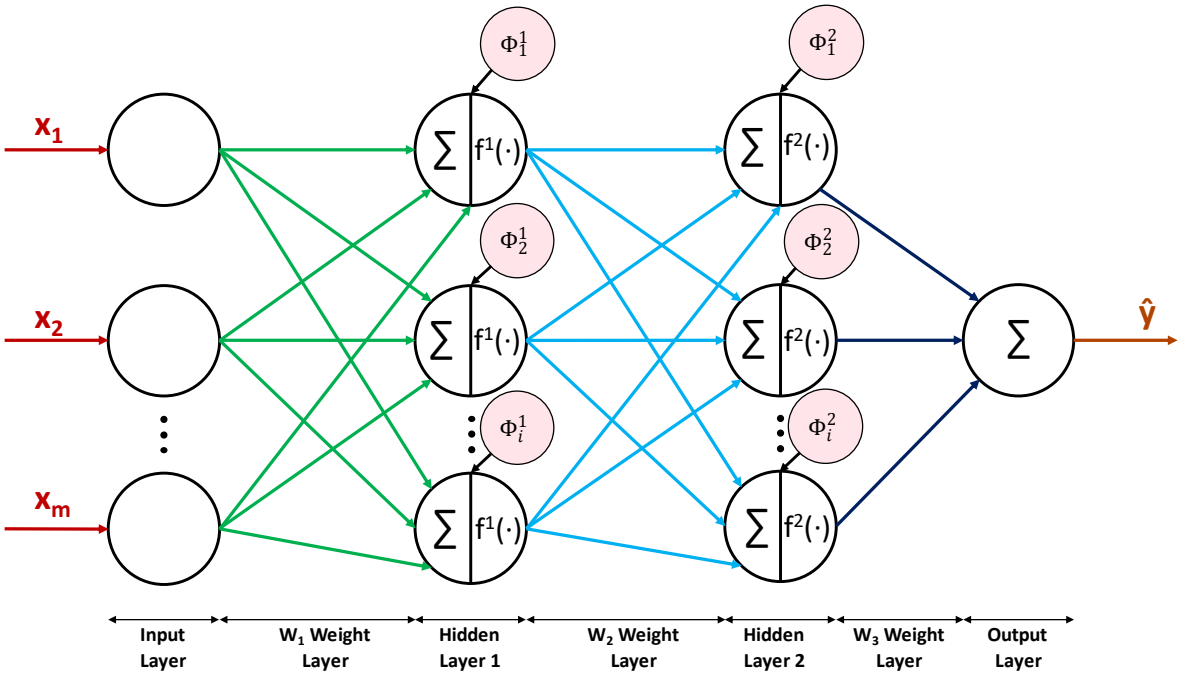


**Figure 8: General structure of feed forward neural networks**

A typical feedforward neural network for regression consists of an input layer, one or more hidden layers, and an output layer. Each layer is composed of artificial neurons, where each neuron computes a weighted sum of its inputs followed by a nonlinear activation function. The neural network nodes combine inputs from raw data with different coefficients or weights, amplifying or dampening the effect of the input variable, thereby assigning real meaning or effect to the input variables in terms of the output variable or variables to be determined. The progressing of data with a simple structured neural network consists of the following three steps:

- Step 1: weighting;
- Step 2: adding a constant (bias);

- Step 3: use of activation function.

If we consider neuron *m* in layer *l* in general, the equation of the output of neuron *m* can be defined as shown in Equation 13:

$$y_m^l = f^l\left(\sum_{j=1}^{n^{l-1}} w_{mj}^l a_j^{l-1} + \phi_m^l\right) \tag{13}$$

Where $y_m^l$ is the output of neuron *m* in layer *l*, which could be denoted as $a_j^l$ as well, $f^l$ is the activation function, $w_{mj}^l$ is the weight from input neuron *j* in layer *l-1* to to neuron *m*, $a_j^{l-1}$ is output or activation from neuron *j* in the previous layer *l-1*, $\Phi_m^l$ is bias term for neuron *m* in layer *l* and $n^{l-1}$ is the number of input features or neurons in layer *l-1*.

Artificial Neural Networks are able to recognize and model the nonlinear and complex relationships between input and output data [55]. However, they also have significant drawbacks, such as high computational requirements and a tendency to overfit, which necessitates careful monitoring.

#### 2.2.3.5 Support Vector Machine

Support Vector Machines (SVMs), originally designed for classification tasks, have been successfully modified for regression uses under the method known as Support Vector Regression (SVR). SVR is a powerful supervised learning approach aimed at constructing a predictive model by identifying a function that closely matches the connection between input features and a continuous output variable, with an emphasis on reducing model complexity. The general illustration of Support Vector Regression can be seen in Figure 9.

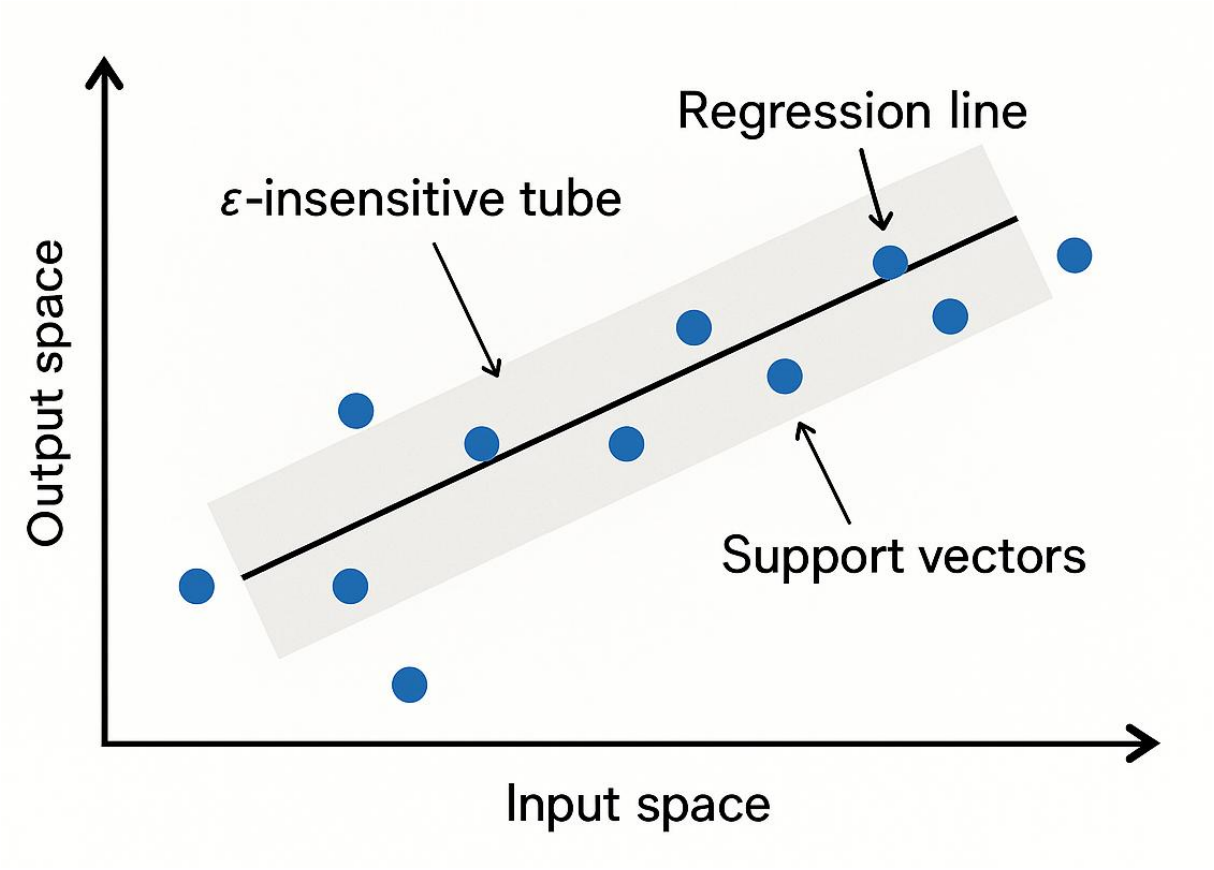


**Figure 9: Illustration of Support Vector Regression**

The objective of SVR is to determine a function f(x) that deviates from the true target values y by no more than a specified margin ε while maintaining maximal flatness. It is generally assumed that the function f(x) behaves as outlined in Equation 14 [56]:

$$f(x) = \langle \boldsymbol{w}, \boldsymbol{x} \rangle + \Phi \tag{14}$$

where w is the weight vector, Φ is the bias term, and $\langle \cdot , \cdot \rangle$ denotes the dot product. To ensure robustness and generalization, SVR optimizes the following convex optimization problem according to Equation 15-18:

$$\min_{w, \Phi, \xi, \xi^*} \frac{1}{2}\|w\|^2 + \lambda \sum_{i=1}^{N} (\xi_i + \xi_i^*) \tag{15}$$

Subject to:

$$y_i - \langle w, x_i \rangle - \Phi \leq \epsilon + \xi_i \tag{16}$$

$$\langle w, x_i \rangle + \Phi - y_i \leq \epsilon + \xi_i^* \tag{17}$$

$$\xi_i, \xi_i^* \geq 0 \tag{18}$$

where $\xi, \xi_i^*$ are slack variables that facilitate adaptability and permit the acceptance of deviations within the ± ε boundary, while λ serves as a regularization parameter mediating the balance between model complexity and the tolerance of errors exceeding ε. A key feature of SVR is the use of kernel functions, enabling the algorithm to compute inner products in the high-dimensional feature space effectively, without needing an explicit transformation. Widely utilized kernel types include linear, polynomial, and Gaussian, also known as Radial Basis Function (RBF) kernels.

SVR is a very common and frequently used machine learning tool that can be used for property prediction [57], biological sludge reactor operation, management and model prediction tasks [58], or for predictive modeling of biomass gasification [59].

#### 2.2.3.6 Gaussian Process Regression

Gaussian Process Regression (GPR) represents a robust and non-parametric probabilistic approach to regression tasks. It relies on Gaussian processes (GPs), which extend multivariate normal distributions. The main aim of GPR is to forecast a continuous target variable, while also quantifying prediction uncertainty. This capability is particularly advantageous for scenarios demanding accurate predictions along with a measure of confidence in the model's results. The illustration of Gaussian Process Regression can be seen in Figure 10.

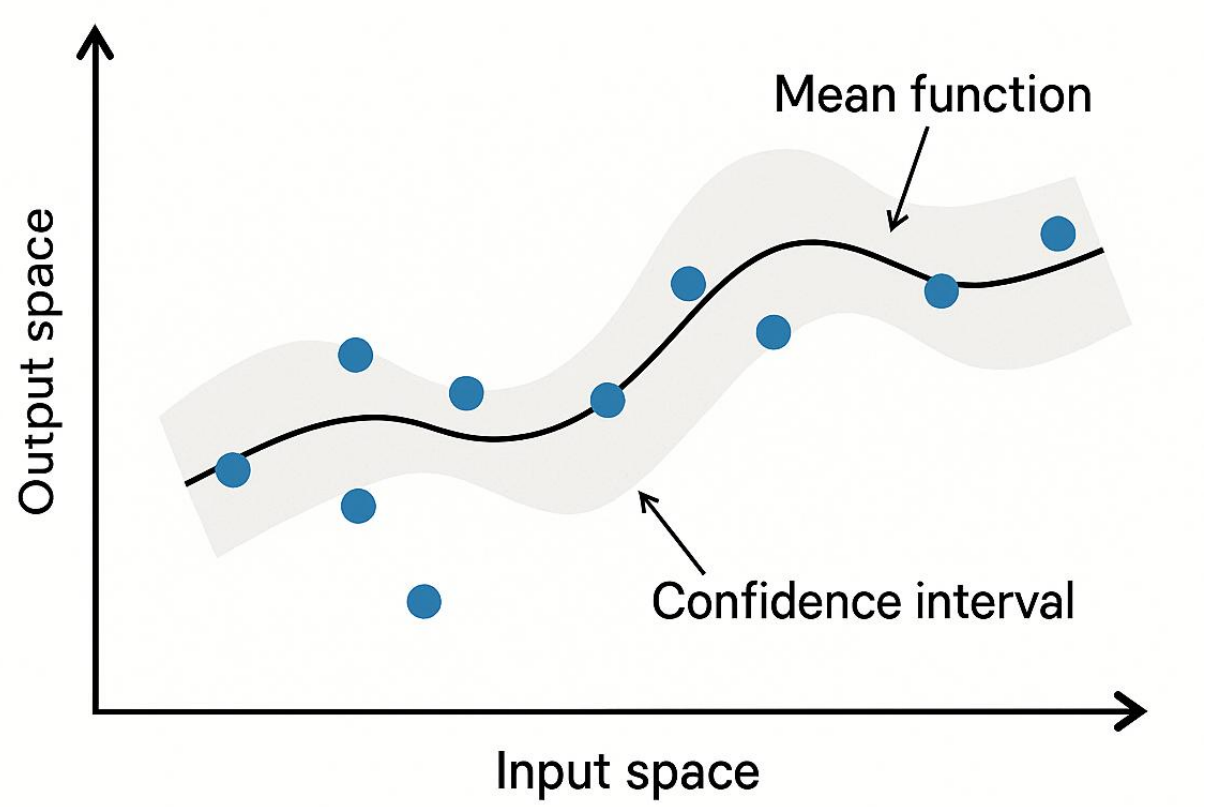


**Figure 10: Illustration of Gaussian Process Regression**

Gaussian process regression (GPR) models represent a category of probabilistic learning models suitable for a range of regression applications. Fundamentally, a Gaussian process represents a random process in which the probability distribution across possible functions is predicted. In mathematical terms, the GPR model asserts that the function *f*, which connects the inputs and outputs, is derived from a Gaussian process described by the mean function $\mu$ and the covariance function *k*. The distribution of *f* at test points $\boldsymbol{X}^*$ is represented as Equation 19:

$$f(x^*) \sim GP((\mu(\boldsymbol{X}), k(\boldsymbol{X}, \boldsymbol{X}^*)) \tag{19}$$

In Gaussian Process Regression, it is assumed that the observed data are generated by function f(x), then affected by Gaussian noise as represented in Equation 20:

$$y_i = f(x_k) + \varepsilon_k \quad \varepsilon_k \sim N(0, \sigma_n^2) \tag{20}$$

where $\varepsilon_k$ represents the error, which is presumed to be independent and identically distributed with a mean of zero and a variance denoted as $\sigma_n^2$ and *N* refers to normal distribution.

The mean function and covariance function are typically defined using kernel functions using the following Equation 21.

$$k(x_i, x_j) = \Phi(x_i)^T \Phi(x_j) \quad (21)$$

Where $\Phi(\cdot)$ is the feature mapping function. For a given training data set, the GPR regression will predict the distribution of f what is the most likely generated the training data. This involves calculating the distribution of „f" for the given the data, defined as Equation 22:

$$p(f|x, y) = \frac{p(y|x, f) \cdot p(f)}{p(y|x)} \quad (22)$$

where $p(f|x, y)$denotes the posterior distribution, which represents the new belief about the function f subsequent to observing the data x and y,

$p(y|x, f)$ is the likelihood of the data, indicating how likely the observed data y is, given the inputs x and the function values f,

$p(f)$ is the prior distribution which represents our initial assumptions regarding the function f prior to observing the data, and

$p(y|x)$ is the model evidence, also known as the marginal likelihood, which serves as the normalization constant.

A principal advantage of GPR is its capability to quantify uncertainty in predictions. In contrast to numerous other regression models, GPR yields not only point estimates but also confidence intervals, which express the uncertainty linked to these predictions. This attribute holds particular value in disciplines such as scientific research and engineering, where comprehending the degree of confidence in a model's output may be as critical as the prediction itself. An additional notable benefit is the adaptability of the model. Being a non-parametric method, GPR can capture highly nonlinear patterns without needing a specified functional form. The kernel function plays a crucial role in GPR by integrating domain-specific prior knowledge, which may include assumptions regarding smoothness, periodicity, or other intrinsic characteristics of the function. This adaptability makes GPR particularly suitable for situations where the actual underlying the function is either not known or complex.

Although GPR has many advantages, it also comes with significant challenges. A key limitation is its scalability, as the computational complexity renders it less appropriate for very large datasets. This becomes a critical constraint when handling large-scale problems or high-dimensional input spaces. In such scenarios, it is frequently necessary to utilize approximations or sparse versions of GPR, which may potentially compromise certain advantages associated with accuracy and the quantification of uncertainty.

An additional challenge lies in the selection of the kernel function and its associated hyperparameters. Although the adaptability of the kernel constitutes a significant advantage of GPR, it requires careful adjustment. An inappropriate choice of kernel or hyperparameters may result in suboptimal performance, manifesting as overfitting or underfitting, thereby decreasing the model's predictive precision. Furthermore, the hyperparameter optimization process can be computationally intensive, often necessitating comprehensive cross-validation or alternative optimization methodologies.

#### 2.2.3.7 LASSO – method

The abbreviation "LASSO" stands for Least Absolute Shrinkage and Selection Operator, a statistical mathematical model for parameterizing data models and selecting parameters that characterize the system [60]. The method was used for the input data to detect whether collinearity is present between the input parameters, as the parameter shrinkage can detect such collinearity.

LASSO – method returns fitted least-squares regression coefficients for linear models of the independent data X and the response y. Each model corresponds to the regularization coefficient λ which is highlighted in Equation 23. With the changing of the λ parameter, the shrinkage of input parameters can be modified and controlled, furthermore, the LASSO – method is able to identify redundant input variables and with the shrinkage of these parameters, their order of importance also can be identified.

The LASSO – method is a relatively high-precision modeling technique, a type of linear regression that uses shrinkage and parameter selection, i.e. it varies the parameters of the model by shrinking the regression coefficients, reducing some coefficients to zero, thereby reducing variance and minimizing bias. The LASSO – method generally performs well even when relatively few observations and many characteristics are available. The accuracy of the solution is strongly dependent on the so-called regularisation parameter (λ), which is in fact a factor controlling the degree of shrinkage and the number and value of the coefficients. The LASSO – method performs the extremum search task according to the objective function in Equation 23, also called L1 regularisation.

$$\min_{\beta_j} \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y_i})^2 + \lambda \sum_{j=1}^{p}|\hat{\beta}_j| \tag{23}$$

Where $\beta_j$ is the regression coefficient associated with feature j and λ is the regularization parameter. By varying the parameter λ, the second term of Equation 23, the penalty factor can

be varied, and by choosing the appropriate value of λ, the solution of the model can be optimized. The λ, as mentioned earlier, gives the degree of shrinkage, which can be observed as follows:

- if λ = 0, no parameter is eliminated. The solution of the LASSO – method will thus be equivalent to the solution of a linear regression;
- as λ is increased, more and more variables are zeroed and eliminated (in theory, all parameters are zeroed when λ = ∞;
- as λ is increased, the bias increases.

The typical selection phase comes after the shrinkage, where all non-zero values are selected for further use in the model. The LASSO – method is a special type of regression that is well suited for models that exhibit a high degree of multicollinearity or when we want to automate certain parts of the model selection, such as variable selection/parameter elimination. Multicollinearity refers to a strong linear dependence between more than two explanatory variables in a multiple regression, in other words, multicollinearity refers to a high correlation between more than two explanatory variables.

Alongside the LASSO – method, I have also employed linear regression, stepwise linear regression, and neural network models, with their outcomes also presented.

## 2.3 Mathematical methods to improve the performance and explainability of Machine Learning models

### 2.3.1 Feature selection methods

In numerous contemporary machine learning tasks, determining the most representative features (i.e., inputs or attributes) of the observed data – known as feature or variable selection – is essential for reducing modeling errors and lowering the resource demands of the modeling processes. If we consider a given input data with *n* number of features and one or more independent variables, then the feature selection task can be described as to find from the n-dimensional feature space a subspace, where the subspace can optimally characterize the independent value. Optimal characterization can be formulated in general terms, but – as mentioned earlier – it is often defined as the minimization of resource demands and complexity together, along with the maximization of performance of machine learning models. In practice, all of the feature selection methods presented function as filter methods, meaning they are used before training the machine learning models and do not depend on the specific training

algorithm chosen. In this chapter, I provide a general introduction to the feature selection methods employed throughout my thesis work in Table 3.

**Table 3: Comparison of different feature selection methods**

| Method / Criterion | F-test | MRMR | RReliefF | Correlation-based |
|---|---|---|---|---|
| **Working principle** | Tests linear dependency between each feature and the target using F-statistics | Maximizes relevance to target while minimizing redundancy between selected features | Estimates importance based on local differences and target variation | Measures linear correlation between feature and target |
| **Statistical methods** | Linear regression & hypothesis testing (F-distribution) | Information theory (mutual information) | Instance-based distance estimation | Correlation |
| **Computational complexity** | Low | Medium | High | Low |
| **Data requirement** | Low | Low | High | Low |
| **Feature interaction identification** | No | Yes | Yes | No |

Minimum Redundancy Maximum Relevance (MRMR) is a popular feature selection technique that aims to find a small set of features that are strongly related to the target variable while being minimally redundant with one another. The fundamental principle of MRMR is to synchronize two opposing goals: increasing the statistical dependency between the selected features and the predicted parameter (maximum relevance), and decreasing the statistical dependency among the chosen features themselves (minimum redundancy) [61]. In most cases,

relevance and redundancy are measured using mutual information. The algorithm then adds features step by step by optimizing a criterion that blends these two components, either as a difference (relevance minus redundancy) or as a ratio (relevance divided by redundancy). This greedy selection strategy makes MRMR computationally efficient and well-suited for high-dimensional data [62]. MRMR is widely used in areas like bioinformatics [63], text classification [64], image and signal processing [65] and others, where datasets typically include many highly correlated features. By cutting down on feature redundancy while maintaining prediction capability, MRMR can enhance model generalization, limit overfitting and lower computational costs in many different applications.

F-test feature selection is an univariate statistical method – meaning it only analyzes a single variable at a time to describe its characteristics – used to evaluate how relevant each individual feature is in relation to a predicted variable [66]. F-test is widely used in supervised learning problems – especially in classification and regression – to reduce the number of features and enhance model interpretability and computational efficiency [67]. The technique relies on analysis of variance (ANOVA) concepts and evaluates whether the mean values of a feature differ significantly across target classes (for classification) or show a linear relationship with the predicted variable (for regression) [68]. F-test feature selection identifies important features by quantifying how much of the variability of the predicted variable can be explained by each feature through an univariate linear model. For each feature, it compares the explained variance by the fitted regression model to the residual variance not explained by the model. Features with larger F-statistics are interpreted as having stronger linear explanatory power with respect to the target variable. F-test is widely used in high-dimensional settings such cancer prediction [69], heart disease prediction [70], forecasting global oil production [71] and detecting malwares [72] due to its computational efficiency, however, because F-test evaluates features independently, it does not capture feature interactions.

The Relief methods consist of distance-based feature selection algorithms that assess feature importance by leveraging local neighborhood relationships in the input space. The core idea shared by all Relief methods is that an informative feature should have similar values for measurements with similar outputs, and clearly different values for data pairs with different outputs, especially within their nearest neighbors [73]. Initially, the method was proposed for binary classification, the Relief framework has been generally improved to handle multi-class classification tasks (ReliefF), regression problems (RReliefF), and to improve robustness

against noise and missing data values [74]. RReliefF is therefore a robust, distance-based feature selection, which is generally applied for regression related tasks.

RReliefF selects the most relevant features by repeatedly drawing samples from the dataset and comparing each one with its nearest neighbors. For each data instance, the algorithm locates neighbors that share similar output values as well as those with different output values, and then adjusts feature weights according to how strongly each feature explains these similarities and differences. Features that change little for different instances and have similar outputs, but they differ significantly for other instances with different outcomes are assigned higher relevance scores. In regression problems, this mechanism is extended by scaling feature differences in proportion to the size of the differences in target values, enabling RReliefF to capture continuous relationships between features and the target. RReliefF in the past decades have been used in various industries for weather forecasting [75], for solar radiation forecasting [76], for energy consumption predictions [77] or for general feature importance analysis tasks [78].

### 2.3.2 Genetic Algorithms

Genetic algorithm (GA) [79], which is a global extreme value searching algorithm that generates different generations of a so-called population during process, while varying their properties towards the optimal solution. Genetic algorithms generally operate on the principle of natural selection, whereby the solving algorithm iteratively and continuously varies the population of solutions and is suitable for solving linear, nonlinear or discontinuous tasks as well. Genetic algorithms by their nature are perfect solutions in order to find the global optima of different predicted values with the use of the trained and hyperparameter optimized machine learning models. The algorithm works according to the following steps:

- creation of a random initial population as Equation 24:

$$\boldsymbol{P} = \{p_1, p_2, \dots, p_n\}, \qquad p_i \in \mathbb{R}^n \tag{24}$$

- generation of several successive populations, using the fittest individuals to generate the next populations according to the steps below:
    1. calculation of the fitness score of each member of the population, which is the raw fitness score as shown in Equation 25:

$$\boldsymbol{F_t} = f(\boldsymbol{p_t}) \tag{25}$$

Where $F_t$ is the fitness value of the t-th solution of each candidate.

2. scaling the fitness values, the scaled values are the expectation values
3. the members with the best fitness values are selected as elite from the population and are passed on unchanged to the next generation
4. parent members are selected from the population, for example with roulette wheel selection according to Equation 26:

$$\boldsymbol{q_t = \frac{F_t}{\sum_{j=1}^{n} F_j}} \tag{26}$$

Where $q_t$ is the probability that individual $p_t$ will be selected, $F_t$ is the fitness value of the t-th solution of each candidate.

5. training children from parent members, with two different methods: either by randomly mutating the parent to create a child, or by different crossover functions through the combination of two parents according to Equation 27, which describes arithmetic crossover, a possible function for crossover:

$$p_{child} = \alpha \cdot p_A + (1 - \alpha) \cdot p_B \tag{27}$$

Where $p_A$ and $p_B$ describe two parent solutions, while $\alpha$ is the crossover weight or blend factor. Mutation of members are descibed by Equation 28:

$$\boldsymbol{p_{t,new} = p_t + \delta_t} \tag{28}$$

Where $\delta_t$ is mutation parameter for the t-th solution of each candidate, a small random value, causing mutation to the population members. Mutation can be characterized by mutation probability and mutation operators (e.g. Real-valued, Uniform, Gaussian, etc.) in order to be able to calculate the final value of the population members.

6. a new generation is formed from the children and elites created

- the algorithm continues the above steps until the stopping criteria is met.

The main steps in the GA process are shown in Figure 11 below.

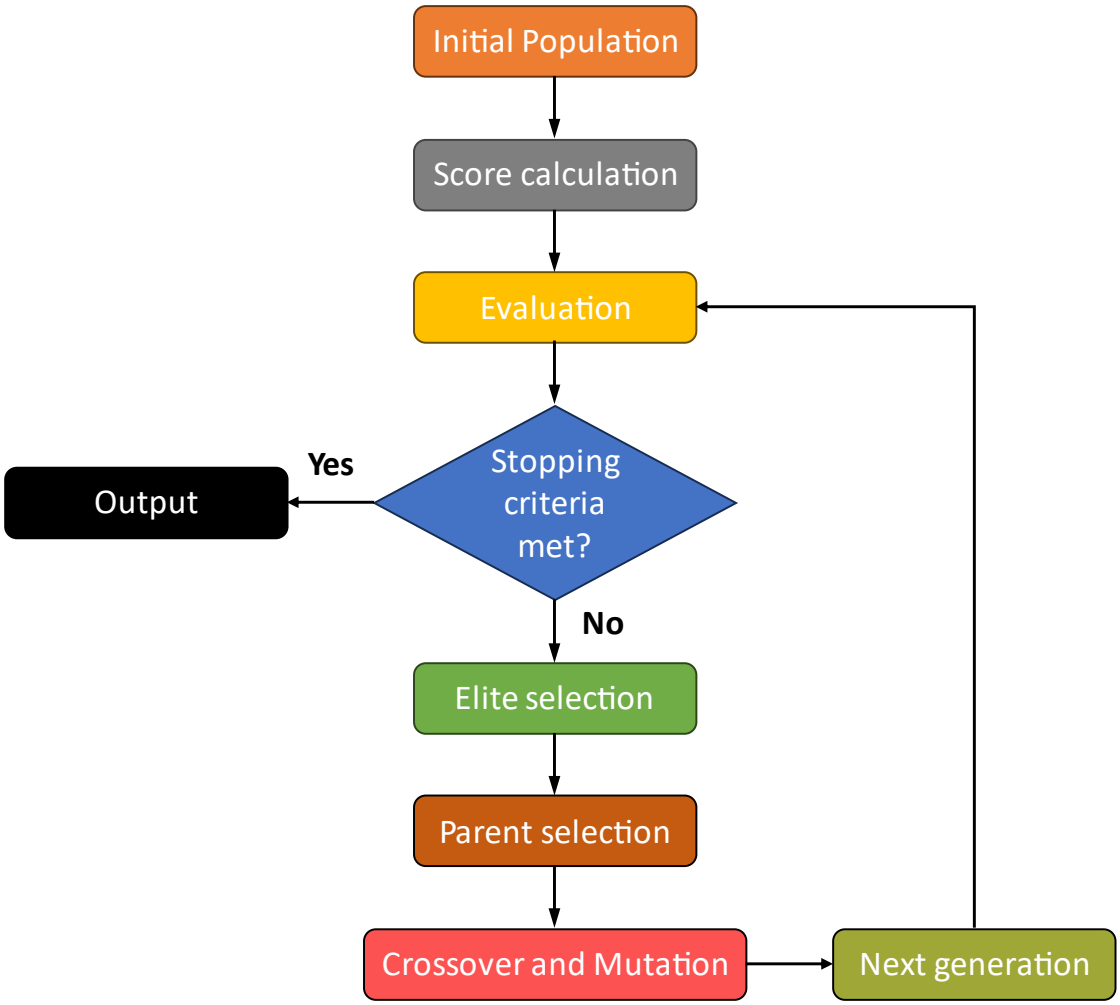


**Figure 11: Main steps of a Genetic Algorithms in general**

### 2.3.3 Partial Dependence Plot

During Machine Learning model development, it is common to seek an understanding of how the model's output is influenced by the predictors, while considering the effect of other predictors are marginalized. A partial dependence plot (PDP) demonstrates the marginal impact of one or more predictors or features on the predicted outcome of a trained machine learning model. It represents the average relationship between a feature and the model's prediction by marginalizing the influence of all other variables. PDPs are beneficial for understanding model operations, veryfing feature expectations, and explaining model behavior. A Partial Dependence Plot (PDP) can be calculated as Equation 29 below:

$$\boldsymbol{PD}_S(\boldsymbol{x_S}) = \mathbb{E}_{\boldsymbol{X_C}}[f(\boldsymbol{x_s}, \boldsymbol{X_C})] = \int f(\boldsymbol{x_s}, \boldsymbol{X_C})\, p(\boldsymbol{X_C})\, d\boldsymbol{X_C} \qquad (29)$$

Where **PD**(·) is the Partial Dependence function, $\mathbb{E}_{X_C}(\cdot)$ is the expected value function, p($\mathbf{X_C}$) is the marginal probability of $\mathbf{X_C}$, which is: p($\mathbf{X_C}$) $\approx \int p(\boldsymbol{x_s}, \boldsymbol{X_C})\, d\boldsymbol{X_S}$. Assuming that each observation is equally likely, and the dependence between $\mathbf{x_S}$ and $\mathbf{X_C}$ and the interactions of $\mathbf{x_S}$ and $\mathbf{X_C}$ in responses is not strong, Partial Dependence estimates the partial dependence by using observed predictor data as follows in Equation 30:

$$\boldsymbol{PD}_S(\boldsymbol{x_S}) \approx \frac{1}{N} \sum_{i=1}^{N} f\left(\boldsymbol{x_s}, \boldsymbol{X_C^i}\right) \tag{30}$$

where N is the number of observations and $\mathbf{X_C^i}$ is the i-th observation for features in $\mathbf{X_C}$. The **PD**(·) function indicates the average marginal effect on the prediction for specified feature values in a set, which is equivalent to averaging all the Individual Conditional Expectations (ICE) in the dataset [80]. This averaging can potentionally mask heterogeneity, particularly when the influence of the predictor is not uniform across the population. This is where ICE plots become useful; they illustrate the conditional relationship for each individual observation as a function of the predictor of interest, and they become useful when it is neccesary to investigate heterogeneities of partial dependence originating from different observations. Essentially, instead of averaging across all $\mathbf{X_C}$, an ICE plot presents individual curves according to Equation 31:

$$\boldsymbol{ICE}_S^i(\boldsymbol{x_S}) = f\left(\boldsymbol{x_s}, \boldsymbol{X_C^i}\right) \tag{31}$$

Where $\mathbf{ICE_S^i}$ is the individual conditional expectation function for i-th observation for features in $\mathbf{X_C}$. An ICE plot offers a detailed and more detailed perspective on how the model output changes with a predictor for specific observations, going beyond the generalized effect depicted by PDP. Some of the significant insights one can obtain include:

- **Heterogeneity or interactions between predictors:** If the ICE curves are not parallel – which is indicating divergence or intersection – it implies that the impact of the predictor $\mathbf{x_S}$ interacts with other variables. Different predictors might be visualised with varying slopes: some being steeper, others flatter, or even having an opposite direction. A PDP by itself could possibly mask this variability by averaging.
- **Outliers or extreme individual responses:** Uncommon ICE trajectories might indicate outliers or measurements where the model behaves unexpectedly concerning that predictor.

- **Nonlinear or piecewise effects:** The shape of individual curves, such as flat regions, sudden shifts and inflection points can indicate nonlinearity or thresholds in the predictor-response relationship that may not be visible in the averaged PDP
- **Average trend vs. variation:** By visualizing the PDP on ICE curves, it can assessed whether the average accurately reflects the overall trend or if there are notable variations among individual curves

### 2.3.4 Shapley – values

The Shapley value, developed by Lloyd S. Shapley in 1953 as part of cooperative game theory, provides a structured method to fairly share the total advantages or costs of a collaborative effort among its individual members. In recent years, this concept has gained significant importance in machine learning explainability, particularly for quantifying the individual contributions of each feature to the dependent variables [81,82]. The attributions can be calculated according to Equation 32:

$$\boldsymbol{\psi_i}(\boldsymbol{\upsilon}) = \sum_{\boldsymbol{S}\subseteq\{\boldsymbol{1},\ldots,\boldsymbol{N}\}\{\boldsymbol{i}\}} \frac{|\boldsymbol{S}|!\,(\boldsymbol{N}-|\boldsymbol{S}|-\boldsymbol{1})!}{\boldsymbol{N}!}[\boldsymbol{v}(\boldsymbol{S}\cup\{\boldsymbol{i}\}-\boldsymbol{v}(\boldsymbol{S})] \tag{32}$$

Where $\boldsymbol{\psi_i}(\boldsymbol{\upsilon})$ is the Shapley value of feature *i*, υ is the characteristic function, formally υ: $2^N \rightarrow \mathbb{R}$, $2^N$ is the power set of N, i.e. all the possible subsets or coalitions of features, S is any subset not including feature *i*, |S|! number of ways to create a coalition *S*, (N – |S| - 1)! is the number of ways to connect features after feature *i*, N! is the number of ways to form a coalition of N parameters, υ(S) is the value of coalition *S* and $[v(S\cup\{i\}-v(S)]$ is the marginal contribution of feature *i* to coalition *S*. In predictive modeling, the function υ(S) represents the expected model when solely the subset of features S is considered, with all other features being marginalized out.

The Shapley value $\boldsymbol{\psi_i}$, quanitifes the average contribution of feature *i* to the prediction by considering all potential interactions between features, establishing Shapley value as an essential element of modern interpretability frameworks.

# 3 Isocyanates in the chemical industry

Isocyanates, characterized by the functional group –N=C=O, are a class of highly reactive organic compounds that play a central role in the production of polyurethanes. Their industrial importance arises from their ability to react readily with hydroxyl groups to form urethane linkages, enabling the synthesis of a vast range of polymeric materials. This reactivity underlies their extensive use in manufacturing flexible and rigid foams, adhesives, sealants, coatings, and elastomers, making them indispensable across sectors including automotive, construction, furniture, and electronics.

## 3.1 Raw materials and importance

In contrast to more commonly produced and used polymers like polystyrene or polyethylene, polyurethanes are synthesized from a diverse array of monomers and thus represent a category of polymers rather than a single compound. Polyurethane polymers, because they contain two types of monomers, polymerise in sequence to form copolymers. This chemical diversity gives polyurethanes their different physical properties, which allows them to be used in a wide range of applications, including rigid and flexible foams, varnishes, coatings, adhesives and other uses. The adaptability of isocyanates originates from their molecular composition, which allows for fine-tuning of polymer properties through the selection of different diisocyanates (e.g., MDI, TDI) and polyols. This adaptability enables the development of materials with tailored mechanical, thermal, and chemical resistance properties, supporting applications from lightweight structural components to insulation and biomedical devices.

The production of polyurethanes involves the reaction of isocyanates with polyols, with the majority of polyols being derived from propylene, which is the most commonly used base material. Figure 12 presents a block diagram illustrating the production process for each isocyanate and propylene-based polyol:

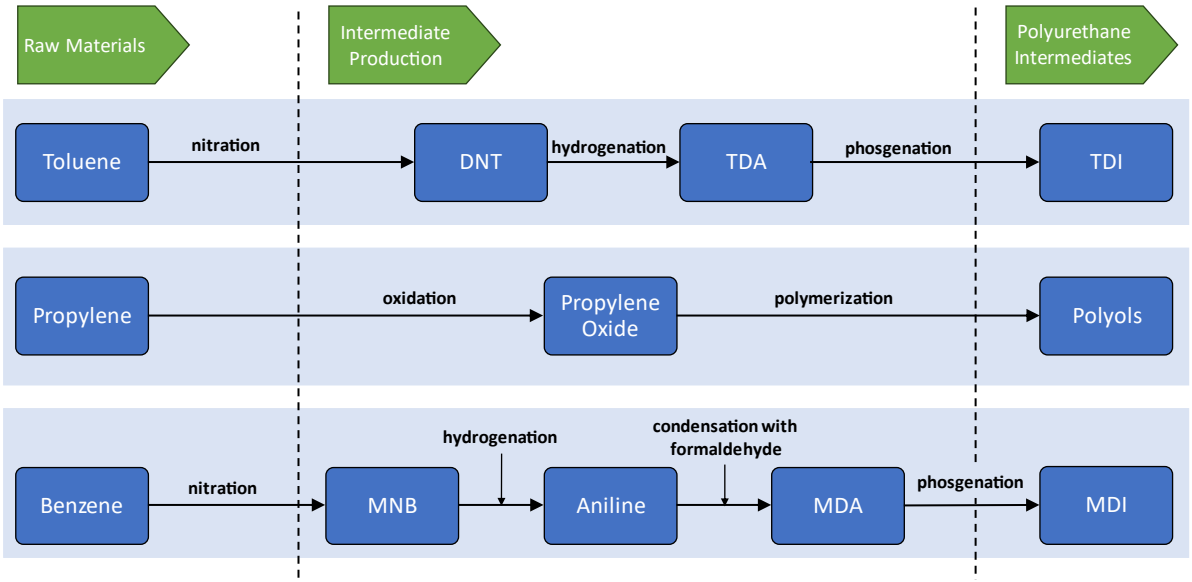


**Figure 12: Simplified block diagram of MDI, TDI and polyol production from raw materials to products**

MDI, an aromatic isocyanate produced in the highest volumes worldwide, acts as a key precursor for a wide range of polyurethane products. The reaction system that involves MDI is complex, featuring multiple reactions, side reactions, and by-products that differ in amounts and properties, thus complicating analytical identification and monitoring. As a result, a sufficiently precise kinetic model discussing MDI synthesis is not yet present in scientific literature.

MDA, in contrast, is not an end product but functions as an intermediate step in synthesizing MDI. It is produced through the condensation reaction of aniline and formaldehyde under acidic conditions. The precise settings of this reaction critically affect the quantitative and qualitative properties of the resulting MDI mixture. Thus, developing a model to explain the formation of MDA is a crucial stage in the process.

## 3.2 Synthesis of Methylenediphenyl diisocyanate

The process of producing MDI involves a multitude of complex reactions and side reactions, making it difficult to identify root causes, side reactions, and by-products that affect product quality. Our goal is to overcome these difficulties using explainable machine learning, utilizing insights obtained to improve real-world industrial applications. By incorporating the outcomes from the machine learning model into the manufacturing workflow through the development of soft-sensors, we aspire to offer data-driven assistance for optimizing processes. The entire machine learning model is constructed using industrial data, and its outcomes have been verified against this data to confirm their reliability and relevance. Owing to the intricate

nature of the material system, various properties must be observed during production to attain an ideal product mix, and numerous unwanted side reactions and by-products influence the quality of MDI products [83]. The three recognized isomers of MDI include 2,2'-MDI, commonly referred to as ortho-ortho-MDI; 2,4'-MDI, known as ortho-para-MDI; and 4,4'-MDI, characterized as para-para MDI, as depicted in Figure 13 below.

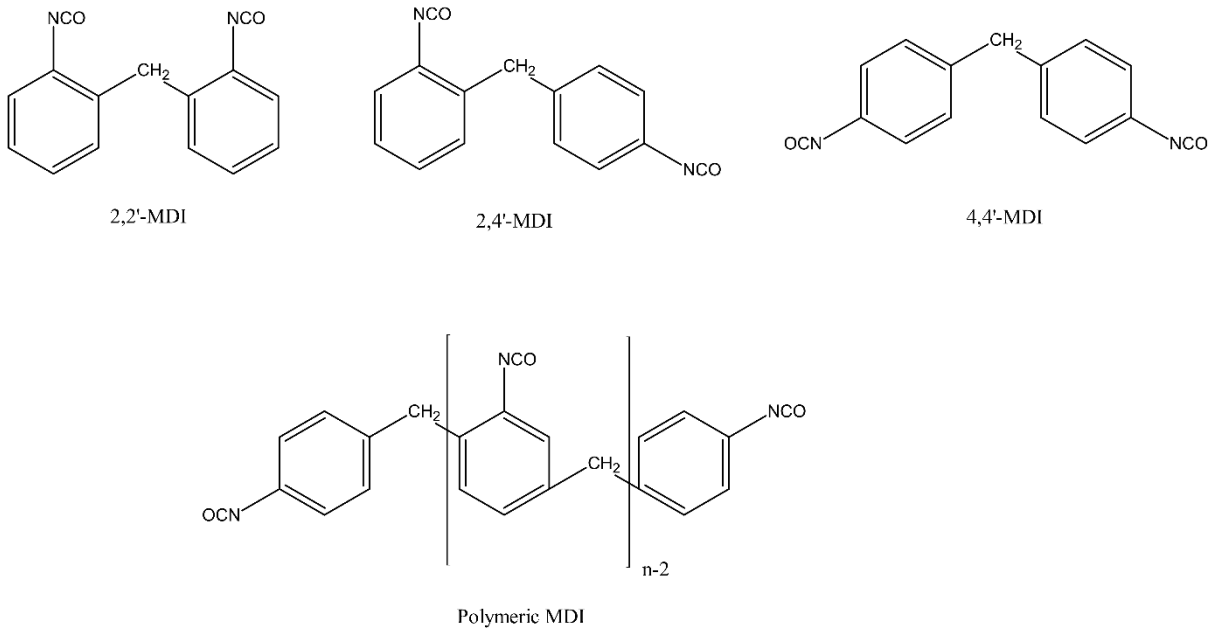


**Figure 13: Different MDI isomers and general formula of polymeric MDI**

MDI isomers are synthesized in forms comprising two or multiple rings during production processes; nevertheless, the 2-ring 4,4'-MDI, designated as "pure MDI," is the molecule produced in the largest quantities and most extensively used. The quantity and quality parameters of the resultant MDI product mixture can be precisely regulated within specific ranges by accurately understanding, defining, and controlling the production parameters of industrial systems. MDI isocyanate mixtures are not directly marketed as finished products; however, they are available for sale as raw materials for subsequent applications. The characteristics of these mixtures can exhibit substantial variability, thereby affecting the quality of the resultant polyurethane products. Although toluene diisocyanate (TDI) mixtures are predominantly marketed in their monomeric form, various MDI mixtures are available, distinguished by their ring distribution and isomeric properties [84,85]:

- Crude MDI, which can also be called a crude product as it contains both monomeric and polymeric MDI components
- Pure MDI, which is the 4-4'-MDI isomer of MDI produced in the largest quantities
- Polymeric MDI or PMDI, containing mostly higher ring number MDI products

- Modified or prepolymer MDIs, which may be made from monomeric or polymeric MDI components produced by catalysed partial reactions

The horizontal product chain of MDI starts with a benzene molecule, which is nitrated to mono-nitrobenzene by mixed acid [86]. The nitrobenzene mixture is subsequently hydrogenated to yield high-purity aniline [87] which can then be directly utilized in the synthesis of MDA [4,88,89]. By adjusting the synthesis parameters, an MDA mixture can be obtained that meets market requirements, through the reaction of formaldehyde and aniline in the presence of solid heterogeneous or liquid homogeneous acids. One of the advantages of MDA production via heterogeneous catalysis, and a significant driving force behind the development of these materials, is the reduction in the volume of effluent from the brine solution. This effluent also contains aromatic hydrocarbons (such as aniline, MDA, and other contaminants), depending on their solubility [90,91].

The raw materials of MDA synthesis are aniline and formaldehyde, which – in case of industrial systems - react in the presence of HCl solution according to the variation of the production parameters, to form the quantitative and qualitative properties of the MDA mixture. This reaction is a so-called condensation reaction, which produces water. The neutralisation of the HCl molecules remaining in the system is taken care in the neutralisation, washing subunit [88]. In the neutralisation reaction, NaOH solution is used in excess to neutralise all HCl molecules. Separation of the organic and aqueous phases is key, so several steps are performed to separate the phases and wash the organic mixture with water to ensure that as little NaCl as possible remains in the organic phase. The neutralised and water washed organic material then needs to be purified in a distillation column. This MDA purification step aims to remove aniline and water from the MDA mixture.

The produced and purified MDA is then transferred to the Phosgenation block, where it is dissolved in ortho-dichlorobenzene (ODCB), the solvent used for MDI production. The purpose of using ODCB is to carry out the phosgenation of the amine compounds, i.e. MDA, in the liquid phase. Phosgene for the phosgenation process is continuously produced from the reaction of CO and $Cl_2$ with a fixed bed catalyst [92–95], then the freshly synthesized phosgene is immediately absorbed in ODCB with deep cooling. The reaction can be seen in Figure 14 below.

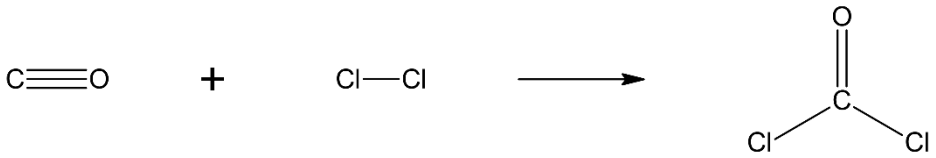


**Figure 14: Formation of phosgene from CO and $Cl_2$**

The MDA-ODCB and PHG-ODCB mixtures are mixed together under intensive conditions to ensure optimal mixing and heat distribution of the reagents, thus reducing the occurrence of side reactions. After mixing the reagents, a multistep phosgenation process is carried out by heating the reaction mixture and ensuring an appropriate residence time, with HCl gas formation as a by-product. The resulting HCl gas after purification can be used for the production of HCl solution to provide the amount of acid catalyst required for the MDA synthesis process [95–97].

To synthesize isocyanates, the most direct and prevalently employed method for isocyanate synthesis entails the reaction of phosgene with either aliphatic or aromatic amines [98]. To synthesize MDI, a widely used route involves first reacting aromatic amines to form the corresponding diamine (MDA), which is then phosgenated to yield the desired diisocyanate (MDI). In the initial step of the reaction, the amine group reacts with phosgene in a highly exothermic process, yielding carbamoyl chloride and releasing hydrochloric acid (HCl). This is followed by the further release of an additional HCl molecule at elevated temperatures, leading to the formation of the isocyanate group. In the case of MDI, the amine molecule involved as raw material is methylenedianiline (MDA), the synthesis and kinetics of which have been extensively studied. The ring distribution of the final MDI product can be influenced by the selection of specific synthesis parameters and molar ratios during the formation of MDA, consequently, MDA plays a critical role in the production of MDI [4,88,90]. The general reactions for the formation of MDA mixtures with different ring distributions is shown in Figure 15.

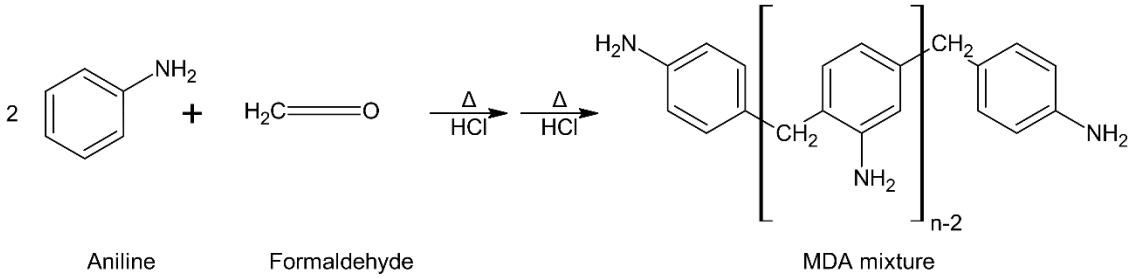


**Figure 15: Formation of MDA mixtures from the reaction of aniline, formaldehyde with HCl as catalyst**

The amine-based route is commonly employed in industry; however, isocyanates can also be synthesized from nitrene intermediates using halogenated compounds. Additionally, non-phosgenic reaction pathways are known, which have regained attention in recent years due to the adverse environmental and toxicological impacts associated with phosgene and chlorine [98].

Subsequent to its formation, during the phosgenation process the MDA mixture undergoes a reaction with a phosgene synthesized from carbon monoxide and chlorine gas. Specifically, this involves the phosgenation of the amine groups through a series of sequential steps. The steps of phosgenation of MDA mixtures are shown in Figure 16.

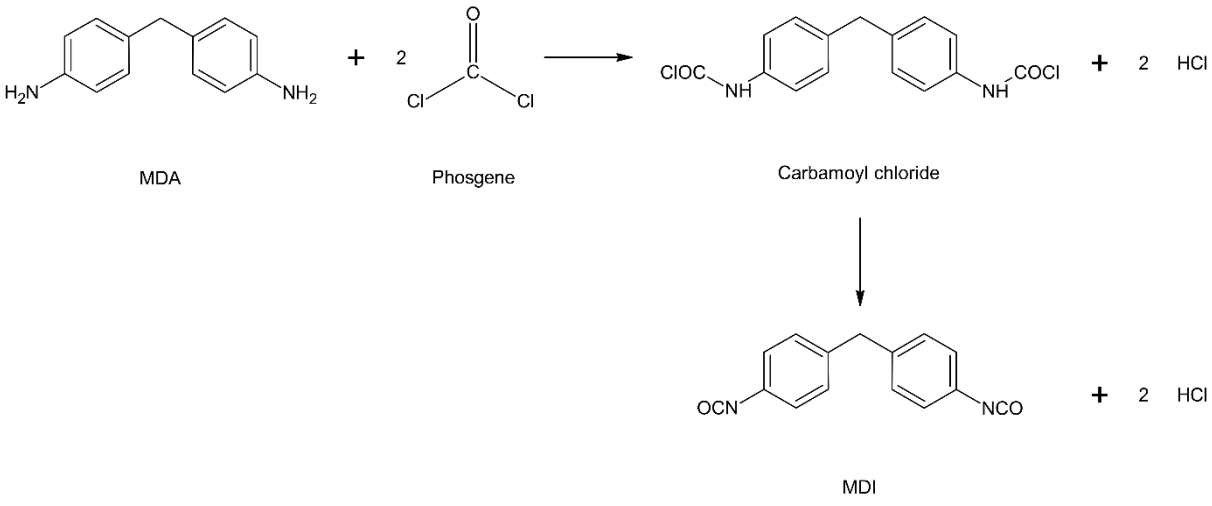


**Figure 16: Steps of phosgenation of MDA to MDI**

The produced by-product HCl gas is recycled either as a raw material for HCl solution production and therefore as a catalyst for MDA synthesis, or for adjusting pH levels within the saline wastewater treatment facility. This treatment procedure encompasses the decomposition of contaminating organic substances and the retrieval of NaCl salt, which may subsequently be

reutilized as a valuable precursor in the synthesis of NaOH and chlorine through electrolysis operations.

After the completion of the phosgenation process, the reaction mixture is purified by pressure reduction to recover excess and unreacted phosgene and to remove other volatile components from the mixture (e.g. HCl gas). The purified reaction mixture is then passed to a stripping system where all remaining volatile components are completely stripped, usually by stripping with HCl gas or nitrogen. The purified MDI - ODCB mixture is then sent to the MDI block, where the solvent, i.e. the ODCB is recovered and the different MDI isomers and MDI molecules with different ring distribution are separated by boiling point differences through distillation, crystallisation, etc.

Figure 17 demonstrates the order of principal operations carried out within each MDI plant block, accompanied by a brief representation of the various material flows connecting the respective blocks.

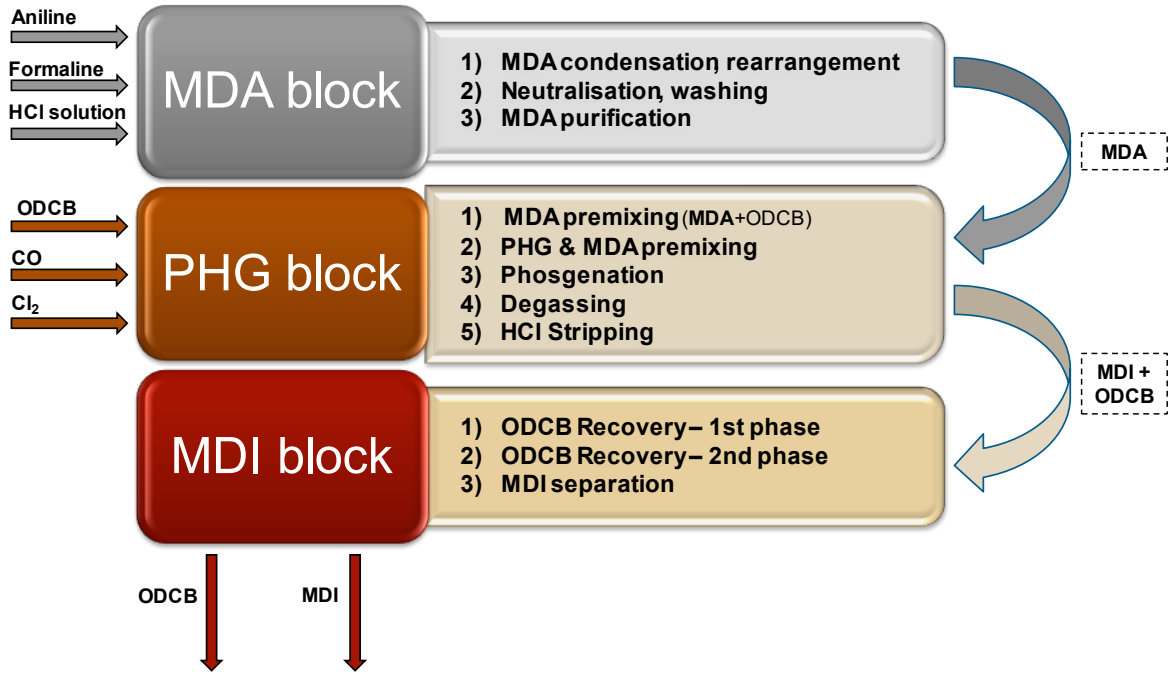


**Figure 17: Block diagram of an industrial MDI production system**

# 4 Exploration and model-based analysis of MDA formation reactions

## 4.1 Importance of MDA formation

The synthesis processes of methylenedianiline (MDA) exhibit considerable complexity due to the multitude of closely related components within these material systems, which present analytical challenges regarding identification and differentiation. Furthermore, the consideration of various potential reaction pathways exacerbates the difficulty in accurately modeling the system. Currently, the precise mapping of the reaction matrix, along with the components involved in reactions related to the generation of methylene diphenyl diisocyanate (MDI), particularly the synthesis of the intermediate MDA, remains a significant challenge. The primary reactions involved in the formation of MDA and MDI are illustrated in Figure 18, with particular emphasis on the formation of MDI through the phosgenation of the 4-4'-MDA isomers.

