Multi-Objective Evolutionary Algorithms For Product Design

Presented by:

Bilal Hasan Aslan

Submitted to the Department of Computer Science at the University of Cape Town in fulfillment of the academic requirements for a Master of Science by coursework and dissertation degree in Computer Science

June 4, 2024

Declaration

1. I understand what plagiarism is.
2. This dissertation titled, 'Multi-Objective Evolutionary Algorithms For Product Design' is my own work.
3. The APA7 convention for citation and referencing has been followed in this dissertation. Every contribution and quotation from the work of other individuals has been appropriately attributed, cited, and referenced.

Signature: 

Bilal Hasan Aslan

Date: 11 February 2024

Acknowledgments

I would like to express my heartfelt gratitude to my supervisor, Prof. Geoff Nitschke, for his unwavering guidance and support throughout the duration of my research. His expertise, valuable insights, and encouragement have been instrumental in shaping the direction and success of this dissertation.

List of publications

Papers related to this thesis

1. *A Computational Method to Support Chemical Product Design Based on Multi-objective Optimisation and Graph Transformers.* da Silva, F.S.C., Aslan, B. and Nitschke, G., 2023, July. In ALIFE 2023: Ghost in the Machine: Proceedings of the 2023 Artificial Life Conference. MIT Press.

2. *Multi-objective Evolution for Automated Chemistry.* Aslan, B., da Silva, F.S.C. and Nitschke, G., 2023, December. In 2023 IEEE Symposium Series on Computational Intelligence (SSCI) (pp. 152-157). IEEE.

3. *Multi-Objective Evolution for Chemical Product Design.* Aslan, B., da Silva, F.S.C. and Nitschke, G., 2024, July. In GECCO 2024: Proceedings of the Genetic and Evolutionary Computation Conference.

Additional Papers, not related to this thesis

1. *Automating Robot Design with Multi-Level Evolution.* Aslan, B., Nitschke, G. and et al, 2024, July. In WCCI 2024: Proceedings of the World Congress on Computational Intelligence.

2. *Morpho-Material Evolution for Automated Robot Design.* Aslan, B., Nitschke, G., 2024, July. In GECCO 2024: Proceedings of the Genetic and Evolutionary Computation Conference.

Abstract

Identifying chemical compounds with optimal properties for specific applications presents a fundamental challenge in materials science. Traditional methods, based on trial-and-error, are inefficient and costly. This thesis introduces an innovative integration of Computational Chemistry and Machine Learning (ML) with Evolutionary Multi-Objective Optimisation (EMOO) techniques to streamline compound design. This approach automates the design process by leveraging ML to accurately predict compound properties and using EMOO to select compounds that meet various criteria. The significance of this work lies in its potential to transform the traditional development process, facilitating the creation of chemical products that fulfill multiple objectives more efficiently. This study not only demonstrates the synergy between advanced ML and optimisation techniques but also presents a comprehensive comparison of the Multi-Objective Covariance Matrix Adaptation Evolution Strategy (MO-CMA-ES) and Non-dominated Sorting Genetic Algorithm II (NSGA-II), including two novel meta-heuristics for enhanced molecular exploration. Our findings reveal that MO-CMA-ES, especially when combined with an extended search meta-heuristic, excels in exploring molecular spaces, establishing it as a preferred method for compound synthesis. This research promises to accelerate compound development specifically for detergent compounds, offering significant implications for product design across various industries.

Contents

List of Figures

List of Tables

Abbreviations

BERT	Bidirectional Encoder Representations from Transformers
CMA-ES	Covariance Matrix Adaptation Evolution Strategy
CV	Computer Vision
EAs	Evolutionary Algorithms
ETKDG	Experimental-Torsion basic Knowledge Distance Geometry with Merck Molecular Force Field optimisation
GANs	Generative adversarial networks
GPT	Generative Pre-Trained Transformer
InChI	International Chemical Identifier
MACCS	Molecular ACCess Systems keys fingerprint
MDS	Multidimensional Scaling
ML	Machine learning
MO-CMA-ES	Multi-Objective Covariance Matrix Adaptation Evolution Strategy
MOEAs	Multi-Objective Evolutionary Algorithms
MOO	Multi-Objective optimisation
MRL	Molecular Representation Learning
NLP	Natural Language Processing
NSGA	Non-dominated Sorting Genetic Algorithm
NSGA-II	Non-dominated Sorting Genetic Algorithm II
PSO	Particle Swarm optimisation

QSAR	Quantitative Structure-Activity Relationships
SELFIES	Self-referencing Embedded Strings
SMILES	Simplified Molecular Input Line Entry System
VAEs	Variational Autoencoders
ViT	Vision Transformer
XlogP	Extended Partition Coefficient

Chapter 1

Introduction

Chemical product design has experienced a pivotal transformation in recent years, shifting from conventional trial-and-error synthesis methods to the adoption of sophisticated computational techniques. These techniques, rooted in the identification and optimisation of compounds for specific chemical properties, have redefined the design landscape, hastening cycle times and streamlining design iterations (Chen & et al., 2018; Schneider, 2018; Winter et al., 2019b).

The transition from traditional trial-and-error synthesis to computer-aided product design has profoundly influenced the field of material design. This paradigm shift, driven by an iterative methodology, involves continuous refinement of compounds to achieve specific chemical attributes. The advantages of this methodology are evident across various applications, from solvents and ionic liquids to polymers and medications (Ng & Gani, 2019). Beyond merely synthesizing new products, it addresses overarching goals such as reducing aquatic toxicity and enhancing synthetic accessibility (Mayr et al., 2016; Winter et al., 2019b).

Yet, the chemical design space remains vast and intricate. Some estimates suggest the presence of over 10^{200} organic compounds, creating a significant computational challenge (Reymond, 2015). Bridging this vastness with computational agility has been possible due to the combination of computational chemistry and Machine learning (ML). Evolutionary Algorithms (EAs), in particular, have become powerful tools, optimizing the search within this colossal space (Keith & et al., 2021). This synergy is especially evident in areas like *de novo* molecular design, spawning breakthroughs across fields from drug discovery to materials science (Bender & Cortes-Ciriano, 2021; Paul, 2021; Tkatchenko, 2020).

Recent research in generative deep learning has added further momentum. Techniques ranging from auto-encoders (Gómez-Bombarelli & et al., 2018; Jin et al., 2020) to graph-based methods that encode complex molecular structures (De Cao & Kipf, 2018; Lim et al., 2020; Samanta & et al., 2019) have seen success. Remarkably, deep recurrent neural networks now adeptly generate chemically viable molecules (Yuan & et al., 2020).

Yet, challenges persist. The confluence of generative techniques and expansive molecular datasets (Kim et al., 2019a) emphasizes the enduring dependence on human expertise (Curtarolo & et al., 2013; Pyzer-Knapp & et al., 2015), especially in the filtration and selection of optimal solution candidates with targeted properties as shown in Figure 1.1. The automation of this process remains a crucial frontier.

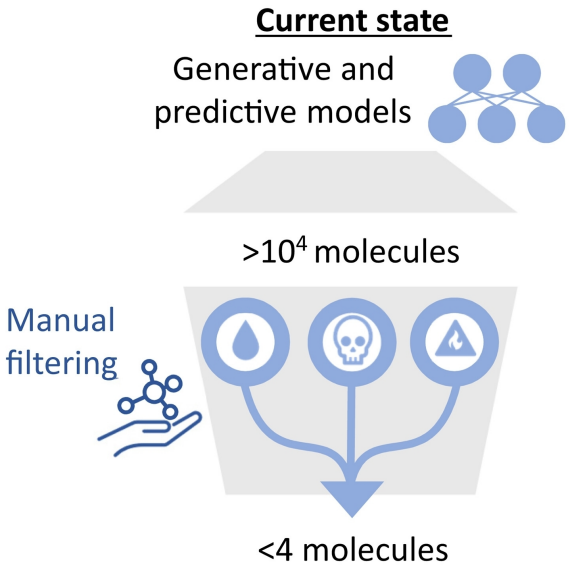


Figure 1.1: Illustrates the workflow of molecule selection using generative and predictive models in the current state of the field. Initially, these models generate a large set ($> 10^4$) of potential molecular candidates. Following this, a process of manual filtering is applied, significantly reducing the number of candidates to fewer than four viable molecules. This figure highlights the disparity between the initial high-throughput computational prediction and the intensive manual selection required to identify the most promising molecules for further development. (Bilodeau et al., 2022a).

Researchers have demonstrated the efficacy of evolutionary algorithms in achieving competitive outcomes within the realm of chemical product design (Jensen, 2019; Kwon & et al., 2021; Leguy & al., 2009; Varela & Santos, 2022; Yoshikawa & et al., 2018), particularly when the chemical design space is defined through specific molecular encodings. Notably, these algorithms prioritize a single property, rather than multiple properties for selection and mutation processes. It was quickly realized that molecules have many properties with trade-offs, such as molecular weight, which impacts solubility and absorption; molecular complexity, affecting synthesis and stability; XlogP, influencing bio-availability; and toxicity, which is critical for safety profiles. Balancing

these properties while keeping the structure of the compound similar to what is known to work is crucial for the successful design of new chemical entities ("similar structure, similar property" (Brown, 2009)), presenting a complex optimisation challenge that necessitates a multi-objective approach.

The multifaceted nature of molecular property optimisation necessitates algorithms capable of effectively navigating the trade-offs between various objectives. MO-CMA-ES (Loshchilov & Hutter, 2016) and NSGA-II (Deb et al., 2002) are particularly well-suited for this task, having been specifically designed for multi-objective optimisation challenges. MO-CMA-ES is adept at adapting its search strategy to the contours of complex optimisation landscapes, a feature that is indispensable when seeking to optimize the interrelated properties of molecules. NSGA-II, with its fast sorting algorithm and crowding distance mechanism, ensures a diverse and well-distributed representation of solutions along the Pareto front (set of non-dominated solutions), making it an excellent tool for discerning the optimal balance between competing molecular characteristics.

