Development of an AI-Driven Methodological Framework for Enhancing Environmental Sustainability in Manufacturing

Zur Erlangung des akademischen Grades eines

DOKTORS DER INGENIEURWISSENSCHAFTEN (Dr.-Ing.)

von der KIT-Fakultät für Maschinenbau
des Karlsruher Instituts für Technologie (KIT)

angenommene

DISSERTATION

von

M.Sc. Lucas Greif

Tag der mündlichen Prüfung:	02.12.2025
Hauptreferent:	Prof. Dr. Dr.-Ing. Dr. h. c. Jivka Ovtcharova
Korreferent:	Prof. Dr.-Ing. Marcus Geimer

Kurzfassung

Der globale Fertigungssektor steht vor einer doppelten Herausforderung: die Produktivität durch Digitalisierung zu steigern und gleichzeitig seinen ökologischen Fußabdruck im Einklang mit internationalen Nachhaltigkeitszielen drastisch zu reduzieren. Obwohl Künstliche Intelligenz leistungsstarke Werkzeuge zur Prozessoptimierung bietet, wird ihre Anwendung in der realen Fertigungspraxis oft durch Datenknappheit und mangelnde Modelltransparenz erschwert, was das Vertrauen untergräbt und die Einführung behindert. Diese Dissertation entwickelt und validiert ein KI-gestütztes methodisches Rahmenwerk, das darauf ausgelegt ist, diese Herausforderungen zu überwinden, indem es die ökologische Nachhaltigkeit in der Fertigung systematisch verbessert. Das Rahmenwerk ist dateneffizient konzipiert, was den Einsatz in Umgebungen mit begrenzten und teuren experimentellen Daten ermöglicht. Ebenso integriert es Techniken der erklärbaren Künstlichen Intelligenz, um sicherzustellen, dass die Empfehlungen für Fachexperten transparent und umsetzbar sind. Die Wirksamkeit wird in einem zweistufigen Validierungsprozess mit drei Anwendungsfällen gezeigt. Stufe 1 prüft die Wahl der Optimierungsverfahren anhand offener Datensätze. Stufe 2 demonstriert das vollständige Rahmenwerk in zwei industriellen 3D-Druck-Fallstudien. In Fallstudie 1 zielte die Methodik auf die Verringerung des produktbezogenen CO_2-Fußabdrucks und erzielte eine Reduktion um 31,9 %, von 31,1 g CO_2e auf 21,2 g CO_2e pro Bauteil. In Fallstudie 2 stand der Energieverbrauch im Fokus und wurde um 70,0 % gesenkt, von 20,0 Wh auf 6,0 Wh. Zugleich reduzierte sich der produktbezogene CO_2-Fußabdruck um 26,7 %, von 20,6 g CO_2e auf 15,1 g CO_2e. Zusätzlich zeigten die Ergebnisse Zielkonflikte zwischen den ökologischen Zielsetzungen und der mechanischen Bauteilqualität.

Abstract

The worldwide manufacturing industry is confronted with a twofold challenge: enhancing productivity via digital transformation while significantly diminishing its environmental impact to comply with global sustainability objectives. Although Artificial Intelligence offers powerful tools for process optimization, its application in real-world manufacturing practice is often complicated by data scarcity and a lack of model transparency, which undermines trust and hinders adoption. This dissertation develops and validates a novel methodological framework, driven by Artificial Intelligence, that is designed to overcome these challenges by systematically improving environmental sustainability in manufacturing. The framework is designed to be both data-efficient, which allows for its use in environments with limited and expensive experimental results and understandable, by incorporating methods from Explainable Artificial Intelligence to guarantee that the suggestions are clear and actionable for domain experts. Its effectiveness and adaptability are demonstrated through a two-stage validation process comprising three application cases. In the first stage, the optimization strategy selection logic is empirically tested on diverse open-source datasets. In the second stage, the complete framework is applied in two industrial case studies in the field of 3D printing. In Use Case 1, the methodology targeted the product carbon footprint and achieved a 31.9 % reduction, from 31.14 g CO_2e to 21.19 g CO_2e per part. In Case 2, it targeted energy consumption and reduced energy usage by 70.0 %, from 20.0 Wh to 6.02 Wh, while also lowering the product carbon footprint by 26.7 %, from 20.6 g CO_2e to 15.1 g CO_2e. Across both cases, the results showed clear trade offs between ecological objectives and mechanical part quality.

Preface

First, I want to thank my primary reviewer and supervisor, Prof. Dr. Dr.-Ing. Dr. h.c. Jivka Ovtcharova. Her mentorship marked by visionary insight, trust, and encouragement challenged me to elevate both the scientific rigor and practical relevance of this work.

Second, I wish to thank Prof. Dr.-Ing. Marcus Geimer, my second reviewer, whose constructive criticism and steadfast support helped refine both the breadth and depth of my analyses.

Third, I am grateful to Prof. Dr.-Ing. Anne Meyer and Prof. Dr.-Ing. Arne Rönnau, now heads of the IMI department, for granting me the freedom and resources to complete this thesis and for challenging me to scrutinize every aspect of my work.

Fourth, I express my gratitude to each member of the IMI - from fellow researchers and administrative staff to master's and bachelor's students and student assistants - whose camaraderie and collaborative spirit drove this research forward in countless ways.

Fifth, I am grateful for the lifelong encouragement of my parents Werner and Ramona Greif, who supported me through every stage of my education and nurtured my curiosity for scientific inquiry from an early age.

Finally, to my girlfriend Jennifer Siegfried: thank you for your patience, understanding, and unwavering belief in me throughout this process. Your support has been my anchor.

Contents

Acronyms and symbols

Acronyms

AI	Artificial Intelligence
AITT	AI-assisted Technology Transfer
AM	Additive Manufacturing
ANOVA	Analysis of Variance
AUC	Area Under the Receiver Operating Characteristic Curve
AutoML	Automated Machine Learning
BEP	Break-Even Point
BIC	Bayesian Information Criterion
BMWK	German Federal Ministry for Economic Affairs and Climate Action
BO	Bayesian Optimization
C2C	Cradle-to-Cradle
CASH	Combined Algorithm Selection and Hyperparameter Tuning
CE	Circular Economy
CMA-ES	Covariance Matrix Adaptation Evolution Strategy
CNN	Convolutional Neural Network

CSRD	Corporate Sustainability Reporting Directive
CV	Cross-Validation
DED	Directed Energy Deposition
DLP	Digital Light Processing
DoE	Design of Experiments
DS	Direct Search
EBM	Explainable Boosting Machine
e-BEP	Environmental Break-Even Point
EFB	Exclusive Feature Bundling
EI	Expected Improvement
ELA	Exploratory Landscape Analysis
EM	Expectation Maximization
EMS	Energy Management Systems
FDC	Fitness Distance Correlation
FDM	Fused Deposition Modeling
FFD	Fractional Factorial Design
GA	Genetic Algorithm
GHG	Greenhouse Gas
GMM	Gaussian Mixture Model
GOSS	Gradient-Based One-Side Sampling
GP	Gaussian Process
GPU	Graphics Processing Unit

GR&R	Gage Repeatability and Reproducibility
HIF-Set	High-Impact Feature Set
ICE	Individual Conditional Expectation
IIoT	Industrial Internet of Things
IoT	Internet of Things
IQR	Interquartile Range
KDD	Knowledge Discovery in Databases
KPIs	Key Performance Indicators
LCA	Life Cycle Assessment
LHS	Latin Hypercube Sampling
LOCF	Last Observation Carried Forward
LOOCV	Leave-One-Out Cross-Validation
LS	Learning System
LSTM	Long-Short-Term Memory
MAD	Median Absolute Deviation
MAR	Missing at Random
MCAR	Missing Completely at Random
MES	Manufacturing Execution System
MICE	Multiple Imputation by Chained Equations
MNAR	Missing Not at Random
MTConnect	MTConnect
OEM	Original Equipment Manufacturer

OPCUA	OPC Unified Architecture
PCA	Principal Component Analysis
PCF	Product Carbon Footprint
PD	Partial Dependency
PDP	Partial Dependence Plot
PIC	Partial Information Content
POI	Probability of Improvement
PRISMA	Preferred Reporting Items for Systematic reviews and Meta-Analyses
PRNGs	Pseudo-Random Number Generators
PSO	Particle Swarm Optimization
RBF	Radial Basis Function
RL	Reinforcement Learning
RSM	Response Surface Methodology
SDG	Sustainable Development Goals
SDR	Sample-to-Dimension Ratio
SGD	Stochastic Gradient Descent
SHAP	SHapley Additive exPlanations
SLA	Stereolithography
SMAC	Sequential Model-Based Algorithm Configuration
TBL	Triple Bottom Line
TFDV	TensorFlow Data Validation

TPE	Tree-structured Parzen Estimator
TSR	Tail Symmetry Ratio
UCB	Upper Confidence Bound
VDI	German Association of Engineers
VRP	Vehicle Routing Problem
XAI	Explainable Artificial Intelligence
XRL	Explainable Reinforcement Learning

Latin symbols and variables

C	Cluster set
D	Dataset
d	Dimensionality
E	Energy consumption
$\mathbb{E}$	Expectation operator
EF	Emission factor
$f(\cdot)$	Objective function
$\hat{f}$	Surrogate of f
$h(\cdot)$	Base learner
$\mathcal{L}$	Loss function
m	Mass
N	Number of samples
$p(\cdot)$	Probability density

$P(\cdot)$	Event probability
PCF	Product Carbon Footprint
R^2	Coefficient of determination
S_j	First-order Sobol index for feature j
S_{T_j}	Total Sobol index for feature j
$\mathbf{X}$	Feature matrix $N \times d$
$\mathbf{x}$	Feature vector in domain Ω
$\mathbf{x}^*$	Optimal feature vector
$\mathbf{x}^+$	Best-so-far feature vector
x_j	j-th feature value
Y	Response variable
y	True target
$\hat{y}$	Predicted target
$\bar{y}$	Mean of true targets

Greek symbols and variables

$\alpha(\cdot)$	Acquisition function
β_j	Regression coefficient of feature j
γ_{ik}	Responsibility $p(k \mid i)$ in GMM
ϵ	Stochastic error term
η	Learning rate
θ	Parameter vector

κ	Exploration–exploitation trade-off parameter
λ	Transformation parameter
μ	Distribution mean
μ_k	Mean vector of component k
π_k	Mixture weight of component k
ρ	Sample-size–to-dimension ratio N/d
σ	Standard deviation
σ^2	Variance
σ_n^2	Noise variance
τ	Autocorrelation length of a fitness landscape
ϕ_j	Shapley value for feature j
$\Phi(\cdot)$	Standard normal CDF
Ω	Feasible search domain
$\mathbf{\Sigma}$	Covariance matrix
$\mathbf{\Sigma}_k$	Covariance matrix of component k

Operators and math symbols

$\arg\min$	Argument minimizing the following expression
$\mathcal{N}(\mu, \sigma^2)$	Normal distribution with mean μ and variance σ^2
$\sqrt{\cdot}$	Square root
$\sum$	Summation operator

General deep indexes

$(-i)$	Index indicating exclusion of sample i
i	Index for data points/samples
j	Index for features/dimensions/coefficients
k	Index for clusters/components/folds

1 Introduction

This chapter situates the thesis within its broader context. It first motivates the work by outlining why improving the environmental performance of manufacturing with the help of artificial intelligence (AI) matters. Then it characterizes the industrial landscape to delimit the problem space and derive the guiding objectives and research questions. The chapter concludes with a brief outline of the structure of the thesis.

1.1 Motivation

Manufacturing and production are foundational to the modern economy. At the same time, they are major contributors to environmental impacts. Collectively, these sectors are responsible for approximately one-fifth of the world's greenhouse gas (GHG) emissions and one-third of its total energy consumption [1]. This impact is highly concentrated in industrialized nations. In the European Union, manufacturing releases an estimated 880 million tons of carbon dioxid(e equivalents annually [2], while in the United States it represents 24% of direct GHG emissions in 2020 [3]. This immense environmental footprint, which includes not only GHG emissions but also water consumption, resource depletion, and waste generation, has placed the industry at the center of the global push for net zero manufacturing.

In response, a confluence of regulatory and societal forces is driving a fundamental shift towards greater environmental accountability. At the global level, the United Nations Sustainable Development Goals (SDGs) provide the

ethical foundation for international efforts and guidelines, with SDG 9 advancing industry, innovation and infrastructure and SDG 12 promoting responsible consumption and production [4]. The wedding cake model of these SDGs, illustrated in Figure 1.1, visually emphasizes the dependence of social and economic welfare on the maintenance of the biosphere, underscoring environmental protection as a foundational prerequisite for all other goals [5].

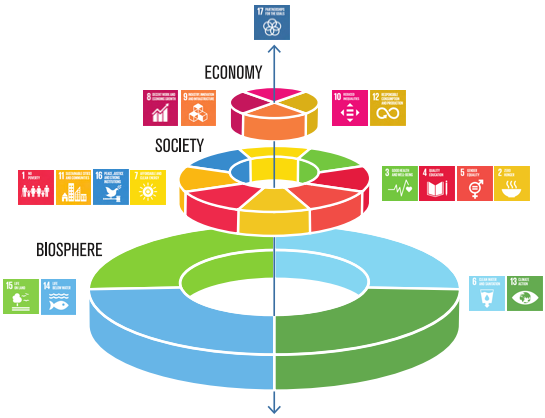


Figure 1.1: The wedding cake model for the SDGs, as presented by the Stockholm Resilience Center, taken from [5].

At the policy level, this strategy is operationalized through European Union-wide benchmarks. The Net Zero Industry Act sets an overall benchmark for the manufacturing capacity of the Union in net zero technologies to reach or exceed at least 40 % of the annual deployment needs by 2030 [6]. Clear terminology is essential for the transition. Emission reduction denotes the targeted reduction within existing processes, including efficiency improvements and leak control. Decarbonization refers to the substitution of high-carbon inputs and processes with low- or zero-carbon technologies. A net-zero strategy is broader because it integrates both approaches and addresses residual, hard-to-abate, emissions through capture or durable removal. Implementing this strategy requires a complete reconfiguration of the value chains [7]. Achieving

net zero requires translating high-level objectives into production paradigms that align technological capabilities with societal objectives. In parallel, the Industry 5.0 vision of the European Commission extends the Industry 4.0 focus on automation by explicitly centering human well-being, resilience, and environmental stewardship [8]. It calls for production systems that enable close collaboration between the workforce and emerging technologies, including AI, while actively mitigating environmental impacts throughout the value chain. In this context, intelligent production and operations provide the operational lever that links strategic ambitions to day-to-day execution. Through an integrated and synergistic suite of digital tools that raise transparency, strengthen traceability, and improve the efficient use of resources, these systems make the goals of Industry 5.0 actionable at scale [9].

Taken together, the industry 5.0 orientation and net zero ambitions translate into a multidimensional challenge for industrial enterprises, as shown in Figure 1.2. Externally, they must respond to stringent regulatory imperatives. Internally,

Figure 1.2: Key challenges for sustainable industrial manufacturing.

they face the costs, operational complexity and risks inherent in the adoption of innovative production techniques or the retrofit of existing systems. The challenge extends beyond the factory floor, requiring the decarbonization of entire industrial supply chains and the integration of the principles of circular economy [10, 11]. Addressing these dual pressures requires a systematic and

intelligent approach to optimize energy use, minimize material waste, and make manufacturing processes more sustainable [12].

1.2 Scope of the Thesis

Within the broader realm of industrial sustainability research, this dissertation focuses on environmental impacts arising specifically from production processes. By production processes, this work refers to the operational and managerial activities that transform raw materials into final products, including aspects such as material consumption and energy usage. A fundamental principle in sustainable engineering is that early stage product development exerts a disproportionate influence on the lifecycle impact of a product. As has been widely established in the literature, design decisions regarding material selection, geometric complexity, and functional requirements effectively "lock in" a large fraction of the total manufacturing costs and the analogous share of environmental burdens [13]. A truly holistic optimization would therefore need to begin at this conceptual stage.

However, this dissertation deliberately delimits its scope to the subsequent production phases. In most industrial contexts, there is a clear organizational separation between the design and production functions [14]. The manufacturing department typically receives a finalized product blueprint and must operate under constraints imposed by client directives or pre-existing technical limitations. Within this operational reality, the most effective and often only actionable levers to improve resource efficiency and ecological impact are the process parameters themselves [15]. This focus is further reinforced by methodological considerations. The optimization of product design is an inherently multi-objective problem, often involving complex, qualitative trade-offs [16]. In contrast, the optimization of a defined production process is a more tractable problem, focused on tuning a specific set of operational variables against a quantifiable objective. By focusing on the latter, this thesis maintains a clear

methodological focus, ensuring the development of a verifiable framework for data-driven improvement.

To operationalize this focused scope, the research is empirically grounded in parameter- and sensor-based data sets from additive manufacturing (AM). This approach deliberately concentrates on quantitative metrics that directly inform both cost and environmental impact, notably CO_2 emissions, material consumption, and energy usage. The selection of these metrics is twofold. First, they are readily measured or inferred from production logs, ensuring that research identifies actionable improvements with immediate practical utility. Second, this focus aligns with the operational realities of industrial collaborations, which prioritize concise outcomes within the manufacturing sphere where engineers have the most direct influence.

Although this work focuses on production, the connection to product design remains decisive. Insights from process optimization, such as identifying material- or energy-intensive steps, should inform subsequent design cycles. Although this feedback loop is not empirically traced in the present work, it represents a key strategic recommendation for organizations seeking a deeper long-term integration of sustainability objectives into their product development roadmaps [17].

1.3 Goal of the Thesis

The primary objective of this dissertation is to develop and validate a data-driven framework for process optimization that is grounded in the practical constraints of modern manufacturing while explicitly prioritizing environmental sustainability. To be effective, this framework must be comprehensive and adaptive, guiding the entire optimization lifecycle from the foundational stages of experiment planning and data acquisition to the final validation of an optimized solution.

Central to this objective is the explicit integration of environmental sustainability metrics. Methodologically, this is achieved by formally treating environmental metrics, such as energy consumption or CO_2 emissions, as the objective function within the optimization framework, rather than as secondary constraints. The research then aims to quantify the achievable reductions in ecological impact and, through the use of explainability analysis, to identify and rank the key process levers that govern this impact.

To ensure that this framework is not only theoretical but also designed for practical applicability, another objective is to design it for real-world industrial environments. To this end, the methodology must address two pervasive constraints: it must be highly sample efficient to minimize the need for costly physical experiments, and it must be robust to metric latency by handling both rapid, sequential data acquisition and slow, batch-based data collection. Building upon this principle of practical application, the final objective is to ensure that the framework is modular and generalizable, thus maximizing its potential impact. Guided by this objective, the thesis is framed by the following overarching research question:

How can data-efficient, explainable optimization be operationalized in industrial environments to advance sustainable manufacturing?

To address this question, the framework adopts a modular architecture in which the core components can be adapted to suit different manufacturing technologies. By designing for this adaptability, the ultimate aim is to bridge the gap between academic research and industrial practice by translating state-of-the-art AI techniques into a transparent and deployable methodology that delivers tangible real-world value.

1.4 Structure of the Thesis

The thesis is composed of six core chapters that build logically on each other to guide the reader from motivation through validation to the final outlook. The sections and their objectives are summarized in Figure 1.3.

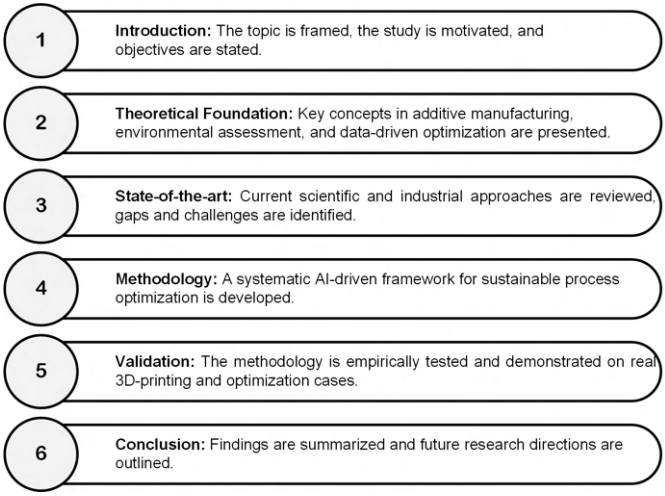


Figure 1.3: Overview of the structure of the thesis.

This chapter 1 introduces the dissertation by motivating the work from macroeconomic, regulatory, and technological perspectives. It defines the research problem, delineates the scope, and establishes the need for an AI-supported methodological framework aimed at ecological optimization of manufacturing processes. Chapter 2 lays the theoretical foundation in three interrelated domains: AM, environmental assessment, and data-driven process optimization techniques. Chapter 3 synthesizes the current state of research using a systematic literature review guided by the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guided literature review, evaluating contemporary approaches to sustainable, data-driven production and revealing the research gap that this dissertation addresses. Chapter 4 presents the

methodological framework developed to reduce the environmental impact of manufacturing processes. It elaborates the derivation of target variables and system boundaries, the experimental design and data collection strategies, the algorithmic pipeline from preprocessing to sequential and offline optimization, the embedding of explainable AI (XAI) for decision support, and the criteria that determine industrial readiness such as robustness and break-even analysis. Chapter 5 validates the framework in a two-stage process comprising three validation cases. Stage 1 empirically tests the selection logic of the optimization strategy on various open-source data sets. Stage 2 applies the complete methodology to AM in two industrial case studies: one aiming to minimize the carbon footprint of the product under data-rich conditions, and the other focusing on reducing energy consumption under severe data constraints. Chapter 6 summarizes the main contributions, reflects on limitations, and delineates avenues for future research.

2 Theoretical Foundations

This chapter establishes the theoretical foundations for the dissertation. It begins in Section 2.1 by detailing the principles of AM, the core production technology under investigation. Section 2.2 then introduces the frameworks for environmental assessment, providing the metrics and strategies needed to quantify and improve sustainability. Finally, Section 2.3 sets out the theoretical foundations of data-driven process optimization.

2.1 Additive Manufacturing

AM denotes a family of processes that build three-dimensional parts directly from digital models by adding material only where it is needed [18]. In contrast to subtractive methods that shape a component by removing stock from a solid blank, AM routes accumulate material in place, which reduces offcuts and scrap [19]. The technology originated in the 1980s as rapid prototyping for visual models, but modern improvements in precision, material range, and reliability have allowed AM to evolve into a viable option for end-use production in industries such as aerospace, medicine, and automotive [19, 20]. All AM processes share the principle of building objects layer by layer, but differ in how layers are formed and fused [21]. The typical workflow begins with an CAD model that is converted to a mesh and digitally sliced into layers [18]. For each slice, the machine deposits or solidifies material in the cross-sectional pattern of that layer, gradually forming the 3D part as the layers accumulate [18]. A detailed, process analysis was conducted in advance of this thesis. The study traces the 3D printing workflow from feedstock production through fabrication

and application to disposal and recycling, identifying relevant stakeholders, the scope of the product, and the corresponding environmental burdens, as shown in Figure A.1 of the appendix. The study revealed that material selection and process parameter decisions account for most of lifecycle environmental impacts, highlighting them as key leverage points for sustainable design (Hauck and Greif [2]).

In compliance with the ISO/ASTM 52900:2021(E) standard, which serves as the primary framework for terminology and categorization in the realm of AM, all existing commercial methodologies are classified into seven process categories. These categories include material extrusion, vat photopolymerization, powder bed fusion, binder jetting, material jetting, sheet lamination, and directed energy deposition [21]. In the process of material extrusion, commonly represented by fused deposition modeling (FDM), in which a thermoplastic filament undergoes heating before being extruded through a nozzle. This technique involves tracing each individual layer and ensuring that it adheres to the layers beneath it [18]. This process is visually depicted in Figure 2.1.

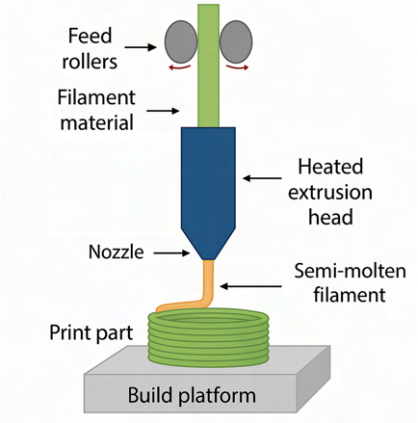


Figure 2.1: Simplified schematic of a FDM process, adapted from [22].

In vat photopolymerization, exemplified by stereolithography (SLA) and digital light processing (DLP), a laser beam or projected image selectively cures a thin layer of liquid photopolymer resin before the build platform indexes to the next layer [21]. Powder bed fusion spreads a thin layer of polymer or metal powder and then fuses the required cross section with a scanning laser or electron beam, repeating the sequence until the part is complete. Binder jetting also spreads powder in layers, but instead of melting it, an inkjet printhead deposits droplets of liquid binder that bind the particles together [21]. Material jetting relies on inkjet-style printheads that eject microscopic droplets of photopolymer or wax, which are immediately cured by ultraviolet light to build each layer [21]. In sheet lamination, sheets of paper, polymer film, or metal foil are successively bonded using adhesive, ultrasonic welding, or brazing and then cut to the desired outline before the next sheet is applied [21]. Finally, directed energy deposition feeds metal wire or powder into the focal zone of a laser, electron beam, or plasma arc, melting material as it is delivered so that the part is built or repaired bead-by-bead [18, 19, 21]. These additive manufacturing processes differ in energy input, feedstock form and attainable microstructure, producing different cost and performance envelopes [21].

As this work focuses on FDM, the choice of material is a primary determinant of printability, mechanical performance, and environmental impact. Table 2.1 summarizes the main advantages and limitations of the commonly used FDM thermoplastics.

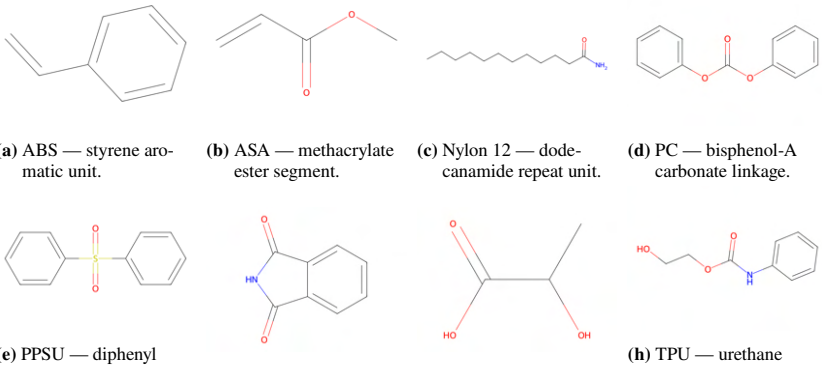
Differences in printability and performance are directly related to polymer chemistry and morphology, which is illustrated in Figure 2.2. Amorphous aromatics such as PC, PEI, and PPSU exhibit high glass transition temperatures and thermal stability, but demand elevated processing temperatures and enclosed printers [23]. Semi-crystalline nylons form hydrogen bonds and crystallize on cooling, delivering fatigue and wear resistance at the cost of moisture sensitivity and shrinkage/warpage [24]. ABS and ASA share a styrenic matrix. ASA replaces butadiene with acrylic rubber, improving UV durability compared to ABS [25]. The aliphatic polyester structure of PLA and its low glass transition

temperatures make it easy to print with a good surface finish, but limit heat and chemical resistance through softening and hydrolysis [26].

Table 2.1: Advantages and limitations of FDM materials based on [27, 28].

Material	Advantages	Limitations
Acrylonitrile–butadiene–styrene (ABS)	Tough; good impact strength; machinable; moderate cost	Warping; odors; requires heated bed/enclosure; poor UV resistance
Acrylonitrile styrene acrylate (ASA)	UV and weather resistant; good mechanical strength; color stability outdoors	Higher cost than ABS; shrinkage/warping risk; odor during printing
Polyamide 12 (Nylon 12)	High fatigue and impact resistance; good chemical resistance; wear resistant	Hygroscopic; warping; higher print temperatures; dimensional drift if not dried
Polycarbonate (PC)	Very high tensile/flexural strength; heat resistant; durable	Very high print temperature; strong warping; enclosure required; hygroscopic
Polyphenylsulfone (PPSU/PPSF)	Excellent chemical and heat resistance; sterilizable; high mechanical strength	Very high processing temperature; expensive; limited printer compatibility
Polyetherimide (PEI, ULTEM)	High strength-to-weight; thermal/chemical stability; flame retardant; biocompatible grades available	Requires high-temperature printers; costly material; limited color options
Polylactic acid (PLA)	Easy to print; low warping; good surface quality; low odor; bio-based feedstock	Brittle; low heat deflection temperature; limited UV/chemical resistance
Thermoplastic polyurethane (TPU)	Highly flexible; excellent abrasion and tear resistance; good impact performance; oil/chemical resistant	Slow printing; stringing; needs tuned extrusion path; reduced dimensional precision

The segmented block architecture of TPU produces high elasticity and resistance to abrasion, but slows the printing and can reduce the precision of the dimensions [29].



(a) ABS — styrene aromatic unit.

(b) ASA — methacrylate ester segment.

(c) Nylon 12 — dodecanamide repeat unit.

(d) PC — bisphenol-A carbonate linkage.

(e) PPSU — diphenyl sulfone linkage (simplified).

(f) PEI, ULTEM — aromatic imide motif.

(g) PLA — lactic acid ester unit.

(h) TPU — urethane carbamate functional group.

Figure 2.2: Simplified representative chemical repeat units or motifs of common FDM thermoplastics oriented on [30]. Structures are schematic and illustrate the dominant chemical functionality influencing material properties.

Concurrently, research is advancing AM to include mineral-based pastes such as clay and concrete [31]. In the realm of 3D printing, clay and concrete systems operate as extrusion processes. However, the underlying material dynamics and mechanisms diverge from those in polymer-based FDM. Rather than introducing a thermoplastic filament into a hot end where it is melted, these systems pump a moist, thixotropic mixture, such as hydrated clay or cement paste, from a reservoir and extrude it through a nozzle with a wider hole [32, 33]. Conventional extruder setups in this context typically employ pneumatic pistons, screw pumps, or auger feeders to propel the paste, in contrast to the gear driven motors used to feed the filament in FDM [33, 34]. As highlighted in recent research, clay printers are designed to compress the clay paste and disperse it in successive layers, notably without employing a heated nozzle [32, 35]. Given that the paste is delivered from a reservoir, often situated at a distance from the nozzle, managing the flow becomes more challenging. Pump-based feeders are subject to latency and hysteresis, necessitating the use of local buffer capacities or pressure feedback systems to ensure a consistent volumetric flow

rate [34, 36]. Furthermore, the heavy weight of the extruder assembly and the extended material feed lines introduce inertia and compliance, requiring motion paths with gentle acceleration and deceleration to prevent pressure fluctuations or nozzle wobbling [36, 34]. Ultimately, the extruded clay or concrete retains its plasticity until it undergoes curing, hardening, either through hydration processes or drying, instead of cooling. Therefore, print mixtures are specifically formulated to have very high viscosity and produce stress for self-supporting capacity, making their buildability dependent on rheological properties such as slump flow and open time [37, 32, 35].

AM has evolved into a multifaceted instrument across diverse industries. Within the aerospace and automotive domains, AM facilitates the fabrication of components that are not only lightweight but also incorporate intricate internal lattice structures or consolidated assemblies of multiple parts. This advancement results in significant weight reduction and enhanced performance metrics [38]. For example, complex fuel nozzles and bracketry have been 3D printed in titanium or nickel alloys, reducing the number of assembly joints and overall mass while maintaining strength. In the medical field, AM is used to fabricate patient-specific implants, such as orthopedic implants and dental restorations and prosthetics with custom geometry that would be impractical to mold or machine [19]. AM also accelerates product development through rapid prototyping, as engineers can quickly iterate design concepts by printing physical models in hours or days, compared to weeks for tooling fabrication in traditional processes [19]. Furthermore, 3D printing has enabled on-demand manufacturing and spare parts production in remote or distributed locations, for example, printing replacement parts on-site in space or in military operations, thereby reducing the need for large inventories of parts [20].

These applications leverage several key benefits inherent to AM. A primary advantage is its extensive design freedom, which allows for the realization of intricate geometries such as internal channels, lattice structures, and organic shapes to optimize performance or functionality [19]. It also enables cost-effective customization, as individual parts can be modified via their digital models without financial or time penalties associated with tooling. This tool-less

nature inherently removes the significant lead times and expenses of processes such as injection molding, rendering low-volume production and rapid design iteration economically viable [39].

Despite its advantages, 3D printing also faces several limitations that currently restrict its broader use in manufacturing. One primary challenge is production speed and throughput. Layer-by-layer fabrication is relatively slow, often yielding only a few cubic inches of material per hour, depending on the process and part size, which is much slower than molding or machining in large quantities [20]. This makes AM less competitive for mass production of simple parts. Another limitation is unit cost at scale because each 3D printer typically produces one part or a batch of parts in its build volume over a given cycle time, scaling up production means adding more machines rather than speeding up a single machine. As a result, the cost per part remains roughly constant and does not decrease substantially with volume, in contrast to other processes, such as injection molding, where high initial tool costs are amortized and per-part costs drop dramatically with volume [39]. Material selection is also more limited in AM. Although the range of printable materials has expanded, including various thermoplastics, photopolymers, metals, and ceramics, these materials often come at a premium price and some have inferior properties compared to their conventionally processed counterparts, for example, fewer fiber-reinforced plastics are available for AM, and some printed metals may have higher porosity or residual stress [19]. Additionally, the layer-wise surface finish of as-printed parts is usually rougher. Achieving a smooth surface or tight dimensional tolerances may require post-process machining or polishing [19]. There are also size constraints, because each printer has a finite build volume, therefore very large parts might need to be printed in sections and joined [39]. Lastly, quality consistency and certification remain concerns as parts intended for critical applications must meet rigorous standards, and ensuring each printed part is defect-free can be challenging without extensive testing or process validation [20]. The key aspects of AM are summarized in Table 2.2.

Table 2.2: Summary of applications, benefits, limitations of AM.

Aspect	Key Points	Examples/Notes
Applications	Aerospace and automotive lightweight structures; patient-specific implants and prosthetics; rapid prototyping; on-demand spare parts in remote or distributed locations	Fuel nozzles, titanium brackets, dental restorations; spare parts for space missions and military use
Benefits	Design freedom for complex geometries; customization without extra cost; tool-less manufacturing; feasible low-volume production; rapid iteration	Customized orthopedic implants, tailored consumer goods, fast design testing
Limitations	Slow production speed and low throughput; high per-unit cost at scale; limited and expensive materials; rough surface finish, need for post-processing; build volume restrictions; challenges in ensuring quality consistency and certification	Build rates of only a few cubic inches/hour; fewer fiber-reinforced materials; porosity issues in printed metals; certification complexity in aerospace/biomedical sectors

2.2 Environmental Assessment and Reduction Strategies in Manufacturing

This thesis proposes a methodology to reduce the environmental footprint of a product. To establish the proper context, this chapter will first outline the fundamentals of environmental assessment and discuss general impact reduction potentials within manufacturing processes. Following this overview, it will specify which of these aspects the present research will address. Over the past decades, a variety of frameworks have been developed to provide a systematic and scientific basis for environmental assessment [40]. These methodologies vary in scope and focus, ranging from holistic, multi-impact evaluations to

detailed tracking of specific material or energy flows [40]. This section provides a detailed exposition of the two most prominent frameworks: Life Cycle Assessment (LCA) and Carbon Footprint Analysis.

LCA is acknowledged as the most thorough and methodologically strict framework for assessing the prospective environmental impacts related to a product, process, or service [41]. This approach provides a structured and detailed analysis to quantify and evaluate the environmental implications systematically [42]. A feature of its system is its comprehensive viewpoint that includes the entire life cycle. This begins with the extraction of raw materials and the procurement of energy, continuing through the phases of production and usage [43]. It also extends to end-of-life management, which includes processes such as disposal and recycling [41]. This systematic approach prevents the problem of burden shifting, where a solution in one part of the life cycle, such as the manufacturing plant, inadvertently creates a larger environmental problem elsewhere, for example, in the supply chain or during the disposal of the product [44]. The methodology of conducting an LCA is a framework standardized by the International Organization for Standardization (ISO) in the ISO 14000 series. The core standards governing its practice are ISO 14040:2006 [45], which outlines the Principles and Framework, and ISO 14044:2006 [43], which specifies the Requirements and Guidelines. The process of conducting an LCA is guided by four phases as illustrated in Figure 2.3.

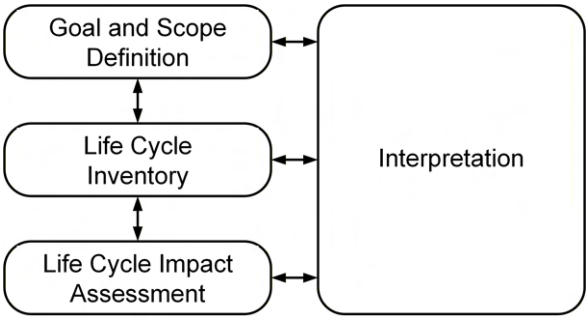


Figure 2.3: The four phases of an LCA, based on [45].

The initial phase, Goal and Scope Definition, establishes the framework and objectives of the study to ensure alignment with its intended purpose [46]. This phase delineates the boundaries and methodological choices that guide the entire assessment [45]. Critical determinations consist of specifying the functional unit, which is a quantified metric that represents the function of the product system and acts as a basis for comparing all associated inputs and outputs [43]. In addition, it involves delineating the system boundaries to identify the particular stages and processes of the life cycle to be incorporated into the analysis [43].

Subsequent to establishing the goal and scope, the Life Cycle Inventory (LCI) phase involves gathering comprehensive data on all environmental inputs and outputs associated with the product or service being studied [45]. This inventory covers all stages of the life cycle, beginning with the extraction and manufacture of raw materials and continuing through the use and eventual disposal phases [47]. To ensure consistency and transparency, the data collection process must adhere to the established boundaries of the system. The LCI is often the most labor- and data-intensive stage of an LCA, with a practical challenge being the availability and quality of data, particularly for complex global supply chains [48].

In the following Life Cycle Impact Assessment (LCIA) phase, the data gathered from the LCI are transformed into possible environmental impacts [45]. This transformation is carried out by organizing the inventory results into different impact categories and calculating them using established indicators [43]. For example, Global Warming Potential (GWP) is a category that assesses the potential contribution of a GHG to global warming over a particular time frame, which is generally considered to be 100 years [49]. The corresponding indicator is expressed in carbon dioxide equivalents (CO_2e), which normalizes the impact of various GHGs relative to CO_2. ISO 14044 outlines optional processes such as normalization, grouping, and weighting, which facilitate the aggregation of multiple categories. Depending on the method selected, this aggregation is capable of producing endpoint results, such as effects on human health, ecosystems, resources, or even a single overall score [43].

The final phase, Interpretation, involves a systematic analysis of the LCI and LCIA results to formulate conclusions, provide recommendations, and communicate the findings to stakeholders [45]. This phase is critical to support informed decision making, identify opportunities to improve environmental performance, and highlight areas for future research [50]. The LCA process is inherently iterative. Insights gained in later phases, particularly during interpretation, often require revisions of the initial goal and scope, creating a feedback loop that refines the accuracy and relevance of the assessment [43]. Therefore, interpretation is not merely the final step, but a continuous process that occurs throughout the study to ensure ongoing alignment and validation [51].

Among the most critical of these choices is the delineation of the system boundaries, which shapes the study's outcome and applicability. The main types of boundary are illustrated in Figure 2.4. The boundary that typifies the most

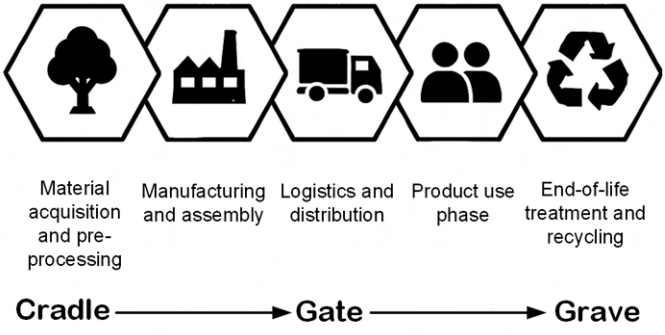


Figure 2.4: Visual representation of LCA system boundaries.

exhaustive linear methodology is referred to as the cradle-to-grave framework. This framework includes every phase of the life cycle of a product, from the initial extraction of raw materials to the final disposal or recycling process of the product [52]. This holistic assessment is essential for robust product comparisons, particularly when use phase or end-of-life impacts differ significantly. Conversely, the cradle-to-gate boundary provides a narrower scope of analysis, encompassing all steps in the production lifecycle, beginning with the extraction

of raw materials and concluding at the stage where the manufactured product exits the manufacturer's facility. This approach is used to quantify a product's embodied environmental impact before it is transported to the consumer and is frequently applied in business-to-business (B2B) contexts, where for example, component data are supplied to an original equipment manufacturer (OEM) for integration into a larger product LCA. The most restrictive boundary is gate-to-gate, which isolates a single facility or process within the value chain. Although this narrow perspective is useful for targeted process optimization and internal decision-making, it carries a high risk of overlooking upstream or downstream impacts, potentially leading to problem-shifting and suboptimal environmental outcomes from a whole-system perspective [53]. Evolving from these linear models, the Cradle-to-Cradle (C2C) concept redefines the end-of-life stage. Instead of a final grave, this framework envisions a closed-loop system where post-use materials become the cradle for new products [54]. This approach distinguishes between biological nutrients that can safely return to the biosphere and technical nutrients that are designed to be continuously re-utilized within industrial cycles without loss of quality [54]. Although these represent the most common frameworks, other specialized system boundaries exist, which can be tailored to the specific questions and contexts of a given LCA.

Although the frameworks described represent common approaches, the LCA methodologies can be adapted for more focused inquiries. A prominent example of such a specialized assessment is the Carbon Footprint Analysis, which concentrates exclusively on a single environmental impact category, climate change [55]. This analysis quantifies the total amount of GHG emissions caused directly and indirectly by an entity and expresses the final result in the standardized metric of CO_2e. Methodologies for calculating carbon footprints are governed by widely recognized standards, most notably the GHG Protocol [56] and ISO 14067 [57]. In corporate carbon accounting, as defined by the GHG Protocol, a key principle involves categorizing emissions into three separate scopes. This categorization is crucial to structuring the emissions inventory and pinpointing specific mechanisms to achieve reductions. The definitions follow the Corporate Accounting and Reporting Standard of the GHG Protocol [56]:

- Scope 1: This category encompasses GHG emissions originating directly from sources owned or under the control of the organization preparing the report. For instance, within a manufacturing entity, such emissions generally arise from the combustion of fuel in industrial furnaces and boilers, as well as from vehicles owned by the company. Additionally, this scope includes process emissions, which are chemical by-products generated during the manufacturing process.
- Scope 2: Indirect emissions from purchased energy arising from the generation of electricity, steam, heat, or cooling consumed by the company. The releases occur at the supplier's facilities but are attributable to the organization's energy use.
- Scope 3: This category broadly encompasses all indirect emissions that arise throughout the value chain of a company, in addition to direct emissions accounted for in other scopes. Often, Scope 3 emissions represent the most substantial and complex set of emissions to quantify, as they entail gathering data beyond the sphere of the organization's direct influence. For manufacturers, Scope 3 encompass emissions associated with the extraction and production processes of acquired materials, i.e., upstream activities, alongside those from transportation and distribution, business travel, and both the utilization and end-of-life management of products sold by the company, i.e. downstream activities.

The difficulty of obtaining high-quality data, especially for activities in Scope 3, is a challenge that often requires extensive supplier engagement and the use of industry average data as a proxy [58].

Table 2.3 provides a comparison that highlights the primary distinctions between performing an LCA and performing a targeted carbon footprint analysis.

Table 2.3: Comparison of LCA and Carbon Footprint Analysis.

Feature	**LCA**	**Carbon Footprint Analysis**
Primary Goal	To evaluate a broad range of potential environmental impacts across a product's entire life cycle	To quantify the total GHG emissions associated with a product, process, or organization
Governing Standard(s)	ISO 14040, ISO 14044.	GHG Protocol, ISO 14067
Typical Scope / Boundary	Cradle-to-Grave, Gate-to-Gate, Cradle-to-Gate.	Organizational Scopes 1, 2, 3
Key Output	Multi-indicator environmental profile.	A single metric expressed in CO_2e
Typical Application in Manufacturing	Comparing alternative product designs or materials; identifying environmental hotspots in a value chain	Corporate sustainability reporting; product labeling; identifying primary sources of GHG emissions for reduction efforts

Following the quantification of environmental impacts, the subsequent step is their systematic reduction. The manufacturing sector is currently experiencing a paradigm shift, moving beyond reactive end-of-pipe pollution control to proactive strategies that embed sustainability into the core of business operations [59]. This section explores the dominant high-level paradigms that reshape industrial activity and details the specific operational strategies that translate these visions into tangible improvements in environmental performance. These approaches range from the reconceptualization of material flows throughout the economy, as seen in CE [60], to the process-level integration of best practices in green and sustainable manufacturing, all of which are supported by targeted interventions in high-impact areas such as energy consumption and waste generation [61].

The CE represents an ambitious systemic shift away from the traditional linear economic model, which is predicated on the perpetual extraction of finite resources, their manufacture into products, and their eventual disposal as waste

[60]. CE is conceived as a regenerative system with the objectives of decoupling economic activity from the consumption of finite resources and eliminating waste entirely [60]. The foundation of CE is based on three principles as described by [62]. The primary goal is to design processes that prevent waste and pollution from arising in the first place. The next goal is to preserve the maximum functional value of products, components, and materials over the longest feasible period. Third, it is important to focus on the regeneration and restoration of natural systems.

The operationalization of these principles is often described through a hierarchy of "R" strategies. This framework extends beyond the conventional "Reduce, Reuse, Recycle" framework to encompass a more comprehensive set of actions, often presented in models such as the 10R framework [62], as illustrated in Figure 2.5.

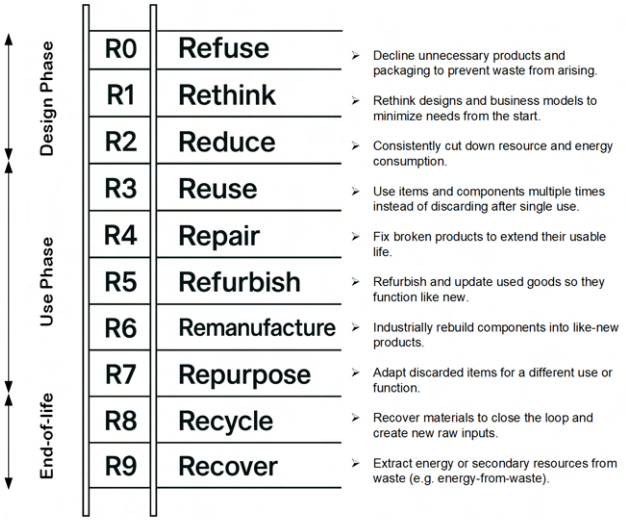


Figure 2.5: The 10R hierarchy for resource management strategies.

A key concept within this hierarchy is the prioritization of inner loops over outer loops [62]. For example, maintaining a product through repair is preferable to remanufacturing it, which is in turn preferable to recycling its constituent materials, as each successive outward loop tends to lose more embedded value and requires greater energy input. In a manufacturing context, the implementation of a circular model requires profound changes across the value chain. This transformation begins at the product design stage, emphasizing modularity, durability, and design-for-disassembly to facilitate easy repair and component harvesting at the end of life. It also requires the development of innovative business models, such as shifting from selling products to providing them as services, implementing product leasing agreements, or establishing comprehensive take-back schemes [63]. Operationally, CE requires the establishment of robust reverse logistics systems to collect used products and the creation of closed-loop production systems where end-of-life products can serve as high-quality feedstock for new manufacturing [64]. However, the transition is not without obstacles. Key challenges include the substantial investment needed for new technologies and infrastructure, the organizational inertia of established linear business models, the logistical complexity of managing reverse supply chains, and, in some cases, a lack of supportive policies and regulation [63].

Although CE offers an overarching macro-level perspective, sustainable manufacturing emphasizes the incorporation of sustainability principles in a more granular way [61]. Specifically, it focuses on embedding these principles directly at the levels of processes, products, and systems within a single organizational entity. This view aligns with the Triple Bottom Line framework [65], which holds that sustainability depends on a balanced advancement of economic viability, environmental stewardship, and social equity. Achieving this requires integrating these three domains so that progress in one does not undermine the others. Within sustainable manufacturing, eco-design serves as a primary approach that integrates environmental objectives from the beginning of product development, thus increasing durability, reducing energy demand, and facilitating recycling [66]. Currently, resource optimization focuses on

minimizing inputs, such as raw materials, water, and energy, per unit of production, which directly reduces costs and reduces the environmental footprint. Furthermore, it is crucial to acknowledge that the influence of a manufacturer transcends the confines of its individual operations [61]. Consequently, the principle of sustainability within the supply chain includes participating in cooperative efforts with partners. These collaborations are essential to ensure that these associates adhere to the highest environmental and social benchmarks, thus ensuring a comprehensive approach to sustainability that pervades the supply chain network [61]. Despite its clear benefits, the transition to sustainable manufacturing is impeded by practical barriers, most notably the prohibitive initial capital investment for new technologies, the technical complexity of retrofitting legacy factories, and the market uncertainty created by inconsistent regulatory landscapes [67].

The overarching concepts of CE and sustainable manufacturing are constructed on the basis of well-defined, deliberate interventions. Typically, within manufacturing operations, these strategic measures focus primarily on addressing the two primary culprits of environmental degradation: the consumption of energy and the production of waste. The industrial sector is the largest consumer of energy and is one of the leading contributors to global GHG emissions [68]. Therefore, optimizing energy efficiency represents one of the most effective levers for environmental improvement. A comprehensive approach to this begins with foundational energy audits and formal management systems, such as ISO 50001 [69], which systematically assess equipment and processes to establish performance baselines. From this understanding, savings can be achieved through process and parameter optimization, where, for example, adjusting the toolpaths of the machine in the machining can reduce energy use by up to 50 % without major capital investment [70]. These strategies are increasingly amplified by advanced monitoring and control systems enabled by Industry 4.0. The deployment of IoT sensors allows granular real-time energy monitoring, while AI algorithms can leverage these data to optimize production schedules and dynamically adjust process parameters to maintain peak efficiency [70]. Waste in manufacturing is the physical manifestation of inefficiency, representing both

a direct economic loss of materials and an additional cost for disposal. Waste minimization, therefore, is a cornerstone of both economic and environmental performance. A foundational approach is the application of Lean Manufacturing principles [71]. Although primarily focused on eliminating economic waste, the systematic reduction of defects, overproduction, and unnecessary processing inherently curtails the consumption of materials and energy. This philosophy aligns perfectly with the creation of closed-loop systems, a core tenet of CE, wherein processes are redesigned to turn waste streams such as scrap metal or wastewater into valuable internal inputs. A more proactive strategy involves material substitution, which seeks to replace virgin, hazardous, or non-recyclable materials with recycled, renewable, or benign alternatives at the source [72]. These physical strategies are improved by improved process control facilitated by digital technologies. Automated systems can manage material inputs with greater precision to reduce scrap, while real-time monitoring can detect process deviations before they result in defective products, thus improving first-pass yield and overall resource productivity. The preceding discussion has outlined a comprehensive framework that encompasses high-level paradigms such as the CE and sustainable manufacturing, as well as the targeted operational interventions that underpin them. These concepts, from systemic approaches to specific process-level tactics, are deeply interconnected. To visually synthesize the hierarchical nature and relationships between these diverse strategies, an integrated conceptual model is presented in Figure 2.6.

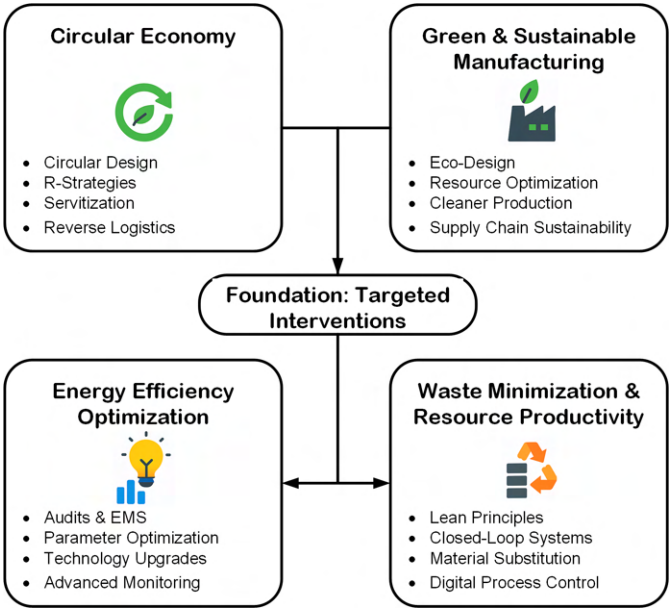


Figure 2.6: Conceptual framework for environmental impact reduction strategies in manufacturing.

2.3 Data-Driven Manufacturing Process Optimization

This section establishes the theoretical foundations of data-driven process optimization in manufacturing. It is organized into four sequential stages that mirror the logical dependencies of the method. First, the methodological requirements define assumptions, objectives, and constraints. Second, data acquisition and preprocessing lay the foundations for building a representative and quality-assured dataset from heterogeneous sensor streams, production logs, and experiments. Third, optimization paradigms introduce the fundamental concepts of mathematical optimization and machine learning used to identify operating regimes that meet the stated objectives under the defined constraints. Fourth,

xai and decision support set out the principles for interpreting model outputs so that recommended actions remain transparent, verifiable, and implementable.

2.3.1 Requirements on Methodologies

A comparative analysis of structured continuous improvement and quality management frameworks for the optimization of industrial processes showed that effective methodologies are characterized by a clearly articulated phase structure with explicit objectives for each phase, supported by systematic procedures and decision gates [73]. In addition, such structure has been shown to facilitate disciplined execution and verifiable progress throughout the optimization lifecycle of the process [73]. Established process modeling frameworks confirm the importance of structuring data-driven projects into well-defined phases. For instance, the Knowledge Discovery in Databases (KDD) methodology organizes the workflow into distinct phases: Problem Specification, Data Understanding, Data Preprocessing, Modeling, Evaluation, and Deployment [74]. Adopting such a phased methodology ensures a systematic and logical progression, where decisions made in earlier phases directly inform later modeling choices. Furthermore, this approach is inherently iterative, allowing for adjustments as objectives or the understanding of the data evolve. The same line of research indicated that high-quality methodologies must embed mechanisms for ongoing incremental progress, thereby institutionalizing continuous improvement rather than treating optimization as a one-off activity [73]. Adaptability was emphasized because methods should be extensible and tool-agnostic so that new analytical techniques and domain-specific instruments can be incorporated as manufacturing technologies evolve [73]. In addition, methodological quality was shown to be enhanced when complementary tools are synthesized to address multiple aspects of improvement in a coordinated manner. However, a potential pitfall was identified, when the overall approach becomes too bulky or complex, the implementation effort and the lead time can escalate, reducing practicality [75]. Building on these foundations, subsequent studies argued that, beyond the traditional criteria of being structured, systematic, and repeatable,

modern methodologies should be technology ready and data-centric [76]. It was shown that seamless integration with manufacturing execution systems, IoT platforms, and broader Industry 4.0 infrastructures is required so that real-time data streams, traceability, and feedback loops can be used for both monitoring and control [76].

2.3.2 Data Acquisition and Preprocessing

The first step is data acquisition, as data is the basis for any optimization or modeling effort [77]. As illustrated in Figure 2.7, manufacturing data can be acquired from a broad spectrum of sources, including sensor and machine monitoring, process execution and production planning records, quality assurance and inspection results, material and supply chain documentation, human-machine interaction logs and outputs derived from experimental studies, simulation models, and predictive analytics frameworks.

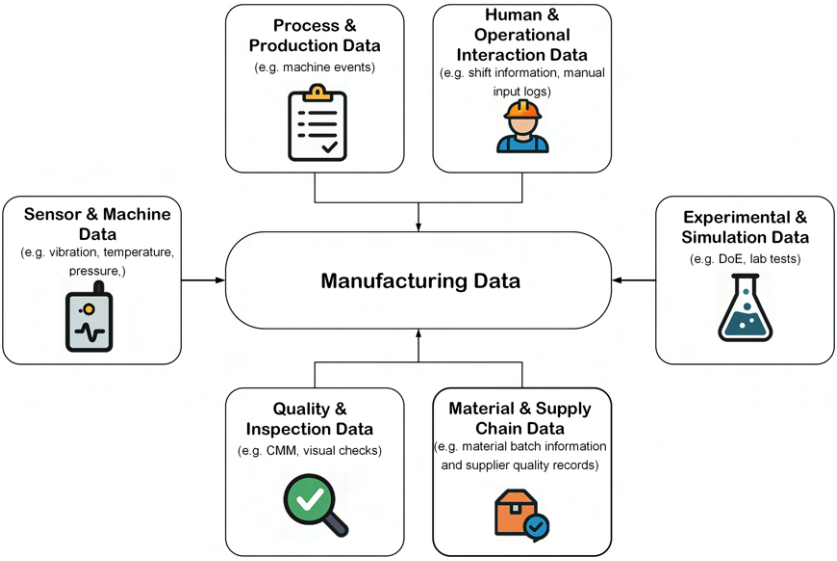


Figure 2.7: Overview of manufacturing data types and sources.

Data acquisition can follow two main pathways: using existing historical data or generating new data through experiments. From an industrial perspective, using historical data is more desirable, as the acquisition of new data always incurs costs [78]. Although historical data require comprehensive extraction and cleaning processes, they are generally much less expensive than gathering new data [78]. However, when generating new data is necessary, a Design of Experiments (DoE) allows maximizing the information gain while minimizing the experimental cost and effort [79]. A fundamental concept in DoE, especially in the context of Response Surface Methodology (RSM), is to estimate the relationship between process factors and a response variable by employing an adaptable empirical model [80]. A polynomial is widely used for this purpose, which represents the response variable Y as a function of the factors $x_1, x_2, \ldots, x_k$. A general form for this model is:

$$Y = \beta_0 + \sum_{i=1}^{d} \beta_i x_i + \sum_{i<j}^{d} \beta_{ij} x_i x_j + \sum_{i=1}^{d} \beta_{ii} x_i^2 + \cdots + \epsilon \tag{2.1}$$

Where:

- Y is the measured response variable.
- x_i is the level of the i-th factor.
- d is the total number of factors studied.
- β_0 is the overall mean or intercept.
- β_i are the coefficients for the main effects, which describe the direct and linear influence of a single factor.
- β_{ij} are the coefficients for the interactions of two factors that describe how the effect of one factor depends on the level of another.
- β_{ii} are the coefficients for the quadratic effects, which model the curvature in the response.

- ϵ is the random unexplained error.

The goal of DoE is to strategically select experimental runs that allow the model's β coefficients to be estimated with high precision and efficiency. The complexity of the design is a function of the number of runs (N) relative to the number of factors (d) and their levels (L). A summary of common DoE strategies to achieve this is provided in Table 2.4.

Table 2.4: Comparison of experimental design methods.

Method	**Principle & Complexity**	**Advantages**	**Disadvantages**	**Ref.**
Full Factorial Design	Evaluates all level combinations across d dimensions. Complexity: $N = L^d$	Enables estimation of all main effects and all orders of interactions	Prohibitive run count as d increases; high cost and times	[81]
Fractional Factorial Design	Uses a selected subset $(1/2^p)$ of the full design. Complexity: $N = L^{d-p}$	Drastically fewer runs; maintains orthogonal estimation of main effects	Confounding: some effects are aliased	[82]
Plackett–Burman Design	Highly fractionated 2-level design, often $N \approx 4(d+1)$	Extremely efficient screening of many factors with few runs	No resolution of interactions	[83]
Taguchi Method	Predefined orthogonal arrays; focuses on robustness using signal-to-noise ratios	Fewer runs; simple tables; focuses on robustness	Ignores most interactions; potential for bias from non-randomized arrays	[84]
Central Composite Design	2-level factorial + center + axial points for quadratic models. $N = 2^d + 2d + n_c$	Fits a full quadratic model; rotatable variance; moderate extra runs	Requires 5 levels per factor; axial points can be impractical	[81]
Box–Behnken Design	3-level quadratic design without corner points; often more efficient than CCD for $d < 5$.	Fewer runs than CCD for quadratics; good rotatability	No exploration of the corner points; limited to 7 factors	[85, 86]
Latin Hypercube Sampling	Stratified random sampling with a flexible number of runs N.	Good marginal coverage; lower variance than random sampling	No guaranteed uniformity in multi-dimensional subspaces; stochastic	[87]
Sobol Sequence	Quasi-random, low-discrepancy sequence for uniform coverage, often with $N = 2^d$	Very uniform space coverage; reproducible; fast integration convergence	less gain in very high dimensions	[88, 89, 90]
D-optimal Design	Algorithmic selection of points to maximize information gain.	Flexible run counts and constraints; suitable for irregular spaces	Model-dependent; requires a set of candidate points; iterative	[91]

Once the data have been acquired, either through experiments or from existing historical records, the phase of data preprocessing begins. This step is indispensable because real-world manufacturing data, regardless of their origin, are frequently compromised by noise, missing values, and inhomogeneity. A recent meta-analysis highlights the significance of this phase, showing that thorough data cleaning enhances model performance. In contrast, utilizing low-quality data deteriorates the data attributes and does not effectively represent data patterns [92].

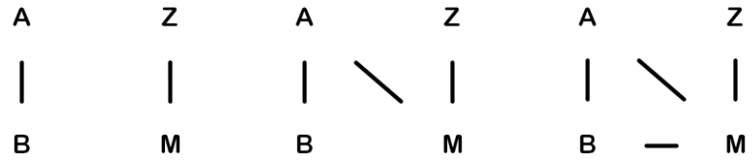
The entire preprocessing pipeline can be broken down into several stages, which are summarized in Table 2.5 and elaborated on in the sections that follow.

Table 2.5: Key data pre-processing stages.

Stage	**Main Tasks**
Data Profiling	Integration and inspection of data types, ranges, and distributions; joining data from multiple sources.
Outlier Treatment	Identification and handling of anomalous data points that can skew models.
Missing-Value Imputation	Imputation of missing entries based on the underlying missingness mechanism.
Scaling & Transformation	Encoding and standardization of numerical features and encoding of categorical variables.
Feature Engineering	Selection and creation of new variables and reduction of data dimensionality.
Data Fusion	Integrating data from multiple sources into one data set.

The pipeline begins with a comprehensive profile of the properties of the data including data types, value ranges, and statistical distributions, which establishes the diagnostic baseline for all downstream choices [93]. Building on this profile, outlier management is undertaken to control anomalous observations that could disproportionately skew model fitting and compromise optimization results. This challenge is well recognized in large-scale data analytics [77].

With outliers addressed, attention shifts to treating missing values. Manufacturing processes often produce incomplete data sets. Although simple methods such as mean or median replacement are available, a robust strategy requires understanding the underlying reason the data are missing. The appropriateness of a method depends on assumptions about why data are missing, formalized as Missing Completely at Random (MCAR), Missing at Random (MAR) and Missing Not at Random (MNAR) [94]. This classification, illustrated in Figure 2.8. The illustration represents a simple causal chain $A \rightarrow B$ together with the missing data indicator M and an exogenous noise term Z.



(a) MCAR **(b)** MAR **(c)** MNAR

Figure 2.8: Taxonomy of missing data, adapted from [95].

- MCAR: Here, the probability of a value being missing is entirely stochastic, independent of both observed and unobserved variables. As shown in Figure 2.8, the missingness indicator M is driven only by the random noise Z. In this scenario, analyses on the remaining complete data can remain unbiased.
- MAR: In the case of MAR, the likelihood that data are missing is contingent upon other observed variables, however it remains unrelated to the actual missing value. For MAR data, likelihood-based methods or multiple imputation can yield unbiased estimates if the model properly includes the predictive observed variables.
- MNAR: Often referred to as non-ignorable missingness, MNAR arises when the likelihood of a value being absent is influenced by the value that

remains unobserved. This mechanism is the most challenging. Standard methods assuming MCAR or MAR will produce biased results unless the missingness mechanism itself is explicitly modeled.

Once missingness has been treated, scale heterogeneity is addressed through normalization and transformation so that variables with large magnitudes do not dominate distance-based models or optimization algorithms [96].

The subsequent stage involves shaping the feature space through feature engineering and dimensionality reduction. Feature engineering refines input variables to better represent the underlying process phenomena. This involves creating composite indicators from raw data. In data sets with numerous highly correlated variables, dimensionality reduction methods are utilized. By truncating the data, a lower-dimensional representation can be created that reduces multicollinearity and noise, simplifying subsequent modeling without information loss [96]. Moreover, diminishing the number of variables is essential for alleviating the curse of dimensionality [97], because both optimization and machine learning algorithms demonstrate enhanced efficiency and robustness in spaces of reduced dimensionality.

Finally, when data originate from multiple systems, data fusion consolidates sources by aligning timestamps, joining unique keys, and enforcing semantic consistency. Proper profiling at the outset and rigorous integration at this stage close the loop, producing a coherent, traceable, and model-ready dataset for subsequent modeling and optimization [93].

2.3.3 Optimization Paradigms: Sequential vs. Offline Approaches

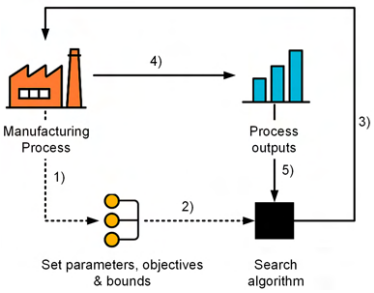
Optimizing process parameters is crucial in modern manufacturing, as stringent quality standards and increasing cost pressures require efficient examination of complex high-dimensional design spaces. In today's manufacturing research,

process parameter optimization is often treated as a derivative-free search problem, where simulating the governing physics is prohibitively costly or the underlying principles are not well-understood enough to reliably utilize gradient information [98]. Two overarching paradigms dominate recent discourse.

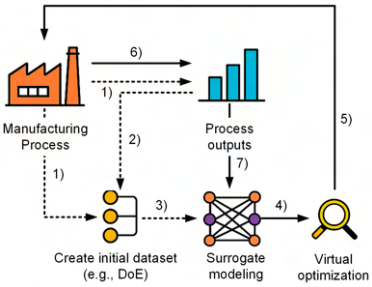
1. Sequential Optimization: These methods operate in an iterative loop directly on the physical process. After each trial, a model is updated, which then guides the selection of the next experimental run [99].

2. Offline Optimization via Static Surrogates: This approach involves two distinct stages. First, a comprehensive data set is generated in a one-time data acquisition campaign, often using DoE. Second, these data are used to train a static surrogate model which then serves as a computationally cheap replacement for the real process. All subsequent optimizations are performed on this static model [100, 101, 102].

Figure 2.9 illustrates, side by side, the two paradigmatic routes for data-driven process parameter optimization.



(a) Sequential optimization workflow



(b) Offline optimization workflow

Figure 2.9: Workflows for data-driven process-parameter optimization. Dashed arrows mark one-off preparatory actions, solid arrows the closed optimization loop.

The sequential optimization workflow, shown on the left, operates as a tight, iterative loop with the real process. After a minimal initial setup, which includes defining variables and objectives, the optimization begins. The solid

arrows underscore that every iteration requires a physical execution of the manufacturing process. The measured outputs are then immediately fed back into the optimizer, which adapts its strategy to select the next promising test point. This paradigm trades a higher number of sequential, individual experiments for the ability to learn and adapt in real time, with the aim of finding the optimum with a minimal total number of physical tests [99]. In contrast, the offline optimization workflow on the right includes a two-stage process. Here, the dashed arrows represents an upfront effort: a comprehensive data set needs to be generated. This data set is then used to train a static surrogate model. All subsequent optimization occurs virtually and exclusively on this static model [100], as indicated by the optimizer interacting with its own surrogate. There is no iterative loop with the real manufacturing process. The final output is a single, theoretically optimal parameter set derived from the model, which can then be validated in a final physical run. This approach front-loads the entire experimental cost to create a global process model, avoiding a lengthy sequence of iterative trials.

