Autonomous Multi-Objective Design Optimization using Bayesian Approaches and Active learning

by

Alexandros Ntagiantas

MTN 2410

Submitted
in partial fulfilment of the requirements for the degree of

Master of Artificial Intelligence

at the

UNIVERSITY OF PIRAEUS

April 2026

II-MSc “Artificial Intelligence”

April 20, 2026

Certified by

Prof. Christoforos Rekatsinas

Supervisor

Certified by

Panagiotis Krokidas

Research Associate

Member of the Examination Committee

Certified by

Prof. Georgios Vouros

Member of the Examination Committee

Abstract

The discovery and design of advanced functional materials often involve extremely large combinatorial design spaces, making exhaustive exploration through traditional computational or experimental approaches impractical. In recent years, machine learning and data-driven optimization techniques have emerged as promising tools for accelerating materials discovery by enabling efficient exploration of complex design spaces. Among these approaches, Bayesian Optimization (BO) has gained significant attention due to its ability to optimize expensive black-box functions using a limited number of evaluations.

However, the performance of BO can deteriorate when applied to very large search spaces, as the algorithm must allocate evaluations across many potentially unpromising regions. To address this challenge, this thesis investigates the integration of Active Learning–based space reduction with Bayesian optimization for efficient materials design.

Specifically, the proposed framework first applies Bayesian Optimization to explore the initial design space and gather informative samples. Subsequently, a classification-based Active Learning method (DAGS) is used to identify and filter low-performing regions of the design space, effectively reducing the search domain. Bayesian Optimization is then re-applied within the reduced design space in order to improve optimization efficiency. The proposed pipeline is evaluated in both single-objective and multi-objective optimization settings.

Experimental results demonstrate that the proposed space-reduction strategy can significantly improve optimization efficiency by focusing the search on high-potential regions of the design space while maintaining competitive solution quality. The findings highlight the potential of combining Active Learning and Bayesian Optimization as a unified framework for data-efficient exploration of large materials design spaces, particularly in the context of nanoporous materials such as Covalent Organic Frameworks (COFs).

Περίληψη

Η ανακάλυψη και ο σχεδιασμός σύγχρονων λειτουργικών υλικών αποτελούν ένα ιδιαίτερα απαιτητικό πρόβλημα, καθώς ο χώρος πιθανών συνδυασμών δομικών και χημικών χαρακτηριστικών είναι εξαιρετικά μεγάλος. Η εξαντλητική διερεύνηση αυτού του χώρου μέσω παραδοσιακών υπολογιστικών ή πειραματικών προσεγγίσεων καθίσταται στην πράξη ανέφικτη, λόγω του υψηλού υπολογιστικού και χρονικού κόστους.

Τα τελευταία χρόνια, η αξιοποίηση μεθόδων μηχανικής μάθησης και βελτιστοποίησης έχει συμβάλει σημαντικά στην επιτάχυνση της διαδικασίας ανακάλυψης υλικών. Οι προσεγγίσεις αυτές επιτρέπουν μια πιο στοχευμένη και αποδοτική εξερεύνηση πολύπλοκων χώρων σχεδιασμού. Σε αυτό το πλαίσιο, η Bayesian Optimization (BO) έχει αναδειχθεί ως μία από τις πλέον κατάλληλες τεχνικές για προβλήματα όπου η αξιολόγηση των υποψήφιων λύσεων είναι δαπανηρή, καθώς μπορεί να εντοπίσει υψηλής ποιότητας λύσεις με περιορισμένο αριθμό αξιολογήσεων.

Παρά τα πλεονεκτήματά της, η απόδοση της BO επηρεάζεται αρνητικά όταν εφαρμόζεται σε πολύ μεγάλους χώρους αναζήτησης, καθώς σημαντικό μέρος των διαθέσιμων αξιολογήσεων κατανέμεται σε περιοχές χαμηλού ενδιαφέροντος. Για την αντιμετώπιση του προβλήματος αυτού, η παρούσα εργασία εξετάζει τον συνδυασμό της Bayesian Optimization με τεχνικές Active Learning, με στόχο τη μείωση του χώρου αναζήτησης και τη βελτίωση της αποδοτικότητας της διαδικασίας βελτιστοποίησης.

Συγκεκριμένα, προτείνεται ένα πλαίσιο στο οποίο η Bayesian Optimization εφαρμόζεται αρχικά στον πλήρη χώρο σχεδιασμού, προκειμένου να συλλεχθούν αντιπροσωπευτικά και πληροφοριακά δείγματα. Στη συνέχεια, αξιοποιείται μια μέθοδος Active Learning βασισμένη σε ταξινόμηση (DAGS), η οποία επιτρέπει τον διαχωρισμό του χώρου σε περιοχές υψηλού και χαμηλού δυναμικού. Μέσω αυτής της διαδικασίας, απομακρύνονται υποψήφιες λύσεις με χαμηλή απόδοση και προκύπτει ένας σημαντικά μικρότερος και πιο εστιασμένος χώρος αναζήτησης. Η Bayesian Optimization εφαρμόζεται εκ νέου στον μειωμένο αυτό χώρο, επιτρέποντας ταχύτερη σύγκλιση και αποδοτικότερη αξιοποίηση των διαθέσιμων αξιολογήσεων.

Η προτεινόμενη μεθοδολογία αξιολογείται τόσο σε μονοκριτηριακά όσο και σε πολυκριτηριακά προβλήματα βελτιστοποίησης. Τα αποτελέσματα δείχνουν ότι η μείωση του χώρου αναζήτησης μπορεί να οδηγήσει σε σημαντική βελτίωση της αποδοτικότητας, χωρίς ουσιαστική απώλεια στην ποιότητα των τελικών λύσεων. Επιπλέον, αναδεικνύεται ότι ο συνδυασμός Active Learning και Bayesian Optimization μπορεί να αποτελέσει ένα αποτελεσματικό και γενικεύσιμο πλαίσιο για την εξερεύνηση μεγάλων χώρων σχεδιασμού, ιδιαίτερα σε εφαρμογές που αφορούν νανοπορώδη υλικά, όπως τα Covalent Organic Frameworks (COFs).

Acknowledgments

I would like to express my sincere gratitude to my supervisor, Prof. Christoforos Rekatsinas, for his guidance, support, and valuable feedback throughout the course of this thesis. His insights and expertise were instrumental in shaping the direction and quality of this work.

I would also like to sincerely thank Research Associate Panagiotis Krokidas for his valuable support, insightful discussions, and contribution to the development of this work.

Finally, I would like to thank my family and friends for their continuous encouragement and support during my studies.

Table of Contents

1 INTRODUCTION ..... ERROR! BOOKMARK NOT DEFINED.

1.1 MOTIVATION 11

1.2 CHALLENGES IN MATERIALS DISCOVERY 13

1.3 MACHINE LEARNING FOR MATERIALS DESIGN 15

1.4 THESIS CONTRIBUTIONS 17

1.5 THESIS STRUCTURE 18

2 BACKGROUND ON MATERIAL DISCOVERY AND OPTIMIZATION 20

2.1 MATERIALS DISCOVERY PARADIGM 20

2.2 NANOPOROUS MATERIALS 21

2.3 COVALENT ORGANIC FRAMEWORKS (COFS) 22

2.4 APPLICATIONS OF COFS 23

2.5 CHALLENGES IN COF DESIGN 24

3 BAYESIAN OPTIMIZATION AND ACTIVE LEARNING 27

3.1 BAYESIAN OPTIMIZATION 27

3.2 SURROGATE MODELS 28

3.3 ACQUISITION FUNCTIONS 30

3.4 MULTI-OBJECTIVE BAYESIAN OPTIMIZATION 31

3.5 ACTIVE LEARNING 33

3.6 ACTIVE LEARNING IN MATERIALS DISCOVERY 34

4 PROPOSED METHODOLOGY 36

4.1 PROBLEM FORMULATION 36

4.2 BASELINE BAYESIAN OPTIMIZATION 37

4.3 DESIGN SPACE REDUCTION VIA ACTIVE LEARNING 38

4.4 PROPOSED OPTIMIZATION PIPELINE 40

4.5 MULTI-OBJECTIVE EXTENSION 42

5 EXPERIMENTAL SETUP 44

5.1 DATASETS 44
5.2 PREPROCESSING 45
5.3 BAYESIAN OPTIMIZATION IMPLEMENTATION 46
5.4 DAGS IMPLEMENTATION 47
5.5 EVALUATION METRICS 49
5.6 INITIALIZATION STRATEGY 51
6 RESULTS 52
6.1 SINGLE-OBJECTIVE RESULTS 52
6.2 MULTI-OBJECTIVE RESULTS 57
6.3 COMPARATIVE ANALYSIS 62
7 DISCUSSION 64

Table of Figures

FIGURE 1: PROPOSED OPTIMIZATION PIPELINE 41
FIGURE 2: CONVERGENCE COMPARISON 53
FIGURE 3: BEST VALUE FOUND 54
FIGURE 4:TOP-1 FOUND 55
FIGURE 5:TOP-10 FOUND 55
FIGURE 6:TOP-100 FOUND 56
FIGURE 7: COMPARISON OF BAYESIAN OPTIMIZATION PERFORMANCE IN THE FULL AND REDUCED DESIGN SPACES. 57
FIGURE 8: HYPERVOLUME CONVERGENCE 58
FIGURE 9: PARETO HITS VS TOTAL EVALUATIONS 59
FIGURE 10: 3D PARETO POINTS FOUND 60
FIGURE 11: PARETO RETENTION AFTER DAGS 61
FIGURE 12: FINAL COMPARISON SUMMARY 62

1. Introduction

This chapter introduces the fundamental motivation and context of the present work. It outlines the main challenges in materials discovery, discusses the role of machine learning in addressing these challenges, and summarizes the key contributions and structure of the thesis.

One of the most important and widely developed directions in contemporary science and technology is the discovery and design of new materials, with applications ranging from energy storage and gas capture to catalysis and environmental technologies. Particular attention has been given to porous materials such as Covalent Organic Frameworks (COFs), which have attracted significant research interest due to both their physicochemical properties and their structural tunability, allowing them to be tailored to specific applications [16,21–23]. This structural flexibility arises from the reticular synthesis of organic building blocks into crystalline frameworks, enabling systematic control over pore size, topology and chemical functionality. As a result, COFs have been widely investigated for applications including gas storage and separation, catalysis, sensing and energy-related technologies.

However, a major challenge in their study arises from the enormous number of possible combinations of structural parameters and chemical components, which leads to an exponential growth of the design space. This makes the evaluation of all candidate materials extremely difficult and, in practice, often infeasible — a well-recognized issue in computational materials discovery [12,14,24]. Traditional high-throughput screening and computational workflows can partially alleviate this problem but still struggle with the combinatorial explosion of candidate structures without additional optimization strategies [24,25].

Traditional approaches rely either on large-scale computational simulations or on experimental trial-and-error procedures, both of which are time-consuming and computationally expensive. In recent years, and especially with the rapid development of machine learning methods, it has become possible to accelerate the materials discovery process through property prediction, optimization of existing materials, and more

targeted exploration of design spaces [2,11,12,24–26]. Data-driven and knowledge-driven frameworks that integrate uncertainty-aware learning, optimization and experimental design are increasingly viewed as key enablers of autonomous materials discovery. These approaches form part of a broader shift toward autonomous materials discovery platforms, where machine learning, optimization algorithms and high-throughput computations are integrated into closed-loop design cycles aimed at accelerating scientific discovery [24,26].

In this context, Bayesian Optimization (BO) has emerged as a particularly effective strategy for optimization problems where the evaluation of candidate solutions is expensive or limited [1,6,16]. BO combines probabilistic surrogate modeling with sequential decision strategies to balance exploration of uncertain regions and exploitation of promising solutions. BO has been successfully applied across multiple materials science domains, including nanoporous materials, alloys, electronic and functional materials, and chemical reaction optimization [14,17,18,27–31]. Despite its encouraging success, Bayesian Optimization faces important challenges when applied to realistic materials design problems. The large volume of data, the complexity of material systems, and the need to simultaneously optimize multiple, often conflicting objectives hinder efficient exploration of the search space [12,14,24]. Open algorithmic issues include high-dimensional and mixed variable spaces, multi-fidelity data and robust surrogate modeling, which are particularly pronounced in materials applications [14,24,29,30].

For example, properties such as adsorption capacity and selectivity often exhibit conflicting behavior. In gas separation applications, maximizing the adsorption capacity of a specific species (e.g., species A) does not necessarily lead to high selectivity of that species over others (e.g., species B), resulting in inherent trade-offs in material performance. Similarly, improvements in adsorption properties may come at the expense of structural stability or synthesis feasibility. This makes it necessary to adopt Multi-Objective Bayesian Optimization (MO-BO) frameworks [2,8–10]. In such settings, the goal is not a single optimal solution but a set of Pareto-optimal solutions that represent different trade-offs among the selected objectives. Such trade-offs frequently arise in

materials systems where improving one property may significantly degrade another, making simultaneous optimization particularly challenging.

In materials science, MO-BO has been used to discover alloys, shape-memory materials, and electronic compounds with concurrent optimization of mechanical, functional, and stability criteria [2,8,18,32,33]. Nevertheless, MO-BO is computationally more demanding, as it requires sufficient coverage of the solution space and reliable uncertainty estimation across multiple objective dimensions [2,14,18,32]. Recent studies highlight that hypervolume-based acquisition functions such as qEHVI can significantly improve the quality of the learned Pareto front under tight evaluation budgets [18,32,33].

At the same time, the increasing availability of large materials databases creates the need for methods that can selectively focus on the most informative regions of the space. In this direction, Active Learning provides a strategy for guiding data selection based on model uncertainty or expected information gain [4,5,24,26]. This enables a reduction of the effective search space and helps avoid unnecessary evaluations in low-interest regions, thereby reducing computational cost. Active learning has already shown promising results in the discovery of COFs and other nanoporous materials for gas storage and separation applications, as well as in multi-objective alloy and manufacturing optimization [7–9,19,24,32,34].

The present work focuses on the development of an efficient framework for design space exploration in materials discovery by combining Bayesian Optimization, Multi-Objective Optimization, Active Learning, and systematic search-space reduction techniques. A primary application of the proposed methodology is the discovery of high-performing COFs, while its multi-objective extension is also evaluated on a separate composite material design problem. The central idea is that optimization efficiency can be significantly improved when the initial design space is restricted to high-potential subregions before applying BO, an approach increasingly explored in both nanoporous materials design and broader materials optimization settings [15,20,31,35].

More specifically, this work investigates how Active Learning techniques can be used for space reduction, filtering large-scale materials spaces while retaining regions likely to

contain high-quality solutions. This perspective differs from conventional BO workflows that assume a fixed design domain and instead treats the search space as an object that can be refined prior to optimization. Bayesian Optimization is then applied in both single-objective and multi-objective settings, using Gaussian Process surrogate models to guide the search in a data-efficient manner, while scalable machine learning models such as XGBoost are employed within the Active Learning stage for classification-based space reduction [6,12,17,18,20,27]. In line with recent work on adaptive and uncertainty-aware optimization for high-dimensional materials problems, the framework emphasizes acquisition strategies suited to complex, multi-criteria design spaces [24,28–30].

The main contribution of this work lies in the integration of design space reduction, Active Learning, and Bayesian Optimization into a unified pipeline for materials design, together with its extension to multi-objective optimization settings. In addition to evaluating the effect of space reduction on single-objective optimization performance, the study also investigates its impact on the recovery of high-quality Pareto-optimal solutions in a multi-objective setting. In this way, the thesis empirically examines the extent to which the search space can be significantly reduced without substantial loss of solution quality, contributing to the broader effort toward autonomous and data-driven materials discovery across different design domains [20,24,26,31,35].