The synthesis processes of methylenedianiline (MDA) present significant complexity due to the multitude of closely related components within these material systems, which pose analytical challenges in terms of identification and differentiation. Moreover, the consideration of various potential reaction pathways further complicates the accurate modeling of the system. Presently, the precise mapping of the reaction matrix and the components engaged in the reactions pertinent to the generation of methylene diphenyl diisocyanate (MDI), particularly the synthesis of the intermediate MDA, remains a formidable challenge. The primary reactions involved in the formation of MDA and MDI are depicted in Figure 18, where the formation of MDI is specifically emphasized through the phosgenation of the 4-4'-MDA isomers.

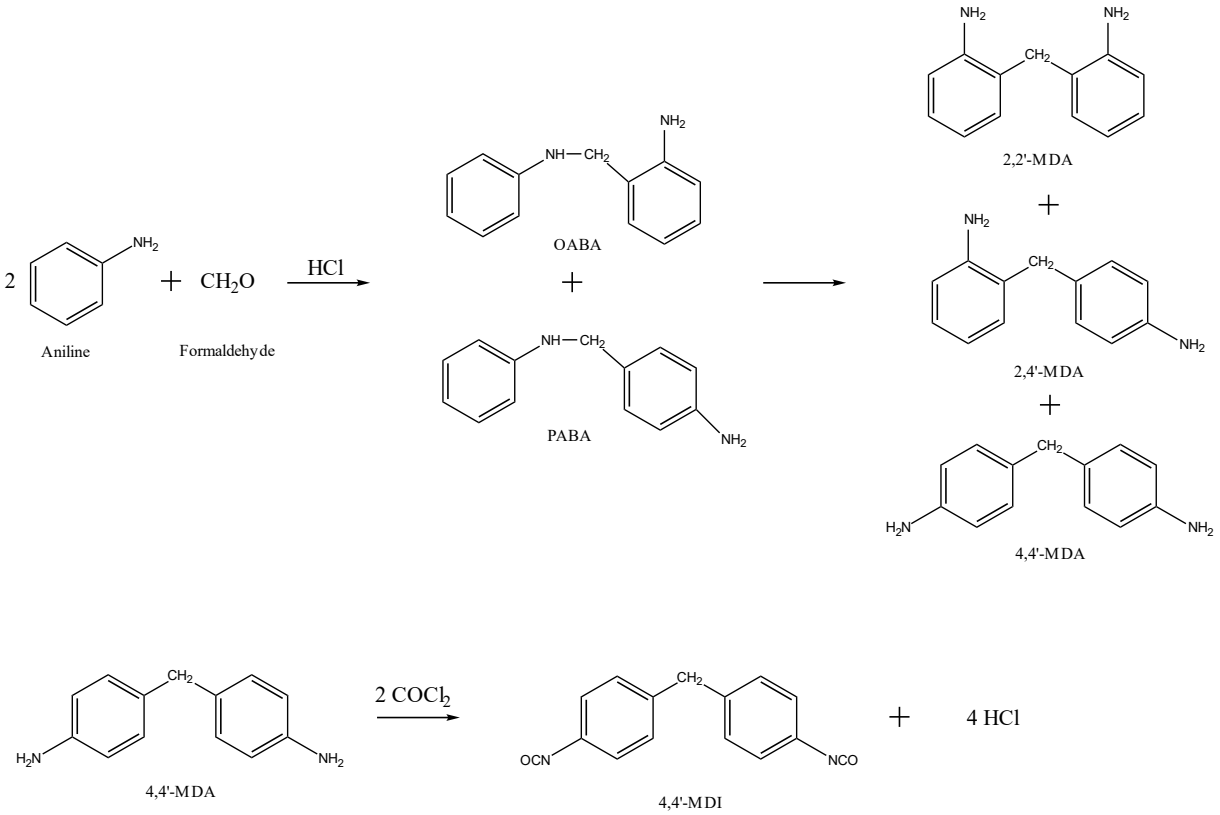


**Figure 18: Formation of different MDA isomers from aniline and formaldehyde in the presence of HCl and the formation of 4,4'-MDI from the phosgenation of 4,4'-MDA**

Few scholarly articles explore this subject, and, in many cases, specific components are excluded from the material system, leading to information loss. Consequently, the resulting kinetic model is difficult to apply to real-world industrial production processes. Our aim is to expand the current literature about MDA related reaction networks to provide a more comprehensive and accurate reaction scheme.

MDA molecules function as key intermediates in the production of MDI. The synthesis of MDA is accomplished through the reaction of aniline with formaldehyde, facilitated by either solid heterogeneous catalysts or homogeneous acids. Within industrial practices, formalin, an aqueous solution of 37 wt% formaldehyde, is commonly utilized. The operational parameters implemented during the synthesis of MDA, in conjunction with the quality parameters of the resulting MDA product mixture, significantly impact the properties of the resulting MDI product mixture. These properties encompass ring distribution, isomer ratio, and other essential key performance indicators (KPIs). Therefore, the synthesis of MDA constitutes a fundamental aspect in the technology of MDI production, as it defines the key characteristic parameters of the final MDI product from diverse perspectives.

In the synthesis of MDA, a two-ring MDA molecule is derived essentially from the combination of two aniline molecules and one formaldehyde molecule. Initially, within the reaction sequence, a condensation reaction occurs in an acidic medium to produce various aminobenzylaniline isomer molecules (o-/p-ABA) as intermediates. These ABA molecules, following a relatively gradual rearrangement process, ultimately transform into MDA molecules. In contrast, in the absence of an acidic component, aniline and formaldehyde molecules engage in a condensation reaction resulting in the formation of aminal. The resultant oligomeric MDA molecules, however, may possess three, four, or even more aromatic rings. The product mixture might include minor quantities of by-products, such as N-methylated products and quinazolines, which are undesirable due to their potential to deplete raw materials. The distribution of the aromatic rings, as well as the quantity and quality of by-products, together with the design and definition of various parameters that characterize the product mixture, can be regulated by adjusting the subsequent operational conditions: for instance,

- aniline/formaldehyde ratio (A/F);
- HCl/aniline ratio (H/A);
- condensation reaction temperature;
- rearrangement reaction temperature;
- reaction or residence time.

The condensation reaction takes place at a lower temperature of about 60-100 °C, while the acid-catalyzed rearrangement reaction, in which O-ABA and P-ABA molecules are converted to MDA molecules, takes place at about 100-160 °C [90]. These MDA molecules are 2,2'-MDA, 2,4'-MDA, and 4,4'-MDA, with the largest amounts of 4-4'-MDA being produced because of stability reasons [89]. The other isomers are also produced in significant quantities, and controlling the proportion of these isomers is very important in the reaction to meet market needs. For homogeneous catalytic processes, any strong mineral acid (e.g. HCl, $H_2SO_4$, $HNO_3$, etc.) can be used. Some researchers also investigated the application of different ionic liquids, but overviewing industrial applications, only the HCl solution is used as catalyst for practical reasons [99]. HCl gas is a by-product of MDI production, which can be absorbed in water and, recycled back to the production of MDA in the form of a solution [100]. Utilization of HCl solution as a catalyst necessitates an additional neutralization step, wherein the neutralization reaction yields a substantial quantity of brine effluent. This effluent is subject to strict regulatory conditions and must undergo treatment processes prior to environmental discharge. Furthermore, the economic implications of the considerable volume of NaOH solution

employed cannot be overlooked [88]. In the context of heterogeneous catalysis, the aforementioned negative effects are not relevant, and these beneficial characteristics have consequently become the focus point of numerous researches in recent decades.

A principal advantage of heterogeneous catalysis, and a fundamental motivation for the advancement of these materials, lies in the reduction of brine effluent volumes, which contain aromatic hydrocarbons (hundreds of ppm of aniline, MDA, and other pollutants), contingent upon their solubility [101]. Heterogeneous catalysts can be reused and thus simplify production processes [102]. In homogeneous catalytic processes, the employment of costly nickel alloys with high nickel content as structural materials is inevitable due to the presence of highly corrosive HCl. This requirement can be circumvented in heterogeneous catalytic processes, leading to a considerable reduction in industrial investment costs associated with equipment and piping. Nevertheless, the efficiency of heterogeneous catalysts can be enhanced by elevating the reaction temperature or increasing their specific surface area and consequently the number of pores. However, these pores pose a risk of solid by-products generated by polymerization reactions obstructing them and thereby rapidly deactivating the catalysts, which constitutes a significant disadvantage [103]. Moreover, while raising the temperature can offset the initially lower and gradually diminishing activity of heterogeneous catalysts in comparison to HCl solution over a brief period, it simultaneously leads to an increase in by-product formation as the reaction temperature rises [104]. Thus, while evaluating the performance, conversion rates, selectivity, and durability of these catalysts, the following factors should be taken into consideration:

- provide a valuable conversion, yield, of valuable products compared to homogeneous catalysis;
- the composition of the product mixture can be controlled - as well as by varying the HCl/aniline ratio;
- have a reasonably long and industrially manageable catalyst lifetime;
- and be able to regenerate to minimise the environmental impact of the process

The most commonly used catalysts are zeolites (e.g. Y, ZSM-5, β, etc.). Over the past few decades, the use of various catalyst forms in MDA production has been explored, focusing on typical catalyst parameters like Si/Al ratio, acidity, specific surface area, and pore distribution. While these have yet to be implemented in industrial applications, all industry producers currently employ HCl as a homogeneous catalyst. Often, the catalysts show high

activity, acceptable conversion rates, and selectivity even at an industrial level; however, they deactivate rapidly, losing over 50% of their activity after merely 8 hours Time On Stream (TOS), which presents a significant challenge for industrial application [105].

Understanding the mechanisms in MDA production, identifying reaction pathways and elements, combined with the creation and refinement of kinetic or population models, allows for accurately assessing the manufacturing process, product mixture quality measures, and crucial performance metrics. Utilizing modeling tools enhances both energy and material efficiency in MDA and MDI production, assists in comprehensive production planning, enables conducting sensitivity analyses, examines alternate operating conditions, and thereby supplies robust, model-based support for industrial-scale production.

This section presents a comprehensive mapping of the reactions and components involved in the MDA synthesis. We identified the kinetic parameters of the system and developed a highly accurate model for these calculations. The model is applicable in examining and supporting production processes on an industrial scale. It also aids in enhancing energy and material efficiencies and improving the safety and stability of operations, thereby ensuring the secure and competitive production of MDI products.

## 4.2 Review of kinetics and mechanisms of MDA formation

Up until now, limited research has focused on the formation of MDA with the goal of identifying the kinetic parameters of reactions in homogeneous catalytic processes. For instance, a study by Ogata *et al.* examined the rate of the condensation reaction involving aniline, formaldehyde, and hydrochloric acid, along with the impact of each component and the pH level, influenced by HCl concentration [106]. The condensation reaction was determined to be of second order with respect to aniline, while exhibiting first-order kinetics concerning formaldehyde and the resulting methylene anilines. Furthermore, the study explored the influence of various inorganic salts on the rate of the condensation reaction. It was observed that an increase in salt concentration leads to an enhanced reaction rate, an effect attributed to both primary and secondary salt effect [107]. The studied equilibrium condensation reactions are described as can be seen in Figure 19:

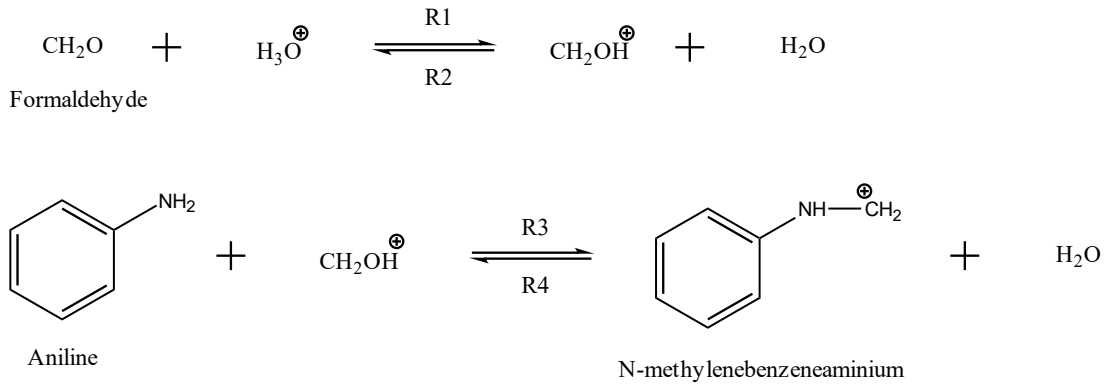


**Figure 19: Condensation reaction of aniline and formaldehyde in acidic environment [106]**

The findings indicated that raising the aniline concentration enhanced the rate of the condensation reaction R3, while in contrast, altering the formaldehyde concentration did not notably impact the R3 reaction rate. However, a higher concentration of HCl reduced the speed of the R3 reaction. When both aniline and formaldehyde concentrations were increased, the rate of the condensation reaction notably surged.

As shown in Figure 20, Nayar *et al.* also examined and presented findings on the reaction rates between aniline and formaldehyde in their study [108]. The reaction rate constant was additionally characterized as a pseudo-first order constant, with findings from Nayar *et al.* closely aligning with those reported by Ogata *et al.* Both studies explored the underlying reaction mechanisms and proposed potential pathways for both heterogeneous and homogeneous catalysis.

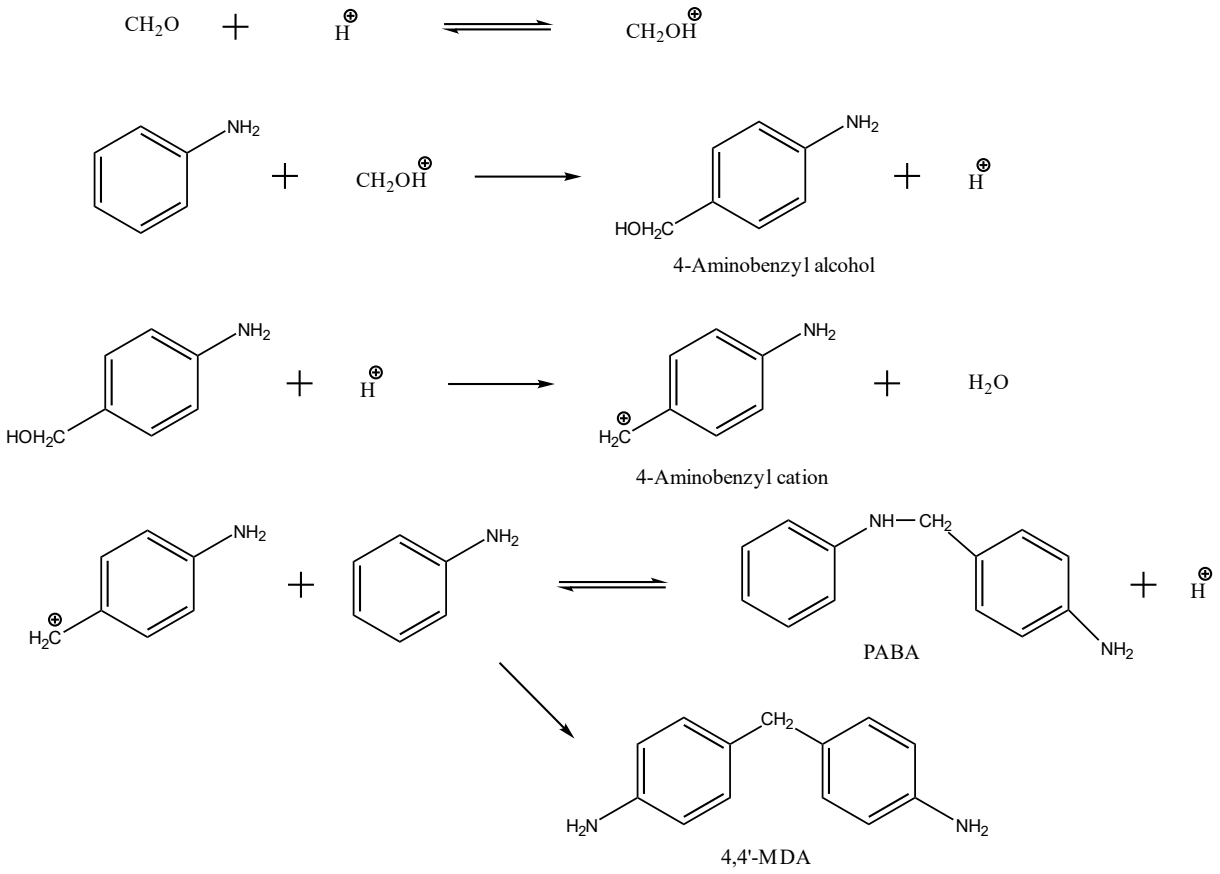


**Figure 20: Suggested reaction mechanism of aniline and formaldehyde in acidic medium [108]**

According to the predicted mechanism, the initial intermediate produced is 4-Aminobenzyl alcohol, resulting from the reaction of aniline with formaldehyde when the latter attacks the para position. In the presence of an acid catalyst, 4-Aminobenzyl alcohol transforms into the 4-Aminobenzyl cation. Subsequently, this cation is converted into 4,4'-MDA through a gradual reaction, with PABA maintaining equilibrium with the 4-Aminobenzyl cation.

Within the literature about heterogeneous catalysis, several mechanisms vary to some extent concerning the components and reaction pathways involved in MDA synthesis. One such mechanism is detailed in the article by Corma *et al.* [105]. In their study, the authors examined delaminated zeolite catalysts that showed higher activity, extended lifetime, and precisely managed selectivity compared to conventional zeolites, which did not undergo delamination. Their findings indicated that even the most promising β-zeolites among the tested catalysts were limited by diffusion, with only a portion of the acid sites being available to the reactants. Furthermore, they suggested a mechanism for homogeneous catalytic systems as illustrated in Figure 21. Differing from previous theories, Corma *et al.* identified the initial step as the reaction between aniline and the acidic agent, leading to the formation of ammonium chloride

molecules. These molecules act as intermediates in the MDA production process, eventually transforming into product molecules following a neutralization step.

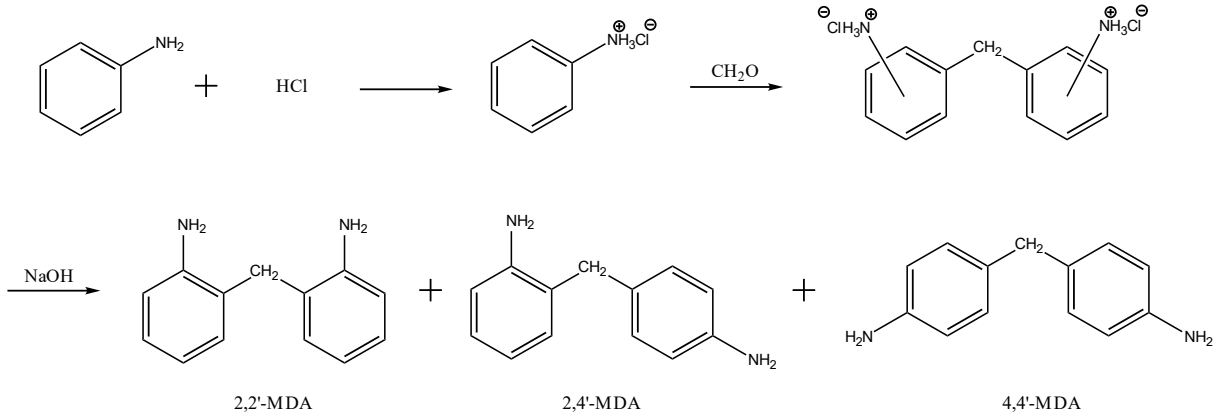


**Figure 21: Proposed mechanism for industrial synthesis of MDA with HCl as homogenous catalyst [105]**

Unlike the conventionally described mechanism for heterogeneous catalysis, Corma *et al.* propose a new reaction mechanism and pathways, visualised in Figure 22. This newly suggested mechanism highlights a key role for a new component, named the aminal, in the MDA synthesis process. Under acidic conditions, the aminal is converted into aminobenzyl-aniline molecules, which subsequently rearrange through multiple steps into various MDA isomers. Kugita *et al.* explored the activity and selectivity of different zeolite catalysts, specifically Y-, β-, and ZSM-5 variants [109]. The influence of varying Si/Al ratios on the behavior of β-zeolites was also investigated. The findings reveal that while Y-zeolites exhibited reduced activity when compared to β-zeolites, they demonstrated significantly greater selectivity for 4,4'-MDA. Furthermore, the study confirmed that the reaction temperature, the aniline-formaldehyde ratio, and the quantity of catalyst employed greatly impact both the yield of MDA and the isomer distribution.

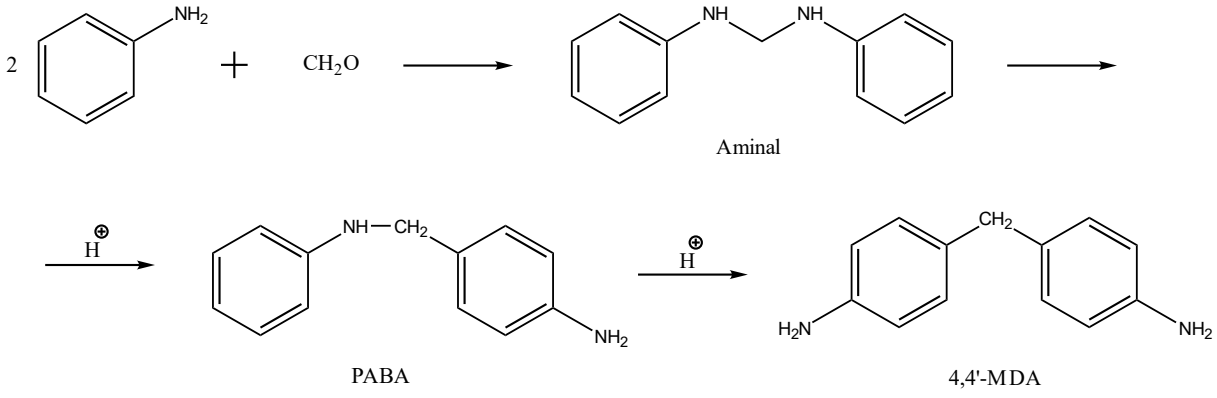


**Figure 22: Aminal formation and MDA synthesis on solid acid catalysts [105]**

Delaminated zeolite catalysts were also investigated by Botella *et al.* [110]. The ITQ-2, ITQ-6, and ITQ-18 catalysts exhibited notably high activity and extended lifetimes in comparison to traditional zeolites. The engineered surface design of delaminated zeolites facilitates precise control of product mixture isomer distribution, providing a significant advantage over conventional catalysts. The study also explored the impact of the external surface area, micropore dimensions, and both Bronsted and Lewis acidity of the catalysts to gain a more detailed understanding of the parameters influencing reaction rate and selectivity. As anticipated, catalysts with easily accessible large external surfaces demonstrated higher activity than conventional ones, though their isomer selectivity was notably influenced by the treatment procedures applied to them. Altering reaction conditions, such as increasing the reaction temperature or the amount of catalyst, can enhance catalyst activity, yet these adjustments also significantly affect the resulting isomer distribution and parameters like the volume and nature of by-products. Their reaction scheme emphasized mono-methyl-MDA (MMM) or N-methyl-MDA as a key by-product, which is formed through a somewhat selective Friedel-Crafts alkylation process and may serve as evidence for quinazoline formation. The anticipated reaction system is illustrated as depicted in Figure 23.

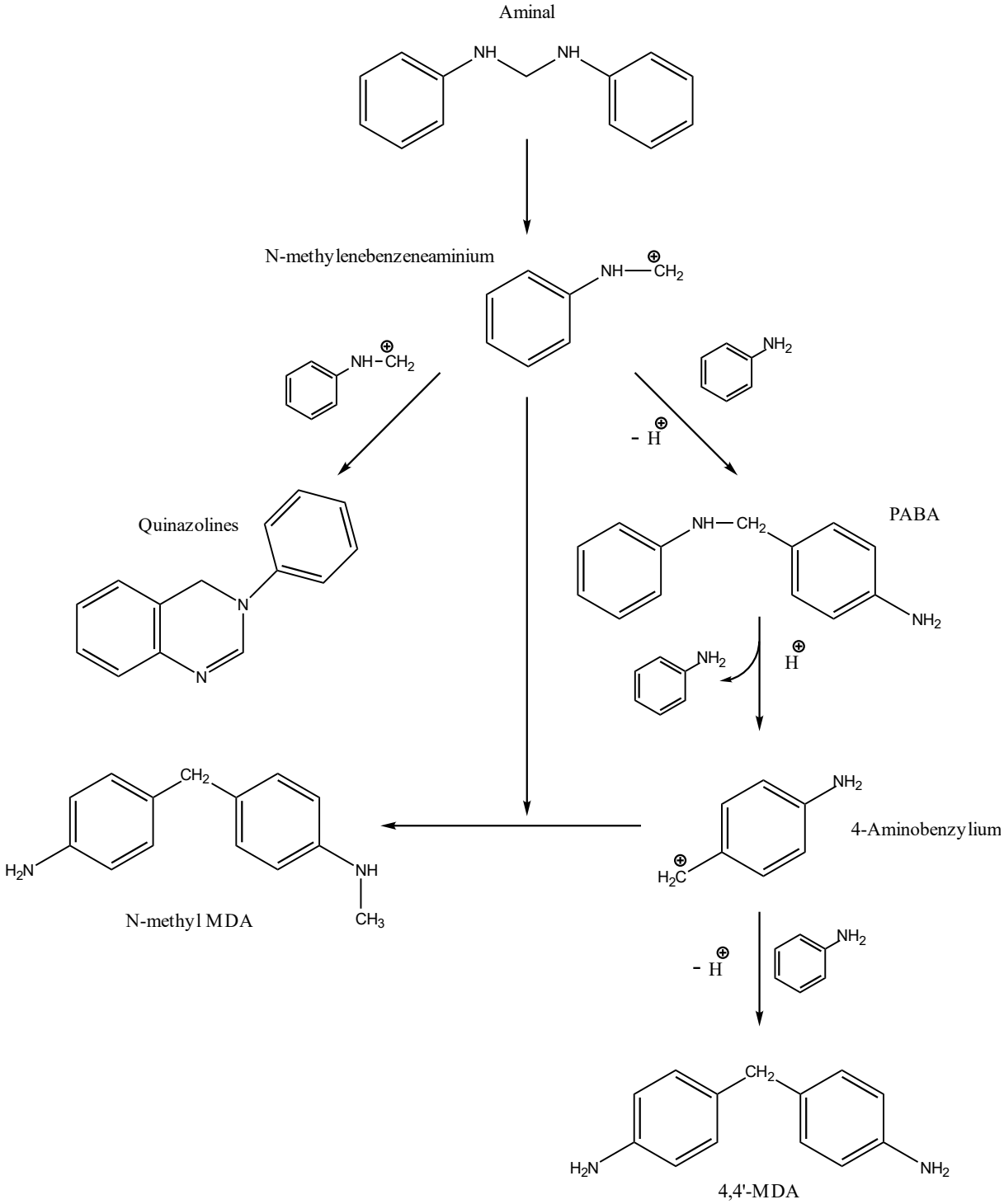


**Figure 23: General reaction scheme for the isomerization and rearrangement of aminal molecules in neutral environment on zeolitic catalysts [110]**

The mechanism of 4,4'-MDA formation was investigated by Wang *et al.* [111]. The findings indicate that the reaction probably proceeds through bimolecular nucleophilic substitution, known as the $S_N2$ mechanism. In the study, they also examined the impact of the aniline/formaldehyde molar ratio on both the quantitative and qualitative properties of the

product mixture. A detailed reaction matrix was established to discover the mechanisms, highlighting HCl as the catalyst, as demonstrated in Figure 24. The reactions were segmented into two sections for enhanced clarity.

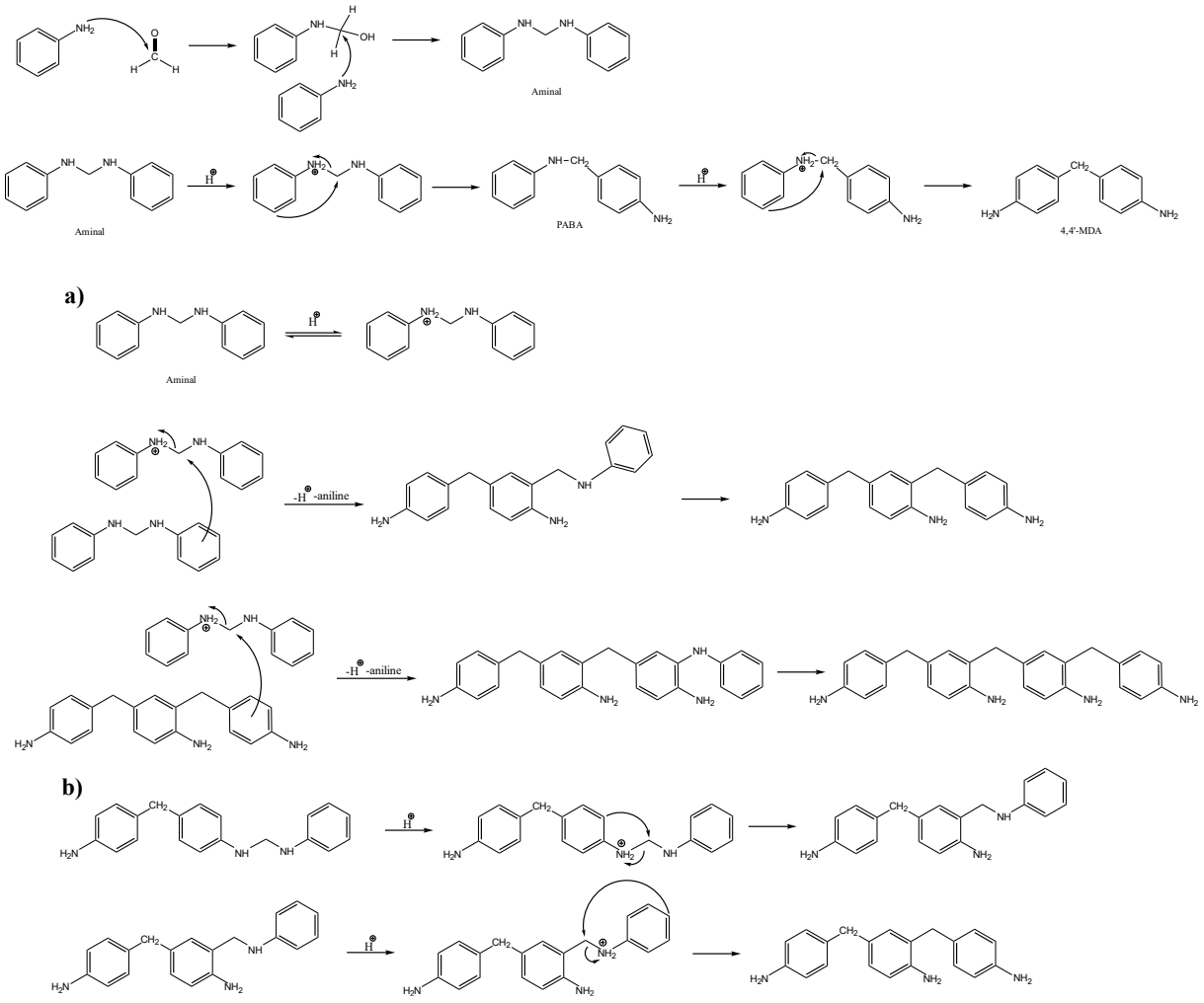


**Figure 24: Possible reaction mechanism of HCl-catalyzed condensation reaction of aniline and formaldehyde [111]. (a) mechanism of chain reactions and (b) MDA molecule further transformation to other derivatives**

The components produced can differ with changes in reaction conditions. Under neutral conditions with an aniline/formaldehyde ratio approximately equal to two, an aminal molecule is initially formed. By introducing acid into this mixture, the molecule undergoes several rearrangement steps, resulting in an MDA molecule. If the aniline/formaldehyde ratio is lower than two, the process tends to yield oligomeric N,N'-methylene compounds, and upon acidifying the solution, these are rearranged into oligomeric MDA molecules. Figure 25 illustrates the reaction pathways and conditions.

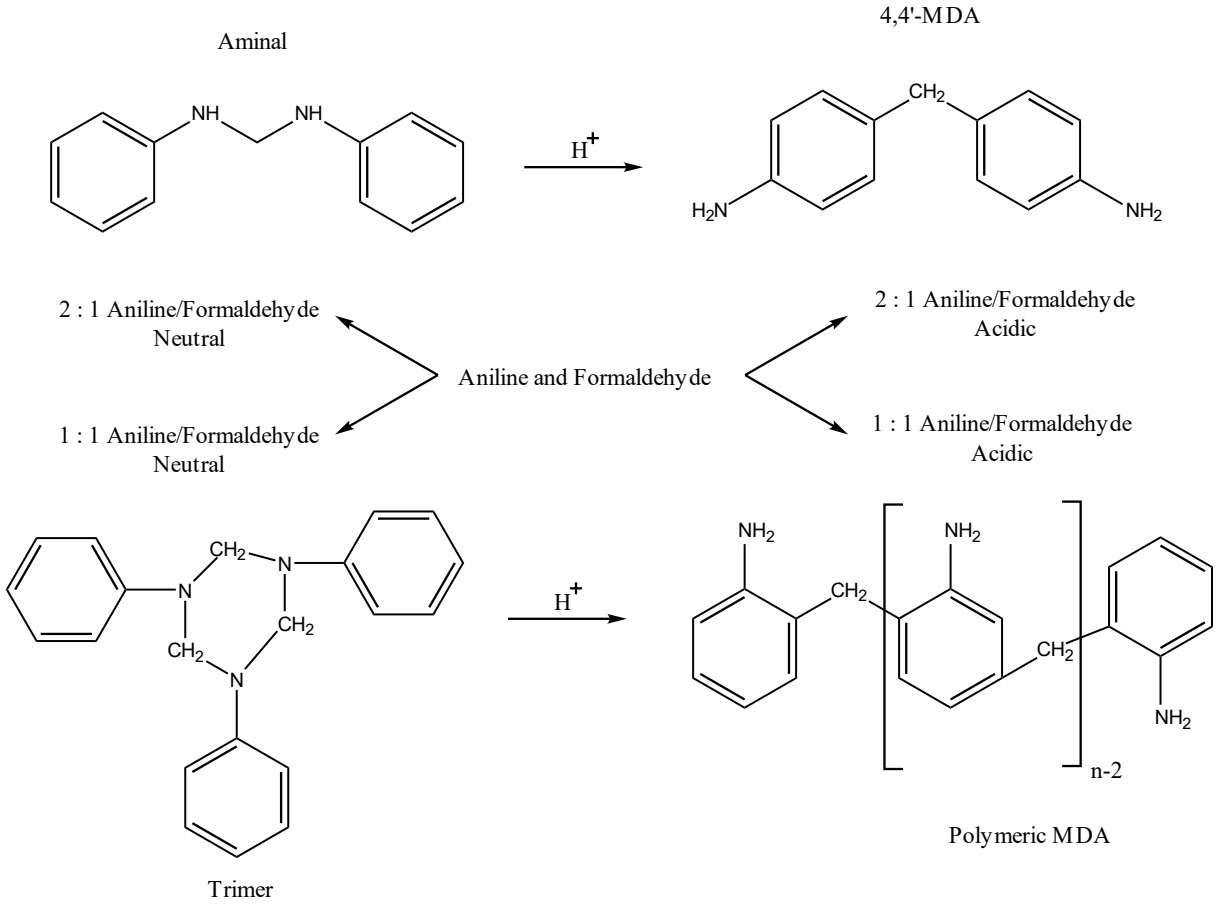


**Figure 25: Alternative reaction pathways for the reaction of aniline and formaldehyde in different reaction conditions and with different reagent ratios [111]**

Salzinger *et al.* investigated Y-type zeolites as well. For the β-zeolites they examined, it was determined that the initial stage of the reaction, involving the formation of amino-benzyl-aniline molecules from aminal molecules, is constrained by pore diffusion. In contrast, the considerably slower subsequent stage, where these molecules are rearranged into MDA isomers, is governed by reaction kinetics [112,113]. In mapping the mechanism driving these reactions, the anticipated pathways for the cleavage of aminal molecules, as well as the associated chain growth and rearrangement reactions for each molecule, were proposed. According to the kinetic data, a reaction of the $S_N2$ type was hypothesized, which is illustrated in Figure 26.

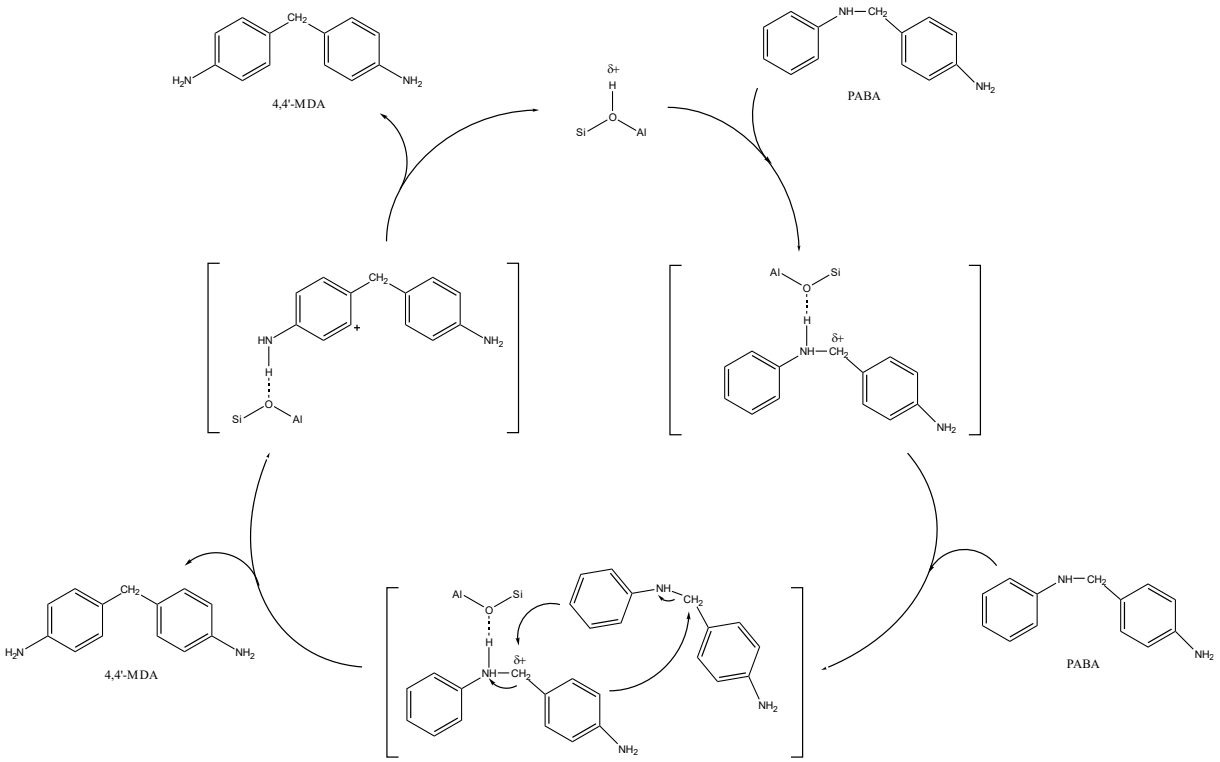


**Figure 26: Proposed $S_N2$ reaction mechanism for the formation of PABA and MDA molecules [112]**

Following an investigation into the reaction mechanisms, a simplified reaction network was developed, containing 8 components and 15 reactions and the corresponding differential

equations were employed to characterize the changes in concentration for each component. The entire specified reaction scheme is illustrated in Figure 27.

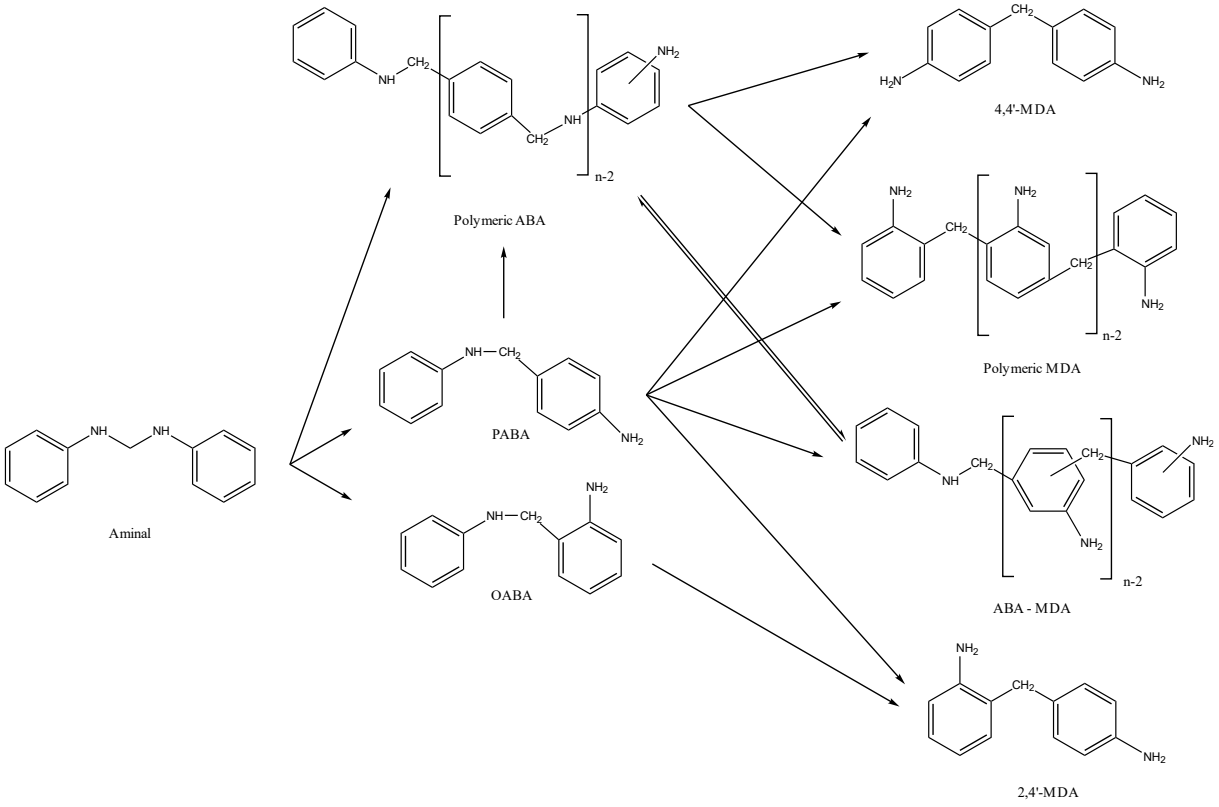


**Figure 27: The simplified reaction network of MDA formation used for nonlinear parameter fitting proposed by Salzinger *et al.* [112]**

Employing this simplified model resulted in a strong correlation with the concentrations of the components measured, namely PABA, OABA, 4,4'-MDA, poly-PABA, and MDA-PABA.

The reaction mechanism, its modelling and the determination of the kinetic parameters have also been addressed by Haus *et al.* [114]. Extensive batch experiments were performed to explore the system, its various elements, and their interactions. To describe the material system, both a kinetic model and a population balance model were utilized, with the utilisation of an amorphous silica-alumina catalyst. In a simplified reaction matrix which was proposed in order to accurately characterize the system, Haus *et al.* identified the material system as consisting of 9 components and 11 reactions, with reaction details shown in Figure 28.

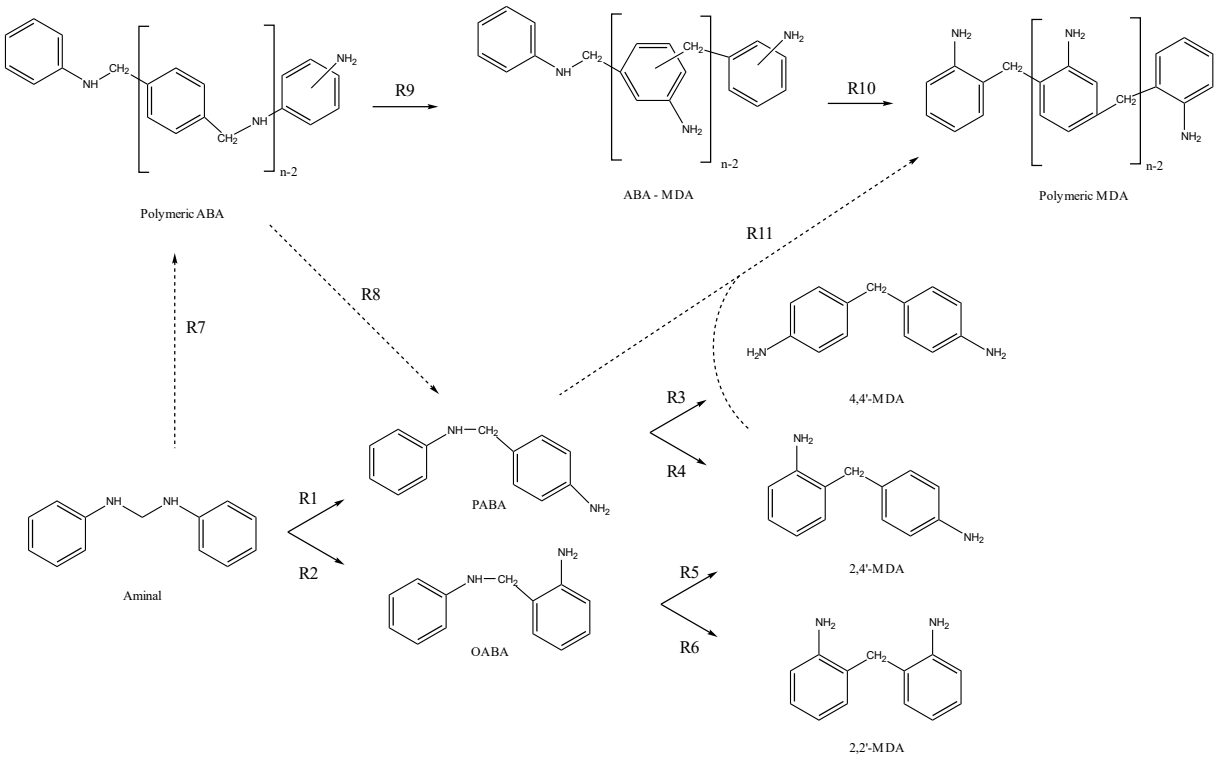


**Figure 28: The simplified reaction network of MDA formation used for fitting empirical kinetic model parameters by Haus *et al.* [114]**

Various components can be classified based on their characteristics. The unwanted by-product is 2,2'-MDA, while the desired outcomes include polymeric MDA molecules, 4,4'-MDA, and polymer ABA molecules. ABA-MDA molecules serve as oligomeric intermediates in the procedure.

Haus *et al.* employed a standard Arrhenius equation to illustrate how the reaction rate constant depends on temperature. The changes in concentrations were characterized by ordinary differential equations. The kinetic parameters found in Equations 33 and 34, namely activation energies and preexponential factors, were identified to describe the process.

$$\frac{\partial c_\kappa}{\partial t} = \sum_{r=1}^{11} \nu_{\kappa,r} \cdot k_r \cdot \prod_{\delta \in \kappa} c_i^{n_{r,\delta}} \cdot \omega_{catalyst} \tag{33}$$

$$k_r = k_{0,r} \cdot e^{-\frac{Ea_r}{RT}} \tag{34}$$

Where $\frac{\partial c_\kappa}{\partial t}$ rate of change of concentration of component κ, $\nu_{\kappa,r}$ is the stoichiometric coefficient for component *k* in reaction *r*, $k_r$ is the reaction rate constant of reaction *r*, $c_i^{n_{r,\delta}}$ is the concentration of component *i* involved in reaction *r* raised to the respective reaction order $n_{r,\delta}$ and $\omega_{catalyst}$ is the catalyst weight fraction.

Based on the measured results the fitted kinetic parameters for each reaction by Haus *et al.* are shown in Table 4 below.

**Table 4: Fitted parameters of the empirical model by Haus *et al.* [114]**

| Reaction | ln($k_0$) [-] | Ea [kJ/mol] |
|---|---|---|
| **R1** | 16.3 ± 2.8 | 62.3 ± 8.3 |
| **R2** | 16.7 ± 2.8 | 71.0 ± 8.2 |
| **R3** | 24.1 ± 0.4 | 97.0 ± 1.1 |
| **R4** | 27.7 ± 0.4 | 113.5 ± 1.3 |
| **R5** | 27.0 ± 1.2 | 103.2 ± 4.0 |
| **R6** | 25.8 ± 2.3 | 117.9 ± 7.5 |
| **R7** | 16.6 ± 2.9 | 66.8 ± 8.4 |
| **R8** | 22.2 ± 2.0 | 98.4 ± 6.6 |
| **R9** | 20.2 ± 7.0 | 88.6 ± 23.0 |
| **R10** | 22.2 ± 7.4 | 100.1 ± 24.0 |
| **R11** | 24.0 ± 0.7 | 96.8 ± 2.4 |

The accuracy of the model can be described with defining KPIs as well. With the use of these KPIs the MDA product quality and the performance of the catalysts can be described. These KPIs in case of MDA mixtures are usually:

- Remaining Reactants (RR): sum of all 2-ring ABA molecules
- Isomer Ratio (IR): molar ratio between 4,4'-MDA and 2,4'-MDA molecules
- Oligomer Fraction (OF): sum of all oligomeric molecules (3≤ ring molecules)
- N-substituted by-products (NS): sum of all N-substituated components

Despite the relatively large uncertainties of the parameters in Table A1, the calculated KPI values showed a good fit compared to the measured values, but the results were not compared in terms of the concentration of each component.

## 4.3 Proposition of an extended reaction network

The analysis of the literature indicates that the complexity of the material system, coupled with uncertainties in the reaction pathways and the numerous components involved, leads to varied simplifications across different studies. These simplifications, while improving tractability, inevitably diminish both accuracy and the range of extractable information. Simplified models of material systems often struggle to align calculated concentrations with experimental data due to omitted reaction pathways. Additionally, neglecting the roles of

aniline and formaldehyde when starting with the aminal molecule itself further obscures information, specifically preventing determination of the aniline/formaldehyde mole ratio. Frequently, N-substituted by-products are also overlooked. As neither the components nor the pathways are fully defined, tracking or predicting by-product formation becomes unfeasible.

In the course of enhancing the reaction system, we took the following factors into account:

- the difference between measured and calculated concentrations should be minimal;
- the difference between measured and calculated KPI values should be minimal;
- the number of components and reactions are as low as possible to ensure simplicity.

For consistency and comparability, the data collected by Haus *et al.* were utilized in the model development. The objective was to reduce model error by reestimating the reaction rate constants for each reaction. When accurately determined, this will enable the calculated values to closely match the measured concentrations and KPIs with satisfactory precision.

In order to achieve this objective, we have also outlined several smaller essential tasks within our identification strategy:

- the model must be capable of managing the molar ratio of aniline to formaldehyde as a crucial factor in MDA production
- Together with the KPIs, the computed concentrations for each component must also demonstrate a strong alignment with the observed data.
- the model must be capable of processing molecules that possess an odd quantity of rings, implying that not just molecules with a pair of rings are able to undergo oligomerization

Based on the results measured by Haus *et al.* and, in order to resolve the uncertainties in the constructed model and to be able to meet the key tasks defined earlier, we propose a new, more detailed model that more accurately describes the mechanism of MDA formation, as shown in Figure 29.

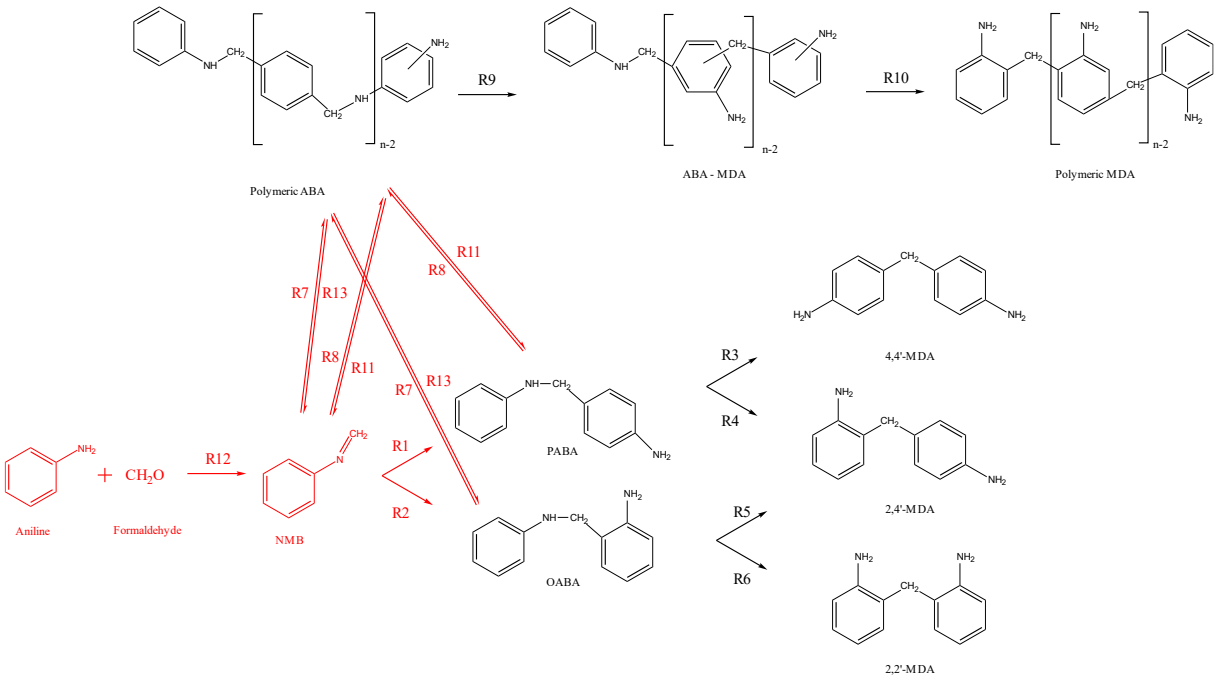


**Figure 29: The suggested optimized reaction network of MDA formation used for fitting the extended empirical kinetic model, newly added or redefined reactions and molecules are highlighted with red color**

The reaction system has been expanded to include three additional components: aniline, formaldehyde, and N-methylbenzene (NMB). The previously included aminal component has been removed and replaced with definitions for the aniline and formaldehyde components, enabling the model to accommodate the mole ratio of aniline to formaldehyde. Furthermore, the reaction mechanisms and pathways have been reevaluated. With the introduction of the NMB component, it is now possible to produce an odd number of rings during oligomerization, as oligomeric structures can form by linking the 1-ring NMB molecule to ABA molecules.

Utilizing the new material system allowed for the determination of the reaction rate constants. However, the activation energies were not determined because measurements were only conducted at a single temperature [114].

## 4.4 Kinetic parameter identification strategies

The study and optimization of the material system were conducted through four different phases, each representing a separate case. In every case, we determined and adjusted the reaction rate constants to align with the observed data. An overview and description of each case are provided in Table 5.

**Table 5: Summary of Cases and objective functions used for reaction rate identification based on Table A1**

| Case name | Remarks |
|---|---|
| Initial state | Simulation without parameter identification |
| A | Simulation & parameter identification |
| B | Simulation & parameter identification with extended reaction scheme |
| C | Simulation & parameter identification with extended reaction scheme iteratively |

- Initial state: simulation of the material system with the reaction rates obtained from Table 4 with no parameter identification defined.
- Case A: uncertainties of kinetic parameters identified and summarized in Table A1 were used as the lower bounds (LB) and upper bounds (UB) to define parameter identification task in order to fit reaction rates with interior-point method and with Equation 37.
- Case B: proposed extended reaction scheme with reidentified reaction rate constants, fitted with Equation 37.
- Case C: extended reaction scheme, where literature & proposed and modified reactions were fitted iteratively with Equation 37, upper and lower bounds were defined based on Table A2 data. The bounds defined for Case A and B were used for the unmodified or literature reactions, while the bounds defined for Case C were used for the proposed or modified reactions.