Scalability constitutes another decisive differentiator. The effectiveness of sequential optimization depends on the chosen algorithm. Sequential search methods, typically degenerate at roughly 20 decision variables [99]. In contrast, offline optimization scalability is primarily limited by the ability to train a robust surrogate model in a high-dimensional space, where deep learning frameworks generally outperform traditional models, provided that the training set is sufficiently large and diverse [101]. In manufacturing, static process assumptions are rarely tolerated, making adaptivity crucial. As sequential optimization constantly probes the real process, it can naturally adapt to slow process drifts. However, major perturbations to boundary conditions or material batches may still require a re-initialization. The static models used in offline optimization are not inherently adaptive to process changes post-creation. Any larger drift requires a new data acquisition campaign and retraining of the surrogate, although techniques such as transfer learning can reduce this cost [102]. The benefits derived from modeling are also demonstrated in interpretability. Sequential optimization often returns a sparse cloud of observations and a single

optimum, offering limited global insight. The explicit surrogate models central to offline optimization, particularly transparent structures such as polynomials or decision trees, admit sensitivity analysis, improving stakeholder confidence [103]. Finally, the approaches differ in their risk of model error. By grounding each decision in a recent, real-world observation, sequential optimization is less vulnerable to large-scale surrogate model bias, although it remains susceptible to experimental noise. Offline optimization, by relying entirely on its surrogate after the initial data acquisition, introduces a non-negligible risk of model misprediction. This requires careful validation of the final recommended parameters prior to deployment in a facility [101].

The success of both paradigms has been demonstrated in the literature. For the offline approach, a greater than twofold reduction in finite element analysis for textile draping was achieved by implementing a deep neural network surrogate [100]. In contrast, the power of sequential optimization was showcased by achieving laser process convergence in approximately thirty experiments via Bayesian optimization (BO), which represented an order of magnitude improvement over exhaustive searches [101]. Table 2.6 summarizes the key differences of both approaches.

In the field of sequential optimization, there are a multitude of algorithms available. These encompass traditional direct search techniques, including pattern search and Nelder-Mead, as well as population-based heuristics, such as Genetic Algorithms (GA) and Particle Swarm Optimization (PSO) [104]. Despite their adaptability, several of these approaches exhibit low sample efficiency [105]. Because they often lack an explicit model to guide their search, they can probe the objective landscape inefficiently, requiring many evaluations to find an optimum. This is a drawback when each evaluation corresponds to a costly physical experiment.

To overcome this limitation, BO has emerged as the state of the art method for expensive black-box problems [105]. The superiority of this approach is well documented. Numerous empirical studies and benchmarks consistently show that BO achieves equal or better performance with an order of magnitude

Table 2.6: Comparison of sequential and offline optimization paradigms.

Aspect	Sequential Optimization	Offline Optimization
Process Knowledge	Iteratively learned during the search; model is often local or implicit.	Encoded in a global, static surrogate model from an initial DoE.
Evaluation Type	A sequence of real experiments, where each step informs the next.	A one-time batch of real experiments, followed by purely virtual evaluations on the surrogate.
Data Requirements	Low initial requirement; data is gathered one point at a time.	High initial requirement for a comprehensive DoE to build the model.
Computational Effort	Low per iteration; dominated by the duration of the physical experiment.	High upfront cost for model training; virtual evaluations are then very fast.
Adaptivity	High; can adapt to process drift in real-time as it constantly probes the process.	Low; the static model is not adaptive to drift and requires complete retraining.
Interpretability	Low; provides an optimal parameter set but limited global process insight.	High; the global model allows for sensitivity analysis.
Risk of Model Error	Low risk of global model bias, but susceptible to measurement noise.	High risk of surrogate model inaccuracy; the model might not represent the true process

fewer evaluations compared to GAs, PSO, CMA-ES, and uninformed grid or random searches, especially under the tight experimental budgets typical in engineering and manufacturing [106, 107, 108]. Given this convergent evidence, this dissertation adopts BO as the primary algorithm for implementing the Sequential Optimization paradigm. BO is a state-of-the-art method for the sample-efficient optimization of black-box functions that are expensive to evaluate [109].

By building a probabilistic model of the objective function and using it to intelligently select where to sample next, BO can find near-optimal settings with an order of magnitude fewer trials than traditional methods such as grid search

or evolutionary algorithms [105]. This strategy was first popularized to optimize costly engineering simulations in the classic EGO algorithm [110]. The strength of BO derives from the synergy of two fundamental elements: a probabilistic surrogate model and an acquisition function [105]. The surrogate model serves as a cost-effective statistical representation of the expensive objective function. With every new observation, the surrogate is refined, progressively yielding a more precise approximation of the true function as the optimization advances [111]. The acquisition function uses the predictions of the surrogate model to quantify the utility or desirability of evaluating any given point in the search space [111]. In each cycle, BO employs the acquisition function to determine the next sampling location. The main function of the acquisition function is to balance the exploration-exploitation trade-off [112]:

- Exploitation: Directing the search towards regions where the surrogate model predicts a high objective value.
- Exploration: Directing the search towards regions where the surrogate model is most uncertain, as a surprisingly good function value might be discovered there.

The iterative nature of this process is illustrated in Figure 2.10. The graphs depict the mean and confidence intervals as estimated by a probabilistic model of the objective function. Although the objective function is illustrated, it is typically unknown in practice. In addition, the graphs include the acquisition functions in the shaded lower plots, where the peaks indicate candidates that either suggest a high objective value or are located in areas where the surrogate exhibits uncertainty [105].

Algorithm 1 formalizes this workflow. The algorithm begins with an initial dataset and then repeatedly performs the following steps: first, it fits the surrogate model to the current dataset. Subsequently, it maximizes the acquisition function to determine the subsequent query point, where it evaluates the objective function and incorporates the new result into its dataset. This cycle repeats

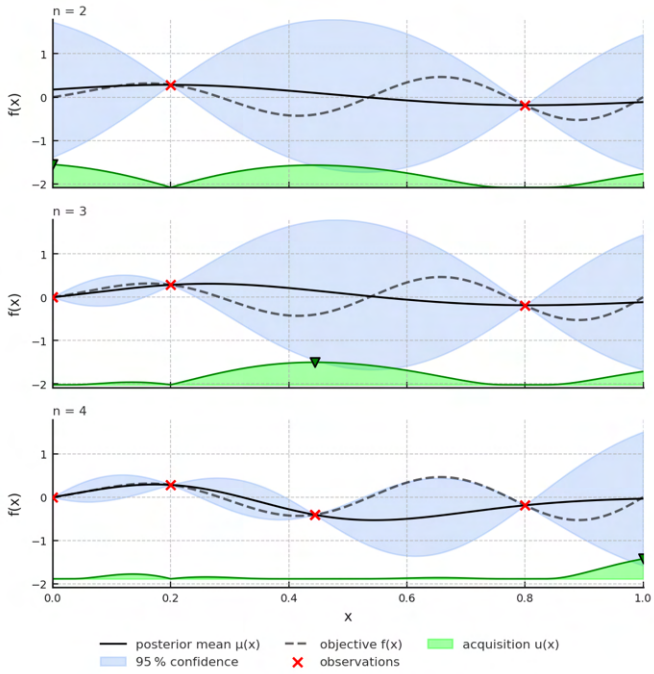


Figure 2.10: Schematic of the BO procedure, as described by [105]. The variable n denotes the iteration step, corresponding to the number of observations collected so far.

until a stopping criterion is met, such as a predefined budget or an improvement threshold [106, 113].

The second paradigm, offline optimization, takes a different approach. Instead of relying on iterative physical trials, it seeks to replace the real process with a data-driven surrogate model, which is then exploited for virtual optimization. This workflow consists of two stages. First, in a one-time data acquisition campaign, a comprehensive dataset is gathered. A supervised machine learning model is then trained on these data to create a static surrogate that accurately maps the input of the process to the output [114, 115]. Once trained, this surrogate serves as a computationally inexpensive substitute for the real system. Optimization algorithms, can then rapidly explore the model's response surface

Algorithm 1 BO Loop

1: **Input:** Initial dataset $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{n}$, search space $\mathcal{X}$, surrogate model class $\mathcal{M}$, acquisition function $\alpha(\cdot; \mathcal{D})$
2: **while** termination criterion not met **do**
3: **Update surrogate:** Fit surrogate model $m \in \mathcal{M}$ to $\mathcal{D}$ to obtain predictive mean $\mu(\mathbf{x})$ and uncertainty $\sigma(\mathbf{x})$
4: **Acquisition maximization:**

$$\mathbf{x}_{\text{next}} \leftarrow \arg\max_{\mathbf{x} \in \mathcal{X}} \alpha(\mathbf{x}; \mathcal{D})$$

5: **Evaluation:** Observe objective value $y_{\text{next}} = f(\mathbf{x}_{\text{next}})$
6: **Dataset augmentation:** $\mathcal{D} \leftarrow \mathcal{D} \cup \{(\mathbf{x}_{\text{next}}, y_{\text{next}})\}$
7: **end while**

to find optimal candidates. These candidates are only validated on the shop floor at the very end. This surrogate-assisted approach is well-established in the literature, both for accelerating complex searches and for modeling physical processes in manufacturing [116, 100]. The general workflow for training the surrogate model is summarized in Figure 2.11.

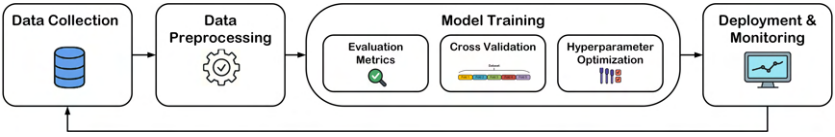


Figure 2.11: A typical supervised learning workflow.

The theoretical pipeline for supervised learning is structured into seven foundational stages.

A reliable supervised learning system is built upon a data collection plan. The process can be decomposed into four interlocking tasks [117]:

1. Sampling Design: This involves identifying the sources, temporal frequency, and coverage of data to ensure that the sample space matches the intended inference domain. When generating new data, rather than

using historical records, this requires selecting a formal experimental design, such as those detailed previously in Table 2.4, to ensure maximum information gain.

2. Instrumentation: This technical phase includes calibrating sensors, logging code, defining fail-safe fallback values, and verifying time-stamp synchronization between distributed data collectors.

3. Storage & Versioning: To guarantee reproducibility and enable rollbacks, every batch of raw data should be committed to an immutable object store along with an explicit schema.

4. Labeling & Quality Assurance: This involves creating provisional labels, often through programmatic or weak supervision, followed by expert adjudication and audit sampling to eliminate systematic errors.

Regarding production stability, large-scale postmortems at Google revealed that more than 60% of major incidents were caused by unvalidated data changes rather than errors in the model code [118]. Concerning label quality, research has shown that corrupting just 5% of labels in a benchmark data set can halve the accuracy of a model, highlighting the extreme sensitivity to annotation errors [119]. Finally, in terms of efficiency, active learning surveys indicate that intelligent sampling strategies can reduce annotation budgets by up to 70%, demonstrating the immense value of a well-designed sampling plan [120].

However, even well-collected data are rarely available in a suitable format for machine learning algorithms, which leads to the next stage in the pipeline: Data Preprocessing, which was already explained in Section 2.3.2.

Once the preprocessing steps are complete, the result is a clean and structured dataset. This data set forms the foundation for supervised learning and consists of a set of input-output pairs $(\mathbf{x}_i, y_i)$. Each $\mathbf{x}_i \in \mathbb{R}^d$ is a feature vector that describes the example of i^{th} and y_i is its associated target. Every instance z has been documented as a pair (x, y) consisting of an input x and a scalar outcome y [121]. The objective is to find a function $f_\theta(\mathbf{x})$, parameterized by θ, that

predicts y from $\mathbf{x}$. This is typically achieved by minimizing the empirical loss on the training pairs.

The most common metrics for evaluating the magnitude of prediction errors are the Mean Squared Error (MSE) and the Mean Absolute Error (MAE) [122]. The Mean Squared Error (MSE) is defined as:

$$\text{MSE} = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2 \tag{2.2}$$

Due to the squaring term, MSE heavily penalizes large errors, making it sensitive to outliers. The RMSE is the square root of the MSE, which is often preferred, as it returns the error to the original units of the target variable.

$$\text{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2} \tag{2.3}$$

Due to the squaring term, both MSE and RMSE heavily penalize large errors, making them sensitive to outliers. In contrast, the Mean Absolute Error (MAE) grows linearly with the error and is therefore more robust [123].

$$\text{MAE} = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_i| \tag{2.4}$$

The choice between the MSE, MAE and RMSE can be informed by assumptions about the error distribution. RMSE is optimal under Gaussian noise, whereas MAE is optimal under Laplacian noise; for this reason, it is common practice to report both [122]. When the relative magnitude of the error is more important than the absolute value, percentage-based or logarithmic metrics are used. The Mean Absolute Percentage Error (MAPE) scales errors by the true value, making it intuitive for percentage-based evaluation [124].

$$\text{MAPE} \;=\; \frac{100}{N}\sum_{i=1}^{N}\left|\frac{y_i - \hat{y}_i}{y_i}\right| \tag{2.5}$$

However, it is undefined when the true value y_i is zero and is known to assign a disproportionately high penalty to errors on small target values [124]. When target values are strictly positive, the Root Mean Squared Logarithmic Error (RMSLE) is often preferred because it measures the relative deviation between predictions and targets.

$$\text{RMSLE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\big(\log(1+y_i) - \log(1+\hat{y}_i)\big)^2} \tag{2.6}$$

The RMSLE inherently penalizes underestimation more than overestimation, making it suitable when prediction shortfalls are more critical than overshoots [125]. The coefficient of determination, R^2, differs from error metrics as it quantifies the proportion of variance in the dependent variable that can be explained by the independent variables [126].

$$R^2 = 1 - \frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{N}(y_i - \bar{y})^2} \tag{2.7}$$

R^2 provides a measure of how well the model explains the variability of the data, with a value of 1 indicating a perfect fit. A large R^2 indicates that most variability is explained, but R^2 can be misleading because it never decreases as predictors are added [126].

The performance metrics discussed above are only truly informative if they estimate how well the model will perform on new unseen data, a property known as generalization. To obtain a robust estimate of this generalization performance and to diagnose issues such as overfitting, various model validation strategies can be employed. The most common family of such strategies is cross-validation (CV), which systematically partitions the data to simulate training and testing on different subsets [127]. A standard and widely adopted implementation of this principle is k-fold CV. The procedure is illustrated in Figure 2.12.

The data set $\mathcal{D}$ is divided into k disjoint folds $\mathcal{D}_1, \ldots, \mathcal{D}_k$, each containing approximately N/k samples. For each fold $j \in \{1, \ldots, k\}$, the model f_θ is

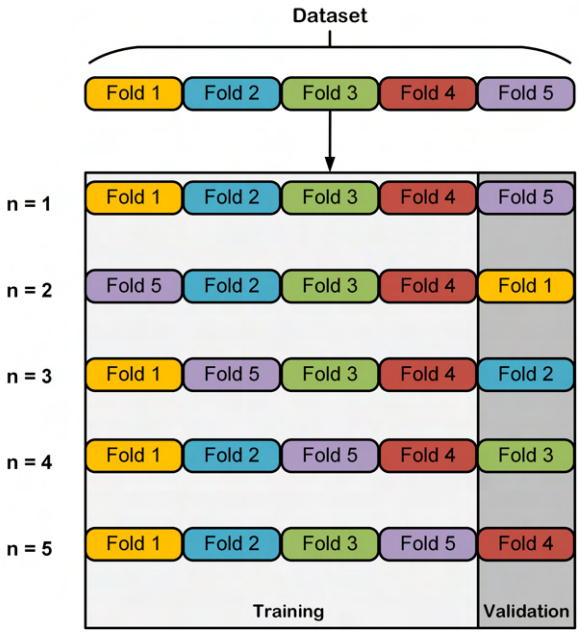


Figure 2.12: Schematic of k-fold CV using k = 5 as an example.

trained on the combined data $\mathcal{D} \setminus \mathcal{D}_j$ and validated on $\mathcal{D}_j$. The resulting CV error $\overline{\mathcal{L}}_{\mathrm{CV}}$ is computed as [128]:

$$\overline{\mathcal{L}}_{\mathrm{CV}}(\theta) \;=\; \frac{1}{k} \sum_{j=1}^{k} \mathcal{L}\big(\theta; \mathcal{D}_j\big), \tag{2.8}$$

where $\mathcal{L}\big(\theta; \mathcal{D}_j\big)$ is the loss incurred by f_θ on the validation fold $\mathcal{D}_j$ [128]. This procedure is repeated for each fold so that every data instance is used exactly once for validation. Compared to a single train-validation split, k-fold CV provides a more stable and reliable estimate of generalization performance, which is particularly valuable for smaller datasets. The choice of the number of folds, k, involves a well-known bias-variance trade-off.

A special, extreme case of this method is Leave-One-Out CV (LOOCV) [129], where k is set to be equal to the total number of samples, N. In this procedure, the model is trained N times, each time using a single data point for validation and the remaining $N - 1$ points for training. The LOOCV error is the average of the errors from these N runs [129]:

$$\overline{\mathcal{L}}_{\mathrm{LOOCV}}(\theta) = \frac{1}{N} \sum_{i=1}^{N} \mathcal{L}\big(\theta; \mathcal{D}_i\big) \tag{2.9}$$

Although LOOCV provides a nearly unbiased performance estimate, it can have high variance and is computationally very expensive. Therefore, in practice, values of $k = 5$ or $k = 10$ are commonly recommended, as they offer a robust compromise between bias and variance for most applications [129, 128]. However, a caveat applies when CV is used in conjunction with Hyperparameter Optimization. The standard k-fold CV provides a reliable performance estimate for a model with a fixed configuration. If hyperparameters are tuned using the same CV procedure that is also used for final performance reporting, the resulting estimate will be optimistically biased. This is because the hyperparameter selection process has effectively seen all the data, and the model is no longer being tested on truly independent folds [130]. To obtain an unbiased performance estimate in such cases, a more advanced procedure known as nested CV is required. This method uses an outer loop to estimate the generalization error and a separate inner loop to tune the hyperparameters for each outer fold, thus preventing information leakage [129]. When the target variable is heavily imbalanced, a stratified CV is advisable. In this variant, each fold $\mathcal{D}_j$ preserves the overall proportion of each class, thus reducing biases in training and validation and producing a more realistic measure of the model's generalization capability [131].

Once a validation strategy is defined, the task of Hyperparameter Optimization (HPO) can be addressed. This is a meta-optimization problem where the goal is to find the optimal configuration for the learning algorithm itself [132]. Let $\Theta \subseteq \mathbb{R}^m$ represent the hyperparameter search space. The aim is to identify the

optimal set of hyperparameters $\mathbf{h}^*$ that minimizes the projected generalization error:

$$\mathbf{h}^* = \arg\min_{\mathbf{h}\in\Theta} \overline{\mathcal{L}}_{\mathrm{CV}}(\mathbf{h}) \tag{2.10}$$

Solving this optimization problem can be approached with methods of varying sophistication. Foundational approaches include exhaustive Grid Search and Random Search, which are simple to implement but often sample inefficient as they do not learn from past evaluations [113]. More advanced strategies treat this as a sequential optimization problem. Methods such as BO are typically more efficient [113].

In supervised learning, performance is an estimate of generalization under uncertainty rather than a fixed constant. Even with CV, fluctuations arise from finite sample noise and the randomness injected by common procedures such as shuffling, resampling, stochastic optimization, and randomized model components. Computationally, this randomness is generated by pseudo-random number generators (PRNGs), which are deterministic algorithms that map an initial state to a statistically well-behaved sequence of numbers [133]. A random seed is a finite integer that initializes this state. By fixing the seed, the subsequent pseudo-random stream becomes deterministic for a given algorithm and software configuration [133]. To manage this variability, setting a random seed is essential. It ensures that the results are reproducible, which is critical for fair comparison, debugging, and verification. Moreover, averaging results over multiple random seeds or experimental runs is a common practice to account for performance variability and to provide more robust and reliable performance estimates [134].

If the surrogate model demonstrates satisfactory predictive performance, it is deployed for virtual optimization. Engineers or data scientists can then probe this fast-to-evaluate surrogate to identify parameter settings that are predicted to optimize the target metric [100].

$$\mathbf{x}^* \;=\; \arg\min_{\mathbf{x}\in\Omega} \hat{y} = \arg\min_{\mathbf{x}\in\Omega} f_\theta(\mathbf{x}), \tag{2.11}$$

where $\Omega \subset \mathbb{R}^d$ is the feasible design space.

2.3.4 Explainable AI and Decision Support

In modern engineering and production settings, advanced AI models such as ensemble learners and deep neural networks are increasingly being used for tasks ranging from design optimization to predictive maintenance. These models often reach high levels of predictive accuracy, but are generally considered black-boxes due to the difficulty humans face in interpreting their internal decision-making processes [135]. This opacity presents a problem in engineering contexts, where comprehending the rationale behind the prediction for a model is essential to establish trust, ensure safety, achieve regulatory compliance and extract new scientific insights from the data [136]. XAI is the field devoted to addressing this challenge. It provides methods to extract human-interpretable explanations from complex models, helping engineers to validate the behavior of the model and gain confidence in the implementation of AI solutions [137]. The aim is to improve transparency through XAI and help users understand and trust AI decisions, ensure compliance with regulations, and overcome the black-box problem [138]. A distinction in XAI is the classification of models into three categories - white box, black box, and gray box [139], which is based on their internal transparency, as illustrated in Figure 2.13.

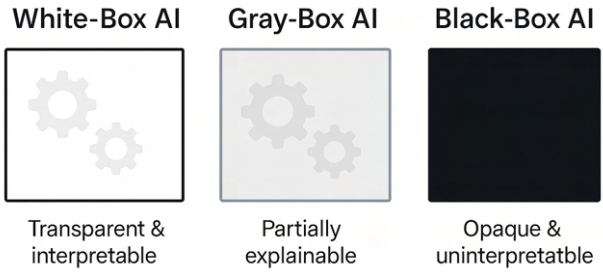


Figure 2.13: A comparison between white-box, gray-box, and black-box models, as outlined by [139].

White-box models are those with a transparent internal structure that is directly understandable to humans, such as linear regression and rule-based systems [140]. This transparency allows for direct inspection of model components, such as analyzing the weights of a linear regression to understand the importance of features. Their primary advantage lies in this inherent interpretability, which simplifies the explanation and aids in adherence to regulatory requirements [139]. However, this transparency frequently results in a decrease in predictive capability, as such models are often overly simplistic to adequately represent complex, non-linear interactions within the data. Moreover, while technically a white box, the interpretability of a linear model with thousands of interaction terms can diminish at scale, rendering it effectively opaque to human understanding [141]. By contrast, black-box models, such as deep neural networks, are highly complex algorithms whose internal workings are obscured by millions of interacting parameters and nested non-linear transformations [142]. The key benefit of these models is their ability to achieve high predictive performance by automatically learning intricate patterns from high-dimensional data [139]. Their drawback is their inherent opacity. This opacity, in turn, makes post-hoc XAI techniques indispensable: The behavior of the model is explained by analyzing input–output relations rather than inspecting its internal structure. Without these supplementary methods, black-box models are difficult to trust or troubleshoot, potentially obscuring hidden biases or errors [139, 143]. Gray-box models occupy an intermediate position by integrating mechanistic domain knowledge with data-driven learning [144]. For instance, a model for a chemical reactor might use known differential equations for reaction kinetics but employ a neural network to learn a complex, unmodeled heat transfer coefficient from operational data. The advantage is that the model structure is partially grounded in established theory, which enhances trust and can reduce data requirements. However, this hybrid approach often leads to increased design complexity. Moreover, while the mechanistic part is interpretable, the data-driven components remain opaque, meaning the final system is not fully transparent and may still require XAI techniques to explain the learned relationships [141]. In many engineering domains, deriving good white-box models is infeasible, making black-box models indispensable for accurately fitting complex phenomena from

data. Although these models offer high predictive accuracy, their inherent lack of interpretability poses a challenge, as engineers are reluctant to trust or deploy systems that they cannot diagnose. This dilemma has motivated the development of post hoc XAI techniques, which are designed to explain the predictions of any pre-trained model, thus bridging the gap between high performance and the need for transparency [145].

2.4 Summary of Theoretical Foundations

This chapter established the conceptual foundation on which the remainder of the dissertation is based. It began by introducing AM, which details the layer-by-layer fabrication principle that distinguishes it from traditional subtractive methods. The second section introduced frameworks for environmental assessment. Two primary methodologies were detailed: the holistic LCA, a cradle-to-grave approach standardized by ISO 14040 and ISO 14044 for multi-impact analysis, and the focused Carbon Footprint Analysis, which quantifies GHG emissions according to scopes 1, 2, and 3. Based on these assessment tools, dominant strategies for impact reduction were introduced. These ranged from the systemic vision of the CE, with its hierarchy of resource management strategies, to the operational focus of green and sustainable manufacturing on improving energy and material efficiency at the process level. Furthermore, the chapter introduced the necessary stages for data-driven optimization. Two primary optimization paradigms were described: Sequential optimization, an iterative online approach, and offline optimization, which relies on training a static surrogate model from a comprehensive upfront dataset. Finally, to address the need for trust and safety in complex algorithms, the chapter concluded by introducing the principles of XAI, a field dedicated to making the decisions of black-box models transparent and understandable.

3 State of the art

Having established the theoretical foundations for data-driven, sustainable manufacturing, this chapter analyzes the state of the art to identify the specific research gaps that motivate this dissertation. The review examines current approaches in both academic literature and industrial practice, focusing on their intersection to pinpoint key deficiencies and opportunities for innovation. To structure this analysis, a systematic review methodology is employed, as detailed in the following subsection. The review concentrates on a representative subset of the literature to highlight the most relevant use cases and models. This approach allows for a clear formulation of the existing research gaps that this thesis aims to address.

3.1 Methodology of Literature Review

The approach used to conduct the literature review is in line with the PRISMA guidelines [146]. However, given the broad scope of the topic and the primary objective of identifying research gaps rather than providing an exhaustive overview, the PRISMA methodology was applied in a simplified manner. Following this methodology, the state of the art analysis is structured into two main parts. The first part provides a broad overview of comparable methodologies and applications in data-driven sustainable manufacturing. It examines general approaches, drawing a distinction between those prevalent in academic research and those implemented in industrial practice. This highlights the current landscape and the common gap between theoretical potential and practical application. The second part conducts a more focused technical review. It

identifies the specific state of the art for each of the key theoretical foundations established in the previous chapter.

The methodology underpinning this review of the literature is shown in Figure 3.1. This process ensures a reproducible and comprehensive analysis, guiding the review from the initial search to the final synthesis and identification of research gaps.

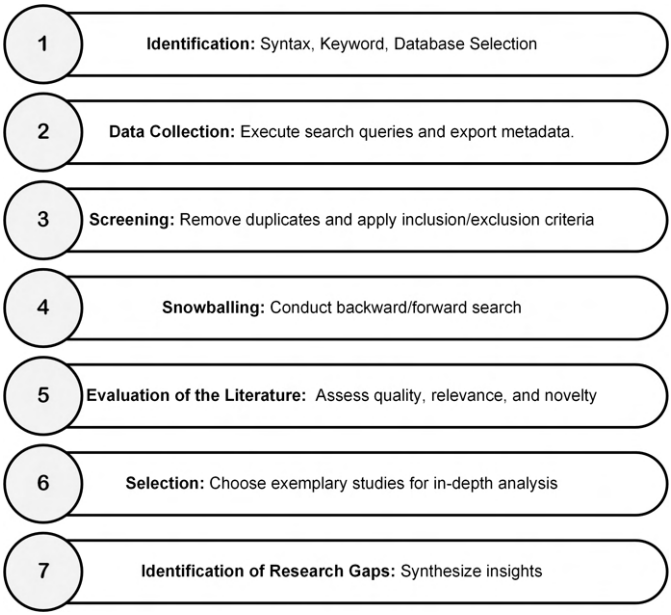


Figure 3.1: Methodological structure of the literature review.

Moreover, the analytical framework developed in this dissertation builds on and complements two prior literature reviews: Greif et al. [1] and Greif et al. [2], which examined the interaction between AI and environmental sustainability from both a strategic and methodological point of view. The first review provides a strategic assessment of the transformative role of AI in advancing sustainability through a SWOT analysis, while the second offers a structured

synthesis of AI methods applied in the context of the SDGs. In conjunction with the present domain-specific investigation into data-driven manufacturing optimization, these three reviews form a coherent analytical foundation. Their combined insights enabled a triangulated gap analysis, thus guiding the identification of unresolved challenges and shaping the methodological scope of this dissertation.

The basis of this review is a well-defined search query, constructed from the terms detailed in Table 3.1. This search was performed in multiple scientific databases including Scopus, Web of Science, and IEEE Xplore and was systematically restricted to English-language publications from January 2015 to May 2025 to capture contemporary advances.

Following the database search, all the records retrieved were exported to the bibliographic management software. The initial screening step involved removing duplicate entries to create a consolidated master bibliography. Subsequently, the titles and abstracts of the unique records were screened against a set of predefined inclusion and exclusion criteria, as detailed in Table 3.2, to assess their relevance.

To ensure comprehensive coverage, this pool of articles was then expanded through snowballing, a process involving both backward checks of reference lists and forward tracking of citations to capture influential studies missed by the initial search. Finally, the full texts of the remaining articles were subjected to a detailed quality assessment based on the checklist presented in Table 3.3. This step evaluated the suitability, methodological transparency, and validity of each study, and low-quality articles were excluded from the final synthesis.

Finally, the selected corpus of high-quality articles was subjected to detailed analysis and synthesis. To structure this review, the literature was organized into three primary thematic sections. The analysis first examines the core methodological frameworks for data-driven optimization, then reviews the specific applications of these methods in manufacturing, and finally assesses their impact on environmental sustainability. This structured synthesis allows for a

Table 3.1: Composite search strategy integrating technology domains, method categories, and their associated terms.

Layer	Category	Search Strings
Technology	Additive Manufacturing	("Additive Manufacturing" OR "3D printing" OR "rapid prototyping")
	General Manufacturing	("Manufacturing" OR "Production Process" OR "Industrial Process")
Method	Experimental Design	("Design of Experiments" OR "experimental planning" OR "sampling strategy")
	Data Preprocessing	("Data Preprocessing" OR "Data Cleaning" OR "Data Preperation")
	Machine Learning	("Machine Learning" OR "predictive analytics" OR "data-driven modeling")
	Optimization	("Optimization" OR "Improvement" OR "Efficiency" OR "Parameter Tuning")
	Bayesian Optimization	("Bayesian Optimization" OR "probabilistic optimization" OR "Bayesian modeling")
	Explainable AI	("Explainable AI" OR "XAI" OR "interpretability" OR "transparency")
Goal	Sustainability	("Sustainability" OR "Environment" OR "Green")
	Robustness	("Robustness" OR "Variance" OR "Noise" OR "Stability")
Template	—	(Technology AND Method) OR (Technology AND Goal) OR (Technology AND Method AND Goal)

Table 3.2: Inclusion and exclusion criteria for title and abstract screening.

Inclusion Criteria	Exclusion Criteria
Focus on data-driven optimization in a manufacturing or industrial context.	Purely theoretical works or applications outside manufacturing.
Publication type is an original research article.	Publication is a review, meta-analysis, editorial, or opinion.
Abstract indicates that a specific methodology was applied and data were used.	Abstract lacks information on the methods or data used.

Table 3.3: Quality assessment checklist for full-text screening. To be included, studies must satisfy at least two of the three assessment items for each criterion.

Criterion	Assessment Item
Suitability	• Direct application to manufacturing process optimization • Explicit focus on environmental or sustainability outcomes • Clear description of use case context (industry sector, process type)
Methodological Transparency	• Detailed description of data sources (origin, size, measurement protocols) • Clear specification of algorithms or experimental protocols • Disclosure of parameter settings and tuning procedures
Validity	• Use of appropriate evaluation metrics • Presence of validation or replication • Discussion of limitations and threats to validity

systematic identification of converging findings, current limitations, and, ultimately, the pressing research gaps that this dissertation aims to address.

A parallel but distinct approach was taken for the industry-oriented review. Although the criteria in Tables 3.2 and 3.3 were strictly applied to the academic literature, these were deliberately relaxed for industry sources to accommodate the limited availability of peer-reviewed documentation. For these sources, such as white papers, technical briefs, and reports from leading firms, the focus shifted to capture representative use cases. Industry sources were included if they:

- Were published by recognized industry players or industry consortiums
- Presented specific examples of data-driven optimization implementations with stated sustainability or efficiency outcomes,
- Provided sufficient descriptive detail regarding technology, method, and environmental impact to permit categorization and thematic analysis.

Quality thresholds such as detailed methodological specifications and formal validation procedures were considered desirable but not mandatory for industry sources, given the non-academic context of these communications.

3.2 State of the Art in Academic Research and Industrial Practice

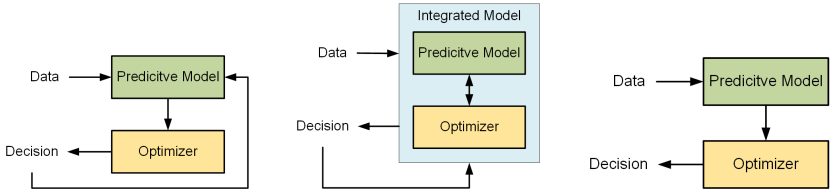
This section presents a state-of-the-art analysis of data-driven, sustainable manufacturing. The review is structured into two complementary parts to provide a comprehensive overview: First, a review of the academic research landscape, which examines the paradigms and trend in data-driven optimization, common applications, and the extent to which sustainability has been integrated into these studies. Second, an analysis of the industrial perspective, which identifies

how these technologies are being implemented in real-world production environments and the practical challenges encountered. By contrasting these two domains, this analysis aims to systematically identify the key research gaps that exist at the intersection of theoretical innovation and practical application.

3.2.1 Research

3.2.1.1 Paradigms, Trends and Applications of Data-Driven Optimization

Recent academic literature on data-driven optimization in manufacturing reveals several key trends, particularly in the development of core optimization paradigms and the deep integration of explainability. In the reviewed literature, data-driven optimization approaches are broadly classified into three categories, as illustrated in Figure 3.2.



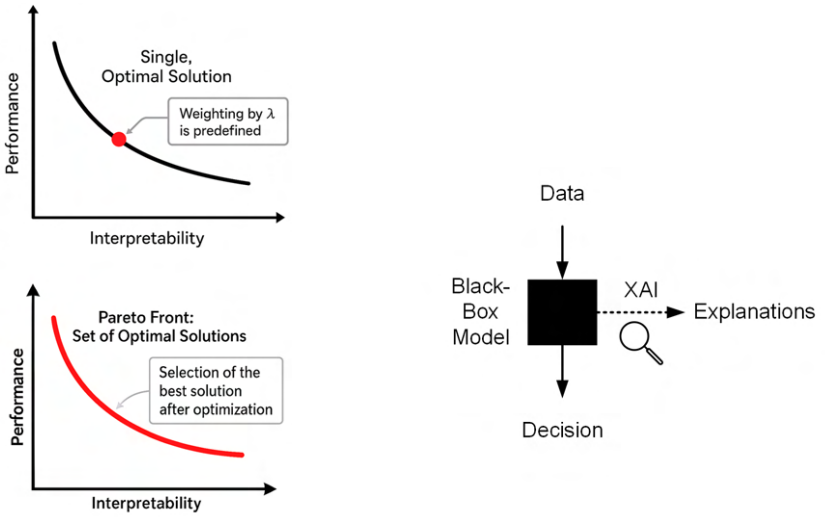
(a) Decision-focused Learning **(b)** End-to-End Learning **(c)** Predict-then-Optimize

Figure 3.2: Overview of Data-Driven Optimization Approaches.

The classic predict then optimize approach first trains a predictive model and then uses it as a surrogate for a separate optimization algorithm [147]. In contrast, end-to-end approaches embed the optimization task directly into the predictive model's training by creating a differentiable optimization layer, thereby training the model to maximize the actual decision quality rather than just predictive accuracy [148]. Decision-focused learning is a related paradigm that modifies the training loss of the predictive model to explicitly account for how its predictions will affect the downstream optimization task [149]. Alongside

these, a key trend is the use of meta-learning for automated algorithm selection. This approach uses machine learning to learn from the characteristics of past optimization problems to automatically choose the most appropriate algorithm for a new unseen problem. This enables systematic performance improvements by taking advantage of experience from a library of problems and reduces the need for manual comparisons of trial-and-error algorithms [150].

Within the manufacturing sector, there is a strong emphasis on not only improving performance but also ensuring that these optimized solutions offer reliability and valuable insight for practitioners [137]. Two primary methodologies exist for this purpose, as depicted in Figure 3.3.



(a) Interpretability by Design **(b)** Post-Hoc Explainability

Figure 3.3: Architectures for XAI in Data-Driven Optimization.

The first approach involves embedding explainability directly into the model's structure and objective function. This is often done in two distinct ways: One strategy involves augmenting the classical objective function with a regularization term. This term penalizes complexity or rewards similarity to a simpler

model, effectively combining performance and interpretability into a single, weighted objective. By pre-defining this weight, the user specifies their preference a priori, and the algorithm returns a single solution that reflects this fixed trade-off [151]. In contrast, the second strategy formulates interpretability as a separate, competing objective, which is common in multi-objective evolutionary algorithms. Instead of finding a single solution, this approach generates a pareto front, which is a diverse set of solutions where each represents an optimal trade-off. This method allows the user to analyze the entire spectrum of trade-offs after the optimization run and to select the solution that best meets their specific needs for both performance and transparency [152]. An example of such an inherently interpretable system is the InterOpt algorithm developed for shale gas production. It uses XAI to quantify the importance of various operational parameters to guide the optimization of drilling settings. By analyzing the importance of the features, InterOpt was able to recommend adjustments that led to a nearly 10 % cost reduction, while also providing transparency about which parameters were the most influential on the results [153].

The second approach involves using post-hoc XAI techniques to provide insights into complex, often black-box, models after they have been trained. Several recent comprehensive applications in manufacturing exemplify this trend. These methods often rely on established tools to analyze model behavior. In the context of manufacturing, for example, Shapley Additive Explanations (SHAP) were used to interpret a machine learning model linking 3D printing process parameters to printed part characteristics, thus identifying which input factors have the greatest impact on part quality [154]. Other works have proposed model-agnostic systems that combine tools such as Local Interpretable Model-agnostic Explanations (LIME) and counterfactual analysis to assist operators in tuning machine tool parameters for optimal production quality and failure reduction, with the system explaining its suggestions in human-understandable terms [155, 153]. These XAI tools are increasingly integrated into comprehensive frameworks for sustainability applications. The 'Sustain AI' framework, for instance, employs multimodal deep learning to minimize the carbon footprint of manufacturing operations from streaming Industrial IoT (IIoT) data.

A notable feature is its use of SHAP and LIME to make recommendations transparent. If the system suggests rescheduling an energy-intensive task, it can highlight the input factors that most influenced the suggestion. The experiments demonstrated reduced energy consumption and carbon emissions, and the clear rationales provided for each recommendation led to greater reported trust in the AI system [156]. Another powerful post-hoc technique is the use of surrogate explanation models. In one study addressing a CE problem, an optimization based on Reinforcement Learning (RL) found complex policies for a degrading manufacturing system that balances production, maintenance, and quality control. To make these policies understandable, simpler decision tree models were subsequently trained on the RL-generated data to explain the optimal policy in terms of human-readable rules. By following a tree path, it becomes clear how factors such as machine condition or backlog influence the recommended actions, thus building trust [157].

A framework designed to optimize sustainable manufacturing processes integrated deep RL, evolutionary algorithms, and transfer learning. It embedded XAI modules in the decision-making layer to provide transparency into how the AI is balancing trade-offs, such as how a suggested process adjustment saves energy at the cost of a small increase in cycle time. The application of this framework across three separate manufacturing environments demonstrated an average decrease of 27.4% in energy usage and an enhancement of 18.2% in the efficiency of material use [158]. Such interpretable machine learning techniques help domain engineers understand the why behind model predictions and optimization recommendations, bridging the gap between black-box models and actionable process knowledge. The practical selection of such XAI tools is also an active area of research, with recent benchmarks comparing the robustness and real-time feasibility of methods such as SHAP and LIME for industrial applications [159]. As XAI surveys in Industry 4.0 settings highlight, introducing transparency and user trust is crucial when deploying AI-driven optimization in resource-constrained manufacturing environments [160].

3.2.1.2 Data-Driven Optimization in Environmental Sustainability

The data-driven optimization paradigm has emerged as a powerful tool for addressing complex environmental challenges that are often characterized by high-dimensional, nonlinear, and stochastic systems. By interfacing ML models with advanced optimization algorithms, ranging from BO and metaheuristics to RL, these approaches can navigate vast decision spaces to improve system performance and policy outcomes. However, a critical review of the recent literature reveals a clear spectrum of maturity in the adoption of XAI. Although the efficacy of data-driven optimization has been widely demonstrated, the degree to which these systems provide transparent, trustworthy, and interpretable results varies between domains. The following analysis explores this spectrum, from fields with emerging XAI integration to those where mature implementations exists.

In domains with mature data infrastructure, such as energy systems and water resource management, XAI is transitioning from a concern to an integral component for validating solutions and de-risking the deployment of autonomous systems. In the energy sector, a framework for building design optimization has been developed that couples a LightGBM surrogate model with a multi-objective evolutionary algorithm (NSGA-II) to navigate trade-offs between energy demand, CO_2 emissions, and thermal discomfort. SHAP analysis were incorporated to decompose the predictions of the surrogate model [161]. In parallel, an explainable reinforcement learning (XRL) approach has been applied to dynamic control problems such as power grid reconfiguration, in which the agent's learned policy not only optimizes network efficiency but also allows the extraction of rules-based policies that are human-intelligible [162]. A similar trajectory is observable in sustainable water management. To mitigate the risks associated with black-box controllers, RL has been applied to optimize urban drainage systems. To address the black-box nature of the RL agent, an interpretability framework was introduced in the study, where a decision tree surrogate is trained to mimic the agent's policy. This pedagogical surrogate provided a transparent and logical approximation of the complex policy, allowing

human experts to visualize how sensor inputs map to specific control actions [163].

In contrast, other environmental domains are characterized by a performance-centric application of data-driven optimization, where algorithmic efficiency is prioritized over the transparency of the decision-making process. This is particularly evident in pollution control, where, as highlighted in a recent review, metaheuristic algorithms such as genetic algorithms (GA) and particle swarm optimization (PSO) are used to navigate the complex, non-convex search spaces typical of bioremediation processes [164]. Although these methods effectively optimize parameters for the removal of contaminants, the resulting solutions are often implemented without a clear understanding of their underlying causal mechanisms. The literature indicates a gap in scientific insight and stakeholder trust [164]. A similar gap is evident in waste management. In studies on municipal waste collection, for example, predictive models of waste generation have been coupled with route optimization algorithms to address variants of the Vehicle Routing Problem (VRP), reporting double digit percentage reductions in collection costs [165]. However, the decision logic remains opaque. The causal drivers influencing the optimized routes are seldom investigated or reported. In both domains, the focus remains on the optimization outcome, while the reasoning behind the optimal solution is unaddressed. Beyond post hoc analysis, the high-stakes domain of climate science, defined by its profound societal implications and the imperative for scientific discovery, presents a compelling case for a paradigm shift toward inherently interpretable models. In this field, data-driven optimization is crucial for calibrating computationally intractable, physics-based simulations. For example, a GP BO has been applied to tune parameters of the WRF regional climate model. This successfully minimized the discrepancy between model outputs and observed data for extreme heat events, improving the fidelity of the model [166]. However, the why behind the optimized parameter set remain elusive. Addressing this, the argument has been made that the goal should not be to simply explain black-box AI models, but to design gray-box or physics-informed AI architectures that embed known physical constraints and causal relationships [167]. This perspective signals

an open research frontier: developing climate optimization algorithms that not only yield accurate solutions, but also provide mechanistic insights. Such systems would need to quantify epistemic uncertainty while offering transparent reasoning that aligns with established domain knowledge, a prerequisite for trusted adoption in climate policy. This entire spectrum of XAI adoption, driven by application risk, is conceptually summarized in Figure 3.4.

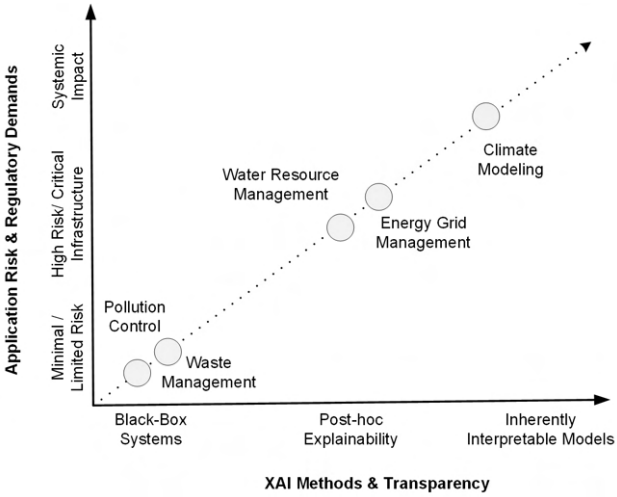


Figure 3.4: Risk-Driven Selection of XAI Strategies in Environmental Optimization.

This evolution towards greater explainability also aligns with emerging regulatory landscapes, most notably the European Union's AI Act [168]. This regulation introduces a risk-based approach, classifying AI systems into tiers with corresponding obligations for transparency and robustness. Many of the applications discussed in environmental optimization fall into these categories. AI systems employed in the management and operation of essential infrastructure, explicitly encompassing energy grids and water systems, are categorized as high-risk. This designation mandates transparency, robustness, and human oversight, providing a strong regulatory driver for the adoption of post-hoc XAI techniques. Conversely, systems where the primary risk is economic, such

as optimizing waste collection routes, would likely be classified as 'minimal' or 'limited' risk, thus having fewer formal explainability requirements. This regulatory distinction helps explain the observed gap, where XAI is often absent in lower stakes applications. Furthermore, while domains such as climate modeling may not fall into the same operational risk category, their profound impact on policy and society places them in a de facto class of systemic risk. This justifies the scientific community's demand for a standard beyond mere compliance, the call for inherently interpretable models, to ensure scientific validity and public trust in high-stakes decisions. Therefore, the selection of an optimization and XAI strategy is increasingly a function of this risk-based calculus, shifting the paradigm from one focused purely on performance to one grounded in transparency, safety, and regulatory alignment.

3.2.1.3 Artificial Intelligence for Sustainable Additive Manufacturing

Building upon the general paradigms and application trends, AM serves as a particularly illustrative case study where a multitude of AI techniques are being applied to enhance sustainability across the entire product lifecycle.

AI is being applied from pre-process design to in-process control and system-level management to reduce material use and eliminate waste. The optimization of sustainability in AM begins in the preprocess phase, long before fabrication starts. A primary area of impact is in design optimization, where AI-driven generative design and topology optimization algorithms automatically evolve part geometries. These methods explore vast design spaces to meet performance requirements such as stress and stiffness with minimal material usage, often creating complex lattice or bio-mimetic structures that are uniquely suited for AM [38].

Complementing this is the use of AI for predictive quality assurance. By training machine learning models on historical build data, it is possible to predict the likelihood of success for a given geometry and orientation before committing

material and energy to the print, guiding users toward setup changes that improve first-time success rates [20].

During the manufacturing process itself, AI is essential for navigating the high-dimensional parameter space that governs quality, energy consumption, and build time. The challenge lies in optimizing numerous controllable parameters, such as laser power, scan speed, or extrusion temperature, where traditional tuning is time consuming and often suboptimal. To address this, the predominant approach in academic research involves surrogate-based optimization. Computationally cheap data-driven surrogate models are trained to approximate the complex input-output relationships of the AM process. For example, an active learning framework that uses GP-based BO has been shown to iteratively identify optimal printing parameters to improve geometric accuracy with only a few hundred data points [169]. To handle the complex and dynamic nature of processes such as DED, more complex surrogates such as long-short-term memory (LSTM) networks have been integrated with BO algorithms to dynamically optimize laser power time series and achieve target material properties [170]. Their approach maximized heat treatment time within 50 BO evaluations.

Beyond setting optimal initial parameters, AI is also crucial for real-time, in-process monitoring, and quality control to reduce waste from failed builds. High-resolution cameras and sensors are used to monitor each layer as it is deposited, and computer vision models, particularly convolutional neural networks (CNNs), are trained to recognize anomalies such as porosity, delamination, or geometric deviations from the intended toolpath. For example, a CNN-based machine vision system was demonstrated to detect defects during multiaxis 3D printing, allowing the process to be stopped or adjusted immediately rather than continuing a flawed build to completion [171]. By avoiding print failures or correcting them on the fly, such AI systems reduce the scrap rate, conserving the material and energy that would otherwise be lost.

Finally, at the system level, AI contributes to the sustainable operation of all AM facilities. In settings with multiple printers, AI-driven scheduling algorithms can intelligently batch and sequence print jobs to maximize energy efficiency,

for example, by clustering parts with similar material requirements to avoid frequent thermal cycling of machines [172, 173].

Furthermore, predictive maintenance, powered by machine learning models that analyze sensor data for indicators of degradation, ensures that machines operate at peak efficiency and avoids the production of faulty parts due to machine drift or failure. Although indirect, these factory-level optimizations are integral to a holistic approach to sustainable AM [20].

As illustrated in Figure 3.5, AI amplifies the sustainability of AM by enabling material efficient designs through generative algorithms, improving the quality through intelligent process optimization, and minimizing waste through predictive quality control and real-time monitoring.

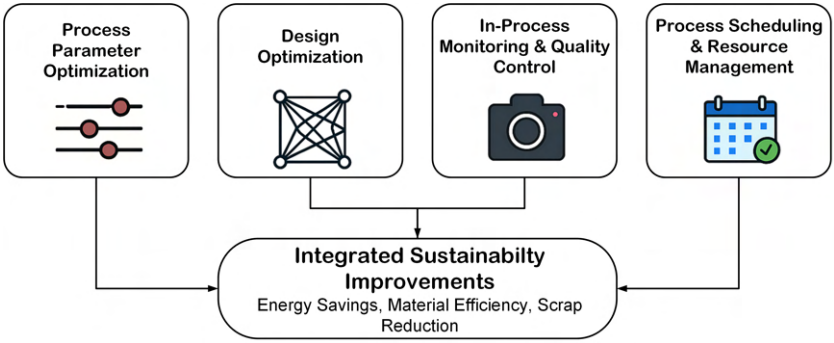


Figure 3.5: Overview of AI-driven contributions to Sustainable AM across the product and system lifecycle, from pre-process design to in-process control and facility-level management.

3.2.2 Industry: State of the Art in Sustainable Manufacturing Optimization

The contemporary industrial landscape is increasingly characterized by the integration of digital technologies into manufacturing processes. From software suites that simulate entire production lines to cloud-based IIoT platforms that

gather real-time data from thousands of assets, industry leaders are building powerful ecosystems for data-driven optimization.

The foundation for modern sustainable optimization is a robust data and simulation infrastructure. Large industrial software providers are embedding AI and sustainability metrics directly into their offerings. Siemens, for example, incorporates green scheduling techniques in its Opcenter APS to reduce energy use during production planning, while its Tecnomatix Plant Simulation tools allow companies to create digital twins of their facilities. These simulations enable the virtual testing of numerous production scenarios to reduce development time by 20 to 30 % [174, 175]. Similarly, AspenTech's adaptive process control uses AI (Aspen DMC3) to improve plant efficiency, increasing yields by up to 3 % while reducing energy consumption by up to 10 % in process industries [176]. Alongside these established suites, specialized IIoT platforms from providers such as Sight Machine and Uptake are enabling real-time analytics by ingesting streaming plant data and applying machine learning to reduce downtime and improve quality [177, 178].

One of the most mature application areas is cloud-based energy management. Bosch's Energy Platform, now deployed in over 120 of its facilities and offered commercially, exemplifies this. The platform aggregates real-time data from machines and building systems to pinpoint inefficiencies. At its plant in Qingdao, China, this system saved approximately 380 MWh in a single year by optimizing resource use [179]. Furthermore, by integrating an AI model to autonomously regulate a chilled water system based on load forecasts, Bosch achieved an additional 10 % energy reduction for that specific system [179]. Similarly, a major automotive manufacturer applied machine learning in its paint shop operations, correlating patterns in temperature, humidity, and airflow to paint quality. This data-driven approach cut energy consumption by 15% while simultaneously improving the paint finish [180].

AI-driven quality control has a direct impact on sustainability by reducing scrap, rework, and wasted materials. The initiative at Bosch provides a powerful case study. In its Bamberg plant, an AI analytics platform ingests roughly one million

data signals per day from assembly stations to automatically flag any deviation from standard process signatures, enabling real-time quality control that virtually eliminates waste [181]. In the automotive sector, computer vision is a key enabler. One manufacturer applied machine learning to inspect printed circuit boards, catching 50 % more defects while reducing human inspection effort by 25 %, leading to higher first-pass yields and less waste [182]. Predictive maintenance uses AI to forecast equipment failures, preventing unplanned downtime and associated energy waste and scrap. The BMW Group used computer vision and machine learning to monitor its robotic assembly lines, analyzing more than 18,000 images per minute to predict robot failures with approximately 90% accuracy. This system prevents over 200 unplanned downtime incidents annually, saving millions and reducing waste from sudden breakdowns [183].

On a large industrial scale, Siemens Energy's Connected Factory platform serves as another notable example. By linking numerous assets at 18 different worldwide factories through an AWS cloud-based infrastructure, this IIoT system employs analytics to oversee production and forecast maintenance requirements. The reported outcomes include up to 25% lower maintenance costs and 15% higher machine availability, demonstrating a direct link between improved asset performance and more efficient, sustainable operations [184]. In summary, the industrial state-of-the-art demonstrates a consistent, two-tiered approach where a foundational layer of data and technology infrastructure enables targeted optimization in key application areas. This integrated architecture is conceptually summarized in Figure 3.6.

Beyond foundational platforms, industry is increasingly deploying a portfolio of advanced AI methods to tackle complex optimization problems that were previously intractable. This new wave of adoption moves beyond predictive analytics toward more autonomous systems that can model, optimize, and control processes with greater precision and efficiency. A primary challenge in manufacturing is optimizing processes that are too expensive, slow, or complex for exhaustive physical tests. To address this, manufacturers are widely adopting surrogate models, data-driven approximations of physical processes—to accelerate R&D and optimization. In the aerospace and automotive sectors,

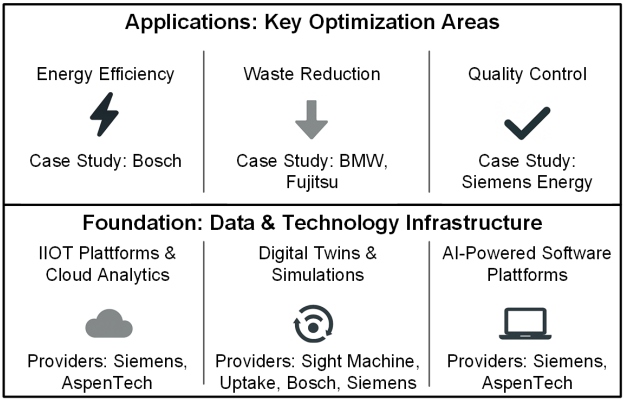


Figure 3.6: A two-layer architecture for industrial data-driven optimization.

high-fidelity digital twins of processes such as machining or injection molding are used to generate data, from which computationally cheap surrogates are trained to predict outcomes such as energy use or product quality. These surrogates can then be rapidly explored using optimization algorithms to find optimal settings with minimal waste or energy [174]. Another example is Google's DeepMind. It used ML to cut the cooling energy of its data center by 40 %[185]. As powerful but often opaque ML methods take on greater decision-making roles, their adoption hinges on trust and transparency. Leading firms such as Bosch emphasize the need for "safe, robust, and explainable" AI for industrial use [186].

3.3 Focused Review of state of the art Technologies, Methods and Future Directions

Having reviewed state of the art applications and methodologies in industry and research, this section provides a focused analysis of the specific technologies and algorithms that enable these solutions. First, the key algorithms and methods

employed in the application-focused papers cited in the preceding sections will be distilled, offering a top-down perspective grounded in documented practice. Second, this is supplemented with a bottom-up review of foundational and benchmark literature to identify the current state of the art algorithms for each core component of a data-driven optimization system.

3.3.1 Data Preprocessing

The success and reliability of any data-driven application is critically dependent on the quality of its input data. Data preprocessing is a prerequisite for building robust models. However, the rigor with which these steps are applied and reported varies across the literature. Table 3.4 provides a high-level summary of the techniques used in the application studies.

Table 3.4: Data preprocessing techniques across selected manufacturing optimization papers.

Paper	Outlier Detection	Missing-Data Imputation	Feature Scaling	Feature Selection
[187]	Yes (residuals obtained during linear regression analysis)	No	No	Yes (feature importance)
[188]	Yes (removed tool-breakage outlier cuts)	Yes (linear interpolation)	Yes (standardized)	Yes (random & tsfresh selection)
[189]	No	Yes (linear/poly/KNN/MICE/MissForest)	Yes (normalized)	No
[190]	Yes (pixel-based anomaly detection)	No	No	No
[170]	Yes (not specified)	No	Yes (not specified)	Yes (not specified)

The literature demonstrates a variety of approaches, from statistical methods to domain-specific knowledge. For example, in a study on Li-ion battery electrode manufacturing, a trimmed mean criterion was used to remove gross outliers from sensor readings, while residuals from a linear regression analysis were used to flag anomalous data points [187]. A more domain-specific approach was taken in a tool wear prediction study, where entire failure cases corresponding to broken tool outliers were explicitly discarded from the dataset, leveraging expert knowledge of the process [188]. For predictive maintenance of semiconductor equipment, an anomaly detection method was presented in which a median reference image was used to calculate a pixel-wise outlier index to identify early wear patterns [190].

Missing values are another challenge in real-world manufacturing data sets. Although simpler methods such as linear interpolation between adjacent measurements are common [188], more sophisticated approaches have been shown to yield superior performance. In a study conducted using the SECOM semiconductor dataset, where 65% of the data was absent, an extensive evaluation of five distinct imputation techniques was performed. The methods analyzed included linear and polynomial interpolation, k-nearest neighbors (with k = 5), multiple imputations by chained equations (MICE), and the nonparametric MissForest algorithm. MissForest was found to have the best downstream classification performance, resulting in a 15% improvement in the F1 score over the other approaches [189]. This highlights the impact that the choice of imputation strategy can have on the model outcomes.

With the high dimensionality of sensor data in modern manufacturing, feature scaling and selection are crucial to improve model training efficiency and performance. A simple yet effective method is to use the built-in feature importance scores of tree-based models to rank and select inputs [187]. Another workflow was proposed for the prediction of wear of CNC tools using a two-stage pipeline. This process first used a random subset search and then applied statistical tests to reduce more than 200 raw and time series features to a core set of 25, which accelerated model training by 60% without compromising predictive power [188].

Despite the demonstrated importance of these techniques, a significant portion of the literature on data-driven optimization omits preprocessing steps or does not report them sufficiently in detail [155, 191, 169]. In simulation-based studies, data are often clean by design, while in some experimental studies, the focus on surrogate modeling overshadows the need for systematic data cleaning. Even when preprocessing is mentioned, descriptions are often generic, for example, 'data cleaning', lacking the specific algorithms, parameters, or thresholds needed for reproducibility. This gap is a critical issue, as neglecting these basic steps can degrade model accuracy, slow convergence, and compromise the reliability of optimization results, especially under the tight data budgets common in manufacturing.

To bridge the gap between the often ad-hoc preprocessing reported in application papers and the standards required for reproducible research, it is essential to detail the state of the art algorithmic approaches for these tasks. The data preparation workflow begins with the identification and handling of outliers, as these can disproportionately influence model training and compromise optimization results. The choice of method depends on whether outliers are defined based on statistical distributions, model performance, or algorithmic anomalies. For industrial data, which often deviates from idealized Gaussian assumptions, methods grounded in robust statistics are preferred over classical approaches based on mean and standard deviation. A principal technique utilizes the Median Absolute Deviation (MAD), whose high breakdown point (nearly 50%) ensures its influence function remains bounded even with significant data contamination [192, 193]. An observation $x_{j,i}$ is flagged as an outlier based on its absolute deviation from the feature median, $r_{j,i} = |x_{j,i} - \mathrm{median}(\mathbf{x}_j)|$, whenever the following condition is met:

$$r_{j,i} > \kappa \, \mathrm{MAD}(\mathbf{x}_j), \qquad \text{where} \quad \mathrm{MAD}(\mathbf{x}_j) = \mathrm{median}\big(r_{j,\cdot}\big). \tag{3.1}$$

The threshold κ is typically chosen in the range $[2.5, 3.5]$, corresponding to a cutoff from the χ^2 distribution [194]. By normalizing with the MAD, robust z-scores can be computed [195],

$$z_{j,i}^{(\text{rob})} = \frac{x_{j,i} - \text{median}(\mathbf{x}_j)}{b \cdot \text{MAD}(\mathbf{x}_j)}, \tag{3.2}$$

which, unlike classical z-scores, are not affected by the presence of the outliers themselves. The constant b is a scaling factor to make the MAD a consistent estimator for the standard deviation under Gaussianity [195].

Alternatively, outliers can be defined as observations that are poorly predicted by a model trained on the majority of the data. This model-based approach assumes that outliers are observations that do not conform to the relationship between the features and the target variable learned from the majority of the data. As used by [187], a regression model is first trained on the data set to act as a representation of the normal behavior of the process. The model prediction errors are then used as an anomaly score, with the logic that large errors indicate points that the model could not explain based on the learned patterns. Given an observation y_i and a model prediction $\hat{f}(x_i)$, the centered absolute residual r_i is calculated [187]:

$$e_i = y_i - \hat{f}(x_i), \tag{3.3}$$

$$r_i = \big|\, e_i - \text{median}(\{e_k\}) \big|. \tag{3.4}$$

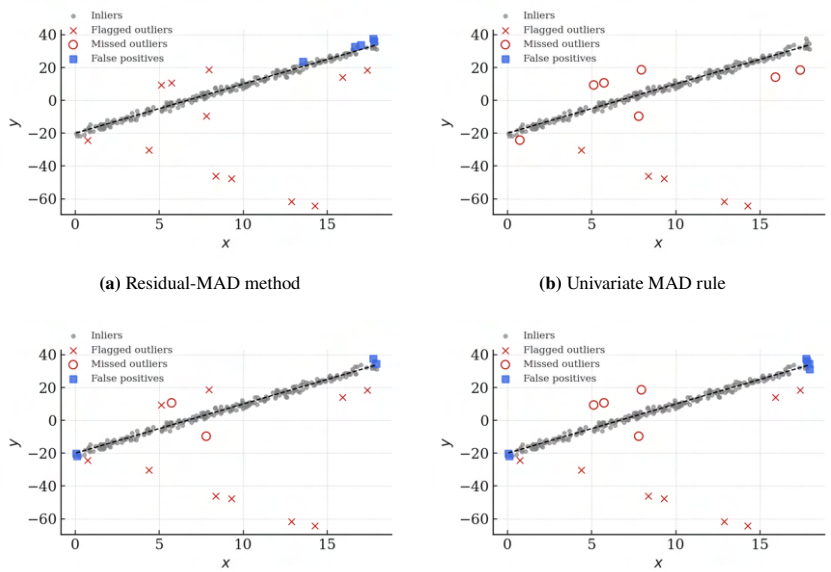
An observation is flagged if its residual exceeds a threshold based on the MAD of all residuals: $r_i > \kappa \cdot \text{MAD}_e$ [194].

Complementing these methods, the Isolation Forest algorithm offers a non-parametric, algorithmic approach. It isolates anomalies in high-dimensional data by recursively partitioning the feature space [196]. Observations that require fewer partitions to be isolated receive a higher anomaly score $s(x)$, and are flagged as outliers if they exceed a user-defined threshold τ [196]:

$$s(x) \geq \tau. \tag{3.5}$$

A related method, particularly suited for streaming data, is the Robust Random Cut Forest (RRCF) algorithm [197]. Similar to the Isolation Forest, RRCF uses an ensemble of trees to isolate points. However, its anomaly score is not based on path length, but on the expected change in model complexity, specifically the number of leaves upon removing a point from a tree. Anomalies, being far from other data, cause a larger structural change when removed, and thus receive a higher score, making the method robust and effective for online anomaly detection.

Figure 3.7 illustrates the behavior of these methods applied to a synthetic data set characterized by a linear trend $y = 3x - 20$. The data was seeded with three types of anomaly: vertical outliers, bad-leverage outliers, which represent large deviations at high x, and genuine inliers.



(a) Residual-MAD method

(b) Univariate MAD rule

(c) Robust Random Cut Forest

(d) Isolation Forest

Figure 3.7: Comparison of anomaly-detection methods applied to a synthetic data set.

The MAD rule primarily highlights vertical outliers but misses leverage-type anomalies, because their distances from the feature median remain moderate. The residual MAD approach, which relies on model errors, additionally captures the bad-leverage points, yet leaves high-leverage observations lying exactly on the regression line untouched, as their residuals are small. The isolation forest isolates vertical and leverage anomalies, but introduces a few false positives near the data boundaries. The RRCF reproduces this pattern while reducing the false-positive rate through its complexity change scoring scheme. Overall, the experiment confirms that univariate robust statistics suffice for strongly axis-aligned anomalies, whereas model-based methods better handle multivariate structures, albeit at the cost of a slightly higher false positive rate.

Subsequent to addressing outliers, the next task is to manage missing values. The selection of an appropriate imputation method is determined by the mechanism of the missing data, which can be categorized as MCAR, MAR, or MNAR [198]. To empirically determine the most plausible mechanism, there are several analyses. First, Little's omnibus test χ^2 probes the null hypothesis that the data are MCAR, whereas a non-significant result supports this assumption [199]. The test statistic pools deviations of pattern-specific means from the overall mean [199]:

$$\chi^2_{\text{Little}} = \sum_{r=1}^{R} n_r \left(\bar{\mathbf{y}}_r - \bar{\mathbf{y}} \right)^\top S^{-1} \left(\bar{\mathbf{y}}_r - \bar{\mathbf{y}} \right), \tag{3.6}$$

where R is the number of distinct missing data patterns, n_r the sample size of the pattern r, $\bar{\mathbf{y}}_r$ the vector of observed means for that pattern, $\bar{\mathbf{y}}$ the general mean vector and S the pooled covariance matrix estimated from complete cases. In MCAR, χ^2_{Little} follows an asymptotic χ^2 distribution with degrees of freedom $\sum_r (p_r - 1)$, where p_r denotes the number of fully observed variables within the pattern r.

Furthermore, to investigate the MAR hypothesis, the dependence of the missingness on fully observed data can be modeled by fitting a logistic regression

where R indicates a missing value and $\mathbf{X}$ represents the set of fully observed predictors [200]:

$$\Pr(R = 1 \mid \mathbf{X}) = \text{logit}^{-1}\big(\alpha + \mathbf{X}\beta\big) \tag{3.7}$$

Substantial predictive power, measured by a pseudo-R^2 or the Area Under the Receiver Operating Characteristic Curve (AUC), supports the MAR assumption [200].