1.1 Motivation

The rapid expansion of materials design spaces, particularly in systems such as Covalent Organic Frameworks (COFs), has fundamentally altered the nature of materials discovery problems. Instead of being limited by data scarcity, modern research is increasingly constrained by the sheer scale and complexity of candidate materials. This shift has

transformed materials discovery into a problem of efficient search and decision-making under limited computational and experimental resources.

In this context, the central motivation of this work arises from the need to develop methods that can navigate extremely large and high-dimensional design spaces while minimizing the number of expensive evaluations. Traditional high-throughput approaches, although effective at exploring large datasets, often rely on exhaustive or semi-exhaustive screening strategies, which quickly become impractical as the size of the search space grows. Consequently, there is a growing demand for intelligent optimization frameworks that can identify promising regions of the design space early in the search process.

Bayesian Optimization offers a principled solution to this challenge by enabling data-efficient exploration through probabilistic modeling and uncertainty-aware decision-making. However, its performance is highly dependent on the structure and size of the search space. When applied directly to large-scale materials datasets, BO may suffer from slow convergence and increased computational overhead, particularly in high-dimensional or heterogeneous domains.

At the same time, recent advances in Active Learning suggest that it is possible to significantly reduce the effective size of the search space by selectively focusing on informative and high-potential samples. By filtering out regions that are unlikely to contain optimal solutions, Active Learning can improve both the efficiency and the stability of optimization algorithms. This idea is particularly appealing in materials discovery, where a large portion of the design space often consists of low-quality or irrelevant candidates.

The motivation of this thesis is therefore grounded in the hypothesis that combining search-space reduction with optimization can lead to substantial improvements in efficiency without compromising solution quality. Rather than treating the design space as fixed, this work explores the idea of dynamically refining the search domain prior to and during optimization. This perspective introduces an additional layer of decision-

making, where not only the next sample is selected, but also the region in which the search is performed.

Furthermore, many real-world materials design problems involve multiple competing objectives, such as maximizing performance while maintaining stability or minimizing cost. This introduces additional complexity, as the goal is no longer a single optimal solution but a set of trade-off solutions. The integration of multi-objective optimization within a reduced and adaptively refined design space constitutes a key aspect of this work.

Overall, the motivation of this thesis is to address the limitations of existing optimization approaches in large-scale materials discovery by proposing a unified framework that combines Bayesian Optimization, Active Learning and search-space reduction. The ultimate goal is to enable more efficient, scalable and autonomous exploration of complex materials design spaces, with a particular focus on nanoporous materials such as COFs.

1.2 Challenges in Materials Discovery

Despite the significant progress achieved in computational materials science and data-driven methodologies, materials discovery remains a fundamentally challenging problem. These challenges arise not only from the intrinsic complexity of material systems, but also from computational, methodological and practical limitations that hinder efficient exploration and optimization.

A primary challenge lies in the vastness of the design space. In systems such as Covalent Organic Frameworks (COFs), the number of possible combinations of building blocks, topologies and functional groups grows combinatorially. Even when restricting the search to physically meaningful configurations, the number of candidate materials can easily reach tens or hundreds of thousands. This renders exhaustive evaluation strategies infeasible, particularly when each evaluation requires expensive simulations or experimental validation.

In addition to scale, the relationship between material structure and properties is highly complex and often nonlinear. Material properties are influenced by multiple interacting factors, including geometric, chemical and electronic characteristics. As a result, small changes in structure can lead to disproportionate or unexpected changes in performance. This lack of simple, interpretable relationships makes it difficult to design materials through intuition alone and necessitates the use of advanced predictive models.

Another critical challenge is the high computational cost associated with evaluating candidate materials. Accurate estimation of properties such as adsorption capacity, stability or selectivity often requires quantum mechanical simulations or detailed molecular modeling. These processes are computationally intensive and limit the number of candidate materials that can be realistically explored. Consequently, efficient sampling and evaluation strategies become essential.

Furthermore, many materials discovery problems are inherently multi-objective. In practical applications, it is rarely sufficient to optimize a single property in isolation. Instead, multiple criteria must be considered simultaneously, such as maximizing performance while maintaining structural stability, manufacturability or cost-effectiveness. These objectives are often conflicting, leading to trade-offs that must be carefully balanced. This significantly increases the complexity of the optimization problem and requires specialized methods capable of identifying sets of optimal trade-off solutions.

Another important challenge concerns the presence of high-dimensional and heterogeneous feature spaces. Materials are often described by a combination of continuous, discrete and categorical variables, including geometric descriptors, chemical compositions and structural parameters. This heterogeneity complicates the application of traditional optimization techniques and places additional demands on surrogate modeling approaches.

Finally, the availability of large-scale materials datasets introduces both opportunities and challenges. While such datasets provide valuable information for training predictive models, they also require methods that can efficiently extract relevant knowledge without being overwhelmed by irrelevant or redundant data. In many cases, only a small subset of

the data contains high-quality or high-performing materials, making selective sampling strategies crucial.

Taken together, these challenges highlight the need for advanced computational frameworks that can efficiently navigate large, complex and multi-objective design spaces. Addressing these limitations requires the integration of machine learning, optimization and data selection strategies, forming the basis for the approaches explored in this work.

1.3 Machine Learning for Materials Design

The integration of machine learning into materials science has led to a paradigm shift in the way materials are designed, evaluated and optimized. Rather than relying solely on physics-based simulations or experimental trial-and-error processes, machine learning enables the development of predictive models that can approximate complex structure–property relationships and guide the exploration of large design spaces in a data-efficient manner.

At the core of this approach lies the ability of machine learning models to learn patterns from existing data and generalize to unseen materials. By training on datasets that describe materials in terms of structural, geometric and chemical features, these models can predict key properties such as adsorption capacity, stability or selectivity with significantly lower computational cost compared to traditional simulation methods. This capability allows for the rapid screening of large numbers of candidate materials, making machine learning an essential component of modern materials discovery pipelines.

A wide range of machine learning techniques has been applied in this domain, including regression models, kernel-based methods, tree-based ensembles and deep learning architectures. Among these, probabilistic models such as Gaussian Processes are particularly important due to their ability to provide uncertainty estimates alongside predictions. This uncertainty information plays a critical role in guiding optimization strategies, as it enables informed decisions about where to sample next in the design space.

In addition to predictive modeling, machine learning has enabled the development of optimization frameworks that actively guide the search for optimal materials. Instead of passively evaluating randomly generated candidates, these frameworks use learned models to iteratively propose new materials that are likely to improve upon previous results. This approach significantly reduces the number of required evaluations and allows for more efficient exploration of complex search spaces.

Furthermore, machine learning facilitates the integration of multiple sources of information, including experimental data, simulation outputs and domain knowledge. This flexibility allows for the construction of hybrid models that can capture different aspects of material behavior and improve prediction accuracy. In recent years, such approaches have been incorporated into autonomous materials discovery systems, where learning, optimization and evaluation are combined into closed-loop pipelines.

Despite its advantages, the application of machine learning in materials design is not without challenges. The quality of predictions depends heavily on the quality and representativeness of the training data, while model performance can degrade in high-dimensional or sparse regions of the design space. Additionally, many materials problems involve multiple competing objectives, requiring models and optimization strategies that can handle complex trade-offs.

Overall, machine learning provides the foundational tools for enabling data-driven and autonomous materials discovery. By combining predictive modeling with intelligent optimization strategies, it becomes possible to efficiently navigate large design spaces and identify high-performing materials. In this work, machine learning plays a central role in both the modeling and optimization components of the proposed framework, forming the basis for the integration of Bayesian Optimization and Active Learning techniques presented in the following sections.

1.4 Thesis Contributions

The main objective of this thesis is to investigate whether the efficiency of Bayesian Optimization in materials discovery can be improved through prior reduction of the design space using Active Learning. The study focuses on Covalent Organic Frameworks (COFs), where the large number of candidate materials makes exhaustive exploration impractical.

More specifically, the first contribution of this work is the establishment of a baseline optimization setting in which Bayesian Optimization is applied directly to the original design space. This serves as a reference point for evaluating the effectiveness of subsequent search-space reduction strategies and allows the performance of Bayesian Optimization to be assessed under the full complexity of the problem.

The second contribution is the application of an Active Learning methodology for design space reduction. The purpose of this step is to identify and retain the most informative or high-potential regions of the original dataset, while filtering out samples that are less likely to contribute to the discovery of optimal solutions. In this way, the initial search space is reduced before optimization is repeated.

The third contribution of the thesis is the re-application of Bayesian Optimization on the reduced design space produced by Active Learning. This enables a direct comparison between optimization in the original space and optimization in the reduced space, making it possible to evaluate whether space reduction improves efficiency while preserving the quality of the final solutions.

In addition, the thesis examines this process in both single-objective and multi-objective settings. In the multi-objective case, the study investigates whether reduced-space optimization can still recover high-quality trade-off solutions across competing objectives, and whether the resulting Pareto front remains competitive compared to optimization over the full search space.

Finally, the thesis provides a systematic experimental analysis of the overall workflow, comparing the performance of Bayesian Optimization before and after design space

reduction. Through this evaluation, the study explores the extent to which Active Learning can act as an effective preprocessing stage for large-scale optimization problems in materials discovery.

Overall, the contribution of this work lies in demonstrating a practical and structured methodology in which Bayesian Optimization is first applied to the original materials space, then combined with Active Learning-based reduction, and finally re-evaluated on the reduced space in order to assess the impact of this reduction on optimization performance, computational efficiency and solution quality.

Furthermore, the proposed methodology is evaluated on two distinct material design problems, including nanoporous materials (COFs) and advanced composite materials. This allows the assessment of the generality and robustness of the approach across different types of design spaces and optimization settings.

1.5 Thesis Structure

The remainder of this thesis is organized as follows.

Chapter 2 provides the necessary background on materials discovery, with a particular focus on nanoporous materials such as Covalent Organic Frameworks (COFs), as well as advanced engineered materials including composite structures. It also discusses the key challenges associated with their design and optimization across different application domains.

Chapter 3 introduces the theoretical foundations of the methods used in this work, including Bayesian Optimization, surrogate modeling techniques, acquisition functions and Active Learning. Both single-objective and multi-objective optimization frameworks are discussed.

Chapter 4 presents the proposed methodology of this thesis. It describes the overall workflow, including the application of Bayesian Optimization on the original design space, the use of Active Learning for space reduction, and the subsequent re-application of Bayesian Optimization on the reduced space.

Chapter 5 details the experimental setup, including the datasets used, preprocessing steps, implementation details of the optimization methods and the evaluation metrics employed. In particular, experiments are conducted on datasets from different material domains, including nanoporous materials (COFs) and composite materials, allowing for the evaluation of the proposed framework across diverse design problems.

Chapter 6 presents the experimental results, comparing the performance of Bayesian Optimization on the original and reduced design spaces, as well as analyzing the impact of space reduction in both single-objective and multi-objective settings.

Chapter 7 discusses the results, providing interpretation of the findings, insights into the behavior of the proposed approach, and an analysis of its strengths and limitations.

Finally, Chapter 8 concludes the thesis and outlines potential directions for future work.

2 Background on Material discovery and optimization

This chapter provides the necessary background for understanding the problem of materials discovery and optimization, with a particular focus on nanoporous materials and Covalent Organic Frameworks (COFs). It introduces the fundamental concepts underlying modern materials design, highlights the challenges associated with large-scale material spaces, and establishes the context for the optimization and learning-based methods presented in the following chapters.

2.1 Materials Discovery Paradigm

Materials discovery can be viewed as a complex optimization problem, where the goal is to identify materials that satisfy specific performance criteria under a set of constraints. These criteria often include physical, chemical or mechanical properties, such as adsorption capacity, stability, selectivity or conductivity. The central difficulty lies in the fact that the relationship between a material's structure and its properties is typically highly nonlinear, high-dimensional and computationally expensive to evaluate.

Traditionally, materials discovery has followed a forward design approach, in which candidate materials are generated and subsequently evaluated through simulations or experiments. While this approach has led to significant advances, it becomes increasingly inefficient as the size of the design space grows. In many modern applications, the number of possible candidate materials is so large that exhaustive or even systematic exploration becomes impractical.

To address these limitations, recent research has shifted toward inverse design strategies. In this setting, the desired material properties are specified in advance, and the task is to identify structures that achieve these targets. This formulation naturally leads to the use of optimization techniques, where the design space is explored in a guided and data-driven manner rather than through brute-force enumeration.

A key component of this paradigm is the use of computational models to approximate the relationship between structure and properties. These models enable the rapid evaluation of candidate materials and serve as surrogates for expensive simulations or experiments. When combined with optimization strategies, such as Bayesian Optimization, they allow for efficient exploration of large and complex design spaces.

Despite these advances, several challenges remain. The high dimensionality of material representations, the presence of heterogeneous features, and the need to balance multiple competing objectives all contribute to the complexity of the problem. These challenges motivate the development of advanced frameworks that integrate machine learning, optimization and data selection techniques, as explored in this thesis.

2.2 Nanoporous Materials

Nanoporous materials constitute a broad class of materials characterized by the presence of pores with dimensions on the nanometer scale [16,21–23]. These materials have attracted significant attention due to their exceptionally high surface area and their ability to selectively adsorb, store, or transport molecules [21–23]. Such properties make them particularly suitable for applications in gas storage, gas separation, catalysis, and sensing technologies [16,21–23].

The performance of nanoporous materials is largely determined by their structural characteristics, including pore size, pore distribution, surface chemistry and overall topology. These properties directly influence how molecules interact with the material, affecting adsorption capacity, selectivity and diffusion behavior. As a result, even small variations in structural parameters can lead to significant changes in performance.

Several major classes of nanoporous materials have been extensively studied, including zeolites, metal-organic frameworks (MOFs) and covalent organic frameworks (COFs) [16,21–23]. Zeolites are crystalline aluminosilicates with well-defined pore structures, widely used in catalysis and separation processes. MOFs consist of metal ions coordinated with organic ligands, forming highly porous and tunable structures. COFs, which are the

primary focus of this work, are composed entirely of light elements connected through strong covalent bonds, offering high stability and structural flexibility.

A key advantage of nanoporous materials is their tunability. By modifying building blocks, linkers or synthesis conditions, it is possible to design materials with tailored properties for specific applications[36]. However, this flexibility also introduces a major challenge: the combinatorial explosion of possible material configurations. The number of potential structures increases rapidly with the number of design parameters, making systematic exploration extremely difficult.

This challenge is particularly evident in large-scale computational materials discovery, where databases may contain tens of thousands of candidate structures. In such settings, efficient methods are required to identify the most promising materials without evaluating the entire design space[37]. This need for efficient exploration strategies motivates the integration of machine learning and optimization techniques, which can guide the search toward high-performing regions of the space.

2.3 Covalent Organic Frameworks (COFs)

Covalent Organic Frameworks (COFs) are a class of crystalline porous materials composed entirely of light elements such as carbon, hydrogen, boron, nitrogen and oxygen, connected through strong covalent bonds. Since their introduction, COFs have attracted significant attention due to their unique combination of structural order, low density and high surface area, which make them particularly suitable for a wide range of applications in materials science.

A defining characteristic of COFs is their modular and highly tunable structure. They are constructed through the assembly of organic building blocks into periodic frameworks using principles of reticular chemistry. By carefully selecting and combining different molecular units, it is possible to precisely control key structural properties such as pore size, pore geometry, topology and chemical functionality. This level of control enables the

design of materials tailored to specific applications, making COFs an attractive platform for targeted materials engineering.