As highlighted in Case C, we employed an adjusted identification approach aimed at gaining a deeper insight into the behavior of the reaction system. This iterative method involved dividing the reaction system into two segments. The first segment comprised all previously published reactions in our proposed network, without any alterations to these reactions. Reaction rate constants for these, as referenced in prior literature, include R3, R4, R5, R6, R9, and R10. The second segment consists of newly integrated or redefined reactions essential for identification of their reaction rate constants, which include R1, R2, R7, R8, R11, R12, and

R13. In Figure 29, these newly added or redefined reactions are marked in red, whereas unaltered reactions are indicated in black.

In order to perform the kinetic parameter identification with sufficient accuracy and with better convergence, it is necessary to normalise the upper and lower bounds between 0 and 1 [115]. This was done using Equation 35, where $\overline{k_N}$ is the identified parameter, optimized during the extremum search task:

$$\overline{\boldsymbol{k}} = \overline{\boldsymbol{LB}} + (\overline{\boldsymbol{UB}} - \overline{\boldsymbol{LB}}) \cdot \overline{\boldsymbol{k_N}} \tag{35}$$

The model of the reaction system is defined as ordinary differential equations describing the concentration trajectories with Equation 36:

$$\frac{\partial c_i}{\partial t} = \sum_{r=1}^{p} \nu_{i,r} \cdot k_r \cdot \prod_{i=1}^{j} c_i^{\nu_{i,r}} \cdot \omega_{catalyst} \tag{36}$$

Where $\frac{\partial c_i}{\partial t}$ rate of change of concentration of species *i* with respect to time, *p* is the total number of reactions, $\nu_{i,r}$ is the stoichiometric coefficient for component *i* in reaction *r*, $k_r$ is the reaction rate constant of reaction *r*, *j* is the number of considered components in the proposed reaction mechanism, $c_i^{\nu_{i,r}}$ is the concentration of component *i* involved in reaction *r* raised to the respective reaction order $\nu_{i,r}$ and $\omega_{catalyst}$ is the catalyst weight fraction.

The differential equations were solved using the ode23s function [116,117] in MATLAB software and fmincon function for the parameter identification task [118–121]. Nonlinear parameter identification was performed by comparing the measured and calculated results. The sum of the squared error, calculated from the differences between the measured and calculated concentrations were defined as the objective functions to be minimized as shown in Equation 37 for the different cases:

$$MIN(F(k_N)) = \sum_{i=1}^{j} (c_{k,i}^{measured} - c_{k,i}^{calculated})^2 \tag{37}$$

The initial conditions of the reaction rates as well as the lower- and upper bounds for the parameter identification are formalized and are given in Table A2 for each cases.

## 4.5 Results and discussion

Following the methodology presented in Section 4.3 and Section 4.4, we present the original literature and the identification results for the cases defined and systemized in Table 5. Additionally, we display the alignment of computed concentration trajectories, derived from the identified parameters, against observed results, which can be found in Figure 30. The results indicate that the precision of the proposed MDA formation model surpasses that of models derived from existing literature. By leveraging the main identification strategies highlighted, both concentration and KPI fits were achieved with high accuracy. Figure 30 exhibits the results calculated with each model for each case.

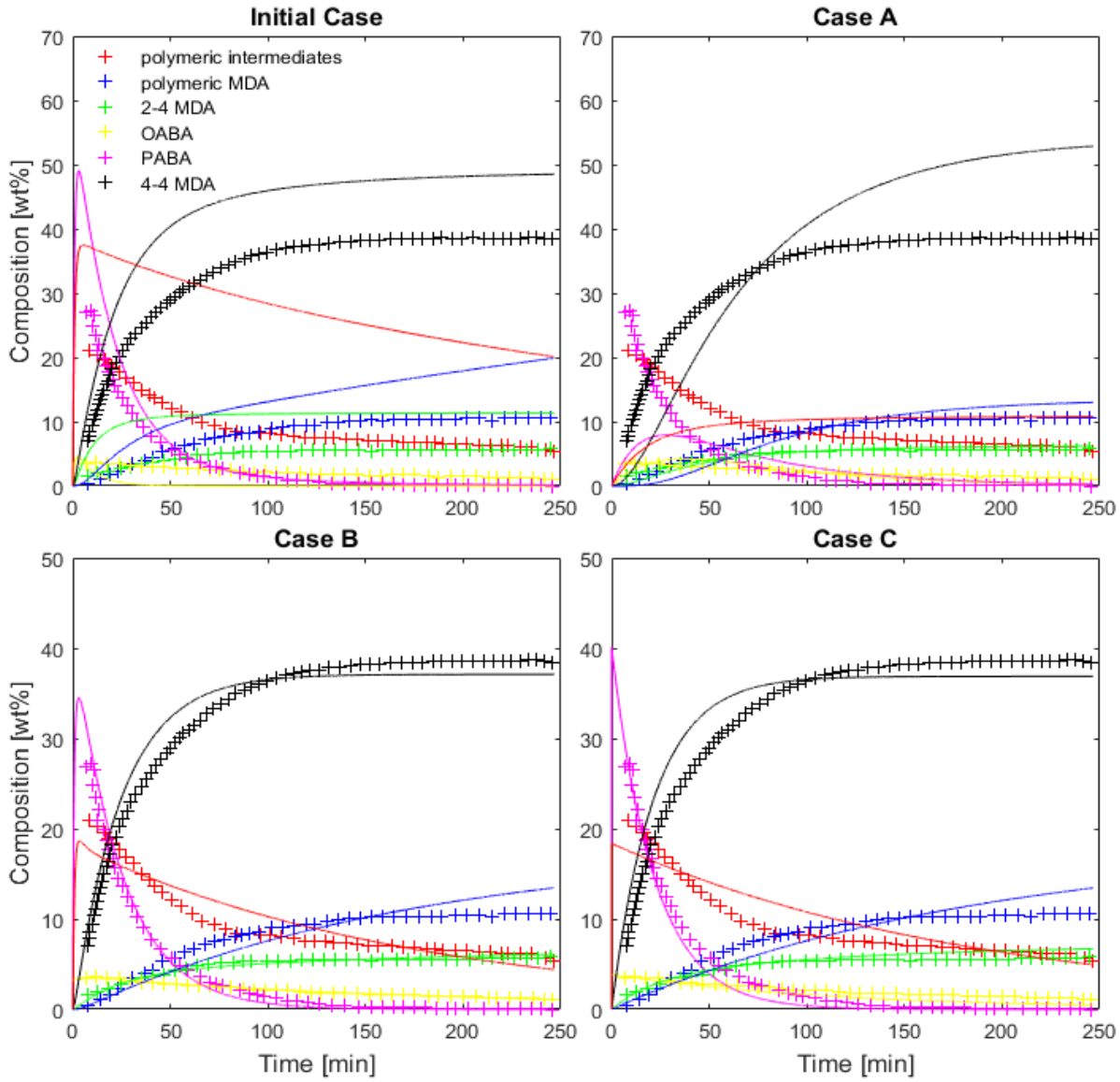


**Figure 30: Comparison of measured and calculated composition of MDA mixture for Cases summarized in Table 5. Continous lines indicate calculated, markers indicate measured values. Reaction conditions: A/F = 3.0, T = 120 °C, t = 4 h, ωcatalyst = 0.16 $g_{catalyst}$ / $g_{solution}$**

In the initial case, the results show that the model is very imprecise when the uncertainties in the kinetic parameters, i.e. the uncertainties in the reaction rate constants defined in Table A1 are neglected. For the concentrations there are significant differences and thus the comparison of measured and calculated KPIs also does not show a good agreement. However, for the variation of concentrations, the characteristics of the individual concentration curves can be observed and are reflected by the model without any refinement.

As a second step in Case A to improve the model and in order to obtain more accurate results, we have considered the uncertainties of the reaction rate constants in the model as lower and upper boundaries for the model parameters fitting as shown in Table A2. This parameter identification task was solved with fmincon function in MATLAB software as previously mentioned.

Case B reflects the results calculated with the optimized model and reaction system. The extended reaction system with our recommended components and reactions were used in order to achieve the best possible fit of the individual concentrations and KPIs compared to the measured data.

As a fourth step in the model development process, the identification of reaction rate constants for the modified and unmodified reactions was performed iteratively. The target of this parameter identification strategy was to halve the number of parameters to be identified in each iteration. In the first step, the reaction rate constants of the modified reactions and, in the second step, the reaction rate constants of the unmodified reactions were identified with the use of the lower and upper bounds defined in Table A2. The bounds defined for Case A were used for the unmodified reactions, while the bounds defined for Case B were used for the modified reactions. The identification was continued step by step until an improvement in the value of the objective function was observed as a result of the re-identified parameters. The identification strategy aimed to reduce the number of parameters to be identified in one iteration step, trying to minimize the correlation between the identifiable parameters. The results of each reaction rate constant after every iteration can be seen in Figure 31.

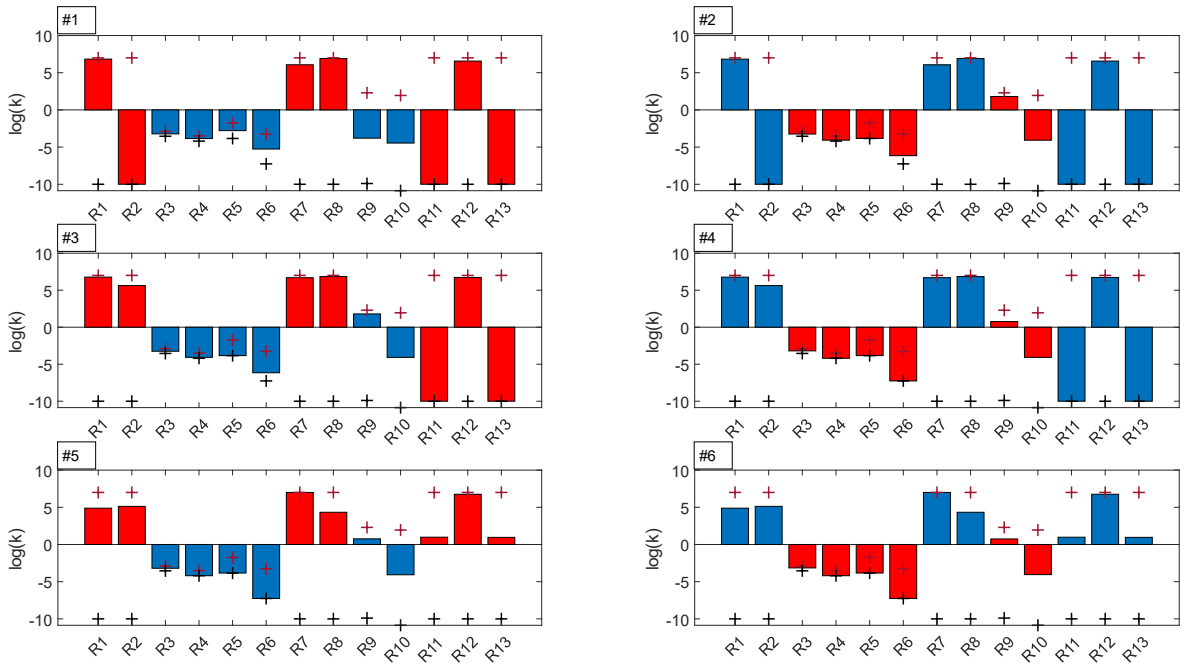


**Figure 31: Identified reaction rate constants after each iteration step. Red bars indicate the reactions that were identified, and blue bars indicate the reactions with fixed reaction rate constants. Brown and black marks highlight the upper and lower brackets of the identification task, respectively**

This iterative identification strategy performed relatively well compared to Case B, but several reactions – for example, R4, R5, and R6 reached their lower boundaries during the identification task, which means these reactions have a lower reaction rate constant than previously identified in the literature and their importance is most probably overqualified. The final value of the objective function was 3.03, which is of the same magnitude compared to the final results summarized in Table 7. In Figure 32, the results of Case B and Case C (i.e. the iterative identification strategy) are compared and plotted.

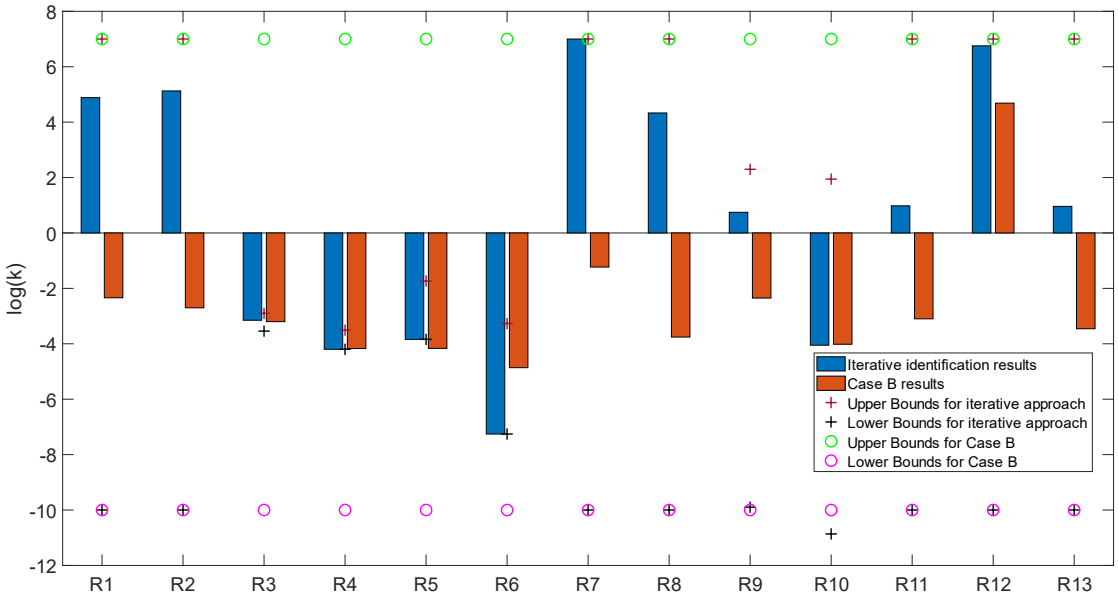


**Figure 32: Reaction rate constants identified for Case B and with the iterative identification strategy (i.e. Case C). Marks and circles are indicating upper and lower bounds for each reaction and identification strategy**

However, it is very important to highlight the differences when comparing some of the reaction rate constants which can affect the final performance and accuracy of the model. In case of reactions R1, R2, R7, R8, and R12 the relatively high *k* values are resulted in the iterative identification procedure can indicate very fast reactions where all the reactants form NMB in the first step, and, as a second step poly-ABA and other intermediates form immediately, which results in very sharp and edgy concentration trajectories, which can be seen in Figure 30, Case C. The identified reaction rate constants obtained by solving the three constrained, non-linear extreme value search problems are collected in Table 6.

**Table 6: Optimized reaction rate constants fitted with component concentrations and with KPIs**

| Reaction | k [$L^{n-1}$ $mol^{1-n}$ $s^{-1}$] | | |
|---|---|---|---|
| | **Case A** | **Case B** | **Case C** |
| **R1** | $1.300 \cdot 10^{-4}$ | $4.570 \cdot 10^{-3}$ | $7.664 \cdot 10^{4}$ |
| **R2** | $5.318 \cdot 10^{-5}$ | $1.989 \cdot 10^{-3}$ | $1.339 \cdot 10^{5}$ |
| **R3** | $1.215 \cdot 10^{-3}$ | $6.299 \cdot 10^{-4}$ | $7.077 \cdot 10^{-4}$ |
| **R4** | $1.999 \cdot 10^{-4}$ | $6.741 \cdot 10^{-5}$ | $6.331 \cdot 10^{-5}$ |
| **R5** | $1.446 \cdot 10^{-4}$ | $6.820 \cdot 10^{-5}$ | $1.446 \cdot 10^{-4}$ |
| **R6** | $5.416 \cdot 10^{-4}$ | $1.373 \cdot 10^{-5}$ | $5.523 \cdot 10^{-8}$ |
| **R7** | $4.484 \cdot 10^{-5}$ | $5.881 \cdot 10^{-2}$ | $1.000 \cdot 10^{7}$ |
| **R8** | $6.671 \cdot 10^{-4}$ | $1.750 \cdot 10^{-4}$ | $2.145 \cdot 10^{4}$ |
| **R9** | $3.272 \cdot 10^{-4}$ | $4.466 \cdot 10^{-3}$ | 5.579 |
| **R10** | $9.864 \cdot 10^{-9}$ | $9.627 \cdot 10^{-5}$ | $8.890 \cdot 10^{-5}$ |
| **R11** | $2.420 \cdot 10^{-3}$ | $7.959 \cdot 10^{-4}$ | 9.537 |
| **R12** | - | $4.854 \cdot 10^{4}$ | $5.673 \cdot 10^{6}$ |
| **R13** | - | $3.487 \cdot 10^{-4}$ | 9.089 |

From the results summarized in Table 6 it can be seen that for the R12 reaction in Case B, and for R7 and R12 reactions in Case C a much higher reaction rate constant was obtained compared to other reactions due to the instantaneous occurrence of these reactions. Comparison of measured and calculated KPIs with each reaction with each model is given in Table 7.

**Table 7: Comparison of measured and calculated KPIs with different reaction rate constants combinations**

| Parameters | Measured | Initial State | Case A | Case B | Case C |
|---|---|---|---|---|---|
| **Remaining Reactants [wt%]** | 2.54 | 0.144 | 0.494 | 1.09 | 0.45 |
| **Oligomeric Fraction [wt%]** | 17.13 | 39.99 | 23.77 | 17.85 | 18.37 |
| **Isomer Ratio [mol·$mol^{-1}$]** | 6.48 | 4.28 | 8.55 | 6.09 | 5.50 |
| **Objective Function** | - | 109.51 | 16.38 | 1.49 | 3.04 |

Beside the used concentration trajectories for parameter identification, Haus *et al.* also published KPIs for different temperatures and A/F molar ratios [114]. Hence, to validate the model developed in Case B we compared the measured KPIs with the calculated ones in different operating conditions. Beside the empirical MDA formation model also published, a so-called population balance model in their referred work, so the measured KPIs can be compared with calculated ones in case the two already existing and published models, and the one proposed in this work (MDA formation model developed in Case B). The comparison was carried out for two of the most important KPIs in terms of product mixture, namely Isomer Ratio (IR) and Oligomer Fraction (OF). The results are shown in Table 8 and Table 9.

**Table 8: Measured and calculated Isomer Ratio values with different models at different A/F molar ratios on constant temperature: T = 120 °C**

| A/F [mol·mol$^{-1}$] | Isomer Ratio [mol·mol$^{-1}$] | | | |
|---|---|---|---|---|
| | **Measured [114]** | **Original empirical [114]** | **Original population [114]** | **Developed model (Case B)** |
| 2.3 | 7.07 | 6.01 | 5.00 | 7.44 |
| 2.7 | 6.73 | - | - | 6.56 |
| 3.0 | 6.48 | 5.66 | 4.50 | 6.09 |

**Table 9: Measured and calculated Oligomeric Fraction values with different models at different A/F molar ratios on constant temperature: T = 120 °C**

| A/F [mol·mol$^{-1}$] | Oligomeric Fraction [wt%] | | | |
|---|---|---|---|---|
| | **Measured [114]** | **Original empirical [114]** | **Original population [114]** | **Developed model (Case B)** |
| 2.3 | 22.92 | 24.62 | 29.18 | 25.91 |
| 2.7 | 17.52 | - | - | 20.72 |
| 3.0 | 17.12 | 21.05 | 21.68 | 17.85 |

Comparing the calculated and measured KPIs, it is clear that a more accurate fit also can be obtained at different operating conditions with the proposed MDA formation model compared to the original empirical and population balance models.

## 4.6 Conclusion

In Section 4, reaction network and reaction kinetics of MDA synthesis was reviewed from literature data for both homogeneous and heterogeneous catalysis. A brief overview of the difficulties encountered in heterogeneous catalytic synthesis, which limit the industrial-scale uptake of heterogeneous catalysts was given.

In order to support and deeply understand the synthesis process, we compared and analyzed the reaction systems proposed in the literature and made several important observations regarding their shortcomings which are resulting loss of information. Based on the information collected and systematized, we performed a reaction system study and developed a model describing it with sufficient accuracy. The difficulties and challenges encountered in using the model were also summarised and, in order to overcome them, further improvements were proposed for the reaction system, the components defined in the model and the reaction rate constants as a result of solving an parameter identification task.

Our goal in the development of the proposed extended MDA formation model was to create one reaction network with robustness, optimal complexity, with all the important reactions, reaction pathways, and components to describe the system with sufficient accuracy. However, due to the limitations of the available and usable measurement data, further measurements and model validation steps and tests are recommended in order to gain a deeper understanding of the proposed reaction network and model performance. In all cases, the values calculated by the developed model achieved a better fit than the original model, enabling the possibility to develop a technological process simulator that can support the industrial-scale MDA synthesis process based on measured data. The proposed MDA formation model should be integrated into an appropiate hydrodynamic model of a real reactor to support the process scale-up or design steps. The model can help in increasing energy and raw material efficiency of the technology and/or speed up the changes in optimal operating conditions to respond for market demands quicker and with relatively low computational requirements. The actual measurement data currently being processed are the results of a laboratory-scale process, but by processing pilot-scale or industrial-scale data, developing an appropriate hydrodynamic model and extending the kinetic model, all the components and reactions which play an important role in MDA synthesis can be described with sufficient accuracy with the developed model. Future researches in this topic can be made to systematically refine the improved model by defining additional componentsand reactions, with a focus on the process of by-product formation (i.e. N-substituted by-products) and the mechanism of formation of higher molecular weight

molecules. The fit of concentration and also the KPIs can be further improved by the extension of the model, which, after the necessary improvements -especially as foreseen in case of the information transfer between laboratory, pilot-scale and industrial level - may be able to resolve the uncertainty in industrial processes, to support and investigate the production process.

# 5 Machine learning model development for MDA formation based on laboratory data

Upon reviewing the existing literature regarding the synthesis of MDA, it is apparent that the influence of individual reaction parameters on the product mixture quality, as well as the strengths of their correlations, remain unexplored. The aim is to discover these correlations and their intensities using the available set of laboratory experimental data, employing relatively straightforward data-driven modeling techniques that can be seamlessly integrated to facilitate industrial-level synthesis processes. Before conducting the laboratory experiments, the synthesis parameters selected for examination during the laboratory synthesis of MDA were those anticipated to exert the most significant impact on product quality. By structuring, preparing, and prefiltering the measured data, we have ensured that our data-driven models will yield realistic and as precise as possible insights into the MDA synthesis process.

The quality properties of MDA mixtures as it was aforementioned include ring distribution, molar ratios of different isomers, quantities of by-products, etc. The operating parameters used for the synthesis of MDA and the quality parameters of the resulting MDA product mixture have a fundamental impact on the properties of the final MDI product mixture, thus the development of mathematical models which can describe the formation of MDA is therefore considered a crucial step. By understanding and mapping the relationships between the independent reaction and dependent quality parameters, controlling and tightly managing the key properties of the MDA mixture, the quantitative and qualitative properties of the final MDI product can be optimised precisely. This allows more flexibility to adapt the production process to market demands, minimise the amount of by-products and maximise the output of higher value products. The developed regression models, based on laboratory experiments, can provide modellers with valuable insights into the direction and strength of correlations between reaction and quality properties, thus helping to understand the technology and to operate it with optimal efficiency with the use of different objective functions.

The key characteristic parameters of the final MDI product are developed during the condensation and rearrangement reactions, making the synthesis of MDA a key step in MDI technology in many respects. The arrangement of the rings, along with the amount and quality of by-products, in addition to the design and specification of other elements defining the product mixture, can be managed by adjusting certain operating conditions, such as the

aniline/formaldehyde ratio (A/F), the HCl/aniline ratio (H/A), temperature, and parameters like residence time.

Currently, there are a limited number of articles available that aim to accurately determine the components, reaction paths and develop a kinetic model for this reaction system and their accuracy is limited mainly due to analytical limitations [4]. Due to the complexity of the reaction system and as shown in the literature, there is no sufficiently accurate and detailed kinetic model available to accurately estimate the effect of the reaction parameters presented on the quality of the product mixture. To overcome the complexity and uncertainties of the system, a new approach could be machine learning, whereby the effects of the presented parameters can be estimated with good accuracy, using appropriate training data. Furthermore, at the industrial level, data from many more operational parameters can be used to estimate the quality of the product mix, thus further refining the machine learning models. Moreover, compared to classical first-principle-based process models, ML models have much lower computational requirements and can be used to monitor technology and product quality in real time, allowing timely intervention if product quality starts to deteriorate, thus saving costs. These costs are shown as follows at the technology level.

In terms of cost reduction, the optimisation of parameters can have a significant impact, since by reducing the A/F ratio, the amount of aniline introduced into the system can be reduced, thus significantly reducing the energy consumption (hot and cold energy) of the purification section where we recover the surplus aniline from the system. By reducing the H/A ratio, the amount of HCl added as catalyst can be reduced, therefore less HCl has to be neutralised with NaOH solution at the end of the condensation and rearrangement reactions, which results in a significant cost reduction and, of course reduced salty wastewater generation. As a result, after the neutralisation section, the washing section can be optimised as well, resulting in less amount of diluted brine formation, which can reduce the energy consumption of the wastewater treatment section. By optimising the reaction temperatures and reducing the residence times, thermal energy and electrical energy can also be saved, or even the plant product output can be increased. By reducing the formation of by-products, the yield of the product mixture can be increased, thus basicly reducing the specific cost of all products. By understanding, exploring and quantifying these effects on the quality of the product mixture and the effects on the operational costs, objective functions can be formed that can determine the optimal set of operational parameters, taking into account product yield, product distribution and energy costs.

In case of a posteriori (also known as black-box) models, the structure of the model does not need to reflect the structure of the real system, its parameter set does not need to have real physical or scientific content, and the relationships between parameters are not described by scientific relationships. In this work, three different modelling methods are presented that can be used to construct black box models and that can be used to describe MDA formation and to estimate the properties of the resulting MDA mixture:

- Linear Regression;
- LASSO – method;
- Artificial Neural Networks.

The aim of this study and the developed models is to support real industrial MDA synthesis process that can be easily deployed as simple tools to be able to achieve an optimal product portfolio and operating cost for the industrial process. Based on the collected laboratory data related to the formation of MDA the mapping of correlations between reaction and quality parameters, reviewing different regression modeling techniques which are developed based on these laboratory experiments, validation and evaluation of their performance in terms of their accuracy in estimating the properties of the resulting MDA product have been performed based on different reaction parameters. By constructing and using these different regression models with appropriate accuracy, it is possible to optimize the production processes with the use of different objective functions, to map and quantify the correlations and their strength between the independent reaction and dependent quality parameters of the MDA formation process, thus to be able to define an optimal parameter set for the synthesis process and to gain an overall deeper understanding. With the mapped and specified correlations, it becomes possible to optimise the quantitative and qualitative properties of the final MDI product itself by precisely defining the optimal reaction parameter set. In this work, our aim is to solve the following tasks:

- collected and analysed the available laboratory data;
- identified and quantified the correlations between independent and dependent variables;
- developed reliable regression models to predict the quality parameters of MDA;
- optimal results were obtained by adjusting the parameters of the regression models and compared the results achieved with each model structure;
- suggestions were identified to further improve the performance of the models.

## 5.1 Methodology

The regression models presented in Chapter 2 are used in order to support the understanding of MDA synthesis process quickly and accurately, increase product selectivity to meet market demands, reduce by-products and energy consumption, and improve the setting of key quality paramteres to expected values. The steps of model development process performed in this chapter are summarised in Figure 33.

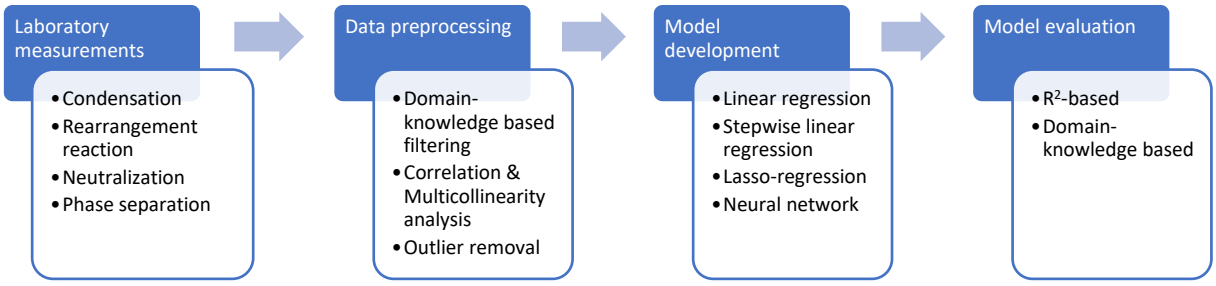


**Figure 33: Process flowchart of the performed data analysis**

## 5.2 Results and discussion

### 5.2.1 Laboratory MDA synthesis process

The laboratory data I used not personally measured by me, but were provided for me to perform the modeling tasks presented in this chapter. The condensation and rearrangement reactions were carried out earlier in a 250 $cm^3$ duplicated glass reactor equipped with an adjustable speed stirrer. The reactor temperature was controlled by a thermostat and the dosing of formaline was carried out with a Metrohm Dozimat 765 programmable dosing unit. The reaction mixture during the condensation, rearrangement, neutralization and phase separation processes is shown in Figure 34-37 and were carried out according to the following steps:

Condensation process:

1. A fixed amount of aniline is introduced into the reactor.
2. The hydrochloric acid was then measured in, resulting aniline hydrochlorid. The released heat was removed with the use of the thermostat. The temperature of the reaction mixture was allowed to rise only for a predefined temperature (Condensation reactor temperature - $T_c$).
3. The exact amount of formalin was measured with the use of the controllable dosing device, where it is also important to maintain the prescribed temperature.
4. Mixing of reaction mixture for a predefined time with temperature control.

Rearrangement process:

1. The reaction mixture was heated up to allow the rearrangement process to take place as the next step of the reaction (Rearrangement reactor temperature - $T_r$).
2. Mixing of reaction mixture for a predefined time with temperature control.
3. Formation of O-, P-ABA and MDA molecules.

**Figure 34: Rearrangement process**

**Figure 35: Neutralization process**

Neutralization process

1. After the rearrangement process, the neutralization of the added acid is necessary. The neutralization process took place on higher temperature with the addition of 49 wt% NaOH solution.
2. Mixing of reaction mixture for a predefined time with temperature control.
3. Phase separation after the mixing.
4. NaOH and NaCl were washed from the neutralized and separated organic phase with warm water.

**Figure 36: Mixing during neutralization**

**Figure 37: Phase separation**

After the reaction was stopped and the reaction system reached steady state, the separation of the organic phase was completed and the analytical tests were carried out by liquid chromatography to determine the ring distribution of the product mixture, while gas chromatography was used to determine the mono-methyl MDA by-product content and identify isomer ratios. The main data of the chromatographs and experiment conditions were as shown in Table 10:

**Table 10: Main parameters of the chromatographs used in the experiments**

| Parameter | Liquid chromatograph | Gas chromatograph |
|---|---|---|
| **Chromatograph** | Waters Alliance 2690 | Agilent 7890 |
| **Column** | 2 pcs Polymer Laboratories PL gel 3 μm 100 Å 300 · 7.5 mm in line | HP-5MS 30m / 0.25mm / 0.25um |
| **Eluent** | THF | Ethanol |

### 5.2.2 Data preprocessing

As a first step of the data preperation, we have preprocessed the laboratory test data available related to the synthesis of MDA with MATLAB, which were carried out and evaluated in the late 2000s. This dataset was measured on laboratory scale with the purpose to better understand and be able to perform deeper analyses on the MDA synthesis process and its main parameters. The measured reaction parameters during the tests, i. e. independent variables, which will be used as 10 input vectors during the model preparation were either:

- molar ratio type quantities (aniline/formaline (A/F), hydrochloric acid/aniline (H/A), water/aniline (W/A));
- temperature type measurements (condensation and rearrangement temperatures);
- residence or reaction times (condensation and rearrangement reaction times), and lastly;
- formaline feeding time and feeding ratios between the reactors.

Among the measured parameters which are describing the quality of the final product, the following categorization can be done for the 9 dependent or output quality variables:

- ring distribution of the MDA mixture (2-ring content, 3-ring content, …>6-ring content);
- isomer ratios (ortho-ortho, ortho-para, para-para);
- by-product quantity (mono-methyl-MDA (MMM)).

The data from a total of 46 individual experiments were used as the basis of the analysis. During the preliminary collection of the data, it was clearly noticable that not all parameters of the dependent variables, i.e. the output data, were recorded for all experiments. Each set of experiments was designed and carried out for a different purpose or to understand different operating parameters better, so they had to be interlinked into the same dataset. The reason for this is that these laboratory experiments have been compiled and summarised from several series of experiments, therefore these laboratory tests were not designed to define the same set of dependent parameters and there was also a difference in time between each series of experiments. After the examination of the data, it became clear that some method of dealing with unmeasured or missing parameters would be necessary during the data analysis task. Table 11 summarizes all the available input and output parameters which were measured during the laboratory experiments.

**Table 11: Available independent and dependent parameters from the laboratory experiments**

| Symbol | Parameter | Type of parameter |
|---|---|---|
| **A/F** | Aniline / Formaline molar ratio | **Independent** |
| **H/A** | Hydrochloric acid / Aniline molar ratio | |
| **W/A** | Water / Aniline molar ratio | |
| **$T_c$** | Condensation reactor temperature | |
| **$T_r$** | Rearrangement reactor temperature | |
| **$t_c$** | Condensation reaction time | |
| **$t_r$** | Rearrangement reaction time | |
| **$F_{R1}$** | Formalin feed ratio to the first reactor | |
| **$F_{R2}$** | Formalin feed ratio to the second reactor | |
| **$t_F$** | Formalin dosing time | |
| **2r** | 2, 3, 4, 5, ≥ 6-ring MDA content, respectively | **Dependent** |
| **3r** | | |
| **4r** | | |
| **5r** | | |
| **>6r** | | |
| **O-O** | Ortho-ortho isomer ratio | |
| **O-P** | Ortho-para isomer ratio | |
| **P-P** | Para-para isomer ratio | |
| **MMM** | Mono-methyl MDA (MMM) by-product content | |

After collecting and structuring the data, the location of the missing data was highlighted. The missing data and their place can be seen in Figure 38, where green color indicates available data and the red color indicates missing data for a defined parameter at a defined location.

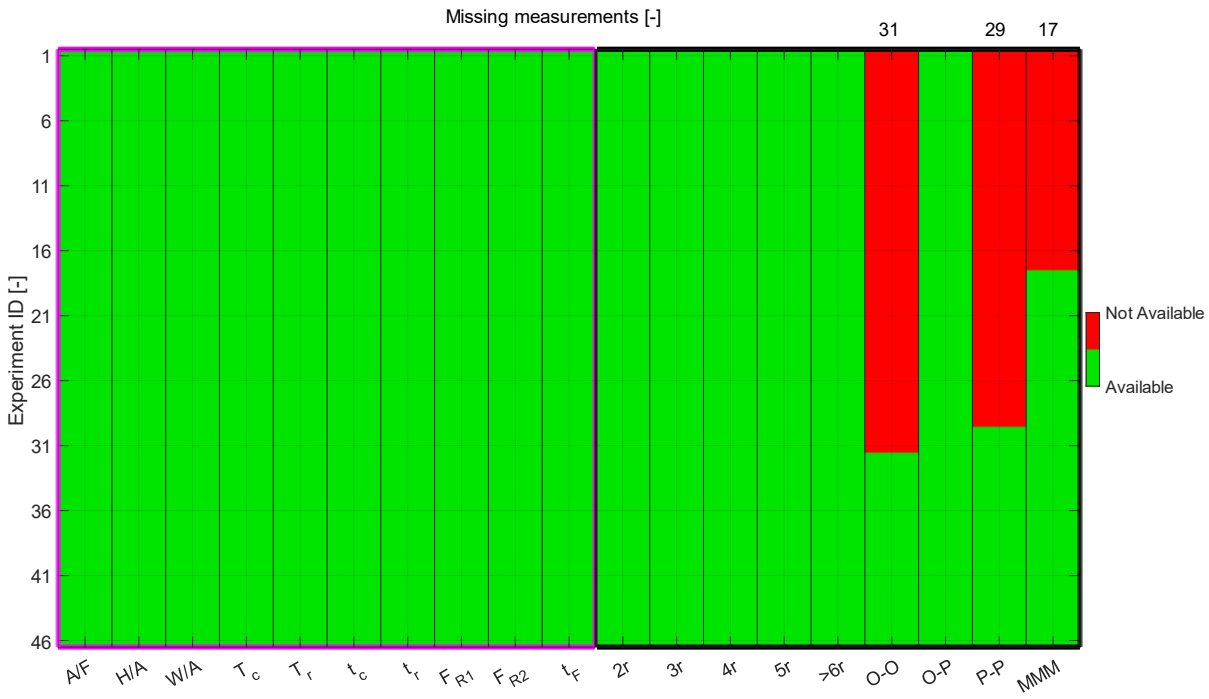


**Figure 38: Locations of available and missing parameters for laboratory experiments**

In Figure 38 the parameters on the X axis from No. 1 (A/F – Aniline / Formaline molar ration) to 10 ($t_F$ - Formaline dosing time) represent the following input parameters, which have been used as independent variables. The dependent variables parameters have been also plotted in Figure 38 from No. 11 (2r – 2-ring content) to 19 (Mono-methyl MDA – MMM content). The numbers at the top of the x-axis represent the number of missing measurements, i.e. the number of laboratory measurements of the given parameter for which the given variable was not measured. This is significant because there are 46 experiments in total, although the number of measurements that can be used for the prediction of different parameters may vary.

To avoid any numerical problems during the data processing, an automated algorithm was implemented to remove any experiments completely where missing values are present from the dataset. For example, if 1 out of the 46 experiments does not contain information on MMM concentration because it was not measured during that test, therefore only the other 45 experiments will be used from the laboratory data as the basis of the data analysis process, the other experiments which contained the missing data will be left out for that training step. This prefiltering step ensures that only those datasets for which the given parameter set is actually available are used in each model training step. By carrying out these preliminary tests, it is also possible to summarize the quantity and quality of the available data.

The dataset were preconditioned, i.e. normalised between 0 to 1, where 0 equals the minimum and 1 equals the maximum value of the measured value set for one parameter. The minimum and maximum values used in the normalisation for each input and output data are not presented for reasons of information protection, and hence the upper and lower bounds on the interpretation ranges of the models that will be created.

With the checking of the available data, it is also possible to gain important information about mean, median, variance and basicly the quality of the data in terms of usability for mathematical modeling by plotting them on violin charts. The data plotted on violin charts can be seen in Figure 39 for each input and output parameters.

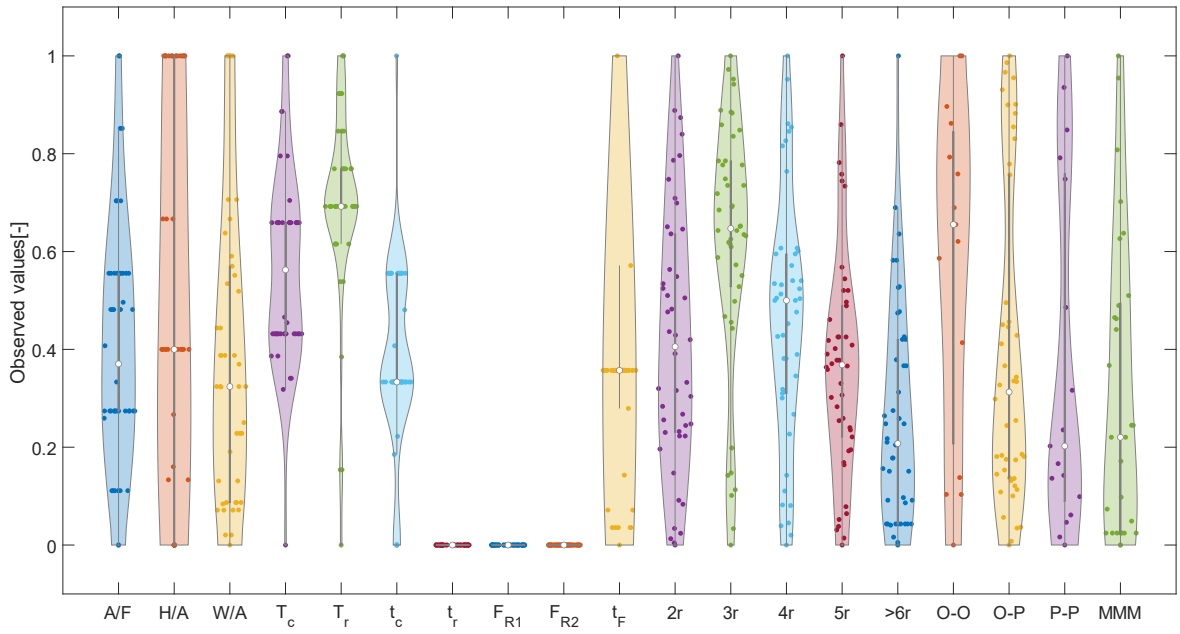


**Figure 39: Availabe dataset from laboratory experiments plotted on violin chart**

After evaluating the available measurement data, it became clear that while the data overall shows adequate variability for use in training data models, there are three specific input vectors—the rearrangement reaction time ($t_r$), the formalin feed ratio for the first reactor ($F_{R1}$) and the formalin feed ratio for the second reactor ($F_{R2}$) — that lack enough variance to serve as effective training inputs. These parameters, since they were not changed during the experiments, do not carry any information on the output parameters and their use is therefore redundant. These three input vectors were deleted before the next data preparation steps. Similarly, by examining the data available as dependent variables, i.e. output vectors, it can be stated that these data are available in sufficient quantity and quality, and therefore all of these vectors have been retained for further analysis.

### 5.2.3 Correlation analysis

Following the initial phases of data collection, which included preprocessing, cleaning, and filtering, a correlation analysis was performed. Pearson correlation coefficients were calculated to assess the relationships between the input and output data. The outcomes of this analysis are illustrated in Figure 40 and Figure 41.

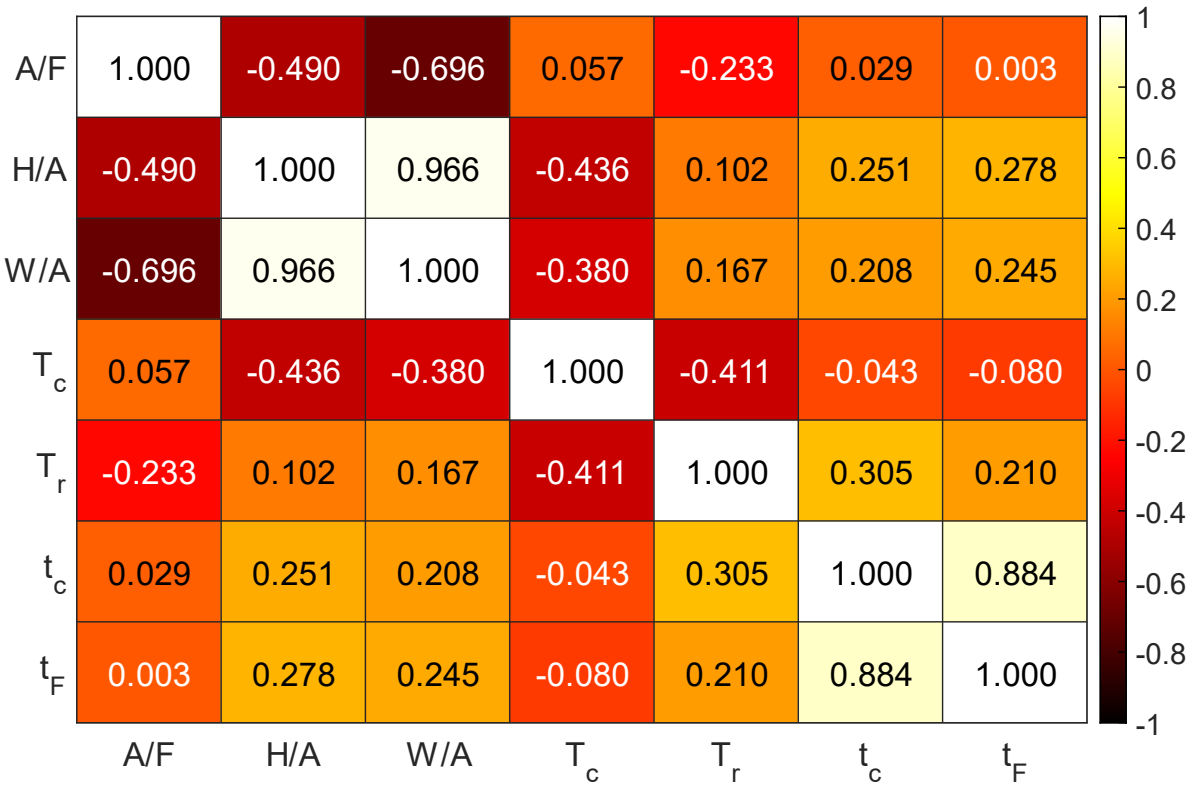


**Figure 40: Correlation coefficients between each independent parameters**

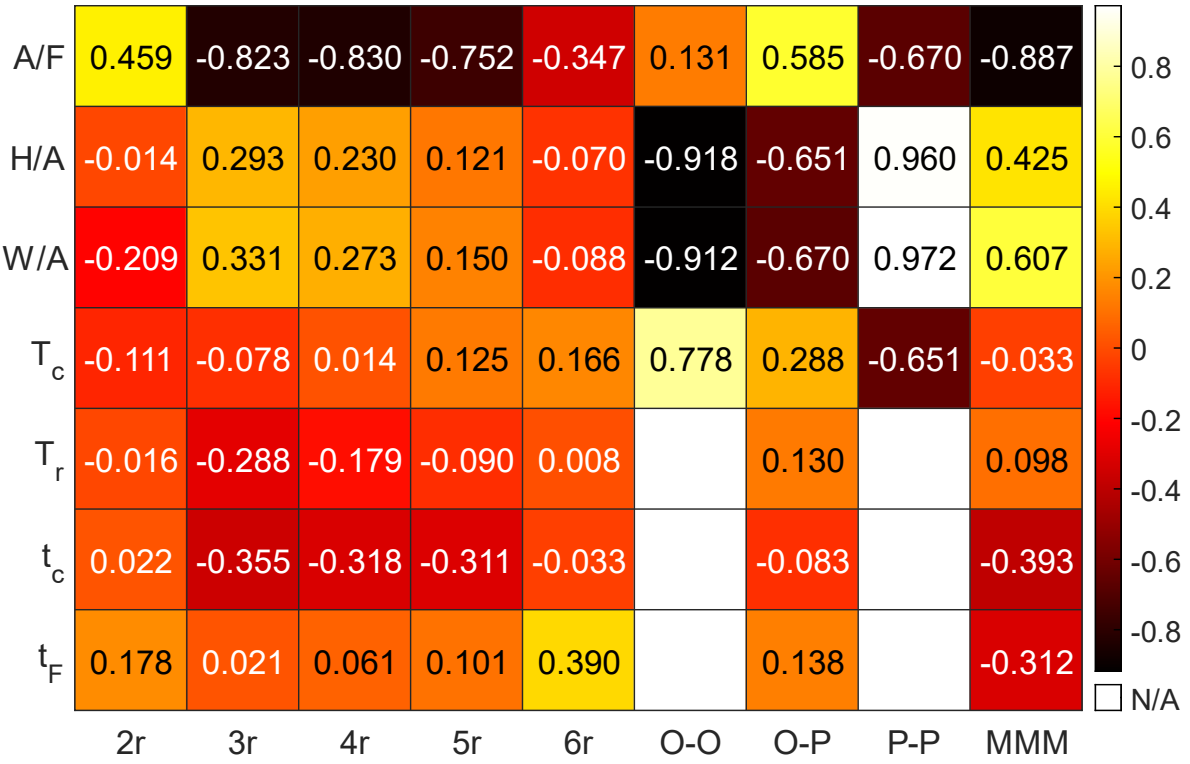


**Figure 41: Correlation coefficients between independent and dependent parameters**

The correlation coefficients in Figure 40 show the direction and strength of the relationships between each input parameters, while Figure 41 show the relationship between each input and output parameters. For some parameters these correlation coefficients cannot be determined with the available dataset. This is due to the fact that the algorithm which was developed to avoid numerical problems deletes experiments with missing parameters that would ensure the variance of the given input vector, i.e. for a given input parameter the given output parameter becomes insensitive, the function is not able to determine correlations between them. That means the input parameters $T_r$, $t_c$ and $t_F$ were kept constant during those experiments, where there are availabe data for O-O and P-P output parameters, i.e. where they were measured and for these six correlations the correlation coefficients cannot be determined.

As shown in Figure 40 and as expected, there is a very strong correlation between the W/A and H/A input parameters, which is due to the fact that most of the total water entering the reaction system is introduced with the HCl solution used as catalyst. The reason why this correlation is not equal to 1 is that in many experiments, in addition to the water introduced with the HCl solution, extra water was added to the reaction system to study and better understand its effect on the quality parameters. To verify the multicollinearity between the W/A

and H/A parameters, the so-called LASSO – method was also applied before preparing the basic regression models for outlier identification.

### 5.2.4 Multicollinearity analysis

In the examination of multicollinearity, the LASSO – method was executed utilizing input data spanning 100 distinct lambda values to evaluate potential multicollinearity among input parameters before the formulation and validation of the final regression models. During the shrinkage process, the model parameters and their fluctuations for each lambda value were documented, along with the MSE values computed during parameter reduction. By observing the parameter shrinkage, it is possible to verify the existence of previously identified collinearity between the W/A and H/A input parameters. This method was applied specifically to the 2-ring content among the output parameters, but the shrinkage was recorded and summarized for all input parameters. Utilizing the LASSO regression with the existing data to predict the 2-ring MDA content resulted in the outcomes illustrated in Figure 42.

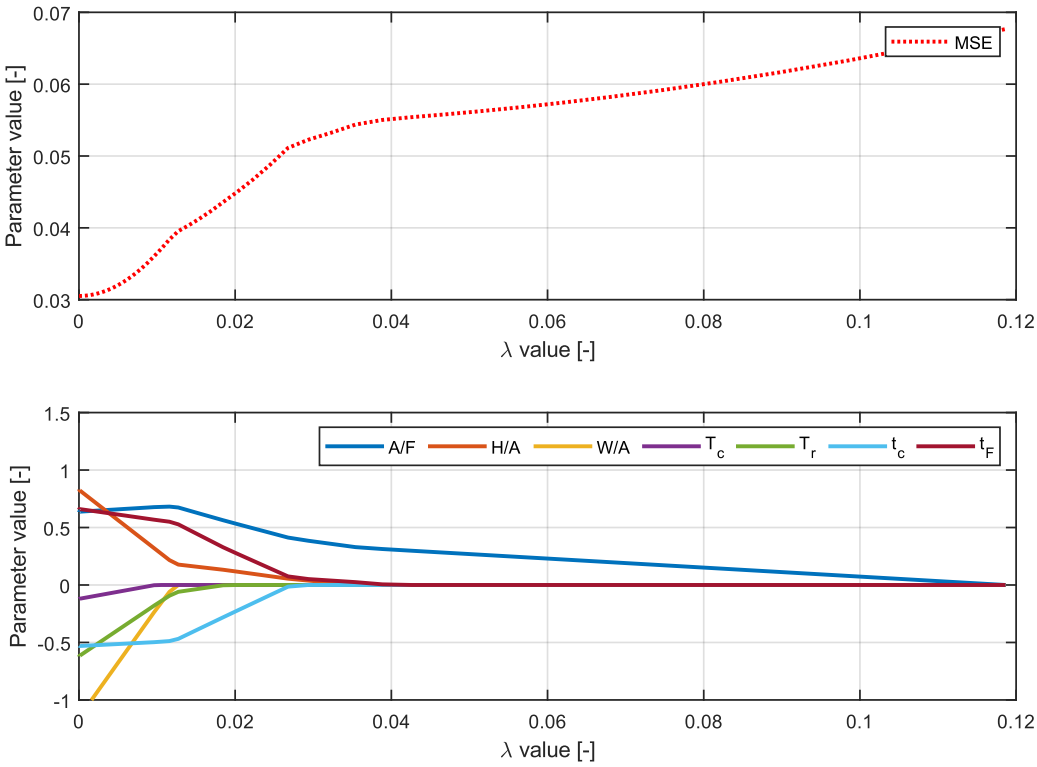


**Figure 42: Result of LASSO regression with different λ values for 2-ring MDA content estimation**

As illustrated in Figure 42, with a continuous increase in the value of the λ parameter, a greater number of parameters are subjected to shrinkage, exhibiting a more pronounced effect. The sequence of parameter shrinkage is detailed in Table 12.

**Table 12: Order of shrinkage performed by LASSO regression for 2-ring MDA content**

| **Order of shrinkage** | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
|---|---|---|---|---|---|---|---|
| **Parameter** | W/A | H/A | $t_c$ | $T_r$ | $T_c$ | $t_F$ | A/F |

As previously expected, using a linear design matrix for the LASSO regression the W/A parameter is shrunk and set to zero first, which confirms the strong correlation with the H/A parameter in Figure 40. However, the studying of this W/A parameter is important because, at the industrial level, the unreacted aniline which is recovered from the process may contain small amounts of water, so understanding the effect of this extra water is relevant, thus the W/A parameter was kept in the data set.

### 5.2.5 Outlier identification

Following the application of the LASSO – method and the assessment of multicollinearity among the input parameters, methods for outlier detection were employed to identify inconsistent data from laboratory experiments, thereby ensuring the optimal performance of the regression model structures and their outcomes.

For identifying potential outliers, the Leave-One-Out validation method was employed utilizing basic regression models, notably without hyperparameter optimization, to discern outliers from experiments that had undergone normalization, prefiltering, and cleaning of missing data. This phase occurred prior to further training steps and before the determination of optimal model structures. The fundamental concept of this cross-validation approach lies in training the model on n-1 experiments while validating it with one omitted experiment, and this process is reiterated for a total of n iterations, ensuring that each experiment serves as a validation sample. While models of this nature may exhibit significant overfitting, the comparison of absolute errors during validation facilitates a credible identification of outliers. To achieve optimal accuracy in outlier identification, the utilization of various models and the inclusion of all available output parameter sets is advocated. This procedure was executed employing three distinct model types: linear regression with a 'linear' design matrix, linear regression with a 'stepwise' design matrix, and an artificial neural network. This was strategic in ensuring that the identified outliers were consistent across these models or to detect any potential discrepancies. LASSO regression, being solely used for multicollinearity analysis and considered a variant of linear regression, was not utilized for outlier identification.

In this work, a three-level hierarchical rule set have been defined to be able to identify experiments as outliers:

- During the training and validation of regression models, a measurement for a given output was identified as an outlier or non-outlier using the IQR method (Output outlier) [122].
- If an experiment contained at least three output parameters which are identified as potential outliers, that experiment for that specific regression model is considered as a potential outlier (Model outlier).
- If an experiment is identified by at least two regression models as an outlier, then the experiment is indeed identified as an outlier and will be deleted from the available dataset (Experiment outlier).

In cases where the three specified outlier detection conditions were not satisfied, the experiment was not classified as an outlier and, consequently, was included in the final dataset. The outcomes derived from the various regression models employed for outlier identification are depicted in boxplots in Figure 43 below.