The integration of ML techniques with evolutionary algorithms is increasingly becoming a possible approach in the fields of property prediction and optimisation. While traditional methods of property prediction are typically deterministic, modern strategies are being increasingly characterized by probabilistic and data-driven approaches. These contemporary methodologies capitalize on the proliferation of data, utilizing extensive datasets to infer predictions and insights, which would be challenging to derive manually. Today's ML models, inherently adaptive, are fine-tuned to predict specific properties like environmental impact. As they access more data, these models continuously refine their predictions, demonstrating the iterative essence of ML. Integrating these approaches not only promises a swifter search process but also enhanced precision in property evaluations (Brown & et al., 2004; Le & Winkler, 2016).

In response to escalating environmental sustainability concerns, the development of automated methods for chemical product design has escalated significantly. To address this, we have embarked on an exploration, leveraging the fusion of deep learning and evolutionary Multi-Objective optimisation (MOO). This novel approach capitalizes on Geometric Deep Learning for molecular property prediction and Evolutionary MOO for navigating the chemical design space towards compounds that simultaneously optimize multiple objectives, such as minimizing aquatic toxicity and maximizing synthetic accessibility.

A cornerstone of this methodology is our strategy to identify the initial pool of candidate compounds, marking the boundaries of our exploration. Coupled with the careful selection of foundational 'seed' compounds, our approach aims for a holistic sweep of the design space, revealing promising compound candidates.

This study primarily focuses on the practical application of our methodology for detergent products: crafting a selection (filtering) algorithm tailored for molecular datasets. Operating under the guiding principle that molecular similarities typically result in analogous behaviors (Mitchell, 2014), our research commences with a foundational reference seed compound. From this starting point, we aim to curate solution candidates that meet the specified property criteria.

In summary, this research not only aspires to refine molecular design's efficiency but also envisions the interplay of ML and evolutionary algorithms in the chemical domain.

1.1 Motivation

The transformative era of chemical product design, led by computational techniques, has revolutionized our approach to molecule synthesis and optimisation. However, while computational methods have streamlined certain processes, they have also unveiled a new array of challenges and gaps that must be addressed to fully realize the potential of this domain.

A primary challenge lies in the vastness and complexity of the chemical design space. With an estimated 10^{200} organic compounds (Reymond, 2015), the task of efficiently navigating and optimizing within this space is monumental. While computational techniques offer promising solutions, the process's manual aspects, particularly the filtration and selection of optimal solution candidates, remain a hindrance (Curtarolo & et al., 2013; Pyzer-Knapp & et al., 2015). The increasing dependence on human expertise limits the scalability and speed of the design process.

Moreover, the increasing environmental concerns underscore the urgency of refining our methodologies. The detrimental impacts of certain chemical products on ecosystems, like aquatic toxicity, have elevated the demand for sustainable and environment-friendly compounds. Efficiently navigating the design space to identify such compounds, without incurring excessive costs or time delays, remains a significant challenge.

Furthermore, the evolving landscape of ML and EAs presents both an opportunity and a challenge. The synergy of these techniques can show remarkable potential, but the path to optimal and harmonized integration and choice of methods remains nebulous. As industries globally move towards automation and computational solutions, the pressure to refine these integrations amplifies.

Given these challenges, our motivation is clear. We aim to devise a method that combines the strengths of deep learning and EAs, and find better EAs approach to address the gaps in the current design process. By crafting an efficient filtering algorithm for molecular datasets, we endeavor to reduce the dependence on manual processes, accelerate the design timeline, and push the boundaries of what is possible in chemical product design.

In the broader context, our research holds the potential to drive significant advancements in diverse industries – from pharmaceuticals to sustainable products. It offers a blueprint for a more efficient, cost-effective, and environmentally-conscious future in chemical product design. By laying the foundation for the next era of innovation, we aspire to bridge the gap between computational promise and practical reality.

1.2 Research Questions

The core of this research interrogates the efficacy of combining advanced computational techniques, specifically Multi-Objective Evolutionary Algorithms (MOEAs) and Molecular Property Prediction Models, in refining the molecular optimisation process for chemical product design. This investigation is structured around a primary question, supported by secondary questions that delve into comparative analyses and methodological efficacies, as delineated in sections pertaining to motivation (1.1) and methodological frameworks (3.1, 3.3, 3.4).

In detail, this research undertakes a comparative analysis of the exploration capabilities and solution quality provided by the Multi-Objective Covariance Matrix Adaptation Evolution Strategy (MO-CMA-ES) and the Non-dominated Sorting Genetic Algorithm II (NSGA-II), as discussed in Sections 3.3 and 3.4 respectively. This examination seeks to answer secondary questions 2.1 and 2.2, focusing on understanding the intricate dynamics of exploration and solution quality when employing MO-CMA-ES or NSGA-II alongside meta-heuristics like direct correlation and extended search (Section 3.2), specifically within the context of optimizing product design.

1. **Primary Research Question:** How do advanced computational methodologies, particularly Evolutionary Optimisation Algorithms combined with Molecular Property Prediction Models, enhance the efficiency and accuracy of molecular optimisation in chemical product design?
2. **Secondary Research Questions:**
    - 2.1. How do the exploration capabilities of Multi-Objective Covariance Matrix Adaptation Evolution Strategy (MO-CMA-ES) and Non-Dominated Sorting Genetic Algorithm II (NSGA-II) compare in the context of identifying optimally designed molecules in molecule space with complex landscapes?
    - 2.2. Between MO-CMA-ES and NSGA-II, which approach yields superior solution sets for molecular optimisation, assessed through both quantitative metrics and qualitative analysis?

The evaluation of the exploratory effectiveness by testing both MO-CMA-ES and NSGA-II, alongside innovative meta-heuristic approaches (Section 3.2), will address the secondary research question 2.1. The analysis of solution sets generated by these MOEAs, in terms of both their qualitative and quantitative aspects, will cater to secondary research question 2.2. Such examinations are designed to pinpoint the most efficacious combination of methodologies and meta-heuristics for exploring and optimizing the molecular space, thereby yielding insights pertinent to the primary research question.

Furthermore, the precision of Molecular Property Prediction models will undergo a thorough evaluation (5.1), as the accuracy of these models significantly enhances the capabilities of MOEAs. This improvement directly feeds into answering the primary research question, illustrating the pivotal role of predictive accuracy in the optimisation process.

1.3 Contributions

The contributions of this research to the field of product design, particularly through the integration of advanced computational methods, are multifaceted and profound. These contributions enhance the practical methodologies employed in the development of innovative products. The detailed contributions are articulated as follows:

1. This study conducts a comparative analysis of two prominent MOO algorithms, the MO-CMA-ES and the NSGA-II, supplemented by an evaluation of various meta-heuristic approaches as detailed in Section 5.2. This comprehensive examination elucidates the adaptability, efficiency, and scalability of MOEAs within complex and dynamic search spaces. The insights derived from this analysis are critical for the ongoing evolution of product design methodologies, providing a method choice to better navigate the complexities of MOO problems in molecular space. This contribution is pivotal in enhancing the capability of chemists to address more sophisticated and nuanced molecule optimisation challenges, thereby pushing the boundaries of what can be achieved in product development.

2. The research identifies and articulates the conditions under which each algorithm outperforms the other, providing a nuanced framework to guide practitioners in selecting the most appropriate computational strategy for their specific design contexts. This strategic guidance is grounded in empirical evidence and detailed analysis, ensuring that the selection of algorithms is not only tailored to the unique requirements of a project but also optimizes the chances of successful outcomes. Such methodological advancements significantly contribute to the decision-making processes in molecule design, facilitating a more informed and strategic approach to the integration of computational methods.

3. Furthermore, by integrating algorithmic objectives with the predictive insights afforded by Molecule Property Prediction Models, the study forges a novel interdisciplinary approach to product design. This innovative methodology not only expands the computational toolkit available to designers but also incorporates a forward-looking perspective into the design process. The capability to predict and utilize molecule properties in product development represents a significant leap forward, offering the potential to achieve more targeted and sophisticated design solutions. This aspect of the research underscores the transformative power of combining computational chemistry and ML in product design, setting a new standard for precision and innovation in the field.

Overall, the contributions of this research are poised to have an impact on both the theoretical frameworks and practical approaches within product design. By offering a deep dive into the capabilities and applications of advanced computational methods, this study not only advances our understanding of complex optimisation challenges but also provides actionable insights for leveraging these technologies in the creation of novel and effective products. As such, it represents a valuable addition to the corpus of knowledge in the field, serving as a cornerstone for future research and development efforts aimed at harnessing computational power to drive innovation in product design.

1.4 Overview

The structure of this thesis is methodically organized into several chapters, each designed to progressively address the research questions and objectives outlined in the introduction. The thesis is structured as follows:

Chapter 2 - Literature Review: This chapter is segmented into four pivotal sections. The initial section delves into the foundational aspects of molecules and molecule optimisation, setting the stage for subsequent discussions on MOEAs and molecular property prediction models. The subsequent sections critically review the current literature on MOEAs and molecular property prediction models, respectively. The final section identifies the research gap, offering a critical analysis of existing studies and their contributions to the field, thereby delineating the research's relevance and potential impact.

Chapter 3 - Methodology: Detailed within this chapter are the methodologies adopted to explore the posed research questions. It elaborates on the selection and application of various evolutionary algorithms and meta-heuristics, providing a comprehensive explanation of their relevance and expected contributions to addressing the research objectives. The chapter ensures a clear understanding of the experimental design and the rationale behind the chosen methods.

Chapter 4 - Experiments: This chapter presents the experimental framework designed to evaluate the research hypotheses. It details the experimental setup, including the configuration of the molecule property prediction model and the application of MO-CMA-ES and NSGA-II algorithms. The chapter aims to transparently convey the methodology for assessing exploration behaviors, solution quality, and diversity, ensuring reproducibility and a clear understanding of the experimental process.