In contrast to other nanoporous materials, such as metal-organic frameworks (MOFs), COFs do not contain metal centers, which contributes to their lower density and potentially improved chemical stability in certain environments. At the same time, their fully organic composition allows for greater flexibility in functionalization, enabling the incorporation of specific chemical groups that enhance interactions with target molecules.

The properties of COFs are strongly influenced by both their structural and chemical characteristics. Parameters such as pore size distribution, surface area, porosity and functional groups play a critical role in determining their adsorption behavior and selectivity. As a result, COFs have been extensively studied for applications including gas storage, carbon capture, molecular separation, catalysis and energy storage.

However, the same structural flexibility that makes COFs highly attractive also leads to a significant increase in design complexity. The number of possible combinations of building blocks, linkers and topologies grows rapidly, resulting in a vast design space that is difficult to explore exhaustively. This challenge is further amplified when considering multiple target properties, as different structural configurations may optimize different performance criteria.

Consequently, the design and optimization of COFs require efficient computational strategies capable of navigating large and complex search spaces. This has motivated the increasing use of machine learning and optimization techniques, which can guide the discovery process by identifying promising candidates without the need for exhaustive evaluation.

2.4 Applications of COFs

Due to their unique structural characteristics, COFs have been widely explored across a broad range of applications in materials science and engineering [16,21–23]. Their high surface area, tunable pore structure and chemical versatility make them particularly suitable for processes that involve adsorption, separation and catalytic activity.

One of the most extensively studied applications of COFs is gas storage and separation. Their porous architecture enables the efficient adsorption of gases such as hydrogen, methane and carbon dioxide, making them promising candidates for energy storage and environmental applications. In particular, COFs have shown strong potential in carbon capture technologies, where selective adsorption of CO_2 is essential for reducing greenhouse gas emissions. The ability to tailor pore size and surface chemistry allows for improved selectivity and capacity, which are critical factors in such applications.

In addition to gas-related processes, COFs have demonstrated significant potential in catalysis. Their well-defined and ordered structures, combined with the possibility of functionalizing their internal surfaces, allow for the design of catalytic sites with high specificity and efficiency [22,23]. This makes them suitable for a variety of chemical reactions, including heterogeneous catalysis and photocatalysis.

COFs have also been investigated in energy-related applications, such as energy storage and conversion. Their structural stability and high surface area make them attractive for use in batteries, supercapacitors and fuel cells. Furthermore, their tunable electronic properties enable their application in optoelectronic devices and sensing technologies.

Despite these promising applications, achieving optimal performance requires careful design and tuning of the material structure. Different applications impose different and often conflicting requirements, such as maximizing adsorption capacity while maintaining structural stability or selectivity. This further emphasizes the need for systematic and efficient optimization strategies in COF design, particularly in the presence of large and complex design spaces.

2.5 Challenges in COF Design

Despite their significant potential and wide range of applications, the design and optimization of Covalent Organic Frameworks (COFs) remain highly challenging. These challenges stem primarily from the complexity of their structural space, the cost of accurate evaluation and the need to balance multiple performance criteria.

A fundamental challenge is the combinatorial explosion of possible COF structures. The large number of available building blocks, linkers and topological configurations leads to an extremely vast design space. Even when restricting attention to chemically feasible structures, the number of potential candidates remains prohibitively large, making exhaustive exploration infeasible. As a result, identifying high-performing materials requires efficient strategies for navigating this space.

Another important challenge lies in the complex relationship between structure and properties. Material performance is influenced by a combination of geometric, chemical and physical characteristics, which often interact in nonlinear and non-intuitive ways. Small variations in structural parameters can lead to significant changes in behavior, making it difficult to predict performance using simple models or intuition alone.

The computational cost of evaluating COFs further complicates the design process. Accurate estimation of properties such as adsorption capacity, selectivity or stability typically requires detailed molecular simulations or quantum-level calculations. These methods are computationally expensive and limit the number of candidate materials that can be evaluated within a reasonable time frame.

In addition, many COF design problems are inherently multi-objective. Practical applications often require the simultaneous optimization of multiple properties, such as maximizing adsorption capacity while maintaining structural stability or minimizing cost. These objectives are frequently conflicting, leading to trade-offs that must be carefully balanced. Identifying optimal solutions in such settings requires methods capable of capturing and exploring these trade-offs effectively.

Finally, the availability of large materials datasets introduces challenges related to data selection and utilization. While such datasets provide valuable information, they often contain a significant number of low-quality or irrelevant samples. Efficiently identifying and focusing on the most informative regions of the design space becomes essential in order to reduce computational effort and improve optimization performance.

These challenges collectively highlight the need for advanced computational approaches that combine machine learning, optimization and data selection techniques. In particular,

they motivate the use of methods such as Bayesian Optimization and Active Learning, which are specifically designed to address problems involving expensive evaluations and large, complex search spaces.

While the discussion in this section focuses on COFs, similar challenges arise in other materials design problems, including composite materials and engineering systems. These shared difficulties further motivate the development of general optimization frameworks capable of handling large, complex and multi-objective design spaces across different domains.

3 Bayesian Optimization and Active Learning

This chapter introduces the theoretical foundations of the optimization and learning techniques used in this work. It presents the principles of Bayesian Optimization, including surrogate modeling and acquisition functions, and discusses how these methods enable efficient exploration of large and complex design spaces. In addition, the chapter introduces Active Learning as a complementary strategy for data selection and search-space reduction, as well as its role in improving the efficiency of optimization processes in materials discovery.

3.1 Bayesian Optimization

Bayesian Optimization (BO) is a sequential optimization framework designed for problems where the evaluation of candidate solutions is expensive, time-consuming or limited [1,6]. It has been widely adopted in domains such as materials science, hyperparameter tuning and engineering design, where each function evaluation may require complex simulations or experimental measurements [6,12].

The main objective of Bayesian Optimization is to identify the global optimum of an unknown objective function using as few evaluations as possible. Instead of directly optimizing the function, BO constructs a probabilistic surrogate model that approximates the underlying objective based on previously observed data. This surrogate model is then used to guide the selection of new candidate points in a data-efficient manner.

At each iteration, Bayesian Optimization follows a sequential process. First, the surrogate model is updated using all available observations. Then, an acquisition function is optimized in order to determine the next point to evaluate. The acquisition function quantifies the utility of sampling a particular point by balancing two competing objectives: exploration and exploitation. Exploration refers to sampling regions of the design space

where uncertainty is high, while exploitation focuses on regions that are predicted to yield high objective values.

One of the key advantages of Bayesian Optimization is its ability to incorporate uncertainty into the decision-making process. By maintaining a probabilistic model of the objective function, BO can explicitly reason about which regions of the design space are uncertain and potentially informative. This makes it particularly suitable for problems where evaluations are costly and must be carefully selected.

In the context of materials discovery, Bayesian Optimization provides an effective mechanism for navigating large design spaces while minimizing the number of required evaluations. By iteratively refining its understanding of the objective function, BO can focus the search on promising regions and avoid unnecessary exploration of low-quality candidates. This property makes it a natural choice for applications involving nanoporous materials such as COFs, where the number of possible structures is extremely large.

Despite its advantages, the performance of Bayesian Optimization depends strongly on the choice of surrogate model and acquisition function, as well as on the characteristics of the underlying problem. High-dimensional spaces, heterogeneous features and complex objective landscapes can pose challenges to the effectiveness of BO, motivating the development of complementary techniques such as Active Learning and design space reduction, which are explored in this work.

3.2 Surrogate Models

A central component of Bayesian Optimization is the surrogate model, which serves as an approximation of the unknown objective function. Since direct evaluations are often computationally expensive, the surrogate model provides a fast and tractable way to estimate function values across the design space. In addition to predictions, an effective surrogate model should also provide information about uncertainty, which is essential for guiding the optimization process.

One of the most widely used surrogate models in Bayesian Optimization is the Gaussian Process (GP). Gaussian Processes are non-parametric probabilistic models that define a

distribution over functions and provide both a mean prediction and a measure of uncertainty at each point in the design space. This uncertainty estimation is a key advantage, as it allows the optimization algorithm to identify regions where predictions are less certain and potentially informative. As a result, GP-based models are particularly well-suited for problems with limited data and expensive evaluations.

However, Gaussian Processes can face scalability issues when applied to large datasets, as their computational complexity grows rapidly with the number of observations. In high-dimensional or large-scale problems, this can limit their practical applicability. To address these limitations, alternative modeling approaches have been explored in the literature, particularly for improving scalability in complex settings.

In the proposed framework of this work, Gaussian Processes are used as the surrogate model within the Bayesian Optimization process. In contrast, XGBoost is not employed as a surrogate model for BO, but is instead used within the Active Learning (DAGS) stage as a classification model. Specifically, XGBoost is part of an ensemble of classifiers used to distinguish promising from non-promising samples in the design space, contributing to the reduction of the search space prior to optimization.

While Gaussian Process models guide the optimization process by modeling the objective function and providing uncertainty estimates, XGBoost contributes to the data selection phase by offering scalability and strong predictive performance in high-dimensional datasets. This separation of roles allows the framework to combine uncertainty-aware optimization with efficient space reduction.

Overall, the selection and use of modeling techniques play a critical role in the effectiveness of the proposed approach. The combination of probabilistic surrogate modeling for optimization and scalable machine learning methods for data filtering is particularly relevant in materials discovery applications, where datasets can be large and complex.

3.3 Acquisition Functions

In Bayesian Optimization, the surrogate model is used in conjunction with an acquisition function to determine the next point to evaluate. While the surrogate model provides an approximation of the objective function, the acquisition function defines a decision strategy that guides the search process. Its role is to quantify the utility of sampling a given point in the design space, enabling the algorithm to select the most informative or promising candidate at each iteration.

A key aspect of acquisition functions is their ability to balance exploration and exploitation. Exploration refers to sampling regions of the design space where the model is uncertain, in order to improve its overall understanding of the objective function. Exploitation, on the other hand, focuses on sampling regions that are predicted to yield high performance based on current knowledge. An effective acquisition function must carefully trade off between these two objectives in order to efficiently locate the global optimum.

Among the various acquisition strategies, Expected Improvement (EI) is one of the most widely used due to its simplicity and effectiveness. The idea behind EI is to select points that are expected to improve upon the best observed value of the objective function. Instead of considering only the predicted mean, EI also takes into account the uncertainty of the surrogate model, favoring points that either have high predicted performance or high uncertainty.

More specifically, Expected Improvement evaluates the expected gain that can be achieved by sampling a candidate point, relative to the current best solution. Points with high EI values are those that offer a strong potential for improvement, either by refining known high-performing regions or by exploring uncertain areas that may contain better solutions.

In practice, the acquisition function is optimized at each iteration to identify the next candidate point. This optimization step is typically much cheaper than evaluating the true objective function, allowing the algorithm to efficiently guide the search process. The

selected point is then evaluated using the true objective, and the surrogate model is updated with the new observation.

In the context of this work, Expected Improvement is used as the primary acquisition function for guiding Bayesian Optimization. Its ability to incorporate both prediction and uncertainty makes it particularly suitable for problems involving expensive evaluations and large design spaces, such as the materials discovery tasks considered in this thesis.

3.4 Multi-Objective Bayesian Optimization

In many real-world optimization problems, including materials design, it is often necessary to optimize multiple objectives simultaneously. These objectives may represent different performance criteria, which are frequently conflicting. For example, improving one property of a material, such as adsorption capacity, may lead to a degradation in another property, such as structural stability. As a result, the notion of a single optimal solution is replaced by a set of trade-off solutions.

This setting gives rise to Multi-Objective Optimization, where the goal is to identify a set of solutions that represent the best possible compromises among the objectives. These solutions are commonly referred to as Pareto-optimal. A solution is considered Pareto-optimal if there exists no other solution that improves one objective without degrading at least one other. The collection of all such solutions forms the Pareto front, which represents the optimal trade-offs achievable within the design space.

Multi-Objective Bayesian Optimization (MOBO) extends the principles of Bayesian Optimization to this setting. Instead of modeling a single objective function, MOBO constructs surrogate models for each objective and uses them to guide the search toward regions that improve the overall quality of the Pareto front. This process requires specialized acquisition functions capable of evaluating the benefit of sampling new points in a multi-objective context.

One widely used approach for handling multiple objectives is scalarization, where multiple objectives are combined into a single objective using a weighted sum. This method simplifies the optimization process and allows the use of standard Bayesian Optimization techniques. However, scalarization requires the selection of weights, which implicitly define a preference over the objectives and may not fully capture the structure of the Pareto front.

An alternative and more principled approach involves directly optimizing the Pareto front using hypervolume-based methods. The hypervolume metric measures the volume of the objective space dominated by the current set of solutions relative to a reference point. A larger hypervolume indicates a better approximation of the Pareto front, as it reflects both the quality and the diversity of the solutions.

In this context, acquisition functions such as Expected Hypervolume Improvement (EHVI) and its batch variant, qEHVI, have been proposed to guide the optimization process. These methods aim to select new points that are expected to maximally increase the hypervolume of the current solution set. Unlike scalarization-based approaches, hypervolume-based methods do not require predefined weights and instead focus on improving the overall shape and coverage of the Pareto front.

Despite their advantages, multi-objective Bayesian Optimization methods are computationally more demanding than their single-objective counterparts. They require accurate modeling of multiple objective functions, as well as efficient computation of the hypervolume and its expected improvement. These challenges become more pronounced in high-dimensional or large-scale problems, motivating the need for techniques that can reduce the complexity of the search space.

In the context of this work, Multi-Objective Bayesian Optimization is applied in order to evaluate the effectiveness of the proposed framework in identifying high-quality trade-off solutions. Both scalarization-based and hypervolume-based approaches are considered, allowing for a comprehensive analysis of their behavior under different design spaces, including both the original and reduced spaces.

3.5 Active Learning

Active Learning is a machine learning paradigm that focuses on improving model performance by selectively choosing the most informative data points for training [4,5]. Instead of using all available data, which may include redundant or low-value samples, Active Learning aims to identify and prioritize the samples that contribute the most to the learning process.

The central idea behind Active Learning is that not all data points are equally useful. In many real-world datasets, particularly in materials discovery, a large portion of the data may correspond to low-performing or uninformative regions of the design space. Training models on such data can lead to inefficiencies, increased computational cost and reduced predictive performance. By contrast, focusing on informative samples allows for more efficient use of computational resources[24].

Active Learning operates through an iterative process. Initially, a model is trained on a subset of the available data. The model is then used to evaluate the remaining data and estimate which samples would be most beneficial if included in the training set. Based on a selection criterion, these samples are added to the training set, and the model is updated accordingly. This process can be repeated until a desired level of performance is achieved.

One of the most commonly used selection strategies in Active Learning is uncertainty-based sampling. In this approach, the model selects data points for which it is least confident in its predictions. These points are considered informative because they help reduce uncertainty and improve the model's understanding of the underlying data distribution. This strategy is particularly effective in settings where labeled data is expensive or where the goal is to improve model performance with a limited number of samples.

In addition to uncertainty-based methods, other selection strategies may include diversity-based sampling and query-by-committee approaches. These methods aim to ensure that the selected samples provide broad coverage of the design space or capture different perspectives of the model. However, uncertainty-based sampling remains one of the most widely used approaches due to its simplicity and effectiveness.

In the context of this work, Active Learning is used not only as a tool for improving model performance but also as a mechanism for reducing the size of the design space. By identifying and retaining the most informative and high-potential samples, it becomes possible to construct a reduced dataset that preserves the essential structure of the problem while eliminating a large portion of irrelevant or low-quality data. This reduced space is then used as the input for subsequent optimization steps, allowing for more efficient exploration and faster convergence of the optimization process.