Another two-stage approach first fits a model on complete cases to predict missing values, $\hat{Y} = f(\mathbf{X})$, and subsequently tests in a second regression whether the propensity to be missing depends on $\hat{Y}$ by testing $H_0 : \gamma_1 = 0$ [201]:

$$\Pr(R = 1 \mid \hat{Y}) = \text{logit}^{-1}\big(\gamma_0 + \gamma_1 \hat{Y}\big) \tag{3.8}$$

A significant relationship strongly indicates an MNAR process [201]. Alternatively, especially with cross-sectional data or longitudinal dropout, selection models such as the Heckman two-step procedure [202] or the Diggle-Kenward model [203] are used. These approaches simultaneously model the outcome and selection processes:

$$\begin{aligned} Y_i &= \mathbf{X}_i\beta + u_i, \\ R_i^* &= \mathbf{Z}_i\gamma + v_i, \qquad R_i = 1[R_i^* > 0], \end{aligned} \tag{3.9}$$

with $\text{Cov}(u_i, v_i) = \rho\sigma_u\sigma_v$. A test showing a non-zero correlation ($\rho \neq 0$) rejects the ignorability assumption and thereby confirms an MNAR mechanism [202, 203].

The outcome of the diagnostics directly informs the selection of a suitable imputation strategy. For data sets with a low percentage of missingness ($< 10\%$) that can be plausibly assumed to be MCAR, simple imputation methods are often

used. These include replacing missing values with the feature's empirical mean for symmetric distributions [204]:

$$\tilde{x}_i = \bar{x} = \frac{1}{n_{\text{obs}}} \sum_{i: x_i \neq \text{NA}} x_i, \tag{3.10}$$

or median for skewed distributions [204]:

$$\tilde{x}_i = \text{median}\big\{x_i : x_i \neq \text{NA}\big\}. \tag{3.11}$$

Although trivial to implement, these single-imputation techniques are known to artificially reduce the true variance of the data and can distort correlation structures [204].

For categorical variables, mode imputation, substituting the most frequent level, is defined as [205]:

$$\tilde{x}_i = \arg\max_c \#\{\, x_j = c\}, \tag{3.12}$$

but also ignores variability.

In time series with brief gaps, the Last Observation Carried Forward (LOCF) method extends the most recent valid measurement [206]:

$$\tilde{x}_t = x_{t'}, \quad \text{where } t' = \max\{\, s < t : x_s \neq \text{NA}\}. \tag{3.13}$$

When the missingness mechanism is more complex or when inter-feature correlations are strong, more sophisticated methods are required. The k-nearest Neighbors (KNN) imputation technique is a non-parametric approach that addresses missing data by substituting each absent value with an aggregate derived from its k closest fully observed neighbors within the feature space [207].

$$\tilde{x}_{i,j} = \frac{1}{k} \sum_{i' \in \mathcal{N}_k(i)} x_{i',j}. \tag{3.14}$$

The Expectation Maximization (EM) algorithm is a powerful model-based approach that assumes that the data follow a specific multivariate distribution. The algorithm iteratively alternates between an "E-step," which computes the expected value of the log-likelihood function [208]:

$$Q(\theta \mid \theta^{(t)}) = \mathbb{E}_{X_{\text{mis}}|X_{\text{obs}},\,\theta^{(t)}}\big[\log f(X \mid \theta)\big], \tag{3.15}$$

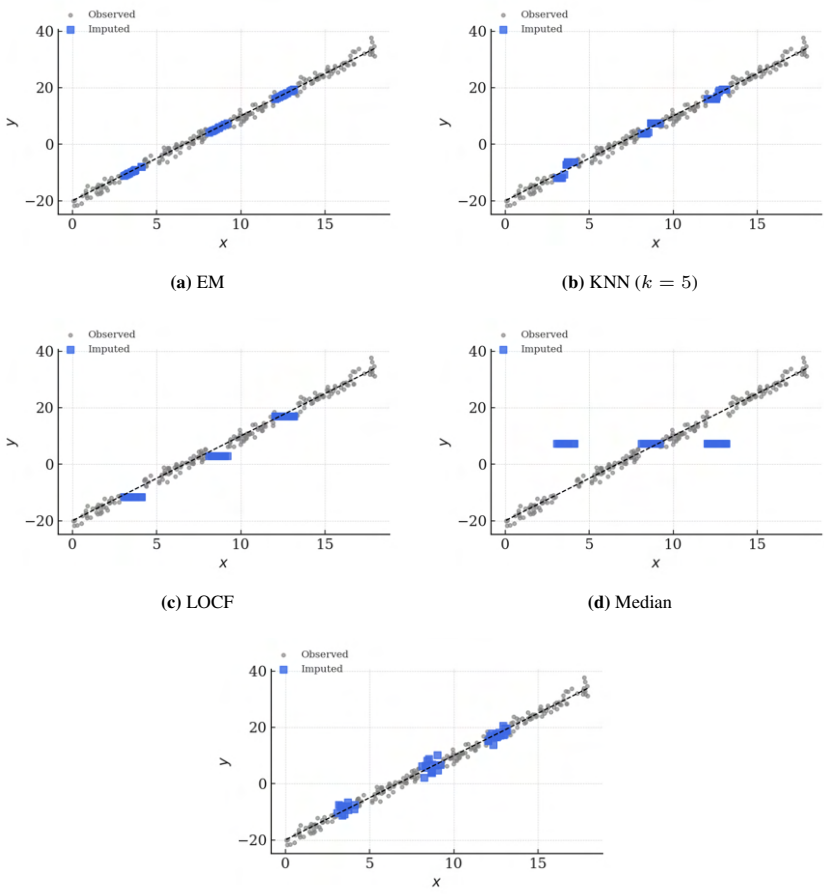
and an "M-step," which updates the distribution parameters by maximizing this expectation [208]:

$$\theta^{(t+1)} = \arg\max_{\theta}\, Q(\theta \mid \theta^{(t)}). \tag{3.16}$$

For a more flexible approach, Multiple Imputation by Chained Equations (MICE) offers a robust framework. MICE generates M completed datasets by cycling through features and sampling from their conditional predictive distributions [198]:

$$\tilde{x}_{j,i}^{(m,\ell)} \sim p^{(\ell)}\big(x_j \mid \tilde{\mathbf{x}}_{-j,i}^{(m,\ell-1)}\big), \quad \ell = 1,\ldots,L,\ m = 1,\ldots,M. \tag{3.17}$$

To demonstrate the influence of standard imputation methods, in Figure 3.8 three MCAR gaps were excised from a synthetic data set governed by the linear equation $y = 3x - 20$.



(a) EM

(b) KNN ($k = 5$)

(c) LOCF

(d) Median

(e) MICE

Figure 3.8: Visual comparison of five imputation strategies applied to identical MCAR gaps.

Simple mean and median imputation filled each gap with a single constant, visibly flattening local variance and producing horizontal plateaus that break the underlying slope. LOCF propagated the preceding value into each gap, introducing staircase artifacts that bias the series along the trend direction. In contrast,

the kNN approach leveraged neighborhood information from both sides of every gap and produced estimates that adhere closely to the ground truth line, although the filled values still underrepresent natural variability. Expectation maximization under a bivariate normal model placed the imputations exactly on the conditional expectation, thereby preserving the slope but shrinking the variance to zero within each gap. MICE restored both trend and dispersion: posterior draws scattered around the regression line and thus conveyed realistic uncertainty. Overall, the experiment shows that single-deterministic strategies risk structural bias, while model-based approaches may oversmooth. In this example, MICE offers the best trade-off between trend fidelity and variance preservation, followed closely by kNN.

The most challenging scenario arises when data is MNAR. Tackling MNAR data necessitates the formulation of specific, unverifiable assumptions regarding the mechanism causing the data to be missing. Three principal frameworks are available to do so. The first, selection models, involves simultaneously modeling the joint distribution of the data and the missingness indicator [203]. A second approach, Pattern-Mixture Models, stratifies the data by the observed missingness patterns and factorizes the joint distribution as $P(Y, R) = P(Y \mid R)P(R)$, where R encodes the missingness pattern [209]. Finally, shared parameter models posit that a latent variable influences both the outcome of interest and the probability that it is missing, enabling joint estimation assuming conditional independence [210]. An MNAR analysis is a workflow, not a single procedure. It should begin with a baseline MAR model for reference, followed by fitting one or more well-justified MNAR models. A step is to conduct extensive sensitivity analyses by varying the key untestable parameters of the MNAR model to assess the robustness of the conclusions. Finally, for any imputation method, a concluding sanity check is essential. This involves comparing the distributions of imputed versus observed values and assessing the sensitivity of the final, downstream model to the chosen imputation strategy. This validation step ensures that the process of handling missing data does not inadvertently introduce bias or distort the conclusions of the subsequent optimization.

Furthermore, manufacturing data sets often include features measured in different units and with varying ranges, which can cause attributes with large magnitudes to dominate distance-based models and optimization routines. To address this, several scaling techniques are commonly employed. Min-Max rescaling linearly transforms each feature into the unit interval [211]:

$$x'_{j,i} = \frac{x_{j,i} - \min_k x_{j,k}}{\max_k x_{j,k} - \min_k x_{j,k}}, \tag{3.18}$$

so that $x'_{j,i} \in [0,1]$ for all i. Although simple and interpretable, this method is sensitive to extreme values. A more robust alternative, especially for data with heavy-tailed distributions, centers each feature by its median and scales by the interquartile range (IQR) [212]:

$$x''_{j,i} = \frac{x_{j,i} - Q_{0.50}(\mathbf{x}_j)}{Q_{0.75}(\mathbf{x}_j) - Q_{0.25}(\mathbf{x}_j)}, \tag{3.19}$$

where $Q_p(\mathbf{x}_j)$ is the empirical p-th quantile of feature j. Another widely used approach is standard scaling, which centers each feature by its mean and scales to unit variance [213]:

$$x^{(\mathrm{std})}_{j,i} = \frac{x_{j,i} - \mu_j}{\sigma_j}, \tag{3.20}$$

where μ_j and σ_j are the sample mean and standard deviation. This normalization is most appropriate for features that are already close to Gaussian. If approximate normality is required by the subsequent model, a monotone power transform can rescale skewed distributions while preserving order. For strictly positive variables, the Box-Cox transformation is adequate [214]:

$$x^{(\lambda)}_{j,i} = \begin{cases} \dfrac{(x_{j,i} + c)^{\lambda} - 1}{\lambda}, & \lambda \neq 0, \\ \ln(x_{j,i} + c), & \lambda = 0, \end{cases} \tag{3.21}$$

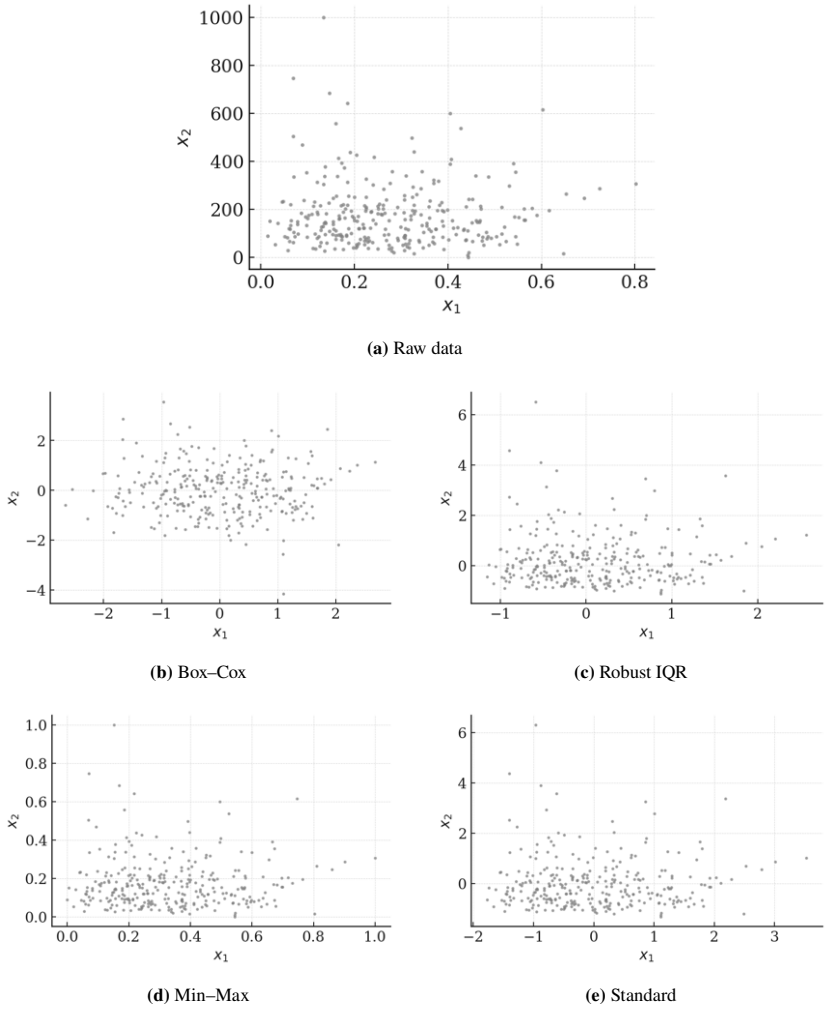
with a shift parameter $c > -\min(x_{j,\cdot})$ ensuring positivity. Whenever a feature contains zeros or negative values, the Yeo–Johnson variant is preferred

because it extends the Box–Cox approach to negative values while retaining monotonicity [215]:

$$x_{j,i}^{(\lambda)} = \begin{cases} \dfrac{[(x_{j,i}+1)^{\lambda}-1]}{\lambda}, & x_{j,i} \geq 0,\ \lambda \neq 0, \\ \ln(x_{j,i}+1), & x_{j,i} \geq 0,\ \lambda = 0, \\ -\dfrac{[(-x_{j,i}+1)^{2-\lambda}-1]}{2-\lambda}, & x_{j,i} < 0,\ \lambda \neq 2, \\ -\ln(-x_{j,i}+1), & x_{j,i} < 0,\ \lambda = 2. \end{cases} \tag{3.22}$$

Both λ and, where applicable, c are estimated by maximizing the log-likelihood under the assumption of normal residuals, ensuring that the transformed feature is as close to Gaussian as possible without sacrificing the rank order of observations.

To illustrate the different scaling techniques, a bivariate synthetic data set was created with heterogeneous feature ranges: a beta-distributed variable in $(0, 1)$ and a log normal variable scaled to $[0, 1000]$. Figure 3.9 shows how the joint distribution changes under each transformation. In the raw space, distance-based models are dominated by the high-magnitude feature. Min–Max scaling compresses both variables into $[0, 1]$, equalizing scale but retaining skewness and outlier leverage. Robust IQR scaling, centered on the median and scaled by the interquartile range, mitigates outlier influence and yields a more balanced spread. Standard normalization removes mean offsets and enforces unit variance, producing a roughly elliptical cloud when the marginal distributions are near-Gaussian, but remains sensitive to heavy tails. Finally, the Box-Cox power transform symmetrizes skewed distributions and stabilizes variance, transforming the log-normal feature into a shape comparable to the beta variable.



(a) Raw data

(b) Box–Cox

(c) Robust IQR

(d) Min–Max

(e) Standard

Figure 3.9: Effect of different scaling techniques on two heterogeneously scaled features.

Converting categorical variables into a numerical form appropriate for ML algorithms requires a set of techniques. The method selected is dependent on

whether the levels of the variable have an intrinsic order. For unordered features, one-hot encoding is the standard approach [216]. It produces a binary indicator matrix $\mathbf{E} = [e_{ik}]$ that avoids imposing a false ordinal relationship [216]:

$$e_{ik} = \begin{cases} 1, & \text{if } x_i = \text{level}_k, \\ 0, & \text{otherwise.} \end{cases} \tag{3.23}$$

Conversely, for ordered process states or grades, a simple integer mapping preserves the monotonic structure of the data [217]. An ordinal map γ : $\text{level}_k \mapsto k$ is defined such that [217]:

$$z_i = \gamma(x_i) \in \{1, \ldots, K\}, \tag{3.24}$$

thereby retaining the crucial rank-order information that a one-hot encoding scheme would obscure.

Following scaling and transformation, feature engineering is the process of using domain knowledge to construct new, more informative variables from the raw data. The goal is to make the relationships between the underlying processes explicit and more accessible to ML algorithms, which can significantly improve the performance of the model. A common example in sustainability applications is the creation of an efficiency metric, such as energy consumed per unit of throughput, which can be a more powerful predictor than either variable alone [218]. This can be expressed as [218]:

$$f_i^{(\text{eff})} = \frac{E_i}{T_i}, \tag{3.25}$$

where E_i is the total energy consumption and T_i is the throughput for the i-th production run.

In the context of high-dimensional datasets obtained from multiple sensors, or in cases where feature correlation is substantial, dimensionality reduction methods are crucial for developing models that are both robust and efficient. PCA offers a linear transformation that maps the data onto a new, lower-dimensional

orthogonal basis [219]. Here, each new feature acts as a linear combination of the original features. These components are ordered such that the first component captures the largest possible variance in the data, the second captures the largest remaining variance, and so on. This is achieved by performing an eigen-decomposition on the sample covariance matrix Σ of the mean-centered data matrix $\mathbf{X} \in \mathbb{R}^{n \times p}$ [219]:

$$\Sigma = \frac{1}{n-1}\mathbf{X}^\top \mathbf{X}. \tag{3.26}$$

The eigenvectors $\mathbf{v}_k$ of Σ form the principal axes, and the corresponding eigenvalues λ_k represent the variance captured by each axis [219]:

$$\Sigma \mathbf{v}_k = \lambda_k \mathbf{v}_k, \quad \text{where } \lambda_1 \geq \cdots \geq \lambda_p. \tag{3.27}$$

The full set of principal components $\mathbf{Z}$ is obtained by projecting the original data onto the matrix of eigenvectors $\mathbf{V}$ [219]:

$$\mathbf{Z} = \mathbf{X}\mathbf{V}, \quad \text{where } \mathbf{V} = [\mathbf{v}_1, \ldots, \mathbf{v}_p]. \tag{3.28}$$

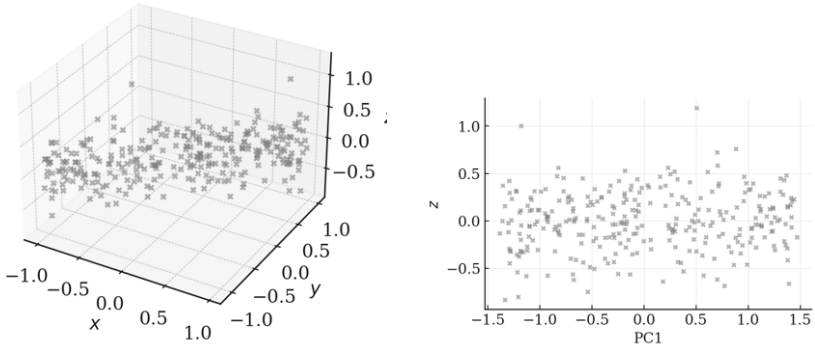
By retaining only the top $k \ll p$ components [219]:

$$\mathbf{Z}_k = \mathbf{X}[\mathbf{v}_1, \ldots, \mathbf{v}_k], \tag{3.29}$$

a large fraction of the total variance, given by $\sum_{i=1}^{k} \lambda_i / \sum_{j=1}^{p} \lambda_j$, can be preserved in a much smaller feature set. This truncation removes multicollinearity and reduces noise in subsequent optimization tasks [220, 221]. In manufacturing contexts, PCA is often used to extract dominant, interpretable modes from sensor data, such as vibration patterns or thermal gradients that explain most of the variability of the process [222].

To illustrate this, a synthetic three–dimensional data cloud was generated such that two coordinates, x and y, are almost perfectly correlated, while z contains independent noise. Figure 3.10 contrasts the original space with the reduced representation. In the left panel, the points form a thin diagonal sheet within

the $x-y$ plane. In the right panel, PCA replaces the redundant pair (x, y) with a single score, PC1, which captures more than 98% of their joint variance, while the uninformative correlation between PC1 and z is eliminated. The transformation thus condenses the feature set from three to two dimensions without loss of information.



(a) Original 3-D feature space **(b)** Reduced space

Figure 3.10: Dimensionality reduction using PCA.

3.3.2 Selection Metrics

Selecting an appropriate optimization strategy is a non-trivial task, often referred to as the algorithm selection problem [223]. The challenge arises from the vast number of available algorithms and the complex and heterogeneous nature of real-world optimization problems. This challenge can be addressed by first conducting a quantitative characterization of the problem that evaluates both the sufficiency of the available data and the structural properties of the underlying landscape.

A first consideration is data sufficiency, which is most directly quantified by the sample-to-dimension ratio (SDR). This metric measures data density as [224]:

$$\rho = \frac{N}{d}, \tag{3.30}$$

where N is the number of labeled observations and d is the dimension. The SDR is an indicator for optimization strategy selection. For example, GP surrogates have been shown to typically require approximately $N \approx 10d$ observations for reliable performance [225]. If a problem has a low SDR, sequential optimization methods that actively acquire data are favored. In contrast, an exceptionally high SDR can make a single offline optimization using a static surrogate particularly efficient.

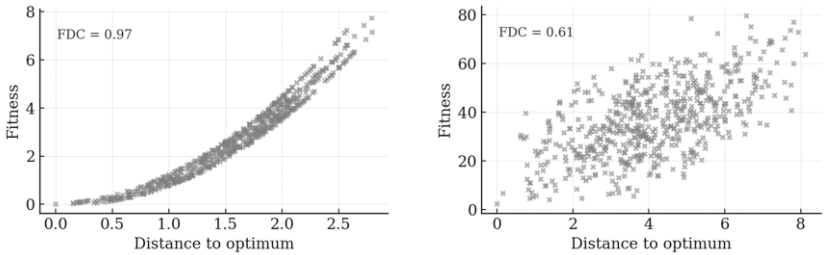
Beyond data sufficiency, the geometric and statistical properties of the landscape of objective functions are determinants of algorithm performance. The discipline of Exploratory Landscape Analysis (ELA) seeks to quantify these characteristics through a variety of statistical measures, thereby offering insights into a problem's modality, separability, and overall complexity [226]. One of the most informative ELA metrics is the Fitness Distance Correlation (FDC), which computes the Pearson correlation between the fitness of a solution, f_i, and its distance to a known global optimum, d_i [227, 228]:

$$\mathrm{FDC} = \mathrm{corr}(f_i,\, d_i). \tag{3.31}$$

A high positive FDC suggests a smooth unimodal landscape that is suitable for local or model-based search methods, while a low or negative FDC indicates a deceptive or rugged landscape that calls for a global search strategy, as contrasted in Figure 3.11.

Another key property is the smoothness of the landscape, which can be quantified by the length of the autocorrelation, τ. This metric measures the speed with which fitness values correlate with distance and is given by [229]:

$$\tau = -\frac{1}{\ln\big(|\rho(1)|\big)}, \tag{3.32}$$

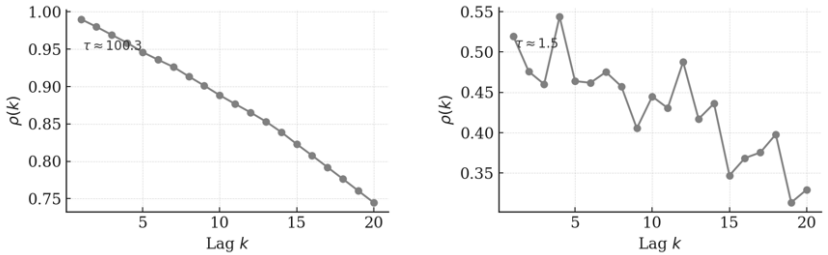


(a) High FDC (smooth landscape)

(b) Low FDC (rugged landscape)

Figure 3.11: Effect of landscape smoothness on fitness-distance correlation.

where $\rho(1)$ is the autocorrelation at a distance of one step. As illustrated in Figure 3.12, a large τ indicates a smooth landscape suitable for surrogate models, while a small τ points to a rugged landscape demanding exploration-focused algorithms.



(a) Large τ (smooth landscape)

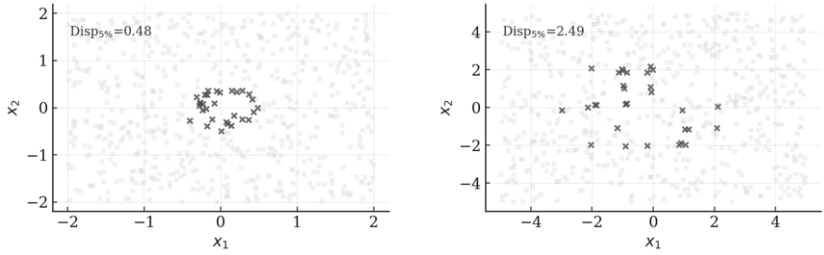
(b) Small τ (rugged landscape)

Figure 3.12: Autocorrelation profiles for two different landscapes.

To understand the modality of the problem, the dispersion metric measures how spread out the best solutions are in the decision space. For the top $\alpha\%$ of solutions, it is calculated as the average pairwise distance between them [228]:

$$\text{Disp}_\alpha = \frac{1}{\alpha n(\alpha n - 1)} \sum_{i<j} \|x_i - x_j\|. \tag{3.33}$$

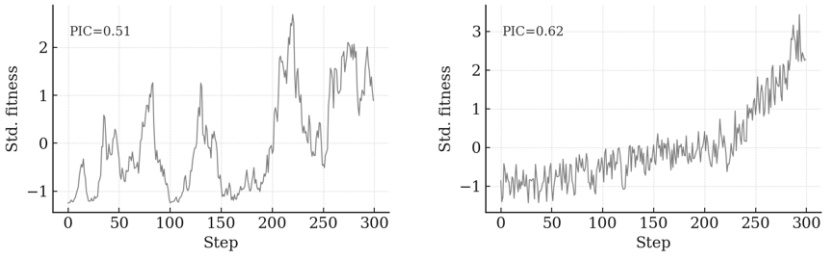
As shown in Figure 3.13, low dispersion suggests a single basin of attraction exploitable by learning-based methods, while high dispersion indicates a multimodal landscape favoring multi-start or evolutionary search.



(a) Low dispersion (single basin) **(b)** High dispersion (multimodal)

Figure 3.13: Spatial spread of the top $\alpha\%$ of solutions.

Finally, the Partial Information Content (PIC) characterizes the number of local optima by analyzing a random walk and encoding its fitness values into a symbolic string [228]. A high PIC value reflects many local optima, favoring global search, whereas a low PIC implies a smoother landscape tractable for surrogate modeling, a concept illustrated in Figure 3.14. Together, this quantitative assessment of the sufficiency of data and the characteristics of the landscape provides a principled foundation for selecting an appropriate optimization strategy, moving beyond simple trial and error.



(a) Low PIC (few optima) **(b)** High PIC (many optima)

Figure 3.14: Symbol-encoded fitness walks for landscapes with low and high modality.

3.3.3 Machine Learning Models

The selection of a predictive model is a critical step in any data-driven optimization workflow. ML models can be divided into three paradigms, as illustrated in Figure 3.15: supervised learning, unsupervised learning, and reinforcement learning [230].

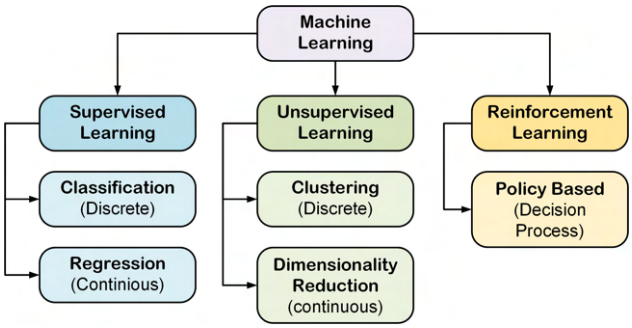


Figure 3.15: Illustration of machine Learning categorization, adapted from [230].

Although unsupervised methods focus on discovering hidden data structures and reinforcement learning targets sequential decision-making, the optimization problems central to this work are framed as prediction tasks using labeled historical data. Consequently, this review will specifically focus on supervised learning techniques.

Furthermore, manufacturing optimization problems are typically characterized by structured tabular datasets and are often constrained by limited data availability due to the high cost of physical experimentation or data acquisition. A review of recent algorithm benchmarks under these specific conditions reveals three main categories of models that consistently achieve state of the art predictive performance: tree-based ensembles, transformer-based deep learning models, and Automated Machine Learning (AutoML) frameworks. Alongside these, a fourth category of inherently interpretable models has emerged as a highly competitive alternative. These models are designed to offer a compelling

balance between high accuracy and full transparency, making them critical for applications where understanding the decision-making process is as important as the outcome itself. These model classes are summarized in Table 3.5.

Table 3.5: Top performing algorithm classes, specific algorithms, and their performance summaries on tabular datasets with fewer than 10,000 samples.

Algorithm Class	**Algorithms**	**Performance Summary**
Tree-based Ensembles	XGBoost, CatBoost, LightGBM, Random Forest (RF)	Consistently the state of the art on small-to-medium tabular datasets, typically outperforming deep learning models in benchmarks [231]. Gradient Boosting variants lead, with RF serving as a robust and powerful baseline [232].
Transformer-based Models	TabPFN, TabNet, TabTransformer, Saint	Deep learning models adapted for tabular data, particularly Transformers, have emerged as strong competitors to tree-based ensembles [233, 234]. The pre-trained TabPFN has shown exceptional performance on small datasets ($N < 1000$), often exceeding tuned ensembles with near-zero tuning cost [235].
AutoML Frameworks	AutoGluon, AutoSklearn, TPOT, H2O	Automated Machine Learning frameworks that automate the entire modeling pipeline consistently produce high-performing models. Leading frameworks such as AutoGluon and AutoSklearn are highly competitive with expert-tuned models across a wide range of benchmarks [236, 232].
Interpretable Models	EBM	Inherently interpretable models, particularly variants of Generalized Additive Models (GAMs) such as the EBM, offer a compelling trade-off. Benchmarks show EBMs achieve performance highly competitive with top-tier black-box models such as CatBoost, while providing full model transparency and editability [237].

The benchmark results indicate a competitive landscape of high-performance algorithms for small to medium-sized tabular data sets. Although the “no free

lunch" principle holds, which implies that no single algorithm is universally superior across all tasks [234], the analysis identifies a subset of model families that achieve consistently strong performance.

To provide context for these findings, the following section reviews the leading approaches by model class and begins with the primitive that underlies widely used ensemble methods. Because both RF and gradient boosting are constructed from decision trees, the discussion first recapitulates the decision tree model and its defining properties. A decision tree represents a non-parametric supervised learning approach that systematically divides the feature space $\mathbb{R}^d$ into hyperrectangles aligned with the axes [238]. At every internal node, there is a selection of a feature j along with a threshold s aimed at optimizing a purity criterion, such as minimizing squared error for regression tasks or enhancing Gini impurity for classification purposes. Formally, every split segregates the dataset into two resulting child nodes [238]:

$$\{\mathbf{x} : x_j \leq s\} \quad \text{and} \quad \{\mathbf{x} : x_j > s\}. \tag{3.34}$$

This process continues until a stopping criterion is met, with each terminal node predicting a constant value, such as the mean response. While a single, deep decision tree is a flexible model capable of capturing complex non-linear relationships, its instability and high sensitivity to the training data make it prone to overfitting. To mitigate the high variance of individual trees, bagging was introduced [239]. Bagging constructs an ensemble consisting of B independent trees, denoted as $\{h_b\}_{b=1}^{B}$, with each tree being trained on a bootstrap sample derived from the original training dataset. The final prediction is formed by taking the average of the predictions made by each tree [239]:

$$\hat{y}(\mathbf{x}) = \frac{1}{B} \sum_{b=1}^{B} h_b(\mathbf{x}). \tag{3.35}$$

This averaging process reduces the variance of the overall predictor. A key advantage of this approach is the out-of-bag (OOB) error estimate [239]. Approximately 37% of the training data is excluded from each bootstrap sample.

These OOB samples allow for the assessment of each tree's performance, offering a reliable estimation of the generalization error without the necessity of a distinct validation set [239].

The RF algorithm improves the bagging approach by adding another level of randomness, which helps to further reduce the correlation between individual trees [240]. At each potential split within each tree, the search for the best feature and threshold is restricted to a random subset of m features, where m is typically set to $\lfloor \sqrt{d} \rfloor$ for classification or $\lfloor d/3 \rfloor$ for regression. This strategy reduces the correlation between the trees in the ensemble, which in turn further reduces the variance of the averaged predictor, typically at the cost of a negligible increase in bias, resulting in a model with excellent predictive performance and robustness.

Complementary to the parallel ensemble method of bagging is boosting, which was formalized as a form of gradient-based functional optimization [241]. Instead of training independent trees, gradient boosting builds an additive model in a sequential, stage-wise fashion. The process begins with an initial constant prediction, $\hat{y}^{(0)}$, and each subsequent iteration t fits a new, typically shallow decision tree h_t to the negative gradient - or pseudo-residuals - of a differentiable loss function $L(y, \hat{y})$ with respect to the current predictions [241]:

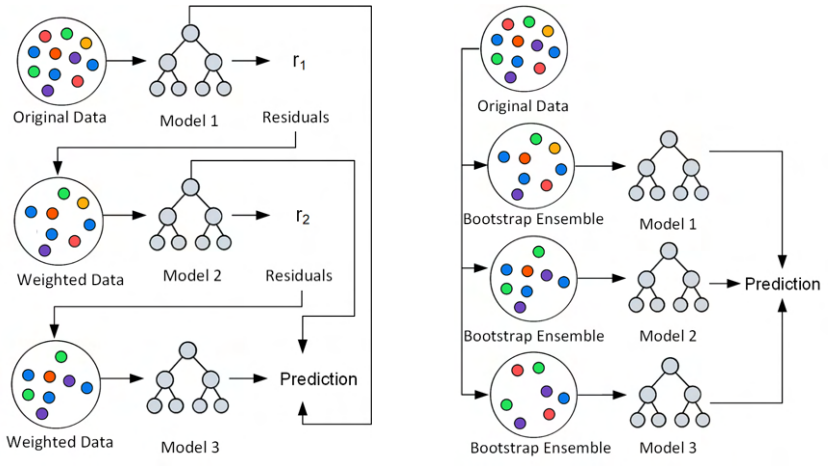
$$r_i^{(t)} = - \left. \frac{\partial L(y_i, \hat{y})}{\partial \hat{y}} \right|_{\hat{y}=\hat{y}^{(t-1)}(\mathbf{x}_i)}. \tag{3.36}$$

The full ensemble is then updated by adding the contribution of the new tree, scaled by a learning rate η [241]:

$$\hat{y}^{(t)}(\mathbf{x}) = \hat{y}^{(t-1)}(\mathbf{x}) + \eta \, h_t(\mathbf{x}). \tag{3.37}$$

The learning rate, $\eta \in (0, 1]$, is a key parameter for regularization. Smaller values of η reduce the influence of each individual tree, improving the generalization performance of the model at the cost of requiring a larger number of boost iterations to converge. This sequential process gradually corrects the errors of the previous iterations, resulting in a highly accurate predictive model.

Figure 3.16 visually summarizes the architectural differences between bagging and boosting.



(a) Boosting **(b)** Bagging

Figure 3.16: Comparison of boosting and bagging.

Modern, highly optimized implementations of gradient-boosted trees include XGBoost [242], LightGBM [243], and CatBoost [244]. Although all three share the same core sequential learning process, they introduce distinct innovations to address key challenges in scalability, regularization, and data handling. XGBoost frames each boosting iteration as the minimization of a regularized objective function. At step t, a new tree f_t is added to the model by minimizing [242]:

$$\mathcal{L}^{(t)} = \sum_{i=1}^{n} L\big(y_i, \hat{y}^{(t-1)}(\mathbf{x}_i) + f_t(\mathbf{x}_i)\big) + \Omega(f_t), \tag{3.38}$$

where the regularizer $\Omega(f) = \gamma T + \frac{\lambda}{2} \sum_{j=1}^{T} w_j^2$ penalizes both the number of leaves T and the magnitude of the leaf weights w_j. By taking a second-order Taylor expansion of the loss L, an approximate objective is obtained [242]:

$$\widetilde{\mathcal{L}}^{(t)} = \sum_{i=1}^{n} \Big[g_i f_t(\mathbf{x}_i) + \tfrac{1}{2} h_i f_t(\mathbf{x}_i)^2 \Big] + \Omega(f_t), \tag{3.39}$$

with gradients g_i and Hessians h_i evaluated at the current predictions. This formulation allows for highly efficient, parallelizable split-finding.

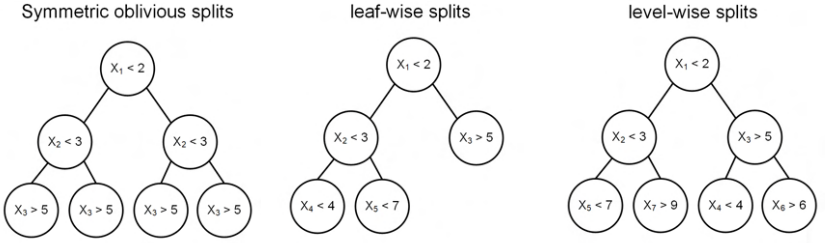
Expanding on this emphasis on efficiency, LightGBM incorporates two innovative methods for managing extremely large datasets [243]. Gradient-Based One-Side Sampling (GOSS) strategically reduces the number of instances by keeping those with significant gradients and randomly selecting from the remaining instances, thereby preserving the most informative data points for identifying splits [243]. At the same time, Exclusive Feature Bundling (EFB) reduces dimensionality by merging sparse, mutually exclusive features into single features without information loss [243].

Addressing the distinct challenge of high-cardinality categorical variables, CatBoost implements a sophisticated ordered, permutation-based target encoding [244]. For a given random permutation π of the training data, the encoding for a categorical value at position π_m is calculated using only the target values of the preceding observations in that permutation [244]:

$$\tilde{x}_{\pi_m} = \frac{\displaystyle\sum_{j=1}^{m-1} \mathbf{1}\big(x_{\pi_j} = x_{\pi_m}\big)\, y_{\pi_j} + \alpha}{\displaystyle\sum_{j=1}^{m-1} \mathbf{1}\big(x_{\pi_j} = x_{\pi_m}\big) + \alpha}, \tag{3.40}$$

where α is a smoothing prior. This method effectively leverages categorical information while preventing the target leakage common in simpler encoding schemes.

Finally, the frameworks differ in their tree growth architecture, as illustrated in Figure 3.17. XGBoost employs a level-wise strategy, growing all nodes at a given depth before proceeding deeper [242]. In contrast, LightGBM uses a leaf-wise approach that prioritizes splitting the leaf with the highest expected gain [243]. CatBoost constructs symmetric or oblivious trees, where all nodes at the same level use the same splitting criterion, a structure that is highly efficient for both CPU and GPU execution [244].



(a) CatBoost **(b)** LightGBM **(c)** XGBoost

Figure 3.17: Comparison of split structures across the three gradient-boosting frameworks.

Adapting the powerful self-attention mechanism of transformers to non-sequential, tabular data presents unique challenges, particularly the lack of inherent feature order and the quadratic complexity of the attention operation with respect to the number of features. To address this, several specialized architectures have been proposed, including TabTransformer [245], SAINT [246], and TabNet [247]. These models are built around the core self-attention layer, defined as [248]:

$$\mathrm{Attn}(Q, K, V) \;=\; \mathrm{softmax}\Big(\frac{QK^{\top}}{\sqrt{d_k}}\Big)\, V, \tag{3.41}$$

which in principle allows every feature token to attend to every other. TabTransformer approaches the problem by first creating robust feature representations. Each categorical feature is converted into a learned vector embedding, and these

embeddings are then processed by standard multi-head attention layers. It employs a self-supervised pre-training objective, where a fraction of input tokens is masked and the network is tasked with reconstructing them, thereby learning powerful, context-aware feature representations [245].

SAINT, on the contrary, modifies the attention mechanism itself by decomposing it into two distinct stages. It first applies column-wise attention to model inter-feature interactions within each sample [246]:

$$H_{\text{col}} = \text{Attn}\big(XW_Q^{\text{col}},\, XW_K^{\text{col}},\, XW_V^{\text{col}}\big), \tag{3.42}$$

followed by row-wise attention across the batch to capture inter-sample dependencies [246]:

$$H_{\text{row}} = \text{Attn}\big(SW_Q^{\text{row}},\, SW_K^{\text{row}},\, SW_V^{\text{row}}\big). \tag{3.43}$$

This decomposition effectively models complex relationships without computing a full, dense attention matrix across all features and samples simultaneously.

TabNet introduces a third paradigm, using sequential attention with a learned sparse mask to mimic the decision-making process of tree-based models [247]. At each of its decision steps, a differentiable masking mechanism, powered by a Sparsemax activation, selects a small subset of the most salient features for processing [247]:

$$p \;=\; \text{Sparsemax}(A\,h), \quad \text{Sparsemax}(z) = \arg\min_{p \in \Delta^{d-1}} \|p - z\|^2. \tag{3.44}$$

This forces the model to reason about features sequentially and sparsely, making its decision process more interpretable. These architectural innovations are critical for computational feasibility. By avoiding a full quadratic self-attention matrix, the training cost of these models can scale nearly linearly in the number of samples, making them viable alternatives to traditional ensembles on modern hardware.

Building on these prior models, TabPFN (Tabular Prior-Data Fitted Network) is a transformer-based foundation model that learns to solve arbitrary small-to-medium tabular tasks in a single forward pass by leveraging in-context learning on millions of synthetic datasets sampled from a structural causal prior [249]. During pre-training, the model parameters ϕ are optimized to minimize the expected negative log-likelihood of test set labels given the training context [249]:

$$\phi^* \;=\; \arg\min_{\phi} \; \mathbb{E}_{\substack{(X_{\text{train}}, y_{\text{train}}, \\ X_{\text{test}}, y_{\text{test}}) \sim \Pi}} \Big[-\log \; q_\phi\big(y_{\text{test}} \mid X_{\text{train}}, y_{\text{train}}, X_{\text{test}}\big) \Big], \tag{3.45}$$

where Π denotes the prior synthetic data. At inference time in a new dataset, the network ingests both labeled training rows and unlabeled test rows as a single sequence of tokens, applies alternating feature and sample self-attention blocks, and outputs a predictive distribution over each test label in one shot. Concretely, each TabPFN layer m computes [249]:

$$H^{(m+1)} \;=\; \mathrm{MLP}\Big(\mathrm{Attn}_{\text{sample}}\big(\mathrm{Attn}_{\text{feature}}(H^{(m)})\big)\Big), \tag{3.46}$$

ensuring invariance to row and column order while efficiently capturing both intrasample and intersample dependencies. By pretraining on diverse synthetic tasks, TabPFN internalizes a powerful, fully learned tabular learning algorithm that requires no gradient updates at inference and achieves state of the art accuracy on datasets up to 10,000 rows and 500 features [235].

Automated Machine Learning (AutoML) frameworks such as AutoGluon [236] and Auto-Sklearn [250] automate both pipeline selection and hyperparameter tuning by searching a joint space of feature preprocessors, learners, and their configurations to directly minimize an empirical validation loss, as illustrated in Figure 3.18.

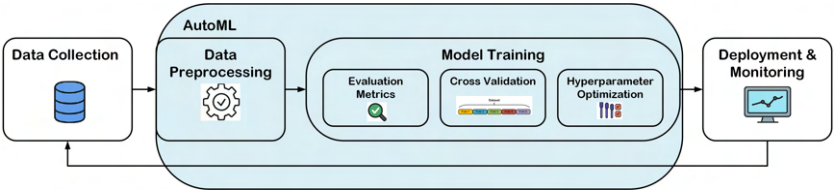


Figure 3.18: Schematic overview of an AutoML system.

The goal is to find the optimal configuration θ^* that minimizes the cross-validated loss over the space of all possible pipelines, Θ:

$$\theta^* \;=\; \arg\min_{\theta\in\Theta} \frac{1}{K}\sum_{k=1}^{K} \mathcal{L}\big(y^{(k)},\, f(x^{(k)};\theta)\big) \tag{3.47}$$

The leading frameworks employ distinct strategies to navigate this vast search space. Auto-Sklearn, for instance, utilizes meta-learning to warm-start a BO over Θ [250]. It iteratively builds a surrogate model of the validation error to intelligently select new configurations to evaluate. After identifying a set of high-performing pipelines $\{\theta_i\}_{i=1}^{M}$, it constructs a final weighted ensemble [250]:

$$F(x) \;=\; \sum_{i=1}^{M} w_i\, f\big(x;\,\theta_i\big), \qquad \sum_{i=1}^{M} w_i = 1, \tag{3.48}$$

with weights w_i typically derived from the cross-validation performance.

In contrast, AutoGluon employs a multi-layer stacking architecture [236]. It begins by training a diverse set of base learners and then iteratively creates new layers of models. Each subsequent layer l learns on an augmented feature set $z_i^{(l)}$ that includes the original features x_i plus the out-of-fold predictions from all models in the preceding layers [236]:

$$z_i^{(l)} = \big[x_i,\; \hat{y}_i^{(1)}, \ldots, \hat{y}_i^{(l-1)}\big]. \tag{3.49}$$

A final greedy selection process then chooses and weights a subset of the best models from all layers to form the final predictive ensemble [236]:

$$F(x) = \sum_{m=1}^{M} w_m \, f^{(l_m)}\big(z^{(l_m)}; \theta^{(l_m)}\big), \quad \sum_{m=1}^{M} w_m = 1. \tag{3.50}$$

Ultimately, both the BO and multi-layer stacking strategies encapsulate the end-to-end automation of the machine learning pipeline—from preprocessing to ensembling—within a single, user-friendly 'fit' operation.

Finally, a class of inherently interpretable models, GAMs, offers a powerful alternative to black-box approaches [251]. GAMs maintain transparency by expressing the expected response as an additive sum of feature-wise shape functions [251]:

$$g\big(\mathbb{E}[Y \mid \mathbf{x}]\big) = \beta_0 + \sum_{j=1}^{d} f_j(x_j), \tag{3.51}$$

where g is a link function and each f_j is a one-dimensional function that can be visualized to understand the effect of each feature on the prediction. To take advantage of this structure with the predictive power of modern machine learning, the EBM was introduced [252]. An EBM is a modern GAM implementation that learns each shape function f_j using a round-robin process with gradient boosting. For each feature, a very shallow decision tree is fit to the current residuals of the model, and its contribution is added to the corresponding shape function, scaled by a learning rate. To capture higher-order effects, this can be extended to include pairwise interaction terms [252]:

$$g\big(\mathbb{E}[Y \mid \mathbf{x}]\big) = \beta_0 + \sum_{j=1}^{d} f_j(x_j) + \sum_{j<k} f_{jk}(x_j, x_k). \tag{3.52}$$

This modular architecture is the key to the EBM's value; it achieves predictive performance that is often competitive with complex ensemble models while remaining fully interpretable, as the contribution of each individual feature and

pairwise interaction can be independently visualized and inspected by domain experts.

The achievement of optimal performance from any selected algorithm is contingent upon the systematic tuning of its hyperparameters. This process can be formalized as an optimization problem [253]:

$$x^* = \arg\min_{x \in \mathcal{X}} L\big(\mathcal{D}_{\text{val}}; \mathcal{A}(\mathcal{D}_{\text{train}}, x)\big), \tag{3.53}$$

where $\mathcal{A}(\mathcal{D}_{\text{train}}, x)$ denotes the model produced by algorithm $\mathcal{A}$ on training data $\mathcal{D}_{\text{train}}$ with hyperparameters x, $L(\mathcal{D}_{\text{val}}; \cdot)$ is the loss function evaluated on validation data $\mathcal{D}_{\text{val}}$, and $\mathcal{X}$ is the hyperparameter search space. To address the computational inefficiency of traditional hyperparameter search methods, such as grid or random search [253, 254], a suite of advanced model-based HPO frameworks has been developed to find optimal configurations with minimal evaluations. These methods accelerate the discovery of optimal configurations by intelligently sampling the search space. Frameworks such as Optuna [255], Hyperopt [256], and SMAC3 [257] operate on the principles of BO, iteratively building a probabilistic surrogate model of the objective function to guide the search towards more promising regions. They differ primarily in their choice of surrogate: Optuna and Hyperopt predominantly use a Tree-structured Parzen Estimator (TPE), while SMAC employs a RF [254]. Optuna's define by run API enables the dynamic construction of complex conditional search spaces at runtime and integrates built-in pruning strategies to stop underperforming trials based on intermediate results [255], while Hyperopt relies on a statically declared search graph and lacks native early stopping, requiring each trial to run to completion [256]. SMAC3, on the contrary, uses its RF surrogate to provide explicit uncertainty quantification and variable importance measures, and uses ensemble-based surrogate updates along with parallel trial evaluation to efficiently navigate high-dimensional and mixed-type hyperparameter spaces [257]. Empirical benchmarks demonstrate that the choice of the optimal framework is task-dependent. For complex Combined Algorithm Selection and Hyperparameter tuning (CASH) scenarios, Optuna has been shown to excel, while

Hyperopt has demonstrated superior performance in neural architecture search tasks. Despite these nuances, these modern HPO frameworks consistently provide significant performance gains over naive search strategies, achieving higher model quality with a fraction of the computational budget [254].

3.3.4 Bayesian Optimization

As described in Section 2.3.3, BO has emerged as a principal methodology for optimizing expensive-to-evaluate black-box functions. The efficacy of this sample-efficient approach hinges on the interplay of its two core components: the probabilistic surrogate model and an acquisition function. This section details the state of the art choices for these two components, beginning with an analysis of the surrogate models.

The most widely used surrogate model in BO is the GP, a non-parametric model that defines a prior distribution over functions [115]:

$$f(\mathbf{x}) \sim \mathcal{GP}\big(m(\mathbf{x}), k(\mathbf{x}, \mathbf{x}')\big), \tag{3.54}$$

where $m(\mathbf{x})$ denotes the mean function and $k(\mathbf{x}, \mathbf{x}')$ represents the covariance function, also known as the kernel function, which incorporates assumptions regarding the smoothness of the function. The behavior of the GP is governed by its kernel, which can be broadly categorized as isotropic or anisotropic. The isotropic kernel depends solely on the Euclidean distance between the points, $\|\mathbf{x} - \mathbf{x}'\|$. A frequently encountered example is the squared exponential RBF kernel [258]:

$$k_{\text{iso}}(\mathbf{x}, \mathbf{x}') = \sigma^2 \exp\left(-\frac{\|\mathbf{x} - \mathbf{x}'\|^2}{2\ell^2}\right), \tag{3.55}$$

where a single length-scale parameter, ℓ, controls the smoothness across all dimensions. For problems where input dimensions have differing sensitivities, anisotropic kernels are more appropriate. A typical anisotropic variant of the

RBF kernel introduces a diagonal matrix of length-scales, $\mathbf{L} = \text{diag}(\ell_1, \ldots, \ell_d)$, allowing the model to capture more nuanced relationships [258]:

$$k_{\text{aniso}}(\mathbf{x}, \mathbf{x}') = \sigma^2 \exp\left(-\frac{(\mathbf{x} - \mathbf{x}')^\top \mathbf{L}^{-1}(\mathbf{x} - \mathbf{x}')}{2}\right). \tag{3.56}$$

Isotropic kernels typically suffice in settings with modest dimensionality, but anisotropic kernels can capture more complex parameter interactions in real-world processes, particularly if some input dimensions have a greater influence on the objective than others [111].

Although GPs are commonly used for BO [110, 105], alternative non-parametric models, especially those grounded in tree ensembles, have demonstrated efficacy as substitutes. RF, for example, serve as surrogate models within frameworks such as Sequential Model-Based Algorithm Configuration (SMAC) [259]. Although RFs do not provide a closed-form posterior distribution such as GPs, uncertainty can be approximated by the empirical variance of the predictions across the different trees in the ensemble [260]:

$$\hat{\mu}_{\text{RF}}(\mathbf{x}) = \frac{1}{B}\sum_{b=1}^{B} f_b(\mathbf{x}), \quad \hat{\sigma}^2_{\text{RF}}(\mathbf{x}) = \frac{1}{B-1}\sum_{b=1}^{B}\big(f_b(\mathbf{x}) - \hat{\mu}_{\text{RF}}(\mathbf{x})\big)^2, \tag{3.57}$$

where $f_b(\mathbf{x})$ is the prediction of the b-th tree. A different strategy is employed by the Tree-Structured Parzen Estimator (TPE), a model-based optimization approach widely used in automated hyperparameter tuning [113]. Instead of directly modeling the conditional distribution $p(y|\mathbf{x})$, TPE inverts the generative process to the model $p(\mathbf{x}|y)$. It partitions the observed data based on a quantile γ of the objective function values into a "good" set and a "bad" set, and then fits two separate non-parametric density estimators, $l(\mathbf{x})$ and $g(\mathbf{x})$, to each set respectively. The acquisition function, which maximizes the Expected Improvement, is then proportional to the ratio of these densities. The next point to evaluate, $\mathbf{x}^*$, is chosen as [261]:

$$\mathbf{x}^* = \arg\max_{\mathbf{x}} \frac{l(\mathbf{x})}{g(\mathbf{x})}. \tag{3.58}$$

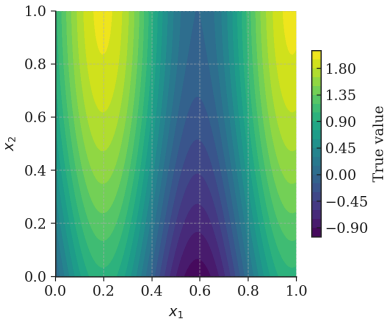
By sampling from regions where the density of high-performing points is large relative to the density of low-performing points, TPE efficiently navigates complex and conditional parameter spaces, making it a powerful alternative to kernel-based methods.

To illustrate the different trade-offs between the surrogate models, Figure 3.19 juxtaposes their posterior surfaces. The chosen ground-truth objective is a sinusoidal function, designed to test the flexibility of each model:

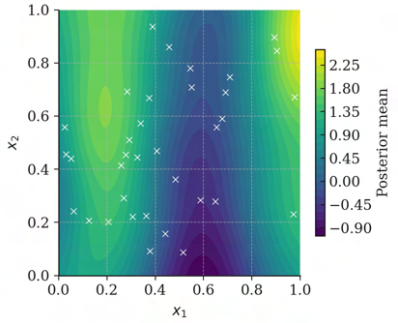
$$f(x_1, x_2) = \sin(8x_1) + \sin(1.5x_2). \tag{3.59}$$

The high frequency along the x_1 axis and the low frequency along the x_2 axis challenge each surrogate's ability to adapt to different length scales. The comparison reveals how different model assumptions interact with these landscape characteristics. The isotropic Gaussian process, constrained to a single length scale, fails to capture the rapid oscillations along the x_1 axis and consequently oversmooths the landscape. In contrast, the anisotropic GP adapts by learning separate length scales for each dimension, allowing it to accurately model the high-frequency structure in x_1 while preserving the smoothness along x_2. Its posterior mean aligns closely with the ground truth, particularly in data-rich regions. Tree-based surrogates offer different representations. The RF captures the non-stationary features without assuming global smoothness, successfully reproducing the function's overall structure. However, it introduces piecewise constant artifacts and lacks a principled, built-in measure of predictive uncertainty, which can be problematic for acquisition functions that rely on variance. The TPE visualizes a completely different concept. Its density ratio heatmap does not reconstruct the objective surface.

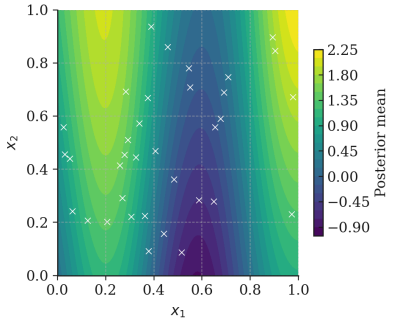
After defining a surrogate model, BO employs an acquisition function, $\alpha(\mathbf{x}; \mathcal{D})$, to strategically determine the utility of sampling a new point $\mathbf{x}$. Several strategies exist to balance the exploration-exploitation trade-off, with three being particularly common.



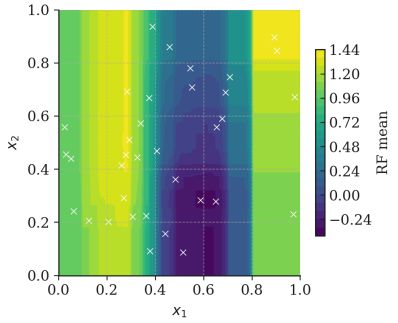
(a) Ground-truth objective



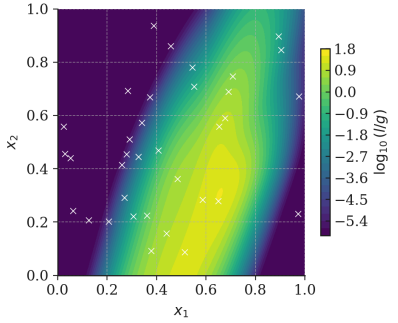
(b) Isotropic GP surrogate



(c) Anisotropic GP surrogate



(d) RF surrogate



(e) TPE surrogate

Figure 3.19: Posterior mean surfaces of four surrogate models compared with the true objective.

The most widely used is Expected Improvement (EI), which quantifies how much improvement over the current best observation, $f(\mathbf{x}^+)$, can be expected at a new point [110]:

$$\alpha_{\text{EI}}(\mathbf{x}) = \mathbb{E}\big[\max\{0, f(\mathbf{x}^+) - f(\mathbf{x})\}\big] \tag{3.60}$$

A simpler and more exploitation-focused alternative is the Probability of Improvement (POI), which calculates the probability that a new point will yield a better value than the current incumbent [105]:

$$\alpha_{\text{PoI}}(\mathbf{x}) = \Pr\big[f(\mathbf{x}) \leq f(\mathbf{x}^+)\big] = \Phi\left(\frac{f(\mathbf{x}^+) - \mu(\mathbf{x})}{\sigma(\mathbf{x})}\right) \tag{3.61}$$

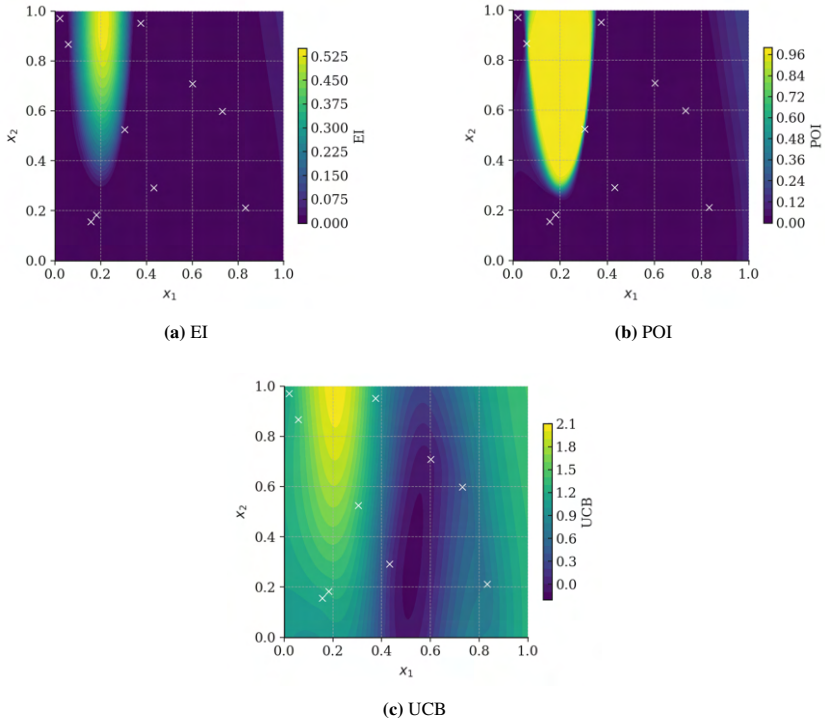
A third prominent strategy, the Upper Confidence Bound (UCB), makes the exploration-exploitation trade-off explicit through a tunable parameter κ. The function directs the search by balancing the surrogate's mean prediction, $\mu(\mathbf{x})$, with its uncertainty, $\sigma(\mathbf{x})$ [262]:

$$\alpha_{\text{UCB}}(\mathbf{x}) = \mu(\mathbf{x}) - \kappa \cdot \sigma(\mathbf{x}) \tag{3.62}$$

where a larger κ encourages more exploration by prioritizing regions of high uncertainty.

Figure 3.20 shows the three acquisition functions after fitting the same anisotropic GP to just ten training points, illustrating their distinct exploration-exploitation strategies. EI, which combines the predicted mean with uncertainty and scales the gain by the survival function, produces a tight, elliptical hotspot. This peak is centered near the best-observed region, and the function's value quickly drops to zero in areas where no substantial improvement is expected. In contrast, the POI exhibits a markedly wider area of interest. Because POI ignores the magnitude of improvement and considers only the probability of exceeding the incumbent, any point with a non-vanishing chance of being better receives a high score. This results in a less selective surface that is more broadly exploratory. The UCB, shown with an exploration factor of $\kappa = 1$, offers a tunable compromise. Its surface, defined by $\mu(\mathbf{x}) + \sigma(\mathbf{x})$, still follows the

predicted ridges of the surrogate, but the one-sigma bonus keeps the regions of high uncertainty elevated. UCB therefore provides a middle ground: it is more exploratory than the focused EI but more selective than the broad POI.



(a) EI

(b) POI

(c) UCB

Figure 3.20: Acquisition landscapes with ten training points.

Although the iterative loop of surrogate modeling and acquisition function optimization is the core of BO, its performance is influenced by the initial set of data points used to train the first surrogate model. This initial sampling phase provides the algorithm with its first view of the objective landscape. To investigate this step, a preliminary study for this dissertation was conducted by Greif et al. [3]. The study was designed to compare the impact of different initial

sampling strategies on BO performance across both benchmark mathematical functions. The following analysis details the key strategies investigated, specifically Latin Hypercube Sampling (LHS), Fractional Factorial Design (FFD), and a baseline with no structured initial sampling, and presents the main findings of this comparative study.

BO may proceed without any dedicated initial design by drawing the very first sample directly from the acquisition function. In this 'cold start' scenario, every point queried by the optimizer serves a dual purpose: it both explores unknown regions of the space and exploits promising areas indicated by the current surrogate. The practical effectiveness of this approach hinges on dynamically tuning the exploration-exploitation balance of the acquisition function, for instance by assigning a higher exploration weight during the early iterations to ensure adequate coverage. A traditional method employed is Factorial Designs, aiming to systematically investigate the extremities of the design space to assess both main effects and interactions among parameters. In a complete factorial design, represented as 2^d, each possible combination of d factors is assessed at two levels, producing experimental runs 2^d. When there are numerous factors, a FFD presents a cost-effective option by selecting a subset of these runs [82]. This reduction is achieved by confounding higher-order interactions that are assumed to be negligible, allowing the study to focus on the most significant effects with fewer trials [263]. The most common strategy in modern computer experiments is LHS, a stratified sampling method that ensures that the full range of each input parameter is evenly represented [87]. For a design with N samples and d dimensions, each dimension is first stratified into N equally probable intervals. Then, for each dimension j, a random permutation π_j of the integers $\{1, \ldots, N\}$ is generated. The i-th sample point $\mathbf{x}_i$ is then constructed such that its j-th coordinate is sampled from the interval corresponding to the i-th element of the permutation for that dimension [87]:

$$\mathbf{x}_i = (x_{i1}, x_{i2}, \ldots, x_{ik}), \quad \text{where } x_{ij} \sim U\left[\frac{\pi_j(i) - 1}{N}, \frac{\pi_j(i)}{N}\right] \tag{3.63}$$

This procedure guarantees that when the design is projected onto any single dimension, each of the N intervals contains exactly one sample point. This avoids the sample clustering common in simple random sampling and ensures a more efficient exploration of the parameter space, leading to a more accurate initial surrogate model.

The key finding from Greif et al. [3], illustrated for a representative case in Figure 3.21, is that the choice of initial design has a significant impact on optimization performance.

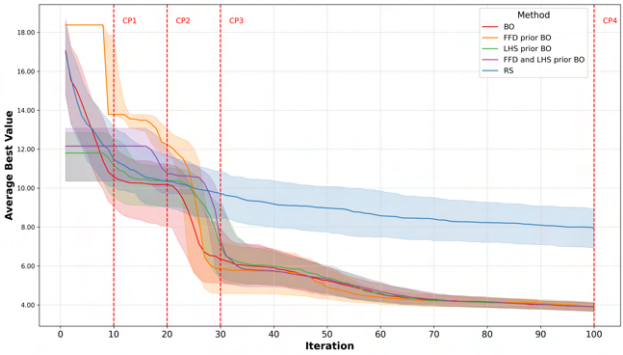


Figure 3.21: Sampling-iteration effects on performance for a representative example. Taken from Greif et al. [3]

Notably, performance discrepancies are most pronounced within the first 30 iterations, where optimization trajectories diverge significantly before converging in later stages. Although no single strategy proved universally superior, as the relative performance of methods such as LHS and FFD is highly dependent on the specific characteristics of the objective function, clear differences in convergence speed were observed. Crucially, in the vast majority of test cases, initiating the search with a structured design provided a warm start leading to statistically significant performance gains within approximately the first 30 iterations when compared to a cold start scenario with no initial sampling. However, the study also demonstrated the robustness of the BO process. Over a longer horizon, typically by the 100th iteration, the adaptive nature of the

acquisition function was able to compensate for a poor or non-existent initial design, with all strategies ultimately converging to a similar performance level. The performance difference between the strategies can be explained by their different approaches to explore the design space, as visualized in Figure 3.22.

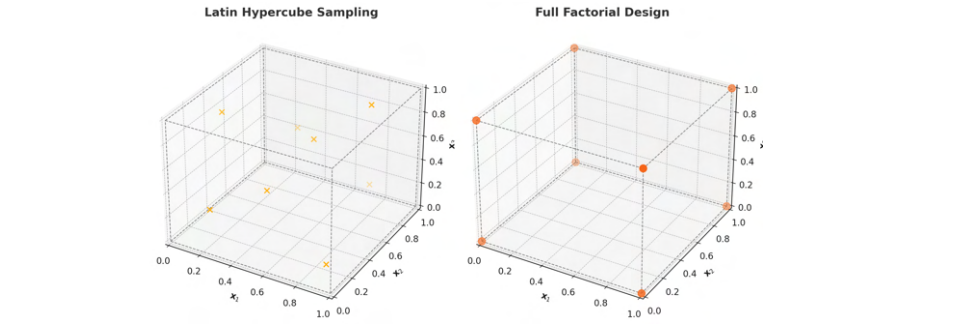


Figure 3.22: Comparison of FFD and LHS sampling strategies on the unit cube (d=3).

An FFD concentrates its sample points at the corners of the unit hypercube, systematically evaluating combinations of extreme factor levels. This approach is particularly valuable when preliminary insights, or even domain intuition, suggest that high-value regions may lie near particular boundaries or when assessing multifactor interactions is essential to refining the search domain. In contrast, LHS provides a more globally representative, space-filling coverage, ensuring that each region is sampled equally.

3.3.5 XAI and Parameter Influence Analysis

To move beyond the predictive accuracy of complex black-box models and into understanding their internal logic, a variety of XAI and parameter influence analysis techniques will be introduced. Although the number of available methods is large, they can be organized into a taxonomy based on the type of insight they provide [264], as illustrated in Figure 3.23. This taxonomy distinguishes between four primary approaches.

- Marginal Effect Plots: These techniques represent the average functional correlation between one or two attributes and the prediction made by the model [265]. They separate the impact of a specific feature by either integrating across the distribution of other attributes or by summing up local gradients, thereby uncovering the overall tendency of the feature's impact.
- Global Feature Importance: This category of methods provides a summary ranking of each feature's overall contribution to the model's predictive performance [89]. By measuring the drop in model accuracy when a feature is shuffled or by aggregating local attribution values, these techniques identify the most influential drivers of the model's decisions across the entire dataset.
- Local Additive Attributions: These approaches elucidate individual predictions by breaking them down into an aggregate of additive components, each corresponding to an input feature [266]. They respond to the inquiry, "What led to this particular prediction?" by attributing credit or criticism to each feature's value for the specific instance, thereby offering detailed, instance-specific transparency.
- Clustering and Grouping: These techniques address explanation heterogeneity by identifying subgroups within the data where the model exhibits systematically different behaviors [267]. By clustering instances based on their local explanations, these methods can uncover interaction effects or conditional relationships that are obscured by global analyses.