Chapter 5 - Results: Here, the outcomes of the conducted experiments are systematically presented. This includes the performance evaluation of the molecule property prediction model, particularly focusing on fish toxicity prediction, then for MO-CMA-ES and NSGA-II algorithms the comparative analysis of exploration behaviors and solution qualities produced by different meta-heuristics. The presentation of results through graphical representations facilitates a nuanced understanding of the findings.

Chapter 6 - Discussion: This chapter provides a critical examination of the experimental results in light of the established research questions. It delves into the implications of the findings, comparing them with existing literature and discussing their relevance to the development of more efficient molecular optimisation strategies. The discussion extends to the potential real-world applications of the research, highlighting its contribution to the field.

Chapter 7 - Conclusions: Concluding the thesis, this chapter synthesizes the research findings, addressing the research questions and evaluating the achievement of research objectives. It reflects on the study's contributions to the knowledge base, discusses limitations, and proposes directions for future research, thereby underscoring the thesis's value and potential impact on the field of molecular optimisation.

Chapter 2

Literature Review

The Literature Review chapter is structured into four key sections, each designed to build a comprehensive background for the study.

Section 2.1 provides an introduction to molecule optimisation, covering fundamental concepts and computational strategies for exploring molecular properties and design.

Section 2.2 examines the role of Evolutionary Algorithms (EAs) in optimisation, detailing their application in complex molecular design problems.

Section 2.3 introduces molecule property prediction models, discussing their development and impact on predicting molecule behaviors and properties accurately.

Section 2.4 synthesizes the literature reviewed, identifying gaps in current research and framing the study's contribution to the field.

2.1 Optimizing Molecules for Product Design

This section presents a high-level overview of molecules and their optimisation. The aim is to highlight its importance, especially using computational methods. Section 2.1.1 defines molecules, their properties, and how these properties are calculated with computational algorithms. Section 2.1.2 shows how molecules are represented. Section 2.1.3 presents the Tanimoto similarity, similarity principle, and its importance on optimisation. Section 2.1.4 shows the significance of the computation algorithms in the chemistry field. Lastly, Section 2.1.5 delves into molecule optimisation and how more optimized molecules could be achieved.

2.1.1 Molecules, Their Properties and Computational Assessment

Molecules, the fundamental units of chemical compounds, are characterized by a unique set of properties that dictate their behavior in various physical, biological, and chemical contexts. These properties, ranging from simple physical attributes to complex toxicological profiles, are integral to understanding the molecule's functionality and potential applications. The connection between a molecule and its properties is profound; it determines how a molecule interacts with its environment, how it can be utilized in product design and its overall suitability for a specific purpose.

This thesis will focus on these properties: molecular weight, molecular complexity, XlogP, reference likeness and fish toxicity. Each has been chosen for its relevance to the environmental and functional performance of the molecules in question, particularly as they pertain to the detergent molecules which will be our benchmark product.

Certain molecular properties can be directly calculated using computational algorithms, leveraging chemical libraries and informatics tools such as RDKit (Bento et al., 2020), which is an open-source cheminformatics software. For instance:

1. **Molecular Weight** is a critical parameter in chemistry and materials science, referring to the total mass of all atoms in a given molecule. Molecular weight plays a pivotal role in product design, especially in the formulation of polymers and other complex materials, as it directly influences properties like viscosity, melting point, and mechanical strength. In the context of detergent molecule design, the molecular weight is particularly crucial. Detergents, essentially surfactants, rely on their molecular architecture to effectively break down and remove dirt and grease (Yangxin et al., 2008). The balance between hydrophobic (water-repelling) and hydrophilic (water-attracting) components in a detergent molecule is heavily dependent on its molecular weight. A well-designed detergent molecule with an optimal molecular weight ensures efficient dirt removal, minimal environmental impact, and cost-effectiveness in production.

2. **Molecular Complexity** refers to the structural intricacy and diversity of a molecule, often characterized by factors such as the number of stereocenters (chiral centers where the spatial arrangement of substituents is non-superimposable on its mirror image), rings (cyclic structures that can significantly influence the chemical behavior and stability of molecules), and functional groups (specific groups

of atoms within molecules that are responsible for the characteristic chemical reactions of those molecules). In product design, particularly in the chemical and pharmaceutical industries, molecular complexity is a crucial consideration as it can impact the synthesis, cost, scalability, and functional properties of the product. For example, more complex molecules might offer enhanced specificity and functionality but could pose challenges in terms of synthesis and cost. In detergent molecule design, molecular complexity is equally vital. The structural complexity of a detergent molecule affects its interaction with different types of soils and stains. A detergent with a carefully designed complex molecular structure can target specific types of dirt and stains more effectively, enhance solubility, and improve biodegradability. Moreover, the complexity can influence the detergent's behavior in different water conditions, such as hard water (water with high mineral content, typically calcium and magnesium, which can reduce soap's effectiveness and cause scaling) or soft water (water with low mineral content, enhancing soap's effectiveness and reducing residue)(Yangxin et al., 2008).

3. **XlogP** a computational algorithm estimating the logarithm of a compound's partition coefficient between n-octanol (a fatty alcohol used as a standard to mimic the lipid-rich environment of biological membranes) and water (log P). Log P is a key descriptor of hydrophobicity, influencing a compound's solubility, absorption, and distribution properties, essential for predicting a substance's behavior in various environments. For detergent molecules, XlogP plays a crucial role in tailoring surfactant properties, balancing hydrophobic and hydrophilic characteristics to enhance cleaning efficiency and environmental compatibility (Rosen & Kunjappu, 2012).

4. **Reference Likeness**: Measures the degree of similarity between a candidate solution and established, effective molecules. This metric is crucial for balancing innovation with proven efficacy, as compounds with similar structures often exhibit comparable behaviors (Brown, 2009). In the optimisation of detergent molecules, leveraging reference likeness ensures that new compounds maintain the desired traits of effective predecessors, thereby streamlining the development process and enhancing the reliability of novel formulations. This approach supports the identification of molecules that are likely to succeed in practical applications, based on their resemblance to known, effective compounds.

These properties are typically computed during the initial stages of product design to screen and select candidates for further development. Other molecular properties, especially those related to toxicology, are not as straightforward to compute due to their complexity. For example:

1. **Fish Toxicity**, the toxicity of a compound to aquatic life, such as fish, is a vital consideration in environmental chemistry. The Global Harmonized System of Classification and Labeling (Winder et al., 2005) provides a framework for identifying chemical hazards, including acute (short-term) and chronic (long-term) aquatic toxicity. These properties are not directly calculable due to their complicated nature, therefore requiring lab tests (Russom et al., 1997). Given the high probability of detergents entering aquatic ecosystems, their impact on aquatic life emerges as a critical factor to consider. In line with our objective to develop a fully automated system, the implementation of a molecule property prediction model becomes essential.

Advanced algorithms and ML models are employed to predict these non-computationally calculable properties. These models are trained on lab tested data to learn the relationships between molecular structures and their biological or environmental effects, often using vast datasets to improve prediction accuracy (Alves et al., 2015; Mayr et al., 2016; Russom et al., 1997)

Optimizing a molecule for product development is a balancing act, particularly when the desired attributes present conflicting interests. Consider the example of an efficient detergent formulation. Ideal characteristics might include high lipophilicity, which ensures the detergent's efficacy in dissolving and removing oily stains. However, environmental considerations impose additional constraints; when such chemicals inevitably make their way into aquatic ecosystems, their potential toxicity to wildlife, such as fish, becomes a pressing concern. A detergent's lipophilicity and its aquatic toxicity are often inversely related; as one increases, the other becomes more problematic (Jimoh & Lin, 2019).

In the domain of product design, the application of molecules can be categorized into three distinct types based on their utilization: structure-based, substructure-based, and reaction-based uses. This thesis will concentrate on structure-based utilization, particularly relevant to the design and formulation of detergents.

Structure-based molecules are defined as the entirety of a molecule's structural composition dictates its physical and chemical properties, which in turn, directly impacts the functionality and efficiency of the end product. In detergents, the structure-based functionality is exemplified by the way the long-chain molecules interact with surfaces and contaminants. The architecture of these molecules, typically characterized by a hydrophobic tail and a hydrophilic head, allows them to insert themselves between dirt and the surface. The hydrophobic part of the molecule adheres to grease and oils, while the hydrophilic part remains interfaced with water, thus effectively detaching grime from surfaces upon agitation in water (Tadros, 2006).

2.1.2 Molecule Representation

Molecule input systems are a cornerstone in computational chemistry and cheminformatics, providing structured representations of chemical structures for computational analysis and modeling. Among these systems, the Simplified Molecular Input Line Entry System (SMILES) (Weininger, 1988a) (shown in Figure 2.1) is one of the most widely used due to its simplicity and ease of implementation. However, SMILES is not without its drawbacks. A notable issue is its potential for ambiguity; a single SMILES string can sometimes represent multiple molecules, which can lead to significant problems. For example, one SMILES representation might correspond to two different molecules, where one could be toxic and the other not, posing a challenge for Molecule Property Prediction Models.

Self-referencing Embedded Strings (SELFIES) (shown in Figure 2.1) were developed to address some of the limitations of SMILES. SELFIES (Krenn et al., 2019) are designed to be 100% robust, ensuring that every string maps to a valid molecular graph. Despite this advantage, SELFIES are computationally more complex and less intuitive than SMILES, which can be a barrier to their widespread adoption in some applications.

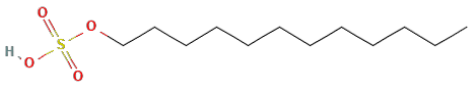
2D and 3D graph representations (shown in Figure 2.1) of molecules provide additional information that is not captured in SMILES strings, such as the spatial arrangement of atoms, which is crucial for understanding molecular interactions and properties. These graph-based representations are often derived from SMILES strings using computational algorithms; thus, they inherit any inaccuracies or ambiguities present in the SMILES representation. Despite this, their use is prevalent, particularly in the training of neural networks for tasks, where the spatial information can be critical (Duvenaud et al., 2015; Zhou et al., 2023).