Overall, Active Learning provides a powerful framework for guiding data selection in large and complex datasets. Its ability to focus on informative samples makes it particularly suitable for applications in materials discovery, where both data size and evaluation cost pose significant challenges.

3.6 Active Learning in Materials Discovery

The application of Active Learning in materials discovery has gained increasing attention in recent years, particularly in problems characterized by large datasets and expensive evaluation processes. In such settings, the ability to selectively focus on the most informative regions of the design space is essential for improving efficiency and reducing computational cost.

In materials science, datasets often contain a large number of candidate structures, many of which correspond to low-performing or irrelevant solutions. Evaluating all candidates using high-fidelity simulations or experiments is typically infeasible. Active Learning addresses this challenge by enabling the identification of samples that are most likely to improve model performance or contribute to the discovery of optimal materials.

A common approach in this context involves training a predictive model on an initial subset of data and then using this model to guide the selection of additional samples. Uncertainty-based strategies are particularly effective, as they prioritize regions of the design space where the model is less confident. By iteratively refining the dataset in this

manner, Active Learning can significantly improve the efficiency of the overall discovery process.

Beyond improving predictive performance, Active Learning can also be used as a tool for design space reduction. By filtering out regions that are unlikely to contain high-quality solutions, it becomes possible to construct a smaller and more focused dataset that retains the most relevant information. This approach is especially valuable in large-scale materials discovery problems, where the majority of candidate materials may not contribute meaningfully to the optimization objective.

In the context of nanoporous materials such as COFs, Active Learning has been applied to identify promising structures for applications such as gas storage and separation. By focusing on high-potential candidates, researchers can significantly reduce the number of required evaluations while maintaining the quality of the discovered materials. Similar approaches have also been explored in other materials domains, including composite materials and engineering design problems, where large and complex design spaces are common.

In this work, Active Learning is employed as a preprocessing step for Bayesian Optimization. The goal is to reduce the size of the original design space while preserving regions that are likely to contain optimal or near-optimal solutions. This reduced space is then used to perform optimization more efficiently, enabling a direct comparison between optimization in the original and reduced design spaces.

Overall, the integration of Active Learning into materials discovery workflows provides a powerful mechanism for improving both data efficiency and optimization performance. By combining data selection with optimization strategies, it becomes possible to address the challenges of large-scale and high-dimensional design spaces in a systematic and effective manner.

4 Proposed Methodology

This chapter presents the methodology proposed in this thesis for improving the efficiency of materials discovery through the integration of Bayesian Optimization and Active Learning. The overall approach is based on a structured workflow in which Bayesian Optimization is first applied to the original design space, followed by a reduction of the search space using Active Learning techniques, and finally a re-application of Bayesian Optimization on the reduced space. This process enables a direct comparison between optimization in the full and reduced design spaces, allowing the impact of space reduction on performance and efficiency to be systematically evaluated.

4.1 Problem Formulation

The problem addressed in this thesis can be formulated as an optimization task over a set of candidate materials, where each material is described by a set of features and associated with one or more target properties. Let the design space consist of a finite set of candidate materials, each represented by a feature vector that captures its structural and physicochemical characteristics.

In the single-objective setting, the goal is to identify the material that maximizes or minimizes a given target property. This can be expressed as the optimization of an unknown objective function defined over the design space. Since direct evaluation of this function is computationally expensive, it is assumed that only a limited number of evaluations can be performed.

In the multi-objective setting, the problem involves optimizing multiple target properties simultaneously. These objectives are often conflicting, meaning that no single solution optimizes all objectives at once. Instead, the goal is to identify a set of Pareto-optimal solutions that represent the best trade-offs among the objectives.

A key challenge in both settings is the large size of the design space, which makes exhaustive evaluation infeasible. Additionally, the cost associated with evaluating candidate materials further limits the number of samples that can be considered. As a

result, efficient strategies are required to guide the search toward high-performing regions of the space.

In this work, the optimization process is carried out in two stages. First, Bayesian Optimization is applied directly to the original design space in order to establish a baseline performance. Then, Active Learning is used to reduce the size of the design space by selecting a subset of informative and high-potential samples. Finally, Bayesian Optimization is applied again on the reduced space, allowing for a comparison between the two settings.

This formulation allows the investigation of whether design space reduction can improve optimization efficiency without significantly degrading the quality of the obtained solutions. The approach is evaluated in both single-objective and multi-objective scenarios, enabling a comprehensive analysis of its effectiveness across different optimization settings.

4.2 Baseline Bayesian Optimization

As a first step of the proposed methodology, Bayesian Optimization is applied directly to the original design space in order to establish a baseline for performance evaluation. This baseline serves as a reference point against which the effectiveness of subsequent space reduction techniques can be assessed.

In this setting, the full dataset of candidate materials is considered without any prior filtering or modification. Each material is represented by its corresponding feature vector, and the objective function is defined based on the target property or properties of interest. The goal is to identify high-performing materials using a limited number of evaluations, simulating realistic scenarios where computational or experimental resources are constrained.

The optimization process follows the standard Bayesian Optimization workflow. Initially, a small set of samples is selected from the design space and evaluated. These initial

observations are used to train a surrogate model, which approximates the objective function across the entire space. Based on this model, an acquisition function is used to iteratively select new candidate points for evaluation.

At each iteration, the surrogate model is updated with the newly acquired data, and the acquisition function is re-optimized to determine the next sampling point. This sequential process continues for a predefined number of iterations or until a stopping criterion is reached. The performance of the method is then evaluated based on the quality of the solutions identified and the efficiency of the search process.

In the multi-objective setting, the same procedure is extended to multiple objectives, where separate surrogate models are constructed for each objective. The optimization is guided either through scalarization techniques or through hypervolume-based acquisition functions, enabling the identification of Pareto-optimal solutions.

The results obtained from this baseline optimization provide a benchmark for comparison with the subsequent stages of the methodology. In particular, they allow for the evaluation of how effectively Bayesian Optimization can explore the original design space and identify high-quality solutions without any form of space reduction.

4.3 Design Space Reduction via Active Learning

A central component of the proposed methodology is the reduction of the design space prior to the application of Bayesian Optimization. Given the large size and complexity of materials datasets, directly applying optimization methods to the full space can lead to inefficient exploration and increased computational cost. To address this limitation, an Active Learning-based approach is employed in order to construct a reduced design space that focuses on the most informative and high-potential candidate materials.

The underlying idea is that, in many large-scale materials datasets, only a relatively small subset of candidates contributes significantly to the discovery of optimal solutions. The majority of samples may correspond to low-performing or less relevant regions of the design space. By identifying and retaining only the most informative samples, it becomes

possible to reduce the size of the dataset while preserving its most important structural and performance characteristics.

In this work, the design space reduction is achieved through an approach based on the DAGS framework, which builds upon previous research in Active Learning developed within the context of materials design. In its original formulation, this methodology has primarily been applied in regression-based settings, focusing on the prediction of continuous material properties.

In the context of this thesis, the method is extended in two key directions. First, it is adapted from a regression-based formulation to a classification-based Active Learning framework. Instead of predicting continuous values, the approach is used to distinguish between high-potential and low-potential regions of the design space.

Second, and more importantly, the method is employed as a mechanism for design space reduction. This represents a key contribution of the present work, as the framework is used to explicitly filter the original dataset and construct a reduced space that concentrates on promising candidate materials. To the best of our knowledge, this use of the method as a systematic preprocessing step for optimization has not been previously explored in this form.

More specifically, a classification model is trained to identify regions of the design space that are likely to contain high-quality solutions. Based on this model, samples that fall within the high-potential region are retained, while the remaining samples are filtered out. This results in a reduced dataset that eliminates a large portion of low-performing candidates while preserving the most relevant structures.

An important aspect of this process is the use of a threshold parameter, which controls the strictness of the filtering. Lower threshold values lead to a more conservative reduction, retaining a larger portion of the design space and ensuring broader coverage. Higher threshold values result in a more aggressive reduction, focusing on a smaller subset of highly promising candidates while increasing the risk of excluding optimal solutions. This trade-off between coverage and selectivity is a key factor in the effectiveness of the approach.

The resulting reduced design space is then used as the input for the subsequent optimization stage. By applying Bayesian Optimization on this smaller and more focused space, it becomes possible to improve computational efficiency and accelerate convergence, while maintaining the ability to identify high-quality solutions.

Overall, this approach extends prior Active Learning methodologies by introducing a classification-based formulation and leveraging it as a tool for design space reduction. This dual contribution plays a central role in the proposed optimization pipeline and distinguishes the present work from existing approaches.

4.4 Proposed Optimization Pipeline

The proposed methodology can be summarized as a structured optimization pipeline that integrates Bayesian Optimization and Active Learning-based design space reduction. The goal of this pipeline is to improve the efficiency of the search process by focusing computational resources on the most promising regions of the design space.

The pipeline consists of three main stages. In the first stage, Bayesian Optimization is applied directly to the original design space in order to establish a baseline performance. This step provides a reference for evaluating the effectiveness of the subsequent space reduction process and allows for the assessment of how efficiently the optimization method can identify high-performing solutions in the full dataset.

In the second stage, Active Learning is used to perform design space reduction. A classification-based model is trained to distinguish between high-potential and low-potential regions of the design space, based on the characteristics of the candidate materials. Using this model, the original dataset is filtered, and a reduced design space is constructed by retaining only the most promising samples. The level of reduction is controlled through a threshold parameter, which determines the strictness of the filtering process.

In the third stage, Bayesian Optimization is re-applied on the reduced design space. Since the search is now restricted to a smaller and more focused set of candidates, the

optimization process can operate more efficiently, potentially leading to faster convergence and improved performance under limited evaluation budgets.

The outputs of the baseline optimization and the reduced-space optimization are then compared in order to evaluate the impact of design space reduction. This comparison focuses on both the quality of the solutions identified and the efficiency of the search process.

The same pipeline is applied in both single-objective and multi-objective settings, allowing for a comprehensive evaluation of the proposed approach across different types of optimization problems. In the multi-objective case, the effectiveness of the method is assessed in terms of its ability to recover high-quality Pareto-optimal solutions and to maintain a well-defined approximation of the Pareto front.

Overall, the proposed pipeline provides a systematic framework for integrating data selection and optimization in large-scale materials design problems. By combining Active Learning-based filtering with Bayesian Optimization, the approach aims to address the challenges associated with large design spaces and expensive evaluations.

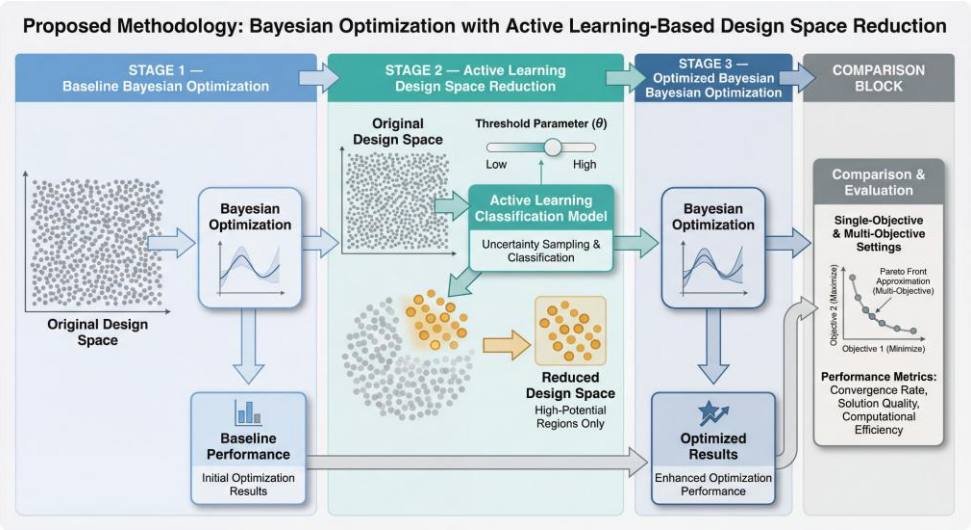

Figure 1: Proposed Optimization Pipeline

4.5 Multi-Objective Extension

The proposed optimization pipeline can be extended to handle multi-objective optimization problems, which are commonly encountered in real-world design scenarios. In such settings, the objective is not to identify a single optimal solution, but rather a set of solutions that represent the best trade-offs among multiple competing objectives.

In the context of this thesis, the multi-objective formulation is not applied to the COFs dataset, which is treated as a single-objective problem. Instead, the multi-objective extension of the pipeline is evaluated on a separate material design problem, namely the Pressure Vessel dataset, which involves multiple conflicting objectives.

The overall structure of the pipeline remains unchanged. Initially, Bayesian Optimization is applied to the original design space in order to establish a baseline performance. This provides a reference for evaluating the impact of design space reduction in the multi-objective setting.

Subsequently, Active Learning is employed to reduce the design space, following the methodology described in the previous section. The classification-based model is used to identify regions that are likely to contain high-quality trade-off solutions, enabling the construction of a reduced dataset that focuses on promising candidates.

Bayesian Optimization is then re-applied on the reduced design space using multi-objective optimization techniques. In this setting, separate surrogate models are used for each objective, and the optimization process is guided by acquisition functions designed for multi-objective scenarios.

Two main approaches are considered for handling multiple objectives. The first is scalarization, where the objectives are combined into a single objective using weighted sums. The second approach is based on hypervolume-driven optimization, using acquisition functions such as Expected Hypervolume Improvement (EHVI) and its batch extension, qEHVI, which aim to directly improve the Pareto front.

The effectiveness of the multi-objective extension is evaluated by comparing the quality of the resulting Pareto front before and after design space reduction. In particular, metrics

such as hypervolume are used to assess how well the reduced-space optimization preserves the structure and coverage of the Pareto front.

Overall, this extension allows the proposed methodology to be evaluated in a more general setting, demonstrating its applicability not only to single-objective materials discovery problems but also to multi-objective optimization tasks in different material domains.

5 Experimental Setup

This chapter describes the experimental setup used to evaluate the proposed optimization framework. It includes the datasets considered, the preprocessing steps applied, the implementation details of the optimization methods, and the evaluation metrics used to assess performance. The experiments are designed to analyze the effectiveness of the proposed pipeline in both single-objective and multi-objective settings, as well as to investigate the impact of design space reduction on optimization efficiency and solution quality.

5.1 Datasets

In this work, the proposed optimization framework is evaluated on two distinct material design problems, each representing a different type of design space and optimization setting. The use of multiple datasets allows for a more comprehensive assessment of the generality and robustness of the approach.

The first dataset consists of Covalent Organic Frameworks (COFs), which are nanoporous materials widely studied for applications such as gas storage and separation. Each material in the dataset is described by a set of physicochemical and structural features, including geometric descriptors and properties related to porosity, density and surface area. The objective in this case is formulated as a single-objective optimization problem, where the goal is to identify materials that optimize a target property of interest. This dataset serves as the primary test case for evaluating the effectiveness of the proposed pipeline in large-scale materials discovery scenarios.

The second dataset corresponds to a composite material design problem, commonly referred to as the Pressure Vessel dataset. This problem involves the optimization of structural design parameters under multiple constraints and objectives, making it suitable for evaluating multi-objective optimization methods. The design variables in this dataset include a combination of continuous and discrete parameters, such as geometric and structural characteristics, which define the configuration of the material system.

Unlike the COFs dataset, the Pressure Vessel problem is formulated as a multi-objective optimization task, where multiple competing objectives must be optimized simultaneously. This dataset is used to assess the performance of the proposed methodology in identifying high-quality trade-off solutions and approximating the Pareto front in complex design spaces.