An established method for elucidating black-box models involves assessing the marginal impact of an input feature on the predicted result. Partial Dependence Plots (PDPs) are a primary method for this analysis, illustrating how the model output changes, on average, as one or two features are varied, while marginalizing the influence of all other features [241]. Given a trained model $\hat{f}(\mathbf{x})$, the partial dependence function for a subset of features S is defined as the expected

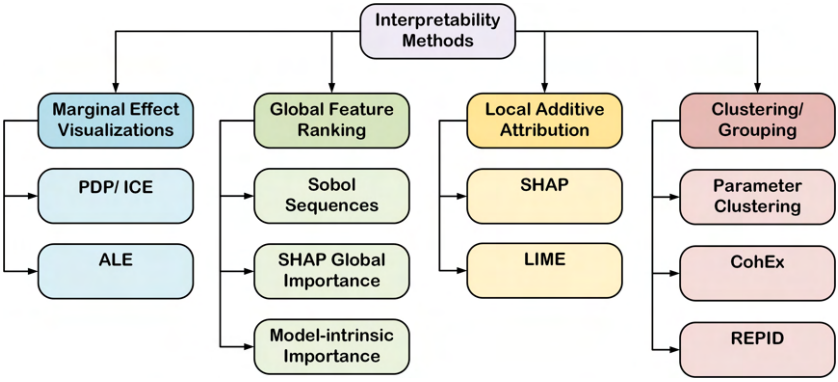


Figure 3.23: Taxonomy of interpretability methods.

value of the model output, where the expectation is taken over the marginal distribution of the complementary features, C [241]:

$$\hat{f}_S(\mathbf{x}_S) = \mathbb{E}_{\mathbf{X}_C}\big[\, \hat{f}(\mathbf{x}_S, \mathbf{X}_C) \,\big] = \int \hat{f}(\mathbf{x}_S, \mathbf{x}_C)\, p_C(\mathbf{x}_C)\, d\mathbf{x}_C.$$

In practice, this integral is intractable and is estimated using a Monte Carlo approach by averaging the model's predictions over the empirical distribution of the training data. For a single feature x_j, the partial dependence at a specific value v is thus approximated by [241]:

$$\hat{f}_{\{j\}}(v) \approx \frac{1}{N} \sum_{i=1}^{N} \hat{f}(x_j = v, \mathbf{x}_C^{(i)}),$$

where $\mathbf{x}_C^{(i)}$ represents the values of all other features for the i-th observation in the data set. By plotting the estimated $\hat{f}_{\{j\}}(v)$ against different values of v, the resulting curve visualizes the average marginal effect of that feature on the output of the model. This can reveal whether a feature has a positive or negative influence and whether its effect is linear, monotonic, or exhibits more complex patterns such as thresholds or saturation points. Figure 3.24 shows a simple

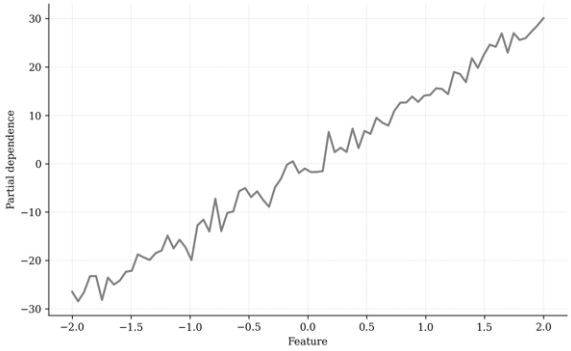


Figure 3.24: Example of a PDP Plot.

example of a PDP plot in which a roughly linear positive influence, the higher values of the feature increase the predicted output of the model.

While PDPs offer an intuitive global summary, their reliance on averaging can be a limitation. By presenting only the mean effect, they may obscure heterogeneous relationships or strong interaction effects present in the data. To address this, individual conditional expectation (ICE) graphs disaggregate this average by displaying a separate prediction curve for each individual data instance [265]. For a given instance i, an ICE curve is produced by altering the feature of interest, x_j, and keeping all other features, $\mathbf{x}_C^{(i)}$, fixed at their observed values. The resulting bundle of curves reveals the conditional relationships specific to the instance. The PDP is simply the mean of all individual ICE curves. The primary utility of ICE plots comes from analyzing the variation within this bundle. If all curves are roughly parallel, this suggests a homogeneous feature effect with minimal interactions, which means that the PDP is a reliable summary for the entire data set. In contrast, if the curves diverge, cross, or form distinct clusters, there is strong evidence of interaction effects, indicating that the influence of x_j depends on the values of other features. In such cases, the PDP alone would be misleading, whereas the ICE plot correctly reveals the underlying heterogeneity in the model's behavior. The key insight is gained by analyzing the geometry of the bundle. If the curves are parallel, it indicates

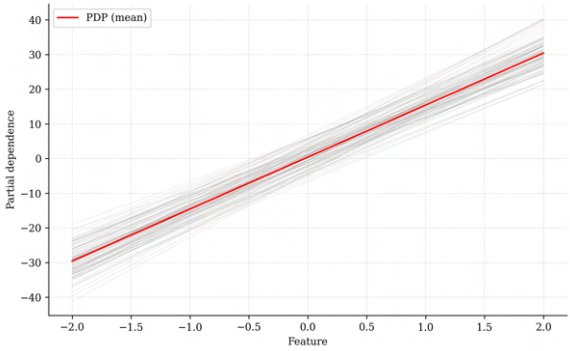


Figure 3.25: Example of an ICE Plot.

the absence of interaction effects, meaning the feature's influence is primarily a homogeneous main effect and the PDP is a reliable summary. However, if the curves are non-parallel, cross, or form distinct clusters, it provides strong evidence of higher-order interaction effects. This heterogeneity signifies that the PDP's global average masks the true conditional nature of the feature's influence, and the model's behavior can only be understood at a local or subgroup level. An example is shown in Figure 3.25. The plot shows the individual ICE curves for a subset of data instances, with their average, the PDP curve, overlaid as a bold red line.

In engineering contexts, marginal effect plots such as PDP and ICE plots serve as powerful tools for model validation and knowledge discovery. A primary application is to verify that the learned relationships of a model align with established domain knowledge. For example, in a study on injection molding quality control, PDPs visualized the marginal effects of machine settings such as injection pressure and melt temperature on the predicted defect rate, supporting the physical plausibility of the learned relationships [268]. The same study used ICE plots to examine heterogeneity at the instance level, delineating sample-specific control ranges and revealing interaction effects that are obscured by PDP averages [268].

While marginal effect plots visualize the shape of a feature's influence, a complementary global approach, Global Sensitivity Analysis, quantifies the magnitude of each input's contribution to the model's output variance. Global sensitivity analysis evaluates how each input variable affects the variability of the model's output, considering the model as a function $y = f(x_1, \ldots, x_d)$, where there are d inputs. The most prominent variance-based method is the calculation of Sobol indices [89]. This technique provides a detailed and interpretable decomposition, estimating first-order indices that represent the fraction of variance caused by a single feature acting alone, as well as total-effect indices that account for the variance caused by a feature's interactions with all other inputs. The core principle of variance-based sensitivity analysis is the decomposition of the total variance of the model output, $\mathrm{Var}(Y)$, into fractions that can be attributed to each input feature and their interactions. The Sobol method, based on the Hoeffding-Sobol decomposition, provides a formal way to achieve this [89]. The most fundamental measure is the first-order Sobol index, S_i, which quantifies the main effect of an input X_i. It represents the fraction of output variance explained by varying X_i alone, while averaging the uncertainty of all other inputs [89]:

$$S_i = \frac{\mathrm{Var}\big(\mathbb{E}[Y \mid X_i]\big)}{\mathrm{Var}(Y)}. \tag{3.64}$$

A large S_i indicates that X_i is highly influential on its own. After accounting for all first-order effects, the remaining variance is due to interactions between variables. The second order index, S_{ij}, measures the variance contribution from the interaction between X_i and X_j that is not captured by their individual main effects [89]:

$$S_{ij} = \frac{\mathrm{Var}(\mathbb{E}[Y \mid X_i, X_j]) - \mathrm{Var}(\mathbb{E}[Y \mid X_i]) - \mathrm{Var}(\mathbb{E}[Y \mid X_j])}{\mathrm{Var}(Y)}. \tag{3.65}$$

This concept extends to higher-order interactions among larger subsets of features. For practical purposes, the most comprehensive and useful measure is often the total-effect index, S_{T_i}. It captures the full contribution of an input

X_i, including its first-order effect and all higher-order interactions in which it is involved. It is efficiently calculated as [89]:

$$S_{T_i} = 1 - \frac{\mathrm{Var}\big(\mathbb{E}[Y \mid X_{\sim i}]\big)}{\mathrm{Var}(Y)}, \tag{3.66}$$

where $X_{\sim i}$ denotes the vector of all input features except X_i. Intuitively, S_{T_i} quantifies the expected output variance that would remain unexplained if the value of X_i were known. Therefore, a value of S_{T_i} close to zero indicates that X_i is a non-influential feature and could potentially be fixed or removed from the model without loss of information.

The estimation of Sobol indices for a black-box model is typically performed using Monte Carlo or quasi-Monte Carlo methods. These approaches involve evaluating the model f at many thousands of input points, sampled from their respective distributions, to numerically approximate the required expectations and variances. Efficient algorithms, often based on clever resampling techniques, have been developed to compute first-order (S_i) and total effect (S_{T_i}) indices with reasonable computational cost [269]. The output of the analysis is a set of indices that fully decomposes the model's output variance. The sum of the first-order indices, $\sum_i S_i$, is less than or equal to one, with the remainder attributed to the cumulative effect of all higher-order interactions. From an interpretability perspective, this variance decomposition provides a powerful and intuitive global summary of the model's behavior. The first-order indices, S_i, rank the input features by their direct influence, allowing engineers to identify the primary drivers of output variance. For example, in a thermal simulation model, Sobol analysis might reveal that material thermal conductivity and ambient temperature are the dominant factors, while other parameters have negligible main effects. Furthermore, the difference between a feature's total effect index and its first-order index ($S_{T_i} - S_i$) quantifies its involvement in interactions. A large gap indicates that a feature's influence is highly context-dependent and coupled with other variables, providing insights for robust design. Figure 3.26

provides an example visualization of these results, allowing a direct comparison of the main effect of each feature versus its total contribution, including all interactions.

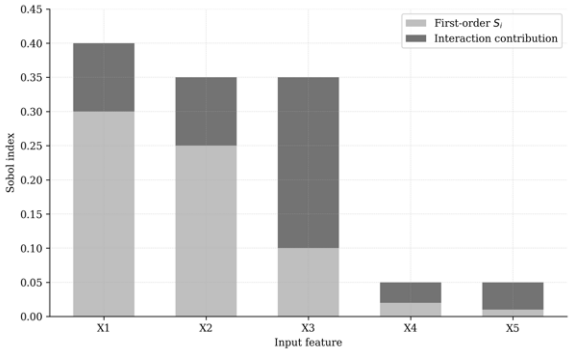


Figure 3.26: First-order and total Sobol indices of an example model.

Figure 3.26 shows an example of a synthetic model. The bar chart decomposes the total Sobol index of each input X_i into its first-order contribution S_i and the residual interaction contribution $S_{T_i} - S_i$. The plot reveals that X_1 and X_2 dominate the variance of the output through strong main effects, while X_3 shows only a modest main effect but a substantial share of interaction, indicating that its influence arises primarily through synergistic terms with other variables. In contrast, X_4 and X_5 show negligible indices, suggesting that they could be fixed or removed without materially affecting the model variance. However, it is important to acknowledge the assumptions and limitations of this method. The resulting Sobol indices are conditional on the chosen probability distributions for the input features, and their values can change if these distributions are modified. The standard formulation also assumes that the model is a deterministic function. Evaluating models exhibiting intrinsic stochasticity necessitates extensions to the methodology. Additionally, determining the indices can pose computational challenges, since the volume of model simulations needed for reliable estimates may escalate swiftly with increasing input dimensions. In engineering design and process control, a primary application of Sobol analysis

is to identify and rank the most important parameters that govern a system's behavior.

For example, in a lifecycle assessment of an AM process, first-order Sobol indices were used to rank input parameters according to their contribution to the variance in environmental impact [270]. Similarly, for an injection molding simulation, the Sobol analysis of a surrogate model quantitatively identified which process and material parameters explained the largest share of variance in pressure output, thus informing quality control strategies [271]. Beyond ranking these individual main effects, Sobol analysis is crucial for uncovering and quantifying interaction effects. By computing total-effect indices, which capture a feature's main effect plus all variance from its interactions, engineers can gain a more complete picture of the system. A large gap between a feature's total effect index (S_{T_i}) and its first-order index (S_i) is a clear indicator that the feature's influence is highly context-dependent and coupled with other variables, an insight that is critical for robust design. Furthermore, this detailed variance decomposition serves as a powerful tool for model validation and debugging. If a feature known to be physically critical exhibits a low Sobol index, it may indicate a flaw in the model's learned relationships or highlight that the training data did not cover the feature's influential range. Conversely, if an ostensibly minor feature displays high sensitivity, it prompts necessary scrutiny into whether this represents a genuine physical effect or a spurious artifact learned by the model. This allows Global Sensitivity Analysis to be used as a "sanity check" to ensure the model's behavior aligns with domain knowledge. Ultimately, Sobol sensitivity analysis provides a rigorous, quantitative framework for decomposing model uncertainty, moving beyond simple feature rankings to offer a detailed, global perspective on what drives the variability of a model's output [89, 269].

While global methods such as PDPs and sensitivity analysis explain a model's average behavior, a common need in engineering is to understand a specific prediction: "Why was this outcome predicted for this particular input?" SHAP is a widely recognized model-agnostic method that tackles the problem by offering localized explanations rooted in cooperative game theory. In the context of a specific prediction, SHAP allocates an importance score to each feature,

reflecting the degree to which the feature's value influenced the deviation of the prediction from a baseline expectation [266]. The fundamental concept involves considering a model's features as participants in a game, with the aim of equitably allocating the 'payout'. This payout represents the discrepancy between a particular prediction $f(\mathbf{x})$ and a baseline expectation, denoted as $y_{\text{base}} = \mathbb{E}[f(\mathbf{X})]$, among these features. For each feature i, SHAP computes a value, ϕ_i, which quantifies its role in the deviation of the prediction, $\Delta y = f(\mathbf{x}) - y_{\text{base}}$. These contributions, called Shapley values, are determined by taking the mean of the marginal contribution of a feature over all possible combinations of the remaining features. The definition of the Shapley value for feature i is [266]:

$$\phi_i = \sum_{S \subseteq \{1,\ldots,d\} \setminus \{i\}} \frac{|S|!\,(d - |S| - 1)!}{d!} \Big(f_{S \cup \{i\}}(\mathbf{x}) - f_S(\mathbf{x}) \Big), \tag{3.67}$$

where the term $f_{S \cup \{i\}}(\mathbf{x}) - f_S(\mathbf{x})$ represents the additional value caused by the feature i when added to a coalition of features S. The formula computes ϕ_i as the weighted average of this marginal contribution on every possible coalition S that does not include i. This formulation is distinctive because it stands as the sole attribution approach that concurrently meets a range of desirable theoretical characteristics. Most importantly, it guarantees efficiency, where the individual feature attributions, ϕ_i, sum exactly to the total prediction deviation. This allows any prediction to be expressed as a simple additive explanation model: $y = y_{\text{base}} + \sum_i \phi_i$.Furthermore, the formulation guarantees symmetry by allocating equal significance to features that have an identical contribution to the model's output, while also applying the null effect by attributing zero significance to features without any influence. Although the exact calculation of Shapley values is NP-hard due to its exponential complexity, the SHAP framework provides highly efficient approximation algorithms, such as TreeSHAP for tree-based ensembles and KernelSHAP as a model-agnostic alternative [266]. The set of Shapley values, $\{\phi_i\}$, for a single prediction can be visualized to provide a clear and additive explanation. A common visualization is the SHAP decision plot, shown in Figure 3.27. This graph illustrates the cumulative trajec-

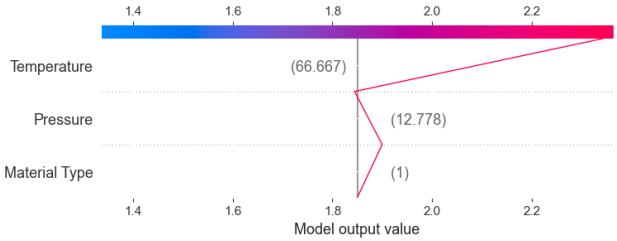


Figure 3.27: Example SHAP Decision Plot.

tory of the SHAP values. It begins with the baseline expectation of the model, denoted y_{base}, and gradually incorporates the contribution of each individual feature until it reaches the final prediction $f(\mathbf{x})$. The plot makes it immediately clear not only the magnitude and sign of each feature's contribution but also their hierarchical impact. In the example, the material type feature provides a minor positive contribution, shifting the prediction slightly from the baseline, which is largely offset by the subsequent negative impact of the pressure. The final, largest positive influence of temperature brings the path to the final output value. Aggregating SHAP values across the complete dataset advances our understanding of the model's global behavior beyond merely interpreting individual predictions. The global significance of a feature can be quantified by calculating the mean absolute value of its SHAP estimates across all instances. In this context, the SHAP beeswarm plot, as illustrated in Figure 3.28, is used. This figure illustrates the distribution of SHAP values across various features,

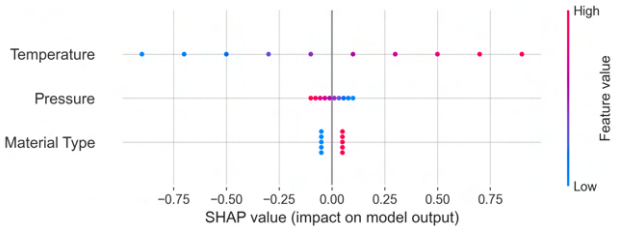


Figure 3.28: Example of a SHAP Beeswarm Plot.

with each point symbolizing an individual instance. The horizontal location of the dot signifies the influence of the respective feature on the prediction, and its color reflects the original value of the feature. This visualization reveals complex relationships: the example shows a clear monotonic trend for Temperature, where low values have a negative impact and high values have a positive impact. Pressure shows an inverse relationship, while the clustered nature of Material Type's SHAP values suggests a non-linear, categorical effect. This dual local-global capability makes SHAP a highly versatile tool in engineering. Analysts can use a decision plot to diagnose a single surprising prediction and then use a summary plot to validate that the model's overall logic aligns with domain knowledge. However, it is critical to acknowledge a key methodological caveat: Standard efficient implementations such as TreeSHAP are based on the assumption that the features are independent. In the presence of highly correlated features, this can lead the model to attribute importance to unrealistic data combinations, potentially producing misleading explanations. Addressing this requires more computationally intensive methods that account for feature dependencies. Despite this limitation, the strong theoretical properties and the intuitive and additive nature of the explanations have established SHAP as a foundational method for the interpretability of the local model [266]. In scientific and engineering applications, SHAP is a primary use for model validation and knowledge discovery. For example, in materials engineering, a study that predicted the fatigue strength of a steel alloy used SHAP to interpret the results [272]. The analysis confirmed that the model's learned relationships were consistent with established metallurgical principles, such as specific tempering temperatures and alloying elements that improve durability, and provided quantitative insight into how much each factor contributed to individual predictions. This demonstrates how SHAP can help verify that a black-box model is learning physically plausible patterns. Beyond model validation, SHAP is a powerful tool for process diagnostics and control. SHAP was employed in an injection molding process to analyze the influence of transient process parameter variations on the ultimate quality of the component [273]. By visualizing SHAP value distributions, engineers could identify which segments of the process,

such as the injection speed profile or cooling time, were most critical in directing the model to predict defects, thereby pinpointing areas for adjustment. This diagnostic capability extends to broader analyses, such as interpreting the importance of features in industry-wide productivity models [274].

Although the preceding methods typically analyze model behavior one feature at a time, a complementary approach seeks to understand the model by identifying distinct behavioral patterns across the entire dataset. The foundational concept stipulates that a globally intricate black-box model can frequently be represented as an ensemble of simpler, easier-to-understand local models, with each one aligning with a particular operational condition or data subgroup. A straightforward application of clustering for model interpretation is to analyze the model's output space. For a model that produces complex, high-dimensional outputs, this involves partitioning the set of output vectors into a predefined number of clusters, K. Each resulting cluster can be interpreted as a different operational mode or type of system failure [275]. By examining the input features that are characteristic of each output cluster, engineers can infer the conditions that lead to specific types of system behavior, which is invaluable for design optimization and scenario analysis. A more sophisticated strategy is to cluster the explanation space, which groups the data instances according to how the model arrived at its predictions [276]. After computing the local attributions for many instances, each instance is represented by its SHAP vector, $\boldsymbol{\phi}^{(n)}$ [276]. Clustering these vectors in the explanation space can uncover different decision-making regimes within the model [277]. A focused variant of this approach is clustering of ICE curves. By grouping curves with similar shapes using an appropriate distance metric, this technique can automatically detect and visualize interaction effects [265]. The result might reveal that a feature has a positive influence for one subgroup of the data but a negative influence for another, providing insights hidden by global analyses.

Figure 3.29 provides an illustration of this process. In the figure, individual data instances are represented by their SHAP values and projected onto a two-dimensional explanation space. The clustering algorithm automatically

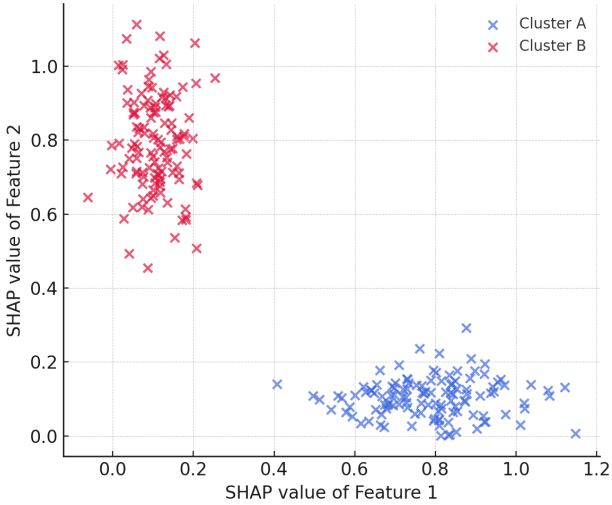


Figure 3.29: Example of clustering instances in explanation space, with each point projected onto its two most important SHAP dimensions.

discovers two well-separated groups: Cluster A, where feature 1 has a high positive contribution, and cluster B, where feature 2 is dominant.

Clustering algorithms vary widely, from straightforward partitioning techniques to more adaptable probabilistic approaches. First, the k-means algorithm functions by iteratively assigning each of the N data points to one of the predefined clusters k [278]. The objective of the algorithm is to identify the set of clusters $C = \{C_1, \ldots, C_k\}$ that reduces the total sum within the cluster of the squared euclidean distances to the centroids of the cluster $\boldsymbol{\mu}_k$ [278]:

$$\arg\min_{C} \sum_{k=1}^{k} \sum_{\mathbf{x}_i \in C_k} \|\mathbf{x}_i - \boldsymbol{\mu}_k\|^2. \tag{3.68}$$

Due to its computational efficiency and the straightforward interpretability of its results, k-means is one of the most widely used clustering approaches in the manufacturing domain [279]. It is often used as a foundational step in larger

data analysis pipelines. For example, a study used k-means to create an initial structured data model from raw machine sensor readings, which was then passed to downstream data mining and scheduling algorithms [275]. However, despite its widespread use, the reliance of k-means on the euclidean distance assumes that clusters are spherical, and its performance can be sensitive to the initial placement of centroids. As confirmed by recent reviews, these assumptions can be a limitation when faced with more complex data distributions [279].

An alternative approach is hierarchical clustering, which builds a nested tree of clusters without requiring the number of clusters to be specified in advance. The widely used form, agglomerative clustering, initiates the process by considering each data point as an individual cluster, progressively merging the nearest cluster pairs [280]. The methodology for measuring inter-cluster distance is guided by a linkage criterion. Among the prevalent criteria are single linkage, complete linkage, and Ward's method, which aims to minimize the overall growth in within-cluster variance. In the case of single linkage, the distance between two clusters A and B is defined as [280]:

$$d_{\text{single}}(A, B) = \min_{x \in A,\, y \in B} d(x, y), \tag{3.69}$$

where $d(x, y)$ denotes the distance between two points x and y. For complete linkage, the distance is defined as [280]:

$$d_{\text{complete}}(A, B) = \max_{x \in A,\, y \in B} d(x, y). \tag{3.70}$$

Ward's method merges the two clusters that result in the minimum increase in total variance within the cluster [281]:

$$\Delta(A, B) = \frac{|A||B|}{|A| + |B|} \|\bar{x}_A - \bar{x}_B\|^2, \tag{3.71}$$

where $|A|$ and $|B|$ are the number of points in clusters A and B, and $\bar{x}_A$ and $\bar{x}_B$ are their respective centroids. This flexibility makes the method highly suitable for manufacturing applications where natural groupings exist but are

not known beforehand. For example, hierarchical clustering has been used to group product CAD geometries to automate the selection of the most appropriate manufacturing technology, showing a high degree of agreement up to 90% with expert evaluations [282]. In production optimization, it has also been applied to tool planning in flexible manufacturing systems, where jobs are grouped by their tooling requirements to minimize costly machine changeovers [283].

A more powerful and flexible approach is offered by Gaussian Mixture Models (GMMs), which models the data as a weighted sum of multivariate Gaussian distributions [284]. Unlike k-means, which is based on euclidean proximity and assumes spherical clusters, a GMM allows for elliptical cluster shapes, provides soft probabilistic assignments, and explicitly models the covariance structure of the data. GMM defines a density function over the input space as a linear superposition of K Gaussian components [284]:

$$p(\mathbf{x}) = \sum_{k=1}^{K} \pi_k \, \mathcal{N}(\mathbf{x} \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k), \tag{3.72}$$

where π_k are the mixing coefficients that sum to one, and each $\mathcal{N}(\cdot)$ is a multivariate Gaussian with a mean vector $\boldsymbol{\mu}_k$ and a covariance matrix $\boldsymbol{\Sigma}_k$. The model parameters $\{\pi_k, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\}_{k=1}^{K}$ are estimated using the EM algorithm, which was already described in Equation 3.15. A study in tool wear monitoring demonstrates the practical power of this approach [285]. In this work, physically motivated features were extracted from high-frequency machine force sensor data. The GMM was then applied to cluster these features in an unsupervised manner to identify different tool wear states. The GMM achieved a classification accuracy of approximately 96% in predicting the current wear phase, a result validated on both proprietary milling experiments and a public benchmark dataset. The automatically discovered clusters corresponded directly to the physical life phases of the tool, such as initial use, uniform wear, and accelerated end-of-life wear. The quality of a clustering solution can be assessed by measures that balance cohesion and separation or penalize model complexity, among which the Silhouette coefficient and the Bayesian Information Criterion

(BIC) are widely used. The Silhouette coefficient quantifies for each object i the degree to which it lies closer to its own cluster than to the nearest neighboring cluster [267]. The calculation involves determining $a(i)$, the mean distance between point i and all other points within its cluster, alongside $b(i)$, which represents the smallest mean distance from i to any different cluster. The silhouette value is subsequently defined by [267]:

$$s(i) \;=\; \frac{b(i) - a(i)}{\max\{a(i),\, b(i)\}}, \quad -1 \le s(i) \le 1. \tag{3.73}$$

Values of $s(i)$ near unity indicate that the point is well matched to its cluster and poorly matched to neighboring clusters; values near zero signify ambiguous assignments at cluster boundaries; and negative values suggest possible misclassification. By averaging these individual scores over all n data points, one obtains the mean Silhouette coefficient [267]:

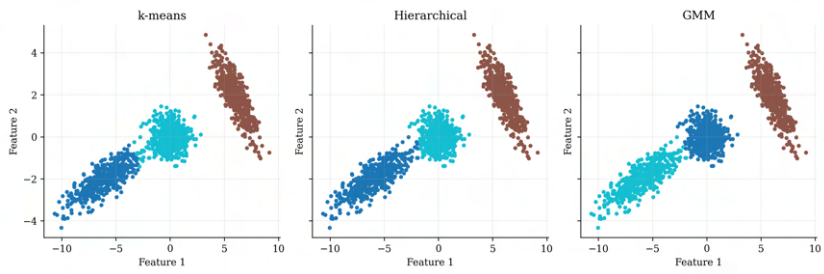
$$\bar{s} \;=\; \frac{1}{n} \sum_{i=1}^{n} s(i), \tag{3.74}$$

where higher values reflect more coherent and well-separated clusters. In contrast, the BIC provides a likelihood-based approach to model selection that incorporates a penalty for the number of free parameters [286]. Given a statistical model fitted to data X with maximized likelihood $L(\hat{\theta} \mid X)$, and p estimated parameters, the BIC is defined as [286]:

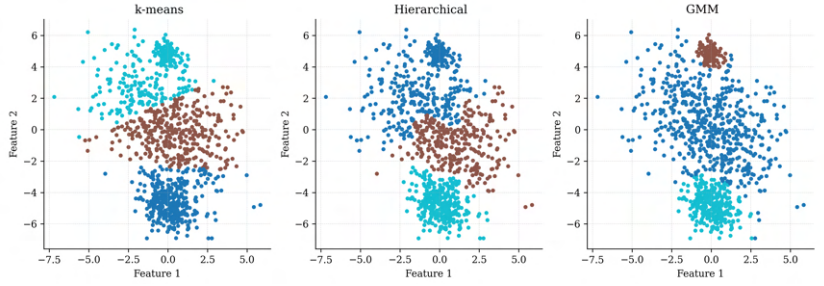
$$\mathrm{BIC} \;=\; -2\, \ln\!\big(L(\hat{\theta} \mid X)\big) \;+\; p\, \ln(n), \tag{3.75}$$

where n denotes the number of observations. In the context of Gaussian mixture models, p includes the weights, means, and covariances of the k components. A lower BIC indicates a preferable trade-off between goodness of fit and model complexity. To determine the optimal number of clusters k, one may therefore compute $\bar{s}(k)$ and $\mathrm{BIC}(k)$ across a range of candidate values. The Silhouette approach selects the value of k that maximizes $\bar{s}$, highlighting the clustering with the highest average cohesion and separation. The BIC approach selects the

value of k associated with the minimum BIC, favoring parsimonious models that explain the data well. When both criteria point to the same k, confidence in the choice increases, as it reflects both a clear geometric separation in the feature space and an optimal balance between statistical fit and complexity. In Figure 3.30 the three clustering methods are applied to two synthetic 2D data sets. In panel (a), where the clusters are roughly spherical and well separated, the three methods recover the true groupings equally well. However, in panel (b), the data exhibit noise, anisotropy, and uneven densities. Here, k-means and hierarchical clustering produce fragmented or misaligned clusters, while the GMM captures the underlying structure most accurately.



(a) Simple, well-separated data points



(b) Noisy, anisotropic data points

Figure 3.30: Comparison of cluster algorithms.

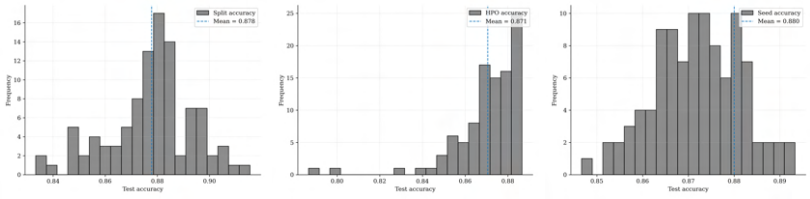
3.3.6 Robustness Testing

The robustness of a machine learning model refers to its ability to maintain consistent and reliable performance despite minor changes in training conditions or input data. Therefore, a robustness check is a fundamental component of model evaluation to ensure the generalizability and trustworthiness of the results [287].

A primary factor that impairs the reproducibility of model results is stochastic variance, which arises from random processes during training. These processes include random initialization of weights, random ordering of training data, and random splitting into training and test sets. The significant impact of this randomness on model performance has been demonstrated in multiple studies. For example, [287] showed that simply changing the random seed can affect the consistency of the models with respect to both precision and attribution of features. As a countermeasure, an averaging of model weights from multiple runs was proposed, which was found to reduce performance variance by up to 72% [287]. In this approach, the random seed was treated as a tunable hyperparameter and it was confirmed that averaging weights from multiple training runs with different seeds significantly stabilizes the results. The selection of the final model after an extensive hyperparameter search represents another potential source of insufficient robustness. Selecting the single best-performing run carries the risk that it is a statistical outlier whose high performance is not reproducible. To improve reproducibility, approaches combine models that exhibit diversity in both their weights and their hyperparameters. It is reported by [288] that a so-called hyperdeep ensemble composed of models from a random hyperparameter search achieves state of the art accuracy and good uncertainty calibration. This approach outperforms the conventional optimization of a single model. Similarly, [134] emphasizes that to trust performance gains, training must be repeated in various sources of variation, as these factors can significantly impact the results of comparisons.

To demonstrate these effects, Figure 3.31 presents the influence of stochastic variance on a synthetic binary classification task. The data set comprises 1,500

rows and 20 features, of which 15 are informative. Two features are linear combinations of these informative features, making them redundant and correlated, and the remaining three features introduce noise. The class distribution is relatively balanced. An RF classifier was trained independently for each source of variance. During training, one source of variance was altered while the other two were kept constant. This experiment was carried out 100 times per source, and the histograms depict the variability in accuracy.



(a) Data split-variance. **(b)** Hyperparameter optimization variance. **(c)** Model-seed variance.

Figure 3.31: Histogram comparison of stochastic variance.

The robustness of a model can also be compromised by the strategies chosen to handle outliers or missing values. Especially with complex or incomplete datasets, dedicated tests are essential to evaluate sensitivity to these preprocessing steps. The use of multiple methods to address missing data is advised by [289] in order to facilitate a comparison of the results. Moreover, it is emphasized that identifying and addressing outliers plays a vital role in the analysis of gathered data. The impact of these decisions can be quantified. For example, the effect of different imputation strategies on a medical prediction model was compared by [290]. This comparison revealed small but non-negligible differences in the R^2 metric. It was concluded that selecting the most appropriate imputation method requires an understanding of the missing data mechanism and the objectives at hand.

When model interpretations serve as the basis for decisions, it is important to verify whether different explanation techniques provide a coherent picture. Recent work by [291] has highlighted the disagreement problem in XAI. It

describes how popular post-hoc explanation methods often produce different rankings of feature importance. This can lead practitioners to unknowingly rely on misleading or contradictory explanations. In a user study, it was reported that 84% of data scientists have encountered such inconsistencies, often resorting to ad hoc heuristics to resolve them [291]. To address this issue, [291] proposed using agreement metrics to measure the consensus between explanations. Should the agreement between two valid explanation methods drop below a specified threshold, it may suggest problems such as collinearity or data leakage.

To perform a final robustness assessment, examining the learning curves for both training and validation is essential to detect overfitting. Overfitting is indicated when the generalization gap, that is, the disparity between the training error and the validation error, substantially exceeds the intrinsic variability of the validation error at the end of the training process [292].

3.4 Research Gap

The preceding state of the art review reveals that while data-driven optimization is a rapidly advancing field, its application to sustainable manufacturing is marked by several critical and persistent gaps. The analysis of both academic literature and industrial practice highlights a disconnect between the potential of advanced algorithms and the practical constraints of real-world manufacturing environments. This dissertation is motivated by three primary deficiencies identified in the current landscape. A significant portion of the academic literature, particularly studies focused on deep learning or complex simulations, implicitly assumes the availability of large, clean datasets. This often fails to account for the common industrial setting in which optimization is constrained by small, costly, and error-prone data sets. The high cost of physical experimentation or the difficulty of instrumenting legacy equipment means that many real-world manufacturing problems are data-scarce. Adding to this issue is the widespread lack of methodological rigor in data preparation. As evidenced by the review,

many studies either omit preprocessing steps entirely, especially those based on simulations where data is clean by design, or fail to report them in sufficient detail. Generic descriptions such as data cleaning without specifying the algorithms, parameters, or thresholds used are common. This gap undermines reproducibility and ignores the fact that neglecting these foundational steps can severely degrade the accuracy of the model and compromise the reliability of optimization results. The review demonstrates a clear and persistent divide between achieving high predictive or optimization performance and ensuring that the resulting models are transparent and trustworthy. Across both academia and industry, explainability is treated as a secondary concern, rather than a core design requirement of the optimization system. The analysis shows that a clear majority of documented studies and applications do not incorporate any XAI methods at all. Of those that do, most rely on feature importance analysis. Truly interpretable-by-design systems or those that use more advanced explanation techniques remain exceptional cases. This indicates a lag between the advocacy for XAI and its actual adoption in optimization practice. Consequently, many state of the art systems produce black-box recommendations, hindering user trust and the willingness of engineers to implement AI-driven strategies in critical systems.

Synthesizing the first two gaps reveals the primary motivation for this dissertation. There is currently no established, holistic framework that systematically addresses the dual challenge of data scarcity and the need for interpretability in the context of sustainable manufacturing optimization. The literature is fragmented. On the one hand, there are powerful data-efficient optimization methods, such as BO. On the other hand, there is a growing toolkit of XAI methods. However, these are rarely co-designed. Researchers have not yet produced a unified methodology that is simultaneously:

- Data-Efficient: Explicitly designed to work with the small sample sizes common in industrial experimentation.
- Interpretable: Built to provide transparent, actionable insights into its decision-making process.

- Integrated: Combines the predictive modeling, optimization, and explanation components into a single, cohesive workflow.

Addressing this gap is the central objective of this thesis. The goal is to develop and validate an integrated, interpretable and data-efficient optimization framework capable of moving beyond a purely performance-centric paradigm to one that delivers trustworthy and actionable solutions for sustainable manufacturing.

3.5 Summary of State of the Art

The review of state of the art, which examines academic research and industrial practice, reveals a field characterized by rapid progress but also by a critical gap between potential and application. Research on data-driven optimization in manufacturing has advanced in two directions. One direction focuses on sequential, experiment-in-the-loop search, while the other relies on offline surrogate modeling with batch data. Decision justification increasingly uses XAI, both through post-hoc analysis of black-box models and through interpretable-by-design formulations. The evidence indicates consistent gains in energy efficiency and product quality, yet the methodological reporting remains heterogeneous. Data foundations such as preprocessing, handling of missing values, and outlier treatment are often under-specified, which limits reproducibility and robustness. Leading industrial companies such as Siemens, Bosch and BMW report impressive successes, including reductions in energy and maintenance costs in the range of 10 to 40% and prognostic precision greater than 90% in predictive maintenance, which reduces scrap rates. Despite these quantifiable performance gains, most commercial systems operate as opaque black boxes. Process engineers receive optimized results, but are rarely provided with insights into the causal rationales, which hinders trust and broad acceptance. In summary, although there are data-efficient algorithms and a broad spectrum of XAI methods, there is a shortage of established, integrated frameworks that systematically combine both aspects. This discrepancy defines the primary research gap that this dissertation addresses.

4 Methodology

The methodology developed in this work is designed as an integrated, multi-stage process that ensures a systematic procedure from the problem statement to the validated solution. Figure 4.1 provides a schematic visualization of this entire process. These phases provide the structure for the procedures detailed in this chapter.

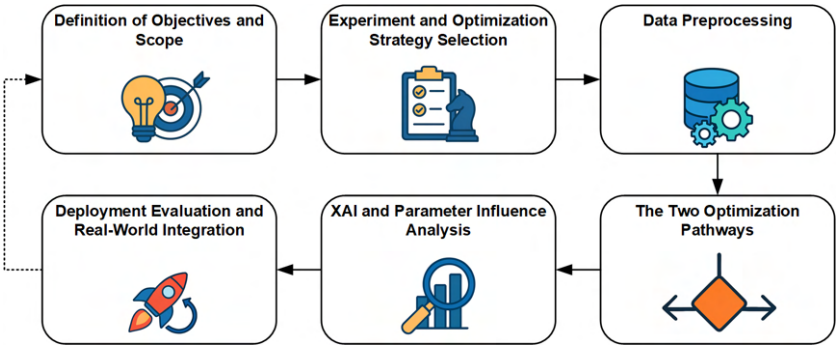


Figure 4.1: Overview of the proposed methodology.

4.1 Design Principles and Scope of the Methodology

The state-of-the-art review in the preceding chapter identified deficiencies in current approaches, namely, the challenges of data scarcity, the gap between model performance and explainability, and the lack of a holistic framework that

integrates these concerns. The methodology detailed in this chapter is designed to address these gaps. To realize this objective in a practical and extensible way, the implementation is guided by two principles: functional modularity and open-source adaptation. Functional modularity dictates that each component of the workflow, from data pre-processing to surrogate modeling and explanation generation, is structured as a distinct module with clearly defined inputs and outputs. This supports advances in specific elements and reduces obstacles for industrial implementation [293]. In parallel, the principle of open source adaptation ensures that each module, where feasible, is built upon mature community-maintained libraries, which accelerates development, reduces vendor lock-in, and enhances reproducibility. To maintain a clear analytical focus, the scope of this methodology is deliberately bounded. Advanced operational concerns such as long-term model drift, cybersecurity, and cross-plant governance are considered beyond the scope of this thesis. These challenges, while important, involve orthogonal control loops and organizational structures that are different from the short-cycle process optimization problem addressed here. The framework's modular architecture, however, provides clear integration points for these capabilities to be added as future extensions without re-engineering the core optimization pipeline.

4.2 Definition of Objectives and Scope

The first phase of the methodological framework is the definition of objectives and scope, a practice anchored in the classical theory of DoE [294]. This initial step is important to ensure that the optimization is well posed, relevant, and aligned with the strategic goals. As illustrated in Figure 4.2, this process involves three sequential stages. It begins with the identification of the relevant product engineering phase to contextualize the problem, followed by the determination of the optimization goals. Finally, these objectives are explicitly aligned with broader environmental sustainability targets.

Figure 4.2: The process for setting objectives and scope.

4.2.1 Identification of the Relevant Product Engineering Phase

Any manufacturing optimization task must be situated within the broader product engineering lifecycle. In this work, the lifecycle is treated as a sequence of eight phases. It begins with strategic planning and concept development and continues through system level design and detail design, as codified for example in Verein Deutscher Ingenieure (VDI) 2221 [13, 295]. After the design focused phases, the process moves to testing and refinement and then to pilot production, where new manufacturing routes are validated at low volume [296]. The lifecycle culminates in production ramp up, which is characterized by steep learning curves, and in series production, where mature process controls are in place [297]. These eight phases are represented as the ordered set $\mathcal{P}$:

$$\begin{aligned}\mathcal{P} = \{ &\text{strategic planning},\ \text{concept development},\ \text{system-level design},\\ &\text{detail design},\ \text{testing \& refinement},\ \text{pilot production},\\ &\text{production ramp-up},\ \text{series production} \}.\end{aligned} \tag{4.1}$$

To operationalize the set in Eq. (4.1), Table 4.1 summarizes observable indicators for the four design oriented phases.

Table 4.1: Phase identification: design phases.

Indicator	**Strategic planning**	**Concept development**	**System level design**	**Detail design**
Typical volume	No builds or few demonstrators	Few proof of concept prototypes	Subsystem prototypes and rigs	Tens of product level prototypes
Historical data	External benchmarks only	Limited measurements and logs	Subsystem data and early integration logs	Component and assembly records
Change freedom	Very high	High	Moderate	Moderate to low
Primary objective	Portfolio and targets	Feasibility and concept screening	Architecture and requirement flow down	Design for manufacture and tolerance allocation
Quality regime	Business gates	Basic engineering tests	Verification plans at subsystem level	Verification and validation at product level
Decision cadence	Stage gate reviews	Sprint based concept reviews	Integration reviews	Design freeze and release reviews

Table 4.2 completes the overview by listing the corresponding indicators for the industrialization and production oriented phases.

Table 4.2: Phase identification: industrialization and production phases.

Indicator	**Testing and refinement**	**Pilot production**	**Production ramp up**	**Series production**
Typical volume	Targeted test series	Short batches at low volume	Rising volumes with changeovers	Stable high volume
Historical data	Structured test logs	Real time logging with short history	High frequency data with growing history	Long history with full traceability
Change freedom	High	Moderate	Low	Very low with formal control
Primary objective	Parameter screening and robustness	Process validation and initial control limits	Achieve rate and stabilize yield	Maintain capability and reduce cost
Quality regime	Engineering acceptance criteria	Pre series control plan and capability indices	Statistical process control and corrective actions	Mature quality system with audits and compliance
Decision cadence	Campaign reviews	Batch reviews and gate approvals	Daily or shift reviews	Routine management by metrics and periodic audits

Using the indicators in Tables 4.1 and 4.2, the phase assignment can be formalized as a maximization over the set $\mathcal{P}$.

$$\widehat{\mathcal{P}} = \arg\max_{\mathcal{P}_i \in \mathcal{P}} \; s(\mathbf{x}, \mathcal{P}_i), \tag{4.2}$$

where $s(\mathbf{x}, \mathcal{P}_i)$ counts the number of indicators that match the signals for phase $\mathcal{P}_i$. The bonds are resolved in favor of an earlier phase when design freedom is high and data are sparse, and in favor of a later phase when production volume is high and quality processes are mature.

The explicit assignment of the process under investigation to a specific phase, $\mathcal{P}_i \in \mathcal{P}$, is a critical step in this methodology. This classification determines the nature of the optimization problem, including the granularity of the decisions to be made, the type and volume of available data, and the appropriate methodological levers. For instance, optimization in early phases, such as system-level design, is typically characterized by assumption-driven, low-fidelity information and focuses on broad design space exploration. In contrast, optimization in late phases, such as series production, benefits from high-frequency sensor streams and extensive historical records, enabling a focus on fine-grained parameter tuning for yield improvement and process stability.

4.2.2 Determination of Optimization Goals

Converting a broad performance or sustainability ambition into a mathematically tractable objective is the next step that links the qualitative problem statement to quantitative optimization. Two families of indicators dominate industrial practice. The first encompasses resource efficiency metrics such as specific energy demand, material yield, and CO_2 equivalent emissions. These variables may be measured directly on the machine or, when the system boundaries are wider, derived from an LCA following ISO 14040/44 [45]. The second family comprises classical key performance indicators (KPIs), for example, tensile strength, costs, scrap rate, geometric conformity or throughput, which safeguard economic competitiveness and often act as hard constraints on sustainability-driven interventions.

In this setting, $\mathbf{x} \in \Omega \subset \mathbb{R}^d$ can be denoted as the vector of input parameters and $f : \Omega \to \mathbb{R}$ defined as the function that assigns each setting its associated optimization indicator. The optimization problem is then formulated as:

$$\min_{\mathbf{x} \in \Omega} f(\mathbf{x}), \tag{4.3}$$

where the feasible region Ω is limited by, for example, machine capacity, physical constraints, safety limits, and specification tolerances. The optimization

is stated using the usual minimization convention, a maximization is handled equivalently. Using a single objective in Equation (4.3), the search space collapses into one quantitative dimension, allowing efficient convergence. When multiple, potentially conflicting objectives are present, they must first be reformulated, through weighting, utility functions, Pareto-ranking proxies, or other aggregation schemes, into a unified objective that fits this framework. Additional performance criteria can still enter the analysis as hard constraints or a penalty term. However, optimization algorithms focus exclusively on the primary objective, ensuring that improvements in the declared key indicator remain interpretable and verifiable on the shop floor.

4.2.3 Alignment of Objectives with Environmental Sustainability

The optimization task is defined by a primary operational indicator $f(\mathbf{x})$, which may target throughput, cost, quality, or energy. Alignment with environmental objectives is ensured by co-evaluating an LCIA-compatible impact $I_c(\mathbf{x})$ for a chosen impact category c in every candidate $\mathbf{x}$. The category c determines the characterization model and its unit u_c. In general,

$$I_c(\mathbf{x}) \;=\; \mathcal{L}_c\big(g(\mathbf{x});\, \boldsymbol{\theta}_c\big) \quad [u_c], \tag{4.4}$$

where $g(\mathbf{x})$ collects measured activity indicators and $\boldsymbol{\theta}_c$ comprises the models and category-specific factors. If the primary indicator is already an LCIA endpoint in category c, no transform is required:

$$I_c(\mathbf{x}) \;=\; f(\mathbf{x}). \tag{4.5}$$

To maintain methodological consistency and enable transfer across sectors, a generalizable framework for quantifying and reducing product-level CO_2e is adopted, building on the lever-based decomposition and sub-process accounting

proposed in the related doctoral study [298]. Consistent with the system-boundary logic of the GHG Protocol and ISO 14067 introduced in Section 2.2, total emissions are expressed as the sum of contributions from fuel and energy supply, the core transformation process, machine operation, auxiliary systems, and logistics, with negative terms where sinks or carbon capture apply. Formally, let $\mathcal{S}$ denote the set of sub-processes; then

$$I_c(\mathbf{x}) \;=\; \sum_{s\in\mathcal{S}} I_{c,s}(\mathbf{x}) \;+\; I_{c,\text{neg}}(\mathbf{x}), \qquad I_{c,\text{neg}}(\mathbf{x}) \le 0. \tag{4.6}$$

Mitigation opportunities are organized as levers grouped into Machine Efficiency, Process Efficiency, Energy Source, Operation Efficiency, Material Efficiency, and CO_2 Capture and Storage. The structure is domain-agnostic and transferable. Dominant contributors vary by context: mobile machinery is typically governed by direct fuel use and utilization patterns, whereas additive and other manufacturing settings are driven by electricity demand, material intensity, and losses arising from idle time or scheduling. Transferability is achieved by retaining the lever taxonomy while adapting the key performance indicators to the relevant functional unit and data sources, for example specific electricity consumption in kilowatt-hours per part or per machine hour, appropriate grid emission factors, and a geometry-independent measure of material intensity based on the proportion of infill mass.

If the indicator is an activity variable, it is translated into LCIA through appropriate emission factors consistent with the IPCC guidelines [49, 299]. For energy-based indicators with energy per functional unit $E(\mathbf{x})$ in kWh, the impact follows from the electricity-mix emission factor EF_E $[\mathrm{kg\,CO_2\text{-}eq\,(kWh)^{-1}}]$:

$$I(\mathbf{x}) \;=\; E(\mathbf{x}) \cdot EF_E. \tag{4.7}$$

For material-based impacts a category-specific factor $EF^{(c)}_{\text{mat}}$ $[u_c\,\mathrm{kg}^{-1}]$ is used for the chosen impact category c:

$$I_c(\mathbf{x}) \;=\; EF^{(c)}_{\text{mat}} \cdot m_{\text{mat}}(\mathbf{x}). \tag{4.8}$$

In the multi-material case with materials i and masses $m_i(\mathbf{x})$, the impact is the sum of component-wise contributions:

$$I_c(\mathbf{x}) \;=\; \sum_i EF^{(c)}_{\text{mat},i}\, m_i(\mathbf{x}). \tag{4.9}$$

If needed, the consumed mass links to batch size and per-unit mass:

$$m_{\text{mat}}(\mathbf{x}) \;=\; n_{\text{batch}} \cdot m_{\text{FU}}(\mathbf{x}). \tag{4.10}$$

Where closed-loop recycling or remelting applies, a credit factor $\gamma_{\text{rec}} \in [0, 1]$ reflects avoided primary production:

$$I(\mathbf{x}) \;=\; EF_{\text{mat}} \cdot \big(1 - \gamma_{\text{rec}}\big) \cdot m_{\text{scrap}}. \tag{4.11}$$

Regardless of the primary target, $I(\mathbf{x})$ is evaluated for each candidate $\mathbf{x}$ to make trade-offs explicit. If a broader set of impacts is relevant, a full LCA in accordance with ISO 14044 [45] is preferable to streamlined estimates, since it captures interdependencies across the value chain [300, 301].

4.3 Experiment Planning and Optimization Strategy Selection

The performance of any optimization strategy, whether statistical or model-based, is bounded by the observations on which it rests. Establishing a high-quality data foundation is therefore the essential prerequisite for any successful learning and optimization cycle. Before an algorithm can reliably recommend set-points, detect drifts, or quantify risk, the underlying measurements must capture the true process variability with sufficient fidelity, granularity, and traceability. Consequently, the initial and often most consequential design decision is related to the strategy to acquire these data. This structured process

is illustrated in Figure 4.3. After collection, this data set subsequently undergoes a pre-processing pipeline which will be outlined in the next section.

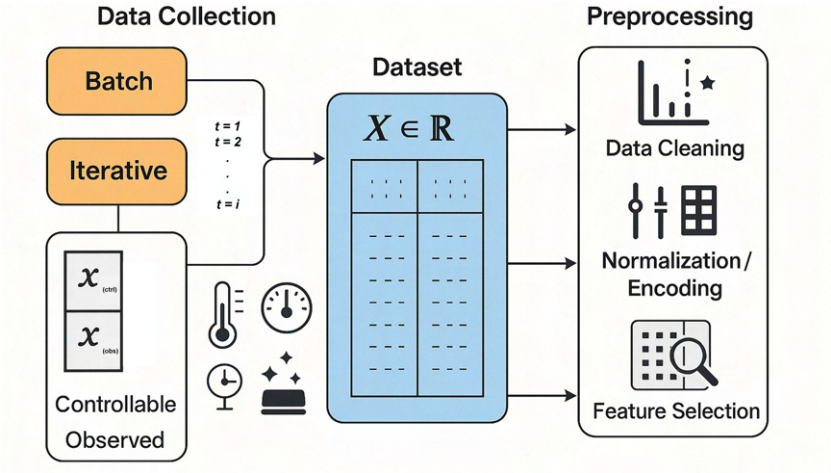


Figure 4.3: The data acquisition and preparation pipeline.

4.3.1 Data Representation and Acquisition Strategies

Regardless of the specific data modalities involved, which can range from high-frequency sensor streams to qualitative quality assessments, the standardized input for any data-driven optimization is a data matrix

$$\mathbf{X} \in \mathbb{R}^{N \times d} \tag{4.12}$$

,where N is the number of observations and d is the number of features. Each row typically represents a distinct experimental run or production cycle, while each column corresponds to a specific process parameter, material property, or environmental variable. The process of populating this data matrix is governed by two strategic decisions: the timing of data collection and the classification of variable types.

The first design choice concerns the data collection protocol, which can be either a batch campaign or a sequential regime. In a batch campaign, the entire set of experiments $\Omega_{\text{batch}} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$ is instantiated in a dedicated sampling window - often a pre-series trial or maintenance shutdown - so that the confounding of day-to-day drift is minimized and the disruption of the shop floor is contained. The resulting data set

$$\mathcal{D}_{\text{batch}} = \left\{ \left(\mathbf{x}_i, y_i\right) \,\middle|\, i = 1, \ldots, N \right\}. \tag{4.13}$$

is then used to train a model in one step. In contrast, in a sequential or iterative regime, the data set evolves as

$$\mathcal{D}_t = \left\{ \left(\mathbf{x}_1, y_1\right), \left(\mathbf{x}_2, y_2\right), \ldots, \left(\mathbf{x}_t, y_t\right) \right\}, \tag{4.14}$$

in which each new configuration $\mathbf{x}_{t+1}$ is selected by an acquisition rule that trades off exploration against exploitation [302].

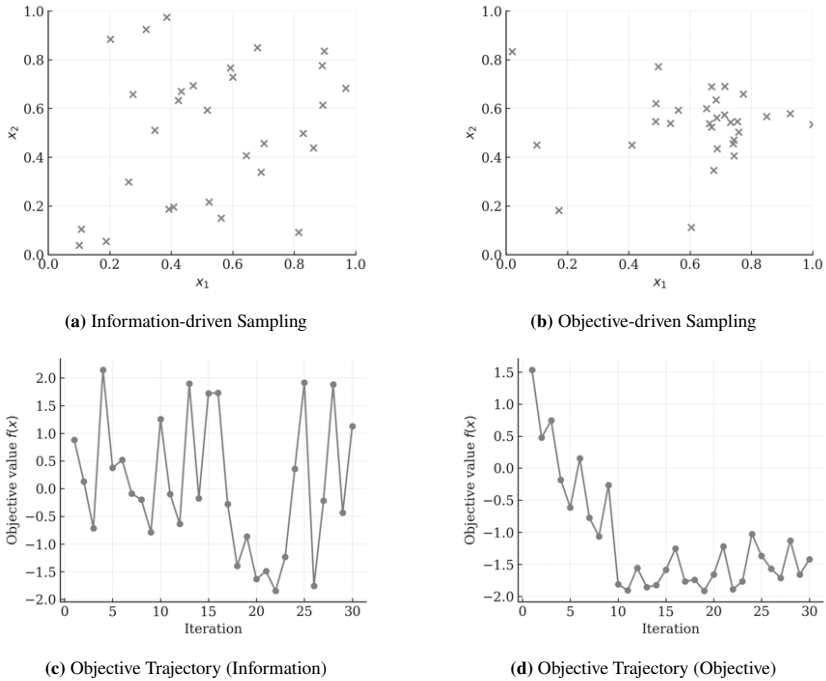
The second design decision is to distinguish between input variables that can be actively manipulated and those that can only be passively observed. Controllable parameters $\mathbf{x}^{(\text{ctrl})} \in \mathbb{R}^{d_c}$ are the levers that the operators can manipulate. The observed covariates $\mathbf{x}^{(\text{obs})} \in \mathbb{R}^{d_o}$ can record the external environment, ambient humidity, feedstock moisture, or a voltage ripple in the grid, whose influence is real but not directly actionable. This distinction influences both modeling and governance: controllables must be varied to avoid aliasing, whereas observed factors require dense enough sampling so that their effect can later be partially accounted for by regression or weighting. When time-resolved sensor streams are available, the observational block is further enhanced by lagged and derived variables to capture dynamic transients [303].

4.3.2 Selection of Optimization Approach and Data Collection Strategy

The performance of any optimization routine is bounded by the information content of the data upon which it is based. To select the most effective strategy for a given problem, a structured top-down decision framework is therefore required. This framework aligns the choice of optimization architecture with the overarching strategic objectives and the practical constraints of the manufacturing environment. Although the ultimate strategic goal is always to improve a specific KPI or environmental metric, the tactical priority for a given optimization campaign or phase can differ. This determines the allocation of the experimental budget and the immediate trade-off between exploitation and exploration, leading to two distinct approaches:

- Focused Exploitation: This is the priority when the mandate is rapid, short-term improvement. The approach is purely objective-driven, concentrating the experimental budget on local refinement within high-potential regions to achieve the steepest possible performance gains. In this mode, building a comprehensive global model is a secondary concern.
- Strategic Exploration: This priority is chosen when establishing a robust, long-term optimization system, for instance, during the initial investigation of a new process or when an existing dataset is insufficient. This information-driven phase is not an end in itself but is a foundational step. The goal is to minimize global model uncertainty by using a space-filling experimental design to build a globally accurate surrogate model, which then enables more effective and reliable KPI optimization over the long term.

Figure 4.4 illustrates the practical consequences of this choice by comparing the two approaches on the same synthetic optimization problem.



(a) Information-driven Sampling

(b) Objective-driven Sampling

(c) Objective Trajectory (Information)

(d) Objective Trajectory (Objective)

Figure 4.4: Comparison of information-driven and objective-driven sampling.

The top row contrasts the resulting sampling patterns: the information-driven strategy (a) produces a space-filling design that covers the entire domain, while the objective-driven strategy (b) rapidly concentrates its samples in a localized region of high potential. The bottom row shows the corresponding optimization trajectories. The exploratory, information-driven approach (c) results in a high-variance objective trajectory as it continues to sample globally to reduce uncertainty, finding a good solution but not necessarily the best one within the limited budget. In contrast, the exploitative, objective-driven approach (d) achieves a much faster convergence to a low objective value by focusing its search, although at the cost of leaving large regions of the design space unexplored.

After establishing the strategic goal, the next factor is the operational tempo, which is determined by the latency of the target metric. This practical constraint dictates whether data can be acquired and processed quickly enough to support an iterative optimization loop, or if a slower, batch-based approach is necessary. To formalize this, a taxonomy of four latency classes can be defined, ranging from near-instantaneous (L0) to long delay (L3), as detailed in Table 4.3. The implications of this classification are profound: an in-line, real-time metric (L0) permits a fully adaptive optimization strategy, whereas a long-delay metric (L3), such as a multi-week corrosion test, mandates a batch-only design. This single factor, therefore, constrains the set of applicable optimization algorithms and shapes the entire experimental protocol. However, the feasibility is bounded by the overall horizon of the campaign. For Table 4.3, the time budget of one week is assumed. For different time budgets, the cadence should be scaled accordingly.

Table 4.3: Impact of target-metric latency on feasible data-collection strategy in industrial manufacturing.

Latency class	**Typical turnaround**	**Feasible sampling mode(s)**	**Manufacturing examples / remarks**
L0 – *in-line*	$\leq$ seconds	Full iterative optimization possible or batch	inline vision control, laser triangulation
L1 – *shift-scale*	minutes to ≤ 8 h	Iterative possible if the line can idle; small daily batches	hardness test after heat treatment, colorimeter reading after paint cure
L2 – *overnight*	8–24 h	Hybrid: nightly batch, next-day model update; 1 iteration / day	Tensile test of heat-treated forgings, composite laminate 3-point bend after overnight cure
L3 – *long-delay*	days to weeks	Batch only; iterative loop impractical	Accelerated salt-spray corrosion of plated parts, thermal-cycling plus pull test of solder joints

With the strategic goal and operational tempo defined, the next step is to assess the project's tactical starting point by evaluating the available resources. This assessment is based on two primary constraints: the volume of existing historical data and the budget for new experimentation. The interplay between these two factors defines four different start-up classes (C0–C3), as detailed in Table 4.4. Each class points to a specific initial action plan, guiding the choice between beginning with immediate optimization of existing data versus prioritizing a dedicated phase of further data collection.

Table 4.4: Start-up classes based on data availability and resource limits.

Class	**Available data**	**Budget for further Data**	**Typical industrial context**
C0	none	<5 d runs	First startup of an expensive machine
C1	$N < 5d$	5–10 d runs	Ramp-up, small R&D study
C2	$N > 10d$	$N \leq 10$d	Sparse legacy historians, pilot production line
C3	$N > 20$d	any	Continuous series production

If available historical data and experimentation budget fall into different classes, the assignment follows a conservative dominance rule that preserves feasibility of the initial plan:

$$c^* \;=\; \min\{\, c_D(N),\, c_B(B)\,\}, \tag{4.15}$$

with $c \in \{\mathrm{C0}, \mathrm{C1}, \mathrm{C2}, \mathrm{C3}\}$ ordered by increasing resource sufficiency. As an override policy aligned with the intended workflow, choose $c^* = c_D(N)$ if no additional experiments are planned, and choose $c^* = c_B(B)$ if a sequential design with new experiments is mandated. When the two classes differ by exactly one level and the expected value of information for the next runs is high, an escalation to the higher class can be justified; otherwise the conservative default c^* should be retained.

Each of the previously defined start-up classes implies a distinct initial action plan, guiding the practitioner toward the most appropriate data collection and

optimization strategy. Table 4.5 outlines these recommended routes, providing a clear decision framework based on the project resource constraints.

Table 4.5: Recommended next data-collection step and optimization approach conditioned on the start-up class.

Class	Data-collection step	Optimization Approach	Why
C0	No dedicated data acquisition is feasible.	Sequential Optimization	No budget for offline experiments; must use live process data.
C1	If budget allows, perform a minimal screening design.	Sequential Optimization or initial Offline Optimization.	Goal is to obtain at least one data point per factor to inform the search.
C2	Leverage existing data; supplement with a targeted DoE if needed.	Sequential Optimization or initial Offline Optimization.	Sufficient data exists for an initial landscape diagnosis.
C3	Prioritize analysis of existing data; new experiments only if analysis reveals gaps.	Offline Optimization via Static Surrogates.	Value of modeling existing data likely exceeds the gain from new runs.

The logic of this framework is clearest at its extremes. A project in class C0, with no data and no experimental budget, must proceed with sequential optimization on the live process, as building a reliable static surrogate is infeasible. In contrast, a project in class C3, with abundant historical data, should first focus on exploiting this existing information through offline optimization via static surrogates before committing resources to potentially redundant new experiments.

For projects in classes C1 or C2, the final strategy selection is refined through a quantitative analysis of the data and the objective landscape. This decision is guided by a set of key metrics that characterize data density, landscape smoothness, and modality, providing an evidence-based choice between surrogate modeling and direct search, as summarized in Table 4.6.

Table 4.6: Metrics to guide the choice between offline optimization via static surrogates and sequential optimization.

Metric	**Offline optimization via static surrogates**	**Sequential Optimization**	**Range & Thresholds**
SDR	$\rho \gtrsim 10$	$\rho \lesssim 10$	Range: $[0, \infty)$; high if $\rho > 10$, low if $\rho < 10$
Missingness	Low to moderate	Potentially high	Noise fraction; low if $< 10\%$, high if $> 30\%$
Autocorrelation length	High	Low	Range: $[0, \infty)$; high if $\tau > 3$, low if $\tau < 1$
PIC	Low	High	Range: $[0, 1]$; low if $M < 0.5$, high if $M > 0.5$
Dispersion	Low	High	Range: $[0, 1]$; low if $D < 0.5$, high if $D > 0.5$

The thresholds presented in the table are based on a consensus of empirical and theoretical results that guide a sequential decision process. First, the sample-to-dimension ratio ρ is assessed. Offline optimization via static surrogates is considered statistically reliable once the sample-to-dimension ratio exceeds roughly ten design points per variable. This $10d$ rule is supported by the study of [304], in which model hyperparameters were shown to stabilize when $\rho = N/d \gtrsim 10$, while lower densities led to an inflated coefficient variance and overfitting. If this data sufficiency condition is met, the next step is to characterize the objective landscape using ELA metrics. A high autocorrelation length $\tau > 3$ indicates a smooth, correlated response surface that is well suited for offline optimization via static surrogates, while $\tau < 1$ suggests severe ruggedness that favors sequential optimization [305]. The multimodality of the landscape is evaluated through the partial information content $M(\phi)$ and the dispersion of elite solutions D. Values of $M(\phi) > 0.5$ and $D > 0.5$ have been

reported to coincide with frequent directional changes and a wide spatial spread of high-quality points, both hallmarks of a multimodal surface that benefits from the adaptive exploration of sequential optimization [306, 307]. Robust estimation of these landscape metrics requires a well-distributed sample. If the existing data are heavily clustered or biased, a small, unbiased pilot probe of approximately 5d randomized runs should first be conducted. Furthermore, while moderate noise or missingness can often be mitigated by regularization and imputation, higher levels of data corruption typically render the static surrogates used in offline optimization unreliable, shifting the preference back to more robust sequential optimization methods.

4.4 Data Preprocessing

Independent of how and when the data are recorded, careful preprocessing is essential to guarantee integrity and comparability between experiments. The state of the art review revealed a preprocessing gap in the literature, where the foundational steps of data preparation are often overlooked or unsystematically reported, undermining the reproducibility and reliability of optimization results. To address this deficiency, this section presents a systematic and prescriptive workflow for data pre-processing.

As illustrated in Figure 4.5, this workflow is organized into six sequential decision stages, moving from initial data assessment to final feature construction. Each stage has a clear objective, a set of recommended techniques drawn from the prior review, and explicit exit criteria to ensure a rigorous and auditable process.

Stage 1: Data Profiling. Before any modification is attempted, the raw tables are profiled to establish a quantitative baseline. This involves computing several key descriptive metrics to inform the subsequent preprocessing strategy.