CCCCCCCCCCCCOS(=O)(=O)O

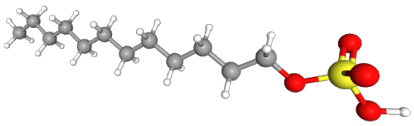
(a) SMILES representation.

[C][C][C][C][C][C][C][C][C][C][C][C][O][S][=Branch1][C][=O][=Branch1][C][=O][O]

(b) SELFIES representation.



(c) 2D Graph representation.



(d) 3D Graph representation.

Figure 2.1: Different representations of widely used detergent molecule Sodium Lauryl Sulphate. Taken from Pubchem database (Kim et al., 2019b)

In summary, while SMILES is the most preferred due to its widespread use and simplicity, its limitations, particularly in terms of potential ambiguity, are significant. SELFIES offer a robust alternative but at the cost of computational complexity. Meanwhile, 2D and 3D graphs provide valuable spatial information but still carry the foundational issues of the SMILES format from which they are often derived. This thesis will focus on SMILES format and 3D graphs that are derived from SMILES format.

2.1.3 The "Similar Structure, Similar Property" Principle and Tanimoto Similarity

The "similar structure, similar property" principle is a foundational concept in computational chemistry, which refers that similar molecules will also tend to exhibit similar chemical properties (Brown, 2009). This principle has been pivotal in developing computational methods for molecular similarity assessment, such as Tanimoto similarity, which quantifies the resemblance between molecular structures (Willett, 2006). Tanimoto similarity, or the Tanimoto coefficient, is derived from the Jaccard index and is specifically adapted for binary or vector-based representations of molecular features.

MACCS are widely used for similarity searching and structure-activity relationship studies. They are particularly effective for rapid screening of molecules with a relatively small and focused set of structural features. For applications requiring rapid screening with a focus on well-established structural features, MACCS with Tanimoto similarity may provide sufficiently accurate results with less computational overhead (Kuwahara & Gao, 2021). For this reasons, in this research we will use MACCS for Tanimoto similarity.

Tanimoto similarity plays a critical role in cheminformatics by providing a metric to gauge the similarity between the structural fingerprints of molecules. This metric has been effectively utilized in the iterative optimisation of molecular structures. By quantifying the extent of similarity, it is possible to incrementally modify a molecule, adjusting its properties in a controlled manner to align with the desired characteristics (Bajusz et al., 2015). Each iteration brings the molecule's properties closer to the optimal parameters, thereby leveraging the underlying principle that structurally similar molecules will possess similar properties.

This concept is vital not just for comparison but for a systematic exploration of design possibilities. In the field of product design, this approach can be likened to the use of design similarity algorithms to identify potential innovations from extensive design databases by screening for features akin to those in successful products. Understanding the complex network of relationships among design elements enables the formulation of effective strategies for applying EAs. These algorithms utilize this principle to navigate through the design space, making 'evolutionary steps' in enhancing specific features or functionalities of a product.

The strategic application of Tanimoto similarity within evolutionary computation is not a matter of observing a correlation between structure and properties. It is an approach that harnesses this relationship to intelligently navigate the molecular landscape. By doing so, we can systematically and efficiently evolve molecules towards a set of desired properties.

2.1.4 Rationale for Computer-Aided Molecular Optimisation

The optimisation of molecules using computational methods has become a cornerstone in modern product design, particularly in sectors like pharmaceuticals, materials science, and product design. The adoption of these computer-based approaches is driven by their unparalleled capacity to manage the immense complexity and diversity inherent in molecular structures and properties, a task that often surpasses the capabilities of traditional experimental methods (Brown, 2009). Through computational techniques, researchers can traverse the vast expanse of chemical space more efficiently and cost-effectively, facilitating the discovery and optimisation of molecules with ideal characteristics (Bilodeau et al., 2022b).

A significant advantage of this computational approach is its ability to dramatically reduce the time and resources needed for development cycles. In the pharmaceutical industry, for example, computational methods can expedite the process of identifying and refining potential drug candidates, compressing a process that typically spans several years into a much shorter time frame (Hughes et al., 2011). This acceleration is made possible by predicting key molecular properties, hence rapidly pinpointing the most viable candidates for further investigation.

Beyond efficiency, the computational approach can enhance the precision of molecular design. It grants the ability to predict properties like toxicity using molecule prediction models which otherwise require expensive and time-consuming lab tests.

In essence, optimizing molecules with computational tools is a strategy that goes beyond mere efficiency and cost-effectiveness; it is a pathway to achieving higher levels of precision and innovation in product design. As these computational methods continue to advance, their influence on molecular optimisation is assured to grow even more integral and transformative.

2.1.5 Current Methodologies in Molecular Optimisation: Approaches and Advances

Optimizing molecules is a complex challenge that necessitates a multidisciplinary approach, blending the expertise of computational chemistry, and computer science, especially when applied to materials science. The core objective is to devise molecules that are not only structurally sound but also functionally adept for their intended use in various products.

In recent years, ML algorithms have become a pivotal element in this optimisation process. Their ability to assimilate and interpret extensive datasets has revolutionized the identification of patterns and relationships within molecular structures that might otherwise remain obscured. Deep learning, a subset of these techniques, has proven particularly effective in predicting molecular behaviors and properties, thereby enhancing the efficiency of the molecule optimisation pipeline. Notably, neural networks have shown promise in predicting the toxicity of compounds, which is a vital concern in molecule development (Butler et al., 2018; Mayr et al., 2016).

Generative adversarial networks (GANs) and Variational Autoencoders (VAEs) stand at the forefront of generative molecular optimisation. These models excel in crafting novel molecular structures by learning from comprehensive databases of existing molecules. GANs, for example, have been utilized to generate molecules with tailored properties, which have been instrumental in aligning molecular designs with predefined criteria (De Cao & Kipf, 2018).

VAEs, on the other hand, offer a slightly different approach. They are particularly renowned for their ability to learn the distribution of data in a latent space, enabling the generation of new molecules by sampling from this space. This feature has proven effective in discovering molecules with specific desired properties, as showcased in a study by Gómez-Bombarelli et al. (2018). The researchers employed VAEs to generate molecules that exhibit high synthetic accessibility.

These advancements highlight a paradigm shift in molecular design, where generative models like GANs and VAEs are no longer confined to theoretical simulations. They are increasingly being integrated into the real world, offering a faster, more cost-effective route to molecular optimisation.

In the realm of molecule optimisation using evolutionary algorithms, significant strides have been made, as evidenced by studies like Douguet et al. (2000), Kwon and Lee (2021), and Winter et al. (2019b). Winter et al. (2019b) showcases the successful optimisation of molecules via Particle Swarm optimisation (PSO) and VAEs across multiple objectives instantaneously. Their methodology innovatively navigates a continuous representation of chemical space generated by VAEs, utilizing a well-structured objective function that combines various prediction models. This approach has proven effective in rapidly identifying molecules with enhanced desirable characteristics.

2.2 Optimisation Algorithms

This section presents an overview of optimisation Algorithms. Section 2.2.1 briefly introduces EAs. Section 2.2.2 presents the concepts of MOO. Section 2.2.3 introduces CMA-ES, widely used optimisation algorithm. Section 2.2.4 presents MO-CMA-ES, adaptation of of CMA-ES for MOO. Section 2.2.5 presents another MOO evolutionary algorithm NSGA-II for benchmarking. Section 2.2.6 analyzes the strengths and weaknesses of MO-CMA-ES and NSGA-II.

2.2.1 Evolutionary Algorithms

EAs represent a subset of evolutionary computation, a field that draws inspiration from biological evolution to solve complex optimisation problems. These algorithms employ mechanisms similar to biological evolution, such as selection, mutation, and crossover, to iteratively evolve solutions to optimisation challenges (Eiben & Smith, 2015).

A primary characteristic of EAs is their versatility in addressing a wide range of problems. Their adaptability is particularly evident in optimisation tasks, where they have shown notable success. Unlike traditional optimisation methods that may struggle with complex landscapes, EAs are adept at navigating multi-modal and non-linear spaces, making them suitable for various real-world applications (Eiben & Smith, 2015).

One prominent application of EAs is in engineering design optimisation, where they have been used to optimize the shapes and materials of components for maximum efficiency and minimum cost (Coello, 2006). Another significant use case is in financial modeling, where EAs assist in portfolio optimisation and risk management (Brabazon & O'Neill, 2006). Moreover, in the field of bioinformatics, EAs have contributed to the understanding of genetic sequences and protein folding, addressing problems that are computationally intensive and intricate (Pal et al., 2006).

The adaptability and robustness of EAs in tackling diverse and complex problems underscore their growing importance in computational research and real-world applications.

2.2.2 Multi-Objective Optimisation

MOO is a critical process in many designs, where multiple conflicting objectives must be considered simultaneously. It is a concept from the field of mathematical optimisation problems, where more than one objective function is optimized concurrently. This area of multiple-criteria decision-making is widely applied in engineering, economics, and logistics, addressing the inherent trade-offs between conflicting objectives (Gunantara, 2018).

In MOO, solutions that are Pareto optimal (non-dominated) represent a balance where improving one objective would worsen another. This concept is visualized as a curve or line on a graph as shown in Figure 2.2, with each point illustrating an optimal

balance among competing objectives. The Pareto frontier is essential in decision-making, marking the most efficient solutions and guiding the identification of the best compromises in situations involving inevitable trade-offs. As the number of objectives increases, the complexity of identifying these optimal solutions also grows. This methodology is particularly significant in product design, allowing for a structured assessment of trade-offs and facilitating the creation of products that effectively satisfy multiple criteria (Gunantara, 2018).