The use of these two datasets enables the evaluation of the proposed framework across different types of material systems, including nanoporous materials and composite materials. This allows for the investigation of how the methodology performs under varying design space characteristics, such as dimensionality, variable types and objective structure.

5.2 Preprocessing

Prior to the application of the optimization framework, a series of preprocessing steps are performed on the datasets in order to ensure consistency, stability and comparability of the results across different experimental settings.

First, feature normalization is applied to the input variables in both the single-objective and multi-objective problems. This step ensures that all features are scaled to a comparable range, preventing variables with larger numerical values from disproportionately influencing the surrogate models. In addition, normalization is also applied to the target variables, which improves numerical stability and facilitates the training of both regression and classification models.

In the single-objective setting, the optimization is performed directly on the target variable of interest, without any transformation or combination with other objectives. In contrast, the multi-objective setting requires additional preprocessing in order to handle multiple competing objectives.

More specifically, in the multi-objective case, scalarization is employed as one of the approaches for combining multiple objectives into a single optimization criterion. This is achieved through the use of weighted sums, allowing the application of standard optimization techniques while capturing different trade-offs between the objectives.

In addition to scalarization, Pareto-based analysis is performed in order to evaluate the quality of the obtained solutions. The Pareto front is used to represent the set of non-dominated solutions, while performance metrics such as hypervolume are computed to quantify the quality and diversity of the solution set.

An important aspect of the preprocessing stage is the application of the DAGS-based filtering mechanism for design space reduction. In the single-objective case, the threshold used for filtering is defined based on the value of the target variable. This allows the classification model to distinguish between high-performing and low-performing samples.

In the multi-objective setting, the threshold is determined based on a combined score derived from the multiple objectives. This enables the classification model to identify candidate solutions that represent desirable trade-offs, rather than optimizing a single performance criterion.

Overall, the preprocessing steps ensure that both datasets are appropriately prepared for the application of the proposed optimization pipeline. They also enable a consistent comparison between single-objective and multi-objective settings, while supporting the effective application of Active Learning and Bayesian Optimization techniques.

5.3 Bayesian Optimization Implementation

The implementation of Bayesian Optimization in this work is based on a combination of established machine learning libraries and custom components, allowing flexibility in both modeling and optimization strategies.

Specifically, the optimization framework is implemented using BoTorch and GPyTorch for probabilistic modeling and acquisition optimization, along with additional components from the scikit-learn library and custom-developed modules. This combination enables the use of both Gaussian Process-based surrogate models and alternative models such as XGBoost, depending on the requirements of the experiment.

The optimization process follows a sequential setting, where a single candidate point is selected and evaluated at each iteration. This corresponds to a batch size of one, ensuring that the model is updated after every evaluation and that each decision is informed by the most recent observations. This sequential approach is particularly suitable for problems involving expensive evaluations, as it maximizes the information gained from each step.

The number of iterations is selected based on the complexity of the problem and the size of the design space. In the single-objective setting, approximately 100 iterations are performed in order to allow sufficient exploration of the search space and accurate identification of high-performing solutions. In the multi-objective setting, a smaller number of iterations, approximately 50, is used due to the increased computational cost associated with modeling multiple objectives and computing hypervolume-based acquisition functions.

For the DAGS-based space reduction stage, a significantly smaller number of iterations is employed, as the goal is not full optimization but rather the identification of informative samples and high-potential regions of the design space. This step focuses on efficient filtering rather than exhaustive search.

At each iteration of the Bayesian Optimization process, the surrogate model is updated with the newly observed data, and the acquisition function is re-optimized to determine the next candidate point. In the single-objective case, Expected Improvement is used as the acquisition function, while in the multi-objective setting, both scalarization-based approaches and hypervolume-based methods such as qEHVI are considered.

Overall, the implementation is designed to balance computational efficiency and modeling accuracy, enabling a consistent comparison between optimization in the original and reduced design spaces across both single-objective and multi-objective scenarios.

5.4 DAGS Implementation

The design space reduction stage is implemented using an Active Learning framework based on the DAGS methodology [3], adapted to operate as a classification-driven

approach. This stage aims to identify and retain high-potential samples from the original dataset, thereby constructing a reduced search space for subsequent optimization.

In both the single-objective and multi-objective settings, the DAGS-based Active Learning framework is formulated as a classification problem. In the single-objective case, the target variable is directly used to define a threshold that separates promising from non-promising samples. In the multi-objective case, the objective values are first normalized and transformed into a single scalar score through scalarization. This score is then used to define the classification boundary between high-potential and low-potential regions of the design space.

The Active Learning process begins with a small, balanced labeled dataset consisting of both positive and negative samples. Typically, this initial set contains a small number of samples, ensuring equal representation of both classes in order to facilitate stable model training.

At each iteration, an ensemble of classification models is trained on the current labeled dataset. The ensemble includes Random Forest, Logistic Regression and XGBoost classifiers, allowing the model to capture different types of decision boundaries and improve robustness. The use of multiple classifiers enables a more reliable estimation of class membership and supports more effective selection of informative samples.

A new sample is selected at each iteration using a custom query strategy designed to balance exploration and exploitation. This selection mechanism incorporates multiple criteria, including feature-space diversity, output-space novelty and density information. By combining these factors, the method prioritizes samples that are both informative and representative of underexplored regions of the design space.

The selected sample is then evaluated using an oracle and added to the labeled dataset with its true label. In addition to oracle-labeled samples, the method also incorporates pseudo-labeling. High-confidence predictions from the classification models are used to assign labels to selected unlabeled samples, which are then included in the training set. This mechanism accelerates the growth of the labeled dataset and improves learning efficiency without requiring additional expensive evaluations.

The Active Learning process continues for a predefined number of iterations, gradually expanding the labeled dataset through both oracle-labeled and pseudo-labeled samples. At the end of this process, the trained ensemble of classifiers is used to identify the set of promising samples within the original dataset.

The reduced design space is then constructed as the union of samples classified as promising by the ensemble. This reduced dataset serves as the input for the subsequent Bayesian Optimization stage, enabling more efficient exploration and improved convergence.

Overall, the implementation of the DAGS framework combines ensemble learning, diversity-aware sampling and pseudo-labeling techniques to achieve effective and scalable design space reduction. This approach plays a central role in the proposed methodology, significantly improving the efficiency of the optimization process. This unified formulation allows the same Active Learning mechanism to be applied consistently across both single-objective and multi-objective optimization settings.

5.5 Evaluation Metrics

The performance of the proposed optimization framework is evaluated using a set of metrics tailored to both single-objective and multi-objective optimization settings. These metrics are selected in order to capture not only the quality of the solutions identified, but also the efficiency of the optimization process.

In the single-objective setting, the evaluation focuses on the ability of the method to identify high-performing solutions within a limited number of evaluations. The primary metric used is the best value found during the optimization process, which reflects the highest-performing solution discovered. In addition, top-k performance is considered, measuring how many of the best-performing candidates in the dataset are successfully identified by the optimization method. This provides a more comprehensive view of the model's ability to rank and retrieve high-quality solutions, rather than focusing on a single optimum.

Another important metric in the single-objective setting is sample efficiency. This metric evaluates how quickly the optimization method converges toward high-performing solutions as the number of evaluations increases. It provides insight into the effectiveness of the search strategy, particularly in scenarios where evaluation cost is a limiting factor.

In the multi-objective setting, the evaluation is based on metrics that assess the quality and diversity of the Pareto front approximation. A key metric in this context is the hypervolume (HV), which measures the volume of the objective space dominated by the obtained solutions with respect to a reference point. A larger hypervolume indicates a better approximation of the Pareto front, as it reflects both improved performance and broader coverage of the objective space.

In addition to hypervolume, the number of nondominated points is used to evaluate the richness of the obtained solution set. This metric reflects how many solutions belong to the approximated Pareto front and provides an indication of the diversity of trade-offs captured by the optimization process.

Furthermore, the number of true Pareto points found is considered, measuring how many of the identified solutions correspond to the actual Pareto-optimal set of the dataset. This metric provides a direct assessment of the accuracy of the method in recovering optimal trade-off solutions.

Finally, the number of evaluations used is reported in order to assess the computational efficiency of the method. This metric is particularly important for comparing performance between optimization in the original design space and in the reduced space, as it reflects the cost required to achieve a given level of performance.

Overall, these metrics provide a comprehensive evaluation of the proposed framework, capturing both the quality of the solutions and the efficiency of the optimization process across different experimental settings.

5.6 Initialization Strategy

In this work, two different initialization strategies are considered for Bayesian Optimization, namely cold-start and warm-start initialization. These strategies determine how the initial training data for the surrogate model is selected and have a significant impact on the efficiency of the optimization process.

In the cold-start setting, the optimization begins without prior knowledge, using a small number of randomly selected samples from the design space. These initial samples are used to train the surrogate model, after which the acquisition function guides the selection of new candidate points. However, since the initial samples are randomly selected, they may lie in non-informative regions of the design space, potentially leading to slower convergence and inefficient use of evaluations.

In contrast, the warm-start setting leverages prior information obtained from the Active Learning stage. Specifically, the DAGS framework identifies promising samples that are more likely to lie in high-performing regions of the design space. These samples are used to initialize the Bayesian Optimization process, allowing the surrogate model to be trained on more informative data from the outset.

As a result, warm-start initialization provides a more favorable starting point for the optimization process, enabling faster convergence and improved sample efficiency. This is particularly important in scenarios where the number of allowed evaluations is limited.

The impact of the initialization strategy is evaluated in both single-objective and multi-objective settings, allowing for a direct comparison between cold-start and warm-start Bayesian Optimization.

6 Results

The results presented in this chapter evaluate the performance of the proposed optimization framework across different experimental settings. The analysis focuses on comparing Bayesian Optimization in the original design space with the proposed reduced-space approach, as well as examining the impact of initialization strategies and design space reduction on optimization efficiency and solution quality.

6.1 Single-Objective Results

The single-objective experiments are conducted on the COFs dataset, with the objective of maximizing the target property (e.g., deliverable capacity). The results compare the performance of Bayesian Optimization in the full design space with the proposed reduced-space approach, as well as the effect of different initialization strategies.

As shown in Figure 2: Convergence Comparison, Bayesian Optimization applied to the reduced design space demonstrates faster convergence compared to the full-space baseline. In particular, the warm-start configuration enables the optimization process to reach high-performing solutions significantly earlier than both the cold-start reduced-space approach and the full-space optimization.

This behavior indicates that the combination of Active Learning-based space reduction and warm-start initialization effectively guides the search toward high-potential regions of the design space, resulting in a more targeted exploration and improved optimization efficiency.

Moreover, the warm-start configuration significantly reduces the number of evaluations required to reach near-optimal performance, highlighting the importance of initializing the optimization process with informative samples.

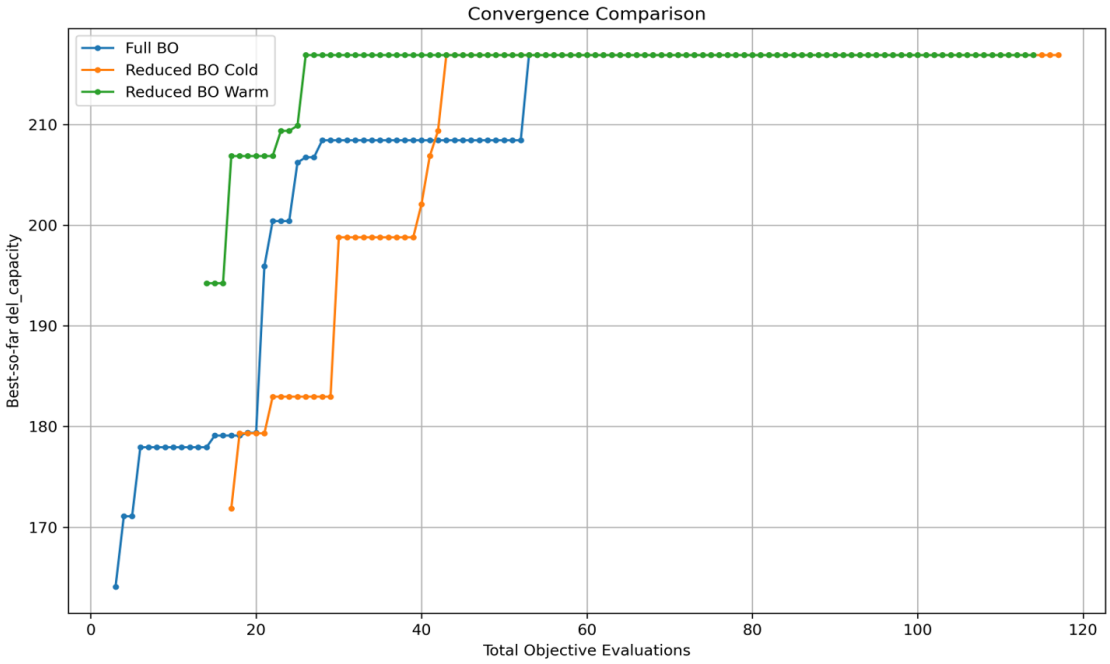

Figure 2: Convergence Comparison

The final best value achieved by the optimization process is comparable across all configurations, indicating that the reduced design space preserves the optimal solution quality. This confirms that the proposed reduction method does not negatively affect the ability of the optimization process to identify high-quality solutions.

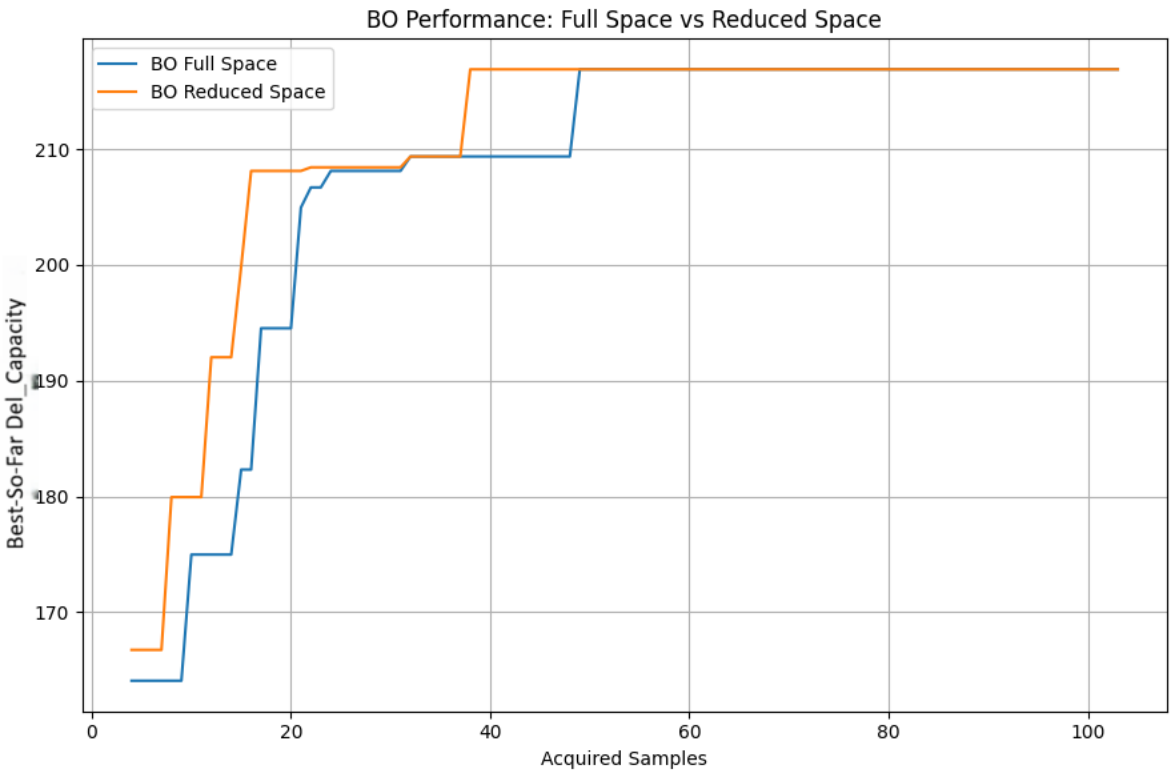

Figure 3: Best value Found

The results for top-1 performance further highlight the efficiency of the reduced-space approach. As shown in Figure 4:Top-1 Found, the reduced-space optimization, particularly in the warm-start configuration, identifies the top-performing candidate significantly earlier than both the cold-start and full-space approaches. A similar trend is observed for top-10 performance. The reduced design space allows the optimization process to identify multiple high-performing candidates more rapidly compared to the full-space baseline. This demonstrates improved ranking capability and more targeted exploration of promising regions of the design space. The results for top-100 performance provide further evidence of the effectiveness of the proposed approach. As illustrated in Figure 6, the reduced-space optimization consistently identifies a larger number of high-quality candidates at earlier stages of the optimization process, indicating better coverage of high-performing regions.