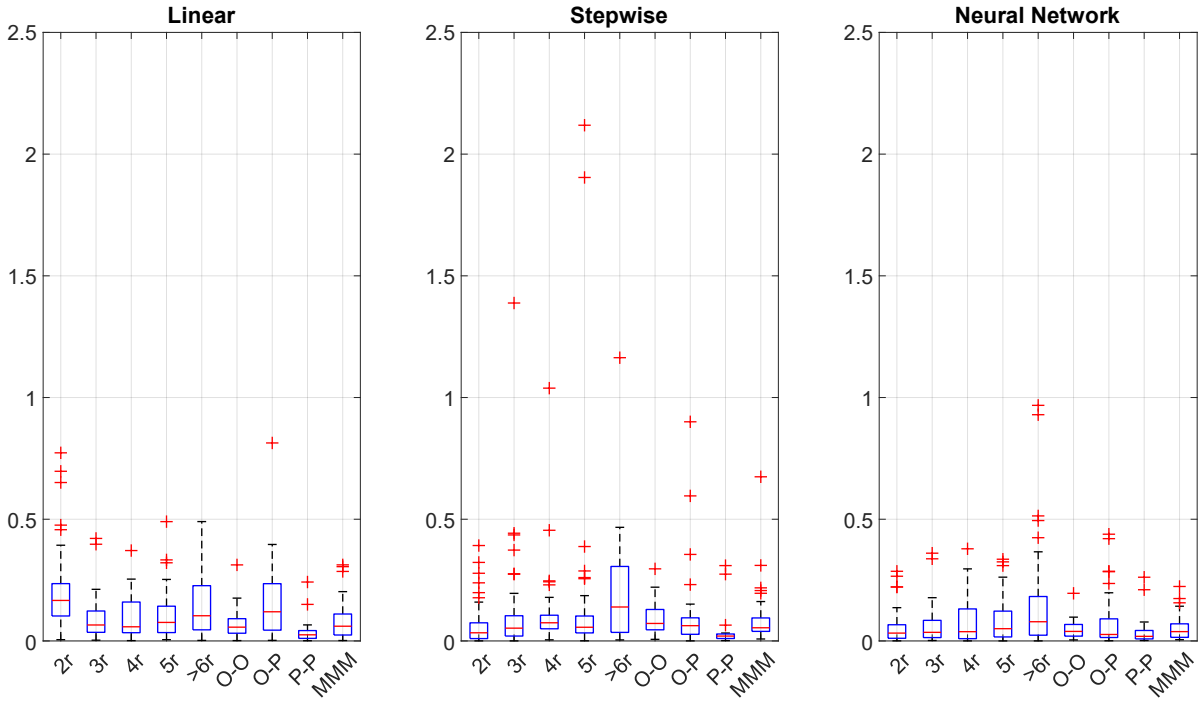


**Figure 43: Absolute errors plotted on boxplot diagrams for different regression models (Neural Network structure: 30-30-30, ReLU, λ =10-4)**

From the analysis of absolute error results depicted in Figure 43, it is evident that while the standard deviation of errors remains relatively low for both the linear design matrix

regression model and the neural network model, the absolute error values exhibit considerable dispersion for the stepwise linear regression model. This observation indicates that the standard deviation for this model type is also comparatively higher relative to the other models. Such an outcome may rise from the possibility that the stepwise linear regression, characterized by an extensive range of hyperparameters, could exhibit heightened sensitivity to overfitting due to its structural components, thereby potentially leading to higher validation errors and greater variance. Figure 44 illustrates not only the total number of outliers but also their specific positions, identified by the serial numbers of potentially anomalous experiments.

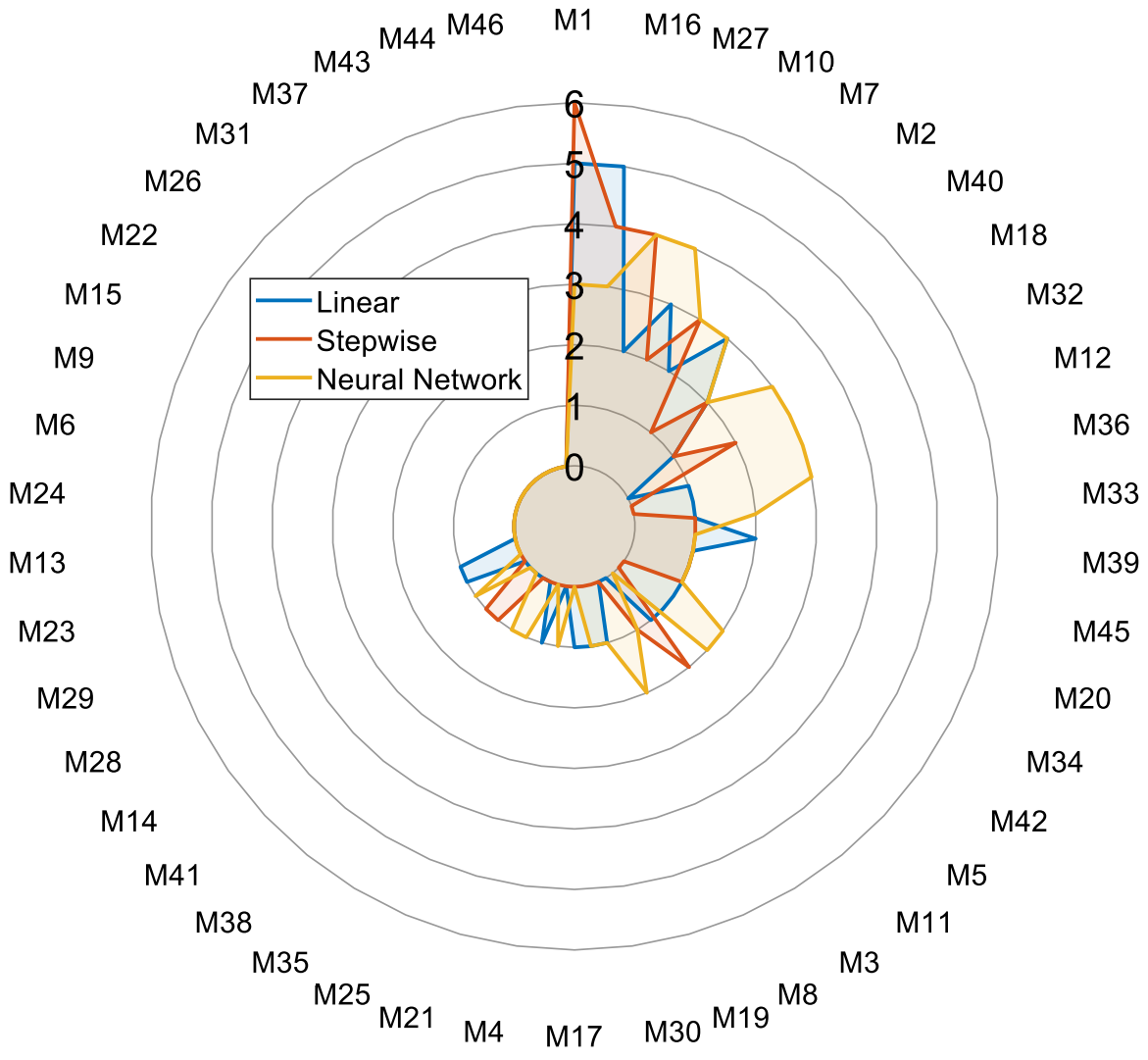


**Figure 44: Number of identified potential outlier outputs for different measurements in case of different regression models**

According to the data presented in Figure 44, it is obvious that the three previously mentioned criteria, specifically output, model, and measurement outlier conditions, are satisfied

by only six experiments. The identifiers for these experiments are M1, M16, M27, M10, M7, and M2, indicating that they have been permanently excluded from the dataset.

## 5.3 Model performance evaluation

Upon comprehensively understanding and preliminarily processing the data with foundational models, it becomes feasible to optimize, evaluate, and conduct comparative analysis of model performance to ascertain the optimal solution and model configuration. Throughout this comparative analysis, additional insights can be collected concerning the reliability of measurement data, the identification of outliers, and model performance, which aids in clarifying the true connections between input and output variables. Consequently, practical expertise and industrial knowledge can be quantified with precise accuracy, assisting in the determination of the optimal parameter set for the manufacturing process. To mitigate the risk of overfitting the trained models, the K-fold cross-validation technique was employed. This method involves partitioning the available data into K segments. From these segments, K-1 portion is designated as the training dataset, while the remaining segment serves as the validation dataset for the trained model. Subsequently, this process is iterated K times, each time selecting a new K-1 portion for training and a new segment for validation, thereby precluding overfitting and facilitating the determination of optimal model parameters. A general 3-Fold cross-validation was applied across all models.

Linear regression models employing a 'linear' design matrix were constructed for the dataset as the most fundamental structured model. Subsequently, optimizable stepwise linear regression models were developed, followed by the preparation of neural network models. Each model was tailored to an individual output parameter vector, thereby necessitating the training and validation of different model for each quality parameter. Following the outlined procedures and data preparation steps, three separate models were successfully designed, trained, and validated.

- Linear Regression with ’**linear’** design matrix;
- Linear Regression with ’**stepwise**’ regression;
- Artificial Neural Network

### 5.3.1 Linear regression

Linear regression models define the relationships between input and output variables using design matrices. The most basic design matrix is the linear model with a constant term, which helps assess the proportion of variance in the dataset that is captured by linear

correlations. In linear systems, outputs change in direct proportion to inputs, whereas non-linear systems exhibit more intricate relationships that cannot be characterized by straightforward linear equations. Table 13 displays the results for all output parameters across the linear regression models.

**Table 13: Average $R^2$ values calculated with linear regression models with linear design matrix for different dependent variables with 3-fold cross-validation**

| Dependent variable | $\bar{R}^2_{training}$ | $\bar{R}^2_{validation}$ |
|---|---|---|
| **2r** | 0.601 | -0.281 |
| **3r** | 0.916 | 0.778 |
| **4r** | 0.909 | 0.658 |
| **5r** | 0.913 | 0.278 |
| **>6r** | 0.681 | -1.705 |
| **O-O** | 0.953 | 0.881 |
| **O-P** | 0.770 | 0.378 |
| **P-P** | 0.976 | 0.803 |
| **MMM** | 0.958 | 0.434 |

The results indicate that a strictly linear model falls short in capturing the connections between the independent and dependent variables; hence, adopting a model that includes non-linear relationships is certainly beneficial. Nonetheless, the most basic linear model can still reasonably accurately describe various dependent variables like O-O, P-P, and 3- and 4-ring content.

### 5.3.2 Stepwise linear regression

To address the limitations of basic linear regression models and develop a model with ideal complexity and satisfactory accuracy for the available data, the approach and effectiveness of stepwise linear regression were explored. Stepwise linear regression is advantageous because it automates the optimization of the model structure based on predefined criteria, thereby creating the optimal model configuration for a specific set of hyperparameters. The algorithm used can optimize various objective functions and hyperparameters such as minimizing the sum of squared errors, enhancing the adjusted $R^2$ value, maximizing the $R^2$ value, or optimizing based on Akaike or Bayesian information criteria. In this scenario, the optimization aimed to maximize the $R^2$ value, and we configured the stepwiselm function to achieve optimal results

for both the training and validation datasets. To accomplish this, the function was allowed to incorporate any terms, such as constant, linear, interactions, quadratic, etc. Additionally, the $P_{Enter}$ and $P_{Remove}$ parameters were meticulously adjusted to identify optimal function parameters, which serve as the threshold criteria for adding or removing terms from the final model structure. The training and validation outcomes for 2-ring MDA content with different $P_{Enter}$ and $P_{Remove}$ settings are presented in Table 14, while results for other dependent variables obtained with various parameter combinations are available in the Supporting Information in Table A3 - A12.

**Table 14: Average $R^2$ values calculated with stepwise linear regression models for 2-ring MDA content with 3-fold cross-validation and different $P_{Enter}$ and $P_{Remove}$ parameters**

| $P_{Enter}$ | $P_{Remove}$ | $\bar{R}^2_{training}$ | $\bar{R}^2_{validation}$ |
|---|---|---|---|
| **0.1** | **0.05** | 0.580 | -0.092 |
| **0.05** | **0.025** | 0.867 | 0.797 |
| **0.04** | **0.02** | 0.867 | 0.797 |
| **0.03** | **0.01** | 0.867 | 0.797 |
| **0.027** | **0.01** | 0.913 | 0.775 |
| **0.025** | **0.01** | 0.913 | 0.775 |
| **0.025** | **0.005** | 0.913 | 0.775 |
| **0.01** | **0.005** | 0.970 | 0.723 |

The analysis indicates that by decreasing the $P_{Enter}$ parameter, the $R^2$ value obtained from the model on the training dataset can be progressively enhanced, as additional terms are incorporated by lowering the threshold, thereby refining the model's accuracy. Concurrently, the validation $R^2$ also increases systematically, and the model's performance on the validation dataset improves due to more effective training. However, upon surpassing a certain threshold, the model exhibits heightened vulnerability to overfitting, resulting in a further increase in the training $R^2$, while the validation $R^2$ begins to decline markedly. Upon review of the results, the parameters are established at $P_{Enter} = 0.03$ and $P_{Remove} = 0.01$ as the optimal settings. Sensitivity analyses were also conducted for the other dependent parameters and various hyperparameter configurations, which are detailed in the supporting information. The superior outcomes of the stepwise linear regression models are presented in Table 15, employing the optimized function parameter set.

**Table 15: Average $R^2$ values calculated with stepwise linear regression models for different dependent variables with 3-fold cross-validation and optimized $P_{Enter}$ and $P_{Remove}$ parameters**

| Dependent variable | $\bar{R}^2_{training}$ | $\bar{R}^2_{validation}$ |
|---|---|---|
| **2r** | 0.867 | 0.797 |
| **3r** | 0.837 | 0.673 |
| **4r** | 0.845 | 0.647 |
| **5r** | 0.919 | 0.484 |
| **>6r** | 0.388 | -0.519 |
| **O-O** | 0.950 | 0.887 |
| **O-P** | 0.925 | 0.416 |
| **P-P** | 0.945 | 0.480 |
| **MMM** | 0.922 | 0.893 |

### 5.3.3 Neural networks

Various activation functions, structures according to the neural network structure shown in Figure 8 and λ parameters were examined during the training phase of neural networks to provide the most precise outcomes in the synthesis of MDA. These evaluations were conducted to represent all dependent parameters with the highest degree of accuracy, subsequently identifying and applying the optimal structures and hyperparameter configurations deemed most effective. It is evident that the accuracy of neural networks can be significantly enhanced through structural and hyperparameterization tests tailored for different dependent variables. Consequently, the results of structure and hyperparameter optimization for additional dependent parameters are comprehensively summarized in the supplementary material. Table 15 illustrates the findings for 2-ring MDA content, while the outcomes for other dependent variables utilizing different parameter sets are detailed in the Supporting Information found in Tables A13-A22.

**Table 16: Performance of neural networks with different hyperparameterizations for 2-ring MDA content with 3-fold cross-validation**

| Structure | Act. Fcn. | λ | $\bar{R}^2_{training}$ | $\bar{R}^2_{validation}$ |
|---|---|---|---|---|
| 10-10-10 | ReLU | $10^{-4}$ | 0.999 | 0.813 |
| 10-10-10 | Tanh | $10^{-4}$ | 0.998 | 0.418 |
| 10-10-10 | Sigmoid | $10^{-4}$ | 0.994 | 0.701 |
| 10-10-10 | ReLU | $10^{-3}$ | 0.996 | 0.886 |
| 10-10-10 | Tanh | $10^{-3}$ | 0.992 | 0.707 |
| 10-10-10 | Sigmoid | $10^{-3}$ | 0.000 | -0.045 |
| 20-20-20 | ReLU | $10^{-4}$ | 1.000 | 0.870 |
| 20-20-20 | Tanh | $10^{-4}$ | 0.998 | 0.626 |
| 20-20-20 | Sigmoid | $10^{-4}$ | 0.994 | 0.703 |
| 20-20-20 | ReLU | $10^{-3}$ | 0.997 | 0.885 |
| 20-20-20 | Tanh | $10^{-3}$ | 0.990 | 0.702 |
| 20-20-20 | Sigmoid | $10^{-3}$ | 0.990 | 0.702 |
| 30-30-30 | ReLU | $10^{-4}$ | 1.000 | 0.845 |
| 30-30-30 | Tanh | $10^{-4}$ | 0.998 | 0.612 |
| 30-30-30 | Sigmoid | $10^{-4}$ | 0.994 | 0.703 |
| 30-30-30 | ReLU | $10^{-3}$ | 0.997 | 0.891 |
| 30-30-30 | Tanh | $10^{-3}$ | 0.991 | 0.712 |
| 30-30-30 | Sigmoid | $10^{-3}$ | 0.000 | -0.045 |
| 40-40-40 | ReLU | $10^{-4}$ | 1.000 | 0.834 |
| 40-40-40 | Tanh | $10^{-4}$ | 0.999 | 0.644 |
| 40-40-40 | Sigmoid | $10^{-4}$ | 0.994 | 0.704 |
| 40-40-40 | ReLU | $10^{-3}$ | 0.997 | 0.911 |
| 40-40-40 | Tanh | $10^{-3}$ | 0.991 | 0.710 |
| 40-40-40 | Sigmoid | $10^{-3}$ | 0.296 | 0.252 |

As displayed in Table 16, the investigation of the optimal neural network structure involved examining four different structures, three types of activation functions, and two λ values. Generally, increasing the number of layers and nodes, thus enhancing the complexity of the network, tends to enhance its accuracy; however, it also raises the risk of overfitting. Adjusting the λ penalty parameter aids in fine-tuning the model to achieve a suitable balance

between learning and validation results. The findings indicate that, even with a straightforward 10-10-10 configuration, a modest increase in the λ parameter (from $10^{-4}$ to $10^{-3}$) significantly enhances validation results without markedly affecting the training outcomes. The sigmoid activation function underperforms compared to the other two functions, while the tanh and ReLU functions show varied performance relative to each other. Following an analysis of the results, a model featuring 40-40-40 nodes with the ReLU activation function and $\lambda = 10^{-3}$ was selected as the optimal structure.

Subsequent to the initial tests, the investigation focused on the performance of various model structures and hyperparameters across all dependent variables. The most favorable outcomes, achieved through the optimization of parameter sets, are detailed in Table 17 below.

**Table 17: Average $R^2$ values calculated with neural network models for different dependent variables with 3-fold cross-validation**

| Dependent variable | $\bar{R}^2_{training}$ | $\bar{R}^2_{validation}$ |
|---|---|---|
| **2r** | 0.997 | 0.911 |
| **3r** | 0.995 | 0.945 |
| **4r** | 0.995 | 0.807 |
| **5r** | 0.989 | 0.704 |
| **>6r** | 0.224 | -0.147 |
| **O-O** | 1.000 | 0.923 |
| **O-P** | 1.000 | 0.871 |
| **P-P** | 0.974 | 0.911 |
| **MMM** | 0.999 | 0.905 |

Analyzing the findings, it is evident that neural network models have achieved a highly accurate fit for various dependent parameters. Comparing the best validation outcomes of each model, it is clear that the performance of the straightforward linear model significantly lags behind that of both the stepwise linear regression and neural network models. By examining the results from the stepwise linear regression and neural network models, one can determine the best models and parameters that can adequately describe the entire set of dependent variables, except for molecules containing six or more aromatic rings.

The comparison of the results alongside the selected model for each dependent variable is illustrated in Table 18 below. It is also possible that the independent variables chosen and

measured do not substantially influence the formation of molecules with >6-ring, as certain variables like residence times might be better to describe these phenomena. Furthermore, upon examining Figure 41, it becomes evident that the factors applying the most substantial influence on the content of molecules with >6-ring are the formalin dosage duration ($t_F$) and the aniline/formalin mole ratio (A/F). The marked relationship between these two parameters and the content of >6-ring molecules can be attributed to the bridging role of the -CH2- functional group in formalin during the coupling of aromatic rings, which is significant in the synthesis of polymeric MDA molecules. As depicted in Figure 41, increasing the amount of formalin (i.e., reducing A/F) can significantly enhance the yield of polymeric MDA products, while extending the formalin dosage duration also proves advantageous for the formation of such polymeric products.

Upon reevaluation and categorization of the available ring distribution data into monomer (2-ring), trimer (3-ring), and polymer (>4-ring) contents, it is observed that while the efficacy of our regression models reaches a minimum, the total quantity of polymer molecules characterized by differing ring numbers can be accurately estimated. Given that the specific quantities of 4-, 5-, and more than 6-ring polymer molecules are not critically significant from the perspective of product mixture quality, they can be collectively grouped and estimated through a process of aggregation. The outcomes of polymer MDA molecules aggregated in this manner are presented in Table 18 as the dependent variable denoted as '>4r'.

**Table 18: Comparison of the validation results for stepwise linear regression and neural network models, selection of optimal model type**

| Dependent variable | $\bar{R}^2_{validation,\ NN}$ | $\bar{R}^2_{validation,\ stepwise}$ | Model selected |
|---|---|---|---|
| **2r** | 0.911 | 0.797 | Neural Network |
| **3r** | 0.945 | 0.673 | |
| **4r** | 0.807 | 0.647 | |
| **5r** | 0.704 | 0.484 | |
| **>6r** | -0.147 | -0.519 | |
| **O-O** | 0.923 | 0.887 | |
| **O-P** | 0.871 | 0.416 | |
| **P-P** | 0.911 | 0.480 | |
| **MMM** | 0.905 | 0.893 | |
| **>4r** | 0.716 | 0.624 | |

The evaluation of results indicates that predicting molecules consisting of >6-ring presents significant challenges due to various factors. One possible issue is that the analytical examination of molecules occurring in minimal quantities may experience substantial relative error. By aggregating the measured data, it is feasible to decrease the number of required neural network models while still acquiring all pertinent and significant information from the data. The dependent parameters derived from the best-performing models, plotted against the measured parameters, are illustrated in Figure 45.

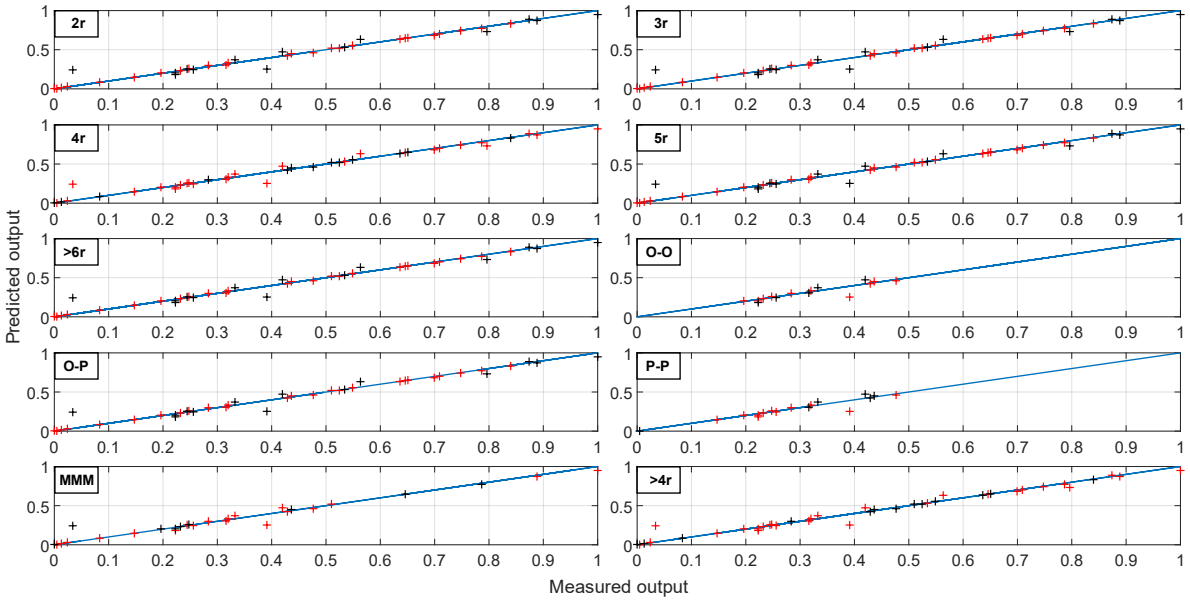


**Figure 45: Predicted output results plotted against measured results. Red markers indicate predicted outputs for training data and black markers indicate predicted outputs for validation data**

## 5.4 Shapley – based explainability analysis

To gain a clearer and explainable understanding of each synthesis parameter, I utilized the top-performing machine learning models to compute and visualize Shapley values for each dependent parameters. The calculated Shapley values for MMM, the most important by-product of the MDA synthesis is visualized in Figure 46, while the Shapley values for all dependent parameters can be seen in Figure A1 – A9.

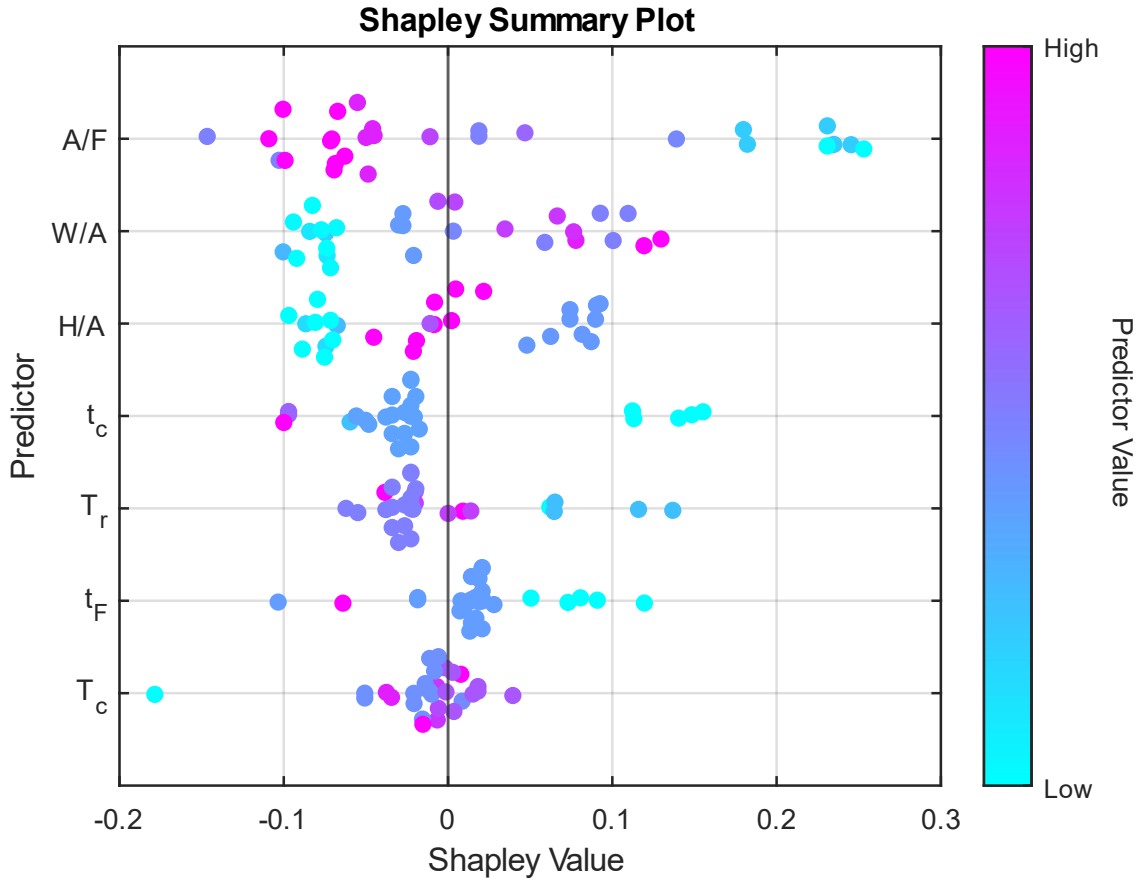


**Figure 46: Shapley values calculated with the best performing machine learning model for MMM formation**

Figure 46 illustrates the ranking of the key production parameters, which was established through the calculation of Shapley values. The parameter A/F exhibits the most significant impact on the quantity of MMM by-product, demonstrating an reciprocal relationship: as the A/F synthesis parameter is raised, the MMM quantity diminishes. Conversely, for W/A and H/A parameters, there is a distinct positive correlation with the MMM amount; the Shapley values suggest that increasing either of these parameters leads to a rise in the MMM quantity.

## 5.5 Conclusion

In Section 5, we assembled and organized laboratory experiments that present the process of MDA formation. The objective of the experimental design was to affect the quality of MDA by altering the production parameters and to describe the direction and magnitude of these correlations. Innovations in machine learning technologies have enabled the quantification of this data, facilitating a deeper comprehension of the chemical synthesis process of MDA. This, in turn, allows for the optimization of ring distribution and isomer ratios based on market

demands, as well as the selectivity of the reaction by modifying synthesis parameters through the application of such machine learning models.

It was discovered that water has a significant effect on the different dependent quality parameters, but a significant part of the water content is introduced into the system by the HCl solution used as catalyst, this means that there is a significant multicollinearity between H/A and W/A parameters, which was also confirmed by using shrinkage – method, i.e. LASSO – method.

The Pearson correlations between input and output parameters were also examined and visualized. Subsequently, a distinct method for detecting outliers was applied to identify specific outliers. This process employed the Leave-One-Out approach as a form of K-fold cross-validation, calculating absolute errors via linear regression and neural network models, which could notably enhance the models' fit.

Upon removing the outliers, the models underwent training and validation. The findings indicate that purely linear models fall short in accurately describing any of the dependent variables. Neural network models provided the best fits; however, stepwise linear regression models also demonstrated commendable performance. The top-performing models were chosen and summarized, enabling the calculation of several key dependent variables, which are vital for assessing product quality, with satisfactory precision. These explainable machine learning models provide straightforward implementation possibility for optimizing MDA synthesis processes, enabling clear quantification of each parameter's contribution through Shapley value analysis.

In future works, it would be valuable to optimize the method of detecting outliers. Currently, the approach eliminates entire experiments where all measured data are available as outliers. However, it is possible to identify only specific subsets of data, such as ring distribution or isomer distribution, as outliers. These subsets could be pinpointed due to reasons like analytical measurement variations (e.g., using gas or liquid chromatography) or errors in sample handling or other factors. Also, incorporating industrial production data into the training dataset could greatly enhance accuracy. This is because industrial production data is abundantly available, thus allowing for substantial improvement in model performance.

A plan for continuously modeling industrial systems and developing machine learning models could be devised. Gathering industrial expertise, practical experience, and insights detailed in this article, alongside results from laboratory experiments, is essential to ensure we

have all necessary foundational knowledge for devising the data analysis plan. Should the results from the ML models prove satisfactory, an optimization strategy, shaped by a cost function, will be needed for the industrial parameters (such as operational data). This cost function should prioritize objectives pertinent to the quality of the product mixture generated within the industrial system, such as enhancing the yield of 2-ring MDA, minimizing by-products, and reducing operational costs. Continuous system control and monitoring are crucial during both operation and adjustments of industrial parameters, based on results from soft sensor models, to assess if the output parameters behave according to the trained models' predictions. In this phase, the models serve as soft sensors. If additional data from industrial systems becomes accessible, the precision of these soft sensors may be enhanced, contingent on the results of the analysis.

Enhancing neural networks with an adaptable structure and tunable hyperparameters allows for improved model accuracy. Additionally, deploying multi-input and multi-output (MIMO) machine learning models enables parallel computation of each dependent parameter. This facilitates the development of an online prediction mechanism using industrial data, optimizing the process in real-time. Through this machine learning approach, by setting the desired product quality via an objective function, one can identify the best configuration of independent parameters within the analysis range. Consequently, the industrial production process can be optimized, ensuring product quality aligns with market requirements.

# 6 Soft-sensor development for product quality estimation with time delay and feature selection in industrial MDI production

The process of producing MDI involves a multitude of complex reactions and side reactions, making it difficult to identify root causes, side reactions, and by-products that affect product quality. Our goal is to overcome these difficulties using explainable machine learning, utilizing insights obtained to improve real-world industrial applications. By incorporating the outcomes from the machine learning model into the manufacturing workflow through the development of soft-sensors, we aspire to offer data-driven assistance for optimizing processes. The entire machine learning model is constructed using industrial data, and its outcomes have been verified against this data to confirm their reliability and relevance.

Owing to the intricate nature of the material system, various properties must be observed during production to attain an ideal product mix, and numerous unwanted side reactions and by-products influence the quality of MDI products. [83]. A significant characteristic of MDI products is the mixture's color, typically presenting as a light brownish red for 4-4' MDI [123], while in polymer products the color turns deeper, darker, and more intense because there are more by-products present. The color of MDI mixtures is crucial for both the manufacturers of the MDI product and the clients who utilize that product as raw material, making it a significant parameter where a higher color value is preferable in MDI production. The color values of different MDI product mixtures can be seen in Figure 47.

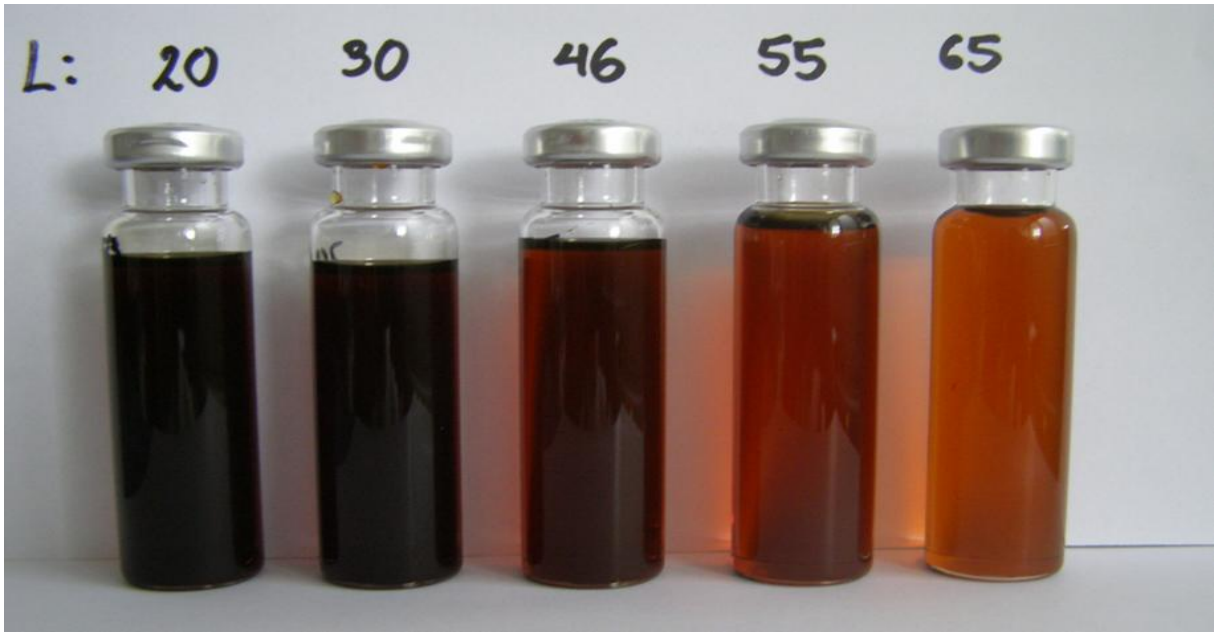


**Figure 47: MDI product mixtures with different L-color quality values**

Once the production processes are finished, the mixture's color may degrade further when subjected to heat or light [124–127]. The darkening of the mixture is associated with the side reactions occurring during the manufacturing process, as well as the quantity and quality of the by-products generated therein. However, in numerous instances, the precise reaction pathways leading to the formation of these by-products remain undiscovered, as does the exact mechanism through which they impact color variation [128,129].

In the chemical industry, the use of various soft-sensor applications is gaining increasing importance, as large volumes of data are generated in chemical processes. Machine learning, deep learning, and other mathematical models provide excellent opportunities for processing large amounts of industrial data, enabling a deeper understanding of processes and the estimation of various quality parameters [130,131]. In the development of a soft-sensor, an additional challenge in any industrial system is the estimation of time delays, which refers to the duration after each process variable begins to influence the output parameter under investigation. Given the scale of a real industrial system, it becomes evident to model developers that each operating parameter does not exert an immediate effect on the parameter under investigation, but rather a delayed impact. These systems are called time-delayed systems [132–134]. For such industrial systems, the material flow is nonlinear and variable, and the system's complexity is significantly high. Due to constant fluctuations in mass flows, recirculations, material bypasses, load levels, and the generally complex product portfolio, a

time delay estimation model has also been developed to enhance the accuracy of machine learning models. When estimating the time delay for each operating parameter, we can classify the methods into static [135], dynamic [136,137] or a combination of both static and dynamic time delay estimation approaches [138,139]. Static time delay estimation assumes a constant delay over time, while dynamic time delay estimation accounts for time delays that vary with changing system conditions. Dynamic time delay estimation methods are particularly suitable for complex systems characterized by frequent load changes or where material movement involves numerous recirculations or feedback loops, while the combined approaches consider both types of delays in the system.

Applications are also feasible where the soft sensor can be used adaptively to adjust model parameters in response to changing conditions [139]. Furthermore, the effect of each operating parameter on the outputs can be estimated directly. However, there are instances in the literature where the parameters exhibit a finite impulse response, meaning that not only do they have a time-shifted effect, but the intensity of their impact also varies continuously over time [136,140].

Currently, there are no a priori or a posteriori models in the literature that comprehensively address and quantify the effects of parameters influencing the color of MDI product mixtures. While several studies have explored side reactions, the presence of certain components and the occurrence of these reactions remain speculative, primarily due to analytical challenges. [91].

Our objective is to enhance the quality of MDI mixtures, with a particular focus on maximizing MDI color, through a continuous closed-loop, model-based analysis of the complex material system. This will be achieved by developing and implementing various a posteriori machine learning models. By utilizing the tools of explainable machine learning, the relationships between production parameters and MDI color can be explored for such a complex system. Additional insights can be gained by investigating these relationships, particularly regarding the presence of side reactions and by-products that are assumed in the literature but have not been analytically proven.

The summary of my research in this field as follows:

- We have gathered the information available from industrial knowledge, experience, and the literature, and subsequently identified the reactions and reaction pathways that could potentially lead to the formation of by-products influencing the color of MDI;

- Based on the defined by-product reaction scheme, we identified the industrial operating parameters most likely to influence changes in MDI color, and then gathered data on these parameters;
- Exploratory data analysis was conducted on the collected data, including correlation analysis, violin plots for variance assessment, time delay estimation, and principal component analysis (PCA);
- Feature selection methods were applied to the resulting dataset to identify the most important input features or parameters for training the machine learning models. The results of the models were then compared and evaluated;
- Based on the results of the best-performing model, we examined how each operating parameter influences the color of MDI. Subsequently, a genetic algorithm was employed to identify the optimal set of operating parameters that maximizes MDI color during production.

## 6.1 Presumed reaction scheme for by-product formation

In this section, I have defined a presumed reaction scheme for side reactions and for by-products formation based on literature and industrial experience in order to identify the key points of the industrial system by understanding the basic processes and reaction system. I have identified the key variables that are most likely to provide the most important and informative information about by-product formation and hence the colour of the final MDI product mixture.

After understanding the industrial processes, the next step before preparing the data analysis plan is to understand, based on the literature, what by-products can be generated during the production of MDI that can affect the colour of the product. It is also important to understand what side reactions can lead to the generation of these by-products and what reaction conditions can promote the occurrence of these side reactions, i.e. how we can influence and how we can reduce the quantity of these by-products.

Based on the literature, the key molecules for the formation of color causing by-products are the different urea compounds [129], which can be formed from the reaction of amine (aniline, MDA compounds) and isocyanate (MDI) function groups. According to Twitchett *et al.* [104], the amide and imidic forms of urea compounds are in equilibrium, with the imidic form being involved in side reactions as shown in Figure 48.

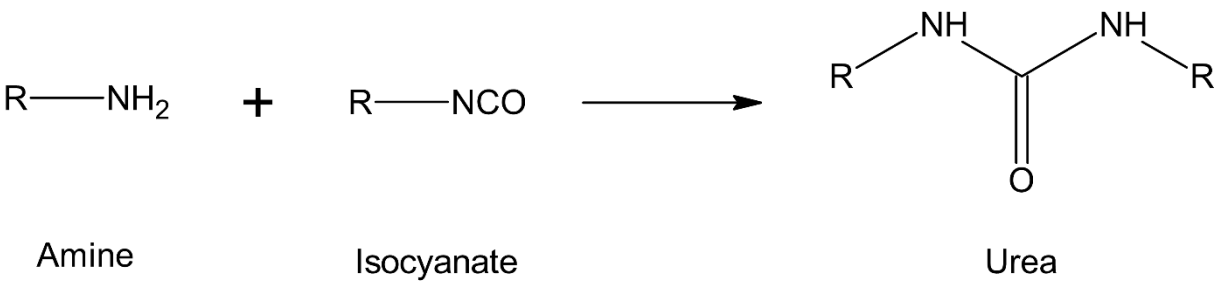


**Figure 48: Formation of urea from the reaction of amines and isocyanates**

The imidic forms of these urea compounds can undergo a series of phosgenation steps to form chloroformamidine and then chloroformamidine-N-carbonyl chloride (CCC) type compounds, which can decompose upon elevated temperatures to form various isocyanate and isocyanide dichloride compounds [104], which have been identified as the cause of the colour change in MDI, as they may can form chlorine radicals. Chlorine radicals can then attack the aromatic compounds present in the system and the methylene bonds between the aromatic molecules to form compounds with conjugated electron structure, thus contributing to the deepening of the color of MDI mixtures [91]. A possible reaction pathway for the formation of dichlorides is shown in Figure 49.

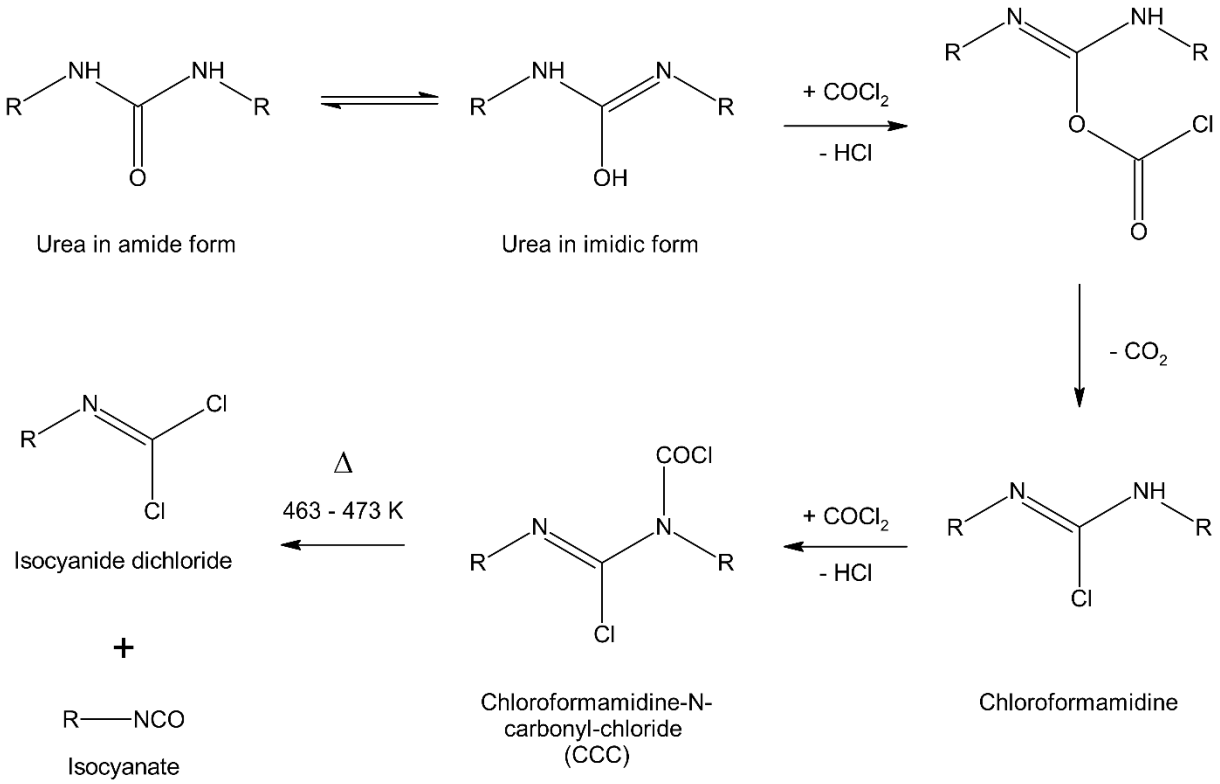


**Figure 49: Suggested reaction route for dichloride formation from urea**

To confirm the literature and to explore the side reactions involved in the synthesis of MDI, Callison *et al.* [91] demonstrated that the multistep phosgenation of urea-type compounds does indeed lead to CCC compounds. It has also been investigated whether the resulting CCC compounds are actually in equilibrium with carbodiimide compounds, an equilibrium that is significantly affected by the storage temperature of the mixture. For the decomposition of CCC compounds to dichlorides, it was found that the decomposition takes place at temperatures above about 130 °C. By studying the decomposition of dichloride compounds, they were able to confirm that different dichloride compounds themselves undergo a color change at the temperatures used in the production of MDI, but this color change is even more significant with the addition of MDI. In polymer-MDI mixtures, the degradation of dichloride compounds causes even more significant colour degradation, presumably due to the enhancement of conjugation at higher molecular weight. Full reaction scheme proposed by Callison *et al.* can be seen in Figure 50.

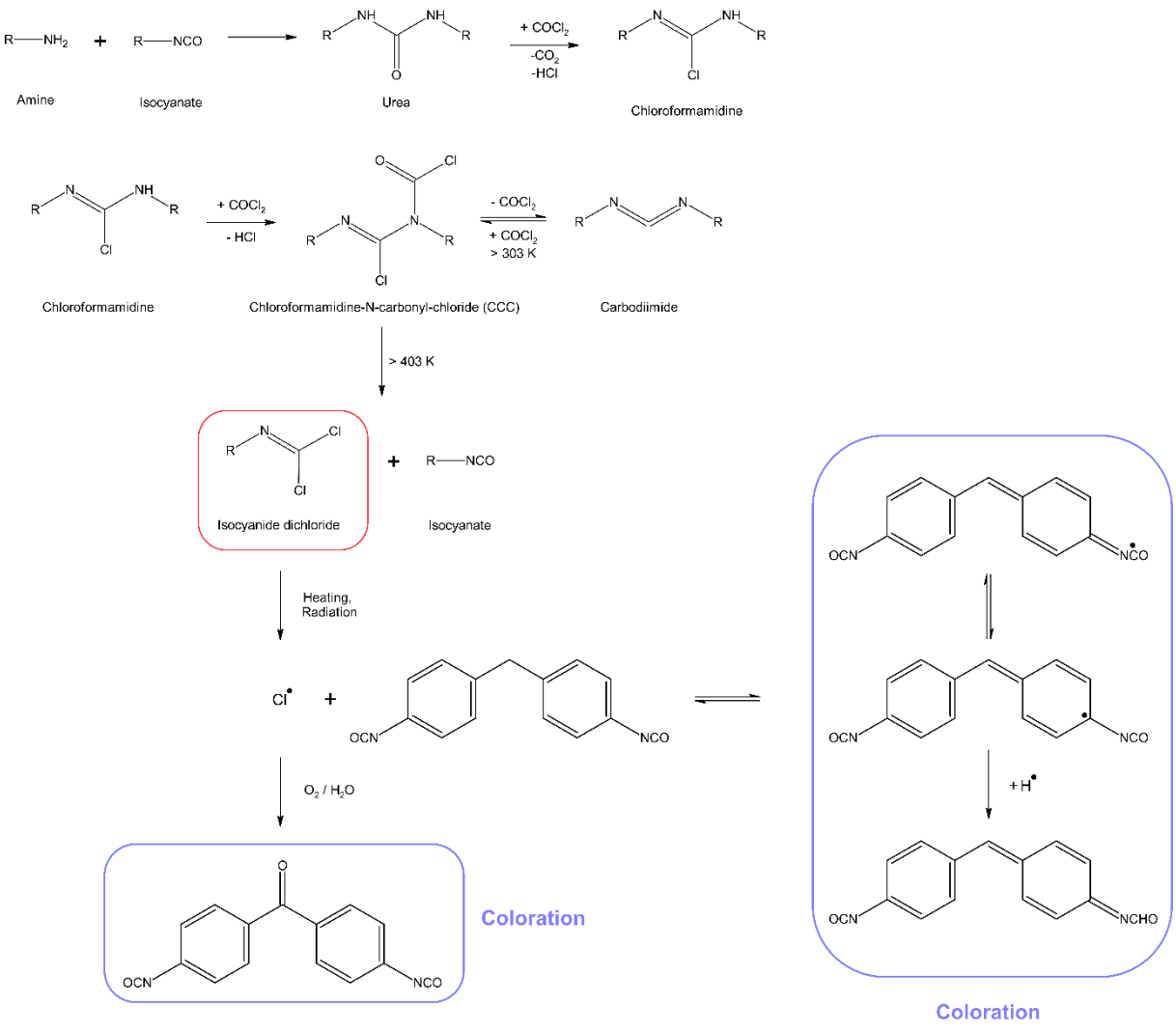


**Figure 50: Suggested full reaction scheme for the formation of dichlorides from urea [91]**

Based on the full reaction scheme, 3 intervention options have been identified for an industrial system to limit the generation of by-products and improve the quality of the MDI product mixture:

- **Avoid the formation of urea compounds:** the formation of urea compounds requires the interaction of -$NH_2$ and -NCO functional groups, so it is necessary to look at the parts of the technology where these two functional groups can potentially meet and the operations where the phosgenation reactions are already taking place. During the phosgenation reactions, -$NH_2$ and phosgene reacts, but the reaction also involves the presence of an unreacted -$NH_2$ group, already converted -NCO group and phosgene in the mixture at the same time. By influencing the parameters of the phosgenation reaction, therefore trying to promote the phosgenation of the -$NH_2$ groups, the amount of urea and indirectly the amount of by-products can be reduced
- **Minimize the formation of CCC compounds:** CCC compounds are formed according to the reactions shown in Figure 50 and once urea is formed during the reactions, it is inevitable that the CCC molecule will form as well in the system. However, we are able to control how much of the total amount of CCC compounds formed will decompose to dichloride and how much will be converted to carbodiimides according to the equilibrium reactions. The conversion from CCC to carbodiimides requires the release of phosgene and the uptake of HCl, so it is crucial to strip the phosgene out of the reaction mixture after the completion of the phosgenation reaction and to provide sufficient HCl during HCl stripping to decompose to carbodiimides
- **Avoid the decomposition of CCC to dichlorides:** the decomposition of CCC molecules to dichlorides is significantly affected by both temperature and residence time, i.e. both operating temperatures and residence times should be limited, especially for units operating at high temperatures well above 130 °C. The highest temperatures for the MDI production are on the ODCB recovery system and MDI separation systems, and it is therefore essential to limit residence times on these systems as low as possible

After the identification of the possible intervention points for the by-products formation, a data analysis plan was prepared to collect data from 52 industrial instruments. The data analysis plan was prepared for the phosgenation and MDI purification blocks, because after collecting and evaluating the preliminary literature and industry knowledge, it became clear that the greatest impact in improving MDI color could be mainly achieved by optimising the operational parameters in these blocks.

## 6.2 Machine Learning model development and optimisation

By implementing the optimal parameter set on the industrial system, the model and its results can be validated, and if necessary, the data collection, preparation, model training and validation steps can be carried-out again by repeating the previous steps, further improving the performance of the developed machine learning models. A summary of the tasks performed in this work is shown in Figure 51 below:

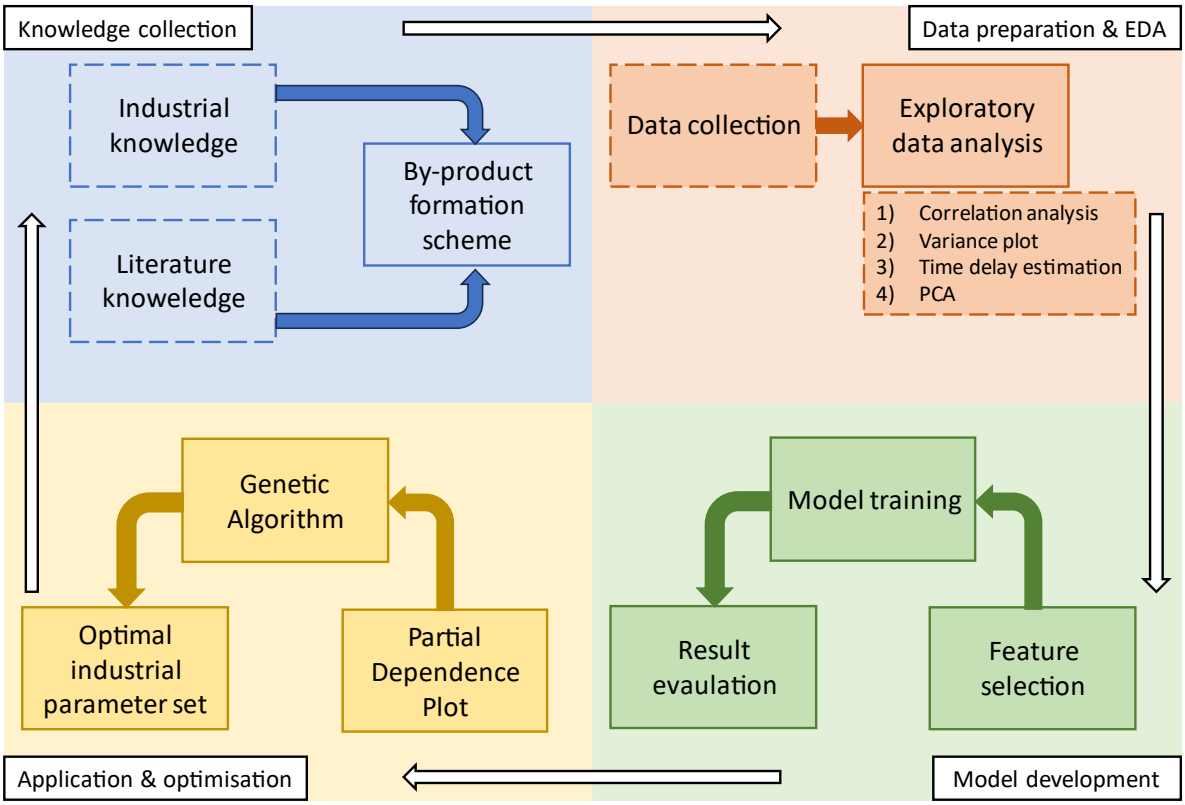


**Figure 51: The tasks performed in this study and their order in the development of machine learning models for estimating MDI color**

According to the limited information available in the literature and industrial experience, the target of this study was to support the industrial MDI production system by performing and refining a circular, continuous offline model-based analysis which later can be developed for online production support. With machine learning models, we are able to maximize the quality of the MDI mixture, in this case the color value, according to the following steps, which are also shown in Figure 51:

- Industrial & Literature knowledge: the mechanisms of side reactions have been explored, summarised and By-product formation scheme has been prepared

- Data collection: identified key devices and systems based on the By-product formation scheme, where side reactions are presumed to be most likely to be inhibited. Based on that, we prepared a data analysis plan to determine which industrial data should be collected and analysed, both in terms of instrumented control system and laboratory sampling – these industrial operational parameters became intervention points and the independent input variables of the machine learning models
- Exploratory Data Analysis: the collected data were pre-analysed by correlation analysis, variance checking using violin plots, i.e. we confirmed that we had the right amount and quality of data. PCA was also used to check the clustering of the data to see whether the data reveals a production state associated with a higher MDI colour
- Feature Selection: a reduction in the number of input parameters was performed using different feature selection methods
- Model Training: trained and validated the machine learning models, then summarised, compared and analysed the results
- Partial Dependence Plot: generated and analysed partial dependence plots to visualise the effect of each operational parameters
- Genetic Algorithm: output optimization based on the best performing model through genetic algorithms
- Optimal industrial parameter set: determination of optimal operational parameters based on GA algorithm results

The learning and validation process of each machine learning model was carried out according to the following steps:

- loading, organising and normalising data
- determining the optimal time delays for each input parameter
- creating training partitions: 5-fold cross validation
- model performance validation, comparison, evaluation
- determinination of the best performing model, optimisation.