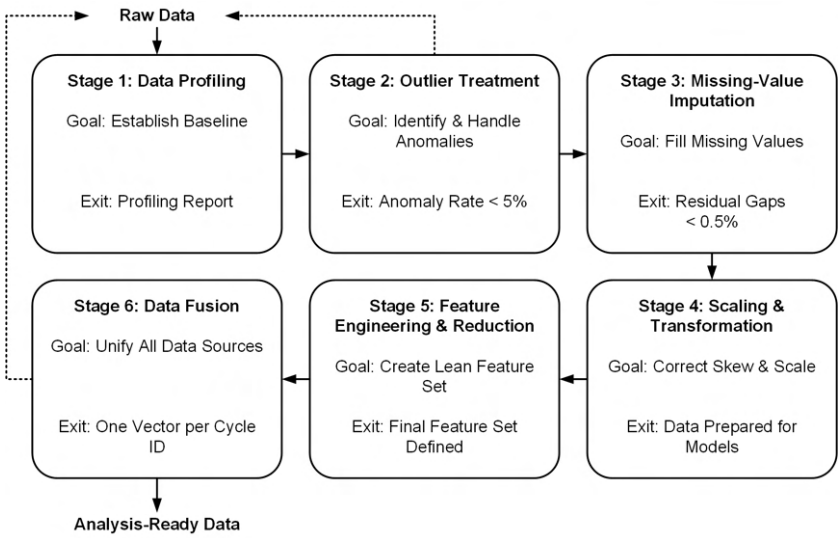


Figure 4.5: The data preprocessing workflow.

The marginal distribution of each feature is analyzed to understand its underlying shape, which is critical to selecting appropriate scaling or transformation methods. To assess the scale, spread, and potential presence of outliers in the data, key empirical quantiles are calculated for each feature. This typically includes the minimum and maximum to establish the range, the median, the interquartile range, i.e., the 25th and 75th percentiles, and the 5th and 95th percentiles to characterize the tails of the distribution. Finally, the rate of missing cells is quantified for each column and for the dataset as a whole, as this metric is the primary determinant for choosing a suitable imputation strategy. These statistics are stored in a versioned report so that all subsequent cleaning and transformation steps can be audited against a stable reference. The stage is considered complete once the summary statistics have been stored in a reproducible format and verified for completeness.

Stage 2: Outlier Treatment. Once a stable baseline has been established, the workflow proceeds to the identification and removal of outliers. The procedure follows a strict outside-in hierarchy: clearly implausible univariate extremes are filtered first because their presence can mask subtler multivariate anomalies. Although the literature shows various effective techniques, such as model-based residual analysis used in some studies [187] or domain-specific filtering [188], this methodology advocates for the non-parametric MAD rule with $\kappa = 3$ as the robust default, given its resilience to the non-Gaussian distributions common in industrial data. For features known to be approximately normal, the classical z-score can be used as a simpler alternative. Following this initial cleaning, and in contrast to many studies that stop at univariate analysis, an Isolation Forest is applied to detect more complex, multivariate anomalies [308]. By measuring the average path length required to isolate a point in an ensemble of random trees, the algorithm captures joint outlierness in moderate-dimensional space without relying on parametric assumptions. This two-step process is governed by a heuristic that is often implicit in the literature: if the outlier fraction for any feature exceeds a high threshold, such as 20%, it is flagged for engineering inspection of a potential systemic fault rather than being statistically treated. The outlier treatment stage is complete when the residual anomaly rate across falls below the predefined ceiling of 5%, at which point the data set is considered sufficiently clean for scaling and encoding in Stage 3.

Stage 3: Missing-Value Imputation. This stage employs a tiered diagnostic workflow to classify the missing data mechanism for each variable according to Rubin's taxonomy. The analysis proceeds sequentially, stopping for any given variable as soon as sufficient statistical evidence is found for a mechanism. This efficient approach ensures that most variables are classified after one or two tests, reserving the more complex diagnostics for truly ambiguous cases. First, Little's omnibus test χ^2 is applied, in which a nonsignificant result supports the assumption of MCAR. Second, If this assumption is not supported, the analysis proceeds to test for a MAR pattern. This is done by fitting a regular logistic regression of the missingness indicator to fully observed

covariates in which substantial explanatory power, that is, pseudo-R^2 or AUC exceeding 0.15 suggests an MAR mechanism. Third, for variables that remain unclassified, a third test directly probes for an MNAR process by evaluating the dependence on the unobserved value itself. In this test, the missingness indicator is regressed on predictions of the latent value that were obtained from complete cases. A significant slope in this regression indicates an MNAR pattern. Finally, in specific situations where longitudinal dropout or selection bias is suspected, a formal selection model is fitted, such as a Heckman two-step or a Diggle-Kenward model. A nonzero inverse Mills ratio or a significant dropout parameter provides conclusive evidence for an MNAR mechanism. The outcome of this sequence is a definitive classification of the missing pattern of each variable, which subsequently dictates the appropriate imputation strategy.

MCAR attributes are imputed with the sample median for static columns or with linear interpolation for time-indexed series. MAR attributes are completed with either MICE or the tree-based MissForest: MissForest is preferred for mixed-type or high-dimensional feature spaces, whereas MICE is preferred when domain-specific parametric forms or detailed convergence diagnostics are required. Attributes identified as MNAR require a specialized handling strategy. The primary approach is to develop a joint selection model that explicitly accounts for the non-random missingness mechanism. If such a model is not methodologically feasible, a decision is made in consultation with process experts: the variable is either excluded from model-critical analyses or imputed using a bespoke, domain-specific surrogate, such as a pre-defined worst-case safety value. To maintain full transparency for subsequent analyses and explainability, a binary flag is added to the dataset, marking every value that has been imputed. The imputation stage concludes with a rigorous quality control and versioning protocol. First, the distributional integrity of the data is verified by comparing the post-imputation distribution of each variable with its original profile using a Kolmogorov–Smirnov test. The process is complete only when two criteria are met: no significant distributional changes are detected and the overall residual missingness of the data matrix falls below the target of 0.5%.

Stage 4: Scaling and Transformation. The transformation and scaling of numerical features are guided by the principle of minimal intervention, designed to preserve the original physical units and enhance the interpretability of the model. Modifications are only performed when justified by distributional properties or the requirements of scale-sensitive algorithms. The methodology follows a two-stage diagnostic process for each feature. First, distributional symmetry is assessed. The tail symmetry ratio (TSR) is computed as follows:

$$\text{TSR} = \frac{Q_{95} - Q_{50}}{Q_{50} - Q_5} \tag{4.16}$$

where Q_p denotes the empirical p-th percentile. Values outside the robust interval of $[0.5, 2]$ indicate a significant imbalance. Second, scale disparity across features is evaluated. A feature is flagged for potential rescaling if its interquartile range (IQR) exceeds that of any other feature by more than three orders of magnitude (10^3). The corrective action is precisely tailored to the diagnosis. For features with asymmetric or heavy tails, a variance-stabilizing transformation is applied; strictly positive records undergo a Box-Cox transform, while variables containing zeros or negative values are treated with a Yeo-Johnson power transformation. For features flagged with scale disparity that are inputs to scale-sensitive algorithms, a suitable scaling method is chosen. Features that are now approximately symmetric are standardized by using their mean and standard deviation. Those still prone to outliers are robustly scaled using their median and IQR. As a special case, sensor readings with known physical limits are linearly mapped to the interval $[0, 1]$ to preserve their inherent constraints. This approach ensures that transformations correct distributional flaws, while scaling is reserved for instances critical to algorithmic performance, yielding a dataset that is both robustly prepared and maximally interpretable.

Stage 5: Feature Engineering and Reduction. The approach to feature engineering and reduction is guided by a principle of parsimony, ensuring that all predictors are directly relevant to the modeling objective. Consequently,

new composite metrics, such as 'energy consumed per part', are engineered conservatively and only when they align with the primary optimization target. The subsequent feature reduction follows a multi-step filtering process, beginning with a test for relevance. Any feature exhibiting a negligible correlation, that is, an absolute correlation below 0.02 with the target variable, is discarded due to its lack of predictive signal. Subsequently, the redundancy from multicollinearity is addressed. When a pair or block of features exhibits very high inter-correlation, i.e., an absolute value above 0.98, a strategic decision is made based on the trade-off between interpretability and information retention.

To identify multicollinearity pairwise associations between continuous variables, Pearson's correlation [309] is used when linearity and approximate normality are plausible.

$$r \;=\; \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}\,\sqrt{\sum_{i=1}^{n}(y_i-\bar{y})^2}}\,. \tag{4.17}$$

Otherwise, monotonic relations are assessed with rank-based measures, and Kendall's τ [310, 311] is adopted as a robust tie-aware alternative. Let C and D denote the numbers of concordant and discordant pairs among all $\binom{n}{2}$ pairs, and let $T_x = \sum_{\ell}\binom{t_\ell}{2}$ and $T_y = \sum_{m}\binom{u_m}{2}$ collect tie contributions for X and Y (with t_ℓ/u_m the sizes of tie blocks). Then

$$\tau_a \;=\; \frac{C-D}{\binom{n}{2}}, \qquad \tau_b \;=\; \frac{C-D}{\sqrt{(C+D+T_x)\,(C+D+T_y)}}\,. \tag{4.18}$$

For maximum interpretability, one of the features is removed on the basis of domain knowledge, preserving the physical meaning of the variables. Alternatively, to retain the maximum shared variance from the correlated set, PCA is used, replacing the original features with the principal components required to explain at least 95% of their combined variance. This stage is complete when the final feature set has been refined for relevance and redundancy, resulting in a lean and robust set of predictors for modeling.

Stage 6: Data Fusion. The final preprocessing stage, data fusion, integrates the various cleaned data sources into a single, analysis-ready dataset. Heterogeneous sources, including high-frequency sensor streams, laboratory inspections, and MES event logs, are first synchronized onto a common time axis using linear or spline interpolation. Subsequently, these time-aligned streams are merged into a unified feature matrix using a master production cycle identifier as the primary key. The success of this stage is clearly defined: There must be exactly one complete and fully populated feature vector for every unique cycle identifier. Before proceeding, a final validation gate confirms that the exit criteria for all previous stages have been successfully met. The data set is only handed over, and the modeling phase is initiated when this comprehensive quality check is passed. To ensure reproducibility, the entire data preparation workflow, including every transformation, parameter, and decision, is codified and version controlled, creating an auditable and executable record. This pipeline incorporates a robust failure handling protocol. If any exit condition cannot be met, for example, if missingness on a key variable remains unacceptably high, the workflow automatically flags the issue and can revert to an earlier stage, such as experimental planning, to trigger a targeted remeasurement. This feedback loop creates a continuous cycle of improving data quality and ensures process integrity.

4.5 Offline Optimization via Static Surrogates

Offline optimization with static surrogates rises and falls with the sampling design. All model inferences and subsequent virtual optimization steps are based on a single data set. Poor designs induce aliasing, ill-conditioned regressors, and blind spots in the domain Ω, whereas good designs control these risks and make the surrogate attractive to optimization. A variety of DoE strategies was already presented in Table 2.4.

Two design domains serve distinct goals and should be selected accordingly. Factorial/effect-oriented designs target identifiable main effects and interactions. They are well suited when factors have discrete levels and when variance components and interpretable response surfaces are required. In contrast, space-filling designs aim for uniform coverage of Ω and are preferable for nonlinear, higher-dimensional, and predominantly continuous parameter spaces in which prediction accuracy and robust interpolation/extrapolation are central. Table 4.7 outlines the primary distinctions and appropriate contexts for each approach.

The selection of a DoE is driven by the use case and the optimization objective. When prediction accuracy over a predominantly continuous, multi-parameter space is the priority, space-filling plans are preferred. When an interpretable response surface is the main target, orthogonal or carefully fractionated factorials are used to preserve identifiability and to control aliasing. While the choice of regime exerts the dominant, first-order influence on generalizability and optimization quality, the within-regime variant mainly serves as a second-order tuning lever which should be aligned with constraints, expected smoothness, noise, and tolerated extrapolation.

The subsequent step involves the development of the static surrogate models. The development of the surrogates follows a four-stage protocol: Data Partitioning, Algorithm Selection, Hyperparameter Optimization, and Model Evaluation. This workflow is illustrated in Figure 4.6. To ensure an unbiased estimate of generalization performance, the data is partitioned into independent datasets for model training, hyperparameter tuning, and final evaluation, with the specific partitioning strategy dictated by the data's underlying structure. For standard tasks, the default approach is a fixed 70/15/15 stratified hold-out split. Stratification is crucial, as it preserves the original class proportions across all subsets, ensuring that each is representative. In cases of small or imbalanced datasets, where a single split could be misleading, a more robust nested k-fold CV is implemented to provide a stable performance estimate while preventing information leakage.

Table 4.7: Comparison of factorial/effect-oriented vs. space-filling designs for offline surrogate modeling.

Aspect	**Factorial / effect-oriented**	**Space-filling**
Goal	Identify main effects and interactions; interpretable response surfaces and variance components.	Uniform coverage of Ω for high prediction accuracy and robust interpolation/extrapolation.
N & scaling with k	Structure-driven; grows combinatorially with levels and factors; See Table 2.4.	Flexible and budget-driven; choose N to meet error targets.
Factors & constraints	Naturally suited for discrete factors; mixing qualitative & quantitative via RSM. Complex feasibility/integrality constraints require pruning or candidate sets.	Best for continuous parameters; categorical factors via blocking/stratification. Constraints handled via constrained LHS, rejection sampling, or D-optimal selection from feasible candidates.
Design properties	Explicit aliasing; corner and center points support curvature and pure-error estimation.	No formal aliasing grid; low correlation, maximin spacing, and low-discrepancy targets. Boundary coverage must be enforced.
Examples	Full/Fractional factorial, Plackett–Burman, Central Composite, Box–Behnken.	LHS, Sobol sequences, D-optimal Design.
Prefer	Prefer for discrete factors and effect attribution at low–moderate k.	Prefer for continuous, nonlinear, higher-k problems with prediction-first goals.
Limitations	Confounding in low-resolution fractions; run inflation; missing curvature.	Coverage holes/poor projections if unoptimized; weak boundary coverage; under-replication in noisy systems.

The resulting data sets serve defined sequential purposes. The training set is dedicated exclusively to learning the model's internal parameters during the optimization process. The validation set is utilized for the iterative process

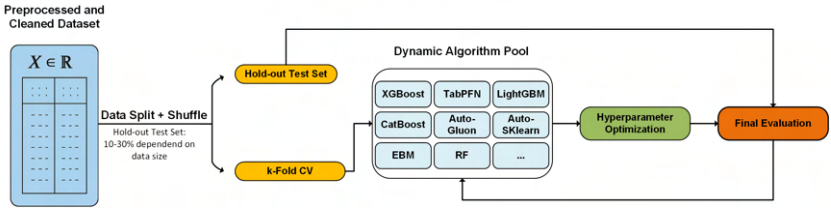


Figure 4.6: The machine learning-based predictive modeling workflow.

of hyperparameter tuning and selecting the final model architecture. Finally, the test set is strictly reserved as a final, unseen hold-out data set as it is used only once to obtain an unbiased estimate of the selected model's generalization performance.

The selection of an appropriate machine learning algorithm for industrial applications transcends the singular objective of predictive accuracy and must incorporate considerations derived from the operational context of the manufacturing domain. Consequently, this methodology commences with the explicit definition of a set of criteria that any candidate model must fulfill.

- Scalability and Implementation Feasibility: Priority is given to algorithms that exhibit low to moderate computational complexity and do not depend on specialized hardware [312]. This ensures feasibility within standard industrial computing environments and enables real-time prediction capabilities where necessary, avoiding the high costs and setup overhead associated with specialized infrastructures [313].
- Public Availability and Maintainability: Preference is given to algorithms from well-supported open source ecosystems with permissive licenses. This strategy mitigates the risk of vendor lock-in, ensures long-term maintainability through community support and public audits, and facilitates collaboration, aligning with the industry-wide shift towards open-source AI solutions [314, 315].
- Model Parsimony and Required Expertise: The complexity of a model is carefully matched to both the available data volume and the level

of tuning expertise required. In the data-limited contexts common to manufacturing, simpler models with fewer parameters are preferred to ensure robust generalization and mitigate the risk of overfitting [316]. This approach also reduces the tuning effort, making such models more accessible and manageable for teams without specialized deep learning expertise.

Applying the aforementioned scalability, maintainability, and parsimony criteria to the candidate algorithms yields a clear profile of their respective strengths and weaknesses in a manufacturing context. Table 4.8 summarizes this evaluation, assessing each algorithm class against key operational indicators such as the minimum sample-to-dimension ratio (ρ) required for robust performance, the necessary computational budget, the level of tuning expertise and the degree of native transparency.

Table 4.8: Practical requirements and characteristics of the algorithm classes.

	Tree Ensembles	**Transformer**	**Auto-ML Systems**	**Interpretable GAM**
Min. ρ	$\gtrsim 10$	$\gtrsim 50$	adaptive	$\gtrsim 5$
Compute Budget	medium; CPU / optional GPU	high; GPU mandatory	medium–high	low CPU
Expertise	medium	high	medium	low
Native Transparency	medium	low	varies	very high
Licence(s)	Apache-2.0 (XGBoost/-LightGBM/-CatBoost), BSD-3 (RF)	MIT (TabNet/-SAINT/Tab-Transformer), Apache-2.0 (TabPFN)	Apache-2.0 (AutoGluon/H2O), BSD-3 (AutoSklearn), GPL-3 (TPOT)	MIT (EBM)

Interpretable GAMs, such as the EBM, are established as the starting point and interpretable baseline. Given their competitive precision in low to moderate data volumes ($\rho < 5$), minimal computational cost, and high transparency,

they are the ideal choice for initial analysis or when regulatory explanation is required. For tasks where predictive performance is paramount, tree-based ensembles occupy the sweet spot. They outperform simpler baselines once a moderate sample-to-dimension ratio ($\rho \gtrsim 10$) is reached and represent the optimal trade-off between accuracy, tuning effort, and computational cost on standard hardware. The most complex models are reserved for specific scenarios. Tabular transformers are considered only in data-rich regimes ($\rho \gtrsim 50$) where their state of the art performance justifies the investment in GPU resources and specialized expertise. Finally, AutoML frameworks serve as a powerful tool to accelerate the model selection process, especially when in-house tuning expertise is limited, with the understanding that their utility is balanced against potentially high computational search costs.

The final hyperparameter optimization follows a protocol designed for robustness, efficiency, and reproducibility. For the diverse and complex model selection scenarios encountered in this thesis, which are analogous to the CASH problems, Optuna was selected as the HPO framework. Its demonstrated strengths in these scenarios, flexible implementation, and robust trial pruning mechanisms make it the most suitable tool for achieving high model quality with a manageable computational budget. A 5x3 nested cross-validation design is employed, where the 5-fold outer loop provides estimates of the generalization error, while the 3-fold inner loop drives the hyperparameter search. Within this inner loop, Optuna explores the search space for a fixed number of 100 trials. To enhance efficiency, Optuna's early-stopping mechanism is activated with a patience of 15 trials, automatically pruning unpromising trials. Subsequently, the top five candidate configurations identified by Optuna undergo stability validation, where each model is re-trained and evaluated 30 times with different random seeds. The configuration that exhibits the best trade-off between high mean performance and low variance is selected as the final optimal setting and is re-analyzed on the entire training dataset to produce the final surrogate model.

The performance of the final model is evaluated using a protocol to provide a multifaceted view of its predictive capabilities. For regression tasks, a set of complementary metrics is reported: the MAE, providing a robust measure

of average error, the RMSE, which penalizes large prediction errors; the R^2, to quantify the proportion of variance in the target variable explained by the model.

Once the predictive accuracy and robustness of the model have been confirmed, it can be used for its primary function as a virtual process model. This surrogate enables highly efficient offline optimization to identify the optimal process parameters without relying on expensive and time-consuming real-world experiments. This optimization process is itself guided by BO. This algorithm iteratively queries the surrogate model by proposing promising input settings and analyzing the model's predicted outputs. Because a single query of the model takes only milliseconds, thousands of potential operating points can be tested virtually in a short amount of time, a procedure that would be infeasible in the physical world. The result of this virtual search is a set of optimal process parameter settings, which can then be validated with a high probability of success in a single, final confirmation run on the real system.

4.6 Sequential Optimization

For sequential optimization tasks where each experiment is costly, this methodology employs BO as the primary sequential search algorithm. A BO procedure is defined by three design choices: the surrogate model that approximates the objective function, the initial sampling strategy to seed the process, and the acquisition function that guides the search. Figure 4.7 summarizes the iterative loop in which these elements interact. The recommended combinations are informed by the pre-study of Greif et al. [3], in which a comprehensive benchmark was conducted on both analytical test functions and industry-derived optimization tasks. The study evaluated three factors: initial sample scheme (no sampling, LHS, FFD), surrogate kernel family (isotropic and anisotropic GP, RF, TPE) and acquisition function (EI, UCB, POI) with varying exploration weights κ and measured performance at multiple iteration checkpoints. The choice of surrogate model is the most critical factor and is dictated by the

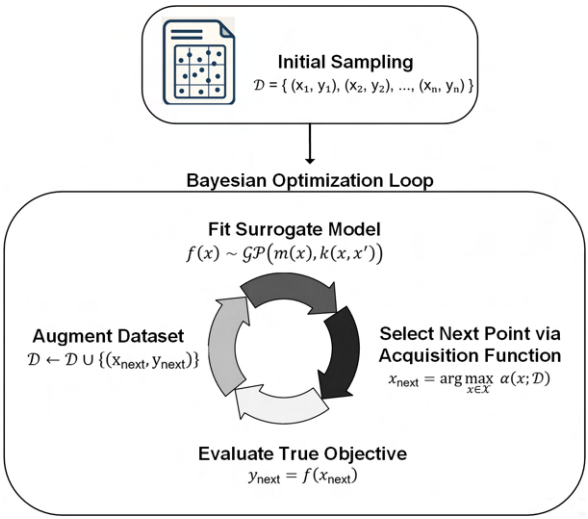


Figure 4.7: The BO Loop.

expected characteristics of the problem's objective function. The findings from the literature [260, 3] show that no single surrogate is universally dominant. Therefore, a problem-specific selection is made based on the recommendations in Table 4.9.

The effectiveness of an initial sampling strategy is dependent on the chosen surrogate model. Based on the interaction effects demonstrated by Greif et al. [3], the sampling strategy is directly coupled to the surrogate selection. When using an anisotropic GP surrogate, no initial sampling is performed. The analysis showed that allowing the acquisition function to guide the search from the very first iteration yields the fastest convergence for this model class. When using tree-based surrogates (RF or TPE), an initial LHS design is used to generate a space-filling set of points. This provides these non-parametric models with good initial coverage of the search space, which was shown to significantly accelerate performance, particularly for complex objective functions. When main effects are of primary interest in low-dimensional problems, an FFD can be used to efficiently seed the model, especially if optima are expected near

Table 4.9: Surrogate model recommendations for BO.

Surrogate Model	**Recommended Use Cases**	**Notes / Limitations**
GP isotropic kernel	Low-dimensional, isotropic, and smooth objective functions. Serves as a simple baseline.	Fast to train but its core assumption of uniform behavior across all dimensions is often violated in practice.
GP ARD kernel	Smooth, medium-dimensional objectives where input variables have differing impacts or scales.	Automatically determines feature relevance; consistently outperforms isotropic GPs in most real-world problems. The default choice for continuous optimization.
RF surrogate	High-dimensional, non-smooth, or discontinuous objective functions.	Highly scalable and robust to irrelevant features. Does not provide principled uncertainty estimates, relying on empirical variance, which can be less effective for exploration.
TPE	Problems with conditional, discrete, or mixed-type parameter spaces (e.g., hyperparameter optimization).	Models density estimators for good/bad trials instead of a direct surrogate. Highly efficient for its target problems, but may converge slower than GPs on simple, smooth functions.

the boundaries. The iterative search is guided by an acquisition function that balances exploration and exploitation.

The strategy for balancing exploration and exploitation in BO is chosen based on a framework that adapts to the dimensionality of the problem (d) and the total evaluation budget ($n_{\max}$). The foundation of this framework is the 10d rule, a principle established in the literature suggesting that a budget of at least $n_{\text{base}} = 10d$ is required to reliably estimate the hyperparameters of a GP surrogate [225]. The relationship between the total budget ($n_{\max}$) and this

baseline sample requirement defines the selection of the optimization strategy in three distinct regimes.

$$\text{Strategy} = \begin{cases} \text{Exploitation-Focused} & \text{if } n_{\max} < 2\, n_{\text{base}} \\ \text{Dynamic Annealing} & \text{if } 2\, n_{\text{base}} \leq n_{\max} \leq 10\, n_{\text{base}} \\ \text{Prolonged Exploration} & \text{if } n_{\max} > 10\, n_{\text{base}} \end{cases} \tag{4.19}$$

In extremely low-budget scenarios, where the surrogate's uncertainty estimates are unreliable, the strategy prioritizes rapid exploitation. This is achieved by using the EI acquisition function, configured with a small trade-off parameter ($\xi = 0.01$) to focus the search primarily on regions with a high probability of immediate improvement over the current best solution. Once the budget is sufficient to establish a reliable surrogate model ($2n_{\text{base}} \leq n_{\max} \leq 10n_{\text{base}}$), a dynamic annealing strategy is applied. This strategy enables a smooth transition from initial global exploration to later local exploitation by adjusting the exploration parameter κ of the UCB acquisition function over time. The decay of κ follows the schedule proposed by Zhang et al. [317]:

$$\kappa(n) = \kappa_{\text{init}} \cdot \gamma^{\max(n - n_{\text{decay}})} \tag{4.20}$$

where $\kappa_{\text{init},c}$ is the initial exploration weight, γ is the decay rate with $0 < \gamma < 1$, and n_{decay} specifies the iteration at which the decay begins. This dynamic approach, illustrated in Figure 4.8, ensures that the search is both globally efficient and locally precise. This schedule is operationalized with specific default parameters: the initial exploration weight κ_{init} is set to 2.58, corresponding to the 99% confidence bound; the decay begins at $n_{\text{decay}} = 2n_{\text{base}}$ and a decay rate γ of 0.95 ensures a steady transition. In high-budget campaigns ($n_{\max} > 10n_{\text{base}}$), the primary risk shifts to premature convergence. Therefore, the strategy involves a prolonged initial exploration phase with a high constant κ value before transitioning to a decay schedule. In high-budget campaigns, where the primary risk is premature convergence, the strategy involves a prolonged initial exploration phase. A high, constant κ value of 2.58 is maintained

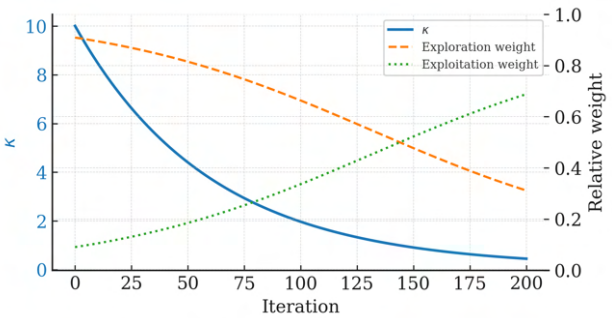


Figure 4.8: Illustration of κ decline, illustrating the shift from exploration to exploitation.

for a substantial portion of the run, until n reaches 30% of the total budget ($n_{\max}$). Only after this extended exploration phase is the same κ-decay schedule initiated, ensuring the search space has been thoroughly mapped before the algorithm focuses on final convergence.

4.7 Explainable AI and Parameter Influence Analysis

To ensure that the surrogate model recommendations are transparent and trustworthy, this methodology integrates the principles of XAI, as detailed in Section 2.3.4. Specifically, a multilayered explainability protocol is employed. This protocol aims to provide a comprehensive understanding of the model's behavior by progressing from a high-level, global analysis to fine-grained, local explanations. The methods employed in this work are organized into four distinct categories based on their explanatory goal, as summarized in Table 4.10.

The analysis commences with a global sensitivity analysis to identify the principal drivers of the model's output. A variance-based decomposition using Sobol indices is the preferred model-agnostic method to quantify the influence of individual features and their interactions [269]. In computationally constrained settings, the importance of rapid permutation features serves as a

Table 4.10: Overview of interpretability method clusters, detailing their purpose, logic, and model agnosticism.

Interpretability Purpose	**Methods**	**Core Explanatory Logic**	**Agnostic vs. Specific**
Global Feature Importance	Sobol' Indices, Permutation Importance, Global SHAP	Ranks features by their overall contribution to the model's output variance or prediction magnitude.	Agnostic: Sobol', Permutation, Kernel SHAP Specific: Native importances, TreeSHAP
Local Prediction Explanation	SHAP (Tree/Kernel), LIME	Decomposes a single prediction into an additive sum of contributions from each input feature.	Agnostic: LIME, Kernel SHAP Specific: TreeSHAP
Behavioral Pattern Discovery	Feature clustering, SHAP vector clustering, ICE clustering	Identifies clusters of instances that share similar structures to find hidden operational regimes.	Agnostic: Feature clustering, ICE clustering Conditional: SHAP clustering depends on the explainer
Feature Effect Visualization	PDP, ICE plots	Visualizes the average functional relationship between a feature and the model's prediction.	Agnostic: PDP, ICE

viable alternative. To ensure that subsequent analyses are not only insightful but also cognitively manageable for domain experts, the size of the High-Impact Feature Set (HIF-Set) is constrained. This constraint is grounded in established cognitive science and user experience research. Classic studies demonstrate that human short-term memory can effectively process only about 7 ± 2 chunks of information simultaneously [318]. More recent work on XAI confirms this, showing that users find explanations more useful when they are based on a very small number of features, with diminishing returns or even reduced clarity

beyond that point [319]. Therefore, this methodology adopts a formal constraint based on these findings. After the initial global importance analysis, the HIF-Set is truncated to include only the top 5 most influential features. This step ensures that all subsequent deep-dive analyzes are focused on a core set of drivers that are both significant and interpretable by human stakeholders. If the total number of influential features is already within this range, no reduction is made. The protocol then moves to the local attribution layer to explain individual predictions. The choice of algorithm here is matched to the surrogate model: TreeSHAP is used for tree-based ensembles due to its computational efficiency [266], while model-agnostic methods such as KernelSHAP [266] are used for other black-box models. The objective of this stage is to derive an additive attribution for each individual prediction, thereby quantifying the precise contribution of each feature to that specific output. Recognizing that manufacturing spaces are often heterogeneous, the analysis proceeds to subgroup discovery. The choice of clustering algorithm is a methodological decision guided by any a priori domain knowledge about the expected cluster geometry and data structure. For instance, if simple, spherical clusters are anticipated, k-Means is selected for its efficiency. When a natural hierarchy or nested relationships are expected, agglomerative hierarchical clustering is directly employed. For potentially elliptical and overlapping clusters, a GMM is used to model covariance and provide probabilistic assignments. Regardless of the chosen algorithm, the optimal number of clusters is determined by applying the elbow criterion alongside the average silhouette coefficient and the BIC to ensure both geometric cohesion and statistical parsimony. In the absence of such prior knowledge, a progressive validation workflow is applied to identify the best model. This process begins with k-Means as a diagnostic baseline. If its resulting clusters lack validity, the analysis proceeds to agglomerative hierarchical clustering, and subsequently to a full-covariance GMM as the most flexible parametric model if a valid structure is still not found.

The final analysis stage focuses on visual shape analysis to create intuitive narratives for domain experts and to understand the functional form of the model's learned relationships. This analysis uses ICE plots [265] and is focused

exclusively on the most important variables in the HIF set. For each feature in this set, individual ICE curves are computed across all data instances to show instance-level effects. These curves are then aggregated to produce a global partial dependency profile, which visualizes the average marginal effect of the feature on the model output.

4.8 Deployment Evaluation and Real-World Integration

Following the modeling and optimization stages, a dedicated evaluation phase is conducted to determine the practical viability of the method under real industrial constraints, as illustrated in Figure 4.9. This phase comprises four interrelated objectives. First, the environmental and economic justification is established by means of break-even analysis. Second, the robustness of the results is examined with respect to the variability of the data, the uncertainty of the model, and the operating conditions. Third, product-level quality checks are performed to verify conformity with the applicable industrial specifications. Fourth, if deficiencies are identified, the method is iteratively refined and re-evaluated until the acceptance criteria are satisfied.

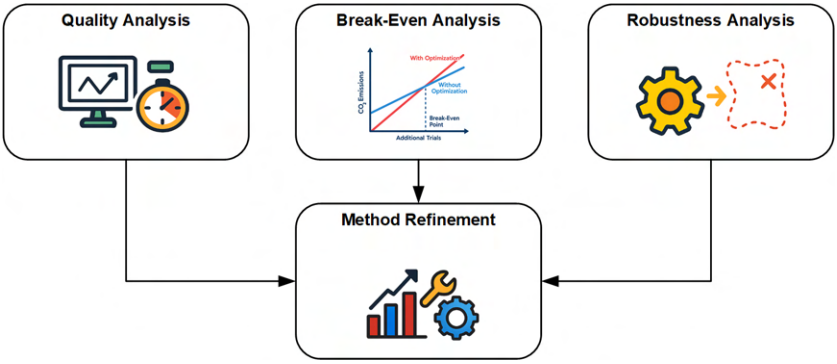


Figure 4.9: The integrated validation workflow.

4.8.1 Break-Even Analysis

To determine the production scale at which the optimization becomes viable, a break-even analysis is performed. This protocol assesses both the Environmental Break-Even Point (e-BEP) [320] and the traditional financial BEP to ensure that environmental and economic objectives converge. To quantify the scale at which any experimental or algorithmic optimization becomes environmentally and economically advantageous, first the cumulative resource burden of the optimization campaign is calculated, covering direct inputs such as energy and materials, as well as indirect contributions from labor, machine depreciation, and computational overhead - whose significance scales with the size of the experimental campaign and can be neglected for small studies - summarized in the resource-cost vector.

$$\mathbf{C}_{\mathrm{opt}} = \left(C_{\mathrm{opt}}^{€},\, C_{\mathrm{opt}}^{\mathrm{CO_2}},\, C_{\mathrm{opt}}^{E},\, \dots \right). \tag{4.21}$$

Next, a baseline is established using the standard or currently employed parameter set. This baseline typically represents a state already refined through prior experience and manufacturer recommendations. Switching from the baseline setting $\mathbf{x}_{\mathrm{base}}$ to the optimized setting $\mathbf{x}_{\star}$ produces for every unit the savings vector:

$$\Delta\mathbf{c}_{\mathrm{unit}} = \left(\Delta c^{€},\, \Delta e^{\mathrm{CO_2}},\, \Delta E,\, \dots \right), \tag{4.22}$$

whose elements quantify how much money, greenhouse gas emissions, energy, and other resources are avoided per cycle, print, batch, or service call. For any individual resource k, the minimum production volume that exactly recoups the optimization effort, is then calculated:

$$N_k^{\mathrm{BE}} = \frac{C_{\mathrm{opt},k}}{\Delta c_{\mathrm{unit},k}} \quad (k \in \{€,\, \mathrm{CO_2},\, E,\, \dots\}). \tag{4.23}$$

Because both costs and savings are typically estimated with uncertainty, a margin of safety is introduced by applying a prudence factor,$\alpha \gtrsim 1$, to derive a conservative threshold [321]:

$$N_k^{\mathrm{BE,safe}} = \alpha\, N_k^{\mathrm{BE}}. \tag{4.24}$$

The value of α is chosen based on the level of uncertainty: $\alpha \approx 1.2$ is adequate for established processes with low volatility, whereas $\alpha \geq 1.5$ is warranted for fluctuating markets, high-risk settings, or technologies with inherent uncertainty [322]. The optimization is considered viable only if the planned production volume, N_{plan}, exceeds all conservative break-even thresholds ($N_{\mathrm{plan}} > N_k^{\mathrm{BE,safe}}$ for all k). This provides the decision-maker with a clear, domain-agnostic criterion: beyond this threshold, every additional unit compounds the benefits; beneath it, the optimization remains a net burden.

4.8.2 Parallel Quality Checks

After production, a vector of quality indicators is captured:

$$\boldsymbol{Q} = \big(q_1,\, q_2, \ldots, q_M\big), \tag{4.25}$$

encompassing metrics such as dimensional accuracy and surface defects. Deviations are detected at an early stage by a streaming evaluation. The depth of inspection is governed by the functional, safety, and aesthetic requirements of the component. As detailed in Table 4.11, a spectrum of quality checks is available to address these varying levels of rigor. For instance, non-critical components such as decorative covers may only require checks from the lower end of the complexity scale, such as visual inspection and dimensional verification. In contrast, structural parts are subjected to the most robust methods, including statistical process control and destructive testing, to ensure the highest level of integrity and performance. The simplest inspection level that satisfies all risk and compliance requirements should be selected.

Table 4.11: Levels of industrial quality checks, with implementation complexity and primary outputs.

Check Type	Implementation Complexity	Primary Output / Metric
Visual Inspection	Low; manual check, minimal equipment.	Conformance to aesthetic and surface integrity specifications.
Dimensional Verification	Moderate; requires metrology tools.	Measurement of geometric deviation from a nominal design; classification of tolerance conformance.
Functional Testing	Variable; requires bespoke test rigs and procedures.	Verification of operational performance against predefined functional requirements.
Statistical Process Control	Moderate; requires integrated sensors and statistical software.	Quantification of process stability and capability; detection of non-random variation.
Destructive Testing	High; consumes samples, requires specialized lab equipment.	Characterization of material properties and failure thresholds.
Automated In-Line Monitoring	High; requires full system integration.	Continuous, real-time classification of process state; automated alerts on specification violations.

Beyond descriptive summaries, statistical tests are used to quantify whether the observed differences in $\boldsymbol{Q}$ between machines, batches, or parameter settings are statistically significant [323]. Pairwise assessments are particularly useful in quality assurance to prioritize targeted corrective actions when subtle but practically relevant changes can be detected [324].

For comparisons involving three or more groups, or for factorial designs with several factors, an omnibus analysis of variance (ANOVA) [325] is preferred

because it first tests a global null before triggering post hoc pairwise investigations and, in factorial form, can assess interaction effects. In a one-way layout with k groups and total N observations, the test statistic is

$$F \;=\; \frac{\mathrm{MS}_{\text{between}}}{\mathrm{MS}_{\text{within}}} \;=\; \frac{\mathrm{SSB}/(k-1)}{\mathrm{SSW}/(N-k)}, \tag{4.26}$$

where SSB and SSW denote the sums of squares between and within the groups, respectively. Under standard assumptions which include independence, approximate normality within groups, and homoscedasticity, F follows an $F_{k-1,\,N-k}$ distribution.

The Kruskal–Wallis test [326] provides a nonparametric counterpart to ANOVA and is preferred under non-normality or heteroscedasticity. With k groups, sizes n_j, total $N = \sum_{j=1}^{k} n_j$, and R_j the sum of ranks in group j, the test statistic is

$$H \;=\; \frac{12}{N(N+1)} \sum_{j=1}^{k} \frac{R_j^2}{n_j} \;-\; 3(N+1), \tag{4.27}$$

with standard tie correction

$$H_c \;=\; \frac{H}{1 - \dfrac{\sum_{s} (t_s^3 - t_s)}{N^3 - N}}, \tag{4.28}$$

where t_s are the sizes of tied rank blocks across all N observations. Under H_0 and for sufficiently large samples, H_c is approximately χ^2-distributed with $k-1$ degrees of freedom.

4.8.3 Robustness Validation

With respect to the robustness of optimization, the methodology adopts a hierarchical evaluation scheme, as visualized in Figure 4.10.

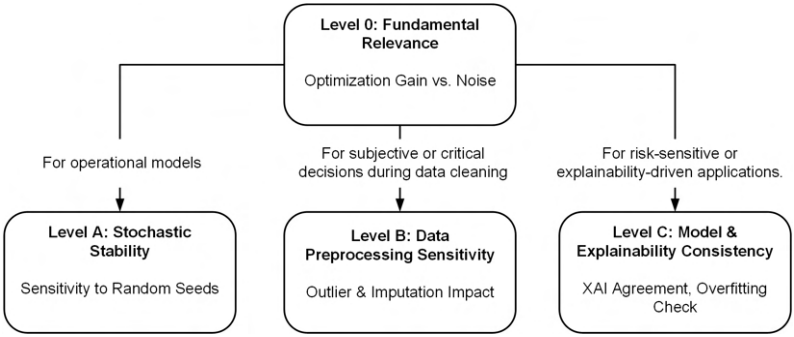


Figure 4.10: Hierarchical protocol for the robustness validation.

In the first tier, the optimization outcome itself is examined in isolation. The absolute change in the target metric, achieved by the tuned configuration, is compared with both the performance of the baseline and the variance expected from the data set noise. Only when this improvement exceeds a predefined use-case-specific relevance threshold, thereby demonstrating that the optimizer has produced a meaningful gain, the subsequent robustness layers are activated.

Level A is mandatory for every model used in offline optimization that is intended for operational use, because random initialization, data partitioning, and hyperparameter search can influence reported performance even when the data set is perfectly curated. To obtain statistically defensible performance estimates, every major source of algorithmic randomness is varied in isolation while the remaining seeds are kept constant. For each factor, $n = 20$ repetitions are performed. For each repetition, the three primary error metrics RMSE, MAE, and R^2 are recorded. The dispersion is then quantified as follows:

$$\mathrm{CoV}_{\mathrm{RMSE}} = \frac{\sigma_{\mathrm{RMSE}}}{\mu_{\mathrm{RMSE}}}, \qquad \mathrm{CoV}_{\mathrm{MAE}} = \frac{\sigma_{\mathrm{MAE}}}{\mu_{\mathrm{MAE}}}, \qquad \sigma_{R^2} = \mathrm{sd}(R^2). \tag{4.29}$$

The optimization process is classified as unstable whenever any one of the following conditions is met: the coefficient of variation (Cov) for RMSE or MAE

exceeds 5%, or the standard deviation of R^2 is greater than 0.05, corresponding to five percentage points on its bounded 0–1 scale. The protocol is summarized in Table 4.12.

Table 4.12: Stochastic stability protocol (Level A).

ID	Randomness source	Escalation trigger
A1	Global seed sensitivity	– (always executed)
A1–m	Model seed	Triggered when $\mathrm{CoV}_{\mathrm{RMSE}} > 5\,\%$ or $\mathrm{CoV}_{\mathrm{MAE}} > 5\,\%$ or $\sigma_{R^2} > 0.05$
A2–h	HPO seed	Same criterion as A1–m.
A3–d	Data-split seed	Same criterion as A1–m.

If the dispersion analysis reveals that random weight initialization or other non-deterministic operations within the learning algorithm contribute more than 5% to the total variance, three complementary lines of action are possible.

First, deterministic back-end settings should be enforced wherever the software stack allows, thereby eliminating pseudo-random effects arising from low-level parallelization or GPU kernel choices. Secondly, variance can be reduced ex post by snapshot or stochastic weight-averaging ensembles: a small set of independently initialized models is trained under identical data and hyperparameter conditions, and their predictions are aggregated by arithmetic averaging. This procedure typically yields both a lower expected error and a narrower confidence band without altering the inference logic of any single model. Third, instability attributable to the interaction between random initialization and model capacity can be counteracted in situ by tightening regularization, for example, by reducing dropout rates, increasing weight decay, or employing early stopping criteria that monitor validation error plateaus. Repeated evaluation after each intervention confirms whether the coefficient-of-variation thresholds are met. If instability persists, weight-space exploration around the converged solutions

can be expanded by a Bayesian model averaging scheme to capture the residual uncertainty.

For randomness attributed to the HPO, several measures are available. First, the trial budget can be amplified, typically doubled, to diminish the stochastic uncertainty inherent in the search process. Secondly, instead of selecting the best single configuration, it is advisable to adopt a median-of-elite strategy in which the component-wise median of the k highest scoring configurations is deployed. This approach reduces the influence of outlier trials and improves reproducibility without incurring additional optimization cost. A complementary cluster analysis of these elite configurations then clarifies whether performance plateaus exist in multiple, qualitatively different regions of the hyper-parameter space. Each regime can subsequently be subjected to a local, low-variance re-optimization to locate the most stable setting within that region. If computational resources permit, the variance can be further reduced by retaining and aggregating an ensemble of the top-k models, thereby smoothing errors while preserving the predictive mean.

Instability attributable to the train–test partition can be analyzed through LOOCV, where observations whose studentized residuals exceed the 90th percentile are flagged for manual inspection. Measurement artifacts identified in this step are either corrected or removed, whereas edge-case observations that represent legitimate but rare operating conditions are retained and their leverage attenuated by reestimating the model with a robust loss function. Should these influential points cluster in specific, underrepresented regions of the feature space, targeted data augmentation - either through additional empirical sampling or simulation - can be used to strengthen the model's generalization capacity in those areas.

Level B is used only when the training data require substantial manual cleaning or when multiple, equally plausible preprocessing options exist. Typical situations include the handling of outliers in small samples, the imputation of non-ignorable missing values, or the reconciliation of divergent data sources. If the data originate from a well-defined measurement protocol and contain

neither extreme values nor systematic gaps, Level B can be skipped. Level B examines whether specific choices in the data preprocessing pipeline alter predictive performance to an extent that may compromise downstream decision making. Two critical decision nodes are evaluated: the treatment of outlying observations and the strategy for imputing missing values. Each investigation keeps the training–test partition and the hyperparameter configuration constant, thereby isolating the effect to the preprocessing steps.

Table 4.13: Decision-sensitivity protocol (Level B).

ID	**Decision node**	**Escalation trigger**
B1	Outlier $\rightarrow$ Impute	- (always executed)
B1-i	Imputation strategy	$\lvert\Delta$error metric$\rvert > 5\,\%$ or paired test $p < 0.05$ when current pipeline is replaced by alternative pipeline.
B1-o	Outlier detection	Same criterion as B1–i.

First, the reference data cleaning pipeline is contrasted with an alternative variant that replaces the existing outlier imputation module by a different detection rule and a correspondingly aligned imputation step. All other preprocessing and modeling components remain identical. If the absolute change in any performance metric exceeds five percent or a paired significance test is performed, the node is classified as decision critical and a second stage investigation is initiated.

Regarding outlier detection, every observation that the alternative variant labels as atypical is examined individually. Influence diagnostics quantify the leverage each point exerts on the fitted model, while contextual meta-data help to distinguish measurement errors from legitimate but rare operating conditions. Observations confirmed as erroneous are removed or corrected. Rare yet valid cases are retained and either down-weighted by a robust estimation scheme or preserved through an explicit indicator so that their information content remains available without dominating the training process.

The impact of alternative imputation schemes is assessed by fitting otherwise identical models and subjecting the error differences to a paired significance test. Evidence of a systematic performance shift implies that the pattern of missingness carries substantive information about the target or the feature distribution. In such cases the imputation method that exhibits the lowest variance and the best average error is selected, and missing values are documented as a key uncertainty driver within the risk register.

Level C becomes relevant once the model predictions are exposed to risk-sensitive decision makers or inform scientific conclusions. In exploratory settings where the model serves only as a screening tool, Level C can be deferred. In regulated domains or whenever feature attributions influence downstream actions, it is executed without exception.

Table 4.14: Model and explainability consistency protocol (Level C).

ID	Consistency criterion	Escalation trigger
C1	Agreement among independent XAI methods	Kendall rank correlation between feature rankings falls below 0.6
C2	Learning-curve overfitting check	Generalization gap $E_{\text{train}} - E_{\text{val}}$ at the largest training size exceeds one standard deviation of E_{val}

The agreement among independent XAI methods evaluates whether several orthogonal explanation techniques converge on the same set of driving variables, thus establishing the credibility of feature-level interpretations. SHAP values computed using the TreeSHAP algorithm, Sobol Index importance scores and model-agnostic or ICE profiles are generated for the final estimator. Each method produces a list of influential features in a ranked order. The consistency among the lists is quantified using the Kendall rank-order correlation coefficient. A coefficient below 0.6 signals a lack of consensus that may be due to multicollinearity, target leakage, or an inappropriate representation of domain knowledge. In such cases, the feature set is reassessed, redundant variables

are merged or removed, and the model is retrained until the rankings exhibit satisfactory alignment.

Learning curves visualize the progression of the training and validation errors as a function of the effective training set size. The difference between the two curves, known as the generalization gap, reveals the capacity of the model to extrapolate beyond the samples it was fitted to. When the gap at the maximum available sample size exceeds one standard deviation of the validation error, the model is deemed to be overfit. Mitigation measures include increasing regularization strength, pruning network depth, or tree complexity, and, where feasible, augmenting the data set with new observations from underrepresented regions of the input space. The learning curve is regenerated after each adjustment to verify that the gap narrows to an acceptable bandwidth.

This cascade ensures that robustness analyses are commensurate with project stakes: a quick Level O filter prevents unnecessary effort on poorly tuned models, Level A is applied universally, while Levels B and C are reserved for cases in which data cleaning or explanatory accountability dictates deeper scrutiny.

4.8.4 Methodological Reassessment and Refinement

Whenever the gap between predicted and observed outcomes exceeds the predefined acceptance limits, the workflow enters a targeted reassessment phase. Three levers are available:

- Surrogate Model Refinement: This involves a reassessment of the surrogate model itself. Potential actions include modifying the feature set, adjusting preprocessing steps, re-optimizing hyperparameters, or selecting an alternative algorithm from the model pool, with the objective of improving predictive accuracy and the reliability of uncertainty estimates.

- Search Policy Modification: For sequential optimization campaigns, this involves adjusting the search policy to improve learning efficiency. Modifications may include altering the acquisition function to change the exploration-exploitation balance, refining parameter bounds, or changing the batch size for subsequent experiments.
- Experimental Investigation: This involves conducting further physical experiments to diagnose the source of the discrepancy. Controlled replications of previous trials are performed to differentiate between model bias and stochastic effects, such as measurement noise or unaccounted process variations.

This iterative refinement process is applied until the discrepancy between predicted and validated outcomes falls within the pre-defined tolerance, ensuring that the overall methodology remains aligned with production constraints and process variability. Combined with robustness diagnostics, this feedback loop aims to ensure that the model performance claims are valid and hold in real-world manufacturing environments.

4.9 Summary of the Methodology

This methodology presents a framework for process optimization in data-constrained manufacturing environments. To address gaps identified in the literature regarding data scarcity, model explainability, and lack of integrated workflows, the framework is built on principles of modularity and open source adaptation. It begins with a problem definition phase, where the optimization task is situated within the product engineering lifecycle, technical objectives are formalized, and these goals are aligned with environmental sustainability targets using LCA principles. A core contribution of this work is a structured decision framework for selecting between two primary optimization paradigms: Offline optimization via static surrogates and sequential optimization. This choice is guided by an analysis of strategic goals, operational constraints, and available

resources. Both pathways are contingent upon a high-quality data foundation, which is established through a comprehensive six-stage data preprocessing protocol that addresses profiling, outlier treatment, missing value imputation, scaling, feature engineering, and data fusion. For the offline path, the methodology specifies a protocol for developing and validating a high-fidelity surrogate model. This includes an algorithm selection process grounded in practical manufacturing requirements, and a nested cross-validation scheme for robust hyperparameter optimization. For the sequential path, it provides a guide for designing an efficient BO campaign by conditionally selecting the surrogate model, initial sampling strategy, and an adaptive exploration-exploitation policy that is dynamically tailored to the problem's dimensionality and the available evaluation budget. To ensure transparency and generate actionable insights, a multi-layered XAI protocol is applied post-hoc to any finalized model, progressing from global feature importance down to local, instance-level explanations and the discovery of behavioral archetypes. Finally, the framework bridges the gap to industrial practice with a deployment evaluation protocol. This final validation stage assesses the real-world viability of any proposed solution through break-even analysis, quality assurance plan, and a refinement loop, ensuring that the resulting optimizations are practically reliable in manufacturing environments.

5 Validation of the Methodology

The validation is organized into two distinct stages. The initial validation step focuses exclusively on the methodology's optimization strategy selection component. The objective is to empirically test the core hypothesis that an adaptive search strategy is more sample-efficient for finding an unknown optimum than a static, one-shot modeling approach. To achieve this, five publicly available material science datasets are used as a testbed. As these data sets are drawn from completed studies, crucial steps such as data acquisition, pre-processing, and initial robustness checks were part of the original work. This allows for a focused and controlled comparison of the optimization strategies themselves, avoiding the redundancy of reapplying the full methodological chain. This stage is crucial to confirm the foundational logic that guides the choice between iterative optimization and static surrogate modeling followed by virtual optimization. Building on the findings of Stage 1, the second stage demonstrates the end-to-end applicability and effectiveness of the entire methodology using two use cases. Both cases are situated in the field of FDM 3D printing, but are designed to test the framework under opposing data-regime conditions: one data-rich scenario and one data-limited scenario. Each of these case studies follows the complete methodological cycle of Section 4. A summary of the three validation use cases and their specific focus is provided in Table 5.1.

Table 5.1: Descriptive overview of the validation use cases.

Stage	Use Case	Key Methodological Focus
Stage 1	Empirical Validation of the optimization strategy selection	This analysis focuses on validating the optimization strategy selection logic by empirically comparing the sample efficiency of adaptive and static strategies. The trade-off between global model accuracy and optimization performance is compared.
Stage 2	CO_2 Reduction of 3D Printing with PLA	This use case demonstrates the full methodology under data-rich conditions. A surrogate model is trained on a factorial experimental design to map the process landscape of FDM printing with PLA. A suite of XAI methods is applied to quantify the influence of process parameters on the PCF and reveal trade-offs between CO_2 reduction and part quality.
	Energy Reduction of 3D Printing with PLA	This use case demonstrates the full methodology under severe data constraints. BO is employed for direct, data-efficient parameter search. Post-hoc XAI analyses are performed to interpret the identified optimum, understand the dominant energy drivers, and validate the solution for pilot-scale implementation.

5.1 Empirical Validation: Adaptive Search vs. Static Modeling

This use case empirically validates the core principles of the optimization framework by comparing the performance of an adaptive search strategy against a static, one-shot modeling approach. The objective is to evaluate which method is more effective in various application scenarios to determine a minimum. To isolate the effect of the sampling strategy, the same surrogate model, a RF, was used for both methods. This approach minimizes the surrogate model gap and ensures that the comparison primarily highlights the difference between an adaptive and a static search. In summary, two different strategies were compared:

- Adaptive Search: Sequential optimization was performed using BO with an RF surrogate and EI as an acquisition function. The search started directly with this setting, foregoing any prior DoE sampling.
- Static Modeling: First, a single, space-filling sample of a predefined size N was generated using LHS. LHS was chosen to maximize the accuracy of the regression (see Table 4.7). Second, an RF model was trained on this static data set. The location of the optimum was then predicted on the basis of this one-shot model.

The central hypothesis is that, for the task of finding a minimum, the adaptive search strategy will be more sample-efficient than the static modeling approach, especially when the total number of evaluations is small.

Five publicly available data sets were selected as a diverse test bed, representing a range of experimental domains, processes, and optimization objectives. The key properties of these data sets are summarized in Table 5.2.

The distributions of the objective values for each data set are shown in Figure 5.1, revealing varied probability densities that underscore the diverse complexity and variability in the target metrics. An upfront analysis of the landscapes, based on the metrics calculated for the complete datasets, reveals the nature of the optimization landscape.

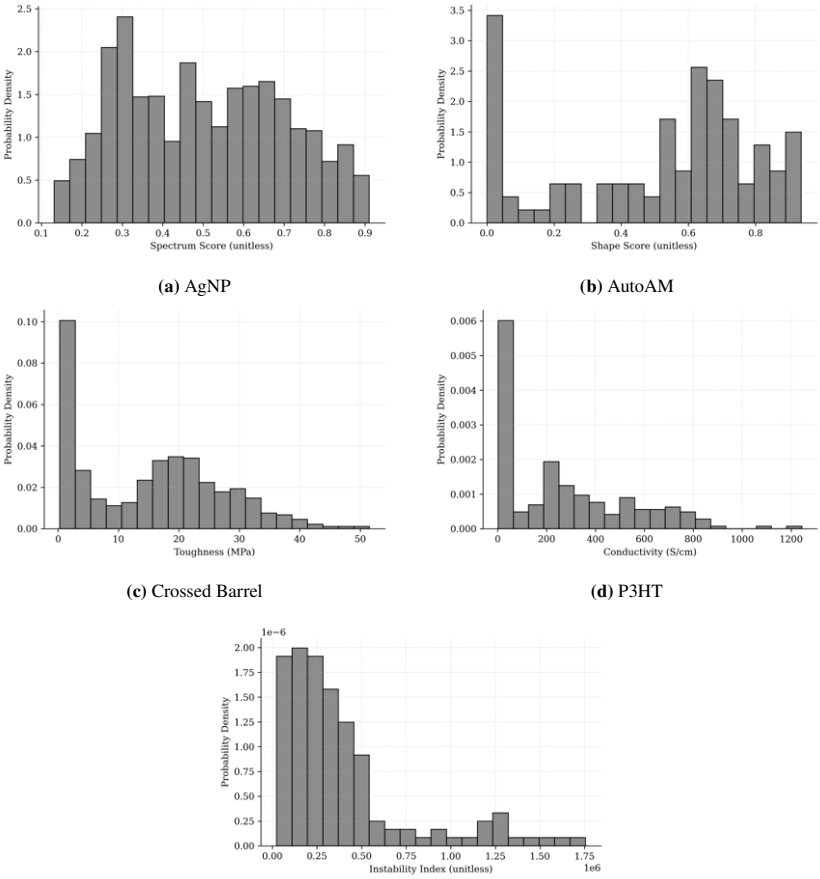
As shown in Table 5.3, the landscapes possess features that are simultaneously favorable and challenging for the static approach. High autocorrelation indicates global smoothness, yet high values for Dispersion and PIC signal local ruggedness and multimodality.

The comparative performance was tracked at various evaluation checkpoints (N), as detailed in Table 5.4. The underlying optimization problem was defined as a minimization task, with the goal of finding the lowest attainable target value. For BO, each checkpoint reports the lowest value observed after N sequential evaluations. For LHS+RF, each checkpoint corresponds to an independent one-shot experiment with a budget of N, after which the trained

model was used to predict the minimum of the response surface. To account for stochastic variability in both the sampling strategies and model training, 20 independent optimization runs were performed for each method. The values reported in Table 5.4 represent the mean and standard deviation of these repeated experiments.

Table 5.2: Material Science Datasets and their properties.

Dataset	d	Domain	Fabrication & Characterization	Objective	Size
AgNP [327]	5	Nanoparticle synthesis	Droplet-based microfluidic synthesis; in-line hyperspectral absorption measurement	Match nanoprisms absorption spectrum	120
AutoAM [328]	4	Autonomous AM	Direct-write 3D printing; image capture and geometry analysis	Minimize geometry error	100
Crossed Barrel [329]	4	Additive metamaterial structures	FDM fabrication of lattice specimens; uniaxial compression testing	Maximize toughness	1800
P3HT [330]	5	Organic semiconductors	Spin-coated thin films; Organic Field-Effect Transistor characterization to extract field-effect mobility	Maximize field-effect mobility	233
Perovskite [331]	3	Hybrid halide perovskites	Spin-coating of thin films; optical monitoring under accelerated aging for stability assessment	Maximize compositional stability	139



(a) AgNP

(b) AutoAM

(c) Crossed Barrel

(d) P3HT

(e) Perovskite

Figure 5.1: Distributions of target property across different material science data sets.

The results provide a confirmation of the hypothesis. The adaptive strategy establishes a performance advantage very early in the search process, often within the first 10-15 evaluations when the SDR is low. This efficiency gap is

maintained or widens as the budget for evaluations increases, highlighting the superiority of an adaptive approach to this task.

Table 5.3: Calculated selection metrics for the five datasets.

Metric	AgNP	AutoAM	Crossed barrel	P3HT	Perovskite
SDR	24.0	25.0	450.0	46.6	46.3
FDC	0.58	−0.51	0.42	0.22	0.65
τ	16.24	6.33	6.93	5.85	13.97
$\text{Dispersion}_{10\%}$	0.78	2.01	1.36	1.19	0.14
PIC	0.66	0.65	0.65	0.73	0.69

Table 5.4: Checkpoint-wise performance comparison for BO and LHS+RF.

N	Method	AgNP	AutoAM	Crossed barrel	P3HT	Perovskite ($\times 10^5$)
5	BO	0.45 ± 0.19	0.00 ± 0.00	3.05 ± 3.74	12.72 ± 2.55	2.43 ± 0.92
	LHS+RF	0.46 ± 0.18	0.15 ± 0.20	4.83 ± 8.62	59.64 ± 69.09	2.97 ± 1.41
10	BO	0.32 ± 0.15	0.00 ± 0.00	1.60 ± 0.81	9.63 ± 2.26	1.75 ± 0.86
	LHS+RF	0.46 ± 0.16	0.04 ± 0.09	5.97 ± 10.14	32.43 ± 38.78	3.04 ± 2.12
15	BO	0.27 ± 0.13	0.00 ± 0.00	1.44 ± 0.85	7.31 ± 3.10	1.47 ± 0.82
	LHS+RF	0.46 ± 0.16	0.05 ± 0.10	4.29 ± 5.98	36.69 ± 13.88	2.63 ± 1.40
20	BO	0.25 ± 0.10	0.00 ± 0.00	1.25 ± 0.52	7.24 ± 3.09	1.40 ± 0.73
	LHS+RF	0.45 ± 0.16	0.08 ± 0.19	5.43 ± 9.03	13.98 ± 11.98	2.16 ± 1.83
25	BO	0.24 ± 0.09	0.00 ± 0.00	1.13 ± 0.53	7.24 ± 3.09	1.18 ± 0.73
	LHS+RF	0.40 ± 0.21	0.00 ± 0.00	1.81 ± 1.07	13.98 ± 11.98	2.51 ± 1.79
30	BO	0.24 ± 0.09	0.00 ± 0.00	1.06 ± 0.43	6.41 ± 3.08	1.16 ± 0.75
	LHS+RF	0.33 ± 0.16	0.00 ± 0.00	2.05 ± 4.12	79.16 ± 81.89	1.22 ± 0.76
35	BO	0.22 ± 0.07	0.00 ± 0.00	0.99 ± 0.36	6.41 ± 3.08	1.08 ± 0.66
	LHS+RF	0.29 ± 0.13	0.00 ± 0.00	1.45 ± 1.22	42.37 ± 32.14	1.45 ± 1.12
40	BO	0.21 ± 0.03	0.00 ± 0.00	0.91 ± 0.30	6.34 ± 3.04	1.03 ± 0.65
	LHS+RF	0.37 ± 0.17	0.02 ± 0.06	1.61 ± 1.00	25.14 ± 46.11	2.17 ± 1.70
45	BO	0.20 ± 0.03	0.00 ± 0.00	0.90 ± 0.31	6.34 ± 3.04	0.80 ± 0.42
	LHS+RF	0.35 ± 0.16	0.03 ± 0.09	3.80 ± 10.17	18.48 ± 19.39	1.91 ± 1.36
50	BO	0.20 ± 0.03	0.00 ± 0.00	0.90 ± 0.31	6.34 ± 3.04	0.73 ± 0.46
	LHS+RF	0.35 ± 0.15	0.01 ± 0.03	3.37 ± 6.70	24.59 ± 46.66	1.68 ± 1.22
100	BO	0.19 ± 0.02	0.00 ± 0.00	0.87 ± 0.29	5.32 ± 3.25	0.60 ± 0.48
	LHS+RF	0.34 ± 0.17	0.04 ± 0.10	1.44 ± 0.34	24.00 ± 45.51	1.61 ± 1.41

An analysis was conducted to illuminate the mechanisms underlying the observed advantage of the adaptive strategy. This investigation focuses on the fundamental trade-off between global model fidelity and local search efficiency within the domain of black-box optimization. The question under consideration is whether achieving superior optimization performance requires a globally accurate model, or whether it can result from targeted sampling in promising areas, thus offsetting a reduction in global predictive precision. Two contending hypotheses emerge from this premise:

- A model trained on a broad, space-filling sample achieves higher global predictive accuracy across the entire design space.
- An adaptive search strategy that concentrates evaluations in promising regions identifies a better optimum value, even when its global model is less accurate.

To directly assess this, both methodologies were re-trained on 80% subsets of randomly chosen data samples and then independently evaluated. For each subset, surrogate models were constructed, and both their overall predictive accuracy and optimization results were analyzed. The procedure was executed a total of 10 times, and the computed mean along with the standard deviation of the outcomes are shown in Table 5.5.

Table 5.5: Validation of the trade-off between global accuracy and optimization performance.

Metric	Model	Perovskite	AgNP	AutoAM	P3HT	Crossed barrel
RMSE	BO	$(2.05 \pm 0.73) \times 10^5$	0.16 ± 0.00	0.22 ± 0.05	173.93 ± 43.18	8.58 ± 1.12
	RF	$(1.22 \pm 0.24) \times 10^5$	0.047 ± 0.00	0.10 ± 0.02	122.98 ± 29.31	6.41 ± 0.38
MAE	BO	$(1.53 \pm 0.53) \times 10^5$	0.11 ± 0.00	0.17 ± 0.03	156.65 ± 26.43	7.84 ± 1.20
	RF	$(0.99 \pm 0.22) \times 10^5$	0.03 ± 0.00	0.07 ± 0.01	78.63 ± 14.68	4.46 ± 0.26
R^2	BO	0.39 ± 0.02	0.26 ± 0.06	0.38 ± 0.13	0.47 ± 0.22	0.47 ± 0.01
	RF	0.88 ± 0.02	0.94 ± 0.01	0.89 ± 0.05	0.77 ± 0.05	0.70 ± 0.04
Best Found y	BO	$(2.37 \pm 0.00) \times 10^4$	0.18 ± 0.07	0.00 ± 0.00	4.74 ± 2.82	0.38 ± 0.00
	RF	$(1.15 \pm 0.38) \times 10^5$	0.22 ± 0.01	0.00 ± 0.00	7.08 ± 0.84	1.23 ± 0.00

The results confirm both hypotheses and reveal the trade-off. First, the static RF models demonstrate superior global accuracy. Across the five datasets, their RMSE and MAE values are consistently lower and their R^2 scores are substantially higher. This confirms that a space-filling static approach is effective in building a globally representative model. In contrast, the data support the second hypothesis. The adaptive BO process consistently discovers a lower minimum value. For instance, on the perovskite data set, BO finds a minimum that is almost five times lower than the best value identified from the globally superior RF model.

The previous analysis showed that an adaptive strategy is superior in finding the minimum. To investigate this, the next analysis examines whether certain landscape features further enhance that advantage. Therefore, a correlation analysis was performed linking the initial landscape metrics, from Table 5.3, to the performance gap measured between the adaptive and static methods.

Table 5.6: Correlation coefficients between problem metrics and average performance difference over all checkpoints for the minimization task. Negative values indicate an advantage of BO.

Metric	**Correlation**
Problem dimension, d	-0.51
FDC	$+0.03$
Autocorrelation length, τ	$+0.48$
10% Dispersion	-0.07
PIC	-0.88

The analysis reveals that there is a strong negative correlation of -0.88 with PIC. This indicates that local ruggedness is the most important feature that determines the advantage of an adaptive search. The more complex the local landscape, the more an unintelligent, static sampling plan is punished and an intelligent, adaptive search is rewarded. The moderate negative correlation with dimension aligns with the known strengths of BO in lower-dimensional spaces. The autocorrelation length τ shows a positive correlation of $+0.48$, suggesting

that long-range dependency slightly reduces the relative performance of BO. The FDC and 10% dispersion metrics show correlations near zero, which implies that these metrics have minimal predictive value for the comparative performance of the two methods of these datasets.

The results presented in this chapter confirm a critical trade-off between optimization efficiency and global process understanding. For the specific task of expensive black-box optimization, an adaptive sampling strategy is superior to a static DoE approach. In complex landscapes, especially those identified as globally smooth but locally rugged, an uninformed space-filling sample is inefficient because it wastes evaluations in uninteresting regions. An adaptive method such as BO, on the contrary, learns from every data point to progressively focus its search, demonstrating superior sample efficiency. However, this targeted efficiency reveals the drawback of the adaptive search approach. Focusing samples in areas with high potential to rapidly locate the optimum can result in the method overlooking the development of a comprehensive global understanding of the entire search space. The resulting surrogate model can be highly accurate near the discovered optimum but is unreliable in unexplored regions, making it less suitable for generating global process insights.