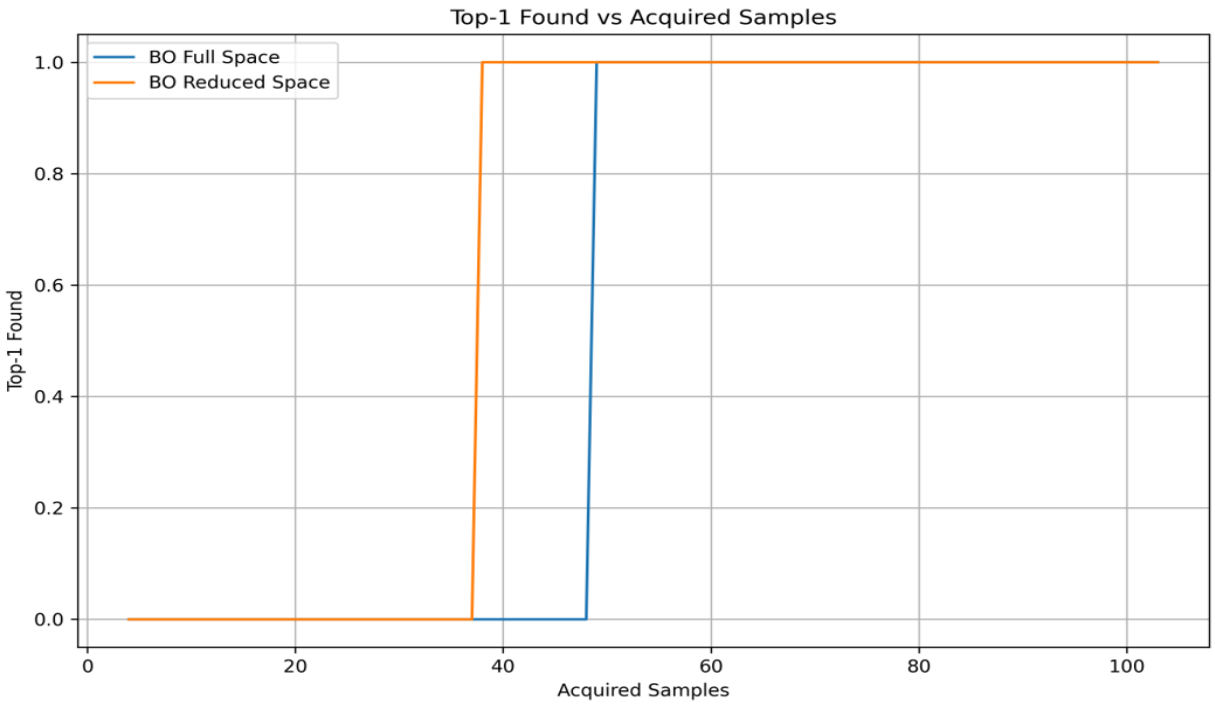

Figure 4:Top-1 Found

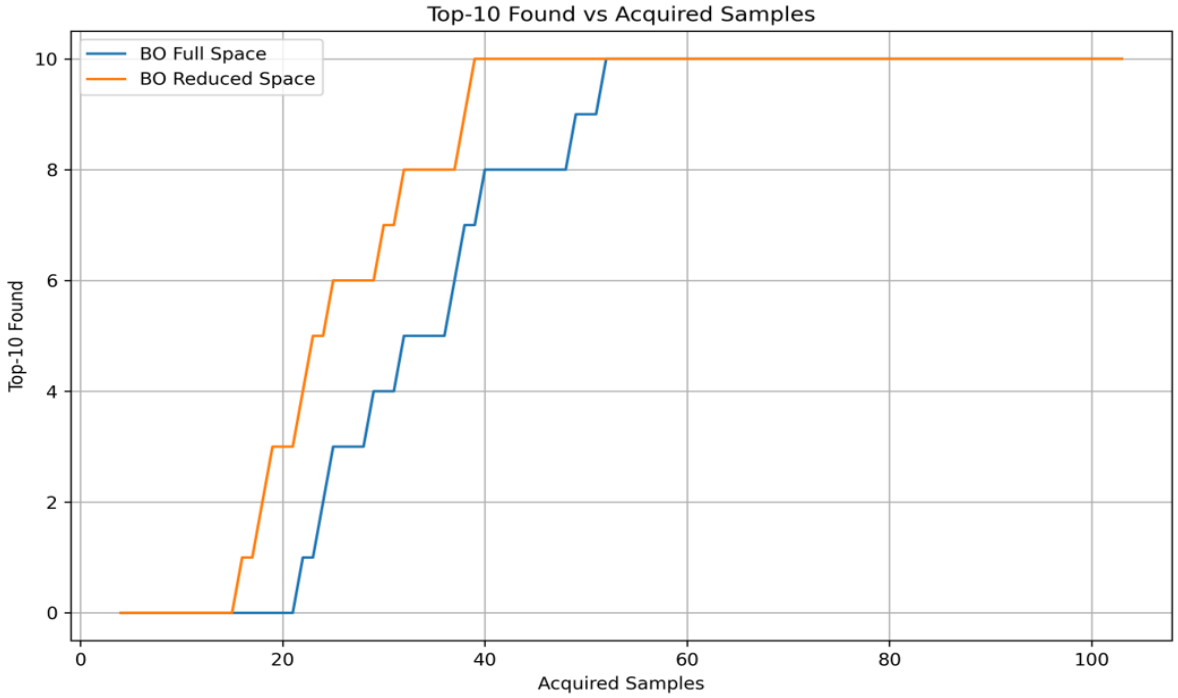

Figure 5:Top-10 Found

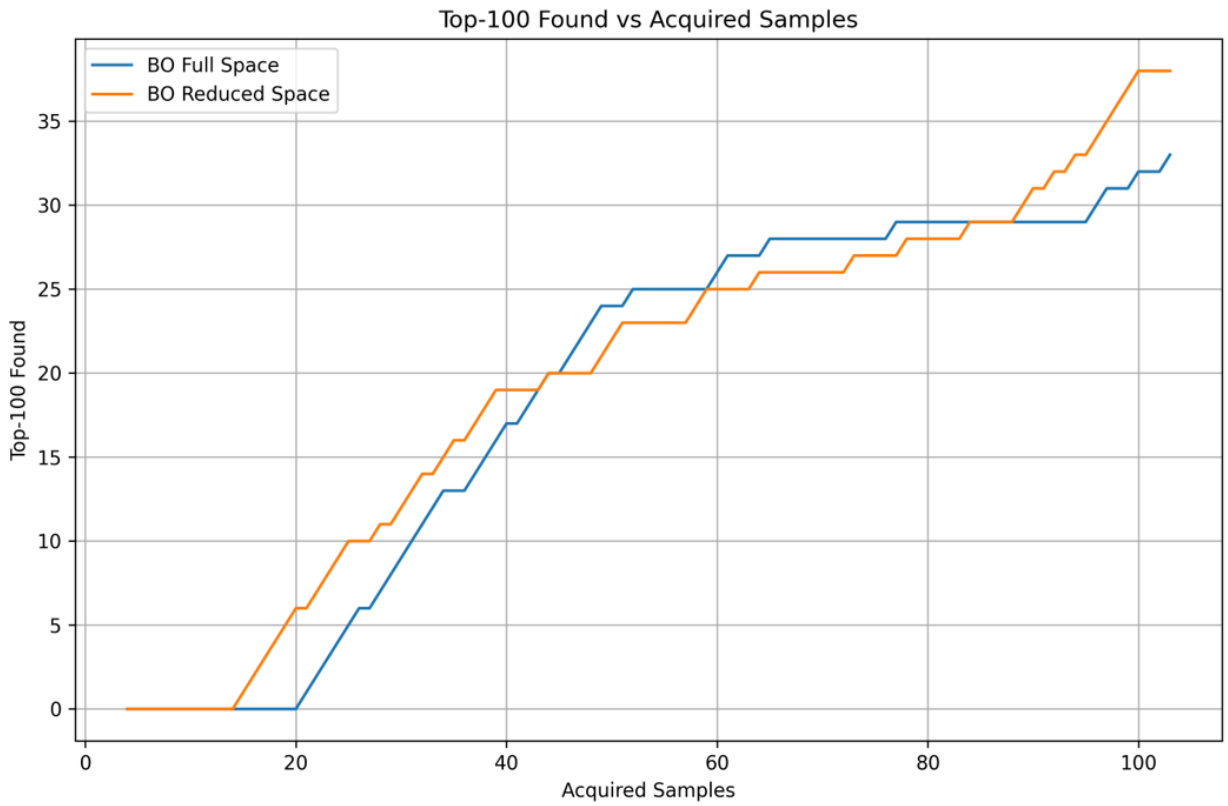

Figure 6:Top-100 Found

A comprehensive comparison of the optimization performance is presented in

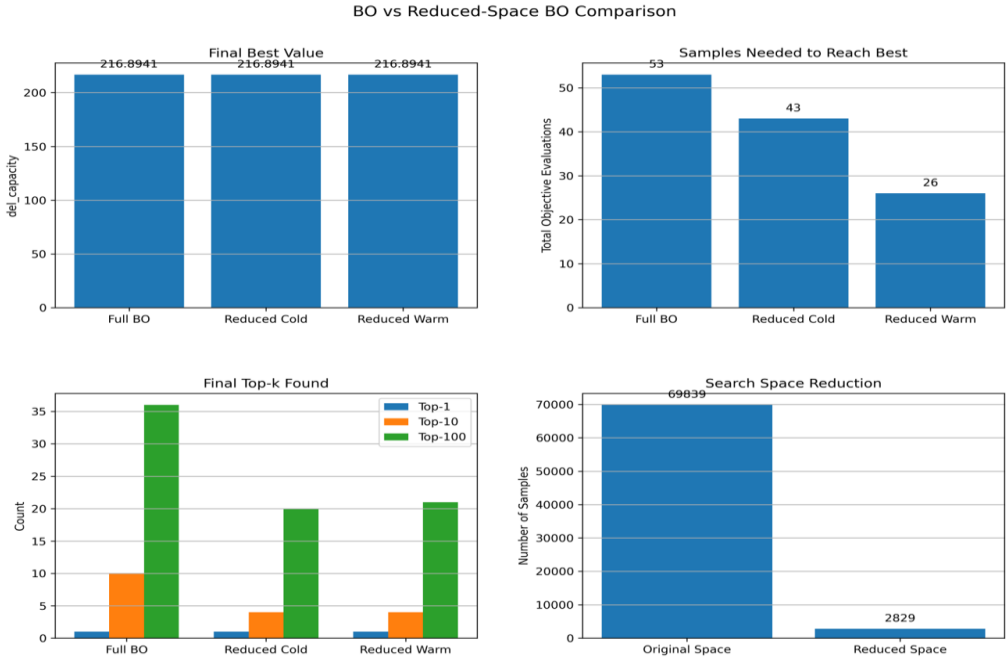

Figure 7. The results show that all approaches achieve comparable final best values, confirming that the reduced design space preserves solution quality.

However, the reduced-space optimization requires significantly fewer evaluations to reach optimal performance. In particular, the warm-start configuration achieves the best solution with substantially fewer evaluations compared to both the full-space and cold-start approaches.

Additionally, the results demonstrate a significant reduction in the size of the search space, highlighting the effectiveness of the proposed Active Learning-based filtering method.

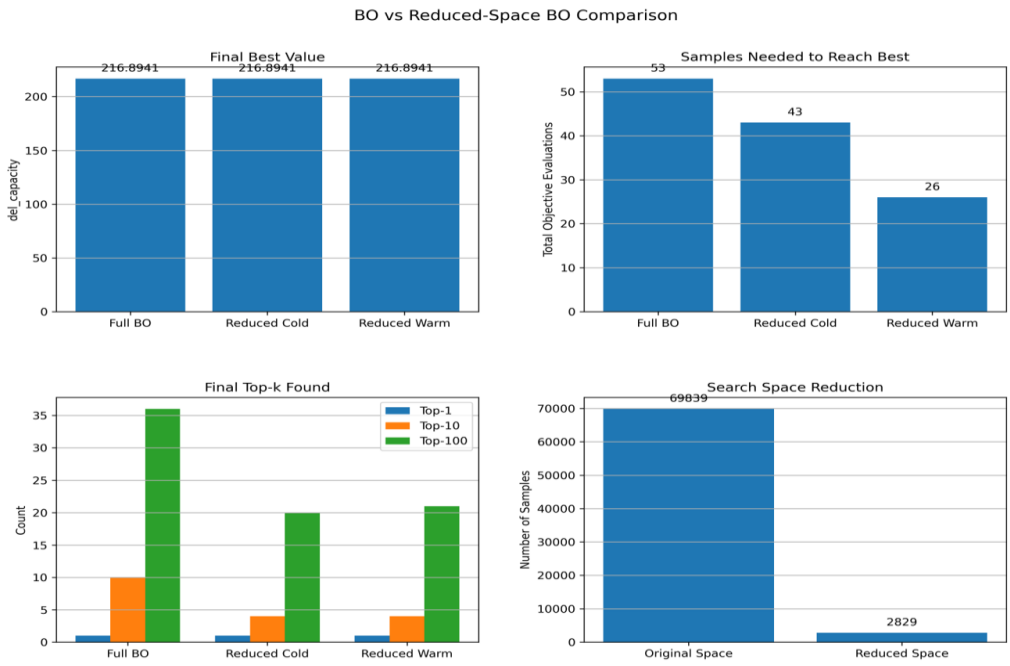

Figure 7: Comparison of Bayesian Optimization performance in the full and reduced design spaces.

6.2 Multi-Objective Results

The multi-objective experiments are conducted on a different dataset from the one used in the single-objective setting. Specifically, while the single-objective experiments are performed on the COFs dataset, the multi-objective experiments are carried out on the Pressure Vessel dataset, which represents a composite material design problem involving multiple competing objectives.

This distinction allows the proposed methodology to be evaluated across different domains and optimization settings, providing insight into its generality and robustness.

The objective in this setting is to approximate the Pareto front and efficiently identify high-quality trade-off solutions. The results compare Bayesian Optimization in the full design space with the proposed reduced-space approach, as well as the impact of cold-start and warm-start initialization strategies.