### 6.2.1 Regression models

The measured target variables can generally be expressed as it was written in Equation 38, accounting for measurement errors $\varepsilon_k$.

$$y_k = f(\boldsymbol{x}_{k-\boldsymbol{d}}, \boldsymbol{\theta}) + \varepsilon_k, k = 1, \dots, N \tag{38}$$

During the development of these models that can be used to predict the color of MDI mixtures, the following five different types of models were used and compared based on their performance:

- linear regression (LR)
- regression tree (RT)
- neural networks (NN)
- support vector machine (SVM)
- gaussian process regression (GPR)

Model predictions, $\hat{y}_k$ can be written in general as in Equation 39:

$$\hat{y}_k = f(\boldsymbol{x}_{k-\boldsymbol{d}}, \boldsymbol{\theta}) \tag{39}$$

Where $\hat{y}_k$ is the model prediction for the output value. To be able to estimate the optimal time delays for each input parameters, we developed a correlation-based time delay estimation model that estimates the optimal time delays based on the absolute strength of Pearson correlations between the output and input parameters as it can be seen in Equation 40.

$$d^* = arg\max_{\boldsymbol{d}} \boldsymbol{c} = \left| \frac{\sum(\boldsymbol{x}_{k-\boldsymbol{d}} - \bar{x}_k)\,(y_k - \bar{y})}{\sqrt{\sum(\boldsymbol{x}_{k-\boldsymbol{d}} - \bar{x}_k)\ \sum(y_k - \bar{y})^2}} \right| \tag{40}$$

Where **c** is the calculated Pearson correlation vector for all input parameters with a size of $1 \times N_x$. The solution ***d**** corresponds to the time delays that maximize the absolute Pearson correlation ***c***, ensuring the strongest relationship between the shifted input variables and the target variable. The constraint for $d_i$ is 12 hours $\leq \mathbf{d} \leq$ 72 hours, which means that a time delay of less than 12 hours is not possible, so all dead times were between 12 –72 hours.

The calculation and minimization of prediction error is performed for all machine learning models according to the cost function in Equation 41, which is in fact the mean squared error function:

$$\min_{\boldsymbol{\theta}} \frac{1}{N_x} \sum_{i=1}^{N_x} (y_k - \hat{y}_k)^2 \tag{41}$$

### 6.2.2 Selection of important features

Performing feature selection was important for two reasons. Firstly, due to the large amount of data, the performance of the models may not be optimal due to low quality, irrelevant or of limited predictive value, which may then degrade the model results. On the other hand,

the aim of our investigations is to identify the most important independent parameters that can be optimised as operational parameters to further improve the MDI color in a real industrial system. It is easy to realize that on an industrial-scale system, modifying 52 parameters in near-one-time is not feasible, while implementing an optimal parameter set of 10-15-20 parameters may be of manageable magnitude. Feature selection has been performed using four different methods and the results of these methods have been compared to determine the feature set that is considered ideal. Our goal was thus to further reduce the size of the input matrix of the original 52 features, which already contained a restricted set of independent data, so that the accuracy of the models would be reduced as little as possible, or not at all, while the number of data required for regression would be reduced significantly and the optimization of the industrial system could be performed. The feature selection methods used were as follows:

- MRMR – Minimum-Redundancy-Maximum-Relevance method
- F-test method
- RReliefF method
- Correlation-based feature selection
- Aggregated method

The MRMR algorithm tries to find the optimal set of features to estimate the output parameters as accurately as possible, while the mutual information between these parameters are minimised (or the dissimilarity of these parameters are maximised). The algorithm quantifies redundancy and relevance using pairwise mutual information of features, and mutual information of a feature and the response variable [63].

The F-test method calculates the importance of each predictor characteristic one by one by calculating the p-values of the data set. As a general formalisation, the F-test method calculates the ratio of the explained variance to the total variance for each data series.

The original Relief algorithm was limited to classification problems with two classes only, but the RReliefF method can be applied to regression problems and for multiple classes. In case of regression problems, the RReliefF method does not use the classification identification, i.e. whether or not an element belongs to a given class, but a probability is calculated instead [141]. The algorithm penalizes predictors and rewards predictors based on their capabilities to predict response values and uses intermediate weights to compute the final predictor weight.

With the correlation-based feature selection method, Pearson correlations were calculated between each predictor and the output parameter, and the absolute values of these correlations were then ranked in their order of importance.

In the process of implementing the Aggregated approach, the previously mentioned outcomes of the four feature selection methods were synthesized to establish and evaluate a combined feature set.

After the model training & development, the different machine learning models can be used to minimize or maximize the output parameter using constrained or unconstrained edge search algorithms. In this work, to maximize the color of MDI, I used a so-called genetic algorithm (GA) [79], which is a global extreme value searching algorithm already presented in Chapter 2.

## 6.3 Results

The objective of this study was to develop a model for tracking, estimating, and optimizing the color of an MDI product mixture produced within a real industrial system. As such, the values of the operational parameters and their associated physical interpretations are not discussed or analyzed within this article. Due to confidentiality concerns regarding the data, system, and technology, all data have been normalized. Only a selection of significant relationships, effects, and their underlying mechanisms between the operational parameters and the MDI color are presented as examples to safeguard the proprietary technology.

### 6.3.1 Data preparation

During the specified three-month period, data was gathered from the industrial system at an hourly sampling interval, along with all laboratory measurements concerning the quality variable to be predicted - the color of MDI. Given the relative infrequency of laboratory measurements, collecting instrument data hourly was seemed sufficiently frequent for prediction purposes. Altogether, more than 2,300 hours of data and over 220 laboratory color measurements were gathered, and then processed and prepared for further analysis and machine learning model training. The collected data was solely utilized for training and validation, with no separate dataset designated for testing.

Before the model development process, data were collected and systematically organized. During this phase, a total of 52 operating parameters related to phosgenation and the MDI block were identified and designated as key parameters based on practical experience and a

preliminary literature review. The data collection and analysis spanned a period of approximately the aforementioned three months, during which the plant underwent a restart following its annual maintenance activities. Consequently, the dataset contains phases of load increases and decreases, as well as transient conditions, which provide a robust foundation and high variance for conducting investigations and training the ML models. The normalized plant production data are presented in Figure 52 below.

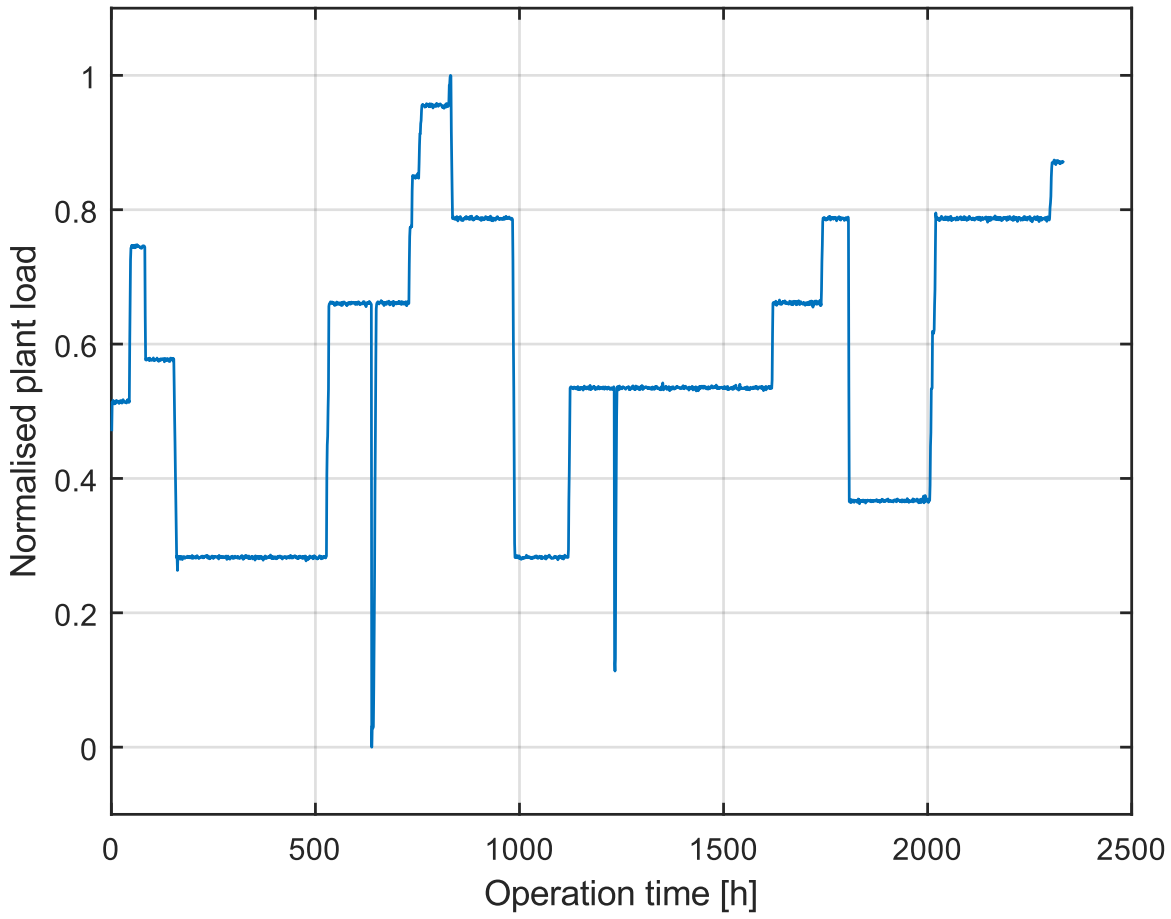


**Figure 52: Normalised load data of the industrial plant**

As illustrated in Figure 52, there are highly transient periods of 1-2 days between each load change, during which load increases or decreases occur. During these intervals, the MDI color exhibited significant changes after preliminary verification, as the system's steady state also underwent substantial fluctuations. Consequently, the load data for these periods, along with the corresponding laboratory L-color measurements, were removed from the dataset. The revised, cleaned load data series is presented in Figure 53 below, and the normalized values of the L-color parameters measured at the respective time points were incorporated following initial outlier detection and removal.

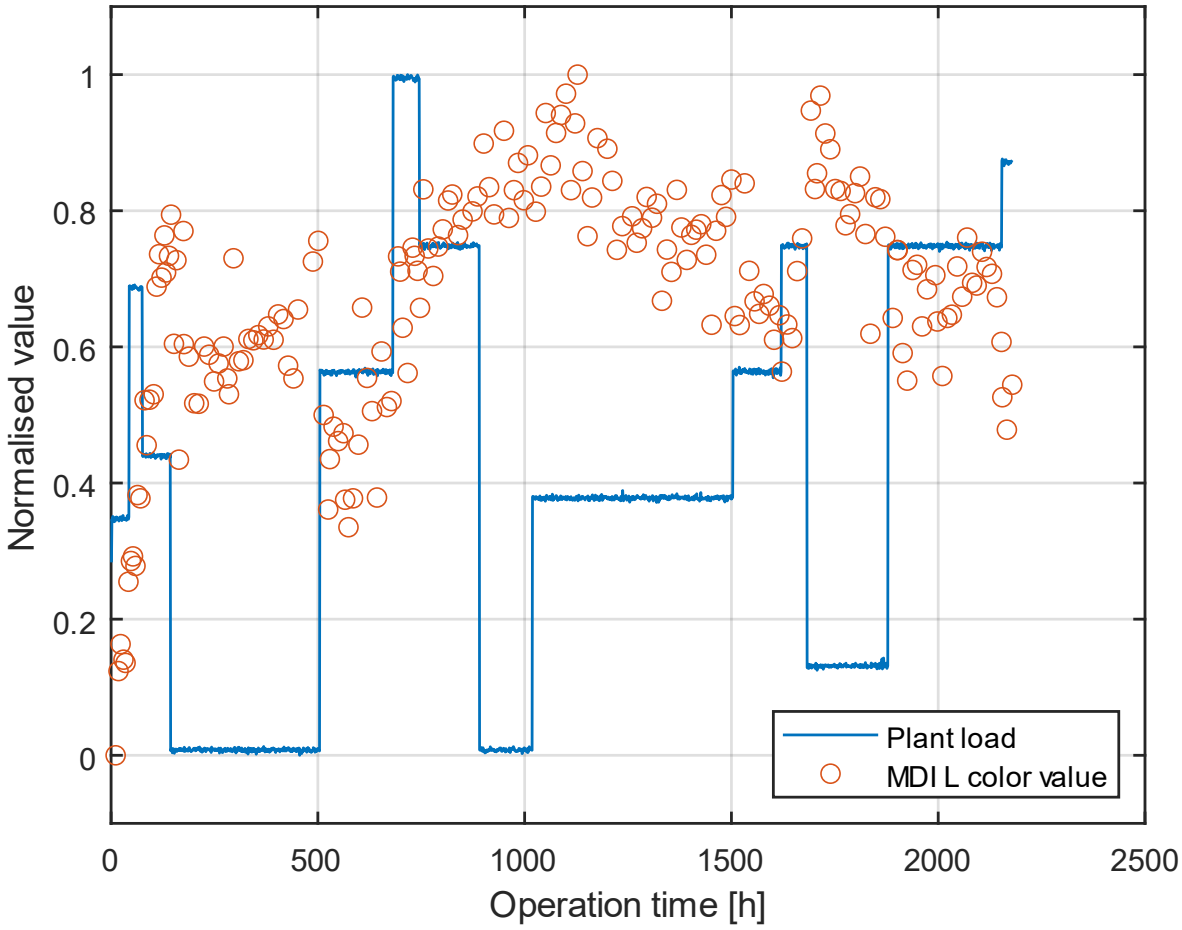


**Figure 53: Normalised and cleaned plant load and MDI color data of the industrial plant**

To assess the variance of each parameter, a violin plot was created. This plot serves to visualize the median and variance of all 52 input parameter vectors, effectively allowing for an evaluation of the quality of the available data. The data for the operational parameters displayed on the violin plot are shown in Figure 54 below.

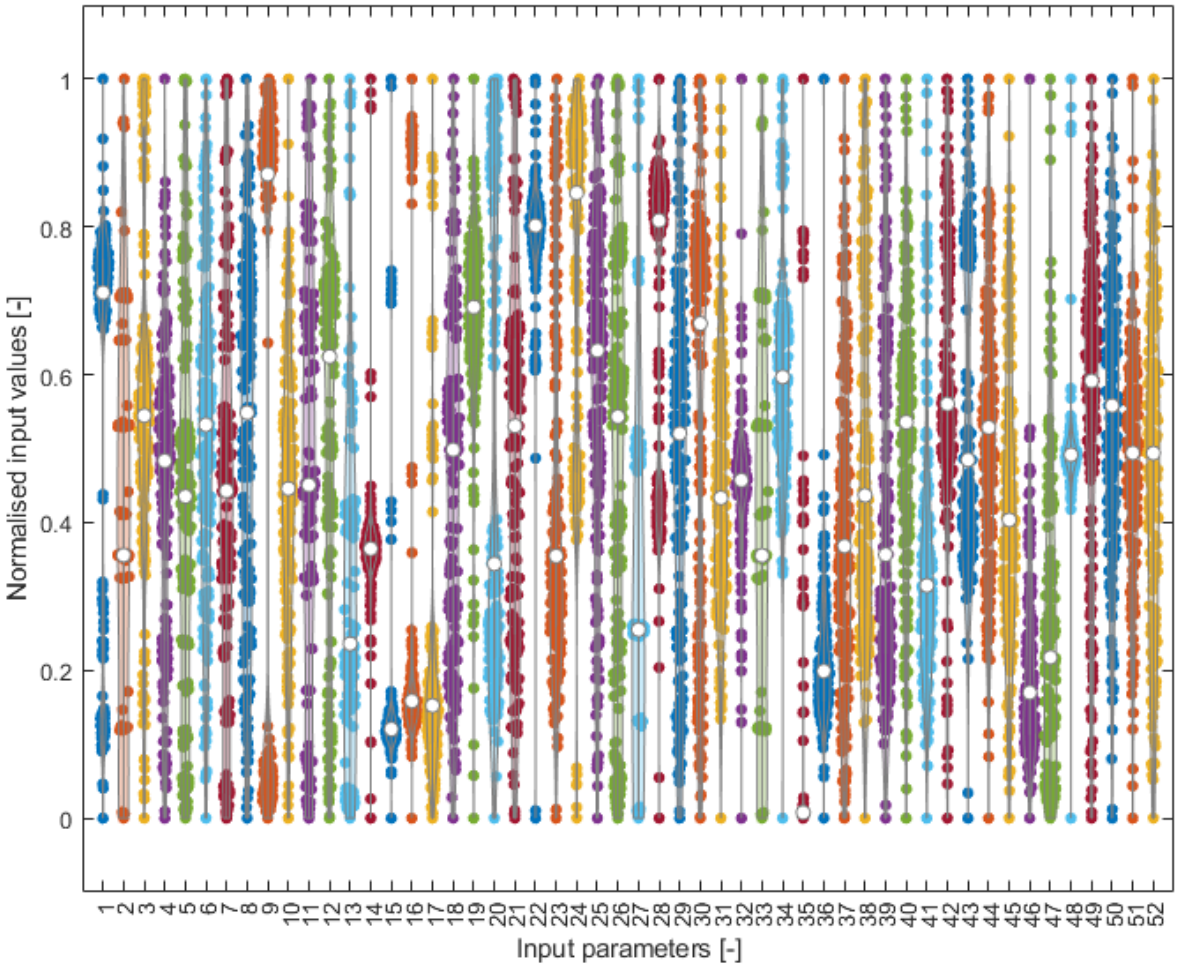


**Figure 54: Input parameters plotted on violin plot**

Figure 54 demonstrates that the data exhibit sufficient variance, making it suitable for the predefined input parameters to be utilized in subsequent analysis.

A Pearson correlation analysis was also conducted, in which correlation coefficients were calculated between the 52 operational input parameters and the MDI color data. The results of this analysis are presented in Figure 55.

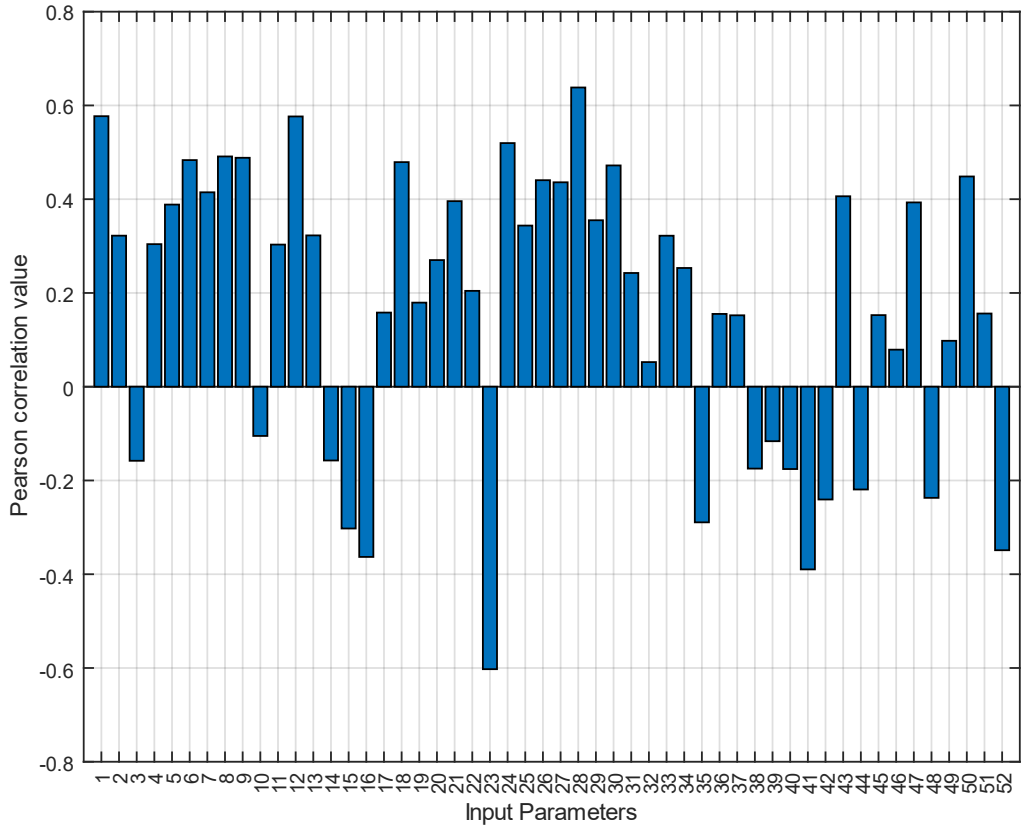


**Figure 55: Pearson correlation between operational parameters and MDI color data**

The positive and negative correlations are linked to changes in the color of the MDI product. Specifically, increasing the positively correlated parameters will lead to an increase in color, while increasing the negatively correlated parameters will decrease the color of the MDI product. These correlations provide valuable preliminary information before the models are trained.

When examining the negative correlations, the relationship between input parameters 15 and 16 is particularly informative, as these two parameters are indicative of the reactor temperatures in the phosgenation process. It is well-established and confirmed that an excessive increase in the temperature of the phosgenation process negatively impacts the color of the MDI product, so the correlation analysis likely reveals useful, explainable and valid insights.

A similar preliminary analysis was conducted for the other parameters exhibiting stronger correlations. However, it is crucial to note that these correlations are only applicable for mapping and quantifying linear relationships. Since the MDI industrial process is inherently nonlinear, these correlations should be approached with caution. For this reason, all operational parameters, including those with weak correlations, were retained prior to training the models,

as the precise impact of these parameters on the final MDI product quality cannot be determined with complete certainty.

### 6.3.2 Time delay

The determination of the optimal time delay to enhance the performance of machine learning models was conducted prior to their training, as described in Section 3 using a correlation-based method. The maximum allowable time delay was set at 72 hours, while the minimum time delay was established at 12 hours for all parameters. These limits were determined based on engineering expertise and insights from operating specialists.

Each operational parameter was then varied within a time window of 12 to 72 hours, and the variation in the absolute strength of their correlation with the output parameter (i.e., MDI color) was calculated. The time delay at which the correlation between the independent and dependent variables was strongest was considered optimal, and this value was recorded as the final time delay for each input parameter in the models.

As an example, the correlation variation for the 15-th operational parameter, which is related to the temperature of the phosgenation reaction (as shown in Figure 16), was analyzed. Industrial experience has demonstrated that this parameter has a negative correlation with the MDI color and is one of the key parameters for refining color. The correlations at different time delays are presented in Figure 56 below.

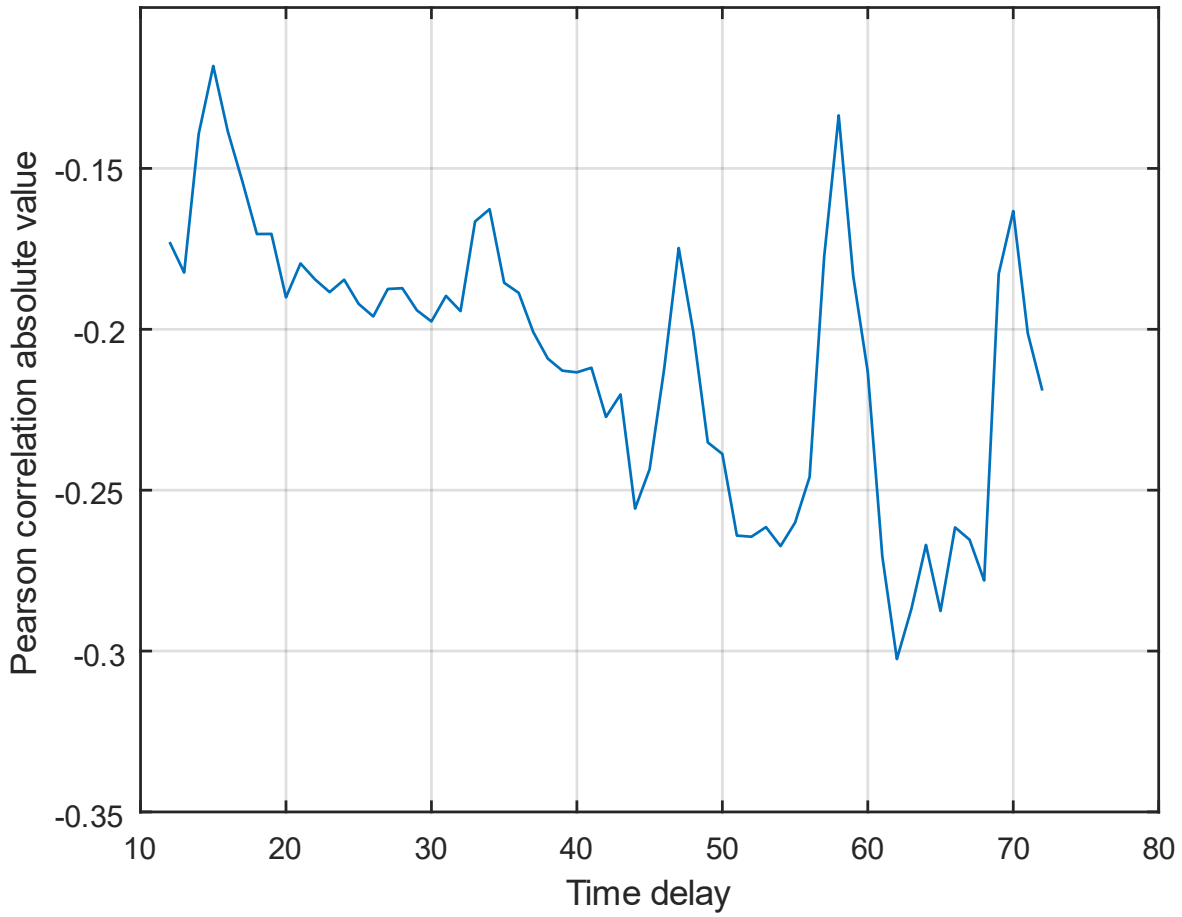


**Figure 56: Calculated Pearson correlation values between 15-th operational parameter and MDI color for different time delays**

Figure 56 shows that the optimal time delay was determined to be 62 hours, as this time delay resulted in the strongest negative correlation with the output parameter.

After selecting and plotting the optimal time delays using the correlation-based method, the determined time delays are presented in Figure 57 below:

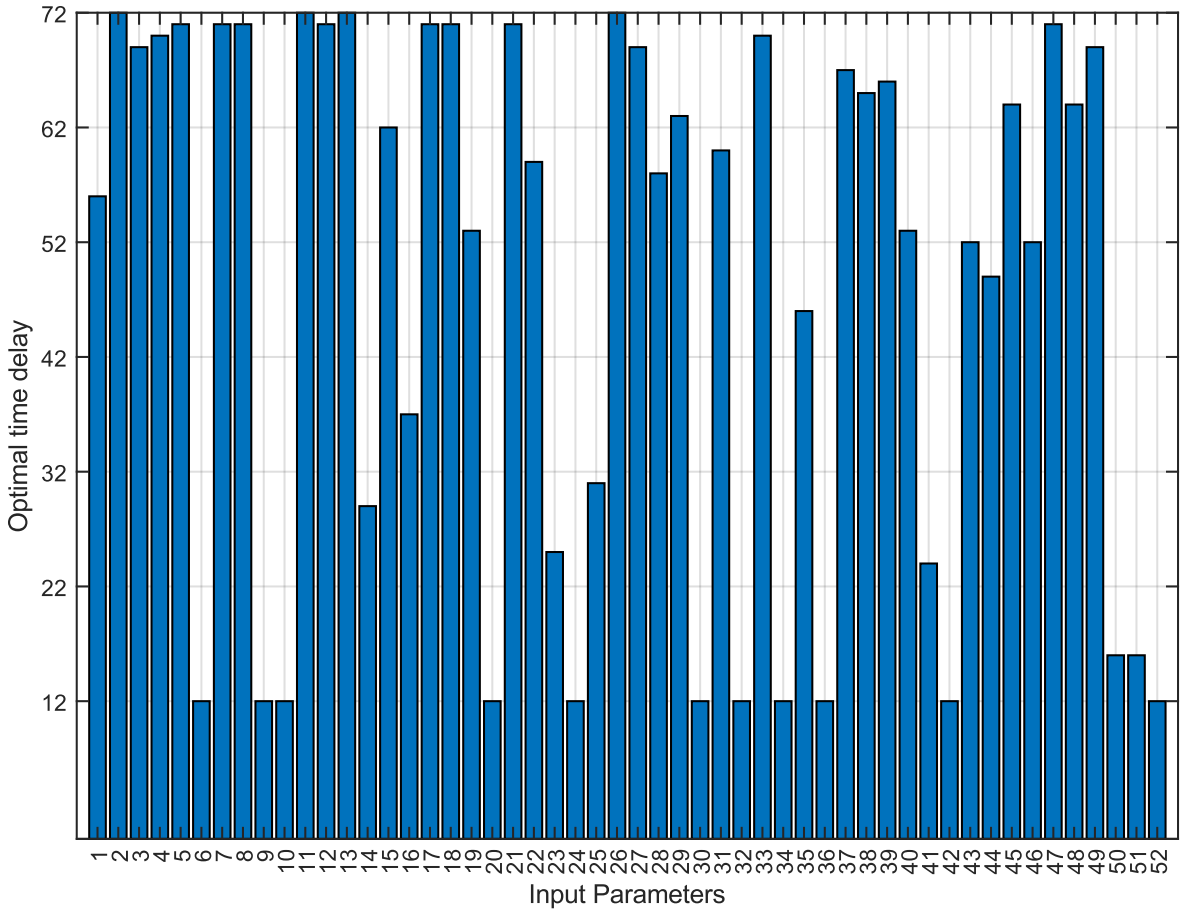


**Figure 57: Calculated optimal time delays for each input parameters**

Figure 57 shows that the optimal time delays varied significantly across different operating parameters. For many parameters, the optimal time delay was found to be greater than 60-70 hours following the correlation analysis, while for some parameters, the optimal time delay was closer to or even exactly at the minimum of 12 hours that was defined. Generally, parameters related to 'fluid level' quantities and 'flow quantities', such as recirculations, exhibited higher model time delays, whereas parameters associated with temperature, pressure, and other thermodynamic equilibria showed lower time delays.

### 6.3.3 Principal Component Analysis

Before training the models and after determining the optimal time delays for the independent parameters, a Principal Component Analysis (PCA) was conducted to gain insights into potential clustering of the data and explore possibilities for dimensionality reduction. The preliminary results of the PCA are presented in Figure 58 below:

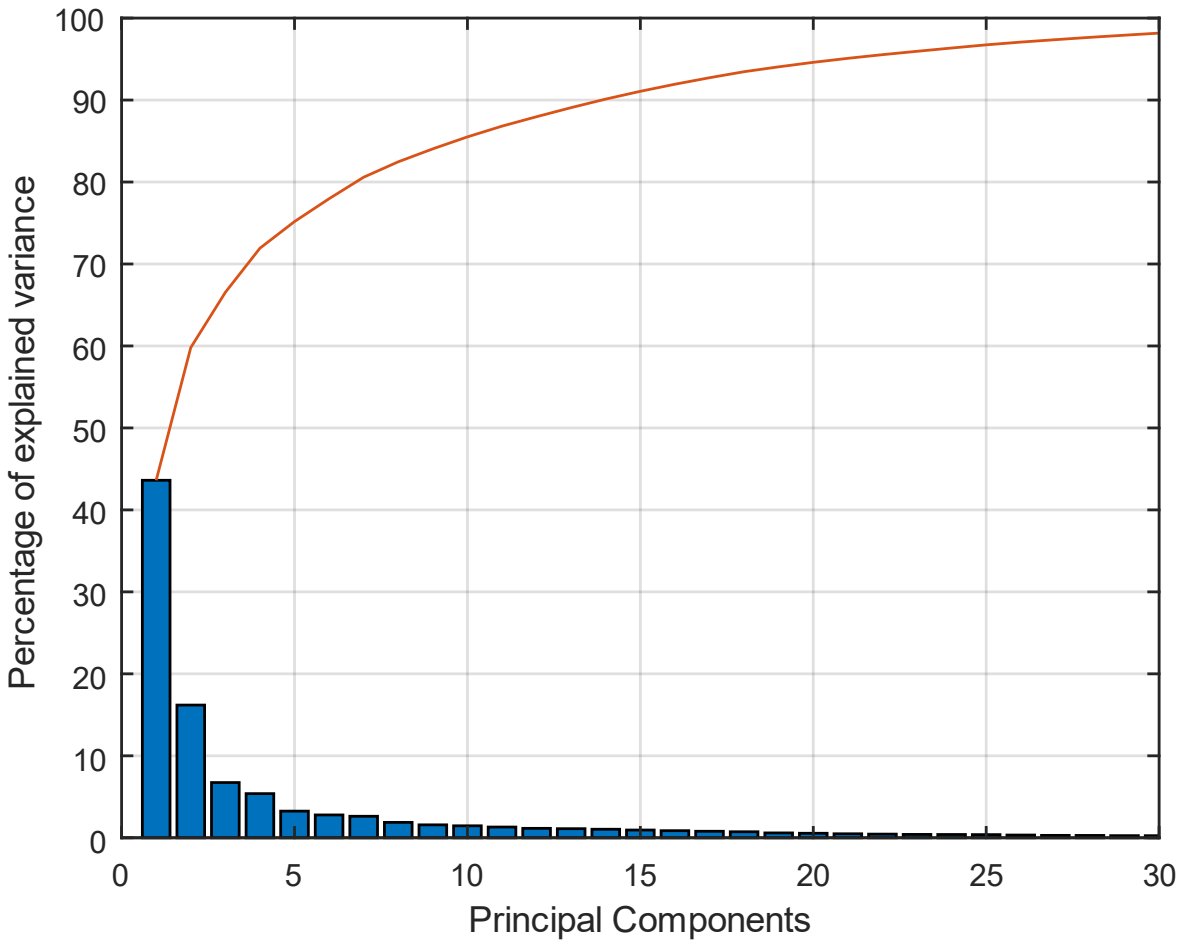


**Figure 58: Results of preeliminary Principal Component Analysis**

As shown in Figure 58, the contributions of each Principal Component are represented by the blue bar chart, while the cumulative sum of the total variance explained is represented by the orange curve. From these results, it is evident that dimensionality reduction could be a viable approach for the modeling process, as approximately 20 Principal Components (PCs) account for 95% of the total variance. This suggests that the number of dimensions could be reduced by more than half without significant loss of information.

By plotting the scores of the first two or three Principal Components, the results shown in Figure 59 and Figure 60 can be visualized. These plots provide valuable insights into potential clustering patterns in the data:

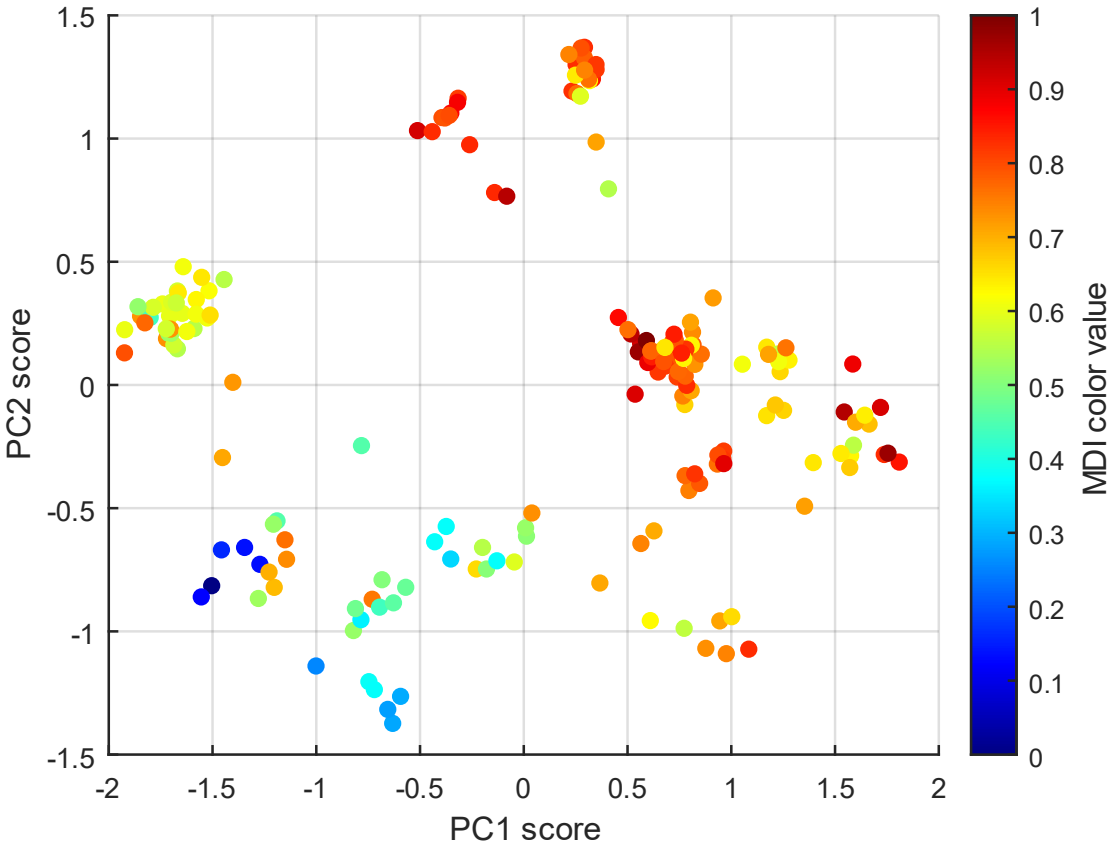


**Figure 59: Plot of results for the first two principal components as coordinate axes. The color of each data point presents the value of the measured MDI color value**

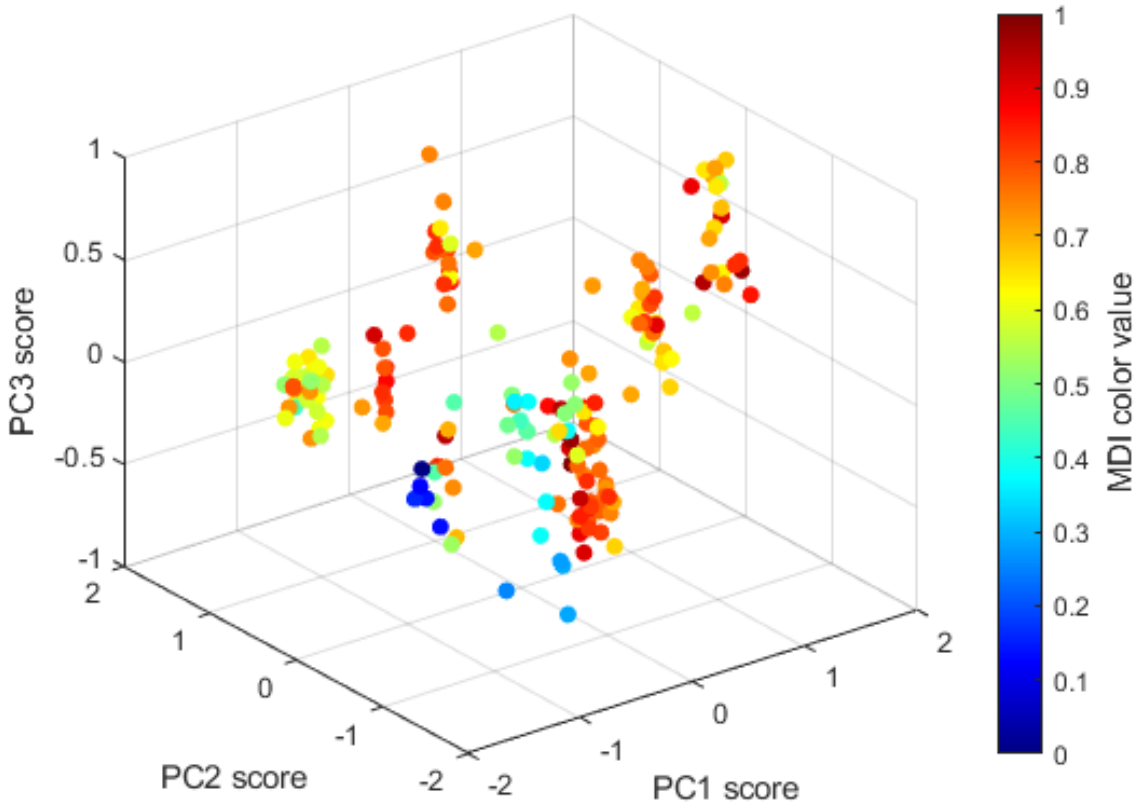


**Figure 60: Plot of results for the first three principal components as coordinate axes. The color of each data point presents the value of the measured MDI color value**

The results shown in Figure 59 and Figure 60 indicate the presence of 3 to 4 significant clusters in the data. Two of these clusters exhibit relatively high MDI color values, while the other two clusters have lower mean color values. These clusters have been identified and are illustrated in Figure 61 below, which is based on the earlier two-dimensional PCA plot in Figure 59.

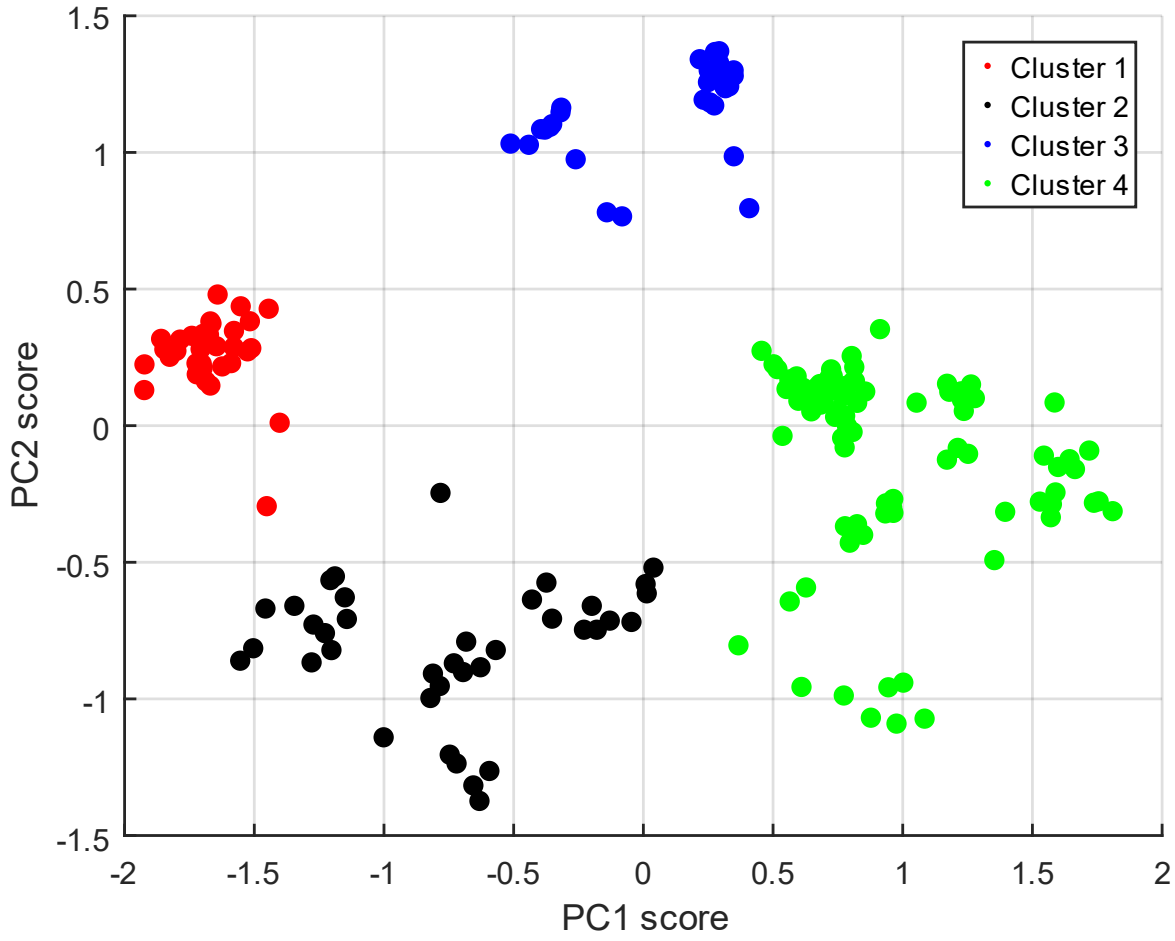


**Figure 61: Plot of results for the first two principal components as coordinate axes. The color of each data point indicates to which cluster the data point belong**

The four clusters shown in Figure 61 exhibit distinctly different mean L-color values and variances, which are summarized in the boxplot presented in Figure 62 below.

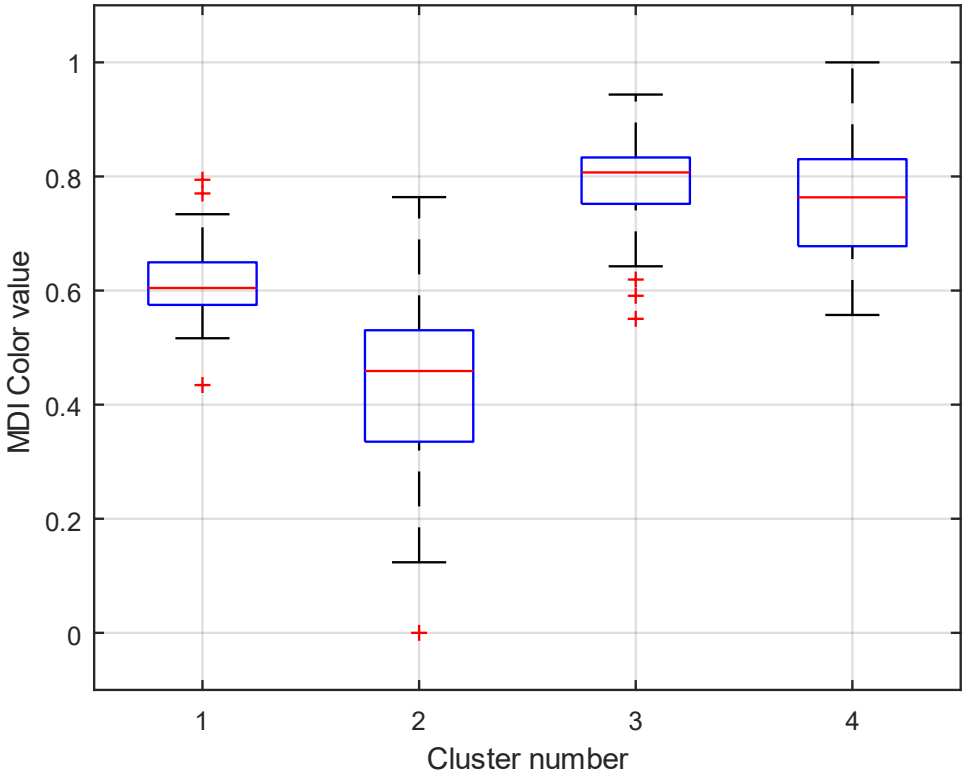


**Figure 62: Boxplot of mean MDI color values and standard deviations for each clusters**

Figure 62 clearly demonstrates that clusters one and two exhibited relatively lower MDI color values, with cluster two also showing a very high standard deviation. Clusters three and four, on the other hand, displayed high mean MDI color values, accompanied by low variance within these periods. From these plots, it is evident that the operational mode considered optimal for MDI product color—and thus likely representing near-optimal operating parameters—can be observed in clusters three and four. Therefore, it is possible to identify and establish an operational parameter set that enables the real industrial system to operate with lower variance and higher average MDI color values.

### 6.3.4 Feature selection

Each feature selection method summarized in Section 6.2.2 was applied to the independent variables, which had already been adjusted with time delays. The results are presented in Figures 63 - 66 below. In these figures, the upper plot displays the independent

parameters in their original order, while the lower plot shows the independent variables arranged according to their order of importance.

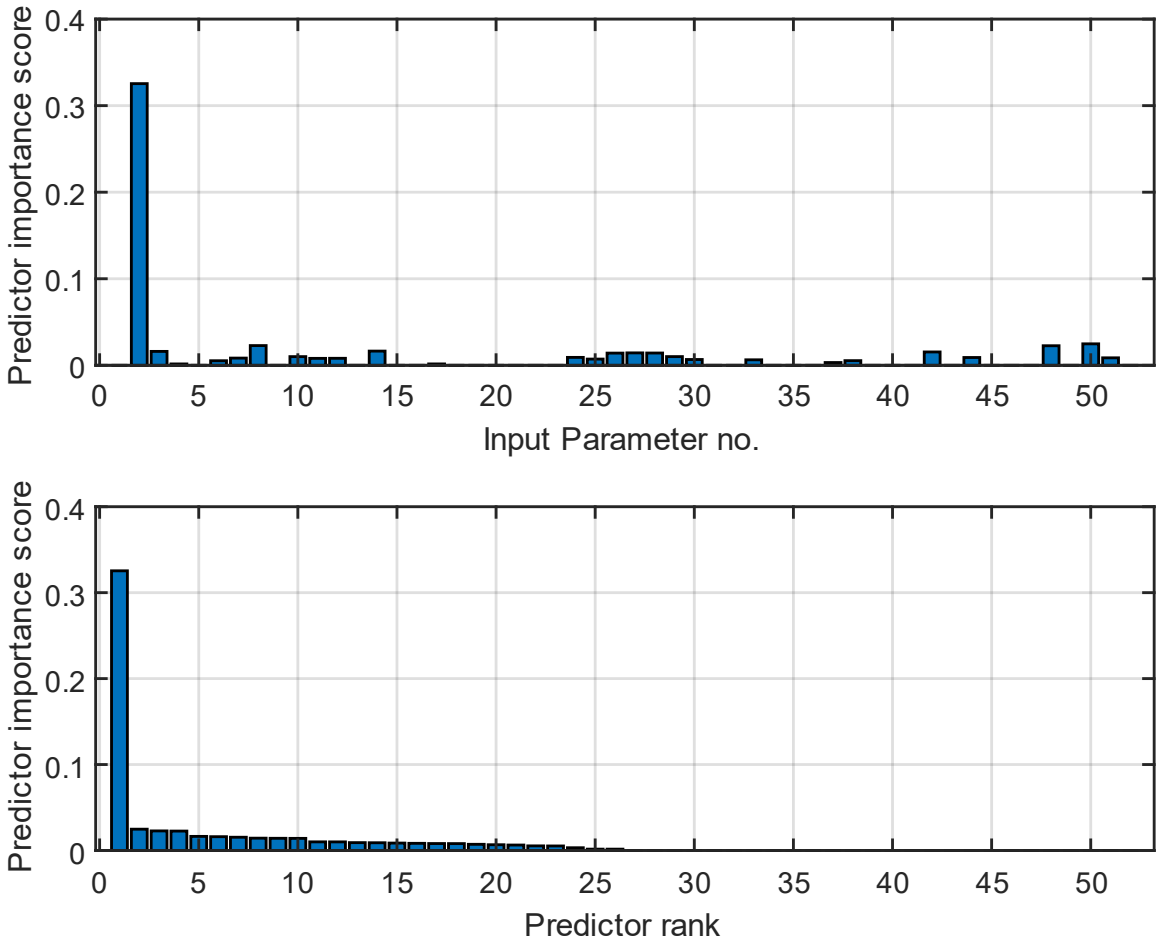


**Figure 63: Results of MRMR method based feature selection. Upper plot shows the independent parameters in order, while the lower plot shows the independent variables in order of importance**

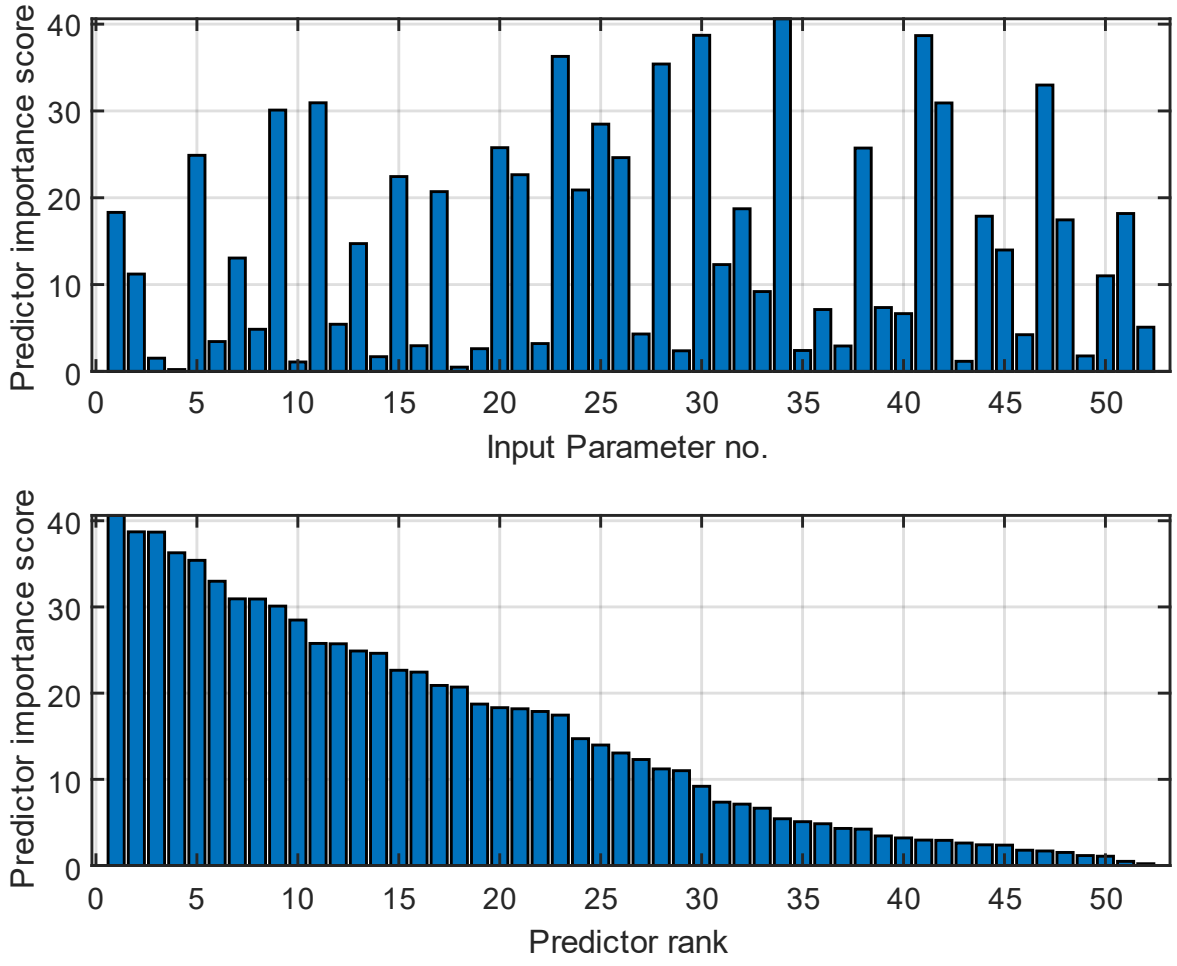


**Figure 64: Results of F-test method based feature selection. Upper plot shows the independent parameters in order, while the lower plot shows the independent variables in order of importance**

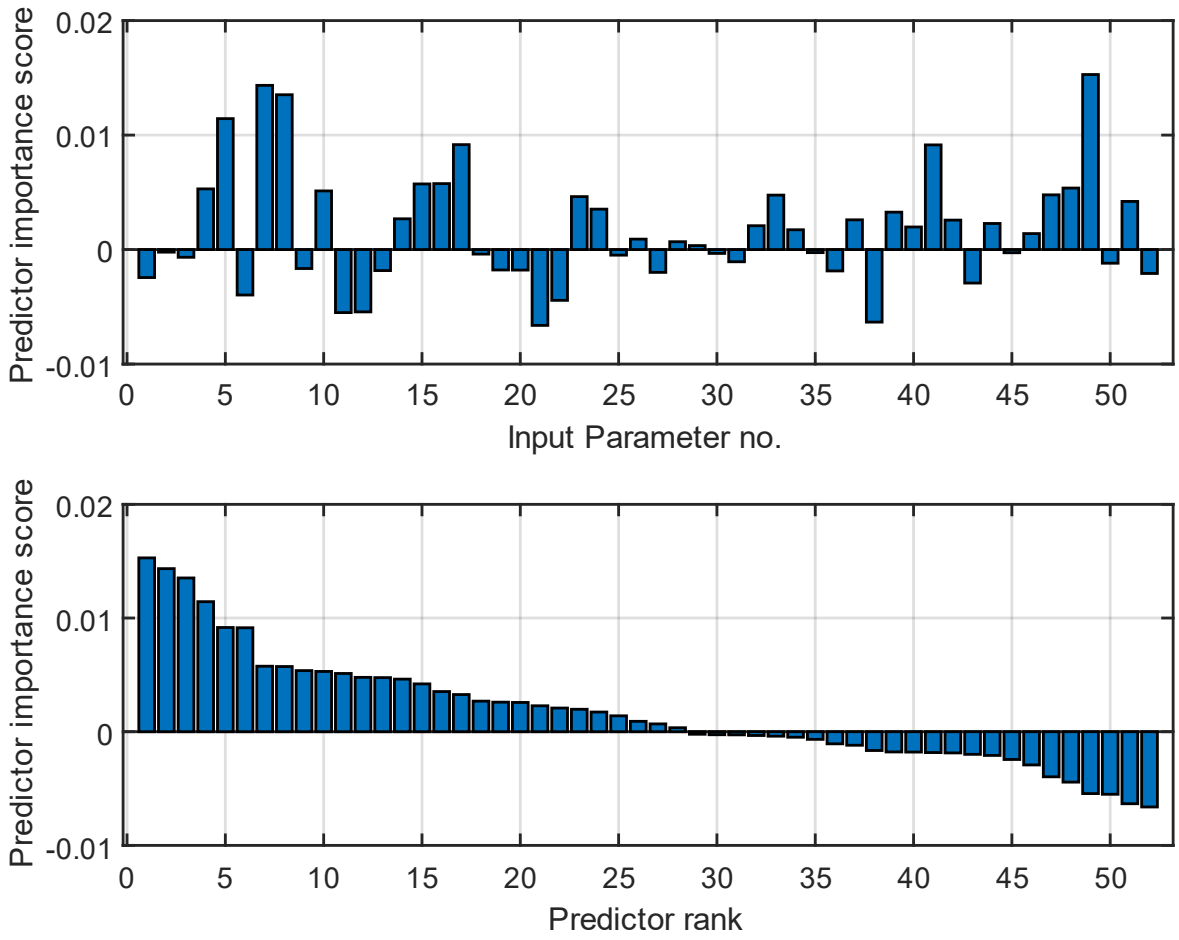


**Figure 65: Results of RReliefF method based feature selection. Upper plot shows the independent parameters in order, while the lower plot shows the independent variables in order of importance**

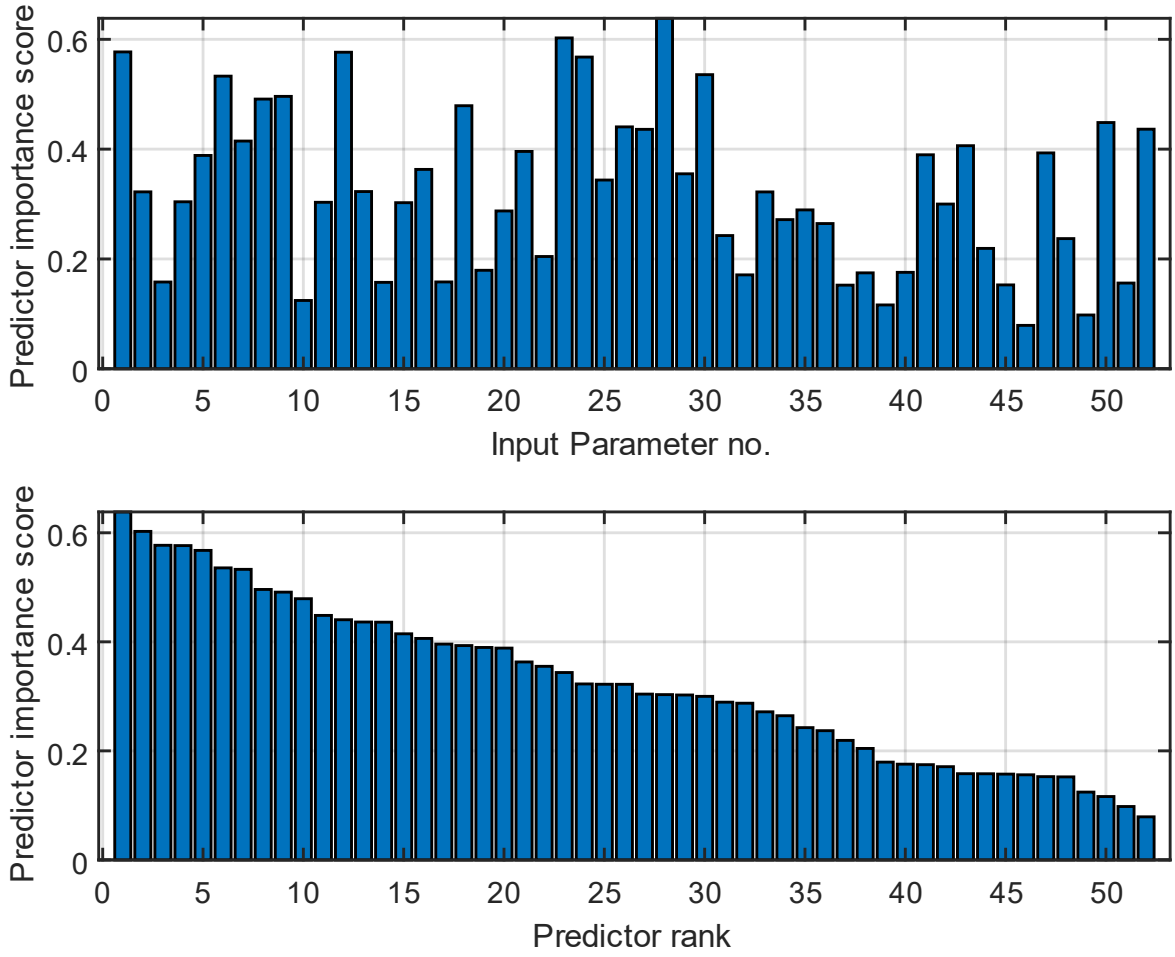


**Figure 66: Results of Correlation-based feature selection. Upper plot shows the independent parameters in order, while the lower plot shows the independent variables in order of importance**

Comparing the results of different feature selection methods reveals discrepancies in which parameters are categorized as highly important, important, or unsuitable for predictive purposes. Due to these differences, we developed a method to compare the results and select a feature set that may be optimal for model training. The determination of the optimal feature set was carried out in two steps.