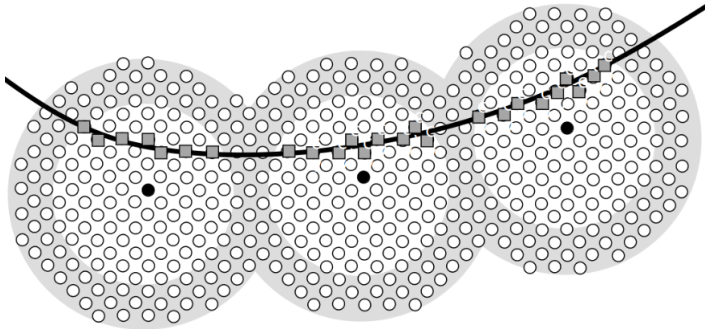


Figure 2.2: Illustration of the Pareto Frontier in Multi-Objective Optimisation. Each dot on the line represents a potential solution evaluated based on multiple objectives. The black line denotes the Pareto Frontier, where the solutions are Pareto optimal; any improvement in one objective would result in the deterioration of the other. The solutions along this frontier offer the most efficient trade-offs between the competing objectives and are considered non-dominated in the context of the given optimisation problem.

Molecule selection inherently aligns with multi-objective paradigms. Traditionally, this domain has often been tackled through a sequential, one-objective-at-a-time approach. However, the advent of multi-objective methodologies, frequently employing a weighted-sum approach (Nicolaou et al., 2007), has emerged as a means to harmonize diverse objectives. Notably, the introduction of Pareto-based methods, exemplified by MoSELECT (Gillet et al., 2002), has sought to address a spectrum of objectives concurrently.

2.2.3 Covariance Matrix Adaptation Evolution Strategy

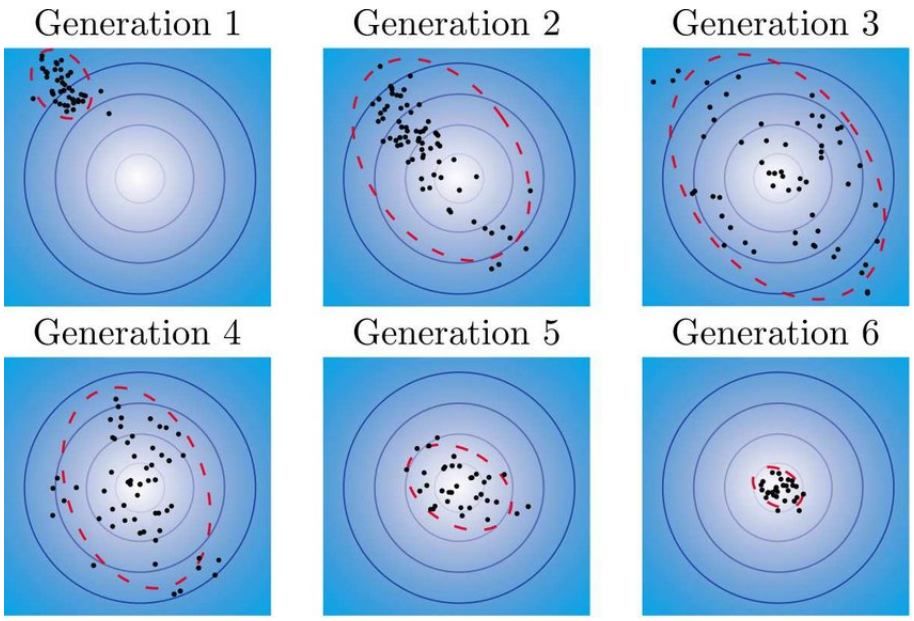


Figure 2.3: Illustration of CMA-ES algorithm (Tan et al., 2019).

The CMA-ES is a widely recognized algorithm in the realm of evolutionary computation. Its development can be traced back to the late 1990s, with its origins rooted in the adaptation of covariance matrices in evolutionary strategies. The seminal work by Hansen and Ostermeier (2001) laid the foundation for CMA-ES, where they introduced an innovative approach for self-adaptation of the strategy parameters (as shown in Figure 2.3) in evolutionary algorithms. This approach marked a significant shift in the evolutionary computation field, as it allowed for more efficient exploration of complex, multi-modal landscapes.

The applications of CMA-ES have spanned various fields, demonstrating its versatility and robustness. In the field of engineering, CMA-ES has been instrumental in optimizing complex design problems, such as aerodynamic shape optimisation (Zhang et al., 2020). In the realm of machine learning, it has been utilized for neural network training, offering an alternative to gradient-based methods, particularly in scenarios where the gradients are difficult to compute (Loshchilov & Hutter, 2016). For the engineering field, specifically for fault diagnosis of wireless sensor networks, CMA-ES has been used to optimize fault diagnosis (He et al., 2018). Additionally, in cloud computing, scheduling of the users' jobs is explored in the research by Emadi et al. (2017) using CMA-ES. These varied applications underscore the algorithm's ability to efficiently navigate complex optimisation landscapes, making it a valuable tool in numerous scientific and engineering endeavors.

The success of CMA-ES in these diverse applications can be attributed to its ability to adaptively tune the search distribution, effectively balancing exploration and exploitation. This adaptability makes it particularly suitable for problems with complex, ill-conditioned, and noisy fitness landscapes. The continual development and refinement of CMA-ES, including its variants, have further expanded its applicability, making it a go-to algorithm in the toolbox of evolutionary computation.

2.2.4 Multi-Objective Covariance Matrix Adaptation Evolution Strategy

The MO-CMA-ES is an extension of the classic CMA-ES, designed to tackle problems involving multiple conflicting objectives. MO-CMA-ES emerges as a sophisticated optimisation approach tailored to address the intricacies of MOO problems. Rooted in the theoretical framework of EAs, MO-CMA-ES integrates the concept of Pareto optimality to navigate through the trade-offs between objectives. This strategy maintains a population of solutions, adapting the covariance matrix to reflect the underlying structure of the Pareto front. The approach allows for a well-distributed set of non-dominated solutions, providing a comprehensive overview of the trade-offs involved (Igel et al., 2007).

In practice, MO-CMA-ES has been applied to case studies across a diverse array of disciplines. For instance, in the field of aerodynamics, it has been utilized for the optimisation of wing shapes to simultaneously minimize drag and maximize lift (Fasel et al., 2017). In finance, MO-CMA-ES has been employed to derive investment portfolios that optimize for both risk and return, offering valuable insights for investors (Lwin et al., 2014). Moreover, in renewable energy, the algorithm has facilitated the design of wind farms, optimizing the placement of turbines to maximize energy production while minimizing environmental impact (Shekar & Shivakumar, 2019).

These applications exemplify the versatility and efficiency of MO-CMA-ES in solving complex optimisation problems where multiple objectives must be considered. The strategy has proven to be particularly beneficial in real-world scenarios where the trade-offs between objectives are not just theoretical considerations but have practical consequences that must be carefully weighed.

2.2.5 Non-dominated Sorting Genetic Algorithm II

The NSGA-II is a seminal algorithm in the field of evolutionary computation, specifically designed for solving MOO problems. Introduced by (Deb et al., 2002), NSGA-II improved upon its predecessor by introducing a fast non-dominated sorting approach and a crowding distance operator, which together enhance the diversity-preserving mechanism and the computational efficiency of the algorithm. The fundamental principle of NSGA-II is to rank the population of solutions based on the levels of non-dominance, with the goal of progressively moving towards the Pareto-optimal front (Deb et al., 2002).

The NSGA-II stands out as a seminal algorithm in the field of MOO, making it a natural choice as a benchmark for various studies. Its superiority in handling MOO challenges has been well-established in various domains (Rahimi et al., 2023).

NSGA-II has been applied across a broad range of disciplines due to its versatility and efficacy. In environmental engineering, it has been utilized for water and waste management optimisation problems (Reed et al., 2013), while in mechanical engineering, it has contributed to the multi-objective design of composite materials (Vo-Duy et al., 2017). These use cases demonstrate NSGA-II capability to effectively balance competing objectives and to find a diverse set of Pareto-optimal solutions.

Comparative studies of NSGA-II with other evolutionary algorithms, such as the original Non-dominated Sorting Genetic Algorithm (NSGA) and the Strength Pareto Evolutionary Algorithm (Zitzler & Thiele, 1998), have illustrated its superior performance in terms of both convergence and diversity. NSGA-II's ability to maintain a diverse set of solutions without specifying any sharing parameters make it a preferred choice in many MOO scenarios (Rahimi et al., 2023).

2.2.6 Comparative Analysis of MO-CMA-ES and NSGA-II

In the literature on evolutionary algorithms, a significant amount of research has been dedicated to the comparative analysis of different algorithms' efficiency and effectiveness in optimisation. The MO-CMA-ES and the NSGA-II are two prominent algorithms that have been frequently contrasted.

MO-CMA-ES is praised for its global search capability and the internal mechanism that adapts the step size, which is highly beneficial in handling complex, multimodal optimisation problems (Hansen & Ostermeier, 2001). It is particularly effective in continuous optimisation spaces and has been shown to converge to global optima with remarkable consistency. However, its computational complexity can be a limitation, especially for problems with a high number of objectives or constraints (Igel et al., 2006).

On the other hand, NSGA-II is renowned for its ability to maintain a diverse set of solutions. It employs a fast non-dominated sorting approach, which makes it particularly effective for problems where the Pareto front needs to be well-defined and represented. NSGA-II has been successful in diverse domains, from engineering design to bioinformatics (Deb et al., 2002). However, its performance can degrade in highly multimodal and complex search spaces, and it may require careful parameter tuning to achieve the best results (Coello, 2007).

Comparative studies suggest that while MO-CMA-ES can be more suitable for problems with complex landscapes and fewer objectives, NSGA-II is generally more diverse and easier to implement for a wide range of multi-objective problems, especially with a larger number of objectives (Zhou et al., 2011). Both algorithms have their unique advantages and limitations, and the choice between them is often dictated by the specific characteristics and requirements of the problem at hand.