The results compare Bayesian Optimization applied to the full design space with the proposed reduced-space approach, as well as the impact of different initialization strategies.

As shown in Figure 8, the hypervolume evolution over the optimization process highlights clear differences between the examined approaches. The reduced-space optimization, particularly with warm-start initialization, achieves higher hypervolume values at earlier stages compared to both the cold-start and full-space configurations.

This behavior indicates that the combination of Active Learning-based space reduction and informed initialization allows the optimization process to focus more effectively on high-quality regions of the objective space, leading to faster convergence toward the Pareto front.

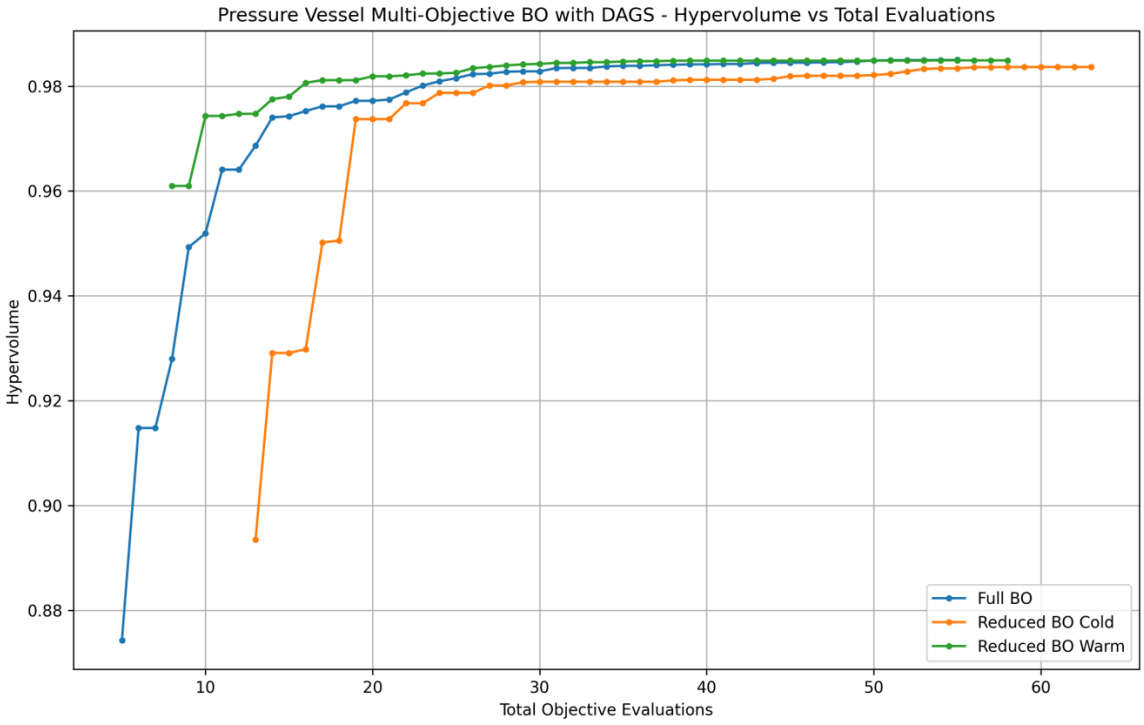

Figure 8: Hypervolume Convergence

The number of true Pareto points discovered during the optimization process provides further insight into the effectiveness of each approach. As illustrated in Figure 9, the warm-start reduced-space method consistently identifies a larger number of Pareto-optimal solutions compared to both the cold-start and full-space configurations.

This demonstrates that initializing the optimization process with informative samples significantly improves the ability to discover high-quality trade-off solutions.

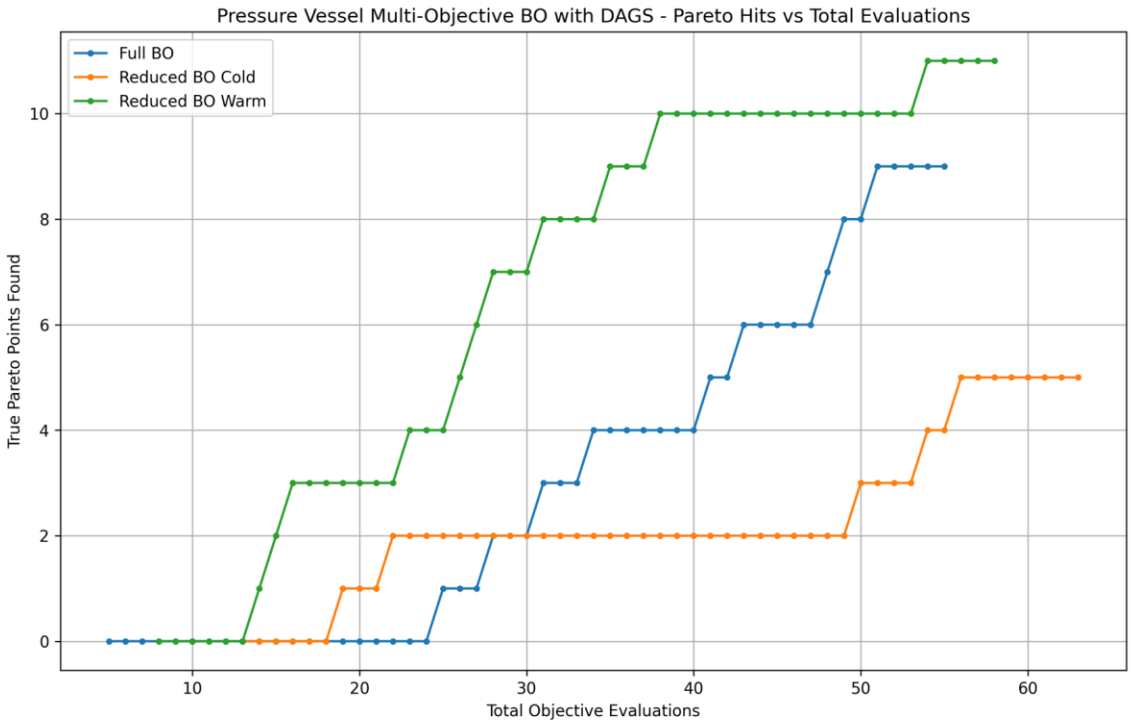

Figure 9: Pareto Hits vs Total Evaluations

A qualitative analysis of the Pareto front approximation is presented in Figure 10. The visualization shows the distribution of Pareto-optimal points identified by each method in the objective space.

The reduced-space approach, especially with warm-start initialization, achieves a broader and more consistent coverage of the Pareto front, closely matching the distribution of the true Pareto-optimal solutions. In contrast, the cold-start approach exhibits more limited exploration and misses several regions of the Pareto front.

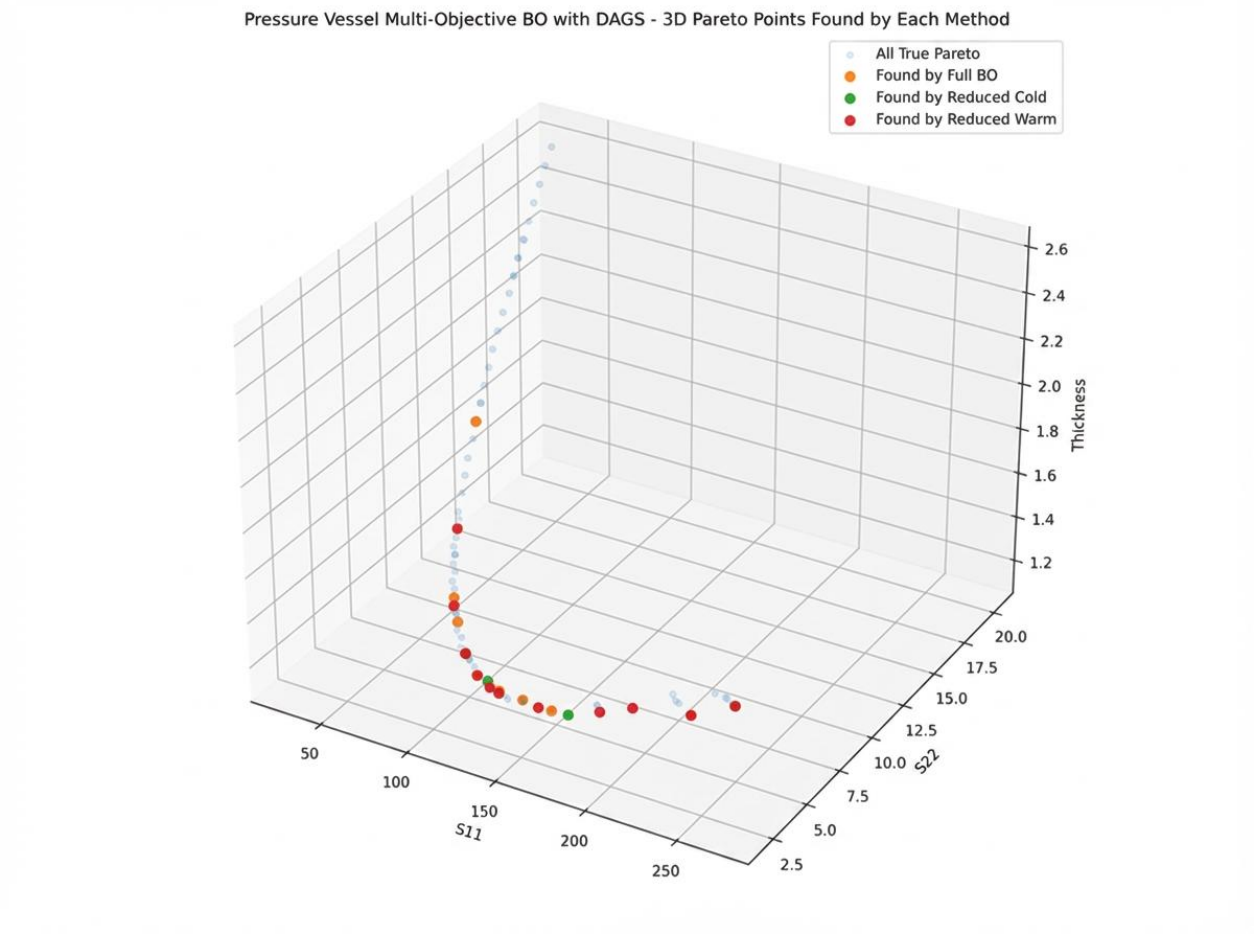

Figure 10: 3D Pareto Points Found

An important aspect of the proposed methodology is its ability to preserve high-quality solutions during the space reduction stage. As shown in Figure 11, a significant portion of the true Pareto-optimal points is retained after applying the DAGS-based filtering.

Although some reduction is expected, the retained subset still captures the essential structure of the Pareto front, ensuring that the optimization process operates on a meaningful and representative subset of the design space.

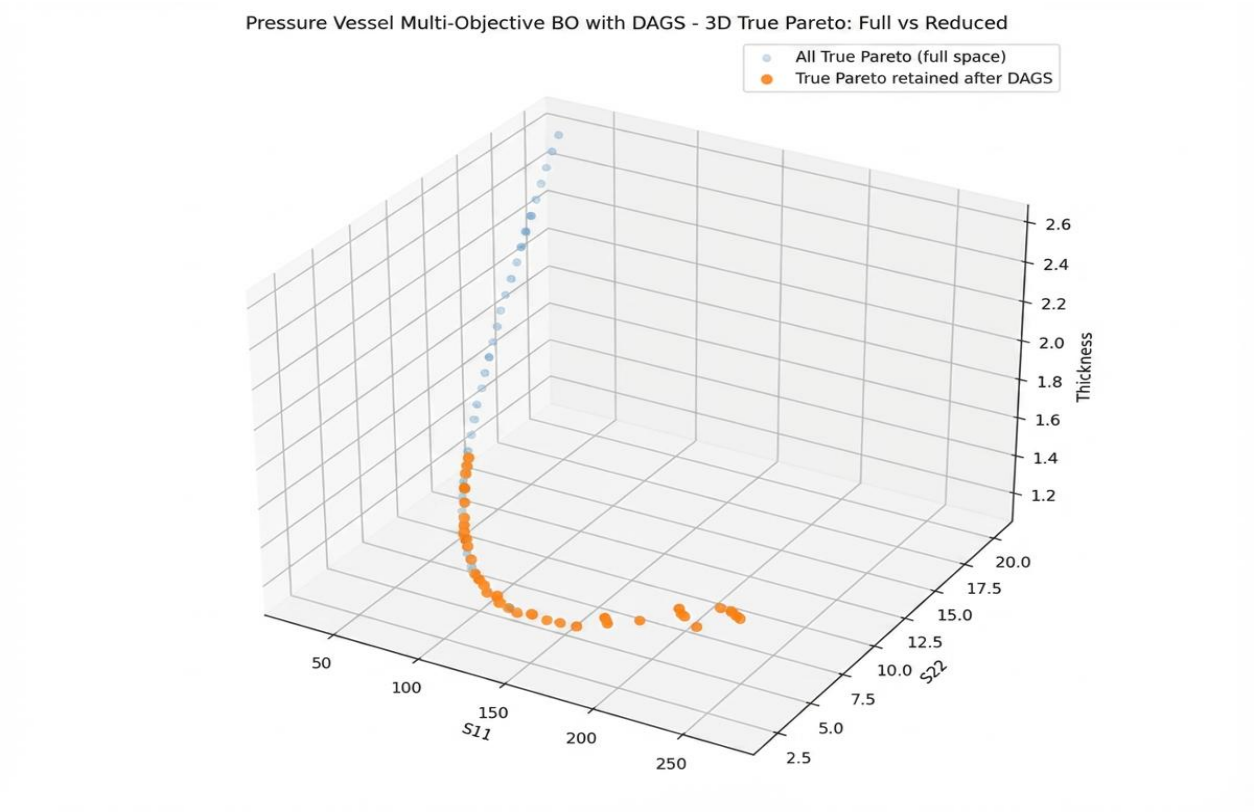

Figure 11: Pareto Retention After DAGS

A summary of the final performance metrics is presented in Figure 12. The results show that all approaches achieve comparable hypervolume values, indicating that the reduced design space preserves solution quality.

However, significant differences are observed in terms of efficiency and solution discovery. The warm-start reduced-space approach identifies more true Pareto-optimal points while requiring fewer evaluations compared to the cold-start and full-space baselines.

This demonstrates that the proposed method successfully balances performance and efficiency, achieving high-quality results with reduced computational cost.