In the first step, we counted how many times each feature was selected by each feature selection method. As shown in Figures 63 – 66, the MRMR method identified 26 features as suitable for prediction, while the ReliefF method identified 28 features with a positive sign. For the other feature selection methods, it is not possible to definitively determine whether a given parameter is suitable for prediction in the same manner, as the less important parameters are simply de-weighted and classified as less important. Based on these results, the 26 most important features identified by the MRMR method were selected for the next step of feature selection across all four methods.

In the second step, we examined which features were identified as the most important 26 by each method in order to compile an aggregated optimal feature set. If an independent parameter was identified as important by more than one model, it is likely that the feature is indeed important for estimating the MDI color. Conversely, if an independent parameter was not selected by any model as one of the top 26 out of all 52, then that parameter is likely to be less important for the regression models.

Figure 67 below shows how many of the 52 independent parameters were selected as part of the top 26 by each model, arranged by the frequency with which they were selected. It also illustrates how many independent parameters were selected four times, three times, etc.

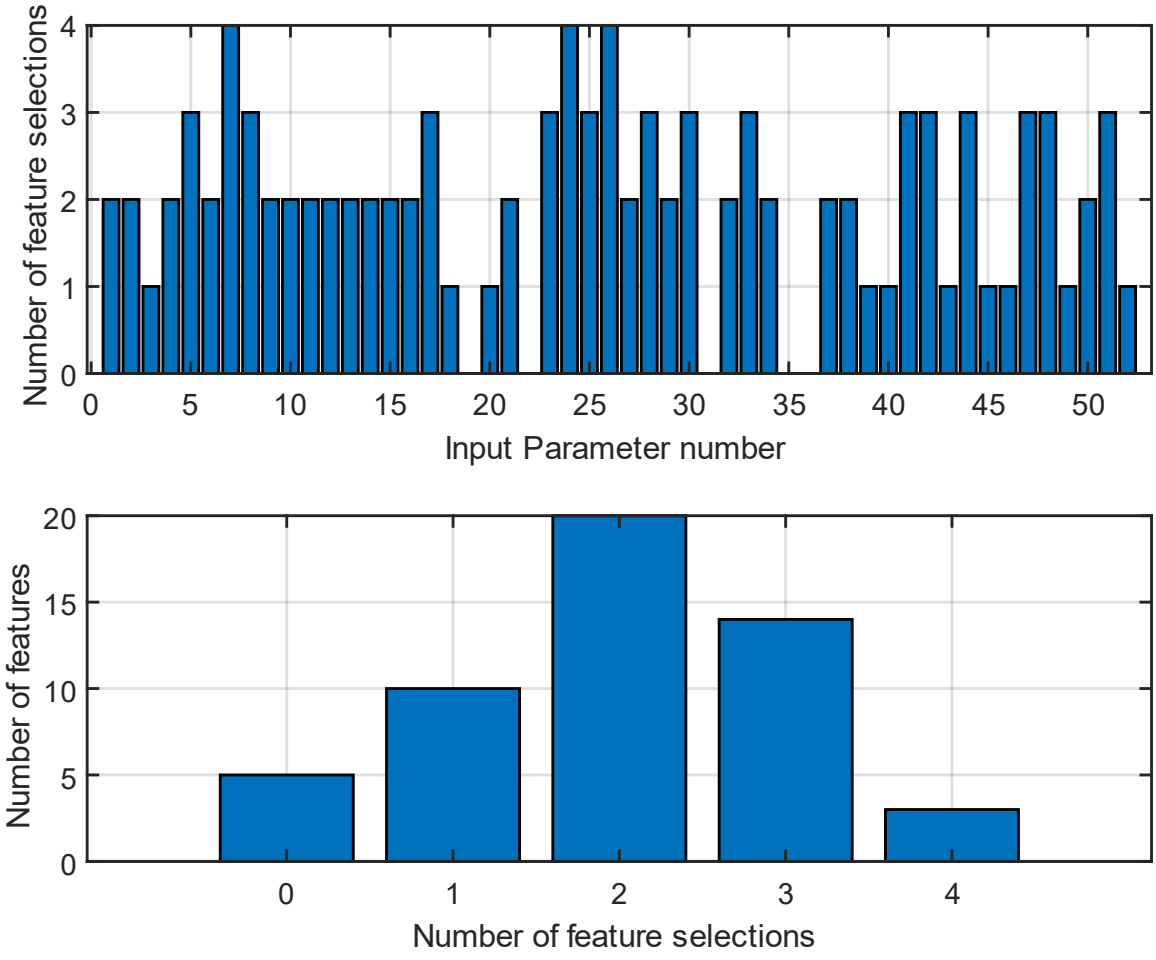


**Figure 67: Cumulative results of the feature selection methods. Upper plot shows how many times an independent parameter was selected as important feature by feature selection methods, while the lower plot shows how many independent parameters were selected by how many feature selection methods**

Of the four feature selection methods tested, all four identified 3 independent parameters as important, while three methods found 14 additional parameters to be important according to

our definition. By choosing a combination of 17 key and optimal features, an aggregated set of 17 different features is created for the feature set of the Aggregated method, thereby providing five distinct feature sets for model training.

The five different feature sets, derived from the various methods, were tested and evaluated by training machine learning models to determine which feature sets would achieve the highest performance. To compare the performance of the models trained using these five feature selection methods, the top 17 parameters out of 52 were selected for each method, as the Aggregated method also identified 17 important features. The performances of the machine learning models trained on these different 17-element feature sets, considered optimal by the five different feature selection methods, were compared. The results were then used to validate which feature set produced the best-performing machine learning models. The results are presented in Table 19 below.

**Table 19: Comparison of the performances of machine learning models trained with optimal feature sets calculated by different feature selection methods**

| Model | MRMR | | F-test | | RReliefF | | Correlation-based | | Aggregated | |
|---|---|---|---|---|---|---|---|---|---|---|
| | $\bar{R}^2_{train.}$ | $\bar{R}^2_{valid.}$ | $\bar{R}^2_{train.}$ | $\bar{R}^2_{valid.}$ | $\bar{R}^2_{train.}$ | $\bar{R}^2_{valid.}$ | $\bar{R}^2_{train.}$ | $\bar{R}^2_{valid.}$ | $\bar{R}^2_{train.}$ | $\bar{R}^2_{valid.}$ |
| **LR** | 0,665 | 0,555 | 0,694 | 0,594 | 0,633 | 0,473 | 0,667 | 0,576 | 0,648 | 0,529 |
| **RT** | 0,632 | 0,271 | 0,683 | 0,421 | 0,755 | 0,292 | 0,779 | 0,412 | 0,737 | 0,348 |
| **NN** | 0,710 | 0,565 | 0,730 | 0,590 | 0,748 | 0,424 | 0,621 | 0,451 | 0,707 | 0,520 |
| **SVM** | 0,627 | 0,645 | 0.670 | 0,664 | 0,595 | 0,595 | 0,642 | 0,631 | 0,626 | 0,615 |
| **GPR** | 0,863 | 0,859 | 0,778 | 0,765 | 0,829 | 0,823 | 0,804 | 0,798 | 0,788 | 0,781 |

As seen from the results in Table 19, all machine learning models performed optimally with the feature set defined by the MRMR method. Among them, the Gaussian Process Regression model achieved the best performance.

### 6.3.5 Model training & optimisation

After preliminarily analyzing, processing, and filtering the data, the normalized and validated datasets were used to train the models listed previously. The performance of these models was tested and compared on three parameters using the feature set derived from the MRMR feature selection method:

- Mean Average Error (MAE)
- Root Mean Square Error (RMSE)
- Coefficient of determination ($R^2$)

In all cases, five-fold cross-validation was applied to the trained models, and where applicable, Bayesian hyperparameter optimization was also performed to maximize model performance (e.g., for Neural Networks, SVM, etc.). The results obtained with the set of 17 independent variables are displayed in Figures 68-70 and summarized in the boxplot figures in Table 20 below.

**Table 20: Results of machine learning models for the optimised parameter set.**

| Model | $\overline{\mathbf{MAE}}_{\mathbf{train.}}$ | $\overline{\mathbf{MAE}}_{\mathbf{valid.}}$ | $\overline{\mathbf{RMSE}}_{\mathbf{train.}}$ | $\overline{\mathbf{RMSE}}_{\mathbf{valid.}}$ | $\overline{\mathbf{R}}^{\mathbf{2}}_{\mathbf{train.}}$ | $\overline{\mathbf{R}}^{\mathbf{2}}_{\mathbf{valid.}}$ |
|---|---|---|---|---|---|---|
| **LR** | 0,093 | 0,105 | 0,117 | 0,134 | 0,665 | 0,555 |
| **DT** | 0,092 | 0,134 | 0,117 | 0,171 | 0,632 | 0,271 |
| **NN** | 0,086 | 0,105 | 0,109 | 0,132 | 0,710 | 0,565 |
| **SVM** | 0,095 | 0,094 | 0,124 | 0,120 | 0,627 | 0,645 |
| **GPR** | 0,059 | 0,059 | 0,075 | 0,075 | 0,863 | 0,859 |

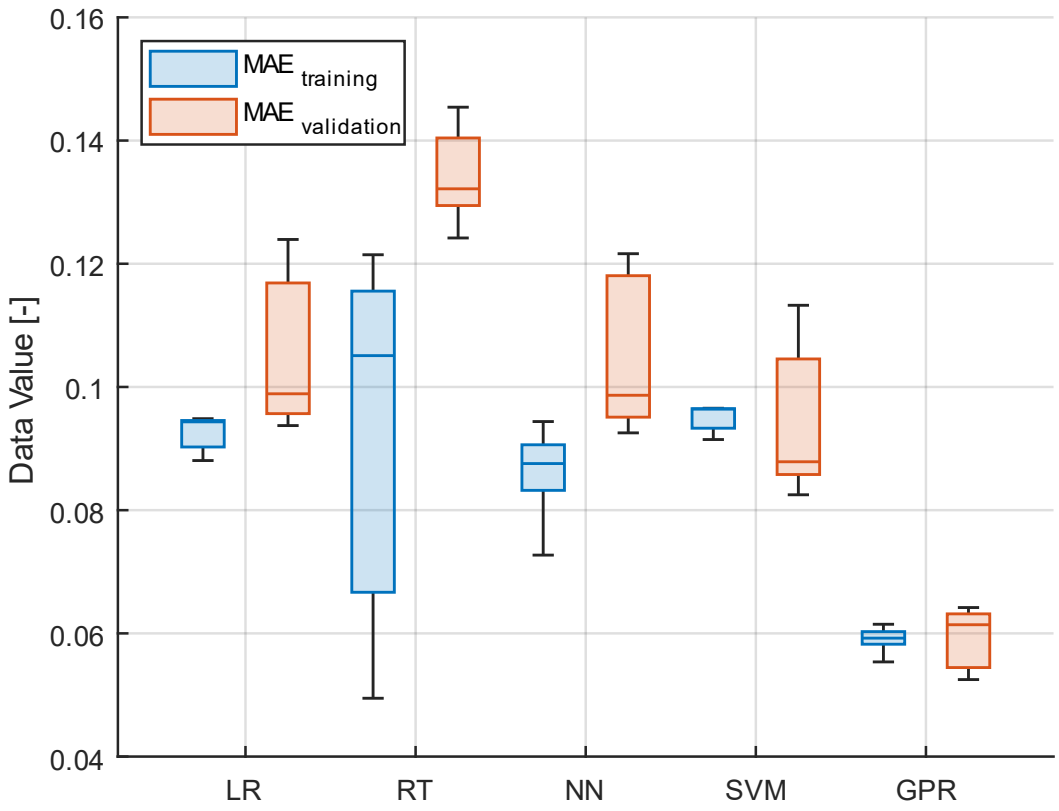


**Figure 68: Calculated Mean Average Errors for all trained models. LR – Linear Regression, RT – Regression Tree, NN – Neutral Network, SVM – Support Vector Machine, GPR – Gaussian Process Regression**

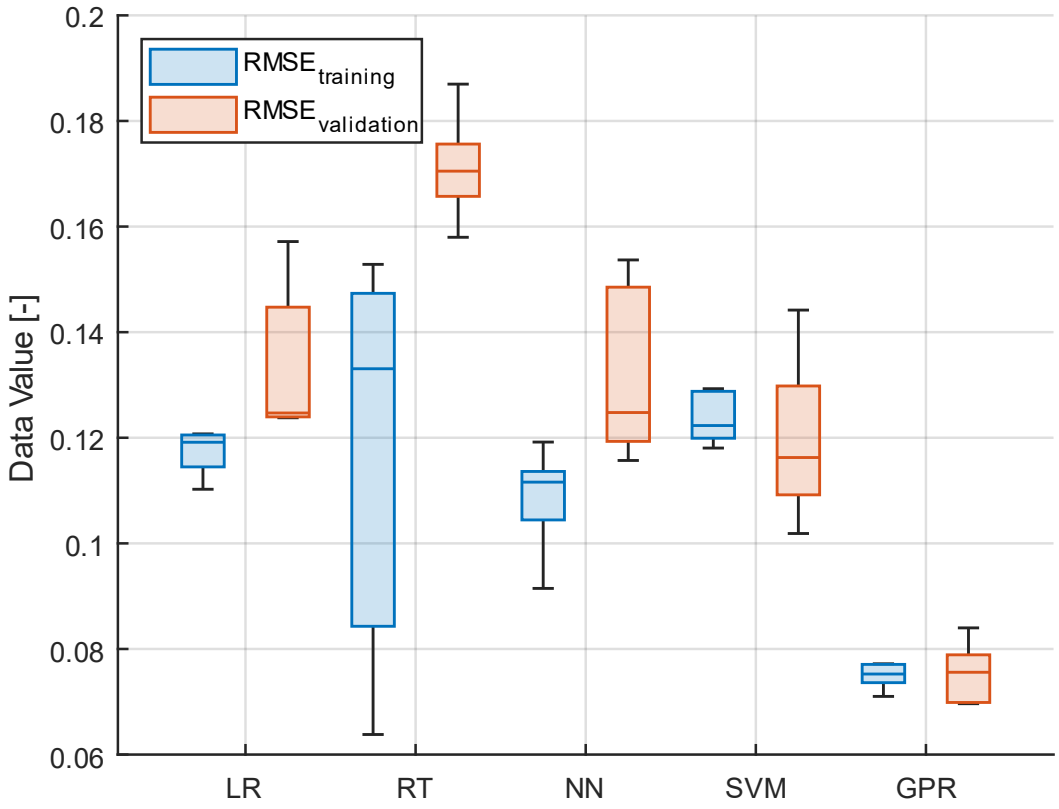


**Figure 69: Calculated Root Mean Square Error for all trained models. LR – Linear Regression, RT – Regression Tree, NN – Neutral Network, SVM – Support Vector Machine, GPR – Gaussian Process Regression**

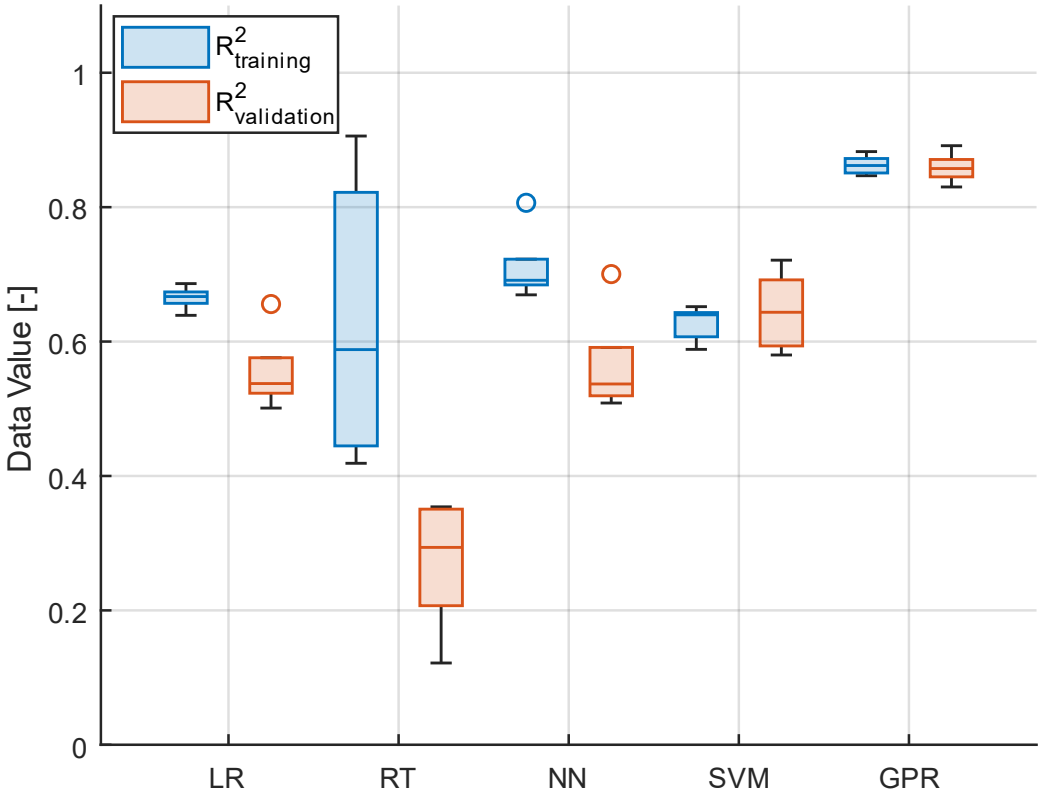


**Figure 70: Calculated R2 values for all trained models. LR – Linear Regression, RT – Regression Tree, NN – Neutral Network, SVM – Support Vector Machine, GPR – Gaussian Process Regression**

From the results shown in Table 20 and Figures 68 – 70, it can be seen that the GPR model clearly performed best, both in terms of training and validation results. The measured and GPR model predicted values are plotted in Figure 71 below.

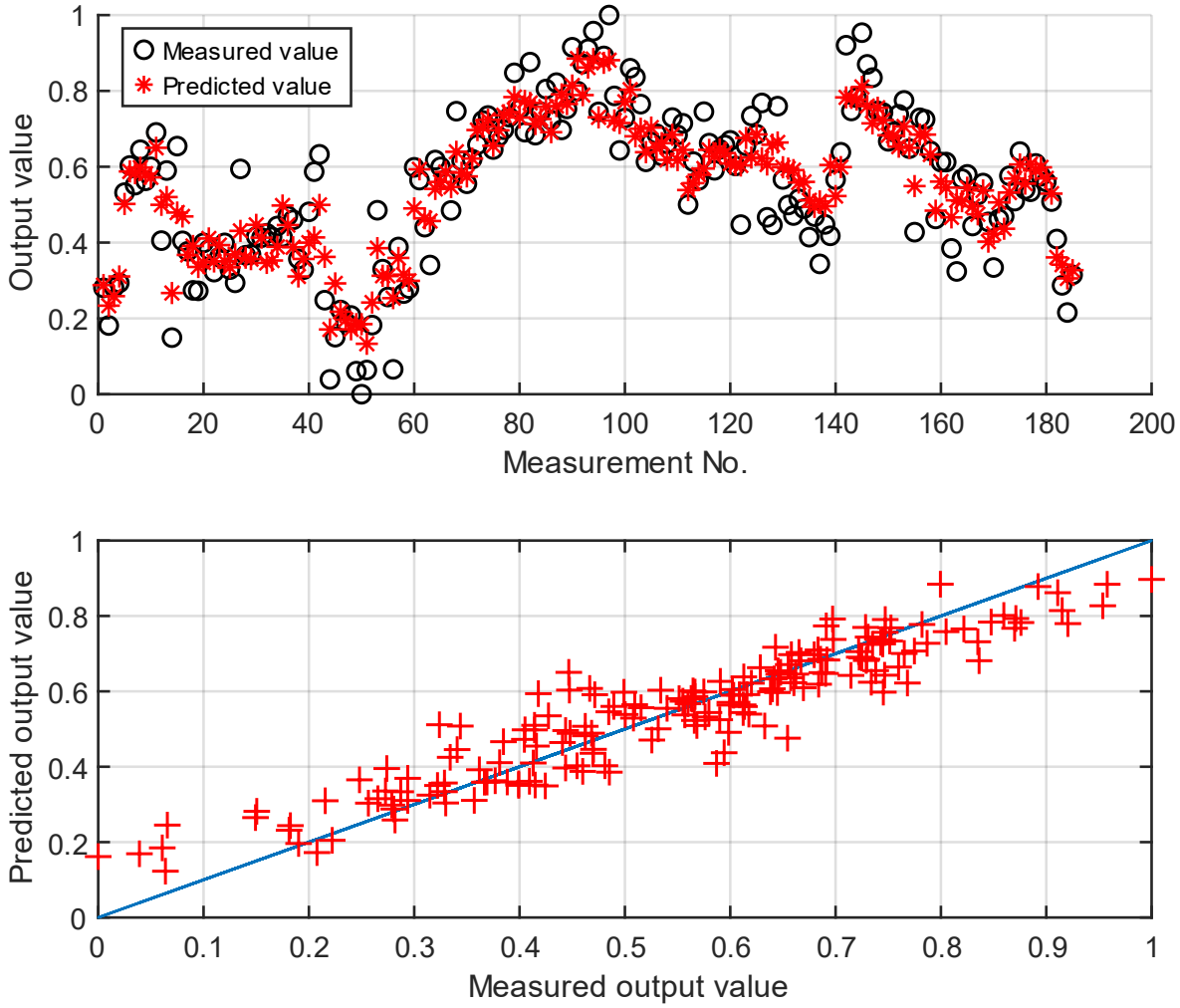


**Figure 71: Comparison of the best performing GPR model predictions with measured output values**

The identified parameters of the proposed GPR-based soft sensor model are presented in Table 21 below:

**Table 21: Identified parameters of the top-performing GPR-based soft sensor model**

| Model parameter | Value or description |
|---|---|
| Kernel Function | Squared Expontential |
| $\sigma_L$ | 0.7007 |
| $\sigma_F$ | 0.2144 |
| Basis Function | Constant |
| $\beta$ | 0.4491 |
| $\sigma$ | 0.0941 |
| Response Transformation | none |

### 6.3.6 Defining explainable optimal operating parameters

After selecting the best-performing machine learning model, the GPR model was combined with an extreme value search algorithm, specifically a genetic algorithm (GA), and additional tests were conducted. An optimal combination of operating parameters was determined based on the model results to maximize the MDI color.

Using the GPR model, we first generated Partial Dependence Plots (PDPs) for each operating parameter, which can be used to visualize the relationship between each feature and the predicted output variable of the model [142]. PDPs illustrates how predictor variables are related to predicted responses in a trained regression model and they can make it easier to understand how changes in each operating parameter affect the MDI color, whether the relationship is linear or nonlinear, and whether the function is monotonic or not. The PDPs generated using the GPR model are shown in Figure 72 below.

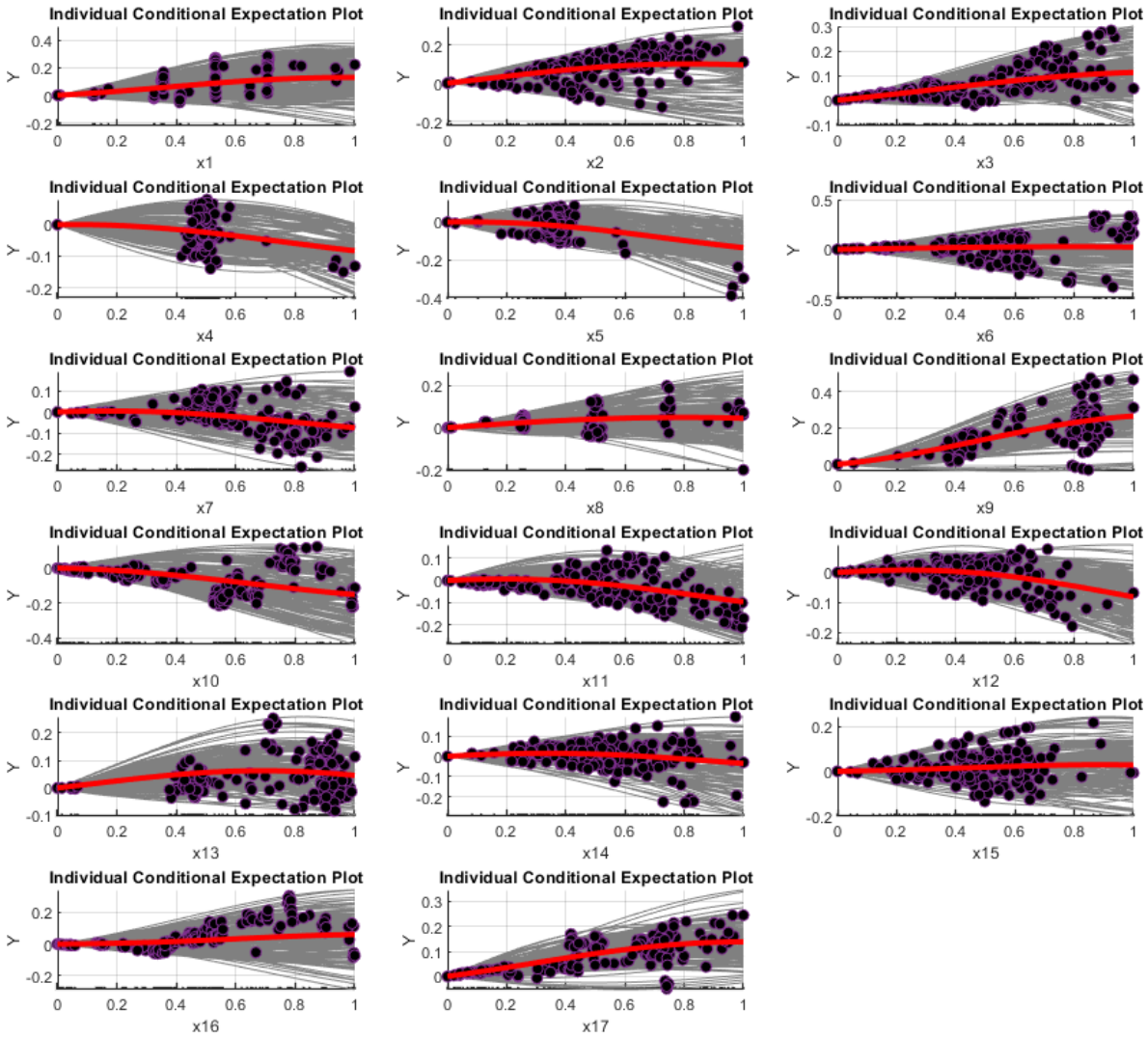


**Figure 72: Partial Dependence Plot calculated for each independent parameters based on the best performing GPR model**

The results in Figure 72 verify the relationships identified by the machine learning model with the domain-based operational experience and knowledge, and the visualisation allows to easily identify the parameters that have a stronger influence on the MDI color.

The findings of the sensitivity analysis indicate noticeable non-linearity in certain parameters, and in some instances, specific key input parameters might have both beneficial and adverse effects on the forecasted variable, such as the MDI color, which is influenced by the system's condition (e.g., $x_8$, $x_{15}$, etc.). Conversely, some parameters exhibit a distinctly positive effect (e.g., $x_9$) or a distinctly negative effect (e.g., $x_5$).

After analyzing the data from the PDPs, we optimized the operational parameters using the genetic algorithm (GA) to maximize the MDI color. Since the GA is fundamentally a minimum-seeking algorithm, the normalized value of the MDI color was assigned a negative

sign to ensure that the algorithm would maximize it. Only lower and upper constraints were defined in the solution; no equalities, inequalities, or nonlinear constraints were specified. The normalized, optimal result for the MDI color, along with the hyperparameters of the model, are presented in Table 22 below:

**Table 22: The normalised, optimal result of MDI color and the hyperparameters of the genetic algorithm.**

| **Parameter** | **Value** |
|---|---|
| Population size | 50 |
| Crossover function | Crossover scattered |
| Crossover fraction | 0.5 |
| Creation function | Uniform individual generation |
| Max generations | 100 |
| Linear inequality constraints | None |
| Linear equality constraints | None |
| Nonlinear constraints | None |
| Lower bounds | 0 |
| Upper bounds | 1 |
| **Result** | **1.057** |

The result of the GA extreme value searching algorithm showed a ~6% increase in the MDI color compared to the maximum color value measured previously when the 17 operating parameters were set to optimal. This result is an extrapolation from the previous data, which were normalized to a 0-1 scale, and should therefore be treated with some caution. However, it clearly demonstrates that a significant improvement can be achieved by optimizing the operational parameters. The normalized values of the operational parameters optimized by the GA are shown in Figure 73 below:

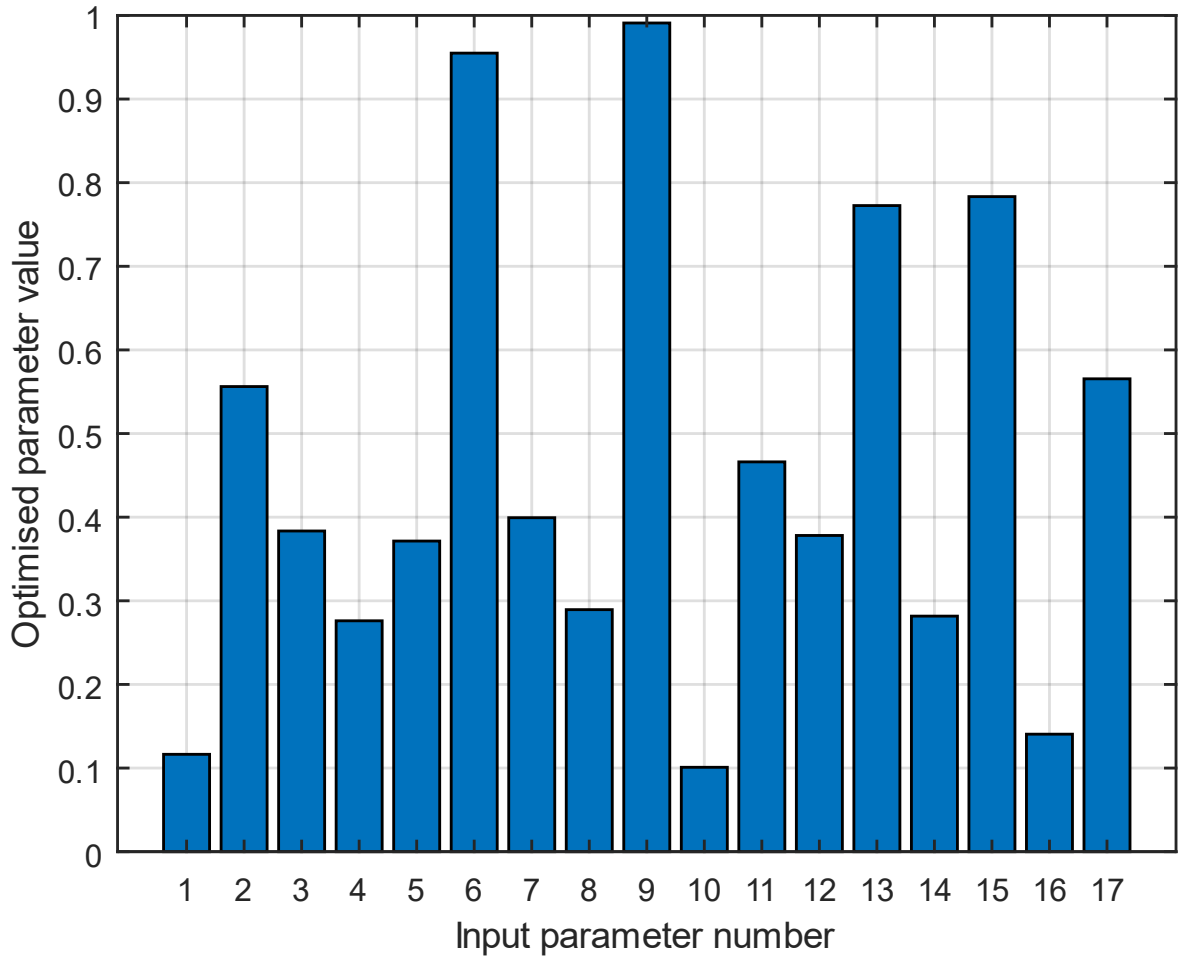


**Figure 73: Optimised operational parameters based on genetic algorithm results**

As shown in Figure 73, each of the 17 parameters to be optimized assumes values between a minimum of 0 and a maximum of 1, due to the strong cross-correlations observed among the operational parameters. The results further substantiate that the individual fine-tuning or adjustment of each operational parameter is unlikely to yield significant improvements in the MDI color. Moreover, owing to the cross-correlations, without a comprehensive understanding of the entire set of operational parameters, it remains indeterminate whether the increase or decrease of any individual parameter will exert a positive or negative influence on the MDI color within the industrial system. However, with the model we have developed, and by considering the complete set of parameters, the MDI color can be estimated with a high degree of accuracy through systematic fine-tuning.

I also examined the robustness of the optimal results predicted by the GA, specifically how changes in the parameters of the optimal parameter set affect the objective function value. During the tests, we varied each operational parameter by ±20% relative to the optimal values originally determined by the GA. This allowed us to gain insights not only into the reliability

and robustness of the optimal solution but also into how deviations of individual parameters from the optimal values influence the objective function value. In other words, we were able to assess which parameters most significantly impact the color of the MDI near the optimum. The results are shown in Figure 74.

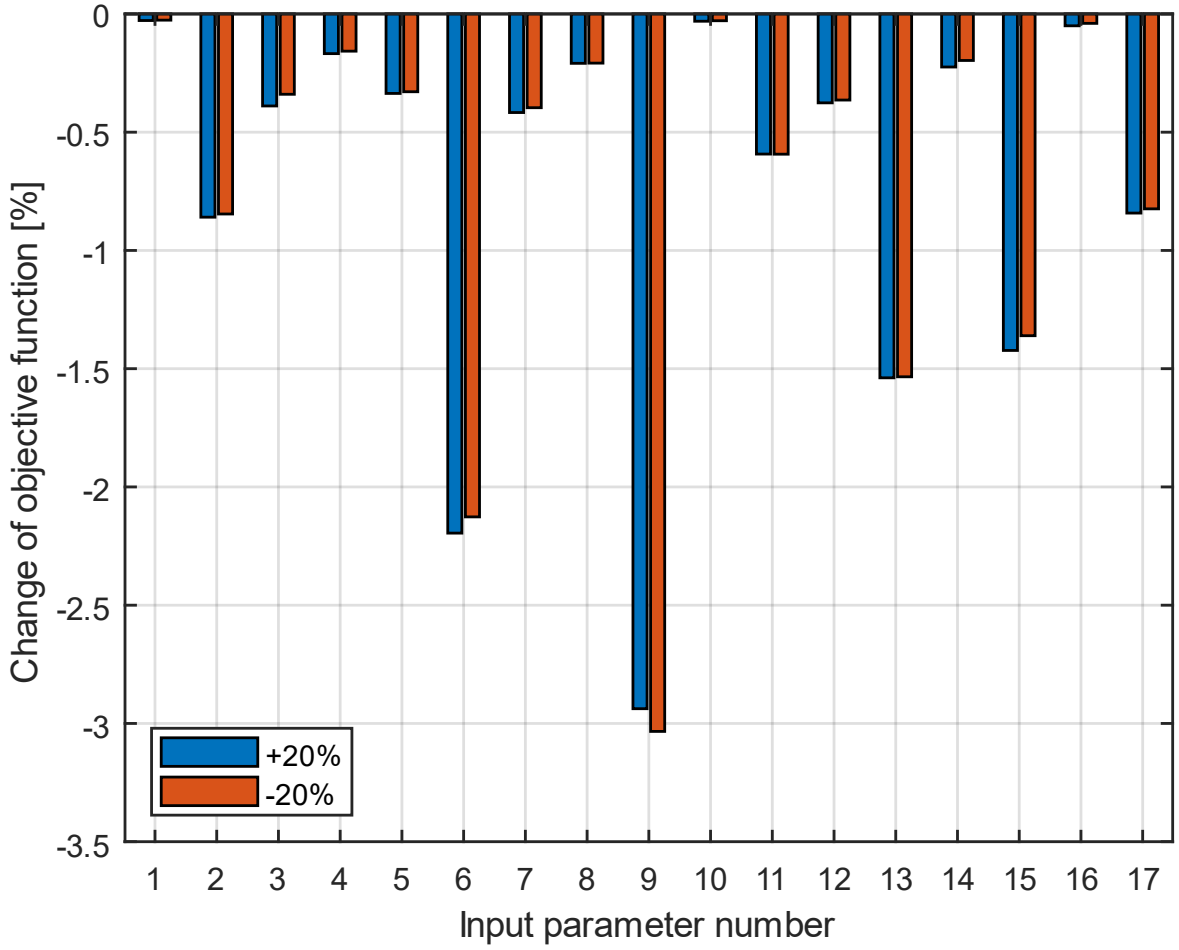


**Figure 74: The relative change in the objective function value compared to the optimal objective function value when the values of different input parameters are varied by ±20%**

As shown in Figure 74 and as expected, when the operating parameters were modified by ±20%, we consistently obtained lower objective function values compared to the optimal objective function value. Furthermore, the results indicate that modifications to the 6th and 9th operating parameters had the most significant impact on the objective function value, while for several other parameters, such as the 1st, 10th, and 16th, no significant degradation in the objective function value was observed compared to the optimal value.

## 6.4 Conclusion

In Section 6, we developed and validated machine learning models on a real industrial system to accurately predict MDI color variation based on real industrial operational parameters. By summarizing and understanding the potential side reactions and by-products forming in the industrial system, we identified key operational parameters for model training and collected operational data. After data collection, cleaning, and sorting, a Pearson correlation-based method was used to estimate the time delays of these parameters. With optimal time delays, we identified the clustering of different operational states using Principal Component Analysis.

To enhance the accuracy of the machine learning models and simplify their transfer to real industrial systems with fewer input parameters, the most important operational parameters were selected by combining the results of five feature selection methods: MRMR, F-test, ReliefF, Correlation-based and an Aggregated feature set based on the other four method's results. This feature selection step reduced the original set of 52 parameters to 17 key parameters. Using the cleaned data, estimated time delays, and optimized feature set, we trained five different machine learning models—Linear Regression, Regression Tree, Neural Network, Support Vector Machine, and Gaussian Process Regression, where the GPR model performed the best.

Based on the GPR model's results, we calculated and visualized partial dependence plots as a form of sensitivity study, which were then compared with real industrial experience and analyzed by industrial experts. To identify the optimal set of operational parameters to maximise MDI color, we applied genetic algorithms based on the GPR model's predictions. The optimized 17 operating parameters were then subjected to a sensitivity analysis in order to assess the robustness of the optimal solution and to evaluate the sensitivity of the objective function to changes in each parameter.

Looking ahead, it is recommended to expand the training domains and data sets of the machine learning model to support operation in all potential operational states. Continuous comparison between measured and predicted color values is also crucial for further validating the accuracy of the predictions and addressing any errors that fall outside the acceptable range. To facilitate the greatest number of comparisons between measured and predicted data, it is also proposed to develop a soft-sensor for online estimation, where the control system of the

industrial system can send operational data to the machine learning model, enabling real-time MDI color predictions and can provide suggestions for operators to further increase MDI color.

Besides the Pearson correlation-based time delay estimation which enhanced the predictive accuracy of the machine learning models in a simple and efficient manner, exploring more intricate and sophisticated time delay estimation methods could further strengthen the predictive capabilities of the models. Therefore, investigating these advanced methods is highly recommended to improve model performance and provide more accurate predictions in complex industrial systems.

# 7 Summary

The dissertation presents a comprehensive, model-driven and data-informed exploration of the industrial production of methylene diphenyl diisocyanate (MDI), with particular emphasis on the synthesis and mechanistic understanding of methylenedianiline (MDA), the principal precursor to MDI. Through the integration of kinetic modeling, machine learning methodologies, and optimization techniques, a both mechanistic and data-driven approaches were developed to enhance the predictability, efficiency, and controllability of MDI manufacturing processes.

A detailed kinetic model was constructed to describe the complex reaction network of MDA synthesis, accounting for critical intermediates, such as ortho- and para-aminobenzylanilines (ABAs), and their transformation into MDA isomers. The model successfully incorporated extended reaction pathways and was validated using both laboratory-scale experiments and available literature data.

The application of Machine Learning models has shown through sensitivity analyses that synthesis parameters – including molar ratios, residence times, and temperature profiles – significantly impact both isomeric distribution and ring structure. These insights allow for the precise identification of operational conditions that promote the formation of preferred isomers, such as 4,4'-MDA, and control by-product formation, ultimately enhancing the quality of downstream MDI products.

To address limitations in conventional modeling approaches, particularly for parameters such as MDI product color, a suite of soft sensors was developed using advanced machine learning algorithms. Techniques including Linear Regression, Regression Trees, Neural Networks, Support Vector Machines, and Gaussian Process Regression were employed. Feature selection methods such as MRMR, ReliefF, F-test, Correlation-based analysis and an Aggregated approach were combined with time-delay estimation to identify the most impactful process variables. GPR emerged as the most accurate and interpretable modeling technique, demonstrating robust performance even under complex, nonlinear conditions.

An optimization process of industrial operational parameters was accomplished through the integration of the best performing trained and validated Gaussian Process Regression model with Genetic Algorithms. The deployment of Genetic Algorithms, based on the superior predictive capacity of the most effective GPR model, enabled the identification of the optimal

operational parameters for a real industrial facility to improve the color quality of the MDI product mixtures.

Altogether, this work exemplifies the synergy between mechanistic and data-driven modeling. It advances the state-of-the-art in chemical process engineering by delivering validated, explainable, and scalable tools for real-time process optimization in industrial MDI production. The methodologies developed herein are adaptable and hold promise for broader application across other complex, multi-step reaction systems within the chemical manufacturing domain.

The outcomes of this research not only contribute to enhanced product quality and process sustainability but also align with the strategic goals of digital transformation and Industry 4.0 within the chemical industry.

# Theses

**Thesis #1.** I expanded the reaction system of existing models reported in the literature for the synthesis of Methylenedianiline by incorporating additional components and reaction pathways, I achieved enhanced model accuracy and reduced information loss.

- I created an extended reaction network by incorporating three additional components – N-methylbenzene, aniline, and formaldehyde – enabling the usage of the aniline/formaldehyde molar ratio, an industrial production parameter, through the kinetic model.
- The incorporation of N-methylbenzene and the extended reaction network enabled to accurately represent the formation of molecules containing an odd number of rings, thereby overcoming the limitation of prior models which were constrained to even-ring structures.
- I demonstrated that the developed extended reaction network including new reaction pathways and components more accurately describe the Methylenedianiline synthesis process, as evidenced by improved fit with both concentration trajectories and key performance indicators.

**Related publications: 1, 4**

**Thesis #2.** I justified that machine learning models efficiently support the understanding of Methylenedianiline synthesis variables on laboratory-scale, aiming to examine the impact of variables on Methylenedianiline quality and to assist large-scale industrial production.

- I determined the key synthesis parameters and their impact with Shapley values for all Methylenedianiline quality parameters, demonstrating that the increase of HCl / aniline molar ratio or water / aniline molar ratio increases by-product and P-P isomer formation, while a higher aniline / formaldehyde molar ratio considerably decreases them. The 2-ring content and O-P isomer content can be increased with higher aniline / formaldehyde molar ratio, lower condensation reaction time and temperature, however, adjusting these parameters in the opposite way results in an increased amount of polymer products
- I implemented a new, three-level hierarchical method for outlier analysis, designed to pinpoint the outliers of laboratory measurements with optimal efficiency.

**Related publications: 2, 5, 6**

**Thesis #3.** Through the proposed intervention strategies, I justified that soft-sensor models can be well applied for Methylenediphenyl diisocyanate color prediction based on real industrial data in an explainable way, illustrated the non-linear behavior of the system and identified the best operational parameters for an industrial Methylenediphenyl diisocyanate production plant.

- I proposed 3 critical intervention strategies within the framework of the industrial production to reduce side reactions: avoid the formation of urea compounds, limit the formation of chloroformamidine-N-carbonyl chloride compounds, and prevent the breakdown of chloroformamidine-N-carbonyl chloride compounds into dichlorides.
- I proved that the accuracy of the machine learning models can be improved by implementing time delays for each industrial operational parameters, which were determined with the use of correlation analysis between input and output data
- I verified the significant non-linearity and complexity of the system with the visualisation of the relationships between each feature on Partial Dependence Plots, which demonstrates that each parameter has local optima depending on the values of other investigated operational parameter and they are highly cross-correlated with each other.
- I identified an explainable and optimal set of operational parameters for the industrial system, achieving an Methylenediphenyl diisocyanate color value close to 1.06 at a normalized level exceeding the initially explored range.

**Related publications: 3, 7, 8**

# Publications related to theses

## Articles in international journals

1. Horváth, G., Kummer, A., Kozár, Z., & Varga, T. (2023). Exploration and model-based analysis of reaction mechanisms related to the formation of methylenedianiline. *Industrial & Engineering Chemistry Research*, *62*(10), 4297-4311.
   Scimago Journal Ranking: Q1, Impact Factor: 4.009
2. Horváth, G., Trujillo, V. J., Réti, J., Kozár, Z., Kummer, A., & Varga, T. (2024). Exploring the essential features influencing the synthesis of methylenedianiline to support industrial processes. *Chemical Engineering Research and Design*, *208*, 626-647.
   Scimago Journal Ranking: Q2, Impact Factor: 4.345
3. Horváth, G., Trujillo, V. J., Réti, J., Kozár, Z., Varga, T., & Kummer, A. (2025). Soft-sensor development for product quality estimation with time delay and feature selection in industrial MDI production. *Chemical Engineering Journal Advances*, *22*, 100751.
   Scimago Journal Ranking: Q1, Impact Factor: 8.093

## Conference abstracts

4. Horváth, G., Varga, T., & Kummer, A. (2022). Exploration and model-based analysis of reaction mechanisms related to the formation of methylenedianiline, *50. Műszaki Kémiai Napok Jubileumi Konferencia: 50th Engineering Chemistry Jubilee Conference*, Veszprém, Hungary, pp. 30
5. Horváth, G., Trujillo, V. J., Réti, J., Kozár, Z., Varga, T., & Kummer, A. (2023). Development and comparison of machine learning models to analyse the synthesis of methylenedianiline, *PhD hallgatók anyagtudományi napja XXIII: Materials science day XXIII*, Veszprém, Hungary, pp. 5
6. Horváth, G., Varga, T., & Kummer, A. (2023). Data-driven modeling of methylenedianiline (MDA) synthesis, *Műszaki Kémiai Napok 2023: Engineering Chemistry Conference 2023*, Veszprém, Hungary, pp. 43
7. Horváth, G., Varga, T., & Kummer, A. (2024). Identification and model-based analysis of color problems in the synthesis of methylenediphenyl diisocyanate products, *Műszaki Kémiai Napok 2024: Engineering Chemistry Conference 2024*, Veszprém, Hungary, pp. 41

8. Horváth, G., Trujillo, V. J., Réti, J., Kozár, Z., Varga, T., & Kummer, A. (2024). Machine learning-based product quality estimation with time delay-based feature selection in industrial MDI production, *PhD hallgatók anyagtudományi napja XXIV: Materials science day XXIV*, Veszprém, Hungary, pp. 13

# References


(1) Rasmuson, A.; Andersson, B.; Olsson, L.; Andersson, R. *Mathematical Modeling in Chemical Engineering*, 1st ed.; Cambridge University Press, 2014. https://doi.org/10.1017/CBO9781107279124.

(2) Bertolini, M.; Mezzogori, D.; Neroni, M.; Zammori, F. Machine Learning for Industrial Applications: A Comprehensive Literature Review. *Expert Syst. Appl.* **2021**, *175*, 114820. https://doi.org/10.1016/j.eswa.2021.114820.

(3) Pehrsson, L.; Ng, A. H. C.; Stockton, D. Industrial Cost Modelling and Multi-Objective Optimisation for Decision Support in Production Systems Development. *Comput. Ind. Eng.* **2013**, *66* (4), 1036–1048. https://doi.org/10.1016/j.cie.2013.08.011.

(4) Horváth, G.; Kummer, A.; Kozár, Z.; Varga, T. Exploration and Model-Based Analysis of Reaction Mechanisms Related to the Formation of Methylenedianiline. *Ind. Eng. Chem. Res.* **2023**, *62* (10), 4297–4311. https://doi.org/10.1021/acs.iecr.2c03450.

(5) Zhang, Q.; Shah, N.; Wassick, J.; Helling, R.; Van Egerschot, P. Sustainable Supply Chain Optimisation: An Industrial Case Study. *Comput. Ind. Eng.* **2014**, *74*, 68–83. https://doi.org/10.1016/j.cie.2014.05.002.

(6) Forbes, M. G.; Patwardhan, R. S.; Hamadah, H.; Gopaluni, R. B. Model Predictive Control in Industry: Challenges and Opportunities. *IFAC-Pap.* **2015**, *48* (8), 531–538. https://doi.org/10.1016/j.ifacol.2015.09.022.

(7) Pal, P. K.; Hens, A.; Behera, N.; Lahiri, S. K. Digital Twins: Transforming the Chemical Process Industry—A Review. *Can. J. Chem. Eng.* **2025**, cjce.25611. https://doi.org/10.1002/cjce.25611.

(8) Danglade, F.; Pernot, J.-P.; Véron, P.; Fine, L. A Priori Evaluation of Simulation Models Preparation Processes Using Artificial Intelligence Techniques. *Comput. Ind.* **2017**, *91*, 45–61. https://doi.org/10.1016/j.compind.2017.06.001.