5.2 CO_2 Optimization of 3D Printing with PLA

The first use case was initially introduced by Hauck et al. [4], and is explored in depth in this section in relation to the proposed methodology. The use case was developed within the scope of the AI-assisted Technology Transfer (AITT) research project, funded by the German Federal Ministry for Economic Affairs and Climate Action (BMWK) under the Leichtbau technology transfer program (TTP). The project's core objective is to leverage AI techniques within the product development process to systematically reduce CO_2 equivalents. The research project focuses on two distinct technological processes: 3D printing and injection molding. In this context, the application explored is 3D printing, using PLA, which was selected for its origin from renewable resources, its

biodegradability, and its low toxicity during printing, as described in Table 2.1. For transferability, the use case investigates the production of standardized tensile test specimens according to ASTM D638 Type I, which also conforms to the ISO 527 and GB/T 1040 standards [21].

5.2.1 Definition of Objectives and Scope

The PLA use case is assigned to the testing and refinement phase in the lifecycle of Eq. (4.1). This assignment follows directly from the observable indicators summarized in Tables 4.1 and 4.2. The geometry and the material are fixed, which places the task after detail design. The primary objective is the screening and refinement of process parameters to reduce CO_2-equivalents subject to basic mechanical acceptance, which matches the purpose of testing and refinement. The data situation consists of structured experimental logs and short run histories rather than long operational records, which is characteristic of testing and refinement rather than pilot production or later phases. Parameter changes are still freely admissible between runs, indicating a high degree of freedom of change. The quality regime uses standardized tensile tests on ASTM D638 Type I (ISO 527, GB/T 1040) specimens [21] to verify mechanical performance, which is consistent with engineering acceptance criteria in testing and refinement. The decision cadence is organized in campaigns with periodic technical reviews rather than daily or shift-based production reviews, which again aligns with testing and refinement. The use case does not qualify as pilot production because there is no pre-series control plan, no initial capability indices, and no batch gate approvals. It is also not part of the design phases because the manufacturing method and the material are no longer open design variables. By these cues, the most plausible phase is testing and refinement, which we formalize as $\widehat{\mathcal{P}} =$ testing and refinement.

The single quantitative objective is to minimize the product carbon footprint (PCF), expressed directly in grams of CO_2-equivalent. The PCF is calculated following the IPCC Guidelines for National GHG Inventory [49], with further

methodological refinements at the process level as described in [332]. The total equivalents ($E_{\text{total},i}$) associated with each printed part i are segmented into Scope 2 and Scope 3:

$$E_{\text{total},i} = E_{1,i} + E_{2,i} \tag{5.1}$$

Scope 2 emissions ($E_{2,i}$) derive from the electrical energy consumption (P_i) associated with the printing process and are calculated using the grid-specific emission factor (EF_2):

$$E_{2,i} = P_i \cdot EF_2 \tag{5.2}$$

According to the German Environment Agency [333], the emission factor for the German electricity grid is $EF_s = 388$ g CO_2-eq/kWh. Scope 3 emissions ($E_{1,i}$) are determined based on the mass of the printed part (m_i) multiplied by the PLA-specific emission factor (EF_1):

$$E_{1,i} = m_i \cdot EF_1 \tag{5.3}$$

The emission factor for PLA ($EF_1 = 3.22$ kg CO_2-eq/kg) is sourced from the EcoInvent database, referencing data from one of the world's largest PLA production plants in Nebraska [334]. In addition, visual inspection and sampled tensile strength tests will be used to assess the quality of the parts.

5.2.2 Experiment Planning and Optimization Strategy

Four parameters of the FDM process were chosen based on recommendations from the literature on key factors that influence the PCF of FDM printing: layer height [335], infill density [335], number of perimeters [336], and nozzle temperature [337].

This use case pursues two objectives: first, to obtain a global, interaction-sensitive understanding of the process landscape, and second, to identify a parameter setting that minimizes the PCF. Because both energy consumption and part mass are available immediately after each print, the target metric falls

into latency class L0 according to Table 4.3. For the part, no historical data exist, and the available budget allows for roughly 100 experiments. Therefore, the scenario is classified as the start class C2 in Table 4.4. Within the framework, an L0–C2 problem can be addressed either by an iterative sequential optimizer or by a single-batch DoE. Both paths satisfy the latency and budget constraints. In this study a full factorial DoE was chosen because the sub-goal of obtaining a comprehensive understanding of parameter dependencies outweighs the efficiency benefits of a purely sequential search. Moreover, a comprehensive factorial plan tests every combination of factor levels, permitting orthogonal and unbiased estimation of both main effects and interactions, something a space-filling design would capture only indirectly and with reduced statistical power. In the complete factorial design, each parameter was systematically varied at three distinct levels, resulting in a comprehensive set of:

$$N = 3^4 = 81 \tag{5.4}$$

experiments which fits the budget of 100 trials. The three levels represent the lowest, median, and highest parameter settings. The levels of the process parameters were selected by first considering the maximum ranges recommended by the material supplier, then cross-referencing the operating limits of the 3D printer (Prusa MK4) and finally confirming the allowed min-max values in the slicing software. The final parameter ranges are shown in Table 5.7.

Table 5.7: Selected Parameter and parameter levels for the second use case.

Level	Layer Height	Infill Density	Perimeters	Nozzle Temp.
01	0.16 mm	15%	2	190 °C
02	0.22 mm	57.5%	4	205 °C
03	0.28 mm	100%	6	220 °C

After performing the complete factorial design, to confirm that the search for the optimum is possible with a static surrogate model the landscape metrics of Table 4.6, are calculated.

Table 5.8: Computed Landscape metrics.

Metric	Value	Interpretation
SDR ρ	20.25	medium-high observations per factor
Autocorrelation length τ	5.11	smooth global landscape
PIC $M(\phi)$	0.52	low-moderate number of local optima
Dispersion D	0.35	clustered optima

The landscape metrics presented in Table 5.8 imply an environment that is reasonably data-rich. The autocorrelation length signifies a smooth global landscape. The moderate values of the partial information content and the low values of the dispersion indicate a landscape with a moderate level of ruggedness, featuring clustered optima. These attributes collectively support the use of a surrogate modeling approach.

5.2.3 Data Preprocessing

The initial phase of the analysis involved a comprehensive data preprocessing workflow designed to ensure the quality and integrity of the data set. This process began with data profiling, where the range and empirical quartiles of all variables were summarized in Table A.1 of the appendix, to establish a baseline understanding of their distributions. The subsequent preprocessing steps, outlier detection and data imputation, were exclusively focused on constructing the cumulative energy consumption. All recorded process settings serve as input features and are complete, so no feature-aggregation of the PCF relied on the raw power trace from the printer, which required outlier treatment and data imputation. The initial phase in processing the power trace data involved identifying and managing outliers. For this task, the MAD rule was used with

a threshold parameter of $\kappa = 3$, followed by the application of an isolation forest algorithm. This procedure identified approximately 3.4% of the samples as outliers. The workflow proceeded to address the gaps within the power trace, where the existing gaps of 1.9% combined with 3.4% of outliers resulted in an overall of 5.3%. According to the workflow in Section 4.4, a prerequisite for selecting an appropriate imputation method is to determine the nature of the missingness. Therefore, a binary missing indicator vector was created to map the location of the gaps. Visual inspection of this map revealed a pattern of only isolated, single-sample gaps scattered uniformly throughout the build time. This observation provided preliminary qualitative evidence suggesting an MCAR environment. To substantiate this hypothesis, Little's MCAR test was performed. The test yielded a p-value of $p = 0.89$. This result provides statistical confirmation that the data satisfy the MCAR assumption. The confirmation of the MCAR mechanism, combined with the small overall fraction of missing data, informed the choice of the imputation strategy. According to Section 4.4, the missing power readings were filled by linear interpolation, limited to runs of at most five consecutive missing samples. Subsequently, no absent entries were identified within the dataset.

Because the experiment employs a full factorial design that provides an almost uniformly sampled design space, no feature scaling or standardization is required, and the original units of the input parameters are retained to preserve their physical meaning.

The original units of the input parameters are preserved to retain physical meaning. The total energy per build was obtained by trapezoidal integration of the discrete measurements.

$$E = \sum_{i=1}^{n} \frac{(t_i - t_{i-1})}{2} \left(P_i + P_{i-1}\right), \tag{5.5}$$

where P_i and P_{i-1} are successive power readings at times t_i and t_{i-1}, respectively. Based on this aggregation, the weight of the parts and the emission factors, the CO_2-equivalent emissions of the parts can be calculated. As shown

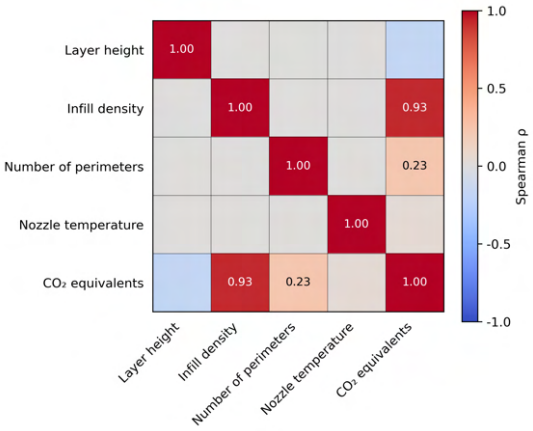


Figure 5.2: Correlation matrix of the process parameter.

in Figure 5.2, the input parameters exhibit negligible interdependencies: only their individual relationships with the target stand out, so no additional feature engineering, dimensionality reduction, or redundancy removal is required.

In the last step, data fusion, the CO_2-equivalents are merged with the corresponding input parameters to obtain the final data set for analysis.

5.2.4 Offline Optimization via Static Surrogates

A series of machine learning models was trained to predict CO_2-equivalent emissions based on the process parameters. To begin with, the full dataset was shuffled using a fixed random seed. The data were then partitioned into training (70 %), validation (15 %) and test (15 %) set. The hyperparameters for each model were optimized using Optuna with an identical budget of 100 trials. During each trial, a three-fold CV was performed on the training and validation splits, with the RMSE on the held-out validation fold serving as the objective. Early stopping mechanisms were applied wherever available to prevent overfitting given the limited sample size. The final performance of the

model was evaluated in an outer loop using a five-fold repeated CV alternately holding out each fold as a test set. To quantify the variability introduced by random initialization and data splits, the entire nested CV procedure was repeated using ten distinct random seeds. All sources of randomness, including data set shuffling, fold assignments, and the Optuna sampler, were controlled by these seeds. For each model, Table 5.9 reports the mean and standard deviation of the outer loop metrics R^2, MAE, and RMSE. In line with the criteria summarized in Table 4.8, tree-based models, AutoML systems, and GAMs were considered well suited to the dataset, and transformer-based architectures were additionally evaluated to validate this selection empirically.

Table 5.9: Predictive Performance Metrics of the Regression Algorithms.

Model	**RMSE**	**MAE**	R^2
XGBoost	0.83 ± 0.64	0.59 ± 0.30	0.96 ± 0.06
CatBoost	0.84 ± 0.60	0.61 ± 0.33	0.96 ± 0.05
LightGBM	0.87 ± 0.57	0.59 ± 0.25	0.96 ± 0.05
RandomForest	1.03 ± 0.63	0.74 ± 0.33	0.94 ± 0.07
AutoSklearn	1.21 ± 0.77	0.74 ± 0.35	0.90 ± 0.16
AutoGluon	1.39 ± 0.86	0.99 ± 0.56	0.92 ± 0.07
TabPFN	2.81 ± 0.44	2.51 ± 0.41	0.66 ± 0.25
EBM	2.94 ± 0.52	2.44 ± 0.54	0.69 ± 0.04
SAINT	5.82 ± 1.11	4.98 ± 1.06	-0.53 ± 1.03
TabNet	31.68 ± 3.62	30.76 ± 4.60	-39.13 ± 18.78

The results show that tree-based techniques are predominant, with XGBoost and CatBoost exhibiting the lowest RMSE and MAE, coupled with the highest R^2. The errors achieved are minimal compared to the absolute CO_2-equivalent emissions, which span approximately 30.4 to 44.2 g per part, corresponding to relative errors of under 2 %. AutoML frameworks deliver robust results. However, they still fall slightly short of the best tree-based algorithms. GAMs, represented by EBM, yield performance on par with TabPFN. They outperform other deep learning methods, such as SAINT and TabNet, which exhibit

markedly poor metrics. Nevertheless, they remain less accurate than the leading models. The moderate to high standard deviations observed suggest that factors such as random initialization and data splitting introduce substantial variability, which will be further evaluated in the following robustness analysis.

5.2.5 Explainable AI and Parameter Influence Analysis

Following the methodology's recommendation, the multi-method explainability strategy was used. First, a global sensitivity analysis was performed using Sobol first- and total-order indices.

Table 5.10: Sobol first-order (S_1) and total-order (S_T) indices.

Feature	S_1	S_T
Infill Density	0.86	0.92
Nozzle Temperature	0.01	0.01
Layer Height	0.11	0.05
Number of Perimeters	0.00	0.07

The Sobol first-order indices reveal that infill density alone explains the majority of variance in predicted CO_2-equivalents. The layer height has a modest direct effect, whereas the number of perimeters contributes only minimally. Finally, the nozzle temperature has essentially no direct influence on the model output.

The total order indices confirm that the total contribution of the infill density remains the largest ($S_T \approx 0.92$), indicating that interactions involving the infill density account for only a small additional fraction of the variance. Layer height and number of perimeters exhibit small increases from S_1 to S_T ($S_T \approx 0.05$ and 0.07), while the nozzle temperature remains negligible ($S_T \approx 0.01$). Furthermore, since only four input parameters were analyzed, the construction of a set of high-impact factors was unnecessary.

Next, the global SHAP summary in Figure 5.3 further revealed both the dispersion and the directional impact of each factor.

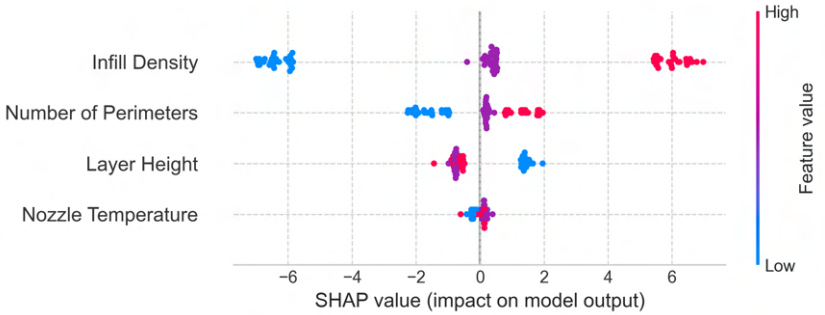


Figure 5.3: SHAP summary plot showing the distribution and sign of feature contributions for all parts.

High infill densities consistently shift predictions upward, indicating increased CO_2-equivalents, whereas low infill densities suppress them. Likewise, the number of perimeters exhibits a comparable effect, where a reduction in the number of perimeters corresponds to a decrease in CO_2-equivalents, while an increase in the number of perimeters results in elevated levels of CO_2-equivalents. In contrast, layer height demonstrates an inverse relationship, whereby reducing layer height results in elevated CO_2-equivalents, whereas increasing layer heights leads to diminished CO_2-equivalents. The clustering of nozzle temperature values tightly around zero reaffirms its insubstantial role within the explored parameter window.

Furthermore, the SHAP decision plot in Figure 5.4 traced the cumulative contributions of the global mean prediction to each specific estimate for each part. Reduced infill densities have the potential to decrease CO_2-equivalents by as much as six units, whereas elevated infill densities can increase them by up to ten units beyond the average. Modifications in the perimeter count and layer

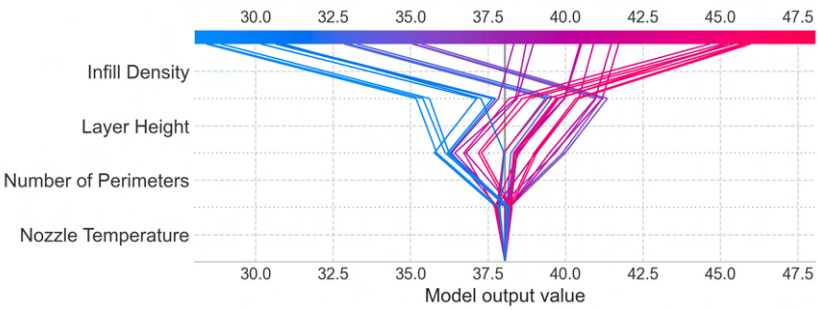


Figure 5.4: SHAP decision plot illustrating how individual feature contributions accumulate from the global mean to each part prediction.

height take place solely subsequent to the primary infill effects, with the near-flat trajectories of nozzle temperature highlighting its minimal impact on the output.

Subsequently, to explore latent groupings within the multidimensional process parameter space, unsupervised clustering techniques were used. As no inherent hierarchical structure was anticipated in the process parameter space, k-means clustering was first tested. However, the algorithm did not produce clearly separated groups. To allow for ellipsoidal cluster shapes and probabilistic membership, a GMM was used. With the GMM, the data points could be satisfactorily grouped. To identify the optimal number of clusters, the BIC and the silhouette coefficient were plotted as functions of the cluster number, as depicted in Figure A.2 of the appendix. Despite a conflicting peak in the silhouette score at $k = 2$, the selection of five clusters was guided by the clear elbow in the BIC curve at $k = 5$ and its associated acceptable silhouette value. To visualize the high-dimensional cluster structure, a PCA was employed. Figure 5.5 displays the resulting clusters in a three-dimensional projection on the first three PCs, which collectively account for approximately 80% of the total variance.

The interpretation of these axes, based on the feature contributions in Table 5.11.

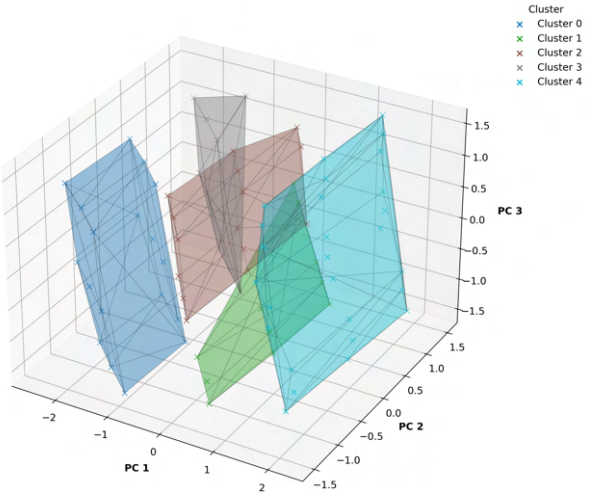


Figure 5.5: Three-dimensional PCA projection colored by the five-component GMM.

Table 5.11: PCA feature contributions to the first three components.

Feature	**PC 1**	**PC 2**	**PC 3**
Layer height	0.12	0.00	-0.99
Infill percentage	-0.68	0.09	-0.16
Perimeters	-0.16	-0.27	-0.04
Nozzle temperature	-0.02	-0.96	-0.00
CO_2 equivalents	-0.71	0.00	0.00
Explained variance	0.40	0.20	0.20

PC 1 alone accounts for 40 % of the total variance and is highly correlated with both the infill density and the CO_2-equivalents. Therefore, the horizontal spread along the PC 1 axis in the scatter plot visualizes the carbon efficiency gradient, where movement along the positive axis leads to prints with a higher PCF. The vertical variation along PC 3 reflects differences in layer height, whereas movement along PC 2 indicates shifts in nozzle temperature and, to a lesser extent, perimeters.

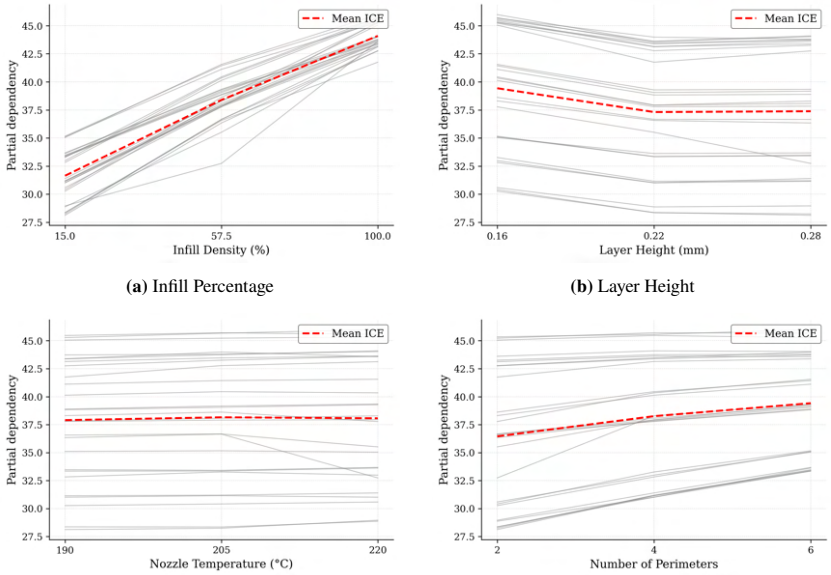
Table 5.12: Cluster sizes, mean carbon footprint and process parameters.

Cluster	n	**CO_2e [g]**	**Layer height [mm]**	**Infill [%]**	**Perimeter [No.]**	**Nozzle temp. [°C]**
0	18	31.60	0.22	15.00	4.00	197.50
1	11	39.60	0.17	57.50	3.80	207.70
2	15	38.10	0.25	57.50	4.30	202.00
3	10	31.70	0.23	19.30	3.80	220.00
4	27	44.10	0.22	100.00	4.00	205.00

The analysis of the characteristics of the cluster summarized in Table 5.12 revealed several different process regimes that align with the previous analysis. Particularly noteworthy is Cluster 4, which represents the highest-emission regime with a mean PCF of 44.10 g. All prints were made with 100% infill using otherwise varying process parameters. In contrast, clusters 0 and 3 are both characterized by a low infill percentage and, consequently, a low PCF. Their primary differentiation lies in the nozzle temperature, in which Cluster 0 uses a cool nozzle, Cluster 3 operates with a hot nozzle. The remaining Clusters 1 and 2 comprise prints with a medium infill percentage and fall into the mid-range for CO_2-equivalents. They differ primarily in their combination of layer height and nozzle temperature. Cluster 1 combines very fine layers with a warm nozzle, while Cluster 2 utilizes coarse layers at a moderate temperature.

Lastly, to analyze the shape and direction of the influence of the process parameters, Figure 5.6 shows the ICE curves for the CO_2-equivalents as each process parameter is varied, while all others remain constant. The vertical grid lines mark the discrete settings used in the factorial design, and the y axis shows the CO_2-equivalents.

From these ICE plots, it is confirmed that infill density exerts by far the strongest influence on CO_2-equivalents. Increasing density from 15 % to 25 % and again from 25 % to 35 % produces mean emission increases on the order of 4–5 g, and the mean ICE curve rises almost linearly across the full 15–100 % range.



(a) Infill Percentage

(b) Layer Height

(c) Nozzle Temperature

(d) Number of Perimeters

Figure 5.6: ICE and mean PDP of CO_2 equivalents of the input parameters.

In contrast, the nozzle temperature does not have an effect, as its ICE curves are nearly flat, and the mean line stays at approximately 37.8 g throughout the 190–230,°C interval. The layer height shows a small negative relationship, with predicted equivalents falling by approximately 2 g as the layer height increases from 0.16 mm to 0.28 mm, indicating that the coarser layers slightly reduce CO_2-equivalents. Finally, the number of perimeters produces a moderate increase in CO_2-equivalents emissions. The tight clustering of individual gray ICE curves for infill density further confirms the Sobol sensitivity analysis, as nearly all observations respond to changes in infill density, producing a consistent upward shift in predicted CO_2-equivalents.

5.2.6 Validation and Real-World Integration

The process of validation and real-world integration commences with conducting the BEP analysis, which is then pursued by an assessment of the part's quality and subsequently, the robustness analysis of the employed machine learning models.

To assess whether the optimization campaign compensates for its resource demands, the initial expenditures with the cost savings per part achieved will be compared. The optimization campaign comprises 81 prints and consumed $E_{\mathrm{Opt}} = 2.20\,\mathrm{kWh}$ of electricity together with $m_{\mathrm{Opt}} = 0.694\,\mathrm{kg}$ of PLA. Using 2024 German industrial tariffs, $c_E = 0.1899\mathrm{EUR}\,(\mathrm{kWh})^{-1}$ [338] for electricity and $c_{mat} = 20\,\mathrm{EUR\,kg^{-1}}$ [339], the direct monetary outlay of the campaign is

$$C_{\mathrm{Opt}} = E_{\mathrm{Opt}}\,c_E \;+\; m_{\mathrm{Opt}}\,c_{mat} = 14.30\,\mathrm{EUR}. \tag{5.6}$$

The associated CO_2 equivalents based on the German grid factor $EF_E = 0.388\,\mathrm{kg\,CO_2\,(kWh)^{-1}}$ [333] and the PLA factor $EF_{mat} = 3.22\,\mathrm{kg\,CO_2\,kg^{-1}}$ [334], is

$$I_{\mathrm{Opt}}^{\mathrm{GWP100}} = E_{\mathrm{Opt}} EF_E \;+\; m_{\mathrm{Opt}}\,EF_{mat} = 3.09\,\mathrm{kg\,CO_2e}. \tag{5.7}$$

The baseline mirrors slicer defaults (layer height $0.22\,\mathrm{mm}$, $57.5\,\%$ infill, four perimeters, nozzle temperature $205\,^{\circ}\mathrm{C}$), while the optimum is the run that minimizes CO_2-equivalents while remaining inside the experimental design space (layer height $0.28\,\mathrm{mm}$, $15\,\%$ infill, two perimeters, $190\,^{\circ}\mathrm{C}$). Switching from baseline to optimal settings reduces per part the electricity demand by $\Delta E = 8.6 \times 10^{-3}\,\mathrm{kWh}$ and the usage of PLA by $\Delta m = 2.06\,\mathrm{g} = 2.06 \times 10^{-3}\,\mathrm{kg}$, giving a monetary savings of

$$\Delta c_{\mathrm{part}} = \Delta E\,c_E \;+\; \Delta m\,c_{mat} = 0.042\,\mathrm{EUR\,part^{-1}}. \tag{5.8}$$

The corresponding PCF reduction is

$$\Delta I_{\text{part}}^{\text{GWP100}} = \Delta E \; EF_E \; + \; \Delta m \; EF_{mat} = 9.96 \times 10^{-3} \, \text{kg} \, \text{CO}_2\text{e}. \quad (5.9)$$

Therefore, the optimization campaign amortizes after

$$N_{\text{BE}}^{\text{cost}} = \frac{C_{\text{Opt}}}{\Delta c_{\text{part}}} \approx 3.35 \times 10^2 \, \text{parts}, \quad (5.10)$$

$$N_{\text{BE}}^{\text{CO}_2\text{e}} = \frac{I_{\text{Opt}}^{\text{GWP100}}}{\Delta I_{\text{part}}^{\text{GWP100}}} \approx 3.10 \times 10^2 \, \text{parts}. \quad (5.11)$$

Applying the prudence factor $\alpha = 1.2$ to the previously calculated break-even points gives conservative thresholds:

$$N_{\text{BE,safe}}^{\text{cost}} = \alpha \, N_{\text{BE}}^{\text{cost}} = 1.2 \times 3.35 \times 10^2 \approx 4.02 \times 10^2 \, \text{parts} \quad (5.12)$$

$$N_{\text{BE,safe}}^{\text{CO}_2\text{e}} = \alpha \, N_{\text{BE}}^{\text{CO}_2\text{e}} = 1.2 \times 3.10 \times 10^2 \approx 3.72 \times 10^2 \, \text{parts} \quad (5.13)$$

These thresholds are consistent with typical batch sizes in prototyping for small and medium sized enterprises. They should be read as cautious targets rather than exact cutoffs. Electricity tariffs in Germany vary over time, and a higher tariff would shift the economic threshold upward, while a lower tariff would move it downward. The cost model focuses on consumables, and adding capital costs, operator time, and post-processing would raise the financial threshold to some extent, which is common in AM studies. The computation for surrogate modeling also consumes energy and money, so including that effort would moderately increase the payback horizons for costs and for carbon. In practice, not every pre-optimization print is scrap. If some early parts are used, the incremental benefits per part are smaller, but the method can still deliver value

across the series. Overall, safe counts remain attainable for the PLA use case and provide a robust planning figure for both cost and PCF savings.

In conjunction with reducing the CO_2 footprint, a comprehensive quality assessment was performed. This evaluation included visual inspection and evaluation of mechanical properties by tensile testing. The visual inspection revealed that parameters involving the lowest nozzle temperature setting exhibited small extrusion issues, such as non-rectangular outer contour or irregularities in the print image itself, as shown in Figure 5.7.

Figure 5.7: Side view of all samples, highlighting extrusion defects on specimens using `N01` (non-rectangular outer contour/curvature (blue) and irregularities in the print image (red)), taken from Hauck et al. [4]

To assess mechanical performance, tensile tests were performed on a random subsample of the printed specimens. The tensile strengths of the nine samples analyzed ranged from 13.18 MPa to 25.57 MPa, indicating a variation of approximately 94%. Table 5.13 lists the tensile strengths and the corresponding process parameters for each specimen.

Table 5.13: Tensile strength values of the analyzed specimens, adapted from Hauck et al. [4].

Experiment No.	Layer Height [mm]	Infill Density [%]	No. of Perimeters	Nozzle Temperature [°C]	Tensile Strength [MPa]
1	0.16	15.0	2	190	16.67
14	0.16	57.5	4	205	19.03
27	0.16	100.0	6	220	25.57
33	0.22	15.0	4	220	15.18
43	0.22	57.5	6	190	18.89
47	0.22	100.0	2	205	24.25
62	0.28	15.0	6	205	13.18
66	0.28	57.5	2	220	17.88
76	0.28	100.0	4	205	21.87

Since there are only minimal dependencies between the input parameters, an analysis of the main effects can be performed. The mean tensile strength was calculated at different levels for each factor. The findings are presented in Table 5.14.

Table 5.14: Main effects: mean tensile strength [MPa] by factor level.

Level	Layer Height	Infill Density	No. of Perimeters	Nozzle Temperature
Minimum	20.42	15.01	19.60	17.78
Median	19.44	18.60	18.69	19.58
Maximum	17.64	23.90	19.21	19.54

When the layer height was increased from 0.16 to 0.28 mm, the mean tensile strength decreased from 20.42 to 17.64 MPa, indicating that finer layers improve interlayer adhesion and reduce stress concentrators. The density of the infill proved to be the most influential factor, with the average strength rapidly increasing from 15.01 MPa at 15 % infill to 23.90 MPa at 100 % infill. This

reflects the increased cross-sectional load and the reduced internal porosity associated with denser fills. In contrast, the number of perimeters had only a marginal effect: strength values clustered tightly around 19 MPa across two to six perimeters, suggesting that once a minimal enclosing shell is established, additional walls contribute little to global stiffness. Nozzle temperature likewise showed no influence, as the mean strength remained effectively constant throughout the 190 - 220 °C range, implying that within these limits the viscosity of the melt and the adhesion changes do not affect the failure stress. In addition, a Kruskal–Wallis test was performed to assess whether the tensile strength distributions differ between the three levels of each factor.

Table 5.15: Kruskal–Wallis test results for process factors on tensile strength.

Factor	H **(statistic)**	p**-value**
Layer Height	0.80	0.67
Infill Density	7.20	0.027
No. of Perimeters	0.00	1.00
Nozzle Temperature	0.40	0.82

This analysis confirms that infill density is the only factor with a robust effect on tensile strength in this data set, while other factors did not produce statistically significant differences at the 5 % level, likely due to small group sizes and limited power.

Following quality analysis, a robustness study was performed to quantify the degree to which the best performing model, XGBoost, is sensitive to the various sources of randomness in the workflow.

First, the overall reduction in CO_2-equivalents is evaluated. Compared to the baseline of the slicer default with a footprint of $31.14\,\mathrm{g\,CO_2e}$ per printed part, the optimized setting lowers the footprint by $9.95\,\mathrm{g\,CO_2e}$, corresponding to a relative reduction of 31.9%. In line with the lever-based and sub-process accounting framework [298], the avoided burden is decomposed as

$$\begin{aligned} \Delta I_{\text{total}} &\approx \Delta I_{\text{mat}} + \Delta I_{\text{el}}, \\ \Delta I_{\text{mat}} &= EF_{\text{PLA}}\,\Delta m, \\ \Delta I_{\text{el}} &= EF_{E}\,\Delta E. \end{aligned} \tag{5.14}$$

The analysis attributes $6.63\,\mathrm{g\,CO_2e}$, equal to 66.6% of the avoided burden, to a reduction of $2.06\,\mathrm{g}$ in PLA consumption in the material provision subprocess. The remaining $3.34\,\mathrm{g\,CO_2e}$, about 33.4%, is the result of $8.6\,\mathrm{Wh}$ less electricity during the machine operation. Under contemporary German emission factors, material savings are the dominant contributor to GHG mitigation, and energy efficiency provides a substantial but secondary contribution. The practical relevance of the $31.9\,\%$ reduction becomes evident when viewed against the predictive uncertainty of the surrogate model. As reported in Table 5.9, the XGBoost model yielded an MAE of $0.59 \pm 0.30\,\mathrm{g\,CO_2}$, which is $1.9\,\%$ of the baseline footprint. Therefore, the optimization signal exceeds the modeling noise by a factor of approximately 17, which confirms that the observed improvement is much greater than the random variation.

Second, robustness to the data-split seed, the model initialization seed, and the Optuna HPO seed was evaluated. For each source, a five-fold CV was performed, repeated on 20 independent seeds. Table 5.16 shows the means and standard deviations for the RMSE, MAE, and R^2.

Table 5.16: Robustness study across different random-seed sources.

Source	**RMSE**	**MAE**	R^2
Data-Split-Seed	0.87 ± 0.34	0.52 ± 0.17	0.96 ± 0.05
Model-Seed	0.88 ± 0.01	0.57 ± 0.02	0.97 ± 0.00
Optuna-Seed	0.86 ± 0.01	0.55 ± 0.03	0.97 ± 0.00

The error metrics' dispersion was determined based on these results, as depicted in Table 5.17.

Table 5.17: Dispersion of the error metrics across 20 repetitions for each seed type.

Source	$\mathbf{CoV}_{\mathrm{RMSE}}$	$\mathbf{CoV}_{\mathrm{MAE}}$	σ_{R^2}
Data-split seed	38.59 %	33.40 %	0.05
Model-initialization seed	0.91 %	4.04 %	0.00
Optuna seed	1.04 %	4.73 %	0.00

Only the data-split seed violates the stability criteria, with a CoV above 30 % for both RMSE and MAE. The standard deviation of R^2 remains just below the 0.05 cutoff point, but the two error metrics already classify this source as decision critical.

Therefore, following the methodology, LOOCV was applied to the XGBoost model, to obtain an unbiased estimate of point-wise prediction error. Figure 5.8 shows the absolute LOOCV errors for all samples, and the dashed red line marks the 90th-percentile threshold.

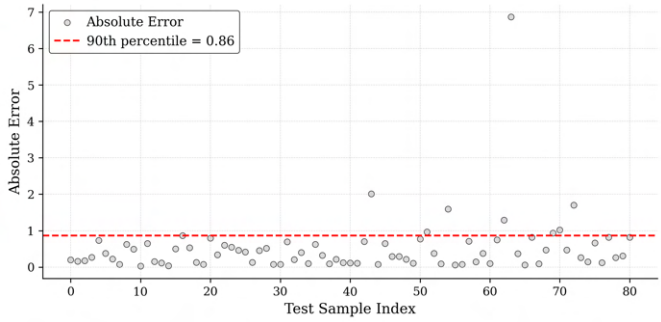


Figure 5.8: Absolute LOOCV errors for each test sample.

Using this procedure, a 90th-percentile absolute error of 0.86 g CO_2e was achieved by the model. The eight samples whose errors exceed this limit are listed in Table 5.18.

Table 5.18: LOOCV samples whose absolute error exceeds the 90th-percentile threshold.

Experiment No.	Layer Height [mm]	Infill Density [%]	Number of Perimeters	Nozzle Temperature [°C]	CO_2e [g CO_2]	Predicted CO_2e [g CO_2]
12	0.16	57.50	2	220	39.12	38.06
44	0.22	100.00	2	190	40.86	42.90
55	0.28	15.00	2	220	29.16	27.63
63	0.28	57.50	2	205	36.99	35.74
64	0.28	57.50	2	220	30.45	37.18
71	0.28	100.00	2	190	43.14	42.13
73	0.28	100.00	2	220	43.86	42.24
81	0.22	57.50	2	220	36.61	35.50

The eight high-error samples share the same shell configuration of only two perimeters, although the full factorial design also included four and six perimeters. This consistent pattern suggests that the model struggle most when the wall thickness is at its minimum. In addition, the layer height of 0.28 mm appears to be problematic, as five of the eight outliers occur in this coarsest setting. The nozzle temperature of 220 °C also coincides with five of the high-error cases. Infill density does not show a single dominant value, since midlevel density and fully solid prints appear among the outliers, but the combination of coarse layers with minimal perimeters and elevated melt temperature clearly drives the largest discrepancies. The four runs with the absolute largest errors all involve two perimeters, a 0.28 mm layer height, and a 220 °C nozzle setting, with infill densities ranging from sparse to completely solid. In particular, experiment 64 under these conditions produced an observed emission of 30.45 g CO_2 versus a prediction of 37.18 g CO_2, a difference of 6.7 g CO_2. In summary, LOOCV reveals higher prediction errors in edge-and-corner combinations because those cells probably lack replicates. Upcoming research should prioritize a few specific targeted repeats within these areas, or a more detailed resolution of continuous factors might lessen regional uncertainty. Globally, the model

remains robust. The RMSE and the 90th percentile error show that the model generalizes well for typical process points.

Moreover, to assess the applicability of the methodology's recommendation rules and verify the sufficiency of data for surrogate modeling, learning curves derived from 50 different seeds were examined. This involved analyzing both the training and validation RMSE regarding the number of training instances.

Figure 5.9 shows that the RMSE rapidly decreases to approximately 30–40 samples and then becomes asymptotically flat, indicating that adding more data beyond this elbow yields diminishing returns for model accuracy.

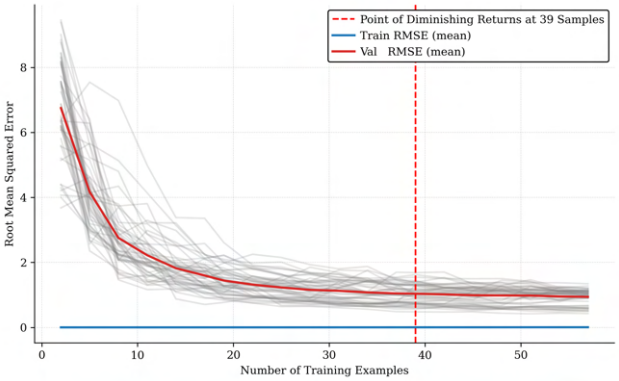


Figure 5.9: Learning-curve for XGBoost across 50 random seeds.

The ensemble of curves also reveals the variability induced by different splits, underscoring the need for around 40 well-distributed samples to achieve stable performance. Furthermore, the tight clustering of validation curves beyond the elbow demonstrates the robustness of the model to different train and test splits. To complement the analysis of the learning curve, the generalization gap was quantified. In the very early regime ($N \leq 10$) a pronounced gap of 2.2 ± 0.4 RMSE units was observed, because the model almost perfectly interpolated the sparsely sampled training points, whereas the validation sets still probed unseen

regions of the input space. As the training pool increased, the validation errors decreased. At the knee point ($n \approx 39$) it had already fallen below 1.00, and beyond $N = 39$ it stabilized around that. This behavior implies that the total error budget at larger N is dominated by irreducible process noise rather than model variance, as collecting more than 40 builds yields only marginal accuracy gains while hardly improving generalization.

To synthesize the insights from the previous two analyses, the following experiment quantifies how strongly out-of-sample performance depends on the location of training samples within the design space. The experimental space of a 3^4 full factorial design can be partitioned into five radial layers defined by the squared distance from the center, namely $r^2 \in \{0, 1, 2, 3, 4\}$. These layers, together with their radii, sample counts, and geometric interpretation, are listed in Table 5.19.

Table 5.19: Layers of the 4-dimensional cube.

No. of ± 1	**Radius r**	**Samples**	**Region**
0	0	1	4D-Centre
1	1	8	centers of the 3-D cubic facet
2	$\sqrt{2} \approx 1.41$	24	midpoints of 2-D square face
3	$\sqrt{3} \approx 1.73$	32	midpoints of 1-D edges
4	2	16	0-D corners

Two sampling protocols are considered to isolate the effect of sample location. For edge prioritization, subsets are formed by selecting points from layer 4 first and then moving inward until the target size of 40 is reached. For center prioritization, the same procedure is applied in reverse order starting with layer 0.

Figure 5.10 reports the distribution of RMSE values obtained from 200 random trainings of the XGBoost model on subsets of 40 prints, each evaluated on the complementary holdout set. All models use identical hyperparameters.

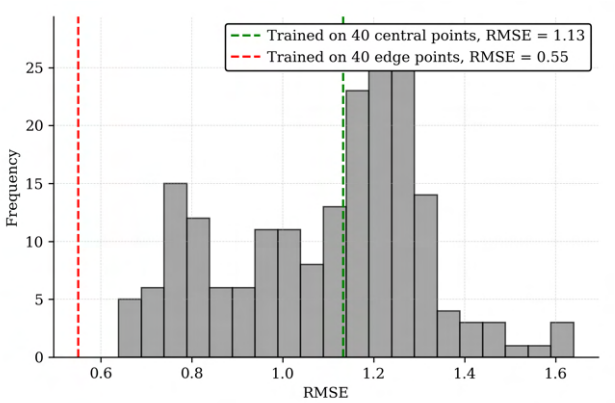


Figure 5.10: Histogram of RMSE for 200 random trainings of the XGBoost model on subsets of 40 samples. The green and red reference lines indicate models trained with center prioritization and edge prioritization.

Training on center layers yields a test error that is comparable to the middle of the random baseline. Training on edge layers reduces the error by roughly a factor of two because the remaining test points are then easier to predict. The broad spread of the histogram indicates that the RMSE can vary by about a factor of two solely due to sample choice. This finding highlights the importance of including edge and corner conditions when the data budget is limited. However, even the worst RMSE observed among the 200 runs remains acceptable, which supports the robustness of the model for practical deployment.

Next, to quantify the robustness of the energy preprocessing pipeline, the raw power traces of all builds were re-analyzed under a series of varied outlier detection schemes. Three alternative families of filters were examined. First, only using the MAD rule with varying thresholds. Second, z-score filtering was applied. Third, only using the isolation forest detector. For every variant, the complete energy reintegration was repeated for all builds, and the resulting totals were expressed as percentage deviations from the reference energies obtained. Table 5.20 displays the outcomes obtained.

Table 5.20: Combined summary of the mean deviation in integrated energy for different outlier-detection settings, relative to the approach used. Positive values indicate higher, negative values lower energy estimates.

Outlier method	Parameter	Mean Δ [%]
MAD	$\kappa = 2.0$	+1.07
MAD	$\kappa = 2.5$	+0.48
MAD	$\kappa = 3.0$	+0.03
MAD	$\kappa = 3.5$	-0.43
MAD	$\kappa = 4.0$	-0.70
z-score	$z = 2.0$	-2.02
z-score	$z = 2.5$	-1.87
z-score	$z = 3.0$	-1.81
IsoForest	contamination = 2 %	-1.63
IsoForest	contamination = 3 %	-1.62
IsoForest	contamination = 5 %	-1.57

Throughout the design space, the mean deviation never exceeded 3 %. Lower κ values admitted more local peaks and therefore produced slightly higher energies, whereas stricter filters, whether z-based or Isolation-Forest, removed additional peaks and reduced total energy consumption. Furthermore, almost all points that IsoForest would discard are already removed by the MAD rule, so the combined scheme is practically equivalent to $\kappa = 3$ alone. Importantly, the inter-build dispersion of these differences remained small, indicating that no subset of experiments was disproportionately affected by any particular filter. Consequently, the modeling results can be regarded as insensitive to reasonable changes in outlier handling, supporting the methodological soundness of the preprocessing workflow.

Lastly, to verify consistency across the explainability methods, the feature rankings from Sobol first order and total order indices, the global SHAP summary, and the qualitative slope magnitudes from the ICE curves were compared. As indicated in Table 5.21, all methods agree on two boundary findings. Infill

density is the dominant driver of CO_2 equivalents. Nozzle temperature contributes negligibly within the explored range. The ordering of the remaining two parameters shows a minor divergence. The Sobol first order index places layer height slightly ahead of the number of perimeters, whereas the Sobol total order index, the SHAP summary, and the ICE slopes indicate a slightly stronger contribution of the number of perimeters than layer height. This pattern is consistent with small interaction effects that are captured by the total order index and reflected in the model-based attributions. The conclusions for practice remain unchanged. Control of infill density should be prioritized, followed by moderate adjustments to perimeters and layer height, while nozzle temperature can be treated as a low impact factor in this window.

Table 5.21: Comparison of feature rankings across methods. Higher rank means stronger influence on CO_2-equivalents.

Method	**Rank 1**	**Rank 2**	**Rank 3**	**Rank 4**
Sobol first order S_1	Infill density	Layer height	No. of perimeters	Nozzle temperature
Sobol total order S_T	Infill density	No. of perimeters	Layer height	Nozzle temperature
SHAP global summary	Infill density	No. of perimeters	Layer height	Nozzle temperature
ICE slope magnitude	Infill density	No. of perimeters	Layer height	Nozzle temperature

A complementary view of the PCA in Table 5.11 supports these findings. The first principal component aligns with infill density and CO_2-equivalents and explains the largest share of variance, while the third and second components are driven by layer height and nozzle temperature. PCA is not a feature importance method for the predictive model, but the alignment of CO_2-equivalents with the infill axis is consistent with the explainability results above.

The robustness investigation yields several practical insights. First, data splitting primarily accounts for the significant standard deviation observed in the models. Although hyperparameter optimization seed and model seed also play a role in

the observed randomness, their impact is considerably less substantial. Second, the learning curve shows that the validation RMSE drops rapidly within the first 40 samples. This means that about 40 well-selected experiments are enough to build a surrogate model that supports optimization and scenario analysis. Third, repeated random subsampling demonstrates that the accuracy of the prediction is strongly dependent on the placement of those 40 experiments. Sets that cover the edge and corner regions of the design space achieve roughly half the RMSE of sets that concentrate on the center. Intentional sampling across all shells of the factorial cube therefore offers the highest return on investment when data are scarce. Fourth, LOOCV reveals eight cases of high errors that share minimal wall thickness, a coarse layer height of $0.28\,\mathrm{mm}$, and a nozzle temperature of at least $220\,^{\circ}\mathrm{C}$. However, in conclusion, surrogate modeling is robust enough for deployment, and no further methodological refinement is required.

Overall, this underscores that the use case outlined by Hauck et al. [4] can be executed following the methodology detailed in this work. While the example presented by Hauck et al. [4] serves as a proof-of-concept, its incorporation into this methodology offers deeper understanding of sustainable 3D printing.

5.3 Energy Consumption Reduction of 3D Printing with PLA

The second use case builds on the study introduced by Greif et al. [3] and is revisited here through the lens of the proposed methodology. The use case originates from the AI-assisted 3D printing for building materials (AIBetOn3D) research consortium funded by the BMWK within the Leichtbau-TTP program. The overarching project seeks data-efficient AI strategies that reduce CO_2-equivalents and increase material and energy efficiency in 3D printing intended for structural components. Clay and concrete extrusion offer promising low carbon manufacturing routes, yet they are experimentally demanding because costs are high and accurate simulation tools are not available. To address these constraints, the present use case employs PLA extrusion as a tractable proxy

that reproduces the essential challenges of clay and concrete printing without the associated costs and safety concerns. Furthermore, this use case mirrors the scenario faced by newer processes in which historical data are unavailable. The designed component, shown in Figure 5.11, is directly derived from a drainage system from AIBetOn3D.

Figure 5.11: The part designed for the case study, taken from Greif et al. [3].

5.3.1 Definition of Objectives and Scope

Given the stated objective and operating conditions, the energy use case is also assigned to the testing and refinement phase in the life cycle of Eq. (4.1). This assignment is based on the indicators summarized in Tables 4.1 and 4.2. The geometry and material are fixed, and the process parameters are freely adjusted between runs. The work is organized in short experimental campaigns with real-time measurement, and there is no pre-series control plan, no batch gate approvals, and no long-term operational history. The primary objective is the screening and refinement of parameter settings to reduce energy consumption according to basic acceptance criteria. These cues match testing and refinement rather than pilot production or later phases. The phase assignment is therefore to $\widehat{\mathcal{P}} =$ testing and refinement.

Compared to the first use case, the present use case adopts a single-objective formulation in which the energy consumption per part is minimized. This change has two rationales. First, a preliminary study showed that adjusting the printing

parameters for the part altered its weight by less than 2 % and, therefore, it would only act as a constant term for the optimization objective. Furthermore, the goal is to illustrate the feasibility of the methodology to allow individual operations to be optimized. However, every part will be weighed and the final CO_2-equivalents will be reported. The PCF calculation follows the calculation reported in the first use case. The quality requirements imposed on the drainage channel demonstrator were intentionally minimal because the printed PLA component serves as a correspondence sample for the forthcoming pilot trials that will employ cementitious and ceramic feedstocks. As PLA manufactured from fused filaments can never reach the compressive strength thresholds mandated for structural concrete in EN 206, destructive strength testing would have not yielded actionable insight. Instead, the decisive functional attribute is the morphology of the internal surface, which must remain sufficiently smooth to ensure unhindered water runoff throughout service. To assess this, a visual inspection will be performed.

5.3.2 Experiment Planning and Optimization Strategy

Drawing on the first use case, the parameters listed in Table 5.22 were selected for three reasons.

Table 5.22: Selected Parameter and parameter bounds for the second use case, adapted from Greif et al. [3].

Parameter	**Bounds**
Infill Mass Proportion	0–30 %
Nozzle Temperature	190–230 °C
Build Plate Temperature	30–60 °C
Layer Height	0.1–0.3 mm
Print Speed	40–150 mm/s

First, the layer height and nozzle temperature were selected to evaluate whether the insights can be transferred to other parts. However, their ranges were

broadened. The maximum nozzle temperature was raised by $10\,^{\circ}\mathrm{C}$. The minimum layer height was reduced by $0.6\,\mathrm{mm}$, and the maximum layer height was increased by $0.2\,\mathrm{mm}$.

Second, rather than prescribing the slicer's infill percentage, the present use case controls the infill–mass proportion:

$$r_{\text{infill-to-mass}} = \frac{m_{\text{infill}}}{m_{\text{tot}}} = \frac{m_{\text{infill}}}{m_{\text{shell}} + m_{\text{infill}}}, \tag{5.15}$$

that is, the ratio of lattice mass to the total filament mass required for the part. Because the shell mass m_{shell} scales with the outer surface area and the fixed shell thickness, whereas the infill mass m_{infill} scales with the interior volume, the ratio $r_{\text{infill-to-mass}}$ normalizes both dependencies. An increase in the surface-to-volume ratio increases m_{shell} and thus the denominator m_{tot}, while a bulky geometry with little surface area increases m_{infill} and likewise m_{tot}. Consequently, two parts with markedly different shapes can share the same $r_{\text{infill-to-mass}}$ even though their slicer infill percentages would need to be set to different values to realize this equality. Therefore, $r_{\text{infill-to-mass}}$ serves as a geometry-agnostic control variable, allowing a fair comparison of mechanical or environmental results between disparate designs.

Third, according to the recommendations of the recent literature, the printing speed [335] and the built plate temperature [340] were added, resulting in 5 input parameters in total.

In this use case, to mirror highly expensive concrete or clay printing, rapid convergence to optimal print settings is more important than process understanding. Similarly to the previous use case, the target metric of energy consumption becomes immediately accessible after each print, aligning with the latency class L0 as depicted in Table 4.3. Regarding the resource limits, this use case is classified as class C0 as detailed in Table 4.4, due to the absence of historical data and a budget constraint of 20 experiments, equivalent to a sample-to-dimension ratio of $\rho = 4.0$. In addition, other metrics essential for guiding the selection process remain uncomputable due to the absence of screening or historical data.

These attributes collectively support the use of an adaptive search approach in the form of a black-box optimization.

5.3.3 Data Preprocessing

In this use case, data preprocessing was performed directly following the collection of each individual design point. The high-frequency power traces were consolidated per part to calculate the total energy consumption for each part, which is the focus of the optimization. The rationale behind cleaning and outlier detection mirrored that of the initial use case, but these processes were executed sample by sample.

Pre-existing gaps affected $0.83 \pm 0.18\,\%$ of each part's power trace (range 0.55–$1.15\,\%$). Point outliers accounted for $1.65 \pm 0.41\,\%$ (range 0.98–$2.32\,\%$). In total, $2.48 \pm 0.46\,\%$ of power trace observations per run were flagged as potentially unusable.

The subsequent step involved assessing the MCAR assumption. Visual inspection of the binary missing value did not reveal any run-wise or time-wise patterns. Little's MCAR test confirmed the MCAR assumption ($p_{\text{min}} = 0.74$, $p_{\text{max}} = 0.97$), indicating that the data loss was random rather than systematic. Across the power traces each run exhibited 2.3 ± 0.9 missing segments (min = 1, max = 4) with an average segment length of 2.1 ± 0.6 samples (range = 1–4). No gap exceeded the predefined safety limit of five consecutive samples.

Therefore, all missing segments were imputed by linear interpolation. After this step, the data set contained 0 % NaNs.

For each run, the energy consumption was then aggregated by trapezoidal integration of the cleaned power trace (Eq. 5.5). Because all points were generated sequentially during optimization, no prior correlation analysis or feature scaling or standardization was performed, thus preserving the original

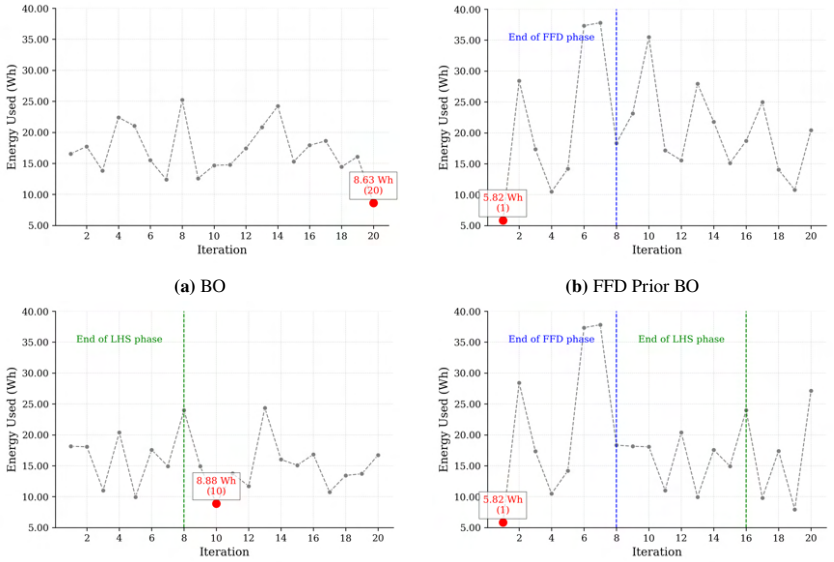
physical units was considered more informative than z scoring such a small, low-dimensional data set. In this scenario, it was not feasible to perform a correlation analysis as the data will be directly collected during the optimization process.

5.3.4 Sequential Optimization

In the BO routine, an anisotropic kernel was selected so that the surrogate model could be adapted to different length-scale sensitivities across the input dimensions. The acquisition function was chosen to be the EI, with a small trade-off parameter $\xi = 0.01$. This was adopted to favor rapid exploitation, given the limited budget of only $N = 20$ trials. To quantify sampling efficiency, four design–optimization hybrids were compared, each limited to a total of $N = 20$ function evaluations:

1. **Pure BO** — all 20 evaluations are adaptively chosen by the BO algorithm.
2. **LHS → BO** — the search is seeded with eight LHS points, after which BO allocates the remaining twelve evaluations.
3. **FFD → BO** — an eight-run quarter FFD serves as an initial screening stage, followed by twelve BO-selected points.
4. **FFD → LHS → BO** — eight FFD runs are executed first, then eight additional points are generated via LHS, and finally four further evaluations are selected by BO.

Figure 5.12 illustrates the energy consumptions of the four optimization strategies side by side.



(a) BO

(b) FFD Prior BO

(c) LHS Prior BO

(d) FFD and LHS Prior BO

Figure 5.12: Energy-consumption trajectories for the four sampling protocols, adapted from Greif et al. [3].

In Figure 5.12a, the energy consumption first fluctuates between approximately 12 and 25 Wh, then gradually decreases to its minimum of 8.63 Wh at iteration 20. In Figure 5.12b, an eight-run quarter FFD immediately finds the global minimum of 5.82 Wh in its first trial. The subsequent BO iterations oscillate above this baseline and never improve on it. In Figure 5.12c started with an eight-point LHS, followed by 12 BO iterations. The best value observed is 8.88 Wh at iteration 10. In Figure 5.12d the search was started with eight FFD runs, then eight LHS points, and finally four BO updates. The same 5.82 Wh minimum appears during the FFD block, and later phases explore higher-energy regions without further gains.

Across the entire set of $n = 80$ experiments, the response and all five process factors span their respective ranges, providing the factual boundaries for subsequent analyses. A detailed summary of the results of data profiling can be found in Table A.2 of the appendix. Specific energy consumption varies from 5.82 Wh in the most economical build to 37.81 Wh in the most demanding. Half of all runs fall between 13.81 Wh and 20.52 Wh, with a median of 16.99 Wh. The layer height ranges from 0.10 mm to 0.30 mm. The interquartile interval lies between 0.161 mm and 0.263 mm, and the median is 0.204 mm. Print speed was tested between $40\ \mathrm{mm\,s^{-1}}$ and $150\ \mathrm{mm\,s^{-1}}$. One quarter of the trials ran below $60.63\ \mathrm{mm\,s^{-1}}$, one quarter above $135.55\ \mathrm{mm\,s^{-1}}$, with a median of $101.47\ \mathrm{mm\,s^{-1}}$. The bed temperature covers $30\,^{\circ}\mathrm{C}$–$60\,^{\circ}\mathrm{C}$, the middle 50 % of observations occupying the band from $37.23\,^{\circ}\mathrm{C}$ to $57.30\,^{\circ}\mathrm{C}$ (median $47.85\,^{\circ}\mathrm{C}$). The nozzle temperature likewise spans its full range, $190\,^{\circ}\mathrm{C}$–$230\,^{\circ}\mathrm{C}$. Half of the builds are located between $193.44\,^{\circ}\mathrm{C}$ and $220.77\,^{\circ}\mathrm{C}$, centered at $202.59\,^{\circ}\mathrm{C}$. Finally, the infill mass proportion extends from 0% to the upper limit of 30%. The first quartile is 2.97%, the third quartile 26.65%, and the median build contains 16.88%.

5.3.5 Explainable AI and Parameter Influence Analysis

Following the guidelines of the methodology, the global sensitivity analysis was initially performed using both the first-order and total-order Sobol indices.

Table 5.23: Monte Carlo Sobol indices for the GP surrogate's prediction of energy consumption.

Feature	S_1	S_T
Layer height	0.72	0.74
Build plate temperature	0.22	0.23
Infill mass proportion	0.04	0.05
Print speed	0.01	0.01
Nozzle temperature	0.01	0.00

Table 5.23 shows that layer height contributes the largest share of the variance of the output, both in its first-order effect and in its total effect, followed by the temperature of the building plate. In contrast, the print speed and the nozzle temperature have negligible influences. The proximity of S_T to S_1 indicates limited interaction effects in the examined window.

A GP regression model with an ARD kernel was trained on the complete BO data set. The hyperparameters were tuned with Optuna. After training, feature attributions were computed with a kernel based SHAP explainer. Figure 5.13 shows the SHAP beeswarm plot for this model.

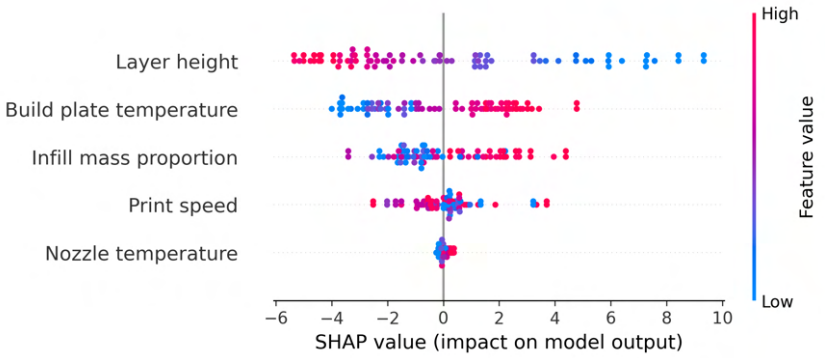


Figure 5.13: Beeswarm representation of SHAP values for the ARD GP regressor predicting energy consumption.

The plot reveals that, within the investigated process window, an increase in layer height consistently lowers the expected energy demand, whereas elevated build-plate temperature and a higher infill-mass proportion drive it upward. Within the examined process window, raising the print speed produces a modest reduction in the predicted energy demand, whereas changes in the nozzle temperature have the weakest effect of all parameters and shift the energy estimate only marginally.

In addition, a clustering analysis was performed. In this scenario, successful segregation was accomplished exclusively through GMM clustering, as hierarchical and k-means clustering failed to produce satisfactory outcomes. To identify the optimal number of clusters, the BIC and the silhouette coefficient were plotted as functions of the cluster number, as depicted in Figure A.3 of the appendix. An elbow was detected at $k = 7$, indicating that 7 clusters were optimal. In Figure 5.14 the clusters are visualized by the first three principal components, which together account for $69\,\%$ of the total variance in the six-dimensional data set.

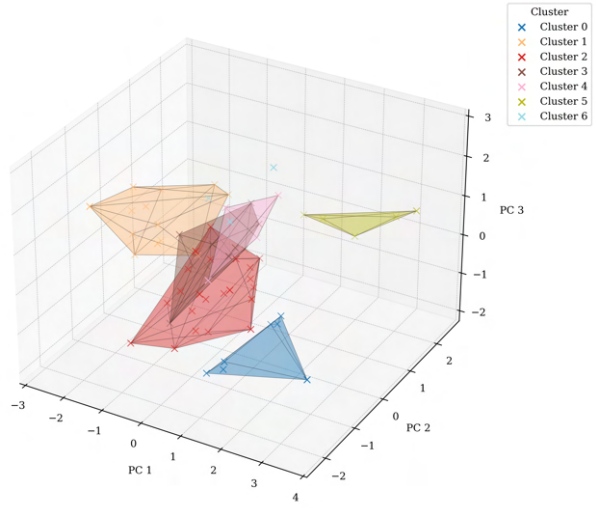


Figure 5.14: Cluster visualization.

Table 5.24 indicates that the first principal component is characterized by a strong positive loading of specific energy consumption and, with opposite sign, by a negative loading of layer height. Accordingly, PC 1 can be interpreted as an energy intensity dimension. Positive scores correspond to settings with thin layers and elevated energy demand, whereas negative scores correspond to coarser layers with lower energy demand. The second principal component

Table 5.24: PCA feature contributions to the first three components.

Feature	PC 1	PC 2	PC 3
Layer height	-0.61	0.09	-0.27
Infill mass proportion	-0.08	0.47	0.66
Print speed	-0.06	0.69	-0.22
Bed temperature	0.37	0.20	-0.64
Nozzle temperature	0.03	-0.49	0.05
Energy consumption	0.69	0.10	0.16
Explained variance	0.31	0.20	0.18

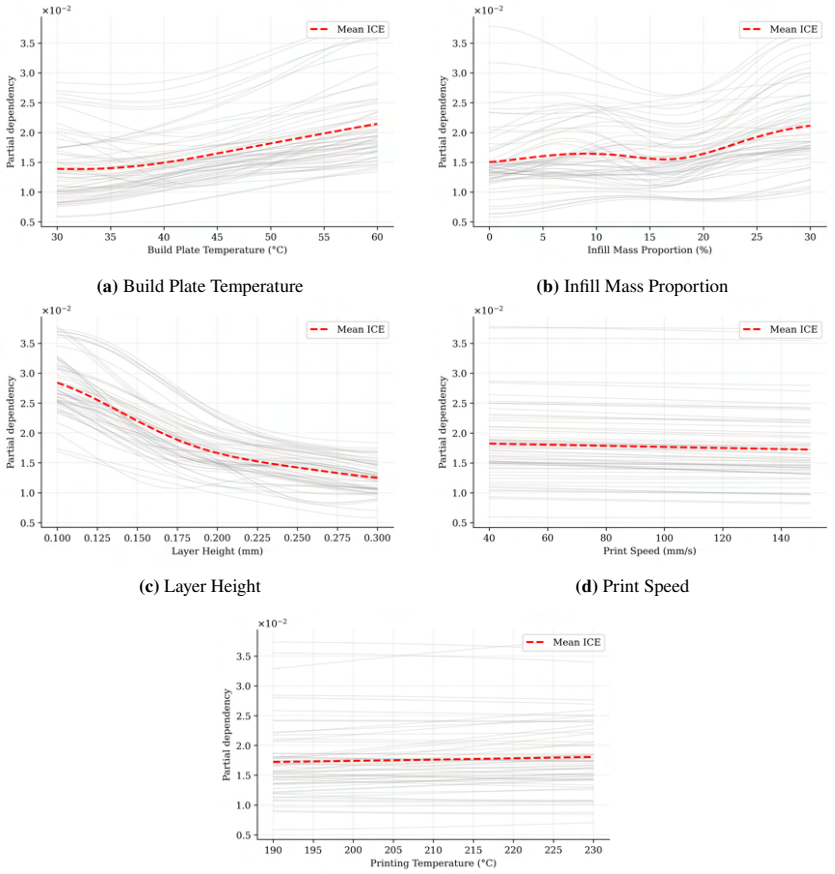
is governed by print speed and infill mass proportion, with signs opposite to nozzle temperature. This component separates regimes with high throughput and dense infill from regimes with lower speeds, cooler process conditions, and more material sparing settings. The third principal component contrasts a high infill fraction with a low build plate temperature and therefore captures a trade off between thermal input through the build plate and consolidation inside the part.

Table 5.25: Cluster sizes and mean process parameters.

Cluster	n	Layer height [mm]	Build plate temp. [°C]	Infill mass proportion [%]	Print speed [mm/s]	Nozzle temp. [°C]	Energy cons. [Wh]
0	8	0.14	57.13	1.90	68.84	223.92	27.09
1	16	0.26	33.75	19.05	100.06	207.14	10.41
2	27	0.20	50.57	8.50	80.36	200.27	15.85
3	12	0.28	54.98	24.26	126.28	216.29	16.14
4	6	0.13	33.37	13.50	132.51	216.90	19.76
5	6	0.14	58.86	28.18	147.37	191.23	30.69
6	5	0.13	34.67	28.38	42.94	203.12	24.70

The numerical cluster profiles in Table 5.25 can be interpreted within the PCA framework. Cluster 1 occupies the region of PC 1 that is associated with low energy demand and coarse layers. It combines this with moderate infill and a moderate nozzle temperature. This combination makes Cluster 1 the most resource efficient group. Clusters 0 and 5 lie at the opposite end. Both use very fine layers. Cluster 5 also applies the highest print speed and a high nozzle temperature. As a result, these two clusters exhibit the largest mean energy demand. Cluster 3 reflects the fast and dense regime highlighted by PC 2. It contains the coarsest layers in the study together with a high infill fraction and the second highest print speed. The resulting energy footprint remains moderate. Clusters 2 and 4 occupy intermediate regions of the PCA space. They differ mainly in build plate temperature and infill fraction. This difference explains their separation along PC 3.