2.3 Molecule Property Prediction

This section presents an overview of Molecule Property Prediction models. Section 2.3.1 delineates the foundational aspects and historical advancements in molecule property prediction. It highlights the pivotal role of predictive models in cheminformatics and drug discovery, underscoring their contribution to reducing experimental costs and time. Section 2.3.2 introduces the concept of representation learning for neural networks. Section 2.3.3 introduces the State-of-the-Art property prediction model, *Uni-Mol*.

2.3.1 Molecule Property Prediction Models

Molecule property prediction models are pivotal in the fields of cheminformatics and drug discovery, offering the ability to forecast a wide range of chemical properties based on molecular structure. These predictive models serve as essential tools, significantly reducing the need for costly and time-consuming experimental procedures. They are typically grounded in Quantitative Structure-Activity Relationships (QSAR) and/or machine learning techniques, which correlate molecular descriptors with biological activities or physicochemical properties (Feinberg et al., 2018; Walters & Barzilay, 2020).

Montavon et al. (2012) provided a novel approach to applying deep learning to predict molecular properties, particularly atomization energy, showcasing the potency of neural networks in computational chemistry. Gómez-Bombarelli et al. (2018) introduced a novel method for molecule design using VAEs, enabling the generation and optimisation of molecular structures in a continuous latent space, a breakthrough for drug discovery and product design. Complementing these developments, Wu et al. (2018) established MoleculeNet, a comprehensive benchmark dataset for molecular ML, facilitating the assessment of diverse ML methods, including graph convolutional networks, in predicting molecular properties. These studies collectively mark a significant evolution in molecule property prediction, highlighting the roles of deep learning, data-driven design, and robust benchmarking in the field.

Despite their successes, molecule property prediction models also face challenges, such as the need for large and diverse datasets to train on, and the difficulty in interpreting the models' predictions, which is critical for scientific validation and understanding (Polishchuk, 2017).

2.3.2 Representation Learning Models

In recent times, the prevalence of representation learning (pretraining methodologies) (Bengio et al., 2013; Hamilton et al., 2017; Zhang et al., 2018) has been manifest across a lot of applications spanning Natural Language Processing (NLP) with models like Bidirectional Encoder Representations from Transformers (BERT) (Devlin et al., 2018) and Generative Pre-Trained Transformer (GPT) (Brown et al., 2020; Radford et al., 2018, 2019), and Computer Vision (CV) with the likes of Vision Transformer (ViT) (Dosovitskiy et al., 2020). A shared characteristic of these applications is the abundance of unlabeled data compared to limited labeled data. This discrepancy is

mirrored in the context of classified molecules within the PubChem dataset (Kim et al., 2019b), where a mere 3% are categorized as toxic to fish. A proposed solution draws inspiration from representation learning paradigms, commencing with a pretraining phase to extract meaningful representations from unlabeled data, followed by a fine-tuning step that capitalizes on the scarcity of labeled data.

Recent advancements in Molecular Representation Learning (MRL) models have substantially elevated their performance in various property prediction tasks (Fang et al., 2022; Rong et al., 2020; Wang et al., 2022; Yang et al., 2019; Zang et al., 2017). Nevertheless, a critical hurdle remains—enhancing performance and expanding the applicability of existing MRL models. Natural molecules inherently possess intricate 3D structures, and their properties are heavily influenced by these spatial arrangements (Crum-Brown & Fraser, 1865; Hansch & Fujita, 1964). However, numerous contemporary MRL approaches initiate by representing molecules as one-dimensional sequential strings (e.g., SMILES (Wang et al., 2019; Weininger, 1988b; Xu et al., 2017) and International Chemical Identifier (InChI) (Handsel et al., 2021; Heller et al., 2015; Winter et al., 2019a)) or 2D graphs (Hu et al., 2019; Li et al., 2021; Rong et al., 2020; Wang et al., 2022; Ying et al., 2021). Such representations potentially constrain the models' capacity to fully leverage 3D information for downstream tasks.

2.3.3 Uni-Mol, State-of-the-Art Molecule Property Prediction Model

Uni-Mol a pioneering Universal 3D Molecular Representation Learning Framework (Zhou et al., 2022), has surged to the forefront as a state-of-the-art molecule property prediction model. Distinguished by its pretraining on 3D structures of 210 million molecules, *Uni-Mol* delves deeper into understanding molecule structures. These 3D structures are crafted using the Rdkit (Landrum et al., 2013) from SMILES, employing the Experimental-Torsion basic Knowledge Distance Geometry with Merck Molecular Force Field optimisation (ETKDG) (Riniker & Landrum, 2015). This pretraining equips the model to predict previously intractable properties of molecules.

Uni-Mol leverages transformers, adopting a Pre-LayerNorm architecture to accommodate 3D spatial data. It integrates invariant spatial positional encoding, pair representation, and an SE(3)-Equivariance coordinate head. To facilitate the preservation of positional encoding under global rotation and translation, *Uni-Mol* employs relative positional encoding. This involves using Euclidean distances between atom pairs and a pair type-

aware Gaussian kernel. Moreover, transformers are adeptly designed to encompass both token-level and pair-level representations. Token-level representations serve as a baseline for downstream fine-tuning, while pair-level encoding encapsulates spatial positions, enabling a more nuanced grasp of 3D spatial relationships.

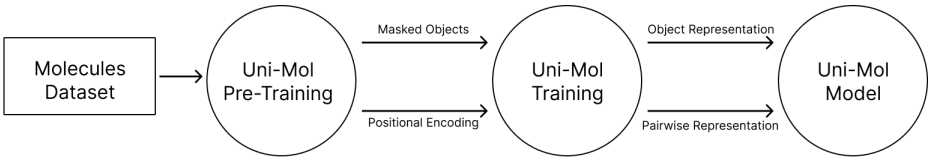


Figure 2.4: *Uni-Mol* graph transformer. Left: Pre-training architecture. Middle: Inputs, including masked objects and spatial positional encoding created by pairwise Euclidean distances are used for training. Right: Pairwise and individual object representations comprise foundations for model.

Self-supervised learning emerges as a potent strategy for molecule property prediction, capitalizing on extensive datasets of unlabeled molecules. *Uni-Mol* employs a masked atom prediction task for self-supervised learning, with a special atom $[CLS]$ representing the entire molecule. However, as the 3D spatial positional encoding encodes pair distances and corresponding atom types, the existing masked atom prediction task falls short of encouraging meaningful learning. To mitigate this, *Uni-Mol* introduces a 3D position denoising task, injecting uniform noise into coordinates and recalculating positional encoding based on the perturbed coordinates. Additionally, two heads are employed to recover accurate spatial positions—a pair-distance prediction head and a coordinate prediction head. The overall pre-training process and architecture of *Uni-Mol* are illustrated in Figure 2.4.

The intricate architecture and pretraining methodology of *Uni-Mol* demonstrate its adeptness at integrating 3D molecular information for enhanced property prediction, laying the foundation for improved molecule design and selection strategies. The incorporation of spatial relationships, self-supervised learning, and transformative architecture positions *Uni-Mol* at the forefront of molecular representation learning, fostering a deeper understanding of intricate molecular properties and interactions.

2.4 Synthesis of Literature and Identification of Research Gap

VAEs and GANs occupy a leading position in the field of generative molecular optimisation. However, the 'black-box' nature of these networks, coupled with their ability to produce a vast number of molecular structures and huge chemical datasets like PubChem (Kim et al., 2019b), necessitates an efficient search algorithm. The research gap lies in developing a method that would filter through the provided datasets to select the most favorable candidates, as depicted in Figure 2.5.

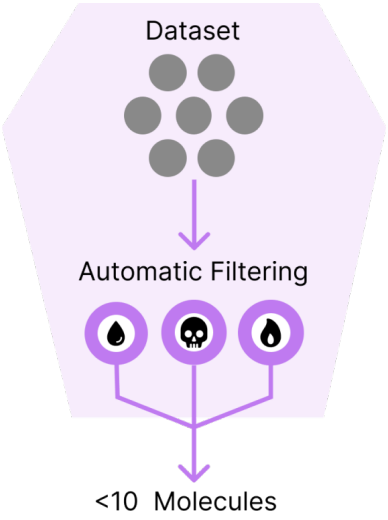


Figure 2.5: This diagram depicts the initial dataset being subjected to an automated filtering mechanism, which effectively reduces the number of candidate molecules. This filtration is to address the research gap where existing methodologies may yield an overwhelming number of potential compounds, thus complicating the selection process for practical applications.

When presented with a molecular graph and a set of attributes, it is intuitive to frame the design challenge as a MOO problem. The objective is to identify a group of molecules that collectively optimize these attributes. The MOO approach is particularly relevant for designing detergents, where efficacy and environmental safety are often at odds. Ideal detergents have high XlogP to break down oils and stains, but this can lead to increased aquatic toxicity. MOO helps in finding a balance, optimizing for effective cleaning while minimizing environmental impact. While single-objective EAs (Guimaraes et al., 2017; Kusner et al., 2017; Winter et al., 2019c; Yang et al., 2019; Zang et al., 2017) and MOEAs (Namasivayam & Bajorath, 2012; Winter et al., 2019c) have been proposed, however yet again these methods lean towards generative strategies.

The dynamic landscape of molecular optimisation (continuously changing scenarios and parameters) necessitates methodologies that can efficiently explore and exploit a multi-dimensional solution space. MOEAs, such as MO-CMA-ES and NSGA-II, offer a significant advantage over current ML methods used in molecular property optimisation. MOEAs excel in handling complex optimisation problems characterized by multiple, often conflicting, objectives (Deb et al., 2002; Hansen et al., 2003). Unlike ML methods that may require predefined data patterns or rely heavily on training data, MOEAs are adept at discovering a diverse set of solutions through evolutionary processes that mimic natural selection and genetics. This attribute makes MOEAs particularly suited for optimizing molecular properties, where the solution space is not only vast but also poorly understood.