(9) Iavarone, S.; Péquin, A.; Chen, Z. X.; Doan, N. A. K.; Swaminathan, N.; Parente, A. An a Priori Assessment of the Partially Stirred Reactor (PaSR) Model for MILD Combustion. *Proc. Combust. Inst.* **2021**, *38* (4), 5403–5414. https://doi.org/10.1016/j.proci.2020.06.234.

(10) Biglari, A.; Sutherland, J. C. An A-Posteriori Evaluation of Principal Component Analysis-Based Models for Turbulent Combustion Simulations. *Combust. Flame* **2015**, *162* (10), 4025–4035. https://doi.org/10.1016/j.combustflame.2015.07.042.

(11) Zhu, B.; Ren, S.; Wang, P.; Si, F. An Inductive Transfer Regression Framework for Small Sample Modeling in Power Plants. *Chem. Eng. Res. Des.* **2024**, *210*, 112–129. https://doi.org/10.1016/j.cherd.2024.08.020.

(12) Bridewell, W.; Langley, P.; Todorovski, L.; Džeroski, S. Inductive Process Modeling. *Mach. Learn.* **2008**, *71* (1), 1–32. https://doi.org/10.1007/s10994-007-5042-6.

(13) Yang, J.; Aldemir, T.; Smidts, C. A Deductive Method for Diagnostic Analysis of Digital Instrumentation and Control Systems. *IEEE Trans. Reliab.* **2018**, *67* (4), 1442–1458. https://doi.org/10.1109/TR.2018.2864630.

(14) Soares Pinto, F.; Zemp, R.; Jobson, M.; Smith, R. Thermodynamic Optimisation of Distillation Columns. *Chem. Eng. Sci.* **2011**, *66* (13), 2920–2934. https://doi.org/10.1016/j.ces.2011.03.022.

(15) Nascimento, C. A. O.; Giudici, R.; Guardani, R. Neural Network Based Approach for Optimization of Industrial Chemical Processes. *Comput. Chem. Eng.* **2000**, *24* (9–10), 2303–2314. https://doi.org/10.1016/S0098-1354(00)00587-1.

(16) Wu, H.; Zhao, J. Deep Convolutional Neural Network Model Based Chemical Process Fault Diagnosis. *Comput. Chem. Eng.* **2018**, *115*, 185–197. https://doi.org/10.1016/j.compchemeng.2018.04.009.

(17) Mohd Amiruddin, A. A. A.; Zabiri, H.; Taqvi, S. A. A.; Tufa, L. D. Neural Network Applications in Fault Diagnosis and Detection: An Overview of Implementations in Engineering-Related Systems. *Neural Comput. Appl.* **2020**, *32* (2), 447–472. https://doi.org/10.1007/s00521-018-3911-5.
(18) Vo, N. D.; Oh, D. H.; Hong, S.-H.; Oh, M.; Lee, C.-H. Combined Approach Using Mathematical Modelling and Artificial Neural Network for Chemical Industries: Steam Methane Reformer. *Appl. Energy* **2019**, *255*, 113809. https://doi.org/10.1016/j.apenergy.2019.113809.
(19) Rojek, I. Technological Process Planning by the Use of Neural Networks. *Artif. Intell. Eng. Des. Anal. Manuf.* **2017**, *31* (1), 1–15. https://doi.org/10.1017/S0890060416000147.
(20) Asadisaghandi, J.; Tahmasebi, P. Comparative Evaluation of Back-Propagation Neural Network Learning Algorithms and Empirical Correlations for Prediction of Oil PVT Properties in Iran Oilfields. *J. Pet. Sci. Eng.* **2011**, *78* (2), 464–475. https://doi.org/10.1016/j.petrol.2011.06.024.
(21) Alhajree, I.; Zahedi, G.; Manan, Z. A.; Zadeh, S. M. Modeling and Optimization of an Industrial Hydrocracker Plant. *J. Pet. Sci. Eng.* **2011**, *78* (3–4), 627–636. https://doi.org/10.1016/j.petrol.2011.07.019.
(22) Zahedi, G.; Lohi, A.; Mahdi, K. A. Hybrid Modeling of Ethylene to Ethylene Oxide Heterogeneous Reactor. *Fuel Process. Technol.* **2011**, *92* (9), 1725–1732. https://doi.org/10.1016/j.fuproc.2011.04.022.
(23) Balabin, R. M.; Smirnov, S. V. Variable Selection in Near-Infrared Spectroscopy: Benchmarking of Feature Selection Methods on Biodiesel Data. *Anal. Chim. Acta* **2011**, *692* (1–2), 63–72. https://doi.org/10.1016/j.aca.2011.03.006.
(24) Jia, M.; Xu, D.; Yang, T.; Liu, Y.; Yao, Y. Graph Convolutional Network Soft Sensor for Process Quality Prediction. *J. Process Control* **2023**, *123*, 12–25. https://doi.org/10.1016/j.jprocont.2023.01.010.
(25) Zheng, S.; Liu, K.; Xu, Y.; Chen, H.; Zhang, X.; Liu, Y. Robust Soft Sensor with Deep Kernel Learning for Quality Prediction in Rubber Mixing Processes. *Sensors* **2020**, *20* (3), 695. https://doi.org/10.3390/s20030695.
(26) Mowbray, M.; Vallerio, M.; Perez-Galvan, C.; Zhang, D.; Del Rio Chanona, A.; Navarro-Brull, F. J. Industrial Data Science – a Review of Machine Learning Applications for Chemical and Process Industries. *React. Chem. Eng.* **2022**, *7* (7), 1471–1509. https://doi.org/10.1039/D1RE00541C.
(27) Shields, B. J.; Stevens, J.; Li, J.; Parasram, M.; Damani, F.; Alvarado, J. I. M.; Janey, J. M.; Adams, R. P.; Doyle, A. G. Bayesian Reaction Optimization as a Tool for Chemical Synthesis. *Nature* **2021**, *590* (7844), 89–96. https://doi.org/10.1038/s41586-021-03213-y.
(28) MacLeod, B. P.; Parlane, F. G. L.; Morrissey, T. D.; Häse, F.; Roch, L. M.; Dettelbach, K. E.; Moreira, R.; Yunker, L. P. E.; Rooney, M. B.; Deeth, J. R.; Lai, V.; Ng, G. J.; Situ, H.; Zhang, R. H.; Elliott, M. S.; Haley, T. H.; Dvorak, D. J.; Aspuru-Guzik, A.; Hein, J. E.; Berlinguette, C. P. Self-Driving Laboratory for Accelerated Discovery of Thin-Film Materials. *Sci. Adv.* **2020**, *6* (20), eaaz8867. https://doi.org/10.1126/sciadv.aaz8867.
(29) Raccuglia, P.; Elbert, K. C.; Adler, P. D. F.; Falk, C.; Wenny, M. B.; Mollo, A.; Zeller, M.; Friedler, S. A.; Schrier, J.; Norquist, A. J. Machine-Learning-Assisted Materials Discovery Using Failed Experiments. *Nature* **2016**, *533* (7601), 73–76. https://doi.org/10.1038/nature17439.
(30) Smith, A.; Keane, A.; Dumesic, J. A.; Huber, G. W.; Zavala, V. M. A Machine Learning Framework for the Analysis and Prediction of Catalytic Activity from Experimental

Data. *Appl. Catal. B Environ.* **2020**, *263*, 118257. https://doi.org/10.1016/j.apcatb.2019.118257.
(31) Sans, V.; Porwol, L.; Dragone, V.; Cronin, L. A Self Optimizing Synthetic Organic Reactor System Using Real-Time in-Line NMR Spectroscopy. *Chem. Sci.* **2015**, *6* (2), 1258–1264. https://doi.org/10.1039/C4SC03075C.
(32) Horváth, G.; Trujillo, V. J.; Réti, J.; Kozár, Z.; Varga, T.; Kummer, A. Soft-Sensor Development for Product Quality Estimation with Time Delay and Feature Selection in Industrial MDI Production. *Chem. Eng. J. Adv.* **2025**, *22*, 100751. https://doi.org/10.1016/j.ceja.2025.100751.
(33) Żurański, A. M.; Martinez Alvarado, J. I.; Shields, B. J.; Doyle, A. G. Predicting Reaction Yields via Supervised Learning. *Acc. Chem. Res.* **2021**, *54* (8), 1856–1865. https://doi.org/10.1021/acs.accounts.0c00770.
(34) Glielmo, A.; Husic, B. E.; Rodriguez, A.; Clementi, C.; Noé, F.; Laio, A. Unsupervised Learning Methods for Molecular Simulation Data. *Chem. Rev.* **2021**, *121* (16), 9722–9758. https://doi.org/10.1021/acs.chemrev.0c01195.
(35) Jia, W.; Sun, M.; Lian, J.; Hou, S. Feature Dimensionality Reduction: A Review. *Complex Intell. Syst.* **2022**, *8* (3), 2663–2693. https://doi.org/10.1007/s40747-021-00637-x.
(36) Kurgan, L. A.; Musilek, P. A Survey of Knowledge Discovery and Data Mining Process Models. *Knowl. Eng. Rev.* **2006**, *21* (1), 1–24. https://doi.org/10.1017/S0269888906000737.
(37) Studer, S.; Bui, T. B.; Drescher, C.; Hanuschkin, A.; Winkler, L.; Peters, S.; Müller, K.-R. Towards CRISP-ML(Q): A Machine Learning Process Model with Quality Assurance Methodology. *Mach. Learn. Knowl. Extr.* **2021**, *3* (2), 392–413. https://doi.org/10.3390/make3020020.
(38) ÇeliK, Ö. A Research on Machine Learning Methods and Its Applications. *J. Educ. Technol. Online Learn.* **2018**, *1* (3), 25–40. https://doi.org/10.31681/jetol.457046.
(39) Department of Computer Science and Informatics, University of Energy and Natural Resources, Sunyani, Ghana; Nti, I. K.; Nyarko-Boateng, O.; Aning, J. Performance of Machine Learning Algorithms with Different K Values in K-Fold CrossValidation. *Int. J. Inf. Technol. Comput. Sci.* **2021**, *13* (6), 61–71. https://doi.org/10.5815/ijitcs.2021.06.05.
(40) Dewancker, I.; McCourt, M.; Clark, S. Bayesian Optimization for Machine Learning : A Practical Guidebook. arXiv 2016. https://doi.org/10.48550/ARXIV.1612.04858.
(41) Shams, M. Y.; Elshewey, A. M.; El-kenawy, E.-S. M.; Ibrahim, A.; Talaat, F. M.; Tarek, Z. Water Quality Prediction Using Machine Learning Models Based on Grid Search Method. *Multimed. Tools Appl.* **2023**, *83* (12), 35307–35334. https://doi.org/10.1007/s11042-023-16737-4.
(42) Moiz, A. A.; Pal, P.; Probst, D.; Pei, Y.; Zhang, Y.; Som, S.; Kodavasal, J. A Machine Learning-Genetic Algorithm (ML-GA) Approach for Rapid Optimization Using High-Performance Computing. *SAE Int. J. Commer. Veh.* **2018**, *11* (5), 291–306. https://doi.org/10.4271/2018-01-0190.
(43) Rosso, M. M.; Cucuzza, R.; Di Trapani, F.; Marano, G. C. Nonpenalty Machine Learning Constraint Handling Using PSO-SVM for Structural Optimization. *Adv. Civ. Eng.* **2021**, *2021* (1), 6617750. https://doi.org/10.1155/2021/6617750.
(44) Paleyes, A.; Urma, R.-G.; Lawrence, N. D. Challenges in Deploying Machine Learning: A Survey of Case Studies. *ACM Comput. Surv.* **2023**, *55* (6), 1–29. https://doi.org/10.1145/3533378.
(45) Shi, Z.; Yao, W.; Li, Z.; Zeng, L.; Zhao, Y.; Zhang, R.; Tang, Y.; Wen, J. Artificial Intelligence Techniques for Stability Analysis and Control in Smart Grids:

Methodologies, Applications, Challenges and Future Directions. *Appl. Energy* **2020**, *278*, 115733. https://doi.org/10.1016/j.apenergy.2020.115733.
(46) Su, X.; Yan, X.; Tsai, C. Linear Regression. *WIREs Comput. Stat.* **2012**, *4* (3), 275–294. https://doi.org/10.1002/wics.1198.
(47) Zhang, Z. Variable Selection with Stepwise and Best Subset Approaches. *Ann. Transl. Med.* **2016**, *4* (7), 136–136. https://doi.org/10.21037/atm.2016.03.35.
(48) Yamashita, T.; Yamashita, K.; Kamimura, R. A Stepwise AIC Method for Variable Selection in Linear Regression. *Commun. Stat. - Theory Methods* **2007**, *36* (13), 2395–2403. https://doi.org/10.1080/03610920701215639.
(49) Cheng, M.; Feng, S.; Li, G.; Lian, H. Greedy Forward Regression for Variable Screening. *Aust. N. Z. J. Stat.* **2018**, *60* (1), 20–42. https://doi.org/10.1111/anzs.12218.
(50) Breiman, L. Random Forests. *Mach. Learn.* **2001**, *45* (1), 5–32. https://doi.org/10.1023/A:1010933404324.
(51) Aziz, N.; Akhir, E. A. P.; Aziz, I. A.; Jaafar, J.; Hasan, M. H.; Abas, A. N. C. A Study on Gradient Boosting Algorithms for Development of AI Monitoring and Prediction Systems. In *2020 International Conference on Computational Intelligence (ICCI)*; IEEE: Bandar Seri Iskandar, Malaysia, 2020; pp 11–16. https://doi.org/10.1109/ICCI51257.2020.9247843.
(52) Zeng, X.; Wu, G. Data-Driven Soft Sensor Modeling for DCT Clutch Pressure Based on Decision Tree and Stepwise Regression. In *2020 7th International Conference on Information Science and Control Engineering (ICISCE)*; IEEE: Changsha, China, 2020; pp 436–442. https://doi.org/10.1109/ICISCE50968.2020.00097.
(53) Ching, P. M. L.; So, R. H. Y.; Morck, T. Advances in Soft Sensors for Wastewater Treatment Plants: A Systematic Review. *J. Water Process Eng.* **2021**, *44*, 102367. https://doi.org/10.1016/j.jwpe.2021.102367.
(54) Mulrennan, K.; Donovan, J.; Creedon, L.; Rogers, I.; Lyons, J. G.; McAfee, M. A Soft Sensor for Prediction of Mechanical Properties of Extruded PLA Sheet Using an Instrumented Slit Die and Machine Learning Algorithms. *Polym. Test.* **2018**, *69*, 462–469. https://doi.org/10.1016/j.polymertesting.2018.06.002.
(55) Specht, D. F. A General Regression Neural Network. *IEEE Trans. Neural Netw.* **1991**, *2* (6), 568–576. https://doi.org/10.1109/72.97934.
(56) Smola, A. J.; Schölkopf, B. A Tutorial on Support Vector Regression. *Stat. Comput.* **2004**, *14* (3), 199–222. https://doi.org/10.1023/B:STCO.0000035301.49549.88.
(57) Deiss, L.; Margenot, A. J.; Culman, S. W.; Demyan, M. S. Tuning Support Vector Machines Regression Models Improves Prediction Accuracy of Soil Properties in MIR Spectroscopy. *Geoderma* **2020**, *365*, 114227. https://doi.org/10.1016/j.geoderma.2020.114227.
(58) Zaghloul, M. S.; Hamza, R. A.; Iorhemen, O. T.; Tay, J. H. Comparison of Adaptive Neuro-Fuzzy Inference Systems (ANFIS) and Support Vector Regression (SVR) for Data-Driven Modelling of Aerobic Granular Sludge Reactors. *J. Environ. Chem. Eng.* **2020**, *8* (3), 103742. https://doi.org/10.1016/j.jece.2020.103742.
(59) Elmaz, F.; Yücel, Ö.; Mutlu, A. Y. Predictive Modeling of Biomass Gasification with Machine Learning-Based Regression Methods. *Energy* **2020**, *191*, 116541. https://doi.org/10.1016/j.energy.2019.116541.
(60) Tibshirani, R. Regression Shrinkage and Selection Via the Lasso. *J. R. Stat. Soc. Ser. B Methodol.* **1996**, *58* (1), 267–288. https://doi.org/10.1111/j.2517-6161.1996.tb02080.x.
(61) Darbellay, G. A.; Vajda, I. Estimation of the Information by an Adaptive Partitioning of the Observation Space. *IEEE Trans. Inf. Theory* **1999**, *45* (4), 1315–1321. https://doi.org/10.1109/18.761290.

(62) Hanchuan Peng; Fuhui Long; Ding, C. Feature Selection Based on Mutual Information Criteria of Max-Dependency, Max-Relevance, and Min-Redundancy. *IEEE Trans. Pattern Anal. Mach. Intell.* **2005**, *27* (8), 1226–1238. https://doi.org/10.1109/TPAMI.2005.159.
(63) Ding, C.; Peng, H. Minimum Redundancy Feature Selection From Microarray Gene Expression Data. *J. Bioinform. Comput. Biol.* **2005**, *03* (02), 185–205. https://doi.org/10.1142/S0219720005001004.
(64) Lee, C.; Lee, G. G. MMR-Based Feature Selection for Text Categorization. In *Proceedings of HLT-NAACL 2004: Short Papers on XX - HLT-NAACL '04*; Association for Computational Linguistics: Boston, Massachusetts, 2004; pp 5–8. https://doi.org/10.3115/1613984.1613986.
(65) Estevez, P. A.; Tesmer, M.; Perez, C. A.; Zurada, J. M. Normalized Mutual Information Feature Selection. *IEEE Trans. Neural Netw.* **2009**, *20* (2), 189–201. https://doi.org/10.1109/TNN.2008.2005601.
(66) Rasmussen, C. E., R. M. Neal, G. E. Hinton, D. van Camp, M. Revow, Z. Ghahramani, R. Kustra, and R. Tibshirani. *The DELVE Manual*; The University of Toronto: Toronto, Ontario, Canada, 1996.
(67) Murtaugh, P. A. Methods of Variable Selection in Regression Modeling. *Commun. Stat. - Simul. Comput.* **1998**, *27* (3), 711–734. https://doi.org/10.1080/03610919808813505.
(68) Omer Fadl Elssied, N.; Ibrahim, O.; Hamza Osman, A. A Novel Feature Selection Based on One-Way ANOVA F-Test for E-Mail Spam Classification. *Res. J. Appl. Sci. Eng. Technol.* **2014**, *7* (3), 625–638. https://doi.org/10.19026/rjaset.7.299.
(69) Dhanya, R.; Paul, I. R.; Akula, S. S.; Sivakumar, M.; Nair, J. J. F-Test Feature Selection in Stacking Ensemble Model for Breast Cancer Prediction. *Procedia Comput. Sci.* **2020**, *171*, 1561–1570. https://doi.org/10.1016/j.procs.2020.04.167.
(70) Pathan, M. S.; Nag, A.; Pathan, M. M.; Dev, S. Analyzing the Impact of Feature Selection on the Accuracy of Heart Disease Prediction. *Healthc. Anal.* **2022**, *2*, 100060. https://doi.org/10.1016/j.health.2022.100060.
(71) Aydin, G. Regression Models for Forecasting Global Oil Production. *Pet. Sci. Technol.* **2015**, *33* (21–22), 1822–1828. https://doi.org/10.1080/10916466.2015.1101474.
(72) Kumar, B. J.; Naveen, H.; Kumar, B. P.; Sharma, S. S.; Villegas, J. Logistic Regression for Polymorphic Malware Detection Using ANOVA F-Test. In *2017 International Conference on Innovations in Information, Embedded and Communication Systems (ICIIECS)*; IEEE: Coimbatore, 2017; pp 1–5. https://doi.org/10.1109/ICIIECS.2017.8275880.
(73) Urbanowicz, R. J.; Meeker, M.; La Cava, W.; Olson, R. S.; Moore, J. H. Relief-Based Feature Selection: Introduction and Review. *J. Biomed. Inform.* **2018**, *85*, 189–203. https://doi.org/10.1016/j.jbi.2018.07.014.
(74) Robnik-Šikonja, M.; Kononenko, I. Theoretical and Empirical Analysis of ReliefF and RReliefF. *Mach. Learn.* **2003**, *53* (1–2), 23–69. https://doi.org/10.1023/A:1025667309714.
(75) Islam, T.; Srivastava, P. K.; Dai, Q.; Gupta, M.; Zhuo, L. Rain Rate Retrieval Algorithm for Conical-Scanning Microwave Imagers Aided by Random Forest, RReliefF, and Multivariate Adaptive Regression Splines (RAMARS). *IEEE Sens. J.* **2015**, *15* (4), 2186–2193. https://doi.org/10.1109/JSEN.2014.2372814.
(76) Acikgoz, H. A Novel Approach Based on Integration of Convolutional Neural Networks and Deep Feature Selection for Short-Term Solar Radiation Forecasting. *Appl. Energy* **2022**, *305*, 117912. https://doi.org/10.1016/j.apenergy.2021.117912.
(77) Son, H.; Kim, C.; Kim, C.; Kang, Y. PREDICTION OF GOVERNMENT-OWNED BUILDING ENERGY CONSUMPTION BASED ON AN RRELIEFF AND SUPPORT

VECTOR MACHINE MODEL. *J. Civ. Eng. Manag.* **2015**, *21* (6), 748–760. https://doi.org/10.3846/13923730.2014.893908.

(78) Sun, Y.; Gong, H.; Li, Y.; Zhang, D. Hyperparameter Importance Analysis Based on N-RReliefF Algorithm. *Int. J. Comput. Commun. Control* **2019**, *14* (4), 557–573. https://doi.org/10.15837/ijccc.2019.4.3593.

(79) Conn, A. R.; Gould, N. I. M.; Toint, P. A Globally Convergent Augmented Lagrangian Algorithm for Optimization with General Constraints and Simple Bounds. *SIAM J. Numer. Anal.* **1991**, *28* (2), 545–572. https://doi.org/10.1137/0728030.

(80) Goldstein, A.; Kapelner, A.; Bleich, J.; Pitkin, E. Peeking Inside the Black Box: Visualizing Statistical Learning With Plots of Individual Conditional Expectation. *J. Comput. Graph. Stat.* **2015**, *24* (1), 44–65. https://doi.org/10.1080/10618600.2014.907095.

(81) Merrick, L.; Taly, A. The Explanation Game: Explaining Machine Learning Models Using Shapley Values. In *Machine Learning and Knowledge Extraction*; Holzinger, A., Kieseberg, P., Tjoa, A. M., Weippl, E., Eds.; Lecture Notes in Computer Science; Springer International Publishing: Cham, 2020; Vol. 12279, pp 17–38. https://doi.org/10.1007/978-3-030-57321-8_2.

(82) Rozemberczki, B.; Watson, L.; Bayer, P.; Yang, H.-T.; Kiss, O.; Nilsson, S.; Sarkar, R. The Shapley Value in Machine Learning. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence*; International Joint Conferences on Artificial Intelligence Organization: Vienna, Austria, 2022; pp 5572–5579. https://doi.org/10.24963/ijcai.2022/778.

(83) *The Polyurethanes Book*; Randall, D., Lee, S., Eds.; Huntsman Polyurethanes] ; New York : Distributed by John Wiley & Sons: [Everberg, Belgium, 2002.

(84) Prof. J.P. Thomé; C. Joaquim-Justo; Prof. Dr. B. Nemery; H. Vanhooren. *Methylenediphenyl Diisocyanate (MDI)*; Federal Public Service Health, Food-Chain Safety and Environment, 2005.

(85) Christensen, F.; H. Nilsson, N.; Jeppesen, C. N.; Clausen, A. J. *Survey of Certain Isocyanates (MDI and TDI), Part of the LOUS-Review*; The Danish Environmental Protection Agency, 2014.

(86) Quadros, P. A.; Oliveira, N. M. C.; Baptista, C. M. S. G. Continuous Adiabatic Industrial Benzene Nitration with Mixed Acid at a Pilot Plant Scale. *Chem. Eng. J.* **2005**, *108* (1–2), 1–11. https://doi.org/10.1016/j.cej.2004.12.022.

(87) Mishra, P.; Ein-Mozaffari, F. Using Computational Fluid Dynamics to Analyze the Performance of the Maxblend Impeller in Solid-Liquid Mixing Operations. *Int. J. Multiph. Flow* **2017**, *91*, 194–207. https://doi.org/10.1016/j.ijmultiphaseflow.2017.01.009.

(88) Benneker, A. M.; van der Ham, L. G. J.; de Waele, B.; Zeeuw, A. J.; van den Berg, H. Design and Intensification of Industrial DADPM Process. *Chem. Eng. Process. - Process Intensif.* **2016**, *109*, 39–50. https://doi.org/10.1016/j.cep.2016.08.009.

(89) Boros, R.; Farkas, L.; Nehéz, K.; Viskolcz, B.; Szőri, M. An Ab Initio Investigation of the 4,4′-Methlylene Diphenyl Diamine (4,4′-MDA) Formation from the Reaction of Aniline with Formaldehyde. *Polymers* **2019**, *11* (3), 398. https://doi.org/10.3390/polym11030398.

(90) de Angelis, A.; Bosetti, A.; Millini, R.; Perego, C. Catalytic Processes for Environmentally Friendly Methylene Diphenyl Diisocyanate Production. In *Chemistry Beyond Chlorine*; Tundo, P., He, L.-N., Lokteva, E., Mota, C., Eds.; Springer International Publishing: Cham, 2016; pp 171–200. https://doi.org/10.1007/978-3-319-30073-3_5.

(91) Callison, J. The Investigation of a Side Reaction Leading to Colour Formation in a Polyurethane Production Chain, 2011. http://theses.gla.ac.uk/2498/.
(92) Shapatina, E.; Kuchaev, V.; Penkovoi, B.; Temkin, M. Kinetics of Catalytic Synthesis of Phosgene. *Kinet. Catal.* **1976**, *17* (3), 559–566.
(93) Shapatina, E.; Kuchaev, V.; Temkin, M. Kinetics of Phosgene Synthesis at Low Chorine Concentrations. *Kinet. Catal.* **1979**, *20* (5), 1183–1187.
(94) Varga, T. Process Models and Data Mining Techniques in Determination and Characterization of Safe Operating Regimes, 2009. https://konyvtar.uni-pannon.hu/doktori/2010/Varga_Tamas_dissertation.pdf.
(95) Boros, R. Z.; Koós, T.; Wafaa, C.; Nehéz, K.; Farkas, L.; Viskolcz, B.; Szőri, M. A Theoretical Study on the Phosgenation of Methylene Diphenyl Diamine (MDA). *Chem. Phys. Lett.* **2018**, *706*, 568–576. https://doi.org/10.1016/j.cplett.2018.06.024.
(96) Gibson, E. K.; Winfield, J. M.; Muir, K. W.; Carr, R. H.; Eaglesham, A.; Gavezzotti, A.; Parker, S. F.; Lennon, D. A Structural and Spectroscopic Investigation of the Hydrochlorination of 4-Benzylaniline: The Interaction of Anhydrous Hydrogen Chloride with Chlorobenzene. *Phys Chem Chem Phys* **2009**, *11* (2), 288–297. https://doi.org/10.1039/B810027F.
(97) Gibson, E. K.; Winfield, J. M.; Muir, K. W.; Carr, R. H.; Eaglesham, A.; Gavezzotti, A.; Lennon, D. A Structural and Spectroscopic Investigation of the Hydrochlorination of 4,4′-Methylenedianiline. *Phys. Chem. Chem. Phys.* **2010**, *12* (15), 3824. https://doi.org/10.1039/b916368a.
(98) Six, C.; Richter, F. Isocyanates, Organic. In *Ullmann's Encyclopedia of Industrial Chemistry*; Wiley-VCH, Ed.; Wiley, 2003. https://doi.org/10.1002/14356007.a14_611.
(99) Tian, J.; An, H.; Cheng, X.; Zhao, X.; Wang, Y. Synthesis of 4,4′-Methylenedianiline Catalyzed by $SO_3$ H-Functionalized Ionic Liquids. *Ind. Eng. Chem. Res.* **2015**, *54* (31), 7571–7579. https://doi.org/10.1021/acs.iecr.5b01519.
(100) Merenov, A. S.; Jewell, D. W.; Gillis, P. A.; Jansma, G. I.; Breed, A.; Anderson, J. J.; Reed, D. J. Process for the Production of Methylene Diphenyl Diisocyanate Isomer Mixtures with Specific Isomer Distributions and New Products Derived Therefrom. EP2619176A1, July 31, 2013. https://patents.google.com/patent/EP2619176A1/ru#patentCitations (accessed 2022-07-09).
(101) Keller, T. C.; Arras, J.; Haus, M. O.; Hauert, R.; Kenvin, A.; Kenvin, J.; Pérez-Ramírez, J. Synthesis-Property-Performance Relationships of Amorphous Silica-Alumina Catalysts for the Production of Methylenedianiline and Higher Homologues. *J. Catal.* **2016**, *344*, 757–767. https://doi.org/10.1016/j.jcat.2016.08.016.
(102) Wegener, G.; Brandt, M.; Duda, L.; Hofmann, J.; Klesczewski, B.; Koch, D.; Kumpf, R.-J.; Orzesek, H.; Pirkl, H.-G.; Six, C.; Steinlein, C.; Weisbeck, M. Trends in Industrial Catalysis in the Polyurethane Industry. *Appl. Catal. Gen.* **2001**, *221* (1–2), 303–335. https://doi.org/10.1016/S0926-860X(01)00910-3.
(103) Cheung, K. Y.; Marquez, C.; Tomkins, P.; Parvulescu, A.-N.; Gordillo, A.; De Baerdemaeker, T.; De Vos, D. Lewis Acid Solid Catalysts for the Synthesis of Methylenedianiline from Aniline and Formaldehyde. *J. Catal.* **2021**, *400*, 114–123. https://doi.org/10.1016/j.jcat.2021.05.031.
(104) Twitchett, H. J. Chemistry of the Production of Organic Isocyanates. *Chem. Soc. Rev.* **1974**, *3* (2), 209–230. https://doi.org/10.1039/CS9740300209.
(105) Corma, A.; Botella, P.; Mitchell, C. Replacing HCl by Solid Acids in Industrial Processes: Synthesis of Diamino Diphenyl Methane (DADPM) for Producing Polyurethanes. *Chem. Commun.* **2004**, No. 17, 2008–2010.

(106) Ogata, Y.; Okano, M.; Sugawara, M. Kinetics of the Condensation of Anilines with Formaldehyde. *J. Am. Chem. Soc.* **1951**, *73* (4), 1715–1717.
(107) Ogata, Y.; Okano, M.; Yamamoto, T. The Salt Effect in the Aniline-Formaldehyde Condensation. *Bull. Chem. Soc. Jpn.* **1952**, *25* (2), 86–88.
(108) Nayar, M. R. G.; Francis, J. D. Kinetics and Mechanism of the Aniline-Formaldehyde Reaction in Acid Medium. *Makromol. Chem. Macromol. Chem. Phys.* **1978**, *179* (7), 1783–1790.
(109) Kugita, T.; Hirose, S.; Namba, S. Catalytic Activity of Zeolites for Synthesis Reaction of Methylenedianiline from Aniline and Formaldehyde. *Catal. Today* **2006**, *111* (3–4), 275–279. https://doi.org/10.1016/j.cattod.2005.10.036.
(110) Botella, P.; Corma, A.; Carr, R. H.; Mitchell, C. J. Towards an Industrial Synthesis of Diamino Diphenyl Methane (DADPM) Using Novel Delaminated Materials: A Breakthrough Step in the Production of Isocyanates for Polyurethanes. *Appl. Catal. Gen.* **2011**, *398* (1–2), 143–149. https://doi.org/10.1016/j.apcata.2011.03.026.
(111) Wang, C. Y.; Li, H. Q.; Wang, L. G.; Cao, Y.; Liu, H. T.; Zhang, Y. Insights on the Mechanism for Synthesis of Methylenedianiline from Aniline and Formaldehyde through HPLC–MS and Isotope Tracer Studies. *Chin. Chem. Lett.* **2012**, *23* (11), 1254–1258. https://doi.org/10.1016/j.cclet.2012.10.001.
(112) Salzinger, M.; Lercher, J. A. Reaction Network and Mechanism of the Synthesis of Methylenedianiline over Dealuminated Y-Type Zeolites. *Green Chem.* **2011**, *13* (1), 149–155.
(113) Salzinger, M.; Fichtl, M. B.; Lercher, J. A. On the Influence of Pore Geometry and Acidity on the Activity of Parent and Modified Zeolites in the Synthesis of Methylenedianiline. *Appl. Catal. Gen.* **2011**, *393* (1–2), 189–194.
(114) Haus, M. O.; Keller, T. C.; Arras, J.; Pérez-Ramírez, J. Advanced Kinetic Models through Mechanistic Understanding: Population Balances for Methylenedianiline Synthesis. *Chem. Eng. Sci.* **2017**, *167*, 317–326. https://doi.org/10.1016/j.ces.2017.02.008.
(115) Till, Z.; Varga, T.; Sója, J.; Miskolczi, N.; Chován, T. Kinetic Identification of Plastic Waste Pyrolysis on Zeolite-Based Catalysts. *Energy Convers. Manag.* **2018**, *173*, 320–330. https://doi.org/10.1016/j.enconman.2018.07.088.
(116) Shampine, L. F.; Reichelt, M. W. The MATLAB ODE Suite. *SIAM J. Sci. Comput.* **1997**, *18* (1), 1–22. https://doi.org/10.1137/S1064827594276424.
(117) Bogacki, P.; Shampine, L. F. A 3(2) Pair of Runge - Kutta Formulas. *Appl. Math. Lett.* **1989**, *2* (4), 321–325. https://doi.org/10.1016/0893-9659(89)90079-7.
(118) Byrd, R. H.; Gilbert, J. C.; Nocedal, J. A Trust Region Method Based on Interior Point Techniques for Nonlinear Programming. *Math. Program.* **2000**, *89* (1), 149–185. https://doi.org/10.1007/PL00011391.
(119) Byrd, R. H.; Hribar, M. E.; Nocedal, J. An Interior Point Algorithm for Large-Scale Nonlinear Programming. *SIAM J. Optim.* **1999**, *9* (4), 877–900. https://doi.org/10.1137/S1052623497325107.
(120) Coleman, T. F.; Li, Y. On the Convergence of Interior-Reflective Newton Methods for Nonlinear Minimization Subject to Bounds. *Math. Program.* **1994**, *67* (1–3), 189–224. https://doi.org/10.1007/BF01582221.
(121) Coleman, T. F.; Li, Y. An Interior Trust Region Approach for Nonlinear Minimization Subject to Bounds. *SIAM J. Optim.* **1996**, *6* (2), 418–445. https://doi.org/10.1137/0806023.
(122) Wan, X.; Wang, W.; Liu, J.; Tong, T. Estimating the Sample Mean and Standard Deviation from the Sample Size, Median, Range and/or Interquartile Range. *BMC Med. Res. Methodol.* **2014**, *14* (1), 135. https://doi.org/10.1186/1471-2288-14-135.

(123) Da Silva, J. E. E.; Alarcon, R. T.; Gaglieri, C.; Magdalena, A. G.; Da Silva-Filho, L. C.; Bannach, G. New Thermal Study of Polymerization and Degradation Kinetics of Methylene Diphenyl Diisocyanate. *J. Therm. Anal. Calorim.* **2018**, *133* (3), 1455–1462. https://doi.org/10.1007/s10973-018-7211-1.
(124) Allen, N. S.; McKellar, J. F. Photochemical Reactions in an MDI-Based Elastomeric Polyurethane. *J. Appl. Polym. Sci.* **1976**, *20* (6), 1441–1447. https://doi.org/10.1002/app.1976.070200602.
(125) Beachell, H. C.; Son, C. P. N. Color Formation in Polyurethanes. *J. Appl. Polym. Sci.* **1963**, *7* (6), 2217–2237. https://doi.org/10.1002/app.1963.070070621.
(126) Beachell, H. C.; Son, C. P. N. Stabilization of Polyurethane to Thermal Degradation. *J. Appl. Polym. Sci.* **1964**, *8* (3), 1089–1096. https://doi.org/10.1002/app.1964.070080306.
(127) Beachell, H. C.; Chang, I. L. Photodegradation of Urethane Model Systems. *J. Polym. Sci. [A1]* **1972**, *10* (2), 503–520. https://doi.org/10.1002/pol.1972.150100216.
(128) Callison, J.; Edge, R.; De Cuba, K. R.; Carr, R. H.; McDouall, J. J. W.; Collison, D.; McInnes, E. J. L.; Van Der Borden, W.; Van Der Velde, K.; Winfield, J. M.; Lennon, D. Origin of Impurities Formed in the Polyurethane Production Chain. 1. Conditions for Chlorine Transfer from an Aryl Isocyanide Dichloride Byproduct. *Ind. Eng. Chem. Res.* **2012**, *51* (6), 2515–2523. https://doi.org/10.1021/ie2013136.
(129) Callison, J.; Betzler, F.; De Cuba, K.; Van Der Borden, W.; Van Der Velde, K.; Carr, R. H.; Senn, H. M.; Farrugia, L. J.; Winfield, J. M.; Lennon, D. Origin of Impurities Formed in a Polyurethane Production Chain. Part 2: A Route to the Formation of Colored Impurities. *Ind. Eng. Chem. Res.* **2012**, *51* (34), 11021–11030. https://doi.org/10.1021/ie300987v.
(130) Yuan, X.; Xu, W.; Wang, Y.; Yang, C.; Gui, W. A Deep Residual PLS for Data-Driven Quality Prediction Modeling in Industrial Process. *IEEECAA J. Autom. Sin.* **2024**, *11* (8), 1777–1785. https://doi.org/10.1109/JAS.2024.124578.
(131) Yuan, X.; Xu, N.; Ye, L.; Wang, K.; Shen, F.; Wang, Y.; Yang, C.; Gui, W. Attention-Based Interval Aided Networks for Data Modeling of Heterogeneous Sampling Sequences With Missing Values in Process Industry. *IEEE Trans. Ind. Inform.* **2024**, *20* (4), 5253–5262. https://doi.org/10.1109/TII.2023.3329684.
(132) Jiang, Y.; Yin, S.; Dong, J.; Kaynak, O. A Review on Soft Sensors for Monitoring, Control, and Optimization of Industrial Processes. *IEEE Sens. J.* **2021**, *21* (11), 12868–12881. https://doi.org/10.1109/JSEN.2020.3033153.
(133) Yuan, X.; Gu, Y.; Wang, Y.; Yang, C.; Gui, W. A Deep Supervised Learning Framework for Data-Driven Soft Sensor Modeling of Industrial Processes. *IEEE Trans. Neural Netw. Learn. Syst.* **2020**, *31* (11), 4737–4746. https://doi.org/10.1109/TNNLS.2019.2957366.
(134) Yuan, X.; Qi, S.; Shardt, Y. A. W.; Wang, Y.; Yang, C.; Gui, W. Soft Sensor Model for Dynamic Processes Based on Multichannel Convolutional Neural Network. *Chemom. Intell. Lab. Syst.* **2020**, *203*, 104050. https://doi.org/10.1016/j.chemolab.2020.104050.
(135) Souza, F.; Santos, P.; Araújo, R. Variable and Delay Selection Using Neural Networks and Mutual Information for Data-Driven Soft Sensors. In *2010 IEEE 15th Conference on Emerging Technologies & Factory Automation (ETFA 2010)*; IEEE: Bilbao, 2010; pp 1–8. https://doi.org/10.1109/ETFA.2010.5641329.
(136) Griesing-Scheiwe, F.; Shardt, Y. A. W.; Pérez-Zuñiga, G.; Yang, X. Soft Sensor Design for Variable Time Delay and Variable Sampling Time. *J. Process Control* **2020**, *92*, 310–318. https://doi.org/10.1016/j.jprocont.2020.07.001.
(137) Hikosaka, T.; Aoshima, S.; Miyao, T.; Funatsu, K. Soft Sensor Modeling for Identifying Significant Process Variables with Time Delays. *Ind. Eng. Chem. Res.* **2020**, *59* (26), 12156–12163. https://doi.org/10.1021/acs.iecr.0c01655.

(138) Wang, W.; Yang, C.; Han, J.; Li, W.; Li, Y. A Soft Sensor Modeling Method with Dynamic Time-Delay Estimation and Its Application in Wastewater Treatment Plant. *Biochem. Eng. J.* **2021**, *172*, 108048. https://doi.org/10.1016/j.bej.2021.108048.
(139) Xiong, W.; Li, Y.; Zhao, Y.; Huang, B. Adaptive Soft Sensor Based on Time Difference Gaussian Process Regression with Local Time-Delay Reconstruction. *Chem. Eng. Res. Des.* **2017**, *117*, 670–680. https://doi.org/10.1016/j.cherd.2016.11.020.
(140) Wang, K.; Shang, C.; Liu, L.; Jiang, Y.; Huang, D.; Yang, F. Dynamic Soft Sensor Development Based on Convolutional Neural Networks. *Ind. Eng. Chem. Res.* **2019**, *58* (26), 11521–11531. https://doi.org/10.1021/acs.iecr.9b02513.
(141) Robnik-Šikonja, M.; Kononenko, I. Theoretical and Empirical Analysis of ReliefF and RReliefF. *Mach. Learn.* **2003**, *53* (1/2), 23–69. https://doi.org/10.1023/A:1025667309714.
(142) Friedman, J. H. Greedy Function Approximation: A Gradient Boosting Machine. *Ann. Stat.* **2001**, *29* (5). https://doi.org/10.1214/aos/1013203451.

# Appendix

Table A1 data was originally published by Haus *et al.* [114] The primary aim of Haus *et al.* work to develop a regular kinetic model, so-called empirical model and a population model to describe MDA formation as accurately as possible. The results of the parameter identification what was performed for the empirical model is summarized in Table A1.

**Table A1: Fitted parameters of the empirical model by Haus *et al.* [114]**

| Reaction | $\ln(k_0)$ [-] | Ea [kJ/mol] |
|---|---|---|
| **R1** | 16.3 ± 2.8 | 62.3 ± 8.3 |
| **R2** | 16.7 ± 2.8 | 71.0 ± 8.2 |
| **R3** | 24.1 ± 0.4 | 97.0 ± 1.1 |
| **R4** | 27.7 ± 0.4 | 113.5 ± 1.3 |
| **R5** | 27.0 ± 1.2 | 103.2 ± 4.0 |
| **R6** | 25.8 ± 2.3 | 117.9 ± 7.5 |
| **R7** | 16.6 ± 2.9 | 66.8 ± 8.4 |
| **R8** | 22.2 ± 2.0 | 98.4 ± 6.6 |
| **R9** | 20.2 ± 7.0 | 88.6 ± 23.0 |
| **R10** | 22.2 ± 7.4 | 100.1 ± 24.0 |
| **R11** | 24.0 ± 0.7 | 96.8 ± 2.4 |

For the parameter identification tasks for each cases was performed with the use of upper and lower boundaries. For Case A, the originally published reaction uncertainties of kinetic parameters identified and summarized in Table A1 were used as the lower bounds (LB) and upper bounds (UB). For Case B we used the same bounds for each reaction to find the optimal solution for the proposed extended reaction network. In Case C, we kept the original boundaries of Case A for the unmodified reactions and kept Case B boundaries for all the modified reactions. The summary of boundaries can be seen in Table A2.

**Table A2: The initial values and lower-, upper bounds defined in the parameter identification tasks, where n is the reaction order**

| Reaction | k [$L^{n-1}$ $mol^{1-n}$ $s^{-1}$] | | | | | |
|---|---|---|---|---|---|---|
| | Case A | | Case B | | Case C | |
| | LB | UB | LB | UB | LB | UB |
| **R1** | $4.82 \cdot 10^{-5}$ | 2.10 | $10^{-10}$ | $10^{7}$ | $10^{-10}$ | $10^{7}$ |
| **R2** | $5.17 \cdot 10^{-6}$ | $2.12 \cdot 10^{-1}$ | | | $10^{-10}$ | $10^{7}$ |
| **R3** | $2.87 \cdot 10^{-4}$ | $1.25 \cdot 10^{-3}$ | | | $2.87 \cdot 10^{-4}$ | $1.25 \cdot 10^{-3}$ |
| **R4** | $6.33 \cdot 10^{-5}$ | $3.12 \cdot 10^{-4}$ | | | $6.33 \cdot 10^{-5}$ | $3.12 \cdot 10^{-4}$ |
| **R5** | $1.44 \cdot 10^{-4}$ | $1.84 \cdot 10^{-2}$ | | | $1.44 \cdot 10^{-4}$ | $1.84 \cdot 10^{-2}$ |
| **R6** | $5.52 \cdot 10^{-8}$ | $5.42 \cdot 10^{-4}$ | | | $5.52 \cdot 10^{-8}$ | $5.42 \cdot 10^{-4}$ |
| **R7** | $1.44 \cdot 10^{-5}$ | $8.14 \cdot 10^{-1}$ | | | $10^{-10}$ | $10^{7}$ |
| **R8** | $1.05 \cdot 10^{-6}$ | $3.25 \cdot 10^{-3}$ | | | $10^{-10}$ | $10^{7}$ |
| **R9** | $1.27 \cdot 10^{-10}$ | $1.98 \cdot 10^{2}$ | | | $1.27 \cdot 10^{-10}$ | $1.98 \cdot 10^{2}$ |
| **R10** | $1.37 \cdot 10^{-11}$ | $8.79 \cdot 10^{1}$ | | | $1.37 \cdot 10^{-11}$ | $8.79 \cdot 10^{1}$ |
| **R11** | $1.37 \cdot 10^{-4}$ | $2.42 \cdot 10^{-3}$ | | | $10^{-10}$ | $10^{7}$ |
| **R12** | - | - | | | $10^{-10}$ | $10^{7}$ |
| **R13** | - | - | | | $10^{-10}$ | $10^{7}$ |

**Table A3: $R^2$ values calculated with stepwise linear regression models for 2-ring MDA content with 3-fold cross-validation and different $P_{Enter}$ and $P_{Remove}$ parameters**

| $P_{Enter}$ | $P_{Remove}$ | $\bar{R}^2_{training}$ | $\bar{R}^2_{validation}$ |
|---|---|---|---|
| **0.1** | **0.05** | 0.580 | -0.092 |
| **0.05** | **0.025** | 0.867 | 0.797 |
| **0.04** | **0.02** | 0.867 | 0.797 |
| **0.03** | **0.01** | 0.867 | 0.797 |
| **0.027** | **0.01** | 0.913 | 0.775 |
| **0.025** | **0.01** | 0.913 | 0.775 |
| **0.025** | **0.005** | 0.913 | 0.775 |
| **0.01** | **0.005** | 0.970 | 0.723 |

**Table A4: $R^2$ values calculated with stepwise linear regression models for 3-ring MDA content with 3-fold cross-validation and different $P_{Enter}$ and $P_{Remove}$ parameters**

| $P_{Enter}$ | $P_{Remove}$ | $\bar{R}^2_{training}$ | $\bar{R}^2_{validation}$ |
|---|---|---|---|
| **0.1** | **0.05** | 0.747 | 0.636 |
| **0.05** | **0.025** | 0.837 | 0.673 |
| **0.04** | **0.02** | 0.854 | 0.665 |
| **0.03** | **0.01** | 0.905 | 0.287 |
| **0.027** | **0.01** | 0.905 | 0.287 |
| **0.025** | **0.01** | 0.913 | 0.236 |
| **0.025** | **0.005** | 0.913 | 0.236 |
| **0.01** | **0.005** | 0.926 | 0.299 |

**Table A5: $R^2$ values calculated with stepwise linear regression models for 4-ring MDA content with 3-fold cross-validation and different $P_{Enter}$ and $P_{Remove}$ parameters**

| $P_{Enter}$ | $P_{Remove}$ | $\bar{R}^2_{training}$ | $\bar{R}^2_{validation}$ |
|---|---|---|---|
| **0.1** | **0.05** | 0.829 | 0.596 |
| **0.05** | **0.025** | 0.829 | 0.596 |
| **0.04** | **0.02** | 0.845 | 0.647 |
| **0.03** | **0.01** | 0.845 | 0.647 |
| **0.027** | **0.01** | 0.875 | 0.601 |
| **0.025** | **0.01** | 0.892 | 0.328 |
| **0.025** | **0.005** | 0.892 | 0.328 |
| **0.01** | **0.005** | 0.917 | -0.015 |

**Table A6: $R^2$ values calculated with stepwise linear regression models for 5-ring MDA content with 3-fold cross-validation and different $P_{Enter}$ and $P_{Remove}$ parameters**

| $P_{Enter}$ | $P_{Remove}$ | $\bar{R}^2_{training}$ | $\bar{R}^2_{validation}$ |
|---|---|---|---|
| **0.2** | **0.1** | 0.621 | 0.031 |
| **0.15** | **0.05** | 0.802 | 0.401 |
| **0.11** | **0.05** | 0.802 | 0.401 |
| **0.1** | **0.05** | 0.836 | 0.476 |
| **0.05** | **0.025** | 0.894 | 0.434 |
| **0.04** | **0.02** | 0.894 | 0.434 |
| **0.03** | **0.01** | 0.919 | 0.484 |
| **0.025** | **0.005** | 0.919 | 0.484 |
| **0.01** | **0.005** | 0.943 | 0.153 |

**Table A7: $R^2$ values calculated with stepwise linear regression models for >6-ring MDA content with 3-fold cross-validation and different $P_{Enter}$ and $P_{Remove}$ parameters**

| $P_{Enter}$ | $P_{Remove}$ | $\bar{R}^2_{training}$ | $\bar{R}^2_{validation}$ |
|---|---|---|---|
| **0.2** | **0.1** | 0.172 | -0.318 |
| **0.15** | **0.05** | 0.344 | -0.657 |
| **0.11** | **0.05** | 0.388 | -0.519 |
| **0.1** | **0.05** | 0.388 | -0.519 |
| **0.05** | **0.025** | 0.601 | -131.399 |
| **0.04** | **0.02** | 0.665 | -173.690 |
| **0.03** | **0.01** | 0.706 | -173.780 |
| **0.027** | **0.01** | 0.706 | -173.780 |
| **0.01** | **0.005** | 0.870 | -172.215 |

**Table A8: $R^2$ values calculated with stepwise linear regression models for O-O isomer MDA content with 3-fold cross-validation and different $P_{Enter}$ and $P_{Remove}$ parameters**

| $P_{Enter}$ | $P_{Remove}$ | $\bar{R}^2_{training}$ | $\bar{R}^2_{validation}$ |
|---|---|---|---|
| **0.1** | **0.05** | 0.924 | 0.837 |
| **0.05** | **0.025** | 0.924 | 0.837 |
| **0.04** | **0.02** | 0.950 | 0.887 |
| **0.03** | **0.01** | 0.950 | 0.887 |
| **0.027** | **0.01** | 0.950 | 0.887 |
| **0.025** | **0.01** | 0.950 | 0.887 |
| **0.025** | **0.005** | 0.950 | 0.887 |
| **0.01** | **0.005** | 0.978 | 0.817 |

**Table A9: $R^2$ values calculated with stepwise linear regression models for O-P isomer MDA content with 3-fold cross-validation and different $P_{Enter}$ and $P_{Remove}$ parameters**

| $P_{Enter}$ | $P_{Remove}$ | $\bar{R}^2_{training}$ | $\bar{R}^2_{validation}$ |
|---|---|---|---|
| **0.2** | **0.1** | 0.538 | -0.030 |
| **0.15** | **0.08** | 0.538 | -0.030 |
| **0.12** | **0.06** | 0.538 | -0.030 |
| **0.1** | **0.05** | 0.577 | -0.020 |
| **0.05** | **0.025** | 0.739 | -0.427 |
| **0.04** | **0.02** | 0.739 | -0.427 |
| **0.03** | **0.01** | 0.817 | 0.232 |
| **0.025** | **0.01** | 0.817 | 0.232 |
| **0.025** | **0.005** | 0.819 | 0.168 |
| **0.01** | **0.005** | 0.925 | 0.416 |
| **0.005** | **0.0005** | 0.990 | 0.351 |

**Table A10: $R^2$ values calculated with stepwise linear regression models for P-P isomer MDA content with 3-fold cross-validation and different $P_{Enter}$ and $P_{Remove}$ parameters**

| $P_{Enter}$ | $P_{Remove}$ | $\bar{R}^2_{training}$ | $\bar{R}^2_{validation}$ |
|---|---|---|---|
| **0.1** | **0.05** | 0.945 | 0.480 |
| **0.05** | **0.025** | 0.945 | 0.480 |
| **0.04** | **0.02** | 0.959 | 0.425 |
| **0.03** | **0.01** | 0.969 | 0.427 |
| **0.2** | **0.1** | 0.945 | 0.480 |
| **0.15** | **0.08** | 0.945 | 0.480 |
| **0.12** | **0.06** | 0.945 | 0.480 |
| **0.025** | **0.01** | 0.969 | 0.427 |
| **0.025** | **0.005** | 0.969 | 0.427 |
| **0.01** | **0.005** | 0.979 | 0.420 |

**Table A11: $R^2$ values calculated with stepwise linear regression models for MMM by-product content with 3-fold cross-validation and different $P_{Enter}$ and $P_{Remove}$ parameters**

| $P_{Enter}$ | $P_{Remove}$ | $\bar{R}^2_{training}$ | $\bar{R}^2_{validation}$ |
|---|---|---|---|
| **0.2** | **0.1** | 0.810 | 0.713 |
| **0.15** | **0.08** | 0.860 | 0.749 |
| **0.12** | **0.06** | 0.905 | 0.779 |
| **0.1** | **0.05** | 0.905 | 0.779 |
| **0.05** | **0.025** | 0.905 | 0.779 |
| **0.04** | **0.02** | 0.922 | 0.893 |
| **0.03** | **0.01** | 0.922 | 0.893 |
| **0.025** | **0.01** | 0.922 | 0.893 |
| **0.025** | **0.005** | 0.922 | 0.893 |
| **0.01** | **0.0005** | 0.975 | 0.456 |

**Table A12: $R^2$ values calculated with stepwise linear regression models for >4-ring polymer MDA content with 3-fold cross-validation and different $P_{Enter}$ and $P_{Remove}$ parameters**

| $P_{Enter}$ | $P_{Remove}$ | $\bar{R}^2_{training}$ | $\bar{R}^2_{validation}$ |
|---|---|---|---|
| **0.2** | **0.1** | 0.645 | 0.519 |
| **0.15** | **0.08** | 0.759 | 0.556 |
| **0.12** | **0.06** | 0.801 | 0.624 |
| **0.1** | **0.05** | 0.801 | 0.624 |
| **0.05** | **0.025** | 0.865 | 0.437 |
| **0.04** | **0.02** | 0.879 | 0.462 |
| **0.03** | **0.01** | 0.901 | 0.488 |
| **0.027** | **0.01** | 0.901 | 0.488 |
| **0.025** | **0.01** | 0.901 | 0.488 |
| **0.025** | **0.005** | 0.901 | 0.488 |

**Table A13: Performance of neural networks with different hyperparameterizations for 2-ring MDA content with 3-fold cross-validation**

| Structure | Act. Fcn. | λ | $R^2$training | $R^2$validation |
|---|---|---|---|---|
| 10-10-10 | ReLU | $10^{-4}$ | 0.999 | 0.813 |
| 10-10-10 | Tanh | $10^{-4}$ | 0.998 | 0.418 |
| 10-10-10 | Sigmoid | $10^{-4}$ | 0.994 | 0.701 |
| 10-10-10 | ReLU | $10^{-3}$ | 0.996 | 0.886 |
| 10-10-10 | Tanh | $10^{-3}$ | 0.992 | 0.707 |
| 10-10-10 | Sigmoid | $10^{-3}$ | 0.000 | -0.045 |
| 20-20-20 | ReLU | $10^{-4}$ | 1.000 | 0.870 |
| 20-20-20 | Tanh | $10^{-4}$ | 0.998 | 0.626 |
| 20-20-20 | Sigmoid | $10^{-4}$ | 0.994 | 0.703 |
| 20-20-20 | ReLU | $10^{-3}$ | 0.997 | 0.885 |
| 20-20-20 | Tanh | $10^{-3}$ | 0.990 | 0.702 |
| 20-20-20 | Sigmoid | $10^{-3}$ | 0.990 | 0.702 |
| 30-30-30 | ReLU | $10^{-4}$ | 1.000 | 0.845 |
| 30-30-30 | Tanh | $10^{-4}$ | 0.998 | 0.612 |
| 30-30-30 | Sigmoid | $10^{-4}$ | 0.994 | 0.703 |
| 30-30-30 | ReLU | $10^{-3}$ | 0.997 | 0.891 |
| 30-30-30 | Tanh | $10^{-3}$ | 0.991 | 0.712 |
| 30-30-30 | Sigmoid | $10^{-3}$ | 0.000 | -0.045 |
| 40-40-40 | ReLU | $10^{-4}$ | 1.000 | 0.834 |
| 40-40-40 | Tanh | $10^{-4}$ | 0.999 | 0.644 |
| 40-40-40 | Sigmoid | $10^{-4}$ | 0.994 | 0.704 |
| 40-40-40 | ReLU | $10^{-3}$ | 0.997 | 0.911 |
| 40-40-40 | Tanh | $10^{-3}$ | 0.991 | 0.710 |
| 40-40-40 | Sigmoid | $10^{-3}$ | 0.296 | 0.252 |