Finally, as shown in Figure 5.15, the energy consumption of the PLA prints exhibits markedly different sensitivities. The proportion of infill mass shows an overall rising trend. Energy consumption increases moderately from 0 % to 10 %, afterwards it stays nearly constant with a small downward trend up to 18 %, from there it climbs more steeply beyond 20 %, increasing from approximately 0.015 kWh in 0 % to roughly 0.021 kWh in 30 %. The nozzle temperature has the least impact, with the mean ICE curve almost flat between 190 °C and 230 °C, rising by only 0.001 kWh over that full range. The build plate temperature shows a modest upward effect on energy use. Energy consumption is nearly flat from 30 °C to about 35 °C (0.0135 kWh), then increases steadily to approximately 0.021 kWh at 60 °C. The layer height plot shows the strongest negative dependency, with the mean energy falling steeply from about 0.028 kWh at 0.10 mm to roughly 0.014 kWh at 0.30 mm.



(a) Build Plate Temperature

(b) Infill Mass Proportion

(c) Layer Height

(d) Print Speed

(e) Printing Temperature

Figure 5.15: ICE plots showing the partial dependency of energy consumption on each process parameter.

Finally, the print speed shows a minimal downward trend, from about 0.0185 kWh at 40 mm/s to about 0.017 kWh at 150 mm/s. In general, the tight grouping of individual curves around the mean in each graph underscores the low predictive variability of the model throughout the design space.

5.3.6 Validation and Real-World Integration

To determine whether the 20-print optimization campaign pays back its initial investment, the total cost and CO_2 equivalents were compared against the savings per part achieved by switching from baseline settings ($E_{\text{base}} = 20\text{Wh}$, $m_{\text{base}} = 4\text{g}$) to the optimized settings were compared.

In the best-case DoE–BO sequence, the campaign consumes

$$E_{\text{Opt},1} = \sum_{i=1}^{20} E_i^{\text{DoE-BO}} \approx 4.15 \times 10^{-1}\,\text{kWh}. \tag{5.16}$$

With a total PLA mass of

$$m_{\text{Opt}} = 20 \times 4\,\text{g} = 8.00 \times 10^{-2}\,\text{kg}, \tag{5.17}$$

its total cost is

$$C_{\text{Opt},1} = E_{\text{Opt},1}\, c_E + m_{\text{Opt}}\, c_{mat} \approx 1.68 \times 10^{0}\,\text{EUR}, \tag{5.18}$$

and its total CO_2e is

$$I_{\text{Opt},1}^{\text{GWP100}} = E_{\text{Opt},1}\, EF_E + m_{\text{Opt}}\, EF_{mat} \approx 4.18 \times 10^{-1}\,\text{kg}\,\text{CO}_2\text{e}. \tag{5.19}$$

Switching from baseline to this best-case optimum ($E_{\text{opt},1} = 5.82\,\text{Wh}$) yields a per-part energy reduction

$$\Delta E_1 = E_{\text{base}} - E_{\text{opt},1} \approx 1.42 \times 10^{-2}\,\text{kWh}, \tag{5.20}$$

with no change in mass, so the per-part savings are

$$\Delta c_{\text{part},1} = \Delta E_1\, c_E \approx 2.69 \times 10^{-3}\,\text{EUR}\,\text{part}^{-1}, \tag{5.21}$$

$$\Delta I_{\text{part},1}^{\text{GWP100}} = \Delta E_1\, EF_E \approx 5.50 \times 10^{-3}\,\text{kg}\,\text{CO}_2\text{e}\,\text{part}^{-1}. \tag{5.22}$$

Hence the break-even points for the best-case campaign are

$$N_{\mathrm{BE}}^{\mathrm{cost},1} = \frac{C_{\mathrm{Opt},1}}{\Delta c_{\mathrm{part},1}} \approx 6.24 \times 10^{2}\,\text{parts}, \tag{5.23}$$

$$N_{\mathrm{BE}}^{\mathrm{CO_2e},1} = \frac{I_{\mathrm{Opt},1}^{\mathrm{GWP100}}}{\Delta I_{\mathrm{part},1}^{\mathrm{GWP100}}} \approx 7.60 \times 10^{1}\,\text{parts}. \tag{5.24}$$

Under the worst-case LHS–BO sequence, the campaign consumes

$$E_{\mathrm{Opt},2} = \sum_{i=1}^{20} E_i^{\mathrm{LHS\text{-}BO}} \approx 3.10 \times 10^{-1}\,\mathrm{kWh}. \tag{5.25}$$

With the same total material usage, the corresponding cost and CO_2e are

$$C_{\mathrm{Opt},2} = E_{\mathrm{Opt},2}\, c_E + m_{\mathrm{Opt}}\, c_{mat} \approx 1.66 \times 10^{0}\,\mathrm{EUR}, \tag{5.26}$$

$$I_{\mathrm{Opt},2}^{\mathrm{GWP100}} = E_{\mathrm{Opt},2}\, EF_E + m_{\mathrm{Opt}}\, EF_{mat} \approx 3.77 \times 10^{-1}\,\mathrm{kg\,CO_2e}. \tag{5.27}$$

Applying a prudence factor $\alpha = 1.2$ yields conservative thresholds:

$$N_{\mathrm{BE,safe}}^{\mathrm{cost},1} = 1.2 \times N_{\mathrm{BE}}^{\mathrm{cost},1} \approx 7.49 \times 10^{2}\,\text{parts}, \tag{5.28}$$

$$N_{\mathrm{BE,safe}}^{\mathrm{CO_2e},1} = 1.2 \times N_{\mathrm{BE}}^{\mathrm{CO_2e},1} \approx 9.12 \times 10^{1}\,\text{parts}. \tag{5.29}$$

Overall, the BE thresholds indicate a favorable and manageable payback. Compared with the previous use case, the cost BEP of $\approx 7.49 \times 10^2$ parts is about $1.90\times$ higher, whereas the CO_2e BEP of $\approx 9.12 \times 10^1$ parts is approximately one-quarter of the earlier value. This divergence can be explained with the strong reduction in energy demand achieved at the optimum, while the part mass remains constant in this setting.

Furthermore, in terms of quality requirements, all manufactured parts were visually examined. No evidence of bead detachment, warpage, surface pitting, or incomplete interlayer bonding was observed when the samples were analyzed.

In addition, no surface smoothness deficiencies that could impede water flow were identified, confirming their suitability for the intended application of the drainage system.

With the BEP results and the quality evaluation contextualized, the analysis proceeds to the robustness testing of the optimization result. Following the methodology, the PCF is decomposed into sub-process contributions as in Eq. (5.6). The baseline has a PCF of $20.60\,\mathrm{g\,CO_2}$ per part, comprising $7.76\,\mathrm{g}$ emitted by electricity consumption during machine operation and $12.90\,\mathrm{g}$ originating from the constant mass of $4.00\,\mathrm{g}$ PLA. Adoption of the best settings found reduces the PCF to $15.10\,\mathrm{g\,CO_2}$, which corresponds to an absolute reduction of $5.50\,\mathrm{g\,CO_2}$ and a relative decrease of $26.70\,\%$. Five replicate prints performed at the optimum consumed on average $6.02 \pm 0.27\,\mathrm{Wh}$. The $\pm 0.27\,\mathrm{Wh}$ variability represents only $1.40\,\%$ of the baseline demand of $20.00\,\mathrm{Wh}$. The optimization therefore reduces the mean energy consumption by $13.98\,\mathrm{Wh}$, corresponding to a reduction of $70.00\,\%$ relative to the baseline. The improvement signal thus exceeds the experimental noise by a factor of approximately 52, confirming the robustness of the observed gain.

In the context of sequential optimization, the robustness evaluation of levels A and B is not accounted for, since it would necessitate conducting additional real-world experiments, which the available budget constraints did not allow. The evaluation of XAI alignment is conducted concerning Level C.

To verify consistency across the XAI methods, the feature rankings from the Sobol first order and total order indices, the global SHAP summary, and the slope magnitudes from the ICE curves were compared. As indicated in Table 5.26, all methods agree on a common ordering. The layer height is identified as the primary driver of energy consumption. The build plate temperature ranks second. The infill mass proportion ranks third. The print speed ranks fourth. The nozzle temperature has the weakest influence within the investigated window. These results suggest that the layer height and the build plate temperature should be prioritized in process control when energy reduction is the objective, while the nozzle temperature can be treated as a low impact factor in this setting.

Table 5.26: Comparison of feature rankings across methods for the energy consumption use case. Higher rank means stronger influence.

Method	Rank 1	Rank 2	Rank 3	Rank 4	Rank 5
Sobol first order S_1	Layer height	Build plate temperature	Infill mass proportion	Print speed	Nozzle temperature
Sobol total order S_T	Layer height	Build plate temperature	Infill mass proportion	Print speed	Nozzle temperature
SHAP global summary	Layer height	Build plate temperature	Infill mass proportion	Print speed	Nozzle temperature
ICE slope magnitude	Layer height	Build plate temperature	Infill mass proportion	Print speed	Nozzle temperature

A complementary perspective from the PCA loadings in Table 5.24 supports this interpretation. The first principal component aligns positively with energy consumption and negatively with the layer height, which is consistent with the strong influence of the layer height observed in the Sobol, SHAP and ICE analyzes.

In addition, to assess whether the Gaussian-process regressor with an ARD kernel constituted the suitable surrogate, the response surface was re-estimated using two alternatives: a Gaussian-process regressor with an isotropic RBF kernel and an RF, and their predictive accuracy was compared with that of the ARD variant.

Table 5.27: Test-set performance of the three regressors evaluated over ten random splits of the BO Experiments data.

Model	R^2	MAE	RMSE
GPR-ARD	0.76 ± 0.12	1.63 ± 0.68	2.41 ± 1.19
GPR-isotropic	0.17 ± 0.67	2.88 ± 0.74	3.97 ± 0.91
RF	0.54 ± 0.45	1.98 ± 0.61	2.80 ± 1.12

The comparative evaluation in Table 5.27 indicates that the GP model with an ARD kernel attains the highest average R^2 and the lowest MAE and RMSE across the ten random splits, whereas the isotropic kernel and the RF surrogate perform noticeably worse. These results suggest that allowing feature specific length scales increases the fidelity with which the response surface is learned. However, the findings should be interpreted with care. The analysis isolates the regression step and does not capture the interaction between the surrogate model, the sampling design, and the acquisition function that would operate in an iterative BO loop. Because this interaction governs how informative new experiments are selected, the relative ranking of surrogates observed here does not necessarily translate to sequential optimization performance. The present comparison therefore serves to assess how accurately each model can fit the underlying response surface given the available data.

Furthermore, spearman rank correlation analysis was performed to uncover monotonic dependencies among process variables and response. Figure 5.16 displays the resulting correlation matrix.

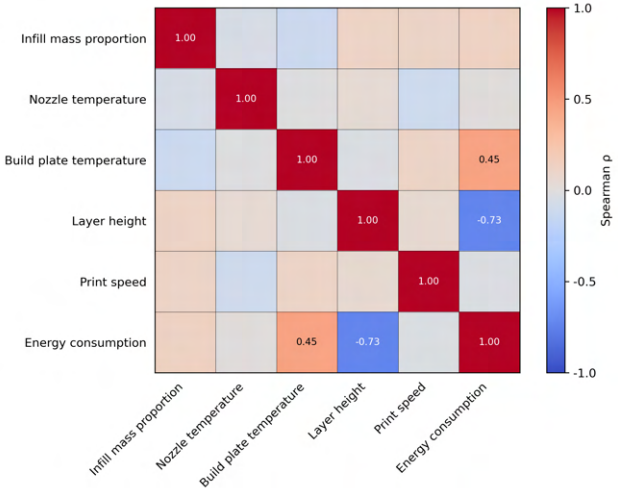


Figure 5.16: Spearman rank-correlation matrix of all process parameters and the response variable.

The matrix confirms that the input parameters are independent of each other, indicating that the experimental design did not introduce appreciable col-linearity. The absence of pronounced interparameter correlations fulfills the conditional independence assumption required for post hoc explainability tools such as ICE curves, ensuring that the effect of each variable can be interpreted in isolation.

Overall, the satisfaction of quality and the energy savings achieved, when comparing the baseline configurations to the optimized parameter settings, do not necessitate a methodological reassessment or refinement for the goal of the use case.

5.4 Discussion

Manufacturing occupies a central position in global sustainable development and is critical to achieving the SDGs of the United Nations. In particular, SDG 9, SDG 12, and SDG 13 require a substantive transformation of industrial processes to achieve greater resource and energy efficiency. Digital technologies and data-driven methods, such as the methodology proposed in this work, enable these goals to be operationalized at the plant level by integrating measurement, modeling, and optimization into routine decision making.

One of the most fundamental findings is that a systematic, data-driven adjustment of process parameters leads to substantial environmental and economic benefits, regardless of the chosen optimization strategy. The demonstrated savings in CO_2-equivalents, energy, and material consumption represent a direct and measurable contribution to the aforementioned sustainability goals.

To systematically analyze this central finding and illuminate its full implications, the following discussion is structured as follows. First, the overarching findings from the industrial use cases are synthesized to identify the key levers for sustainable manufacturing. Building on this, the section discusses the methodological trade-offs that emerge from the empirical validation. From these insights, the concrete implications for industrial practice are then derived,

before the chapter concludes with a critical examination of the methodology's limitation.

The two industrial use cases on the reduction of CO_2-equivalents and energy consumption in PLA FDM printing reveal convergent insights into the determination of process dynamics. Regardless of the chosen optimization objective, be it the PCF or pure energy consumption, two parameter groups proved to be dominant:

1. Material Deposition: The amount of polymer deposited, mainly controlled by the infill density and secondarily by the number of perimeters, is the largest driver of the environmental footprint.

2. Process Time and Thermal Energy: The layer height has a negative influence on energy consumption. Larger layer heights reduce the number of toolpaths and the total printing time, leading to a lower energy demand. Similarly, the heated bed temperature was shown to be a relevant energy driver, while the nozzle temperature had a negligible influence within the investigated process window.

These results point to a generalizable principle for sustainability oriented optimization in FDM. Reducing material consumption and shortening active machine time are the most effective levers to improve resource efficiency. At the same time, the validation exposes a performance trade off. In the energy reduction use case, quality requirements were intentionally low, which permitted optimization without mechanical constraints. In the PCF-oriented use case, a clear tension emerged between environmental objectives and mechanical performance. Tensile strength analysis showed that infill density, the strongest driver of CO_2 reduction, is also the primary determinant of mechanical stability. Within the investigated parameter window, lowering the infill density reduces the tensile strength in a nearly monotonic manner. Consequently, the CO_2 savings achieved through a lower infill must be balanced against the application-specific strength targets.

This type of conflict is typical in numerous manufacturing systems and highlights the necessity to evaluate quality limitations alongside environmental metrics, and conversely, environmental considerations alongside quality measures. Traditionally, process optimization in industry has focused primarily on quality or performance characteristics, such as maximizing strength or minimizing cycle time. However, this represents merely the inverse trade-off. The one-sided pursuit of quality goals often implicitly leads to higher resource consumption and thus to a larger environmental footprint. This work shows that a holistic view is essential to find a conscious and quantified compromise instead of neglecting one of the two aspects.

The empirical validation then revealed a second, equally important trade-off that lies at the core of the developed methodology. The choice between a static approach and an adaptive search is not only technical but strategic. The approach of creating a global surrogate model using a space-filling experimental design aims to achieve a comprehensive understanding of the process landscape. The result is a highly predictive model that maps the entire design space. This offers maximum flexibility as the data can be used to optimize the system for different target variables, for example, first CO_2, then tensile strength, analyze interactions and generate deep process knowledge. The price for this is a higher initial experimental effort. This drawback is exacerbated by two factors. First, the nature of factorial designs, where the number of experiments grows exponentially with the number of factors and levels. Adding just one more parameter or an additional level can overwhelm the budget. Second, as the number of parameters increases, this approach quickly runs into the limits imposed by the Curse of Dimensionality [97]. With each additional dimension, the volume of the search space grows exponentially, causing the data points to become sparse even with large sample sizes. Reliably covering the entire space becomes practically impossible, and interpolation between distant measurement points becomes unreliable. In contrast, the adaptive search is designed to find an optimum for a predefined target variable with a minimum sample size. The validation confirmed the superior sample efficiency of this approach. BO consistently found better optima than the static model with the

same or a smaller experimental budget. However, this advantage comes at the cost of a loss of global knowledge. The generated data are strongly biased towards the optimum, which makes the resulting surrogate model unsuitable for a general process analysis or optimization for alternative target variables. An optimization of tensile strength with the data from the energy reduction use case would therefore have been hardly possible. The methodology developed thus offers a clear decision-making framework. If the focus is on the fast and resource efficient optimization of a stable target criterion, direct optimization is the preferred path. If, however, the goal is to build a deep understanding of the process, explore trade-offs, or flexibly adapt to changing requirements, global surrogate modeling is indispensable.

Furthermore, the validation results have far-reaching implications for the industrial use of AI-supported optimization methods. Break-even analyzes show that the initial effort for a data-driven optimization campaign is amortized with small to medium production volumes. This makes the methodology attractive not only for large-scale production but also for SMEs and prototyping. The ability to achieve cost and resource savings without compromising part quality, as long as the boundary conditions are observed, represents a clear competitive advantage. The methodology should not be understood as an isolated tool but as a strategic component in the product development process. Utilizing different strategies for identical processes at various stages is also feasible. In early phases, a global surrogate model can provide valuable knowledge about the limits and sensitivities of the process. In late phases, direct optimization can be used to quickly and efficiently find the optimal process parameters for a specific application. The application of XAI methods is crucial to acceptance and success in practice. They translate the complex, non-linear relationships of a machine learning model into understandable and actionable insights for the process engineer. This builds trust in the black box, validates the model results against existing domain knowledge, and enables informed human-centered decision making.

As with any framework, the one presented here has its limitations, which also provide opportunities for future development. The methodology relies on landscape metrics to guide the choice of strategy. In situations with extremely scarce data or entirely new processes, it is difficult to calculate these metrics. Although the framework provides a default recommendation for such cases, the initial decision is based more on an educated guess than on hard data. A future enhancement could include mechanisms to dynamically adapt the strategy as more is learned about the process landscape during the initial experiments. Furthermore, the effectiveness of the methodology relies on a partnership between the data-driven framework and human expertise. The selection of suitable initial parameters and constraints, and particularly the interpretation of the results to generate actionable insights, often require expert knowledge. Without this expert-in-the-loop, there is a risk of making suboptimal decisions that could reduce the efficiency of the optimization or lead to impractical solutions. Another aspect is the handling of competing goals. The methodology is very effective at revealing trade-offs, as demonstrated by the example of CO_2 reduction versus mechanical strength. However, currently it does not offer an integrated mechanism to resolve these conflicts directly. The next logical step would therefore be to extend the framework to include multi-objective optimization methods. This would allow users to generate a range of optimal compromise solutions, the so-called Pareto front, and to make a conscious, data-driven decision that considers both quality and sustainability aspects. Finally, while this work presents best practices for the various steps of the methodology drawn from research and application cases, it cannot cover every individual contingency of a specific industrial setting. Every real-world scenario comes with its own unique set of constraints, be it specific machine behaviors, material inconsistencies, or unique quality standards. Therefore, the framework should be seen as a robust and adaptable guide, not a rigid prescription, which practitioners must tailor to their specific context.

5.5 Summary of the Validation

The validation of the proposed methodology is systematically organized into two distinct stages. The initial stage focuses on an empirical validation of the methodology's core principle, specifically testing the hypothesis that an adaptive search strategy, such as BO, exhibits greater sample efficiency in locating an unknown optimum compared to a static, one-shot modeling approach combining a DoE with model training. This was tested using five publicly available material science datasets. The results confirmed that the adaptive BO strategy consistently established a performance advantage, often within the first 10-15 evaluations, and was more effective in finding the true optimal value. This analysis revealed a trade-off. While the static approach produced models with superior global predictive accuracy across the entire process space, the adaptive search was demonstrably more efficient for the specific task of optimization. The advantage of the adaptive strategy was found to be more pronounced in landscapes with high local ruggedness, as indicated by a strong negative correlation with the PIC. Building upon these findings, the second stage demonstrates the end-to-end applicability of the entire methodology through two industrial use cases in FDM 3D printing, designed to test the framework under both data-rich and data-limited conditions. The first use case, a data-rich scenario, aimed to minimize the PCF of a printed PLA part. Following the methodology, a surrogate modeling approach was selected, and machine learning models were trained on data from a full factorial experiment. XAI analysis identified infill density and layer height as the dominant drivers of CO_2 equivalents. A critical conflict between environmental goals and mechanical quality was uncovered as the infill density, the strongest lever for reducing the PCF, was also the most decisive factor for the part's tensile strength. The second use case, representing a data-limited scenario, focused on minimizing energy consumption with a constrained budget of only 20 experiments. Here, a direct black-box optimization strategy using BO was employed. The results showed that initializing the search with a structured FFD was the most efficient strategy, locating the minimum global energy within the initial screening phase. Across

both industrial applications, a consistent principle emerged: reducing material deposition and minimizing active machine time are the most effective levers for enhancing resource efficiency. The validation confirms that the methodology provides a robust framework for navigating the strategic choice between achieving a comprehensive global process understanding and performing rapid and resource-efficient optimization, while also quantifying the inherent trade-offs between sustainability objectives and traditional quality metrics.

6 Conclusion and Future Work

The conclusion of this work consolidates the essential findings and contributions. It begins by revisiting the central research questions and presenting a consolidated summary of the results from each chapter, highlighting their respective roles in achieving the overall research objectives. Building upon these findings, the chapter then provides a forward-looking perspective, outlining several promising avenues for future research that extend and build upon the work presented here.

6.1 Conclusion

Chapter 1 delineates the problem landscape at the intersection of ecological transformation and data-driven manufacturing. In response to global sustainability goals and regulatory pressure, this thesis proposes a data-efficient, AI-driven framework for process optimization under the constraints of data scarcity. The scope is focused on the ramp-up and serial production phases, with AM as a primary application context. This chapter culminates in the central research question: How can data-efficient, explainable optimization be operationalized in industrial environments to advance sustainable manufacturing?

Chapter 2 establishes the theoretical foundations required for this work. It provides a technical overview of the relevant manufacturing processes, including thermoplastic and paste extrusion. Then it details established environmental assessment tools, namely LCA, with an emphasis on system boundaries and impact categories. The chapter proceeds to outline the necessary components

to achieve a complete data-driven optimization pipeline, outlining the choice between sequential optimization and offline optimization with static surrogates. The chapter concludes with the establishment of XAI as a component to bridge the gap between model performance and user trust.

Chapter 3 presents a systematic review of the state of the art. A literature review, conducted according to the PRISMA methodology, reveals that while powerful data-efficient optimization algorithms and a growing repertoire of XAI techniques exist, they are rarely integrated into a single, coherent framework. The analysis identifies a critical research gap: the absence of a holistic methodology that unifies data-efficient, explainable, and practically deployable optimization for sustainable manufacturing.

Chapter 4 details the novel methodology developed to address this identified gap. The proposed framework is modular and sequential, guiding from the definition of the goal to industrial validation. It specifies a systematic process for defining sustainability metrics, a structured framework for choosing between offline and sequential optimization strategies, and a six-stage pipeline for data preprocessing. The methodology integrates multi-layered XAI protocols for global, local, and subgroup analysis, while formal protocols for break-even analysis, quality assurance, and iterative refinement ensure the practical viability of the results.

Chapter 5 provides an empirical validation of the proposed methodology. A quantitative study on open-source material science datasets first demonstrates the superior sample efficiency of the sequential optimization strategy compared to a static surrogate approach. Subsequently, two industrial case studies apply the complete framework to FDM of PLA. The first case minimizes CO_2 emissions and uses XAI to reveal a critical trade-off between the ecological objective and mechanical properties. The second case demonstrates a reduction in energy consumption within a severely limited experimental budget. In both studies, the efficacy and adaptability of the framework in different data availability regimes are confirmed.

Viewed through the wider lens of global sustainability agendas, the dissertation advances three overarching frameworks at once. First, it operationalizes key aspirations of the United Nations 2030 Agenda. By embedding carbon footprint and energy intensity metrics directly in the optimization objective, the work creates an actionable pathway toward SDG 12 on responsible consumption and production and SDG 13 on climate action, while its emphasis on data-efficient AI solutions that can be retrofitted to existing equipment speaks to the call for inclusive and resilient industrial innovation of SDG 9. Quantitative case studies translate these high-level goals into factory floor practice, showing that high reductions in energy demand and material use are possible without incurring prohibitive experimentation costs. Second, the thesis aligns with the European Commission's vision of a CE by transitioning the focus of process optimization from solely maximizing yield to also reducing embodied emissions and minimizing wasted feedstock. The ability of the methodology to illuminate trade-offs, such as the tension between tensile strength and CO_2 production, gives engineers a transparent decision calculus to balance productivity with resource loops, which supports design-for-recyclability and life-cycle thinking at the level of process parameters. Third, and most directly, the framework exemplifies the principles of Industry 5.0, which extend the automation-centric narrative of Industry 4.0 to emphasize human-centricity, sustainability, and resilience. The sample efficient BO reduces the physical and computational burden of experimentation, the multilayer XAI component restores human interpretability to machine recommendations, and the modular architecture ensures that the knowledge gained in one setting can be transferred to the next, building organizational resilience. In short, the dissertation translates the abstract pillars of Industry 5.0 into a concrete, data-driven toolkit that empowers engineers to co-create climate-aligned production systems rather than merely automate existing ones.

6.2 Future Work

The research presented in this thesis establishes a methodological framework and also opens several promising avenues for future investigation. One primary direction is to broaden the framework's application in order to test its generalizability. A second path involves deepening the methodology itself to further enhance its sample efficiency. Finally, a third area of focus is to advance the practical application of XAI through dedicated user-centric studies.

The first step is to validate the generalizability of the proposed framework beyond the initial use case of 3D printing with PLA. The future objective is to demonstrate the domain-agnostic capabilities of the methodology by transferring it to a diverse set of manufacturing technologies. Initial studies have commenced in this field, demonstrating potential in reducing energy consumption and enhancing buildability within injection molding and clay 3D printing processes, respectively. Key research questions will include: How does the optimal surrogate model class change with the higher dimensionality and different noise characteristics of injection molding? Can the dimension-aware BO strategy effectively navigate the complex, non-linear trade-offs in multi-material optimization for metal AM? The assessment of success will encompass not just the final performance attained, but also the efficiency of transfer, defined as the measurable decrease in the experimental budget necessary, compared to initiating optimization from scratch in every new domain.

A key challenge in industrial optimization remains the scarcity of data for new products or processes. Future research will focus on improving the sample efficiency of the methodology by moving from single-task optimization to multitask learning. One promising avenue is the use of transfer learning to warm-start optimization algorithms. Methodologically, this could be implemented by using a surrogate model trained on a data-rich source task, from a well-established product line as a prior for a BO on a new, data-scarce target task. A more ambitious long-term direction is the development of domain-specific foundation

models for manufacturing. This research would involve aggregating large, heterogeneous datasets from multiple production lines, designing a self-supervised pre-training objective that forces the model to learn the underlying physics of the process, and developing efficient fine-tuning strategies for new tasks.

This work demonstrated that detailed post hoc XAI analyses were highly valued by industrial stakeholders. However, this raises the question: What level and type of interpretability is truly necessary and effective in practice for different user roles? To address this, a formal human-computer interaction (HCI) study is needed. This study would involve presenting process engineers and quality managers with standardized optimization diagnostic tasks under different explanation conditions. Key dependent variables would include task accuracy, decision-making time, and subjective ratings of trust and usability. Think-aloud protocols would be used to capture qualitative insights into the users' reasoning processes, providing an evidence-based guide for designing effective XAI interfaces for manufacturing.

Finally, as the industry progresses towards the long-term vision of fully autonomous production, a fundamental question arises: if a manufacturing plant operates without human intervention, does the need for human-centric explainability diminish? This thesis posits that, counterintuitively, the need for robust XAI may become even more critical, opening a vital avenue for future research that goes beyond explaining static models to a user. This future work would explore the role of XAI in the certification and accountability of autonomous systems, where explainability may become a prerequisite for regulatory approval as humans remain legally and ethically responsible. Furthermore, research into XAI is needed as a tool to ensure resilience and enable debugging. When novel failures inevitably occur, explainability methods will be essential for engineers to conduct root cause analysis in these complex, emergent systems. A third compelling research direction is the use of XAI for automated scientific discovery, creating methods to extract novel physical insights from an autonomous agent's behavior, thereby turning the smart factory into an engine for continuous innovation. Ultimately, this research would investigate the evolution of XAI from a tool for human-computer interaction to a fundamental component for

the governance, safety, and innovative capacity of next-generation autonomous systems.

A further potential research direction is the integration of the methodology into an autonomous, agent-based system. The long-term vision would be a self-optimizing production line where a software agent could independently execute the full cycle of data analysis, optimization, and validation. In such a paradigm, the systematic protocols developed in this thesis could serve as the agent's cognitive modules: the data preprocessing pipeline could function as its perception system, the framework for strategy selection as its core decision-making engine, and the optimization and XAI protocols as its mechanisms for action and self-reflection. Realizing this vision would open up new research questions, such as how to formalize the methodology's qualitative decision points into machine-executable rules and how to design effective human-on-the-loop interfaces for supervising such an autonomous agent.

A Appendix

Table A.1: Use Case 2 (PLA dataset): Descriptive statistics for process variables.

Variable	Quantiles						
	0%	5%	25%	50%	75%	95%	100%
Power (W)	59.7	67.5	87.6	101.1	111.0	124.1	139.5
Imputed Power (W)	59.7	68.1	89.7	98.9	109.8	123.2	139.5
Layer Height (mm)	0.16	0.16	0.16	0.22	0.28	0.28	0.28
Infill Density (%)	15.0	15.0	15.0	57.5	100.0	100.0	100.0
Number of Perimeters	2	2	2	4	6	6	6
Nozzle Temperature (°C)	190	190	190	205	220	220	220
Energy Consumption (Wh)	19.54	21.71	25.02	27.35	29.24	32.68	34.17
Material Consumption (g)	6.18	6.39	7.55	8.64	10.01	10.32	10.43
CO_2 Emissions (g)	27.96	28.77	33.40	38.81	43.25	45.39	46.40

Table A.2: Use Case 3 (BO dataset): Descriptive statistics for process variables.

Variable	Quantiles						
	0%	5%	25%	50%	75%	95%	100%
Infill Mass Proportion (%)	0	0	2.97	16.88	26.65	30	30
Nozzle Temperature (°C)	190	190	193.44	202.59	220.77	230	230
Bed Temperature (°C)	30	30	37.23	47.85	57.30	60	60
Layer Height (mm)	0.100	0.100	0.161	0.204	0.263	0.300	0.300
Print Speed (mm/s)	40	40	60.62	101.47	135.55	150	150
Energy Consumption (Wh)	5.82	8.86	13.81	16.99	20.52	35.58	37.81

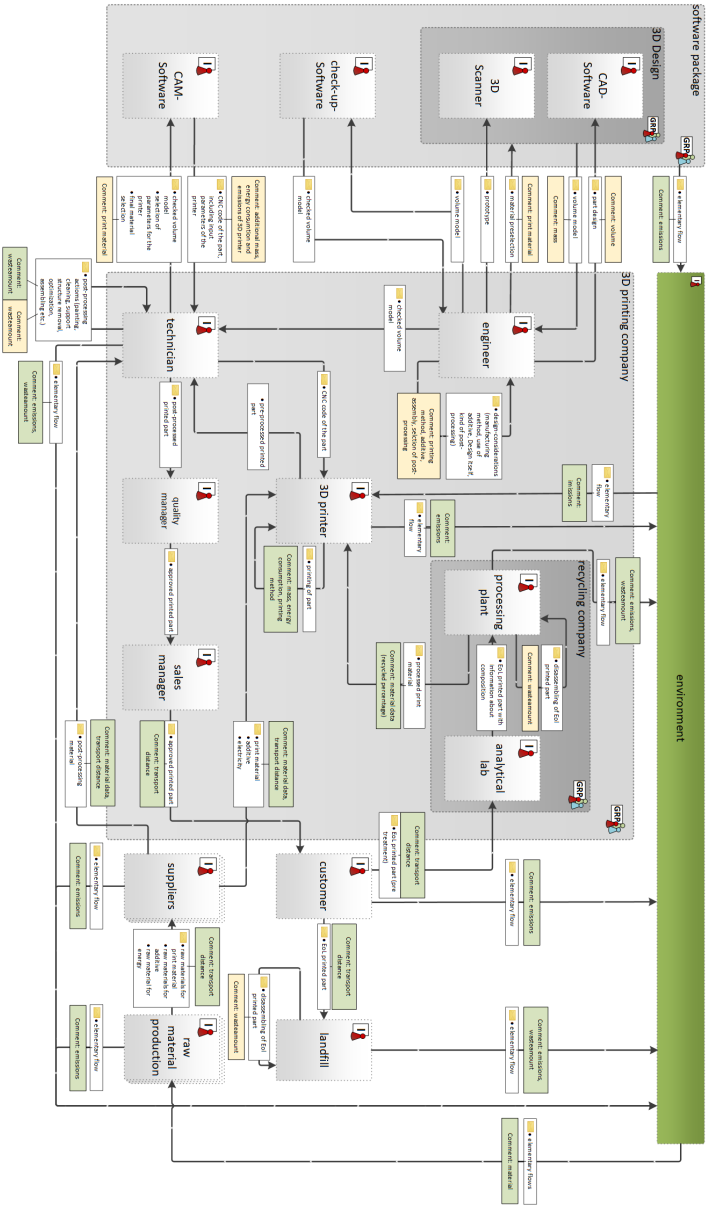
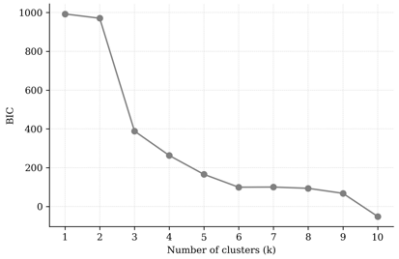
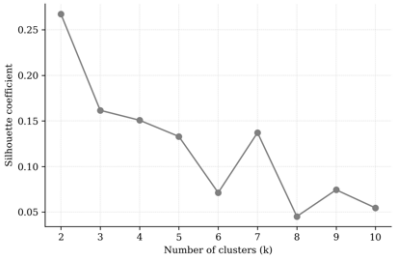


Figure A.1: Subject-oriented process analysis of the 3D-printing workflow, from Hauck and Greif [2].

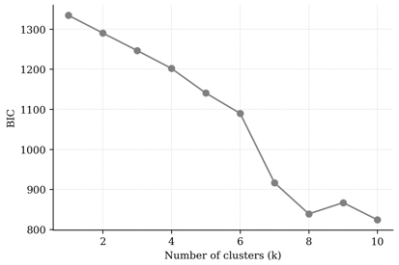


(a) BIC values for different numbers of clusters

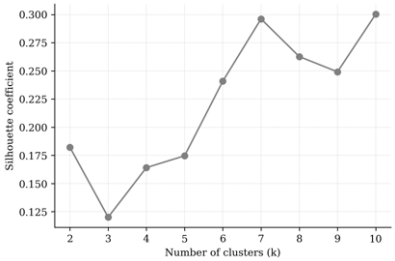


(b) Silhouette coefficient for different numbers of clusters

Figure A.2: Use Case 2 (PLA dataset): BIC and Silhouette analysis to identify the optimal number of clusters.



(a) BIC scores for different numbers of clusters.



(b) Silhouette scores for different numbers of clusters.

Figure A.3: Use Case 3 (BO dataset): BIC and Silhouette analysis to identify the optimal number of clusters.

List of Figures

List of Tables

List of Publications

Journal articles

[1] Lucas Greif, Andreas Kimmig, Sleiman El Bobbou, Paul Jurisch, and Jivka Ovtcharova. Strategic view on the current role of ai in advancing environmental sustainability: a swot analysis. *Discover Artificial Intelligence*, 4 (1):45, 2024. doi: https://doi.org/10.1007/s44163-024-00146-z.

[2] Lucas Greif, Fabian Röckel, Andreas Kimmig, and Jivka Ovtcharova. A systematic review of current ai techniques used in the context of the sdgs. *International Journal of Environmental Research*, 19(1):1, 2025. doi: https://doi.org/10.1007/s41742-024-00668-5.

[3] Lucas Greif, Niklas Hübschle, Andreas Kimmig, Simon Kreuzwieser, Anatole Martenne, and Jivka Ovtcharova. Structured sampling strategies in bayesian optimization: evaluation in mathematical and real-world scenarios. *Journal of Intelligent Manufacturing*, pages 1–31, 2025. doi: https://doi.org/10.1007/s10845-025-02597-2.

[4] Svenja Hauck, Lucas Greif, Nils Benner, and Jivka Ovtcharova. Advancing sustainable additive manufacturing: Analyzing parameter influences and machine learning approaches for co2 prediction. *Sustainability*, 17(9): 3804, 2025. doi: https://doi.org/10.3390/su17093804.

[5] Lucas Greif, Svenja Hauck, Matthes Elstermann, and Jivka Ovtcharova. A human-centric analysis of life cycle assessments: development and validation of a subject-oriented model. *International Journal of Production*

Research, pages 1–26, 2025. doi: https://doi.org/10.1080/00207543.2025.2494081.

[6] Lucas Greif, Svenja Hauck, Andreas Kimmig, and Jivka Ovtcharova. A knowledge graph framework to support life cycle assessment for sustainable decision-making. *Applied Sciences*, 15(1):175, 2024. doi: https://doi.org/10.3390/app15010175.

[7] Jakob Bönsch, Lucas Greif, Svenja Hauck, Simon Kreuzwieser, Anjela Mayer, Felix Longge Michels, and Jivka Ovtcharova. Virtual engineering: Hands-on integration of product lifecycle management, computer-aided design, extended reality, and artificial intelligence in engineering education. *Chemie Ingenieur Technik*, 96(11):1460–1474, 2024. doi: https://doi.org/10.1002/cite.202300169.

[8] Lucas Greif, Jakob Bönsch, Karan Lashkare, Francesco Arcidiacono, Tobias Zuehlke, and Jivka Ovtcharova. Robuste bauteilidentifikation mittels digitaler fingerabdrücke. *Zeitschrift für wirtschaftlichen Fabrikbetrieb*, 120 (s1):232–235, 2025. doi: https://doi.org/10.1515/zwf-2024-0128.

[9] Svenja Hauck and Lucas Greif. Sustainable product development and production with ai and knowledge graphs. *Zeitschrift für wirtschaftlichen Fabrikbetrieb*, 120(s1):107–111, 2025. doi: https://doi.org/10.1515/zwf-2024-0119.

[10] Anjela Mayer, Lucas Greif, Tim Markus Häußermann, Simon Otto, Kevin Kastner, Sleiman El Bobbou, Jean-Rémy Chardonnet, Julian Reichwald, Jürgen Fleischer, and Jivka Ovtcharova. Digital twins, extended reality, and artificial intelligence in manufacturing reconfiguration: A systematic literature review. *Sustainability*, 17(5):2318, 2025. doi: https://doi.org/10.3390/su17052318.

Conference contributions

[1] Arelia McKern, Anjela Mayer, Lucas Greif, Jean-Rémy Chardonnet, and Jivka Ovtcharova. Ai-based interactive digital assistants for virtual reality in educational contexts. In *2024 IEEE 3rd German Education Conference (GECon)*, pages 1–5. IEEE, 2024. doi: https://doi.org/10.1109/GECon62014.2024.10734030.

[2] Svenja Hauck and Lucas Greif. Facilitating the preparation of life cycle assessment through subject-oriented process modeling: A methodological framework. In *International Conference on Subject-Oriented Business Process Management*, pages 173–188. Springer, 2024. doi: https://doi.org/10.1007/978-3-031-72041-3_12.

[3] Lucas Greif and Matthes Elstermann. The role of large language models in evolving digital landscapes: A subject-oriented analysis. In *International KES Conference on Human Centred Intelligent Systems*, pages 111–124. Springer, 2024. doi: https://doi.org/10.1007/978-981-97-8598-8_11.

[4] Svenja Hauck, Lucas Greif, and Kim Skade. A methodological framework for survey design: Investigating perceptions of product sustainability – a case study. In *Conference on Sustainable Energy Education – SEED*, pages 699–707, 2024. doi: https://doi.org/10.4995/SEED2024.2024.19007.

Bibliography

[1] World Economic Forum. Reducing the carbon footprint of the manufacturing industry through data sharing. https://www.weforum.org/impact/carbon-footprint-manufacturing-industry/, 2022. Published 23 Mar 2022 (updated 3 Jun 2025), Last access: 07.08.2025.

[2] Roland Berger GmbH. Climate protection in the manufacturing sector: Challenges and solutions. https://www.rolandberger.com/en/Insights/Publications/Climate-protection-in-the-manufacturing-sector-Challenges-and-solutions.html, 2019. Last access: 07.08.2025.

[3] United States Environmental Protection Agency. Inventory of u.s. greenhouse gas emissions and sinks: 1990–2020. Technical report, U.S. EPA, 2022. URL https://www.epa.gov/ghgemissions/inventory-us-greenhouse-gas-emissions-and-sinks-1990-2020. Last access: 07.08.2025.

[4] United Nations Department of Economic and Social Affairs. *The Sustainable Development Goals Report 2021*. The Sustainable Development Goals Report. United Nations, aug 2021. ISBN 9789210056083. doi: https://doi.org/10.18356/9789210056083.

[5] Stockholm Resilience Centre. The sdgs wedding cake, 2016. URL https://www.stockholmresilience.org/research/research-news/2016-06-14-the-sdgs-wedding-cake.html. Last access: 07.08.2025.

[6] European Commission. Net-zero industry act: Proposal for a regulation of the european parliament and of the council establishing a framework of measures for strengthening europe's net-zero technology products manufacturing ecosystem. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52023PC0161, 2023. Last access: 07.08.2025.

[7] Baozhuang Niu, Nan Zhang, Fengfeng Xie, and Hailun Zhang. Optimisation of carbon emission reduction in a competitive market with varying saturation and eco-conscious consumers. *International Journal of Production Research*, 62(18):6518–6541, 2023. doi: https://doi.org/10.1080/00207543.2023.2245065.

[8] Maija Breque, Lars De Nul, Athanasios Petridis, and European Commission, Directorate-General for Research and Innovation. Industry 5.0 – towards a sustainable, human-centric and resilient european industry. Technical report, Publications Office of the European Union, Luxembourg, 2021. URL https://op.europa.eu/en/publication-detail/-/publication/468a892a-5097-11eb-b59f-01aa75ed71a1. Last access: 07.08.2025.

[9] Foivos Psarommatis, João Sousa, João Pedro Mendonça, and Dimitris Kiritsis. Zero-defect manufacturing: the approach for higher manufacturing sustainability in the era of industry 4.0: a position paper. *International Journal of Production Research*, 60(1):73–91, 2021. doi: https://doi.org/10.1080/00207543.2021.1987551.

[10] S. C. Lenny Koh, Fu Jia, Yu Gong, Xiaoxue Zheng, and Alexandre Dolgui. Achieving carbon neutrality via supply chain management: position paper and editorial for ijpr special issue. *International Journal of Production Research*, 61(18):6081–6092, 2023. doi: https://doi.org/10.1080/00207543.2023.2232652.

[11] Ashish Dwivedi, Md. Abdul Moktadir, Charbel José Chiappetta Jabbour, and Daniel Estima de Carvalho. Integrating the circular economy and industry 4.0 for sustainable development: Implications for responsible footwear production in a big data-driven world. *Technological Forecasting and Social Change*, 175:121335, 2022. doi: https://doi.org/10.1016/j.techfore.2021.121335.

[12] Majid Moradi Aliabadi and Yinlun Huang. Chapter 13 - a decision support framework for sustainable and smart manufacturing. In Masoud Soroush, Michael Baldea, and Thomas F. Edgar, editors, *Smart Manufacturing*, pages 353–376. Elsevier, 2020. ISBN 978-0-12-820027-8. doi: https://doi.org/10.1016/B978-0-12-820027-8.00013-7.

[13] Karl T. Ulrich and Steven D. Eppinger. *Product Design and Development*. McGraw-Hill Education, New York, NY, USA, 4 edition, 2016. ISBN 978-0073101422.

[14] A. Beckers, Tim Hommen, Marco Becker, Mia J. K. Kornely, Eike Reuter, Gonsalves Grünert, Lucia Ortjohann, Jannis Jacob, Philipp Niemietz, S. Barth, and T. Bergs. Digitalized manufacturing process sequences – foundations and analysis of the economic and ecological potential. *CIRP Journal of Manufacturing Science and Technology*, 2022. doi: 10.1016/j.cirpj.2022.09.001.

[15] L. Yi, Moritz Glatt, P. Sridhar, K. Payrebrune, B. Linke, B. Ravani, and J. Aurich. An eco-design for additive manufacturing framework based on energy performance assessment. *Additive manufacturing*, 33:101120, 2020. doi: 10.1016/j.addma.2020.101120.

[16] Reda Nujoom, Qian Wang, and A. Mohammed. Optimisation of a sustainable manufacturing system design using the multi-objective approach. *The International Journal of Advanced Manufacturing Technology*, 96:2539 – 2558, 2018. doi: 10.1007/s00170-018-1649-y.

[17] John Johansen and Jens O. Riis. The interactive firm—towards a new paradigm. a framework for the strategic positioning of the industrial

company of the future. *International Journal of Operations & Production Management*, 25(2):202–216, 2005. doi: https://doi.org/10.1108/01443570510577038.

[18] Ian Gibson, David W. Rosen, Brent Stucker, and Mahyar Khorasani. *Additive Manufacturing Technologies: 3D Printing, Rapid Prototyping, and Direct Digital Manufacturing*. Springer, Cham, Switzerland, 3rd edition, 2021. doi: https://doi.org/10.1007/978-1-4939-2113-3.

[19] Tuan D. Ngo, Ali Kashani, Giuseppe Imbalzano, Kate T. Q. Nguyen, and David Hui. Additive manufacturing (3d printing): A review of materials, methods, applications and challenges. *Composites Part B: Engineering*, 143:172–196, 2018. doi: https://doi.org/10.1016/j.compositesb.2018.02.012.

[20] Sangjin Jung, Levent B. Kara, Zhenguo Nie, Timothy W. Simpson, and Kate S. Whitefoot. Is additive manufacturing an environmentally and economically preferred alternative for mass production? *Environmental Science & Technology*, 57(16):6373–6386, 2023. doi: https://doi.org/10.1021/acs.est.2c04927.

[21] International Organization for Standardization. Additive manufacturing – general principles – fundamentals and vocabulary, 2021. URL `https://www.iso.org/standard/74524.html`. Last access: 07.08.2025.

[22] Sachini Wickramasinghe, Truong Do, and Phuong Tran. Fdm-based 3d printing of polymer and associated composite: A review on mechanical properties, defects and treatments. *Polymers*, 12(7):1529, 2020. doi: https://doi.org/10.3390/polym12071529.

[23] Azamat L Slonov, Azamat A Zhansitov, Diana M Khakulova, Zh I Kurdanova, Ismel V Musov, and S Yu Khashirova. Development of composite materials for 3d printing on the basis of polyphenylenesulfone. In *Materials Science Forum*, volume 935, pages 15–20. Trans Tech Publ, 2018. doi: https://doi.org/10.4028/www.scientific.net/MSF.935.15.

[24] Matteo Arioli, Jordi Puiggalí, and Lourdes Franco. Nylons with applications in energy generators, 3d printing and biomedicine. *Molecules*, 29 (11):2443, May 2024. doi: https://doi.org/10.3390/molecules29112443.

[25] Jiann-Shing Wu, Shu-Chen Shen, and Feng-Chih Chang. Effect of rubber content in acrylonitrile–butadiene–styrene and additional rubber on the polymer blends of polycarbonate and acrylonitrile–butadiene–styrene. *Polymer journal*, 26(1):33–42, 1994. doi: https://doi.org/10.1295/polymj.26.33.

[26] Mateusz Barczewski, Olga Mysiukiewicz, Aleksander Hejna, Radosław Biskup, Joanna Szulc, Sławomir Michałowski, Adam Piasecki, and Arkadiusz Kloziński. The effect of surface treatment with isocyanate and aromatic carbodiimide of thermally expanded vermiculite used as a functional filler for polylactide-based composites. *Polymers*, 13(6):890, 2021. doi: https://doi.org/10.3390/polym13060890.

[27] Jian-Yuan Lee, Jia An, and Chee Kai Chua. Fundamentals and applications of 3d printing for novel materials. *Applied materials today*, 7:120–133, 2017. doi: https://doi.org/10.1016/j.apmt.2017.02.004.

[28] Thirunahary Swetham, Ketham Madhana Mohan Reddy, Akhil Huggi, and M Naveen Kumar. A critical review on of 3d printing materials and details of materials used in fdm. *Int. J. Sci. Res. Sci. Eng. Technol*, 3(2): 353–361, 2017. ISSN 2394-4099.

[29] Nectarios Vidakis, Markos Petousis, Apostolos Korlos, Emmanouil Velidakis, Nikolaos Mountakis, Chrisa Charou, and Adrian Myftari. Strain rate sensitivity of polycarbonate and thermoplastic polyurethane for various 3d printing temperatures and layer heights. *Polymers*, 13(16):2752, 2021. doi: https://doi.org/10.3390/polym13162752.

[30] James E. Mark, editor. *Polymer Data Handbook*. Oxford University Press, Oxford; New York, 2 edition, 2009. ISBN 978-0195181012. doi: https://doi.org/10.1093/oso/9780195181012.001.0001.

[31] Yasong Zhao, Yangyunzhi Gao, Gaofeng Chen, Shujun Li, Amardeep Singh, Xu Luo, Cheng Liu, Jianming Gao, and Hongjian Du. Development of low-carbon materials from ggbs and clay brick powder for 3d concrete printing. *Construction and Building Materials*, 383:131232, 2023. doi: https://doi.org/10.1016/j.conbuildmat.2023.131232.

[32] Shareen S. L. Chan, Ryan M. Pennings, Lewis Edwards, and George V. Franks. 3d printing of clay for decorative architectural applications: Effect of solids volume fraction on rheology and printability. *Additive Manufacturing*, 35:101335, 2020. doi: https://doi.org/10.1016/j.addma.2020.101335.

[33] Sooraj A. O. Nair, Subhashree Panda, Manu Santhanam, Gaurav Sant, and Narayanan Neithalath. A critical examination of the influence of material characteristics and extruder geometry on 3d printing of cementitious binders. *Cement and Concrete Composites*, 112:103671, 2020. doi: https://doi.org/10.1016/j.cemconcomp.2020.103671.

[34] Arnaud Perrot, Alexandre Pierre, Venkatesh N. Nerella, Rob J. M. Wolfs, Elhadj Keita, Sooraj A. O. Nair, Narayanan Neithalath, Nicolas Roussel, and Viktor Mechtcherine. From analytical methods to numerical simulations: A process engineering toolbox for 3d concrete printing. *Cement and Concrete Composites*, 122:104164, 2021. doi: https://doi.org/10.1016/j.cemconcomp.2021.104164.

[35] Yanfang Wu, Junjie Lan, Mingxuan Wu, Wu Zhou, Shaobin Zhou, Hui Yang, Maolin Zhang, and Yue Li. Rheology and printability of a porcelain clay paste for diw 3d printing of ceramics with complex geometric structures. *ACS Omega*, 9(24):26450–26457, 2024. doi: https://doi.org/10.1021/acsomega.4c02543.

[36] Marcin Maroszek, Magdalena Rudziewicz, Adam Hutyra, Paweł Dziura, and Marek Hebda. Evaluation of 3d concrete printing extrusion efficiency. *Applied Sciences*, 14(24):11866, 2024. doi: https://doi.org/10.3390/app142411866.

[37] Nicolas Roussel. Rheological requirements for printable concretes. *Cement and Concrete Research*, 112:76–85, 2018. doi: https://doi.org/10.1016/j.cemconres.2018.04.005.

[38] Runze Huang, Matthew E. Riddle, Dana J. Graziano, Joshua A. Warren, Supratik Das, Sachin U. Nimbalkar, Joseph Cresko, and Eric Masanet. Energy and emissions saving potential of additive manufacturing: The case of lightweight aircraft components. *Journal of Cleaner Production*, 135: 1559–1570, 2016. doi: https://doi.org/10.1016/j.jclepro.2015.04.109.

[39] Eleonora Atzeni and Alessandro Salmi. Economics of additive manufacturing for end-usable metal parts. *The International Journal of Advanced Manufacturing Technology*, 62(9):1147–1155, 2012. doi: https://doi.org/10.1007/s00170-011-3878-1.

[40] Shamraiz Ahmad and Kuan Yew Wong. Sustainability assessment in the manufacturing industry: a review of recent studies. *Benchmarking: An International Journal*, 25(8):3162–3179, 2018. doi: https://doi.org/10.1108/BIJ-08-2017-0214.

[41] Gerald Rebitzer, T. Ekvall, R. Frischknecht, D. Hunkeler, G. Norris, T. Rydberg, W. P. Schmidt, S. Suh, B. P. Weidema, and D. W. Pennington. Life cycle assessment part 1: Framework, goal and scope definition, inventory analysis, and applications. *Environment International*, 30:701–720, 2004. doi: https://doi.org/10.1016/j.envint.2003.11.005.

[42] D.R. Vieira, M.D.T. Casagrande, and L.C.P. da Silva. Life cycle assessment (LCA) applied to the manufacturing of common and ecological concrete: A review. *Construction and Building Materials*, 123:454–462, oct 2016. doi: https://doi.org/10.1016/j.conbuildmat.2016.07.037. URL `https://doi.org/10.1016/j.conbuildmat.2016.07.037`.

[43] International Organization for Standardization. Environmental management – life cycle assessment – requirements and guidelines, 2006. URL `https://www.iso.org/standard/38498.html`. Last access: 07.08.2025.

[44] Marios D Chatzisideris, Nieves Espinosa, Alexis Laurent, and Frederik C Krebs. Ecodesign perspectives of thin-film photovoltaic technologies: A review of life cycle assessment studies. *Solar Energy Materials and Solar Cells*, 156:2–10, 2016. doi: https://doi.org/10.1016/j.solmat.2016.05.048.

[45] International Organization for Standardization. Environmental management – life cycle assessment – principles and framework, 2006. URL `https://www.iso.org/standard/37456.html`. Last access: 07.08.2025.

[46] Mary Ann Curran. *Overview of Goal and Scope Definition in Life Cycle Assessment*, pages 1–62. Springer Netherlands, Dordrecht, 2017. ISBN 978-94-024-0855-3. doi: 10.1007/978-94-024-0855-3_1.

[47] Sangwon Suh and Gjalt Huppes. Methods for life cycle inventory of a product. *Journal of cleaner production*, 13(7):687–697, 2005. doi: https://doi.org/10.1016/j.jclepro.2003.04.001.

[48] Karl G Mueller, Michael U Lampérth, and Fumihiko Kimura. Parameterised inventories for life cycle assessment: Systematically relating design parameters to the life cycle inventory. *The International Journal of Life Cycle Assessment*, 9(4):227–235, 2004. doi: https://doi.org/10.1007/BF02978598.

[49] Simon Eggleston, Leandro Buendia, Kyoko Miwa, Todd Ngara, and Kiyoto Tanabe. 2006 ipcc guidelines for national greenhouse gas inventories. Technical report, Institute for Global Environmental Strategies for the Intergovernmental Panel on Climate Change, Kanagawa, Japan, 2006. URL `https://www.ipcc-nggip.iges.or.jp/public/2006gl/`. Last access: 08.08.2025.

[50] Michael Z Hauschild, Alexandra Bonou, and Stig Irving Olsen. Life cycle interpretation. *Life cycle assessment: Theory and practice*, pages 323–334, 2018. doi: https://doi.org/10.1007/978-3-319-56475-3.

[51] Ólafur Ögmundarson, Markus J Herrgård, Jochen Forster, Michael Z Hauschild, and Peter Fantke. Addressing environmental sustainability of

biochemicals. *Nature Sustainability*, 3(3):167–174, 2020. doi: https://doi.org/10.1038/s41893-019-0442-8.

[52] Göran Finnveden, Michael Z. Hauschild, Henning Ekvall, et al. Recent developments in life cycle assessment. *Journal of Environmental Management*, 91(1):1–21, 2009. doi: https://doi.org/10.1016/j.jenvman.2009.06.018.

[53] David J Jackson and Matthew Brander. The risk of burden shifting from embodied carbon calculation tools for the infrastructure sector. *Journal of Cleaner Production*, 223:739–746, 2019. doi: https://doi.org/10.1016/j.jclepro.2019.03.171.

[54] Michael Braungart, William McDonough, and Anja Bollinger. Cradle-to-cradle design: creating healthy emissions – a strategy for eco-effective product and system design. *Journal of Cleaner Production*, 15(13-14):1337–1348, 2007. doi: https://doi.org/10.1016/j.jclepro.2006.08.003.

[55] Ari Nissinen, Jyri Seppälä, and Tero Heinonen. Make carbon footprints available – and it is not just one value. *Cleaner Logistics and Supply Chain*, 3:100023, 2022. doi: https://doi.org/10.1016/j.clscn.2021.100023.

[56] World Resources Institute (WRI) and World Business Council for Sustainable Development (WBCSD). Ghg protocol corporate accounting and reporting standard (revised edition). Technical report, WRI & WBCSD, 2015. URL `https://ghgprotocol.org/corporate-standard`. Last access: 07.08.2025.

[57] International Organization for Standardization. Greenhouse gases – carbon footprint of products – requirements and guidelines for quantification, 2018. URL `https://www.iso.org/standard/71206.html`. Last access: 07.08.2025.

[58] Maximilian Hettler and Lorenz Graf-Vlachy. Corporate scope 3 carbon emission reporting as an enabler of supply chain decarbonization: A systematic review and comprehensive research agenda. *Business Strategy and the Environment*, 33(2):263–282, 2024. doi: https://doi.org/10.1002/bse.3486.

[59] K. R. Murray and S. Boroń. The necessary paradigm shift in sustainable business practice part 1: The framework. *Journal of Management and Sustainability*, 2019. doi: https://doi.org/10.5539/JMS.V9N1P93.

[60] Julian Kirchherr, Denise Reike, and Marko Hekkert. Conceptualizing the circular economy: An analysis of 114 definitions. *Resources, Conservation and Recycling*, 127:221–232, 2017. doi: https://doi.org/10.1016/j.resconrec.2017.09.005.

[61] Salwa H Abdul Rashid, Stephen Evans, and Philip Longhurst. A comparison of four sustainable manufacturing strategies. *International Journal of Sustainable Engineering*, 1(3):214–229, 2008. doi: https://doi.org/10.1080/19397030802513836.

[62] Denise Reike, Walter J.V. Vermeulen, and Sjors Witjes. The circular economy: New or refurbished as ce 3.0? — exploring controversies in the conceptualization of the circular economy through a focus on history and resource value retention options. *Resources, Conservation and Recycling*, 135:246–264, 2018. doi: https://doi.org/10.1016/j.resconrec.2017.08.027.

[63] Peter Hopkinson, Markus Zils, Philip Hawkins, and Stuart Roper. Managing a complex global circular economy business model: Opportunities and challenges. *California Management Review*, 60(3):71–94, 2018. doi: https://doi.org/10.1177/0008125618764692.

[64] Lu Ding, Tong Wang, and Paul W Chan. Forward and reverse logistics for circular economy in construction: A systematic literature review. *Journal of Cleaner Production*, 388:135981, 2023. doi: https://doi.org/10.1016/j.jclepro.2023.135981.

[65] Shamraiz Ahmad, Kuan Yew Wong, and Shahid I. Butt. Status of sustainable manufacturing practices: literature review and trends of triple bottom-line-based sustainability assessment methodologies. *Environmental Science and Pollution Research*, 30:43068–43095, 2023. doi: https://doi.org/10.1007/s11356-022-22172-z.

[66] Tejendra Singh Singhal, Jinesh Kumar Jain, D Atchuta Ramacharyulu, Alok Jain, Dalael Saad Abdul-Zahra, Manjunatha Manjunatha, and Arun Pratap Srivastava. Eco-design of products and processes: A review on principles and tools for sustainable manufacturing. In *E3S Web of Conferences*, volume 505, page 01033. EDP Sciences, 2024. doi: https://doi.org/10.1051/e3sconf/202450501033.

[67] Marco Garetti and Marco Taisch. Sustainable manufacturing: trends and research challenges. *Production planning & control*, 23(2-3):83–104, 2012. doi: https://doi.org/10.1080/09537287.2011.591619.

[68] Heidi Fuchs, Arian Aghajanzadeh, and Peter L. Therkelsen. Identification of drivers, benefits, and challenges of iso 50001 through case study content analysis. *Energy Policy*, 142:111443, 2020. doi: https://doi.org/10.1016/j.enpol.2020.111443.

[69] International Organization for Standardization. Energy management systems – requirements with guidance for use, 2018. URL `https://www.iso.org/standard/69426.html`. Last access: 07.08.2025.

[70] Walter C. Satyro, José C. Contador, Sônia F. P. Monken, Anderson F. de Lima, and Gilberto G. Soares Jr. Industry 4.0 implementation projects: The cleaner production strategy—a literature review. *Sustainability*, 15(3): 2161, 2023. doi: https://doi.org/10.3390/su15032161.

[71] Rhitwik Sundar, AN Balaji, and RM Satheesh Kumar. A review on lean manufacturing implementation techniques. *Procedia engineering*, 97: 1875–1885, 2014. doi: https://doi.org/10.1016/j.proeng.2014.12.341.

[72] Joshua Thomas Jameson Burd, Elizabeth A Moore, Hesham Ezzat, Randolph Kirchain, and Richard Roth. Improvements in electric vehicle battery technology influence vehicle lightweighting and material substitution decisions. *Applied Energy*, 283:116269, 2021. doi: https://doi.org/10.1016/j.apenergy.2020.116269.

[73] Mirko Soković, Duško Pavletić, and Karmen Kern Pipan. Quality improvement methodologies–pdca cycle, radar matrix, dmaic and dfss. *Journal of Achievements in Materials and Manufacturing Engineering*, 43(1): 476–483, 2010. doi: 10.5604/01.3001.0011.6015.

[74] William J Frawley, Gregory Piatetsky-Shapiro, and Christopher J Matheus. Knowledge discovery in databases: An overview. *AI magazine*, 13(3):57–57, 1992. doi: https://doi.org/10.1609/aimag.v13i3.1011.

[75] Nguyen Van Hop, N Sumate, and N Sitawatch. Modifying integrated model for manufacturing process improvement. In *Proceedings of the International Conference on Simulation and Modeling*, pages 38–47, 2005.

[76] S. Vinodh, Jiju Antony, Rohit Agrawal, and Jacqueline Ann Douglas. Integration of continuous improvement strategies with industry 4.0: a systematic review and agenda for further research. *The TQM Journal*, 33(2): 441–472, 2021. doi: 10.1108/TQM-07-2020-0157.

[77] Fakhitah Ridzuan and Wan Mohd Nazmee Wan Zainon. A review on data cleansing methods for big data. *Procedia Computer Science*, 161:731–738, 2019. doi: https://doi.org/10.1016/j.procs.2019.11.177.

[78] Gary M Weiss and Ye Tian. Maximizing classifier utility when there are data acquisition and modeling costs. *Data Mining and Knowledge Discovery*, 17(2):253–282, 2008. doi: https://doi.org/10.1007/s10618-007-0082-x.

[79] Steven A. Weissman and Neal G. Anderson. Design of experiments (doe) and process optimization. a review of recent publications. *Organic Process Research & Development*, 19(11):1605–1633, 2014. doi: https://doi.org/10.1021/op500169m.

[80] André I Khuri and Siuli Mukhopadhyay. Response surface methodology. *Wiley interdisciplinary reviews: Computational statistics*, 2(2):128–149, 2010. doi: https://doi.org/10.1002/wics.73.

[81] Tijana Rakić, Irena Kasagić-Vujanović, Marko Jovanović, Biljana Jančić-Stojanović, and Darko Ivanović. Comparison of full factorial design, central composite design, and box-behnken design in chromatographic method development for the determination of fluconazole and its impurities. *Analytical Letters*, 47(8):1334–1347, 2014. doi: https://doi.org/10.1080/00032719.2013.867503.

[82] Richard F Gunst and Robert L Mason. Fractional factorial design. *Wiley Interdisciplinary Reviews: Computational Statistics*, 1(2):234–244, 2009. doi: https://doi.org/10.1002/wics.27.

[83] R. L. Plackett and J. P. Burman. The design of optimum multifactorial experiments. *Biometrika*, 33(4):305–325, 1946. doi: https://doi.org/10.1093/biomet/33.4.305.

[84] Saeed Maghsoodloo, Gultekin Ozdemir, Victoria Jordan, and Chen-Hsiu Huang. Strengths and limitations of taguchi's contributions to quality, manufacturing, and process engineering. *Journal of Manufacturing Systems*, 23 (2):73–126, 2004. doi: https://doi.org/10.1016/S0278-6125(05)00004-X.

[85] George E. P. Box and D. W. Behnken. Some new three level designs for the study of quantitative variables. *Technometrics*, 2(4):455–475, 1960. doi: https://doi.org/10.2307/1266454.

[86] Tian-Fang Zhang, Jian-Feng Yang, and Dennis K. J. Lin. Small Box-Behnken design. *Statistics & Probability Letters*, 81(8):1027–1033, 2011. doi: https://doi.org/10.1016/j.spl.2011.02.024.

[87] Michael D. McKay, Richard J. Beckman, and William J. Conover. A comparison of three methods for selecting values of input variables in the analysis of output from a computer code. *Technometrics*, 21(1):239–245, 2000. doi: https://doi.org/10.2307/1271432.

[88] I. M. Sobol′. Distribution of points in a cube and approximate evaluation of integrals. *U.S.S.R. Computational Mathematics and Mathematical Physics*, 7(4):86–112, 1967. doi: https://doi.org/10.1016/0041-5553(67)90144-9.

[89] Ilya M. Sobol'. Global sensitivity indices for nonlinear mathematical models and their monte carlo estimates. *Mathematics and Computers in Simulation*, 55(1–3):271–280, 2001. doi: https://doi.org/10.1016/S0378-4754(00)00270-6.

[90] Marissa Renardy, Louis R. Joslyn, Jess A. Millar, and Denise E. Kirschner. To Sobol or not to Sobol? The effects of sampling schemes in systems biology applications. *Mathematical Biosciences*, 337:108593, 2021. doi: https://doi.org/10.1016/j.mbs.2021.108593.

[91] Jack Kiefer. Optimum experimental designs. *J. Royal Statistical Society, Series B (Methodological)*, 21(2):272–319, 1959. doi: https://doi.org/10.1111/j.2517-6161.1959.tb00338.x.

[92] Ji Zhu, Xiao Zhao, Yang Sun, Giovanni Vessio, and Ezio Frontoni. Relational data cleaning meets artificial intelligence: A survey. *Data Science and Engineering*, 2024. doi: https://doi.org/10.1007/s41019-024-00266-7.

[93] Ziawasch Abedjan, Lukasz Golab, and Felix Naumann. Data profiling: A tutorial. In *Proceedings of the 2017 ACM International Conference on Management of Data*, pages 1747–1751, 2017. doi: https://doi.org/10.1145/3035918.3054772.

[94] Donald B. Rubin. Inference and missing data. *Biometrika*, 63(3):581–592, 1976. doi: https://doi.org/10.1093/biomet/63.3.581.