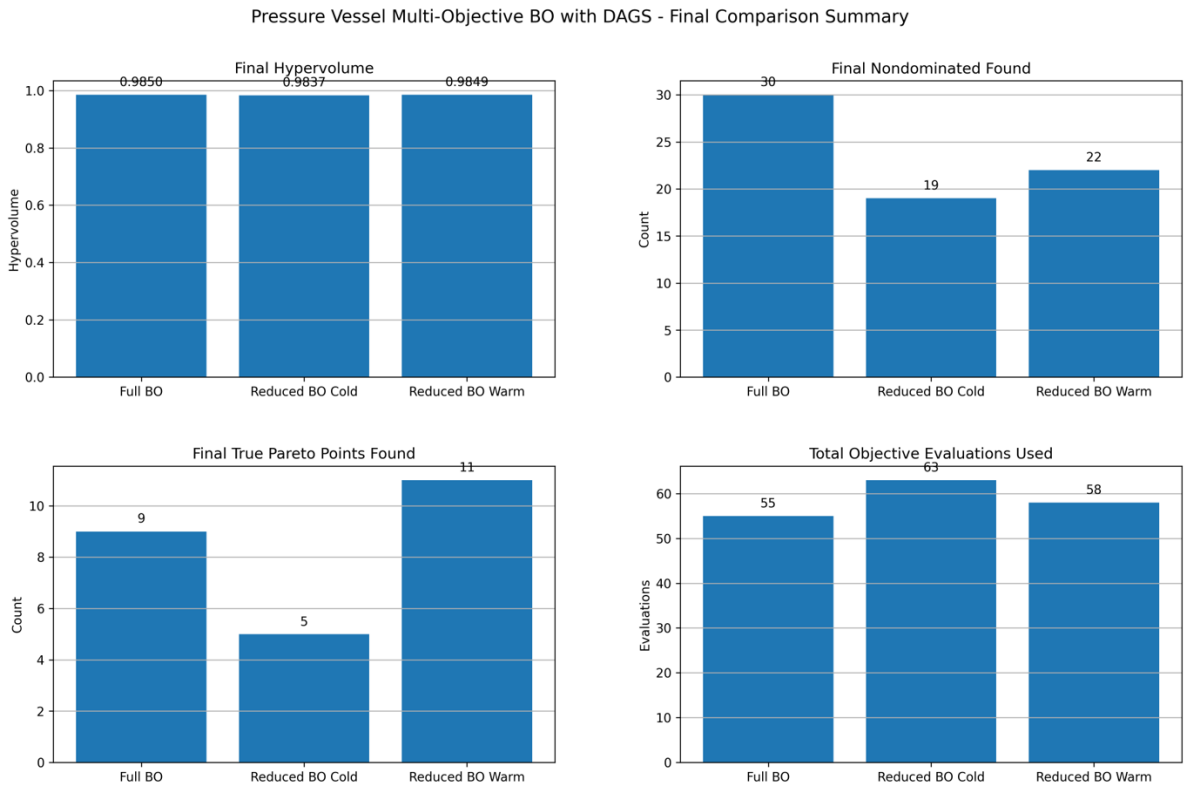

Figure 12: Final Comparison Summary

6.3 Comparative Analysis

This section provides a comprehensive comparison of the proposed methodology against the baseline Bayesian Optimization approach, considering both single-objective and multi-objective settings. The analysis focuses on the impact of design space reduction and initialization strategies on optimization performance, solution quality, and computational efficiency.

A consistent observation across all experiments is that the reduced design space preserves the quality of the final solutions. In both the single-objective and multi-objective cases, the best values and hypervolume metrics achieved by the reduced-space optimization are comparable to those obtained in the full design space. This confirms that the proposed Active Learning-based filtering does not eliminate critical high-performing candidates.

At the same time, the reduced-space approach significantly improves sample efficiency. In the single-objective experiments, the optimization process reaches high-performing solutions with substantially fewer evaluations, particularly when warm-start initialization is used. Similarly, in the multi-objective setting, the reduced-space optimization achieves competitive hypervolume values and identifies Pareto-optimal solutions more rapidly than the full-space baseline.

A key factor contributing to this improvement is the use of warm-start initialization. By leveraging informative samples identified through the DAGS framework, the optimization process begins closer to high-potential regions of the design space. This results in faster convergence and more efficient exploration compared to the cold-start approach, which relies on randomly selected initial samples.

Furthermore, the results demonstrate that design space reduction leads to a more focused and effective search process. By eliminating a large portion of low-performing or irrelevant candidates, the optimization algorithm can concentrate its evaluations on promising regions. This effect is particularly evident in the improved top-k performance and the increased number of Pareto-optimal solutions identified in the multi-objective experiments.

An additional advantage of the proposed approach is its generality. The methodology is shown to perform effectively across different types of design problems, including nanoporous materials such as COFs and composite material design problems such as the Pressure Vessel dataset. This indicates that the combination of Active Learning and Bayesian Optimization can be applied to a wide range of optimization tasks with different characteristics.

Overall, the proposed framework achieves a strong balance between exploration efficiency and solution quality. It enables the identification of high-performing solutions while significantly reducing the computational cost associated with large design spaces. These results highlight the effectiveness of integrating Active Learning-based space reduction with Bayesian Optimization, as well as the importance of informed initialization strategies in complex optimization problems.

7 Discussion

The results presented in this work provide strong evidence for the effectiveness of combining Active Learning-based design space reduction with Bayesian Optimization. The analysis across both single-objective and multi-objective settings highlights several important observations regarding optimization efficiency, solution quality, and the role of initialization strategies.

A key finding of this study is that the proposed space reduction approach is able to significantly decrease the size of the search space without compromising the quality of the final solutions. In the single-objective experiments, the reduced-space optimization consistently achieves the same optimal values as the full-space baseline. Similarly, in the multi-objective setting, comparable hypervolume values are obtained, indicating that the essential structure of the Pareto front is preserved.

This result is particularly important, as it demonstrates that the Active Learning-based filtering process successfully removes low-quality or irrelevant regions of the design space while retaining high-potential candidates. In other words, the reduction process acts as an effective pre-selection mechanism that simplifies the optimization problem without degrading its solution space.

Another important observation concerns the improvement in sample efficiency. Across all experiments, the reduced-space optimization requires significantly fewer evaluations to reach high-performing solutions. This improvement is especially pronounced in the warm-start configuration, where the optimization process is initialized using informative samples obtained from the DAGS framework. By starting closer to high-performing regions, the optimization process avoids unnecessary exploration of low-quality areas and converges more rapidly.

The comparison between cold-start and warm-start initialization further emphasizes the importance of informed starting points in Bayesian Optimization. While cold-start optimization relies on randomly selected initial samples, warm-start initialization benefits

from prior knowledge extracted during the Active Learning stage. This leads to faster convergence, improved top-k performance, and more efficient discovery of Pareto-optimal solutions.

In the multi-objective experiments, the advantages of the proposed approach become even more evident. The reduced-space, warm-start configuration not only achieves competitive hypervolume values but also identifies a larger number of true Pareto-optimal solutions. This indicates that the method is capable of capturing a richer set of trade-offs between competing objectives, which is essential in practical design problems.

An additional insight from the results is the role of search space structure in optimization performance. The ability to restrict the optimization process to a smaller, high-quality subset of the design space reduces the complexity of the problem and allows the surrogate model to better approximate the objective functions. This leads to more accurate acquisition decisions and ultimately improves the overall efficiency of the optimization process.

Despite these promising results, some limitations of the proposed approach should be considered. The effectiveness of the space reduction stage depends on the quality of the Active Learning model and the chosen threshold for defining promising regions. If the filtering process is too aggressive, there is a risk of excluding potentially optimal solutions. Although this issue is mitigated in the present work, it highlights the importance of carefully designing the reduction strategy.

Furthermore, the approach introduces additional computational overhead due to the Active Learning stage. While this cost is offset by the reduction in optimization evaluations, the overall efficiency depends on the balance between these two components. Future work could explore more lightweight or adaptive filtering strategies.

Overall, the results demonstrate that the integration of Active Learning and Bayesian Optimization provides a powerful framework for tackling complex optimization problems with large design spaces. The ability to reduce the search space, leverage informative initialization, and efficiently explore high-potential regions makes the proposed approach particularly suitable for real-world applications in materials design and beyond.

References

[1] Frazier, P. I. (2018). A tutorial on Bayesian optimization. *arXiv preprint arXiv:1807.02811*.

[2] Park, K., Song, C., Park, J., & Ryu, S. (2023). Multi-objective Bayesian optimization for the design of nacre-inspired composites: optimizing and understanding biomimetics through AI. *Materials Horizons*, *10*(10), 4329-4343.

[3] Loutas, T., Oikonomou, A., & Rekatsinas, C. (2025). Bio-inspired discontinuous composite materials with a machine learning optimized architecture. *Composite Structures*, *351*, 118597.

[4] Di Fiore, F., Nardelli, M., & Mainini, L. (2023). Active learning and bayesian optimization: A unified perspective to learn with a goal. *arXiv preprint arXiv:2303.01560*.

[5] Hvarfner, C., Hellsten, E., Hutter, F., & Nardi, L. (2023). Self-correcting bayesian optimization through bayesian active learning. *Advances in Neural Information Processing Systems*, *36*, 79173-79199.

[6] Snoek, J., Larochelle, H., & Adams, R. P. (2012). Practical bayesian optimization of machine learning algorithms. *Advances in neural information processing systems*, *25*.

[7] Khatamsaz, D., Vela, B., Singh, P., Johnson, D. D., Allaire, D., & Arróyave, R. (2023). Bayesian optimization with active learning of design constraints using an entropy-based approach. *npj Computational Materials*, *9*(1), 49.

[8]Khatamsaz, D., Vela, B., Singh, P., Johnson, D. D., Allaire, D., & Arróyave, R. (2022). Multi-objective materials bayesian optimization with active learning of design constraints: Design of ductile refractory multi-principal-element alloys. *Acta Materialia*, *236*, 118133.

[9] Ozaki, R., Ishikawa, K., Kanzaki, Y., Takeno, S., Takeuchi, I., & Karasuyama, M. (2024, March). Multi-objective Bayesian optimization with active preference learning. In *Proceedings of the AAAI conference on artificial intelligence* (Vol. 38, No. 13, pp. 14490-14498).

[10] Martín, L. A., & Garrido-Merchán, E. C. (2021). Many objective bayesian optimization. *arXiv preprint arXiv:2107.04126*.

[11] Zhang, Y., Apley, D. W., & Chen, W. (2020). Bayesian optimization for materials design with mixed quantitative and qualitative variables. *Scientific reports*, *10*(1), 4924.

[12] Zuo, Y., Qin, M., Chen, C., Ye, W., Li, X., Luo, J., & Ong, S. P. (2021). Accelerating materials discovery with Bayesian optimization and graph deep learning. *Materials Today*, *51*, 126-135.

[13] Zhang, H., Chen, W., Iyer, A., Apley, D. W., & Chen, W. (2022). Uncertainty-aware mixed-variable machine learning for materials design. *Scientific reports*, *12*(1), 19760.

[14] Kotthoff, L., Wahab, H., & Johnson, P. (2021). Bayesian optimization in materials science: a survey. *arXiv preprint arXiv:2108.00002*.

[15] Theocharis, I., Krokidas, P., Gkatsis, V., & Giannakopoulos, G. (2024, September). Multi-fidelity bayesian optimization for efficiently sampling the design space of functionalized nanoporous materials. In *Proceedings of the 13th Hellenic Conference on Artificial Intelligence* (pp. 1-4).

[16] Côté, A. P., Benin, A. I., Ockwig, N. W., O'Keeffe, M., Matzger, A. J., & Yaghi, O. M. (2005). Porous, crystalline, covalent organic frameworks. *science*, *310*(5751), 1166-1170.

[17] Chitturi, S. R., Ramdas, A., Wu, Y., Rohr, B., Ermon, S., Dionne, J., ... & Ratner, D. (2024). Targeted materials discovery using Bayesian algorithm execution. *npj Computational Materials*, *10*(1), 156.

[18] Liang, Q., Gongora, A. E., Ren, Z., Tiihonen, A., Liu, Z., Sun, S., ... & Buonassisi, T. (2021). Benchmarking the performance of Bayesian optimization across multiple experimental materials science domains. *npj Computational Materials*, *7*(1), 188.

[19] Tang, H., & Jiang, J. (2022). Active learning boosted computational discovery of covalent–organic frameworks for ultrahigh CH4 storage. *AIChE Journal*, *68*(11), e17856.

[20] Krokidas, P., Gkatsis, V., Theocharis, J., & Giannakopoulos, G. (2025). Navigating materials design spaces with efficient Bayesian optimization: a case study in functionalized nanoporous materials. *Digital Discovery*, *4*(12), 3753-3763.

[21] Jin, Y., & Kumar, P. V. (2023). Bayesian optimisation for efficient material discovery: a mini review. Nanoscale, 15(26), 10975-10984.

[22] Qian, X., Yoon, B. J., Arróyave, R., Qian, X., & Dougherty, E. R. (2023). Knowledge-driven learning, optimization, and experimental design under uncertainty for materials discovery. Patterns, 4(11).

[23] Arróyave, R., Khatamsaz, D., Vela, B., Couperthwaite, R., Molkeri, A., Singh, P., ... & Allaire, D. (2022). A perspective on Bayesian methods applied to materials discovery and design. MRS communications, 12(6), 1037-1049.

[24] Lei, B., Kirk, T. Q., Bhattacharya, A., Pati, D., Qian, X., Arroyave, R., & Mallick, B. K. (2021). Bayesian optimization with adaptive surrogate models for automated experimental design. Npj Computational Materials, 7(1), 194.

[25] Wang, Y., Iyer, A., Chen, W., & Rondinelli, J. M. (2020). Featureless adaptive optimization accelerates functional electronic materials design. Applied Physics Reviews, 7(4).

[26] Chen, J., Ou, P., Chang, Y., Zhang, H., Li, X. Y., Sargent, E. H., & Chen, W. (2025, August). Adaptive Uncertainty-Aware Deep Learning for Materials Discovery With High-Dimensional Design Inputs. In International Design Engineering Technical Conferences and Computers and Information in Engineering Conference (Vol. 89237, p. V03BT03A004). American Society of Mechanical Engineers.

[27] Chen, J., Ou, P., Chang, Y., Zhang, H., Li, X. Y., Sargent, E. H., & Chen, W. (2026). Materials Discovery Using Uncertainty-Aware Constrained Bayesian Optimization With Representation Learning of High-Dimensional Inputs. Journal of Mechanical Design, 148(2), 021707.

[28] Chowdhury, C. (2024). Bayesian optimization for efficient prediction of gas uptake in nanoporous materials. ChemPhysChem, 25(16), e202300850.

[29] Mamun, O., Bause, M., & Ebna Hai, B. S. M. (2025). Accelerated development of multi-component alloys in discrete design space using Bayesian multi-objective optimisation. Machine Learning: Science and Technology, 6(1), 015001.

[30] Solomou, A., Zhao, G., Boluki, S., Joy, J. K., Qian, X., Karaman, I., ... & Lagoudas, D. C. (2018). Multi-objective Bayesian materials discovery: Application on the discovery of precipitation strengthened NiTi shape memory alloys through micromechanical modeling. Materials & Design, 160, 810-827.

[31] Wei, Q., Wang, Y., Yang, G., Li, T., Yu, S., Dong, Z., & Zhang, T. Y. (2025). Discovering novel lead-free solder alloy by multi-objective Bayesian active learning with experimental uncertainty. npj Computational Materials, 11(1), 10.

[32] Coelho, R. C., Alves, A. F. C., Pires, T. N., & Pires, F. A. (2025). A composite Bayesian optimisation framework for material and structural design. Computer Methods in Applied Mechanics and Engineering, 434, 117516.

[33] Rajabi-Kochi, M., Mahboubi, N., Gill, A. P. S., & Moosavi, S. M. (2025). Adaptive representation of molecules and materials in Bayesian optimization. Chemical Science, 16(13), 5464-5474.

[34] Wang, K., & Dowling, A. W. (2022). Bayesian optimization for chemical products and functional materials. Current Opinion in Chemical Engineering, 36, 100728.

[35] Yao, Y., & Oberhofer, H. (2024). Designing building blocks of covalent organic frameworks through on-the-fly batch-based Bayesian optimization. The Journal of Chemical Physics, 161(7).

[36] Yaghi, O. M. (2020). The reticular chemist. Nano Letters, 20(12), 8432-8434.

[37] Lyu, H., Ji, Z., Wuttke, S., & Yaghi, O. M. (2020). Digital reticular chemistry. Chem, 6(9), 2219-2241.