**Table A14: Performance of neural networks with different hyperparameterizations for 3-ring MDA content with 3-fold cross-validation**

| Structure | Act. Fcn. | λ | $R^2$training | $R^2$validation |
|---|---|---|---|---|
| 10-10-10 | ReLU | $10^{-4}$ | 0.999 | 0.860 |
| 10-10-10 | Tanh | $10^{-4}$ | 0.998 | 0.781 |
| 10-10-10 | Sigmoid | $10^{-4}$ | 0.989 | 0.850 |
| 10-10-10 | ReLU | $10^{-3}$ | 0.995 | 0.945 |
| 10-10-10 | Tanh | $10^{-3}$ | 0.982 | 0.859 |
| 10-10-10 | Sigmoid | $10^{-3}$ | 0.302 | 0.293 |
| 20-20-20 | ReLU | $10^{-4}$ | 0.999 | 0.889 |
| 20-20-20 | Tanh | $10^{-4}$ | 0.998 | 0.848 |
| 20-20-20 | Sigmoid | $10^{-4}$ | 0.989 | 0.852 |
| 20-20-20 | ReLU | $10^{-3}$ | 0.997 | 0.794 |
| 20-20-20 | Tanh | $10^{-3}$ | 0.982 | 0.858 |
| 20-20-20 | Sigmoid | $10^{-3}$ | 0.625 | 0.514 |
| 30-30-30 | ReLU | $10^{-4}$ | 1.000 | 0.873 |
| 30-30-30 | Tanh | $10^{-4}$ | 0.998 | 0.831 |
| 30-30-30 | Sigmoid | $10^{-4}$ | 0.989 | 0.862 |
| 30-30-30 | ReLU | $10^{-3}$ | 0.997 | 0.919 |
| 30-30-30 | Tanh | $10^{-3}$ | 0.982 | 0.858 |
| 30-30-30 | Sigmoid | $10^{-3}$ | 0.923 | 0.789 |
| 40-40-40 | ReLU | $10^{-4}$ | 0.999 | 0.893 |
| 40-40-40 | Tanh | $10^{-4}$ | 0.998 | 0.862 |
| 40-40-40 | Sigmoid | $10^{-4}$ | 0.989 | 0.856 |
| 40-40-40 | ReLU | $10^{-3}$ | 0.996 | 0.917 |
| 40-40-40 | Tanh | $10^{-3}$ | 0.969 | 0.852 |
| 40-40-40 | Sigmoid | $10^{-3}$ | 0.923 | 0.790 |

**Table A15: Performance of neural networks with different hyperparameterizations for 4-ring MDA content with 3-fold cross-validation**

| Structure | Act. Fcn. | λ | $R^2$training | $R^2$validation |
|---|---|---|---|---|
| 10-10-10 | ReLU | $10^{-4}$ | 1.000 | 0.691 |
| 10-10-10 | Tanh | $10^{-4}$ | 0.999 | 0.651 |
| 10-10-10 | Sigmoid | $10^{-4}$ | 0.979 | 0.673 |
| 10-10-10 | ReLU | $10^{-3}$ | 0.995 | 0.807 |
| 10-10-10 | Tanh | $10^{-3}$ | 0.966 | 0.727 |
| 10-10-10 | Sigmoid | $10^{-3}$ | 0.310 | 0.189 |
| 20-20-20 | ReLU | $10^{-4}$ | 0.999 | 0.723 |
| 20-20-20 | Tanh | $10^{-4}$ | 0.998 | 0.734 |
| 20-20-20 | Sigmoid | $10^{-4}$ | 0.973 | 0.694 |
| 20-20-20 | ReLU | $10^{-3}$ | 0.995 | 0.803 |
| 20-20-20 | Tanh | $10^{-3}$ | 0.966 | 0.727 |
| 20-20-20 | Sigmoid | $10^{-3}$ | 0.623 | 0.451 |
| 30-30-30 | ReLU | $10^{-4}$ | 1.000 | 0.742 |
| 30-30-30 | Tanh | $10^{-4}$ | 0.999 | 0.804 |
| 30-30-30 | Sigmoid | $10^{-4}$ | 0.973 | 0.696 |
| 30-30-30 | ReLU | $10^{-3}$ | 0.995 | 0.796 |
| 30-30-30 | Tanh | $10^{-3}$ | 0.966 | 0.727 |
| 30-30-30 | Sigmoid | $10^{-3}$ | 0.913 | 0.672 |
| 40-40-40 | ReLU | $10^{-4}$ | 1.000 | 0.782 |
| 40-40-40 | Tanh | $10^{-4}$ | 0.999 | 0.665 |
| 40-40-40 | Sigmoid | $10^{-4}$ | 0.973 | 0.696 |
| 40-40-40 | ReLU | $10^{-3}$ | 0.995 | 0.730 |
| 40-40-40 | Tanh | $10^{-3}$ | 0.966 | 0.727 |
| 40-40-40 | Sigmoid | $10^{-3}$ | 0.913 | 0.671 |

**Table A16: Performance of neural networks with different hyperparameterizations for 5-ring MDA content with 3-fold cross-validation**

| Structure | Act. Fcn. | λ | $R^2$training | $R^2$validation |
|---|---|---|---|---|
| 10-10-10 | ReLU | $10^{-4}$ | 0.999 | 0.523 |
| 10-10-10 | Tanh | $10^{-4}$ | 0.998 | -0.359 |
| 10-10-10 | Sigmoid | $10^{-4}$ | 0.972 | 0.625 |
| 10-10-10 | ReLU | $10^{-3}$ | 0.986 | 0.629 |
| 10-10-10 | Tanh | $10^{-3}$ | 0.964 | 0.587 |
| 10-10-10 | Sigmoid | $10^{-3}$ | 0.308 | 0.107 |
| 20-20-20 | ReLU | $10^{-4}$ | 1.000 | 0.341 |
| 20-20-20 | Tanh | $10^{-4}$ | 0.997 | -0.439 |
| 20-20-20 | Sigmoid | $10^{-4}$ | 0.972 | 0.623 |
| 20-20-20 | ReLU | $10^{-3}$ | 0.990 | 0.609 |
| 20-20-20 | Tanh | $10^{-3}$ | 0.963 | 0.559 |
| 20-20-20 | Sigmoid | $10^{-3}$ | 0.614 | 0.328 |
| 30-30-30 | ReLU | $10^{-4}$ | 1.000 | 0.486 |
| 30-30-30 | Tanh | $10^{-4}$ | 0.997 | -0.678 |
| 30-30-30 | Sigmoid | $10^{-4}$ | 0.972 | 0.623 |
| 30-30-30 | ReLU | $10^{-3}$ | 0.989 | 0.704 |
| 30-30-30 | Tanh | $10^{-3}$ | 0.964 | 0.587 |
| 30-30-30 | Sigmoid | $10^{-3}$ | 0.601 | 0.247 |
| 40-40-40 | ReLU | $10^{-4}$ | 1.000 | 0.446 |
| 40-40-40 | Tanh | $10^{-4}$ | 0.997 | 0.148 |
| 40-40-40 | Sigmoid | $10^{-4}$ | 0.972 | 0.624 |
| 40-40-40 | ReLU | $10^{-3}$ | 0.992 | 0.569 |
| 40-40-40 | Tanh | $10^{-3}$ | 0.963 | 0.559 |
| 40-40-40 | Sigmoid | $10^{-3}$ | 0.613 | 0.329 |

**Table A17: Performance of neural networks with different hyperparameterizations for >6-ring MDA content with 3-fold cross-validation**

| Structure | Act. Fcn. | λ | $R^2$training | $R^2$validation |
|---|---|---|---|---|
| 10-10-10 | ReLU | $10^{-4}$ | 0.998 | -2.542 |
| 10-10-10 | Tanh | $10^{-4}$ | 0.997 | -5.459 |
| 10-10-10 | Sigmoid | $10^{-4}$ | 0.959 | -6.079 |
| 10-10-10 | ReLU | $10^{-3}$ | 0.980 | -2.620 |
| 10-10-10 | Tanh | $10^{-3}$ | 0.911 | -0.945 |
| 10-10-10 | Sigmoid | $10^{-3}$ | 0.224 | -0.147 |
| 20-20-20 | ReLU | $10^{-4}$ | 0.999 | -1.240 |
| 20-20-20 | Tanh | $10^{-4}$ | 0.997 | -4.583 |
| 20-20-20 | Sigmoid | $10^{-4}$ | 0.961 | -6.676 |
| 20-20-20 | ReLU | $10^{-3}$ | 0.984 | -1.173 |
| 20-20-20 | Tanh | $10^{-3}$ | 0.915 | -0.976 |
| 20-20-20 | Sigmoid | $10^{-3}$ | 0.000 | -0.233 |
| 30-30-30 | ReLU | $10^{-4}$ | 0.999 | -0.797 |
| 30-30-30 | Tanh | $10^{-4}$ | 0.997 | -8.297 |
| 30-30-30 | Sigmoid | $10^{-4}$ | 0.862 | -0.433 |
| 30-30-30 | ReLU | $10^{-3}$ | 0.988 | -1.835 |
| 30-30-30 | Tanh | $10^{-3}$ | 0.922 | -1.080 |
| 30-30-30 | Sigmoid | $10^{-3}$ | 0.224 | -0.147 |
| 40-40-40 | ReLU | $10^{-4}$ | 1.000 | -3.019 |
| 40-40-40 | Tanh | $10^{-4}$ | 0.997 | -7.578 |
| 40-40-40 | Sigmoid | $10^{-4}$ | 0.931 | -1.210 |
| 40-40-40 | ReLU | $10^{-3}$ | 0.983 | -0.897 |
| 40-40-40 | Tanh | $10^{-3}$ | 0.914 | -1.006 |
| 40-40-40 | Sigmoid | $10^{-3}$ | 0.000 | -0.233 |

**Table A18: Performance of neural networks with different hyperparameterizations for O-O isomer MDA content with 3-fold cross-validation**

| Structure | Act. Fcn. | λ | $R^2$training | $R^2$validation |
|---|---|---|---|---|
| 10-10-10 | ReLU | $10^{-4}$ | 1.000 | 0.868 |
| 10-10-10 | Tanh | $10^{-4}$ | 0.999 | 0.859 |
| 10-10-10 | Sigmoid | $10^{-4}$ | 0.997 | 0.814 |
| 10-10-10 | ReLU | $10^{-3}$ | 0.998 | 0.891 |
| 10-10-10 | Tanh | $10^{-3}$ | 0.996 | 0.840 |
| 10-10-10 | Sigmoid | $10^{-3}$ | 0.970 | 0.914 |
| 20-20-20 | ReLU | $10^{-4}$ | 1.000 | 0.859 |
| 20-20-20 | Tanh | $10^{-4}$ | 0.999 | 0.886 |
| 20-20-20 | Sigmoid | $10^{-4}$ | 0.997 | 0.804 |
| 20-20-20 | ReLU | $10^{-3}$ | 0.998 | 0.910 |
| 20-20-20 | Tanh | $10^{-3}$ | 0.994 | 0.873 |
| 20-20-20 | Sigmoid | $10^{-3}$ | 0.970 | 0.913 |
| 30-30-30 | ReLU | $10^{-4}$ | 1.000 | 0.923 |
| 30-30-30 | Tanh | $10^{-4}$ | 0.999 | 0.879 |
| 30-30-30 | Sigmoid | $10^{-4}$ | 0.997 | 0.847 |
| 30-30-30 | ReLU | $10^{-3}$ | 0.998 | 0.901 |
| 30-30-30 | Tanh | $10^{-3}$ | 0.995 | 0.847 |
| 30-30-30 | Sigmoid | $10^{-3}$ | 0.997 | 0.847 |
| 40-40-40 | ReLU | $10^{-4}$ | 1.000 | 0.859 |
| 40-40-40 | Tanh | $10^{-4}$ | 1.000 | 0.880 |
| 40-40-40 | Sigmoid | $10^{-4}$ | 0.997 | 0.849 |
| 40-40-40 | ReLU | $10^{-3}$ | 0.999 | 0.899 |
| 40-40-40 | Tanh | $10^{-3}$ | 0.990 | 0.890 |
| 40-40-40 | Sigmoid | $10^{-3}$ | 0.970 | 0.912 |

**Table A19: Performance of neural networks with different hyperparameterizations for O-P isomer MDA content with 3-fold cross-validation**

| Structure | Act. Fcn. | λ | $R^2$training | $R^2$validation |
|---|---|---|---|---|
| 10-10-10 | ReLU | $10^{-4}$ | 1.000 | 0.678 |
| 10-10-10 | Tanh | $10^{-4}$ | 1.000 | 0.091 |
| 10-10-10 | Sigmoid | $10^{-4}$ | 0.992 | 0.401 |
| 10-10-10 | ReLU | $10^{-3}$ | 0.997 | 0.727 |
| 10-10-10 | Tanh | $10^{-3}$ | 0.989 | 0.460 |
| 10-10-10 | Sigmoid | $10^{-3}$ | 0.920 | 0.517 |
| 20-20-20 | ReLU | $10^{-4}$ | 1.000 | 0.871 |
| 20-20-20 | Tanh | $10^{-4}$ | 1.000 | 0.306 |
| 20-20-20 | Sigmoid | $10^{-4}$ | 0.993 | 0.467 |
| 20-20-20 | ReLU | $10^{-3}$ | 0.997 | 0.703 |
| 20-20-20 | Tanh | $10^{-3}$ | 0.989 | 0.466 |
| 20-20-20 | Sigmoid | $10^{-3}$ | 0.858 | 0.446 |
| 30-30-30 | ReLU | $10^{-4}$ | 1.000 | 0.824 |
| 30-30-30 | Tanh | $10^{-4}$ | 0.999 | -0.053 |
| 30-30-30 | Sigmoid | $10^{-4}$ | 0.993 | 0.402 |
| 30-30-30 | ReLU | $10^{-3}$ | 0.998 | 0.736 |
| 30-30-30 | Tanh | $10^{-3}$ | 0.989 | 0.463 |
| 30-30-30 | Sigmoid | $10^{-3}$ | 0.859 | 0.446 |
| 40-40-40 | ReLU | $10^{-4}$ | 1.000 | 0.787 |
| 40-40-40 | Tanh | $10^{-4}$ | 0.999 | -0.093 |
| 40-40-40 | Sigmoid | $10^{-4}$ | 0.994 | 0.467 |
| 40-40-40 | ReLU | $10^{-3}$ | 0.997 | 0.778 |
| 40-40-40 | Tanh | $10^{-3}$ | 0.989 | 0.467 |
| 40-40-40 | Sigmoid | $10^{-3}$ | 0.859 | 0.446 |

**Table A20: Performance of neural networks with different hyperparameterizations for P-P isomer MDA content with 3-fold cross-validation**

| Structure | Act. Fcn. | λ | $R^2$training | $R^2$validation |
|---|---|---|---|---|
| 10-10-10 | ReLU | $10^{-4}$ | 1.000 | 0.583 |
| 10-10-10 | Tanh | $10^{-4}$ | 0.999 | 0.510 |
| 10-10-10 | Sigmoid | $10^{-4}$ | 0.986 | 0.777 |
| 10-10-10 | ReLU | $10^{-3}$ | 0.998 | 0.761 |
| 10-10-10 | Tanh | $10^{-3}$ | 0.974 | 0.900 |
| 10-10-10 | Sigmoid | $10^{-3}$ | 0.974 | 0.900 |
| 20-20-20 | ReLU | $10^{-4}$ | 1.000 | 0.691 |
| 20-20-20 | Tanh | $10^{-4}$ | 0.999 | 0.508 |
| 20-20-20 | Sigmoid | $10^{-4}$ | 0.986 | 0.793 |
| 20-20-20 | ReLU | $10^{-3}$ | 0.998 | 0.851 |
| 20-20-20 | Tanh | $10^{-3}$ | 0.984 | 0.840 |
| 20-20-20 | Sigmoid | $10^{-3}$ | 0.974 | 0.908 |
| 30-30-30 | ReLU | $10^{-4}$ | 1.000 | 0.662 |
| 30-30-30 | Tanh | $10^{-4}$ | 0.999 | 0.333 |
| 30-30-30 | Sigmoid | $10^{-4}$ | 0.986 | 0.803 |
| 30-30-30 | ReLU | $10^{-3}$ | 0.998 | 0.550 |
| 30-30-30 | Tanh | $10^{-3}$ | 0.984 | 0.850 |
| 30-30-30 | Sigmoid | $10^{-3}$ | 0.974 | 0.910 |
| 40-40-40 | ReLU | $10^{-4}$ | 1.000 | 0.604 |
| 40-40-40 | Tanh | $10^{-4}$ | 0.999 | 0.508 |
| 40-40-40 | Sigmoid | $10^{-4}$ | 0.986 | 0.800 |
| 40-40-40 | ReLU | $10^{-3}$ | 0.998 | 0.661 |
| 40-40-40 | Tanh | $10^{-3}$ | 0.984 | 0.842 |
| 40-40-40 | Sigmoid | $10^{-3}$ | 0.974 | 0.911 |

**Table A21: Performance of neural networks with different hyperparameterizations for models for MMM by-product content with 3-fold cross-validation**

| Structure | Act. Fcn. | λ | $R^2$training | $R^2$validation |
|---|---|---|---|---|
| 10-10-10 | ReLU | $10^{-4}$ | 1.000 | 0.862 |
| 10-10-10 | Tanh | $10^{-4}$ | 0.999 | 0.814 |
| 10-10-10 | Sigmoid | $10^{-4}$ | 0.994 | 0.829 |
| 10-10-10 | ReLU | $10^{-3}$ | 0.997 | 0.830 |
| 10-10-10 | Tanh | $10^{-3}$ | 0.992 | 0.810 |
| 10-10-10 | Sigmoid | $10^{-3}$ | 0.958 | 0.780 |
| 20-20-20 | ReLU | $10^{-4}$ | 1.000 | 0.827 |
| 20-20-20 | Tanh | $10^{-4}$ | 0.999 | 0.877 |
| 20-20-20 | Sigmoid | $10^{-4}$ | 0.994 | 0.817 |
| 20-20-20 | ReLU | $10^{-3}$ | 0.997 | 0.821 |
| 20-20-20 | Tanh | $10^{-3}$ | 0.992 | 0.812 |
| 20-20-20 | Sigmoid | $10^{-3}$ | 0.958 | 0.776 |
| 30-30-30 | ReLU | $10^{-4}$ | 0.997 | 0.821 |
| 30-30-30 | Tanh | $10^{-4}$ | 0.999 | 0.905 |
| 30-30-30 | Sigmoid | $10^{-4}$ | 0.994 | 0.816 |
| 30-30-30 | ReLU | $10^{-3}$ | 0.997 | 0.836 |
| 30-30-30 | Tanh | $10^{-3}$ | 0.990 | 0.806 |
| 30-30-30 | Sigmoid | $10^{-3}$ | 0.957 | 0.774 |
| 40-40-40 | ReLU | $10^{-4}$ | 1.000 | 0.868 |
| 40-40-40 | Tanh | $10^{-4}$ | 0.999 | 0.838 |
| 40-40-40 | Sigmoid | $10^{-4}$ | 0.994 | 0.818 |
| 40-40-40 | ReLU | $10^{-3}$ | 0.997 | 0.839 |
| 40-40-40 | Tanh | $10^{-3}$ | 0.991 | 0.803 |
| 40-40-40 | Sigmoid | $10^{-3}$ | 0.957 | 0.773 |

**Table A22: Performance of neural networks with different hyperparameterizations for >4-ring MDA content with 3-fold cross-validation**

| Structure | Act. Fcn. | λ | $R^2$training | $R^2$validation |
|---|---|---|---|---|
| 10-10-10 | ReLU | $10^{-4}$ | 0.999 | 0.684 |
| 10-10-10 | Tanh | $10^{-4}$ | 0.998 | 0.140 |
| 10-10-10 | Sigmoid | $10^{-4}$ | 0.971 | 0.636 |
| 10-10-10 | ReLU | $10^{-3}$ | 0.986 | 0.625 |
| 10-10-10 | Tanh | $10^{-3}$ | 0.964 | 0.629 |
| 10-10-10 | Sigmoid | $10^{-3}$ | 0.306 | 0.144 |
| 20-20-20 | ReLU | $10^{-4}$ | 0.999 | 0.554 |
| 20-20-20 | Tanh | $10^{-4}$ | 0.998 | 0.259 |
| 20-20-20 | Sigmoid | $10^{-4}$ | 0.972 | 0.620 |
| 20-20-20 | ReLU | $10^{-3}$ | 0.964 | 0.629 |
| 20-20-20 | Tanh | $10^{-3}$ | 0.964 | 0.629 |
| 20-20-20 | Sigmoid | $10^{-3}$ | 0.615 | 0.368 |
| 30-30-30 | ReLU | $10^{-4}$ | 1.000 | 0.471 |
| 30-30-30 | Tanh | $10^{-4}$ | 0.998 | 0.150 |
| 30-30-30 | Sigmoid | $10^{-4}$ | 0.971 | 0.636 |
| 30-30-30 | ReLU | $10^{-3}$ | 0.992 | 0.688 |
| 30-30-30 | Tanh | $10^{-3}$ | 0.964 | 0.629 |
| 30-30-30 | Sigmoid | $10^{-3}$ | 0.597 | 0.326 |
| 40-40-40 | ReLU | $10^{-4}$ | 1.000 | 0.658 |
| 40-40-40 | Tanh | $10^{-4}$ | 0.997 | 0.217 |
| 40-40-40 | Sigmoid | $10^{-4}$ | 0.971 | 0.637 |
| 40-40-40 | ReLU | $10^{-3}$ | 0.992 | 0.716 |
| 40-40-40 | Tanh | $10^{-3}$ | 0.964 | 0.629 |
| 40-40-40 | Sigmoid | $10^{-3}$ | 0.614 | 0.374 |

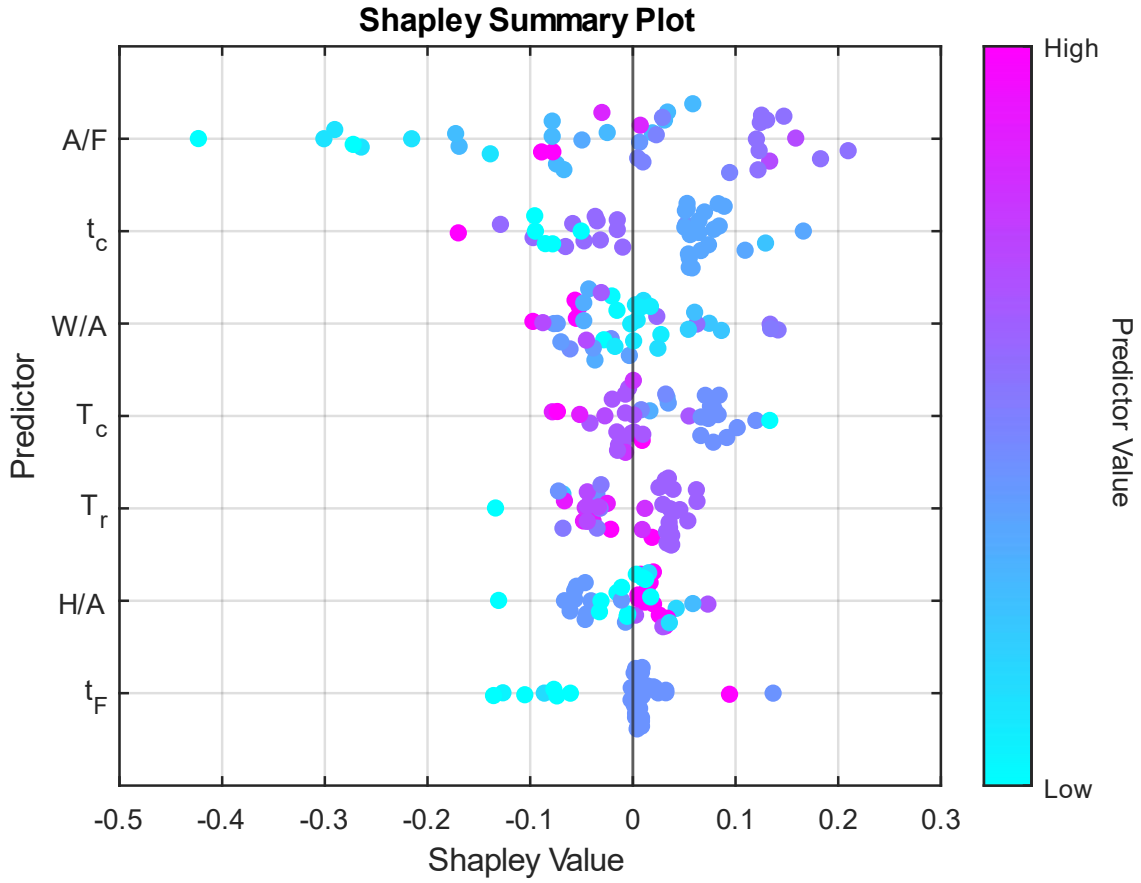


**Figure A1: Shapley values calculated with the best performing machine learning model for 2-ring formation**

Figure A1 displays the calculated Shapley values for the formation of a 2-ring MDA, highlighting that the aniline/formaldehyde molar ratio, the residence time of the condensation reaction and the water/aniline molar ratio are the most critical synthesis parameters. It is important to mention though the water content, i.e. W/A shows significant multicollinearity with the quantity of HCl employed as a catalyst. However, the temperature, HCl to aniline ratio, and formaline dosing time did not significantly impact the 2-ring MDA content.

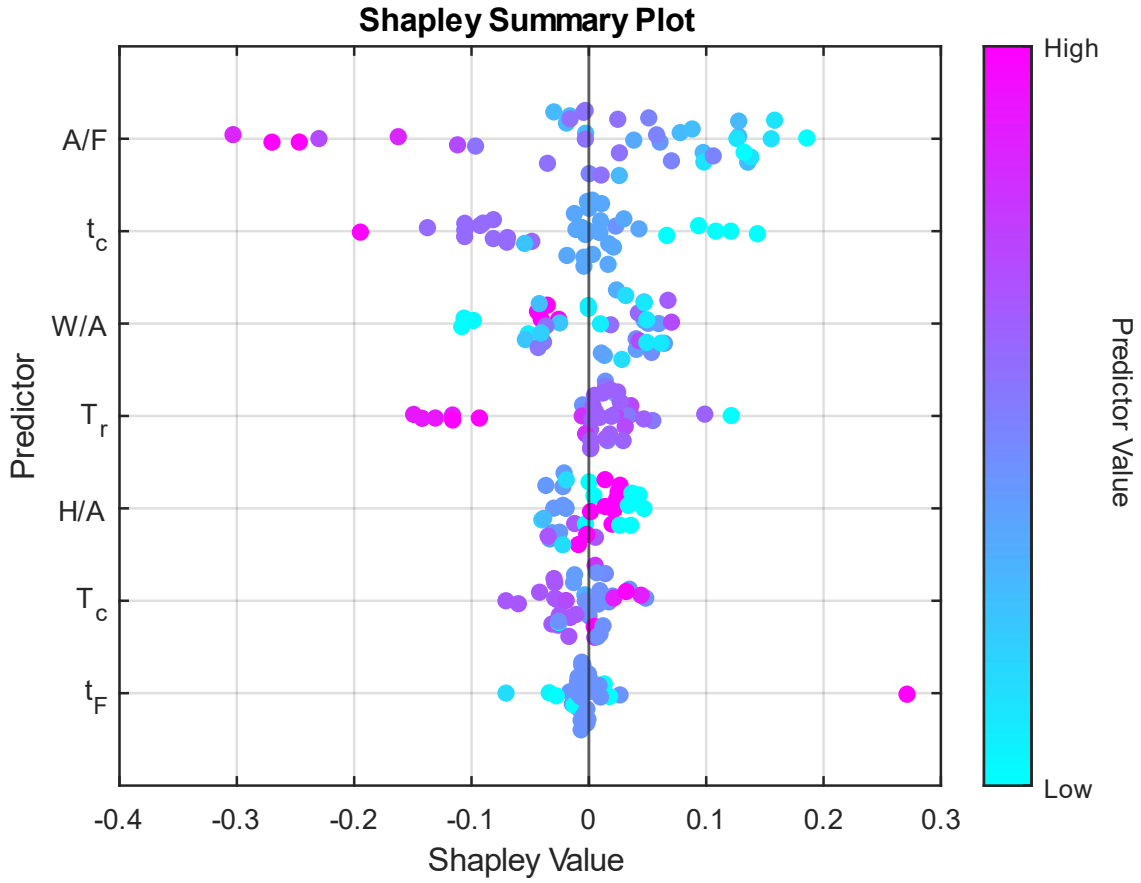


**Figure A2: Shapley values calculated with the best performing machine learning model for 3-ring formation**

Figure A2 displays the calculated Shapley values for the formation of a 3-ring MDA, highlighting that the aniline/formaldehyde molar ratio, the residence time of the condensation reaction and the water/aniline molar ratio are also the most critical synthesis parameters like in case for the 2-ring MDA content. However, these critical synthesis parameters, particularly the A/F parameter, exhibit variations. Increasing the amount of aniline molecules, i.e. A/F ratio enhances 2-ring formation as the activated molecules are more likely to interact with aniline, resulting in 2-ring MDA, rather than reacting other MDA molecules to form multi-ring MDA.

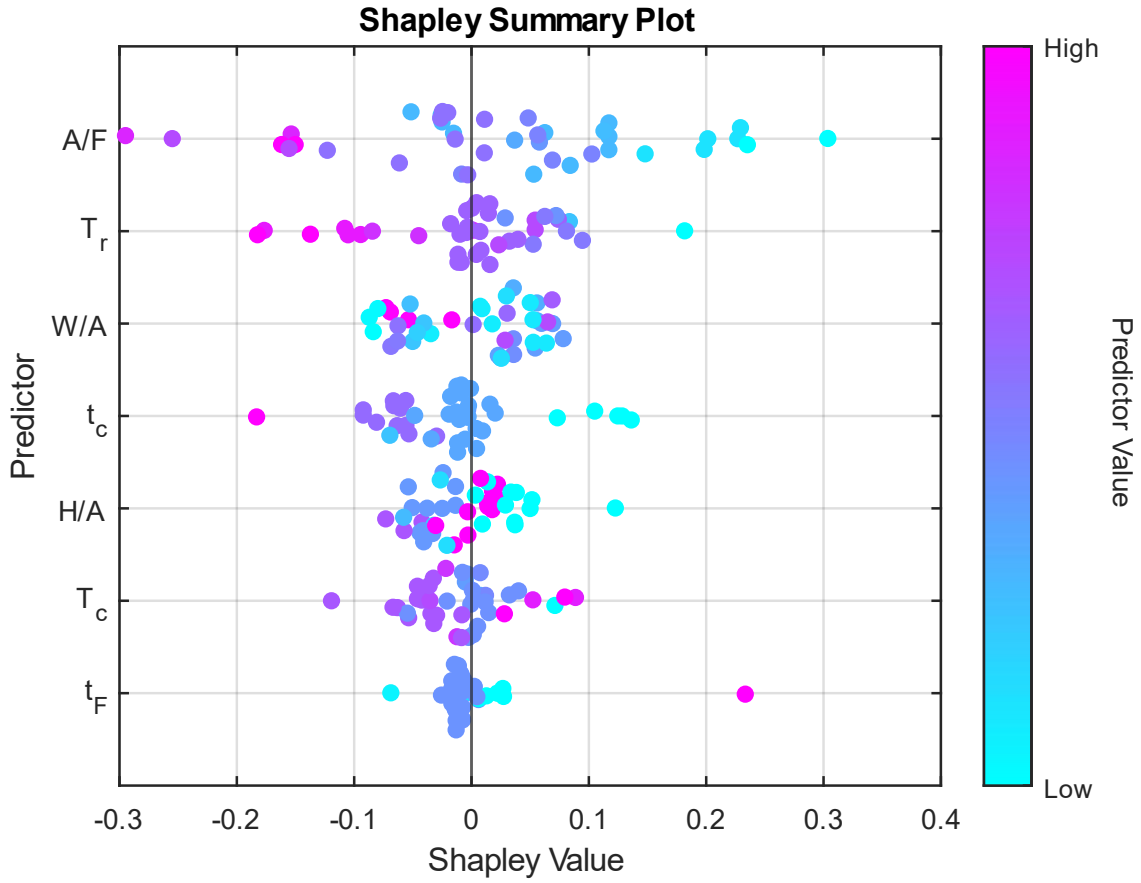


**Figure A3: Shapley values calculated with the best performing machine learning model for 4-ring formation**

Figure A3 displays the calculated Shapley values for the formation of a 4-ring MDA, highlighting that the aniline/formaldehyde molar ratio, the temperature of the rearrangement reaction and the water/aniline molar ratio are also the most critical synthesis parameters. The importance of the condensation reaction residence time decreased, but still has a significant impact on the 4-ring formation and and has the same direction as in case of 3-ring formation. Nonetheless, decreasing the rearrangement reaction temperature has a significantly positive effect on the 4-ring formation.

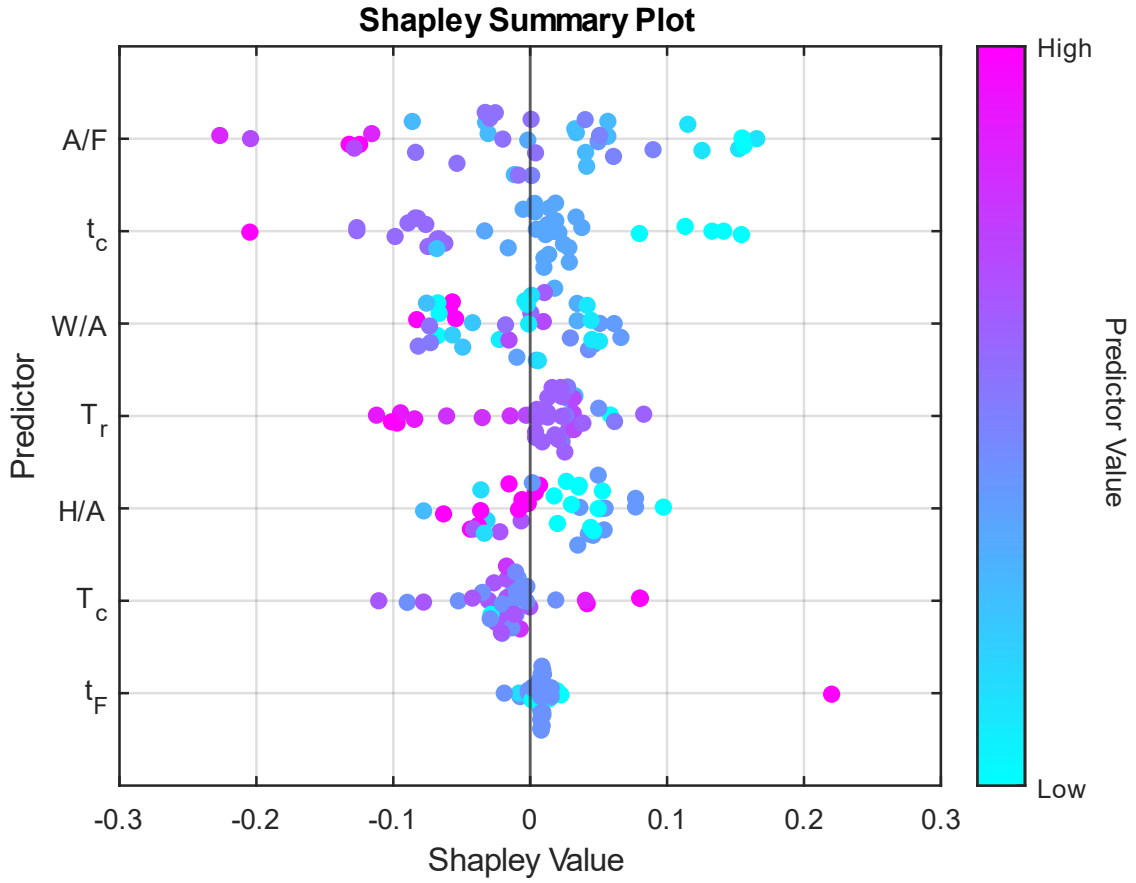


**Figure A4: Shapley values calculated with the best performing machine learning model for 5-ring formation**

Figure A4 displays the calculated Shapley values for the formation of a 5-ring MDA. The significance and influence of synthesis parameters closely resemble those observed in other multi-ring MDA configurations.

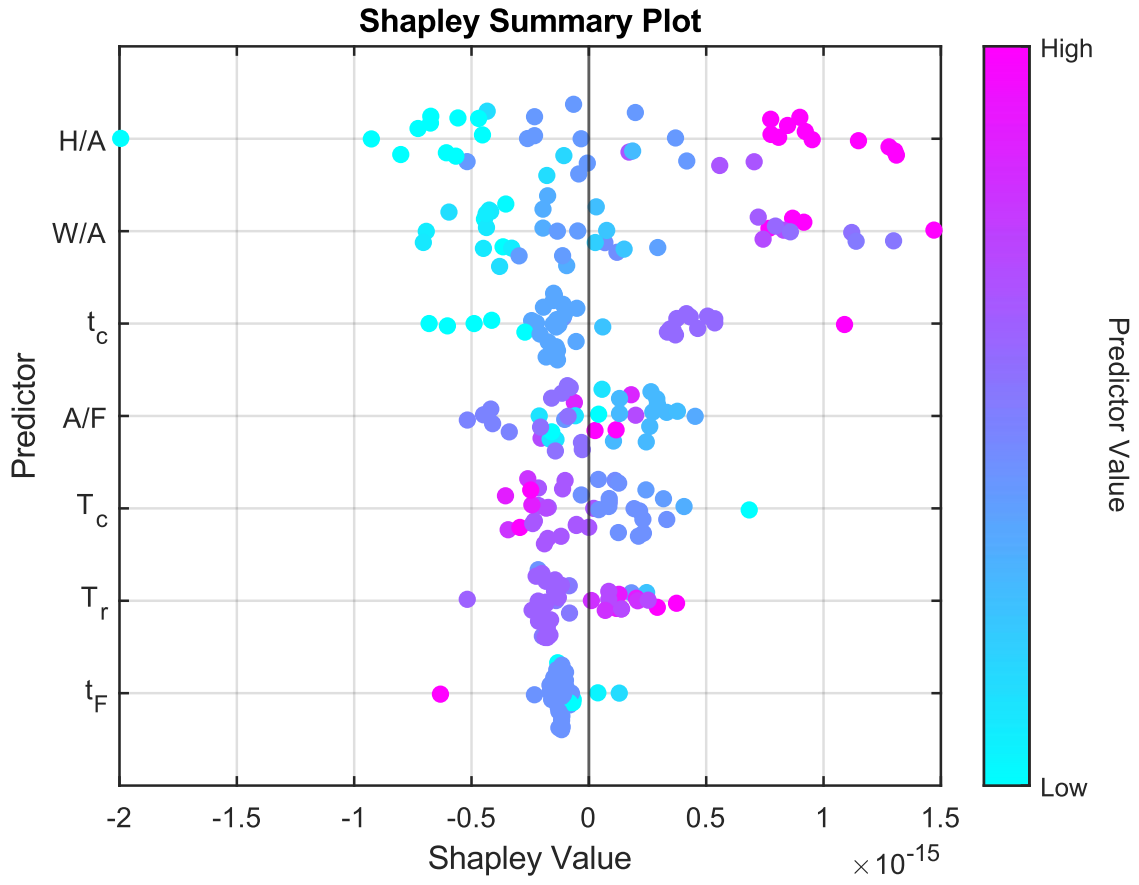


**Figure A5: Shapley values calculated with the best performing machine learning model for 6-ring formation**

Figure A5 displays the calculated Shapley values for the formation of a 6-ring MDA. The impact and role of synthesis parameters differ from those in other multi-ring MDA formations. Nevertheless, it's crucial to highlight that machine learning models demonstrated the poorest performance when predicting solely the quantity of 6-ring formations compared to predictions for other multi-ring formations. In the synthesis of 6-ring MDA, H/A and W/A are became the most important parameters, while A/F and condensation reaction residence time were lesser significant. Meanwhile, reaction temperatures and formaline dosing time were the least significant factors.

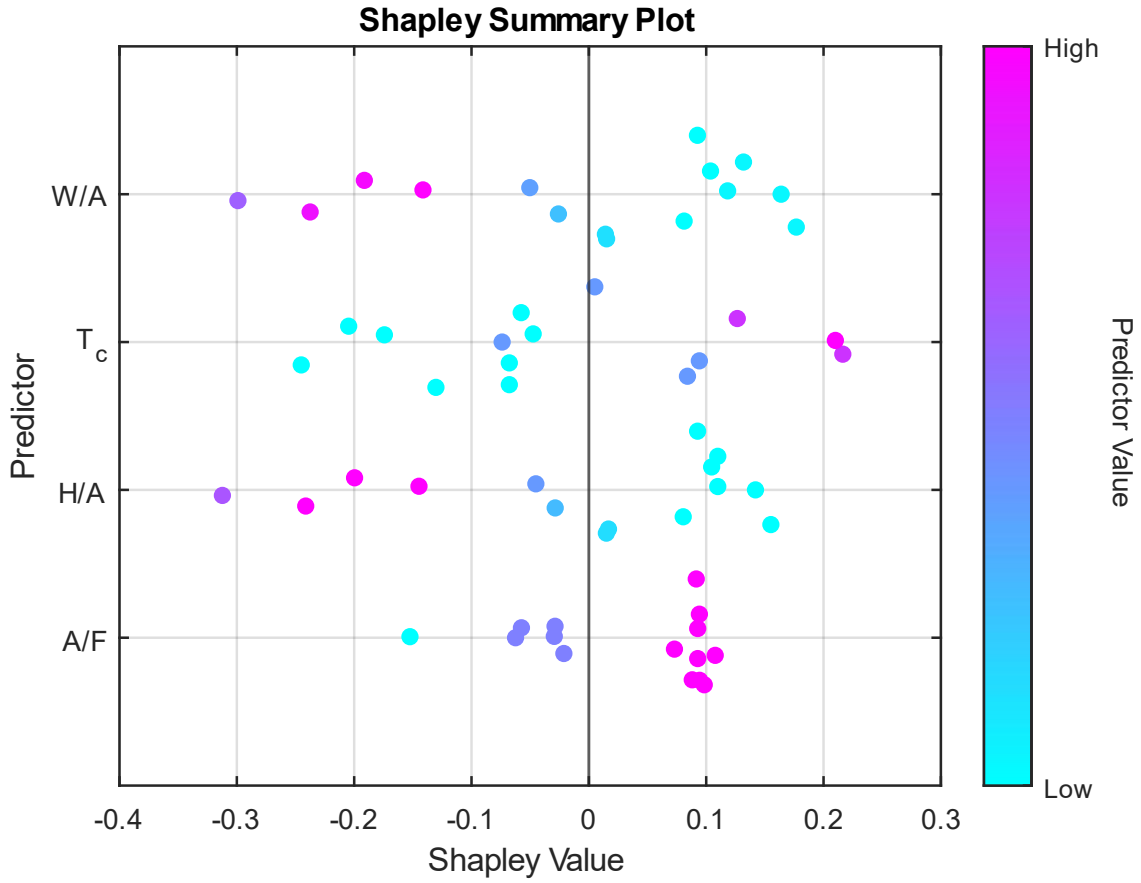


**Figure A6: Shapley values calculated with the best performing machine learning model for O-O isomer formation**

Figure A6 displays the calculated Shapley values for the formation of the O-O MDA isomers. As emphasized in Figure 39, the formation of O-O and P-P isomers has been forecasted based solely on four synthesis parameters, in contrast with other MDA quality parameters. For O-O isomer formation, the W/A ratio and condensation reaction temperature emerged as the key factors, whereas other ratio parameters such as H/A and A/F held less significance.

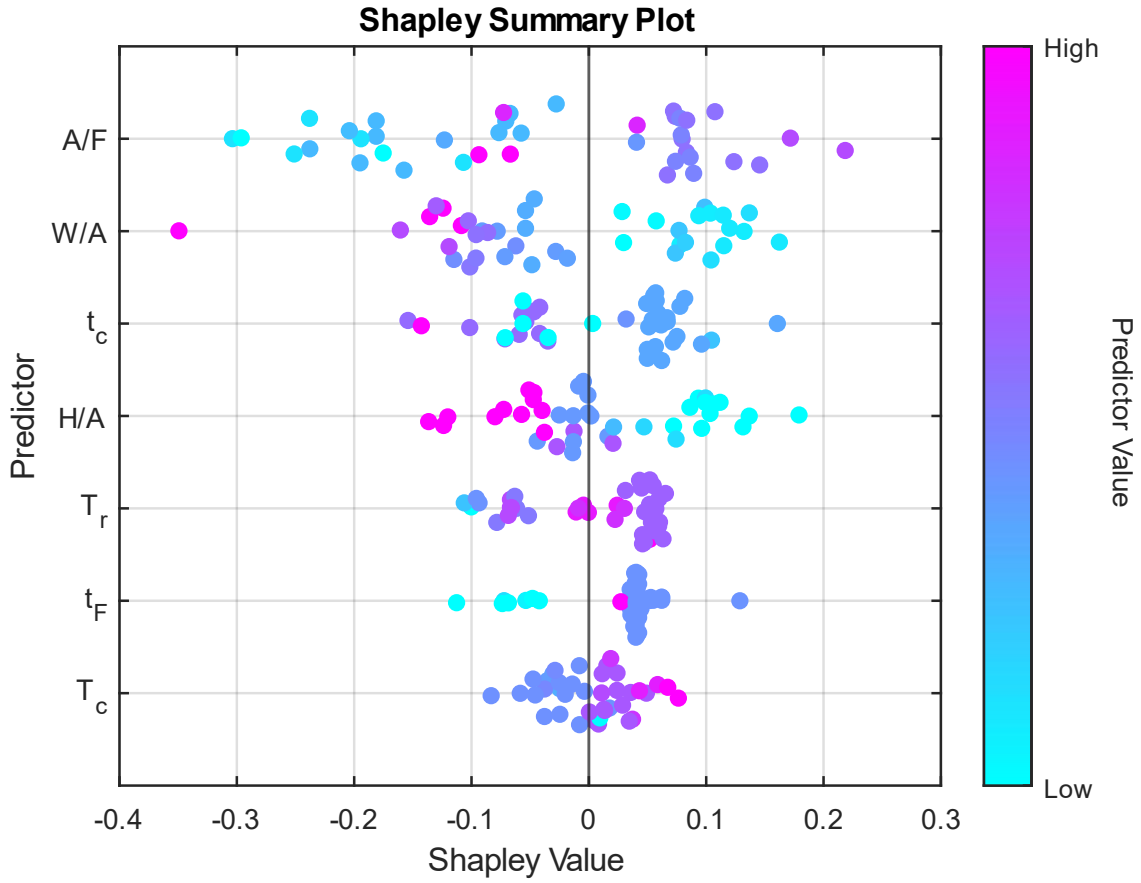


**Figure A7: Shapley values calculated with the best performing machine learning model for O-P isomer formation**

Figure A7 displays the calculated Shapley values for the formation of the O-P MDA isomers. In the case of the O-P isomer ratio, seven synthesis parameters were examined and their order of importance showed similar results to those observed previously during the analysis of ring distributions. It can be stated that the formation of O-P isomers is favored by similar effects as those favoring the formation of 2-ring MDA molecules, differences may also be observed, as in the case of the $T_c$ parameter. For example, a low A/F ratio negatively affects the formation of both 2-ring molecules and O-P isomers, while higher $T_c$ has a positive effect on O-P, but negative effect on 2-ring formation. However, in the case of O-P isomer formation, the Shapley values are found over a wider parameter range, meaning that the individual synthesis parameters influence O-P formation in a similar but differently weighted compared to 2-ring MDA formation. This is clearly observable, for instance, in the case of the H/A ratio: while the Shapley values for O-P range between -0.2 and 0.2, they range only between -0.1 and 0.1 for 2-ring MDA.

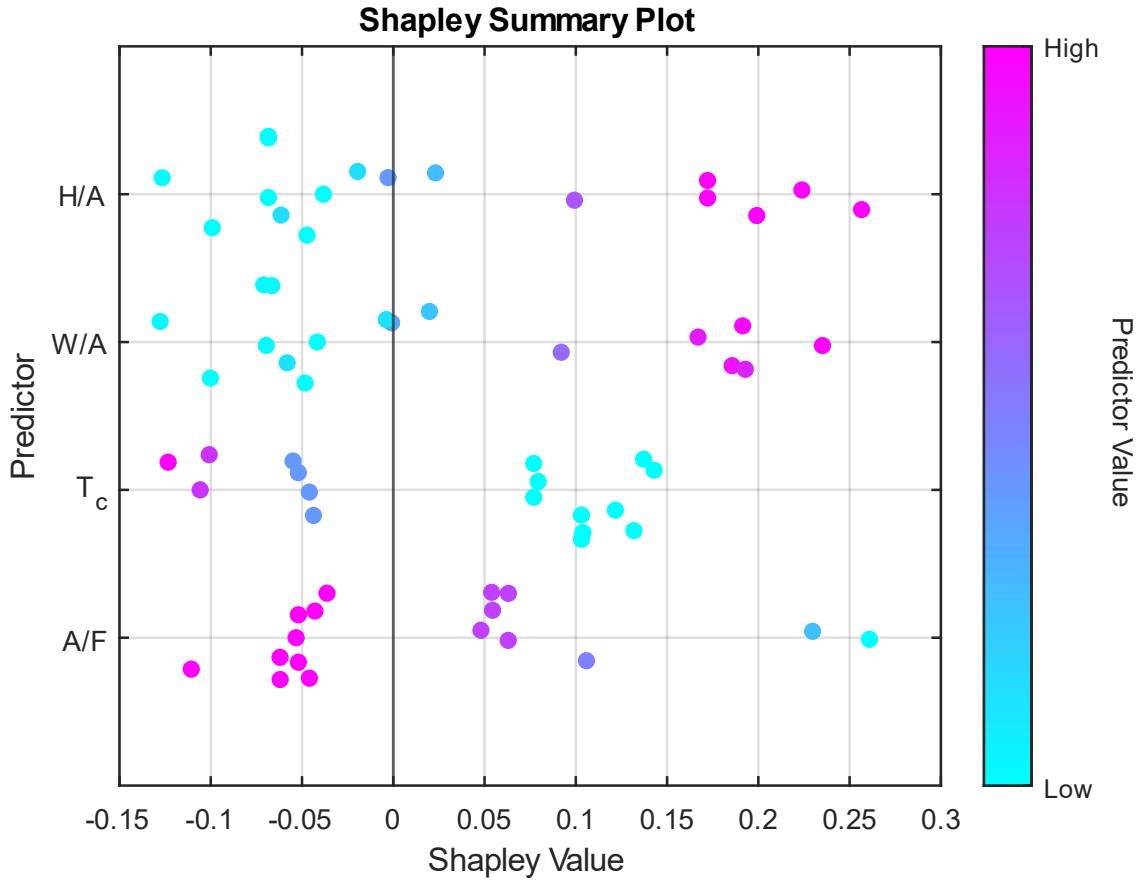


**Figure A8: Shapley values calculated with the best performing machine learning model for P-P isomer formation**

Figure A8 displays the calculated Shapley values for the formation of the P-P MDA isomers. During the formation of the P-P MDA isomer, the order of importance of the synthesis parameters are slightly different to that of the O-O isomer. However, the effects of the synthesis parameters are completely opposite in the two cases. For example, while a lower H/A ratio clearly increases the amount of O-O isomers, a higher H/A ratio favors the formation of P-P isomers.

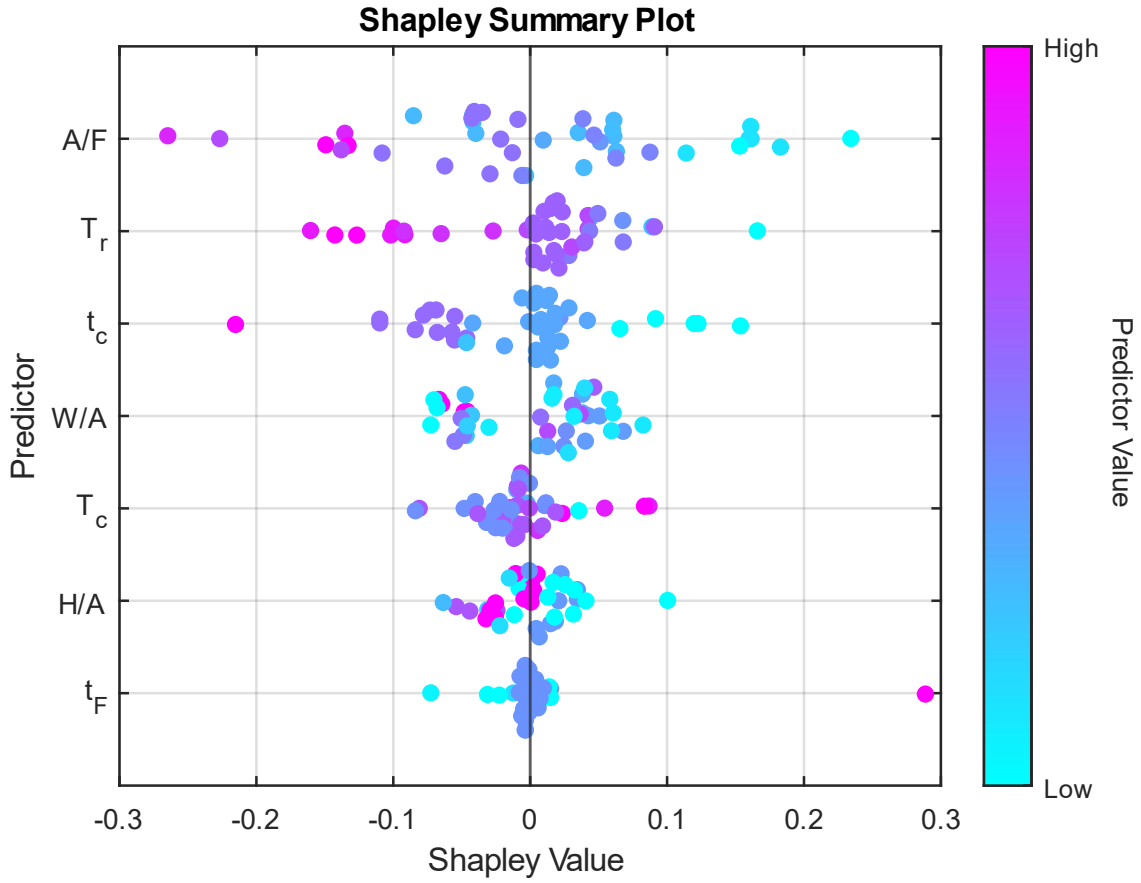


**Figure A9: Shapley values calculated with the best performing machine learning model for >4-ring formation**

Figure A9 displays the calculated Shapley values for the formation of >4-ring MDA molecules. By combining the MDA molecules with four or more rings and calculating the Shapley values, it can be stated that although there are no significant changes in the order of importance, magnitude, or the sign of the parameters, there are cases – such as the A/F ratio and the $T_r$ parameters – where the effects of the parameters can be clearly identified. In contrast, for less important parameters, such as the H/A ratio, the sign of the parameter's effect cannot be reliably estimated.