Employing MOEAs for molecular property design allows for the generation of a Pareto front of optimal solutions, optimized according to a variety of criteria such as cost, efficacy, and toxicity (Coello, 2007; Zitzler et al., 2001). This capability is invaluable, as it provides a spectrum of trade-off solutions, each with its own set of advantages and limitations. For pharmaceutical companies, this means access to a diversified portfolio of molecular compounds, each optimized for a different set of properties. The ability to simultaneously optimize for multiple criteria is indispensable in the pharmaceutical industry, where the development of a new drug requires a delicate balance between efficacy, safety, cost, and regulatory compliance.

This thesis endeavors to explore MOEAs as an initial step towards harnessing their full potential in the field of molecular optimisation. By presenting methodologies that demonstrate the utility of MOEAs in navigating and optimizing complex molecular landscapes, this research aims to lay the groundwork for future explorations in this domain.

In the domain of MOO, MO-CMA-ES and NSGA-II stand out most making them ideal methodology, given the vast, complex, and multi-modal molecular space. They offer a robust promise for identifying optimal solutions that meet diverse criteria, a capability that has been substantiated in numerous applications, as detailed in section 2.2.5 and section 2.2.4.

Although our focus deviates from the broad search scope MO-CMA-ES typically excels in, due to the application of the algorithm with the *Similar Property Principle*, its ability to traverse through complex landscapes remains a compelling rationale for its adoption. While we intend to modify this algorithm to align with our distinct approach, capitalizing on its computational efficiency presents a promising avenue. By tailoring the algorithm to our needs and assessing various modifications, we aim to uncover the adaptations that yield optimal outcomes.

As for NSGA-II, it's robustness in handling multiple objectives and diverse frontiers becomes particularly relevant, as mentioned before in section 2.1.1, product design often involves conflicting objectives. NSGA-II's ability to handle multiple objectives, identifying a range of optimal solutions along the Pareto frontier, makes it a suitable for this work.

As we transition from the synthesis of existing literature to the exposition of our methodological contributions, it becomes apparent that the exploration and application of MOEAs, particularly MO-CMA-ES and NSGA-II, are not merely academic exercises. Instead, they represent a concerted effort to address the challenges of molecular property optimisation. The forthcoming methods section will delineate the specific adaptations and innovations introduced in this study, positioning MOEAs as central to advancing the frontier of molecular design. This approach not only fills the identified research gap but also sets a precedent for the application of EAs in solving complex, MOO problems inherent in molecular science.

Chapter 3

Methodology

This section presents the methodology adopted in this study, comprising two main areas: the development and application of the *Uni-Mol* model for molecule property prediction and the implementation of multi-objective optimisation algorithms for molecular design.

Section 3.1, utilizing the fish toxicity dataset from PubChem, details the training of the *Uni-Mol* model, emphasizing data preprocessing, handling of imbalanced datasets, and fine-tuning strategies. The methodology underscores the exclusion of disconnected structures and duplicates, application of oversampling techniques, and the employment of optimal hyperparameters (found with trial experiments) to ensure the model's robustness and accuracy in predicting molecule properties.

Section 3.2 elaborates on the customization of MOEAs to meet the specific requirements of molecular design. It describes the integration of the *Uni-Mol* model within MOEAs to guide the evolutionary search process, highlighting two novel heuristic approaches: direct correlation and extended search. The methodology focuses on exploring and optimizing the molecular search space, detailing the configuration of algorithms, generation of diverse candidate solutions, and the strategic balance between exploration and exploitation.

Section 3.3 and 3.4 provide comprehensive insights into the implementation of MOEAs, specifically detailing the MO-CMA-ES and the NSGA-II. Each algorithm's implementation is discussed, including the selection of offsprings, building of solution frontiers, varying meta-heuristics and iterative optimisation processes aimed at identifying near-optimal compounds. The algorithms' parameters, constraints, and the iterative refinement process of candidate compounds are meticulously outlined, offering a clear view of the operational framework.

3.1 Molecule Property Prediction Model, *Uni-Mol*

This study employs the fish toxicity dataset acquired from the PubChem database, to train the *Uni-Mol* model. The dataset, primarily in SMILES format, comprises 300,000 non-toxic and 30,000 toxic molecules relative to fish. The toxic subset includes a mix of acute (short-term) and chronic (long-term) toxicity molecules, amalgamated due to the limited availability of toxic fish molecule data.

A thorough analysis of the dataset revealed the presence of a period symbol (“.”), a conventional disconnect symbol in SMILES notation. This symbol indicates disconnected structures within a molecule, representing components that are not chemically bonded. Given the design limitations of *Uni-Mol*, which can process only singular molecular entities and not mixtures, molecules containing this disconnect symbol were excluded from the dataset.

Post exclusion of mixtures and duplicate entries, the refined dataset consisted of approximately 250,000 non-toxic and 3,000 toxic molecules. Training *Uni-Mol* on this imbalanced dataset posed a significant risk of bias towards non-toxicity predictions. To mitigate this, oversampling techniques were applied, as suggested by Mohammed et al. (2020). Furthermore, the optimal hyperparameters for *Uni-Mol*, as recommended in its originating publication, were employed in the model training process.

The same data pre-processing pipeline was employed during fine-tuning to maintain consistency with the pre-training process. For molecules, multiple random conformations can be generated quickly, making it possible to use them as data augmentation during fine-tuning to enhance performance and robustness. In cases where 3D conformations could not be generated, the molecular graph was used as a 2D conformation. Like natural language processing and image analysis, the representation of $[CLS]$, which represents the entire molecule or the mean representation of all atoms, was used with a linear head to fine-tune downstream tasks.

3.2 Multi-Objective Optimisation (Search) Algorithm

The implementation and customization of MOEAs cater specifically to the intricate demands of molecular design. The objectives and constraints aligned with desired molecular properties are defined. The algorithm's configuration should focus on generating a diverse set of promising candidate solutions while respecting predefined constraints. Leveraging Pareto-based techniques facilitates efficient navigation of the multi-objective landscape, enabling the identification of a spectrum of optimal solutions.

The integration of *Uni-Mol* into the MOEAs, as depicted in Figure 3.1, constitutes a pivotal step. This model's predictive capabilities guide the evolutionary search process, aiding in the identification of molecules with higher potential to meet desired property criteria. Strategic measures are devised to strike a balance between exploring new regions within the chemical space and exploiting regions containing promising solutions. Additionally, hybridization strategies synergize the strengths of both algorithms, leading to heightened efficiency and accuracy.

In this study, we propose the integration of two novel heuristic methodologies within the framework of MOEAs for the enhancement of molecular search space exploration, commencing from predetermined seed molecules; direct correlation approach and extended search approach.

1. Direct correlation explores correlations between perturbations on molecular structures and variations in properties of interest, selecting the seed molecules as starting point. Then search space is explored outwards to find optimal compounds. This meta-heuristic is demonstrated in Figure 3.1.

2. Extended search, in turn, initially selects a set of molecules featuring a specified similarity level with respect to the seed molecules, and then uses the selected molecules as starting molecules for direct correlation instead of the original seed molecules. This meta-heuristic is demonstrated in Figure 3.2.

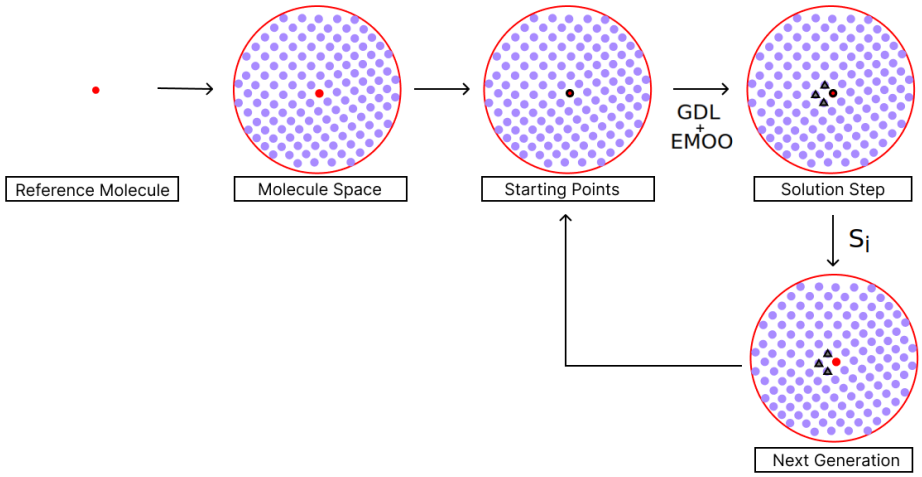


Figure 3.1: Overview of proposed method with meta-heuristic *direct correlation*: given an initial chemical design space, a search space is selected based on MACCS Tanimoto similarity $\hat{T}_0$; from this set, initial offsprings are identified based on Tanimoto similarity $\hat{T}$ to the seed compound; high-risk compounds are removed using *Geometric Deep Learning* (GDL) and optimised compounds are identified using *Evolutionary Multi-objective Optimisation* (EMOO), thus building a solution set; the obtained solution set is used as a new set of initial compounds to iterate the process and build new generations of optimised compounds, until stability is reached; the final result is the set of suggested compounds for consideration for product design.

There are four hyper-parameters that control the behavior of the traversal of the set of candidate compounds for each MOEAs:

1. Initial similarity threshold $\hat{T}_0$: $\hat{T}_0$ is a similarity threshold for creating search space from the dataset. Larger $\hat{T}_0$ ensure that solutions will be similar to the reference compound; M_0 by hard constriction, which can be good when priority is given to preserving the structural properties, at the cost of reducing possibilities for finding more novel compounds.

2. Generic similarity threshold $\hat{T}$: $\hat{T}$ is responsible for finding offspring from parent where each offspring is bigger than the similarity threshold $\hat{T}$. Larger $\hat{T}$ slow down convergence, possibly inducing the method to go through additional cycles to reach stabilization.