[95] Joseph L Schafer and John W Graham. Missing data: our view of the state of the art. *Psychological methods*, 7(2):147, 2002. doi: https://doi.org/10.1037//1082-989X.7.2.147.

[96] Salvador Garc'ia, Juli'an Luengo, and Francisco Herrera. *Data Preprocessing in Data Mining*. Springer, Cham, Switzerland, 2015. doi: https://doi.org/10.1007/978-3-319-10247-4.

[97] Mario Köppen. The curse of dimensionality. In *5th online world conference on soft computing in industrial applications (WSC5)*, volume 1, pages 4–8, 2000. doi: https://doi.org/10.1007/978-0-387-39940-9_133.

[98] Jeffrey Larson, Matt Menickelly, and Stefan M Wild. Derivative-free optimization methods. *Acta Numerica*, 28:287–404, 2019. doi: https://doi.org/10.1017/S0962492919000060.

[99] Tamara G. Kolda, Robert M. Lewis, and Virginia Torczon. Optimization by direct search: New perspectives on some classical and modern methods. *SIAM Review*, 45(3):385–482, 2003. doi: https://doi.org/10.1137/S003614450242889.

[100] Julius Pfrommer, Clemens Zimmerling, Jinzhao Liu, Luise Kärger, Frank Henning, and Jürgen Beyerer. Optimisation of manufacturing process parameters using deep neural networks as surrogate models. *Procedia CIRP*, 72:426–431, 2018. doi: https://doi.org/10.1016/j.procir.2018.03.046.

[101] Tobias Menold, Volkher Onuseit, Matthias Buser, Michael Haas, Nico Bär, and Andreas Michalowski. Laser material processing optimisation using bayesian optimisation: A generic tool. *Light: Advanced Manufacturing*, 5(3):355–365, 2024. doi: https://doi.org/10.37188/lam.2024.032.

[102] Clemens Zimmerling, Christian Poppe, Oliver Stein, and Luise Kärger. Optimisation of manufacturing process parameters for variable component geometries using reinforcement learning. *Materials & Design*, 214:110423, 2022. doi: https://doi.org/10.1016/j.matdes.2022.110423.

[103] Sara Lumbreras and Pedro Ciller. Interpretable optimisation: Why and how we should explain optimisation models. *Applied Sciences*, 15(10): 5732, 2025. doi: https://doi.org/10.3390/app15105732.

[104] Thomas Bartz-Beielstein, Christian WG Lasarczyk, and Mike Preuss. Sequential parameter optimization. In *2005 IEEE congress on evolutionary computation*, volume 1, pages 773–780. IEEE, 2005. doi: https://doi.org/10.1109/CEC.2005.1554761.

[105] Bobak Shahriari, Kevin Swersky, Ziyu Wang, Ryan P Adams, and Nando de Freitas. Taking the human out of the loop: A review of bayesian optimization. *Proceedings of the IEEE*, 104(1):148–175, 2016. doi: https://doi.org/10.1109/JPROC.2015.2494218.

[106] Jasper Snoek, Hugo Larochelle, and Ryan P Adams. Practical bayesian optimization of machine learning algorithms. *Advances in neural information processing systems*, 25, 2012. doi: https://doi.org/10.5555/2999325.2999464.

[107] David Stenger and Dirk Abel. Benchmark of bayesian optimization and metaheuristics for control engineering tuning problems with crash constraints. *arXiv e-prints*, arXiv:2211.02571, 2022. doi: https://doi.org/10.48550/arXiv.2211.02571.

[108] Masahiro Nomura, Shuhei Watanabe, Youhei Akimoto, Yoshihiko Ozaki, and Masaki Onishi. Warm starting CMA-ES for hyperparameter optimization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 9188–9196, 2021. doi: https://doi.org/10.1609/aaai.v35i10.17109.

[109] Bowen Lei, Tanner Quinn Kirk, Anirban Bhattacharya, Debdeep Pati, Xiaoning Qian, Raymundo Arroyave, and Bani K. Mallick. Bayesian optimization with adaptive surrogate models for automated experimental design. *npj Computational Materials*, 7:194, 2021. doi: https://doi.org/10.1038/s41524-021-00662-x.

[110] Donald R Jones, Matthias Schonlau, and William J Welch. Efficient global optimization of expensive black-box functions. *Journal of Global Optimization*, 13(4):455–492, 1998. doi: https://doi.org/10.1023/A:1008306431147.

[111] Peter I Frazier. Bayesian optimization. *Recent Advances in Optimization and Modeling of Contemporary Problems*, pages 255–278, 2018. doi: https://doi.org/10.1287/educ.2018.0188.

[112] James G March. Exploration and exploitation in organizational learning. *Organization science*, 2(1):71–87, 1991. doi: https://doi.org/10.1287/orsc.2.1.71.

[113] James S. Bergstra, Rémi Bardenet, Yoshua Bengio, and Balázs Kégl. Algorithms for hyper-parameter optimization. In *Advances in Neural Information Processing Systems 24 (NIPS 2011)*, pages 2546–2554. Curran Associates, Inc., 2011. doi: https://doi.org/10.5555/2986459.2986743.

[114] Christopher M. Bishop. *Pattern Recognition and Machine Learning*. Springer, New York, 2006. ISBN 978-0387310732.

[115] Carl Edward Rasmussen and Christopher K. I. Williams. *Gaussian Processes for Machine Learning*. MIT Press, Cambridge, MA, 2006. ISBN 978-0262182539.

[116] Yaochu Jin. Surrogate-assisted evolutionary computation: Recent advances and future challenges. *Swarm and Evolutionary Computation*, 1(2): 61–70, 2011. doi: https://doi.org/10.1016/j.swevo.2011.05.001.

[117] Yuji Roh, Geon Heo, and Steven Euijong Whang. A survey on data collection for machine learning: A big data–ai integration perspective. *IEEE Transactions on Knowledge and Data Engineering*, 33(4):1328–1347, 2021. doi: https://doi.org/10.1109/TKDE.2019.2946162.

[118] Neoklis Polyzotis, Sudip Roy, Steven Euijong Whang, and Martin Zinkevich. Data lifecycle challenges in production machine learning: A survey. *SIGMOD Record*, 47(2):17–28, 2018. doi: https://doi.org/10.1145/3299887.3299891.

[119] Curtis G. Northcutt, Lu Jiang, and Isaac L. Chuang. Confident learning: Estimating uncertainty in dataset labels. *Journal of Artificial Intelligence Research*, 70:1373–1411, 2021. doi: https://doi.org/10.1613/jair.1.12125.

[120] Amandalynne Paullada, Inioluwa Deborah Raji, Emily M. Bender, Emily Denton, and Alex Hanna. Data and its (dis)contents: A survey of dataset development and use in machine learning research. *Patterns*, 2(11):100336, 2021. doi: https://doi.org/10.1016/j.patter.2021.100336.

[121] Léon Bottou. Large-scale machine learning with stochastic gradient descent. In *Proceedings of COMPSTAT'2010*, pages 177–186. Springer, 2010. doi: https://doi.org/10.1007/978-3-7908-2604-3_16.

[122] Timothy O. Hodson. Root-mean-square error (rmse) or mean absolute error (mae): When to use them or not. *Geoscientific Model Development*, 15:5481–5487, 2022. doi: https://doi.org/10.5194/gmd-15-5481-2022.

[123] Qi Wang, Conghui Li, Xiaofeng Meng, Jun Zhang, Xuefei Zhang, Yiping Li, and Huaiyu Wan. A comprehensive survey of loss functions in machine learning. *Annals of Data Science*, 9(3):187–212, 2022. doi: https://doi.org/10.1007/s40745-020-00253-5.

[124] Joris de Myttenaere, Benoit Golden, Bruno Le Grand, and François Rossi. Mean absolute percentage error for regression models. *Neurocomputing*, 192:38–48, 2016. doi: https://doi.org/10.1016/j.neucom.2015.12.114.

[125] Casper Solheim Bojer and Jens Peder Meldgaard. Kaggle forecasting competitions: An overlooked learning opportunity. *International Journal of Forecasting*, 37(2):587–603, 2021. doi: https://doi.org/10.1016/j.ijforecast.2020.07.007.

[126] M. J. R. Healy. The use of r2 as a measure of goodness of fit. *Journal of the Royal Statistical Society*, 147:608–609, 1984. doi: 10.2307/2981848.

[127] Hoang Lan Vu, K. T. W. Ng, Amy Richter, and C. An. Analysis of input set characteristics and variances on k-fold cross validation for a recurrent neural network model on waste disposal rate estimation. *Journal of environmental management*, 311:114869, 2022. doi: 10.1016/j.jenvman.2022.114869.

[128] Tadayoshi Fushiki. Estimation of prediction error by using k-fold cross-validation. *Statistics and Computing*, 21:137–146, 2011. doi: https://doi.org/10.1007/s11222-009-9153-8.

[129] Sylvain Arlot and Alain Celisse. A survey of cross-validation procedures for model selection. *Statistics Surveys*, 2010. doi: https://doi.org/10.1214/09-SS054.

[130] Gavin C. Cawley and Nicola L. C. Talbot. On over-fitting in model selection and subsequent selection bias in performance evaluation. *Journal of Machine Learning Research*, 11(70):2079–2107, 2010. URL `http://jmlr.org/papers/v11/cawley10a.html`.

[131] NA Diamantidis, Dimitris Karlis, and Emmanouel A Giakoumakis. Unsupervised stratification of cross-validation for accuracy estimation. *Artificial Intelligence*, 116(1-2):1–16, 2000. doi: https://doi.org/10.1016/S0004-3702(99)00094-6.

[132] Matthias Feurer and Frank Hutter. Hyperparameter optimization. In *Automated machine learning: Methods, systems, challenges*, pages 3–33. Springer International Publishing Cham, 2019. doi: https://doi.org/10.1007/978-3-030-05318-5_1.

[133] Benjamin Antunes and David R. C. Hill. Random numbers for machine learning: A comparative study of reproducibility and energy consumption. *Journal of Data Science and Intelligent Systems*, 2024. doi: 10.47852/bonviewjdsis42024012.

[134] Xavier Bouthillier, Pierre Delaunay, Mirko Bronzi, Assya Trofimov, Brennan Nichyporuk, Justin Szeto, Nazanin Mohammadi Sepahvand, Edward Raff, Kanika Madan, Vikram Voleti, et al. Accounting for variance in machine learning benchmarks. *Proceedings of Machine Learning and Systems*, 2021. doi: https://doi.org/10.48550/ARXIV.2103.03098. arXiv:2103.03098.

[135] Rakesh P, J. R. Shruti, Indraneel Thippeswamy, Nithya B L, and Dheeraj V. A surrogate approach to explainable ai for predictive maintenance: Techniques and applications. *2024 5th International Conference on Circuits, Control, Communication and Computing (I4C)*, pages 160–165, 2024. doi: https://doi.org/10.1109/I4C62240.2024.10748524.

[136] G. Reddy and Y. V. P. Kumar. Explainable ai (xai): Explained. In *2023 IEEE Open Conference of Electrical, Electronic and Information Sciences*

(eStream), pages 1–6, 2023. doi: https://doi.org/10.1109/eStream59056.2023.10134984.

[137] Rudra Tiwari. Explainable ai (xai) and its applications in building trust and understanding in ai decision making. *International Journal of Scientific Research in Engineering and Management*, 2023. doi: https://doi.org/10.55041/ijsrem17592.

[138] Chiranjeevi Bura, Anil Kumar Jonnalagadda, and Prudhvi Naayini. The role of explainable ai (xai) in trust and adoption. *Journal of Artificial Intelligence General science (JAIGS)*, 2025. doi: https://doi.org/10.60087/jaigs.v7i01.331.

[139] Sajid Ali, Tamer Abuhmed, Shaker El-Sappagh, Khan Muhammad, Jose M. Alonso-Moral, Roberto Confalonieri, Riccardo Guidotti, Javier Del Ser, Natalia Diaz-Rodriguez, and Francisco Herrera. Explainable artificial intelligence (xai): What we know and what is left to attain trustworthy artificial intelligence. *Information Fusion*, 99:101805, 2023. doi: https://doi.org/10.1016/j.inffus.2023.101805.

[140] Arun Rai. Explainable ai: From black box to glass box. *Journal of the academy of marketing science*, 48:137–141, 2020. doi: https://doi.org/10.1007/s11747-019-00710-5.

[141] Manju Vallayil, Parma Nand, Weiqi Yan, and Héctor Allende-Cid. Explainability of automated fact verification systems: A comprehensive review. *Applied Sciences*, 13(23):12608, 2023. doi: https://doi.org/10.3390/app132312608.

[142] Sina Nahvi, Marco Schumann, Martin Dix, and Philipp Klimant. Intelligent parameter determination and quality control of an injection molding process. *Procedia CIRP*, 126:721–726, 2024. doi: https://doi.org/10.1016/j.procir.2024.08.297.

[143] Cynthia Rudin and Joanna Radin. Why are we using black box models in ai when we don't need to? a lesson from an explainable ai competition.

Harvard Data Science Review, 1(2):1–9, 2019. doi: https://doi.org/10.1162/99608f92.5a8a3a3d.

[144] Jan Achterhold, Philip Tobuschat, Hao Ma, Dieter Buechler, Michael Muehlebach, and Joerg Stueckler. Black-box vs. gray-box: A case study on learning table tennis ball trajectory prediction with spin and impacts. In *Proceedings of the 5th Annual Learning for Dynamics and Control Conference*, volume 211 of *Proceedings of Machine Learning Research*, pages 878–890, 2023. doi: https://doi.org/10.48550/arXiv.2305.15189.

[145] Prachi Zodage, Hussain Harianawala, Hafsa Shaikh, and Asad Kharodia. Explainable ai (xai): History, basic ideas and methods. *International Journal of Advanced Research in Science, Communication and Technology*, 2024. doi: https://doi.org/10.48175/ijarsct-16988.

[146] Matthew J Page, Joanne E McKenzie, Patrick M Bossuyt, Isabelle Boutron, Tammy C Hoffmann, Cynthia D Mulrow, Larissa Shamseer, Jennifer M Tetzlaff, Elie A Akl, Sue E Brennan, et al. The prisma 2020 statement: an updated guideline for reporting systematic reviews. *bmj*, 372, 2021. doi: 10.1136/bmj.n71.

[147] Toon Vanderschueren, Tim Verdonck, Bart Baesens, and Wouter Verbeke. Predict-then-optimize or predict-and-optimize? an empirical evaluation of cost-sensitive learning strategies. *Information sciences*, 594: 400–415, 2022. doi: https://doi.org/10.1016/j.ins.2022.02.021.

[148] Alan Anis Lahoud, Ahmad Saeed Khan, Erik Schaffernicht, Marco Trincavelli, and Johannes Andreas Stork. Predict-and-optimize techniques for data-driven optimization problems: A review. *Neural Processing Letters*, 57(2):1–28, 2025. doi: https://doi.org/10.1007/s11063-025-11746-w.

[149] Jayanta Mandi, James Kotary, Senne Berden, Maxime Mulamba, Victor Bucarey, Tias Guns, and Ferdinando Fioretto. Decision-focused learning: Foundations, state of the art, benchmark and future opportunities. *Journal of Artificial Intelligence Research*, 80:1623–1701, 2024. doi: https://doi.org/10.1613/jair.1.15320.

[150] Irfan Khan, Xianchao Zhang, Rahman Ali, and Masood Habib. Meta-learning for automated algorithm selection in healthcare: Enhancing clinical outcomes through data-driven recommendation. In *2024 International Conference on IT and Industrial Technologies (ICIT)*, pages 1–6. IEEE, 2024. doi: https://doi.org/10.1109/ICIT63607.2024.10859997.

[151] Alexandre Forel, Axel Parmentier, and Thibaut Vidal. Explainable data-driven optimization: From context to decision and back again. In *Proceedings of the 40th International Conference on Machine Learning (ICML)*, volume 202 of *Proceedings of Machine Learning Research*, pages 10170–10187, 2023. doi: https://doi.org/10.48550/arXiv.2301.10074.

[152] Koen W. De Bock, Kristof Coussement, Arno De Caigny, Roman Słowiński, Bart Baesens, Robert N. Boute, Tsan-Ming Choi, Dursun Delen, Mathias Kraus, Stefan Lessmann, Sebastián Maldonado, David Martens, María Óskarsdóttir, Carla Vairetti, Wouter Verbeke, and Richard Weber. Explainable AI for operational research: A defining framework, methods, applications, and a research agenda. *European Journal of Operational Research*, 317(2):249–272, 2024. doi: https://doi.org/10.1016/j.ejor.2023.09.026.

[153] Yun-Tian Chen, Dong-Xiao Zhang, Qun Zhao, and De-Xun Liu. Interpretable machine learning optimization (interopt) for operational parameters: A case study of highly-efficient shale gas development. *Petroleum Science*, 20(3):1788–1805, 2023. doi: https://doi.org/10.1016/j.petsci.2022.12.017.

[154] Fuguo Liu, Ziru Chen, Jun Xu, Yanyan Zheng, Wenyi Su, Maozai Tian, and Guodong Li. Interpretable machine learning-based influence factor identification for 3d printing process–structure linkages. *Polymers*, 16(18):2680, 2024. doi: https://doi.org/10.3390/polym16182680.

[155] Mandana Kariminejad, David Tormey, Caitr'iona Ryan, Christopher O'Hara, Albert Weinert, and Marion McAfee. Single and multi-objective real-time optimisation of an industrial injection moulding process via a

bayesian adaptive design of experiment approach. *Scientific Reports*, 14: 29799, 2024. doi: https://doi.org/10.1038/s41598-024-80405-2.

[156] Manal Alghieth. Sustain ai: A multi-modal deep learning framework for carbon footprint reduction in industrial manufacturing. *Sustainability*, 17 (9):4134, 2025. doi: https://doi.org/10.3390/su17094134.

[157] Panagiotis D. Paraschos, Georgios K. Koulinas, and Dimitrios E. Koulouriotis. Reinforcement learning-based optimization for sustainable and lean production within the context of industry 4.0. *Algorithms*, 17(3):98, 2024. doi: https://doi.org/10.3390/a17030098.

[158] Lasha Gelovani and Elene Mikeladze. A framework for ai driven optimization of sustainable manufacturing processes and resource efficient production systems. *Innovations in Sustainable Technologies, Environmental Practices, and Policy Development*, 14(12):1–14, 2024. URL `https://owenpress.com/index.php/ISTEPPD/article/view/2024-12-04/9`. Last access: 07.08.2025.

[159] Pamela Hermosilla, Sebastián Berríos, and Héctor Allende-Cid. Explainable AI for forensic analysis: A comparative study of SHAP and LIME in intrusion detection models. *Applied Sciences*, 15(13):7329, 2025. doi: https://doi.org/10.3390/app15137329.

[160] Konstantinos Nikiforidis, Alkiviadis Kyrtsoglou, Thanasis Vafeiadis, Thanasis Kotsiopoulos, Alexandros Nizamis, Dimosthenis Ioannidis, Konstantinos Votis, Dimitrios Tzovaras, and Panagiotis G. Sarigiannidis. Enhancing transparency and trust in ai-powered manufacturing: A survey of explainable AI (xai) applications in smart manufacturing in the era of industry 4.0/5.0. *ICT Express*, 11(1):135–148, 2025. doi: https://doi.org/10.1016/j.icte.2024.12.001.

[161] AS Cruz, LR Caldas, VM Mendes, JC Mendes, and LEG Bastos. Multi-objective optimization based on surrogate models for sustainable building design: a systematic literature review. *Building and Environment*, 266: 112147, 2024. doi: https://doi.org/10.1016/j.buildenv.2024.112147.

[162] Nastaran Gholizadeh and Petr Musilek. Explainable reinforcement learning for distribution network reconfiguration. *Energy Reports*, 11:5703–5715, 2024. doi: https://doi.org/10.1016/j.egyr.2024.05.031.

[163] Zichen He, Wenchong Tian, Jiaying Wang, Hexiang Yan, Kunlun Xin, and Tao Tao. Deep reinforcement learning control as an innovative approach for urban drainage systems: review and prospects. *Water Research*, page 123954, 2025. doi: https://doi.org/10.1016/j.watres.2025.123954.

[164] Jagadeesh Kumar Janga, Krishna R Reddy, and KVNS Raviteja. Integrating artificial intelligence, machine learning, and deep learning approaches into remediation of contaminated sites: A review. *Chemosphere*, 345: 140476, 2023. doi: https://doi.org/10.1016/j.chemosphere.2023.140476.

[165] Carolina Soares de Morais, Tânia Rodrigues Pereira Ramos, Manuel Lopes, and Ana Paula Barbosa-Póvoa. A data-driven optimization approach to plan smart waste collection operations. *International Transactions in Operational Research*, 31(4):2178–2208, 2024. doi: https://doi.org/10.1111/itor.13235.

[166] P Jyoteeshkumar Reddy, Sandeep Chinta, Harish Baki, Richard Matear, and John Taylor. Gaussian process regression-based bayesian optimisation (g-bo) of model parameters – a wrf model case study of southeast australia heat extremes. *Geophysical Research Letterss*, 2024. doi: https://doi.org/10.1029/2024GL111074.

[167] Patrick O'Loughlin, Emily Jones, and Sarah Brown. Toward inherently interpretable ai components in climate modeling. *Environmental Modelling & Software*, 172:105044, 2025. doi: https://doi.org/10.1016/j.envsoft.2024.105044.

[168] Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act), 2024. URL `https://eur-lex.europa.eu/eli/reg/2024/1689/oj`. Last access: 07.08.2025.

[169] Jason E. Johnson, Ishat R. Jamil, Liang Pan, Guang Lin, and Xianfan Xu. Bayesian optimization with gaussian-process-based active machine learning for improvement of geometric accuracy in projection multi-photon 3d printing. *Light: Science & Applications*, 14(1):56, 2025. doi: https://doi.org/10.1038/s41377-024-01707-8.

[170] Vispi Karkaria, Anthony Goeckner, Rujing Zha, Jie Chen, Jianjing Zhang, Qi Zhu, Jian Cao, Robert X Gao, and Wei Chen. Towards a digital twin framework in additive manufacturing: Machine learning and bayesian optimization for time series process optimization. *Journal of Manufacturing Systems*, 75:322–332, 2024. doi: https://doi.org/10.1016/j.jmsy.2024.04.023.

[171] Hao Zhang, Zhaoqi Zong, Yijun Yao, Qing Hu, Mohamed Aburaia, and Heiko Lammer. Multi-axis 3d printing defect detecting by machine vision with convolutional neural networks. *Experimental Techniques*, 47:619–631, 2023. doi: https://doi.org/10.1007/s40799-022-00577-2.

[172] Qiao Wu, Naiming Xie, Shaoxiang Zheng, and Alain Bernard. Online order scheduling of multi 3d printing tasks based on the additive manufacturing cloud platform. *Journal of Manufacturing Systems*, 63:23–34, 2022. doi: https://doi.org/10.1016/j.jmsy.2022.02.007.

[173] Jianjia He, Jian Wu, Ye Zhang, Yaopeng Wang, and Hua He. Large-scale customized production scheduling of multiagent-based medical 3d printing. *Computational Intelligence and Neuroscience*, 2022(1):6557137, 2022. doi: https://doi.org/10.1155/2022/6557137.

[174] Siemens Digital Industries Software. Balancing efficiency and sustainability: How advanced planning and scheduling solutions (aps) transform production planning. White paper, 2024. URL `https://resources.sw.siemens.com/en-US/white-paper-balancing-efficiency-and-sustainability/`. Last access: 05.07.2025.

[175] Siemens Digital Industries. Using a comprehensive digital twin to reduce high-bay warehouse commissioning time by 30 percent. Case study, 2023. URL https://resources.sw.siemens.com/en-US/case-study-ferrero/. Last access: 05.07.2025.

[176] AspenTech. Aspen dmc3 ™: Sustain optimal process performance with ai-enabled adaptive process control. Product page, 2025. URL https://www.aspentech.com/en/products/msc/aspen-dmc3. Last access: 05.07.2025.

[177] Diego Tamburini. Leveraging ai and digital twins to transform manufacturing with sight machine. https://azure.microsoft.com/en-us/blog/leveraging-ai-and-digital-twins-to-transform-manufacturing-with-sight-machine/?msockid=28b26639f2106a872715724df37b6b90, 2019. Last access: 11.08.2025.

[178] Uptake Technologies Inc. https://uptake.com/, 2025. Last access: 11.08.2025.

[179] Bosch Energy and Building Solutions. Bosch location in nashik, complete energy efficiency, record-breaking solar power plant. Web page, 2025. URL https://www.boschbuildingsolutions.com/xc/en/news-and-stories/energy-management-at-nashik-location/. Last access: 05.07.2025.

[180] BMW Group. Things get colourful in debrecen – paint shop becomes first technology to go on-stream at new bmw group plant. Press release, 2024. URL https://www.press.bmwgroup.com/global/article/detail/T0444323EN/things-get-colourful-in-debrecen-%E2%80%93-paint-shop-becomes-first-technology-to-go-on-stream-at-new-bmw-group-plant?language=en. Last access: 05.07.2025.

[181] Bosch Global. Artificial intelligence in manufacturing: The ai analytics platform at bosch bamberg. Web page, 2025. URL https://www.bosch.com/stories/ai-in-manufacturing/. Last access: 05.07.2025.

[182] Fujitsu Laboratories Ltd. Fujitsu develops ai for image inspection to detect abnormalities in product appearance with world-leading precision in key benchmark. Press Release, Mar 2021. URL https://www.fujitsu.com/global/about/resources/news/press-releases/2021/0329-01.html. Last access: 05.07.2025.

[183] BMW Group. Smart maintenance using artificial intelligence. Press release, 2023. URL https://www.press.bmwgroup.com/global/article/detail/T0438145EN/smart-maintenance-using-artificial-intelligence?language=en. Last access: 05.07.2025.

[184] Amazon Web Services. Siemens energy builds industrial iot platform and drives smart manufacturing using aws iot. Case study, 2024. URL https://aws.amazon.com/solutions/case-studies/siemens-energy-video-case-study/. Last access: 05.07.2025.

[185] Richard Evans and Jim Gao. Deepmind ai reduces google data centre cooling bill by 40%, 2016. URL https://deepmind.google/discover/blog/deepmind-ai-reduces-google-data-centre-cooling-bill-by-40/. Last access: 11.08.2025.

[186] Bosch Center for Artificial Intelligence. Applied ai at bosch center for artificial intelligence. Web page, 2025. URL https://www.bosch-ai.com/industrial-ai/#:~:text=We%20apply%20big%20data%20and%20machine%20learning%20to,improve%20quality%20of%20Bosch%20products%2C%20and%20reduce%20costs. Last access: 05.07.2025.

[187] Mona Faraji Niri, Carl J. Reynolds, Luis A. Román-Ramírez, Emma Kendrick, and James Marco. Systematic analysis of the impact of slurry coating on manufacture of li-ion battery electrodes via explainable machine learning. *Energy Storage Materials*, 51:223–238, 2022. doi: https://doi.org/10.1016/j.ensm.2022.06.036.

[188] Tim von Hahn and Chris K. Mechefske. Machine learning in cnc machining: Best practices. *Machines*, 10(12):1233, 2022. doi: https://doi.org/10.3390/machines10121233.

[189] Eunbin Cho, Taewon Chang, and Guhyun Hwang. Data preprocessing combination to improve the performance of quality classification in the manufacturing process. *Electronics*, 11(3):477, 2022. doi: https://doi.org/10.3390/electronics11030477.

[190] Umberto Amato, Anestis Antoniadis, Italia De Feis, Domenico Fazio, Caterina Genua, Irène Gijbels, Donatella Granata, Antonino La Magna, Daniele Pagano, Gabriele Tochino, and Patrizia Vasquez. Predictive maintenance of pins in the ecd equipment for cu deposition in the semiconductor industry. *Sensors*, 23(14):6249, 2023. doi: https://doi.org/10.3390/s23146249.

[191] Sebastian Tillmann, Markus Behr, and Stefanie Elgeti. Using bayesian optimization for warpage compensation in injection molding. *Materialwissenschaft und Werkstofftechnik*, 55:13–20, 2024. doi: https://doi.org/10.1002/mawe.202300157.

[192] R. Hampel, F.˙ The influence curve and its role in robust estimation. *Journal of the American Statistical Association*, 69(346):383–393, 1974. doi: https://doi.org/10.2307/2285666.

[193] J. Rousseeuw, P. and C. Croux. Alternatives to the median absolute deviation. *Journal of the American Statistical Association*, 88(424):1273–1283, 1993. doi: https://doi.org/10.2307/2291267.

[194] Peter J. Rousseeuw and Annick M. Leroy. *Robust Regression and Outlier Detection*, pages 313–329. Wiley Series in Probability and Mathematical Statistics. John Wiley & Sons, New York, 1987. ISBN 9780471725381.

[195] Adriano VB Chagas, Caio SA Felix, Saulo VA Dantas, Adriana S Lima, Francisco AS Cunha, Marcia MC Ferreira, Ricardo B da Silva, Liz O dos Santos, and Sergio LC Ferreira. Use of multi-element robust z-score

and dd-simca methods in the homogeneity study for preparation of a soil laboratory reference material. *Microchemical Journal*, page 114783, 2025. doi: https://doi.org/10.1016/j.microc.2025.114783.

[196] Fei Tony Liu, Kai Ming Ting, and Zhi-Hua Zhou. Isolation forest. In *2008 Eighth IEEE International Conference on Data Mining*, pages 413–422. IEEE, 2008. doi: https://doi.org/10.1109/ICDM.2008.17.

[197] Sudipto Guha, Nina Mishra, Gourav Roy, and Okke Schrijvers. Robust random cut forest based anomaly detection on streams. In Maria Florina Balcan and Kilian Q. Weinberger, editors, *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 2712–2721, New York, New York, USA, 20–22 Jun 2016. PMLR. URL `https://proceedings.mlr.press/v48/guha16.html`.

[198] Donald B. Rubin. *Multiple Imputation for Nonresponse in Surveys*. John Wiley & Sons, New York, 1987. ISBN 9780471658726.

[199] Roderick J. A. Little. A test of missing completely at random for multivariate data with missing values. *Journal of the American Statistical Association*, 83(404):1198–1202, 1988. doi: https://doi.org/10.1080/01621459.1988.10478722.

[200] Craig K. Enders. *Applied Missing Data Analysis*. Guilford Press, New York, 2010. ISBN 9781606236390.

[201] Yaoyao Wang, Suchang Ma, J. Wang, Haifeng Pu, and Pu Wen-Feng. Machine manufacturing process improvement and optimisation. *Highlights in Science, Engineering and Technology*, 2023. doi: https://doi.org/10.54097/hset.v62i.10434.

[202] James J. Heckman. Sample selection bias as a specification error. *Econometrica*, 47(1):153–161, 1979. doi: https://doi.org/10.2307/1912352.

[203] Peter J. Diggle and Michael G. Kenward. Informative drop-out in longitudinal data analysis. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 43(1):49–93, 1994. doi: https://doi.org/10.2307/2986113.

[204] Javier Palarea-Albaladejo. Imputation. In B.S. Daya Sagar, Q. Cheng, J. McKinley, and F. Agterberg, editors, *Encyclopedia of Mathematical Geosciences*, Encyclopedia of Earth Sciences Series. Springer, Cham, 2023. doi: https://doi.org/10.1007/978-3-030-85040-1_424.

[205] Zhongheng Zhang. Missing data imputation: focusing on single imputation. *Annals of translational medicine*, 4(1):9, 2016. doi: https://doi.org/10.3978/j.issn.2305-5839.2015.12.38.

[206] John M. Lachin. Fallacies of last observation carried forward analyses. *Clinical Trials*, 13(2):161–168, 2016. doi: https://doi.org/10.1177/1740774515602688.

[207] Olga Troyanskaya, Michael Cantor, Gavin Sherlock, Pat Brown, Trevor Hastie, Robert Tibshirani, David Botstein, and Russ B. Altman. Missing value estimation methods for dna microarrays. *Bioinformatics*, 17(6):520–525, 2001. doi: https://doi.org/10.1093/bioinformatics/17.6.520.

[208] Arthur P. Dempster, Nan M. Laird, and Donald B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B*, 39(1):1–22, 1977. doi: https://doi.org/10.1111/j.2517-6161.1977.tb01600.x.

[209] Roderick J. A. Little. Pattern-mixture models for multivariate incomplete data. *Journal of the American Statistical Association*, 88(421):125–134, 1993. doi: https://doi.org/10.2307/2290705.

[210] Michael C. Wu and Raymond J. Carroll. Estimation and comparison of changes in the presence of informative right censoring by modeling the censoring process. *Biometrics*, 44(1):175–188, 1988. doi: https://doi.org/10.2307/2531694.

[211] Saichon Sinsomboonthong. Performance comparison of new adjusted min-max with decimal scaling and statistical column normalization methods for artificial neural network classification. *International Journal of Mathematics and Mathematical Sciences*, 2022(1):3584406, 2022. doi: https://doi.org/10.1155/2022/3584406.

[212] scikit-learn developers. *sklearn.preprocessing.RobustScaler: Scale features using statistics that are robust to outliers*. scikit-learn, 2024. URL `https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.RobustScaler.html`. Version 1.6.1 documentation. Last access: 25.04.2025.

[213] Andrew Gelman. Scaling regression inputs by dividing by two standard deviations. *Statistics in medicine*, 27(15):2865–2873, 2008. doi: https://doi.org/10.1002/sim.3107.

[214] Anthony C. Atkinson, Marco Riani, and Aldo Corbellini. The box–cox transformation: Review and extensions. *Statistical Science*, 36(2):239–255, 2021. ISSN 0883-4237. doi: https://doi.org/10.1214/20-STS778. URL `https://doi.org/10.1214/20-STS778`.

[215] I.-K. Yeo and R. A. Johnson. A new family of power transformations to improve normality or symmetry. *Biometrika*, 87(4):954–959, 2000. doi: https://doi.org/10.1093/biomet/87.4.954.

[216] Shuntaro Okada, Masayuki Ohzeki, and Shinichiro Taguchi. Efficient partition of integer optimization problems with one-hot encoding. *Scientific reports*, 9(1):13036, 2019. doi: https://doi.org/10.1038/s41598-019-49539-6.

[217] Chung-Chian Hsu and Shu-Han Lin. Visualized analysis of mixed numeric and categorical data via extended self-organizing map. *IEEE transactions on neural networks and learning systems*, 23(1):72–86, 2011. doi: https://doi.org/10.1109/TNNLS.2011.2178323.

[218] Isabelle Guyon and André Elisseeff. An introduction to variable and feature selection. *Journal of Machine Learning Research*, 3:1157–1182, 2003. doi: https://doi.org/10.1162/153244303322753616.

[219] Hervé Abdi and Lynne J Williams. Principal component analysis. *Wiley interdisciplinary reviews: computational statistics*, 2(4):433–459, 2010. doi: https://doi.org/10.1002/wics.101.

[220] I. T. Jolliffe. *Principal Component Analysis*. Springer, 2nd edition, 2002. doi: https://doi.org/10.1007/b98835.

[221] Richard Bellman. *Dynamic Programming*. Princeton University Press, Princeton, NJ, USA, 1957. ISBN 9780486428093.

[222] Douglas C. Montgomery. *Introduction to Statistical Quality Control*. Wiley, Hoboken, NJ, 7th edition, 2012. ISBN 9781118146811. doi: https://doi.org/10.1002/9781118146811.

[223] John R Rice. The algorithm selection problem. In *Advances in computers*, volume 15, pages 65–118. Elsevier, 1976. doi: https://doi.org/10.1016/S0065-2458(08)60520-3.

[224] Maryam Assafo, Jost Philipp Städter, Tenia Meisel, and Peter Langendörfer. On the stability and homogeneous ensemble of feature selection for predictive maintenance: A classification application for tool condition monitoring in milling. *Sensors*, 23(9):4461, 2023. doi: https://doi.org/10.3390/s23094461.

[225] Jason L. Loeppky, Jerome Sacks, and William J. Welch. Choosing the sample size of a computer experiment: A practical guide. *Technometrics*, 51(4):366–376, 2009. doi: https://doi.org/10.1198/TECH.2009.08040.

[226] Siwei Zhu, Holger R. Maier, and Angus C. Zecchin. Identification of metrics suitable for determining the features of real-world optimisation problems. *Environmental Modelling & Software*, 148:105281, 2022. doi: https://doi.org/10.1016/j.envsoft.2021.105281.

[227] Terry Jones, Stephanie Forrest, et al. Fitness distance correlation as a measure of problem difficulty for genetic algorithms. In *ICGA*, volume 95, pages 184–192, 1995.

[228] Yuan Sun, Saman K Halgamuge, Michael Kirley, and Mario A Munoz. On the selection of fitness landscape analysis metrics for continuous optimization problems. In *7th International Conference on Information and Automation for Sustainability*, pages 1–6. IEEE, 2014. doi: https://doi.org/10.1109/ICIAFS.2014.7069635.

[229] Arnaud Liefooghe, Sébastien Verel, Hernán Aguirre, and Kiyoshi Tanaka. What makes an instance difficult for black-box 0–1 evolutionary multiobjective optimizers? In *International Conference on Artificial Evolution (Evolution Artificielle)*, pages 3–15. Springer, 2013. doi: https://doi.org/10.1007/978-3-319-11683-9_1.

[230] Hooman H Rashidi, Nam Tran, Samer Albahra, and Luke T Dang. Machine learning in health care and laboratory medicine: General overview of supervised learning and auto-ml. *International Journal of Laboratory Hematology*, 43:15–22, 2021. doi: https://doi.org/10.1111/ijlh.13537.

[231] Léo Grinsztajn, Edouard Oyallon, and Gaël Varoquaux. Why do tree-based models still outperform deep learning on tabular data? In *Advances in Neural Information Processing Systems 35 (NeurIPS 2022)*, pages 507–520, 2022. doi: https://doi.org/10.48550/arXiv.2207.08815.

[232] Felix Conrad, Mauritz Mälzer, Michael Schwarzenberger, Hajo Wiemer, and Steffen Ihlenfeldt. Benchmarking AutoML for Regression Tasks on Small Tabular Data in Materials Design. *Scientific Reports*, 12(1):19350, 2022. doi: https://doi.org/10.1038/s41598-022-23327-1.

[233] Ahmed Kadra, Marius Lindauer, Frank Hutter, and Josif Grabocka. Well-tuned simple nets excel on tabular datasets. In *Advances in Neural Information Processing Systems 34 (NeurIPS 2021)*, pages 23928–23941, 2021. doi: https://doi.org/10.48550/arXiv.2106.11189.

[234] Yury Gorishniy, Ivan Rubachev, Valentin Khrulkov, and Artem Babenko. Revisiting deep learning models for tabular data. In *Advances in Neural Information Processing Systems 34 (NeurIPS 2021)*, pages 18932–18943, 2021. doi: https://doi.org/10.48550/arXiv.2106.11959.

[235] Noah Hollmann, Samuel Müller, Lennart Purucker, Arjun Krishnakumar, Max Körfer, Shi Bin Hoo, Robin T. Schirrmeister, and Frank Hutter. Accurate predictions on small data with a tabular foundation model. *Nature*, 637:319–326, 2025. doi: https://doi.org/10.1038/s41586-024-08328-6.

[236] Nick Erickson, Jonas Mueller, Alexander Shirkov, Hang Zhang, Pedro Larroy, Mu Li, and Alexander Smola. Autogluon-tabular: Robust and accurate automl for structured data. *CoRR*, abs/2003.06505, 2020. doi: https://doi.org/10.48550/arXiv.2003.06505.

[237] Sven Kruschel, Nico Hambauer, Sven Weinzierl, Sandra Zilker, Mathias Kraus, and Patrick Zschech. Challenging the performance-interpretability trade-off: An evaluation of interpretable machine learning models. *Business & Information Systems Engineering*, 2025. doi: https://doi.org/10.1007/s12599-024-00922-2.

[238] J. Ross Quinlan. Induction of decision trees. *Machine learning*, 1(1): 81–106, 1986. doi: https://doi.org/10.1007/BF00116251.

[239] Leo Breiman. Bagging predictors. *Machine Learning*, 24(2):123–140, 1996. doi: https://doi.org/10.1007/BF00058655.

[240] Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001. doi: https://doi.org/10.1023/A:1010933404324.

[241] Jerome H. Friedman. Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, 29(5):1189–1232, 2001. doi: https://doi.org/10.1214/aos/1013203451.

[242] Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proc. 22nd ACM SIGKDD Int. Conf. on Knowledge Discovery*

and Data Mining (KDD), pages 785–794, 2016. doi: https://doi.org/10.1145/2939672.2939785.

[243] Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. Lightgbm: A highly efficient gradient boosting decision tree. In *Advances in Neural Information Processing Systems*, volume 30, pages 3146–3154, 2017. doi: https://doi.org/10.5555/3294996.3295074.

[244] Liudmila Prokhorenkova, Gleb Gusev, Aleksandr Vorobev, Anna Veronika Dorogush, and Andrey Gulin. Catboost: Unbiased boosting with categorical features. In *Advances in Neural Information Processing Systems*, volume 31, pages 6638–6648, 2018. doi: https://doi.org/10.1145/3701716.3718479.

[245] Xin Huang, Ashish Khetan, Milan Cvitkovic, and Zohar S. Karnin. Tabtransformer: Tabular data modeling using contextual embeddings. *CoRR*, abs/2012.06678, 2020. doi: https://doi.org/10.48550/arXiv.2012.06678.

[246] Gowthami Somepalli, Micah Goldblum, Avi Schwarzschild, C. Bayan Bruss, and Tom Goldstein. Saint: Improved neural networks for tabular data via row attention and contrastive pre-training. *CoRR*, abs/2106.01342, 2021. doi: https://doi.org/10.48550/arXiv.2106.01342.

[247] Sercan Ö. Arik and Tomas Pfister. Tabnet: Attentive interpretable tabular learning. *CoRR*, abs/1908.07442, 2020. doi: https://doi.org/10.48550/arXiv.1908.07442.

[248] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[249] Noah Hollmann, Samuel Müller, Katharina Eggensperger, and Frank Hutter. Tabpfn: A transformer that solves small tabular classification problems in a second. *CoRR*, abs/2207.01848, 2023. doi: https://doi.org/10.48550/arXiv.2207.01848.

[250] Matthias Feurer, Aaron Klein, Katharina Eggensperger, Jost Tobias Springenberg, Manuel Blum, and Frank Hutter. Auto-sklearn: Efficient and robust automated machine learning. In *Automated Machine Learning*, The Springer Series on Challenges in Machine Learning, pages 113–134. Springer, Cham, 2019. doi: https://doi.org/10.1007/978-3-030-05318-5_6.

[251] Trevor Hastie and Robert Tibshirani. Generalized additive models. *Statistical Science*, 1(3):297–310, 1986. doi: https://doi.org/10.1214/ss/1177013604.

[252] Harsha Nori, Samuel Jenkins, Paul Koch, and Rich Caruana. Interpretml: A unified framework for machine learning interpretability. *arXiv preprint arXiv:1909.09223*, 2019. doi: https://doi.org/10.48550/arXiv.1909.09223.

[253] Katharina Eggensperger, Philipp Müller, Neeratyoy Mallik, Matthias Feurer, René Sass, Noor Awad, Marius Lindauer, and Frank Hutter. HPOBench: A collection of reproducible multi-fidelity benchmark problems for hpo. In *35th Conf. Neural Information Processing Systems (NeurIPS 2021), Datasets and Benchmarks Track*, 2021. doi: https://doi.org/10.48550/arXiv.2109.06716. URL `https://arxiv.org/abs/2109.06716`.

[254] Shashank Shekhar, Adesh Bansode, and Asif Salim. A comparative study of hyper-parameter optimization tools. In *Proc. 2021 IEEE Asia-Pacific Conf. on Computer Science and Data Engineering (CSDE)*, pages 1–6. IEEE, 2021. doi: https://doi.org/10.1109/CSDE53843.2021.9718485.

[255] Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 2623–2631, 2019. doi: https://doi.org/10.1145/3292500.3330701.

[256] James Bergstra, Brent Komer, Chris Eliasmith, Dan Yamins, and David D Cox. Hyperopt: a python library for model selection and hyperparameter

optimization. *Computational Science & Discovery*, 8(1):014008, 2015. doi: https://doi.org/10.1088/1749-4699/8/1/014008.

[257] Marius Lindauer, Katharina Eggensperger, Matthias Feurer, André Biedenkapp, Difan Deng, Carolin Benjamins, Tim Ruhkopf, René Sass, and Frank Hutter. Smac3: A versatile bayesian optimization package for hyperparameter optimization. *Journal of Machine Learning Research*, 23 (54):1–9, 2022. URL `http://jmlr.org/papers/v23/21-0888.html`.

[258] Ngo Anh Vien, Heiko Zimmermann, and Marc Toussaint. Bayesian functional optimization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018. doi: https://doi.org/10.1609/aaai.v32i1.11830.

[259] Frank Hutter, Holger H. Hoos, and Kevin Leyton-Brown. Sequential model-based optimization for general algorithm configuration. In *Learning and Intelligent Optimization (LION)*, volume 6683 of *Lecture Notes in Computer Science*, pages 507–523. Springer Berlin Heidelberg, 2011. doi: https://doi.org/10.1007/978-3-642-25566-3_40.

[260] Q Liang, AE Gongora, Z Ren, A Tiihonen, Z Liu, S Sun, JR Deneault, D Bash, F Mekki-Berrada, SA Khan, et al. Benchmarking the performance of bayesian optimization across multiple experimental materials science domains. *npj Computational Materials*, 7(1):188, 2021. doi: https://doi.org/10.1038/s41524-021-00656-9.

[261] Shuhei Watanabe. Tree-structured parzen estimator: Understanding its algorithm components and their roles for better empirical performance. *arXiv preprint arXiv:2304.11127*, 2023.

[262] Niranjan Srinivas, Andreas Krause, Sham M Kakade, and Matthias W Seeger. Gaussian process optimization in the bandit setting: No regret and experimental design. In *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, pages 1015–1022, 2010.

[263] J. Tsao and Minnie H. Patel. An intuitive design pattern for sequentially estimating parameters of a 2k factorial experiment with active confounding avoidance and least treatment combinations. *Comput. Ind. Eng.*, 66:601–613, 2013. doi: https://doi.org/10.1016/j.cie.2013.08.005.

[264] Alejandro Barredo Arrieta, Natalia Díaz Rodríguez, J. Ser, Adrien Bennetot, S. Tabik, A. Barbado, S. García, S. Gil-Lopez, D. Molina, Richard Benjamins, Raja Chatila, and Francisco Herrera. Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. *Inf. Fusion*, 58:82–115, 2019. doi: https://doi.org/10.1016/j.inffus.2019.12.012.

[265] Alex Goldstein, Adam Kapelner, Justin Bleich, and Emil Pitkin. Peeking inside the black box: Visualizing statistical learning with plots of individual conditional expectation. *Journal of Computational and Graphical Statistics*, 24(1):44–65, 2015. doi: https://doi.org/10.1080/10618600.2014.907095.

[266] Scott M. Lundberg and Su-In Lee. A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, 30:4765–4774, 2017. doi: https://doi.org/10.48550/arXiv.1705.07874.

[267] Peter J. Rousseeuw. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20(1):53–65, 1987. doi: https://doi.org/10.1016/0377-0427(87)90125-7.

[268] Josue Obregon, Jihoon Hong, and Jae-Yoon Jung. Rule-based explanations based on ensemble machine learning for detecting sink mark defects in the injection moulding process. *Journal of Manufacturing Systems*, 60:392–405, 2021. doi: https://doi.org/10.1016/j.jmsy.2021.07.001.

[269] Andrea Saltelli. Making best use of model evaluations to compute sensitivity indices. *Computer Physics Communications*, 145(2):280–297, 2002. doi: https://doi.org/10.1016/S0010-4655(02)00280-1.

[270] Mohamad Kaddoura, Guillaume Majeau-Bettez, Ben Amor, and Manuele Margni. Global sensitivity analysis reduces data collection efforts in lca: A comparison between two additive manufacturing technologies. *Science of the Total Environment*, 975:179269, 2025. doi: https://doi.org/10.1016/j.scitotenv.2025.179269.

[271] Sandra Saad, Alankar Sinha, Camilo Cruz, Gilles Régnier, and Amine Ammar. Towards an accurate pressure estimation in injection molding simulation using surrogate modeling. *International Journal of Material Forming*, 15(6):72, 2022. doi: https://doi.org/10.1007/s12289-022-01717-0.

[272] Chengcheng Liu, Xuandong Wang, W. Cai, J. Yang, and H. Su. Prediction of the fatigue strength of steel based on interpretable machine learning. *Materials*, 16(23):7354, 2023. doi: https://doi.org/10.3390/ma16237354.

[273] Junsang Gim and Lih-Sheng Turng. Interpretation of the effect of transient process data on part quality of injection molding based on explainable artificial intelligence. *International Journal of Production Research*, 61(23):8192–8212, 2023. doi: https://doi.org/10.1080/00207543.2023.2216310.

[274] K. Yamaguchi. Feature importance analysis in global manufacturing industry. *International Journal of Trade, Economics and Finance*, 13(2): 28–35, 2022. doi: https://doi.org/10.18178/ijtef.2022.13.2.719.

[275] Yu Liu, Zhengchao Zhang, Shicao Jiang, and Yunfei Ding. Application of big data technology combined with clustering algorithm in manufacturing production analysis system. *Decision Making: Applications in Management and Engineering*, 7(1), 2024. doi: https://doi.org/10.31181/dmame712024897.

[276] Golshid Ranjbaran, D. Recupero, C. Roy, and Kevin A. Schneider. C-shap: A hybrid method for fast and efficient interpretability. *Applied Sciences*, 2025. doi: 10.3390/app15020672.

[277] Eloi Escriva, Thibault Lefrere, Mehdi Martin, Juliette Aligon, Amanda Chanson, Jean-Baptiste Excoffier, Nicolas Labroche, Chantal Soulé-Dupuy, and Patrick Monsarrat. Effective data exploration through clustering of local attributive explanations. *Information Systems*, 127:102464, 2025. doi: https://doi.org/10.1016/j.is.2024.102464.

[278] John A Hartigan and Manchek A Wong. Algorithm as 136: A k-means clustering algorithm. *Journal of the royal statistical society. series c (applied statistics)*, 28(1):100–108, 1979. doi: https://doi.org/10.2307/2346830.

[279] Yassmin Seid Ahmed and Abbas S. Milani. A review of strategies, challenges, and ethical implications of machine learning in smart manufacturing. *Decision Analytics Journal*, 16:100591, 2025. doi: https://doi.org/10.1016/j.dajour.2025.100591.

[280] Marcel R Ackermann, Johannes Blömer, Daniel Kuntze, and Christian Sohler. Analysis of agglomerative clustering. *Algorithmica*, 69(1):184–215, 2014. doi: https://doi.org/10.1007/s00453-012-9717-4.

[281] Abdelmoula El-Hamdouchi and Peter Willett. Comparison of hierarchic agglomerative clustering methods for document retrieval. *The Computer Journal*, 32(3):220–227, 1989. doi: https://doi.org/10.1108/eb026764.

[282] Baris Ördek, Eric Coatanéa, and Yuri Borgianni. An auto hierarchical clustering algorithm to distinguish geometries suitable for additive and traditional manufacturing technologies: Comparing humans and unsupervised learning. *Results in Engineering*, 25:104418, 2025. doi: https://doi.org/10.1016/j.rineng.2025.104418.

[283] András Darányi, Tímea Czvetkó, Alex Kummer, Tamás Ruppert, and János Abonyi. Multi-objective hierarchical clustering for tool assignment. *CIRP Journal of Manufacturing Science and Technology*, 42:47–54, 2023. doi: https://doi.org/10.1016/j.cirpj.2023.02.002.

[284] Miin-Shen Yang, Chien-Yo Lai, and Chih-Ying Lin. A robust em clustering algorithm for gaussian mixture models. *Pattern Recognition*, 45(11): 3950–3961, 2012. doi: https://doi.org/10.1016/j.patcog.2012.04.031.

[285] Debasish Mishra, Krishna R. Pattipati, and George M. Bollas. Gaussian mixture model for tool condition monitoring. *Journal of Manufacturing Processes*, 131:1001–1013, 2024. doi: https://doi.org/10.1016/j.jmapro.2024.09.038.

[286] Robert E. Kass and Adrian E. Raftery. Bayes factors. *Journal of the American Statistical Association*, 90(430):773–795, 1995. doi: https://doi.org/10.1080/01621459.1995.10476572.

[287] Pranava Madhyastha and Rishabh Jain. On model stability as a function of random seed. In *Proc. of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 929–939, Hong Kong, China, 2019. Association for Computational Linguistics. doi: https://doi.org/10.18653/v1/K19-1087.

[288] Florian Wenzel, Jasper Snoek, Dustin Tran, and Rodolphe Jenatton. Hyperparameter ensembles for robustness and uncertainty quantification. In *Advances in Neural Information Processing Systems, Vol.33 (NeurIPS 2020)*, 2020. doi: https://doi.org/10.48550/ARXIV.2006.13570.

[289] Sung-Min Kwak and Jae-Uk Kim. Statistical data preparation: management of missing values and outliers. *Korean J. Anesthesiology*, 70(4): 407–411, 2017. doi: https://doi.org/10.4097/kjae.2017.70.4.407.

[290] Mackenzie Hurst, Meghan O'Neill, Lief Pagalan, Lori M Diemert, and Laura C Rosella. The impact of different imputation methods on estimates and model performance: an example using a risk prediction model for premature mortality. *Population Health Metrics*, 22(13), 2024. doi: https://doi.org/10.1186/s12963-024-00331-3.

[291] Satyapriya Krishna, Tessa Han, Alex Gu, Steven Wu, Shahin Jabbari, and Himabindu Lakkaraju. The disagreement problem in explainable machine learning: A practitioner's perspective. *Trans. Machine Learning Research (TMLR)*, 2023. doi: https://doi.org/10.48550/ARXIV.2202.01602. arXiv:2202.01602.

[292] Trevor Hastie, Robert Tibshirani, Jerome Friedman, et al. *The Elements of Statistical Learning*. Springer Series in Statistics, 2001. doi: https://doi.org/10.1007/978-0-387-84858-7.

[293] Carliss Y. Baldwin and Kim B. Clark. *Design Rules, Volume 1: The Power of Modularity*. MIT Press, Cambridge, MA, 2000. doi: https://doi.org/10.7551/mitpress/2366.001.0001.

[294] Douglas C. Montgomery. *Design and Analysis of Experiments*. John Wiley & Sons, Hoboken, NJ, 9 edition, 2012. ISBN 978-1118146927.

[295] Verein Deutscher Ingenieure (VDI). Entwicklungsmethodik für technische produkte und systeme - modell der produktentwicklung, 2019. URL `https://www.vdi.de/richtlinien/details/vdi-2221-blatt-1-entwicklung-technischer-produkte-und-systeme-modell-der-produktentwicklung`. Last access: 07.08.2025.

[296] Defense Acquisition University. Glossary of defense acquisition acronyms and terms, 2024. URL `https://www.dau.edu/glossary`. Continuously updated online resource. Last access: 07.08.2025.

[297] Juliette Heraud, Khaled Medini, and Ann-Louise Andersen. Managing agile ramp-up projects in manufacturing–status quo and recommendations. *CIRP Journal of Manufacturing Science and Technology*, 45:125–137, 2023. doi: https://doi.org/10.1016/j.cirpj.2023.06.002.

[298] Isabelle Charlotte Ays. *Development of a CO2e quantification method and of solutions for reducing the greenhouse gas emissions of construction machines*. PhD thesis, Karlsruhe Insitute of Technology, 2020.

[299] Eduardo Calvo Buendia, Kiyoto Tanabe, Andreja Kranjc, Jamsranjav Baasansuren, Masato Fukuda, Sibusisiwe Ngarize, Akio Osako, Yaryna Pyrozhenko, Pavel Shermanau, and Sandro Federici. 2019 refinement to the 2006 IPCC guidelines for national greenhouse gas inventories. Technical report, Intergovernmental Panel on Climate Change, Geneva, Switzerland, 2019. URL `https://www.ipcc.ch/publication/2019-refinement-to-the-2006-ipcc-guidelines-for-national-greenhouse-gas-inventories/`. Last access: 08.08.2025.

[300] Julian M Allwood, Jonathan M Cullen, Mark A Carruth, Daniel R Cooper, Martin McBrien, Rachel L Milford, Muiris C Moynihan, and Alexandra CH Patel. *Sustainable materials: with both eyes open*, volume 2012. UIT Cambridge Limited Cambridge, UK, 2012. URL `https://www.uselessgroup.org/publications/book/chapters`. Last access: 07.08.2025.

[301] Matthias Finkbeiner. Product environmental footprint—ready for the european market? *The International Journal of Life Cycle Assessment*, 19 (3):369–374, 2014. doi: https://doi.org/10.1007/s11367-013-0678-x.

[302] P. Frazier. A tutorial on bayesian optimization. *ArXiv*, abs/1807.02811, 2018. doi: https://doi.org/10.48550/arXiv.1807.02811.

[303] G. Ø. Rønsch, M. Dybdahl, and M. Kulahci. Real-time adjustment of injection molding process settings by utilizing design of experiment, time series profiles and pls-da. *Quality Engineering*, 34(2):215–229, 2022. doi: https://doi.org/10.1080/08982112.2022.2033775.

[304] Jeremy Sadet, Franck Massa, Thierry Tison, El-Ghazali Talbi, and Isabelle Turpin. Deep gaussian process for the approximation of a quadratic eigenvalue problem: Application to friction-induced vibration. *Vibration*, 5(2):344–369, 2022. doi: https://doi.org/10.3390/vibration5020020.

[305] Edmund D. Weinberger. Correlated and uncorrelated fitness landscapes and how to tell the difference. *Biological Cybernetics*, 63(5):325–336, 1990. doi: https://doi.org/10.1007/BF00202749.

[306] Sebastian Muñoz-Herrera and Karol Suchan. Constrained fitness landscape analysis of capacitated vehicle routing problems. *Entropy*, 24(1):53, 2022. doi: https://doi.org/10.3390/e24010053.

[307] Monte Lunacek and Darrell Whitley. The dispersion metric and the cma evolution strategy. In *Proceedings of the 2006 Genetic and Evolutionary Computation Conference (GECCO)*, pages 651–658, 2006. doi: https://doi.org/10.1145/1143997.1144085.

[308] Arvin Jaymes Domingo, Joselito Gabriell Gallardo, Jan Ulrich Nacino, Jonas Sevilla, Patrick Louie Tresmaria, and Jojo Blanza. Outlier detection and clustering of imt-2020 dataset. In *Journal of Computer Science & Computational Mathematics*, volume 14, pages 7–11, 2024. doi: https://doi.org/10.20967/jcscm.2024.02.001.

[309] Karl Pearson. Mathematical contributions to the theory of evolution. iii. regression, heredity, and panmixia. *Philosophical Transactions of the Royal Society of London A*, 187:253–318, 1896. doi: 10.1098/rsta.1896.0007.

[310] Maurice G. Kendall. A new measure of rank correlation. *Biometrika*, 30:81–93, 1938. doi: 10.1093/biomet/30.1-2.81.

[311] Maurice G. Kendall. The treatment of ties in ranking problems. *Biometrika*, 33(3):239–251, 1945. doi: 10.2307/2332303.

[312] Hermann Baumgartl, Josef Tomas, Ricardo Buettner, and Markus Merkel. A deep learning-based model for defect detection in laser-powder bed fusion using in-situ thermographic monitoring. *Progress in Additive Manufacturing*, 5(3):277–285, 2020. doi: https://doi.org/10.1007/s40964-019-00108-3.

[313] Ahmad Aminzadeh, Sasan Sattarpanah Karganroudi, Soheil Majidi, Colin Dabompré, Khalil Azaiez, Christopher Mitride, and Eric Sénéchal. A machine learning implementation to predictive maintenance and monitoring of industrial compressors. *Sensors*, 25(4):1006, 2025. doi: https://doi.org/10.3390/s25041006.

[314] McKinsey & Company, Mozilla Foundation, and Patrick J. McGovern Foundation. Open source technology in the age of ai, April 2025. URL https://www.mckinsey.com/~/media/mckinsey/business%20functions/quantumblack/our%20insights/open%20source%20technology%20in%20the%20age%20of%20ai/open-source-technology-in-the-age-of-ai_final.pdf. Last access: 10.06.2025.

[315] Akshay Antony, Chakradhar Guntuboina, Rishikesh Magar, Lalit Ghule, Ruchit Doshi, Khalid Aman, Sharan Seshadri, and Amir Barati Farimani. Manufacturingnet: A machine learning toolbox for engineers. *SoftwareX*, 23:101478, 2023. doi: https://doi.org/10.1016/j.softx.2023.101478.

[316] Pengcheng Xu, Xiaobo Ji, Minjie Li, and Wencong Lu. Small data machine learning in materials science. *npj Computational Materials*, 9:42, 2023. doi: https://doi.org/10.1038/s41524-023-01000-z.

[317] Zewei Zhang, Guillaume Bellegarda, Milad Shafiee, and Auke Ijspeert. Online optimization of central pattern generators for quadruped locomotion. In *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 13547–13554. IEEE, 2024. doi: https://doi.org/10.1109/IROS58592.2024.10802762.

[318] George A. Miller. The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review*, 63(2):81–97, 1956. doi: https://doi.org/10.1037/h0043158.

[319] Todd Kulesza, Simone Stumpf, Margaret Burnett, Sherry Yang, Irwin Kwan, and Weng-Keen Wong. Too much, too little, or just right? ways explanations impact end users' mental models. In *2013 IEEE Symposium on Visual Languages and Human-Centric Computing (VL/HCC)*, pages 3–10, 2013. doi: https://doi.org/10.1109/VLHCC.2013.6645235.

[320] Ilaria G. Barletta, Melanie Despeisse, and Björn Johansson. The proposal of an environmental break-even point as assessment method of product-service systems for circular economy. *Procedia CIRP*, 72:720–725, 2018. doi: https://doi.org/10.1016/j.procir.2018.03.257.

[321] Charles T. Horngren, Srikant M. Datar, and Madhav V. Rajan. *Cost Accounting: A Managerial Emphasis*. Pearson, 14th edition, 2013. ISBN 9781292363073.

[322] Mark A. J. Huijbregts, Uldis Thissen, Simon Ross, and Joseph B. Guinee. Application of uncertainty and variability in life cycle assessment. part i: A general framework for the analysis of uncertainty and variability in life cycle assessment. *International Journal of Life Cycle Assessment*, 3(5): 273–280, 1998. doi: https://doi.org/10.1007/BF02978772.

[323] Stephen B Vardeman et al. *Statistical methods for quality assurance*. Springer, 2016. doi: https://doi.org/10.1007/978-0-387-79106-7.

[324] Eva JI Hoeijmakers, Bibi Martens, Babs MF Hendriks, Casper Mihl, Razvan L Miclea, Walter H Backes, Joachim E Wildberger, Frank M Zijta, Hester A Gietema, Patricia J Nelemans, et al. How subjective ct image quality assessment becomes surprisingly reliable: pairwise comparisons instead of likert scale. *European Radiology*, 34(7):4494–4503, 2024. doi: https://doi.org/10.1007/s00330-023-10493-7.

[325] Lars St, Svante Wold, et al. Analysis of variance (anova). *Chemometrics and intelligent laboratory systems*, 6(4):259–272, 1989. doi: https://doi.org/10.1016/0169-7439(89)80095-4.

[326] William H. Kruskal and W. Allen Wallis. Use of ranks in one-criterion variance analysis. *Journal of the American Statistical Association*, 47(260): 583–621, 1952. doi: 10.1080/01621459.1952.10483441.

[327] Flore Mekki-Berrada, Zekun Ren, Tan Huang, Wai Kuan Wong, Fang Zheng, Jiaxun Xie, Isaac Parker Siyu Tian, Senthilnath Jayavelu, Zackaria Mahfoud, Daniil Bash, et al. Two-step machine learning enables optimized nanoparticle synthesis. *npj Computational Materials*, 7(1):55, 2021. doi: https://doi.org/10.1038/s41524-021-00520-w.

[328] James R Deneault, Jorge Chang, Jay Myung, Daylond Hooper, Andrew Armstrong, Mark Pitt, and Benji Maruyama. Toward autonomous additive

manufacturing: Bayesian optimization on a 3d printer. *MRS Bulletin*, 46: 566–575, 2021. doi: https://doi.org/10.1557/s43577-021-00051-1.

[329] Aldair E Gongora, Bowen Xu, Wyatt Perry, Chika Okoye, Patrick Riley, Kristofer G Reyes, Elise F Morgan, and Keith A Brown. A bayesian experimental autonomous researcher for mechanical design. *Science advances*, 6(15):eaaz1708, 2020. doi: https://doi.org/10.1126/sciadv.aaz1708.

[330] Johannes M Dieterich and Bernd Hartke. Empirical review of standard benchmark functions using evolutionary global optimization. *Applied Mathematics*, 3:1552–1564, 2012. doi: https://doi.org/10.4236/am.2012.330215.

[331] Shijing Sun, Armi Tiihonen, Felipe Oviedo, Zhe Liu, Janak Thapa, Yicheng Zhao, Noor Titan P Hartono, Anuj Goyal, Thomas Heumueller, Clio Batali, et al. A data fusion approach to optimize compositional stability of halide perovskites. *Matter*, 4(4):1305–1322, 2021. doi: https://doi.org/10.1016/j.matt.2021.01.008.

[332] Yijie Yang, Xiaoyang Zhang, and Dan Wang. Embodied carbon accounting through spatial-temporal embodied carbon models. *arXiv preprint arXiv:2312.06364*, 2023. doi: https://doi.org/10.48550/arXiv.2312.06364.

[333] Petra Icha and Thomas Lauf. Entwicklung der spezifischen treibhausgasemissionen des deutschen strommix in den jahren 1990–2023. Climate Change Report 23/2024, Umweltbundesamt, Dessau-Roßlau, June 2024. URL `https://www.umweltbundesamt.de/sites/default/files/medien/11850/publikationen/23_2024_cc_strommix_11_2024.pdf`. Last access: 08.08.2025.

[334] Ecoinvent Association. Ecoinvent database v3.11, 2023. URL `https://www.ecoinvent.org/`. Last access: 01.05.2025.

[335] Asma Mecheter and Faris Tarlochan. Fused filament fabrication three-dimensional printing: Assessing the influence of geometric complexity and process parameters on energy and the environment. *Sustainability*, 15(16): 12319, 2023. doi: https://doi.org/10.3390/su151612319.

[336] Ge Gao, Fan Xu, Jiangmin Xu, Guanghai Tang, and Zhenyu Liu. A survey of the influence of process parameters on mechanical properties of fused deposition modeling parts. *Micromachines*, 13(4):553, 2022. ISSN 2072-666X. doi: https://doi.org/10.3390/mi13040553.

[337] Moe Elbadawi, Abdul W Basit, and Simon Gaisford. Energy consumption and carbon footprint of 3d printing in pharmaceutical manufacture. *International Journal of Pharmaceutics*, 639:122926, 2023. doi: https://doi.org/10.1016/j.ijpharm.2023.122926.

[338] Eurostat. Electricity price statistics: Electricity prices for non-household consumers. Eurostat database, 2024. URL `https://ec.europa.eu/eurostat/databrowser/view/nrg_pc_205/default/table`. Last access: 08.08.2025.

[339] Vladislav Andronov, Libor Beránek, Vojtěch Krůta, Lucie Hlavůňková, and Zdeňka Jeníková. Overview and comparison of pla filaments commercially available in europe for fff technology. *Polymers*, 15(14):3065, 2023. doi: https://doi.org/10.3390/polym15143065.

[340] Emmanuel U Enemuoh, Stefan Duginski, Connor Feyen, and Venkata G Menta. Effect of process parameters on energy consumption, physical, and mechanical properties of fused deposition modeling. *Polymers*, 13(15): 2406, 2021. doi: https://doi.org/10.3390/polym13152406.