3. Generation size β: For each generation β parents are selected from previous solution space. Larger β increase explorations, subsequently possibly increasing computational time.

4. Generation size λ: For each generation λ offsprings are selected from each parent. Larger λ decrease the randomness in the generation step with the cost of computational time.

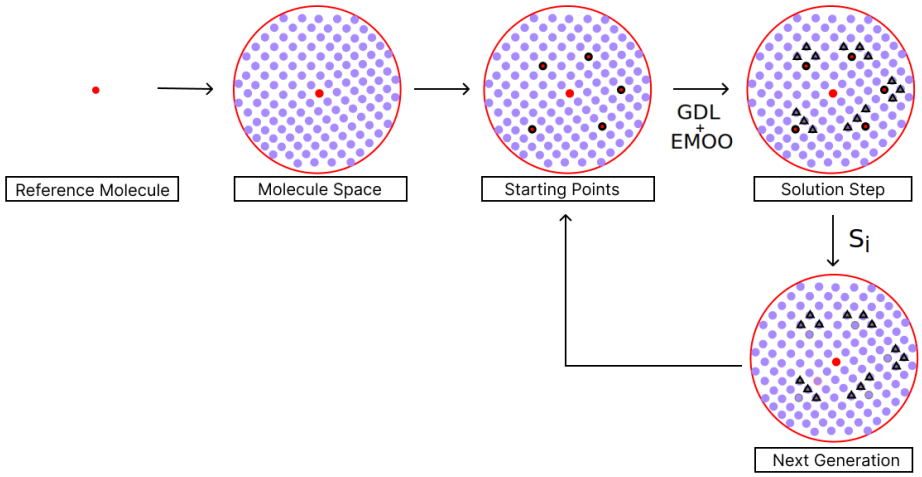


Figure 3.2: Overview of proposed method with meta-heuristic *extended exploration*: given an initial chemical design space, a search space is selected based on MACCS Tanimoto similarity $\hat{T}_0$; then starting compounds are identified using some Tanimoto similarity value to the seed compounds; from this set, initial offsprings are identified based on Tanimoto similarity $\hat{T}$ to the starting compounds; high-risk compounds are removed using *Geometric Deep Learning* (GDL) and optimised compounds are identified using *Evolutionary Multi-objective Optimisation* (EMOO), thus building a solution set; the obtained solution set is used as a new set of initial compounds to iterate the process and build new generations of optimised compounds, until stability is reached; the final result is the set of suggested compounds for consideration for product design.

We also have implemented constraints on properties. There are 2 types of constraints; Generation property constraints ($\mathcal{A}_{constr}$) and Final property constraints ($\mathcal{A}_{constr\_final}$). Generation property constraints control direction of each step as we do not want some property values to exceed some values. It is expected this will occur as there is trade off between properties. Final property constraint is more extreme to get a solution set with the values in range of what we want. The Final property constraint only applied at the end of run, otherwise it will negatively impact exploration of the algorithms causing pre-mature convergence.

Figures 3.1 and 3.2 presents an overview of this automated optimisation (filtration) algorithm's process. After constraining dataset with $\hat{T}_0$, EAs, from starting compounds, explores the space. In each step, selecting offsprings depending on the size β, λ and $\hat{T}$, then evaluating offspring with *Uni-Mol*, generation constraints and other properties by algorithmic calculations.

Broadly, we characterize our problem as an *Information-geometric Evolutionary Multi-objective Optimisation* task:

- **Given** a set $\mathcal{A}$ of relevant properties which can be ascribed to specified compounds (and which are assumed to have domains ranging through real-valued intervals); a subset $\mathcal{A}_{opt} \subseteq \mathcal{A}$ of those properties which must be *optimised*, i.e. either minimised or maximised; a subset $\mathcal{A}_{constr} \subseteq \mathcal{A}$ of those properties which define *constraints*, i.e. such that for each property $A_c \in \mathcal{A}_{constr}$ we have defined two values v_c^{min}, v_c^{max}, where $v_c^{min} \leq v_c^{max}$; and a set of compounds to be considered as candidate solutions for the problem;
- **Find** a set of compounds which are *good enough approximations* of the compounds that optimise the properties in $\mathcal{A}_{opt}$ while ensuring that properties $A_c \in \mathcal{A}_{constr}$ belong to the interval $[v_c^{min}, v_c^{max}]$.

The objectives the implemented methods will try to optimize are molecule weight and complexity, XlogP, and reference-likeness, where reference-likeness is MACCS Tanimoto similarity between the seed compound and offspring. The other objective will be fish toxicity of the molecules which will be evaluated just after offspring selection for computational efficiency.

3.3 Implementation of MO-CMA-ES

Algorithm 1 MO-CMA-ES

function SELECT_OFFPSPRINGS
    Randomly select β molecules from S_{i-1}.
    Select λ molecules using $\hat{T}$ similarity to the each selected
    β molecules from search space.
    Filter out molecules using $\mathcal{A}_{constr}$ and *Uni-Mol*.
end function

function BUILD_FRONTIER
    Remove duplicates from $M_k^c[1]$
    $frontier \leftarrow M_k^c[1]$ ▷ Initialize frontier
    for $i \leftarrow 2$ **to** $length(M_k^c)$ **do**
        Compute boolean array s based on $\mathcal{A}_{opt}$:
        if $\mathcal{A}_{opt}$ is 'max' **then**
            $s \leftarrow M_k^c[i] \geq frontier$
        else if $\mathcal{A}_{opt}$ is 'min' **then**
            $s \leftarrow M_k^c[i] \leq frontier$
        end if
        $domij \leftarrow \text{all}(s)$ ▷ Check if i dominates
        $domedij \leftarrow \text{any}(s)$ ▷ Check if i is dominated
        if $domij$ not empty **then**
            Remove dominated from $frontier$
            Add $M_k^c[i]$ to $frontier$
        else if $domedij$ not empty **then**
            if $M_k^c[i] \equiv frontier$ **then**
                Recompute s with inverted $\mathcal{A}_{opt}$
                $domedij \leftarrow \text{all}(s)$
                if $domedij$ empty **then**
                    Add $sim\_props[i]$ to $frontier$
                end if
            end if
        end if
    end for
end function

function MAIN
    Iteratively build frontiers, adjusting β and λ accordingly.
end function

For the implementation of MO-CMA-ES the strategy to navigate towards near-optimal compounds given specified properties $\mathcal{A}_{opt}$ and $\mathcal{A}_{constr}$ will follow the conceptual framework of MO-CMA-ES, adapted to a non-parametric setting, without mutations meaning with selection from defined search space and modified dynamic approach.

Reference points in the search space are determined using either direct correlation or extended search given M_0 and T_0, thus defining the set M_0^c. By definition, $M_0 \subseteq M_0^c$; the set $\tilde{M}_0^c \subseteq M_0^c$ is then selected based on $\mathcal{A}_{constr}$ and removal of toxic molecules identified using *Uni-Mol*. From these, Pareto optimal solutions are built with the properties calculated, considering $\mathcal{A}_{opt}$, thus assembling the initial Pareto optimal solution set S_0.

Given a generation size determined by β and λ, a randomly selected β parent molecules $\{m_{01}, ..., m_{0\beta}\} \subseteq S_0$, random λ offsprings are selected using T similarity with the respective parent. Offspring are then combined to form M_1 which are candidates for the new solution space S_1. This procedure is repeated to build $S_2, S_3, \ldots$, until a run-time limit is reached or some stability criteria is reached in S_N for some finite N – for example, no change observed in all the frontiers combined after removal of dominated molecules. To help avoid local optima, we also include, following the strategy of MO-CMA-ES, a growth factor $G > 1$ for β and λ: if $\frac{|M_{k+1}|}{|M_k|} < 1$, then both is updated to $\times G$, and if $\frac{|M_{k+1}|}{|M_k|} > 1$, then they are updated to $\times \frac{1}{G}$. The final solution set, combining $\mathcal{S}_1, \ldots, \mathcal{S}_N$ with dominated molecules removed, comprises the molecules suggested as potential solutions.

3.4 Implementation of NSGA-II

Algorithm 2 NSGA-II

function SELECT_OFFPSPRINGS
    Randomly select β pair of molecules from S_{i-1}.
    Select λ molecules from each intersection set created by
    T similarity to the selected β pairs of molecules.
    Filter out molecules using $\mathcal{A}_{constr}$ and *Uni-Mol*.
end function

function NONDOMINATEDSORTING
    for each molecule i in population **do**
        Initialize n_i and S_i
        for each molecule j in population **do**
            if i dominates j **then**
                Add j to S_i
            else if j dominates i **then**
                Increment n_i
            end if
        end for
        if $n_i = 0$ **then**
            Assign rank 1 to i
        end if
    end for
    Extract first front from population
    while population is not empty **do**
        Create next front based on domination counts
    end while
    Calculate crowding distances for each front
end function

function MAIN
    Iteratively build frontiers, adjusting β and λ accordingly.
end function

Following the methodology outlined for MO-CMA-ES in Section 3.3, the implementation of NSGA-II in our framework also aims at identifying near-optimal compounds. However, it diverges by incorporating NSGA-II's robust non-dominated sorting mechanism, coupled with a unique offspring selection process based on the union of parent spaces. This approach aligns with the non-parametric nature of our problem, where the selection is entirely based on the properties of whole molecules.

Initially, we start with a set of candidate compounds similar to MO-CMA-ES. For NSGA-II, the primary focus lies in establishing an efficient non-dominated sorting process, which categorizes compounds based on Pareto dominance considering the properties in $\mathcal{A}_{opt}$. This results in a set of Pareto fronts, with the first front representing the current best set of non-dominated solutions.

For each compound in the population, two entities are initialized:

- *Domination Count* n_i: The number of compounds that dominate the given compound.
- *Dominated Set* S_i: A set of compounds that the given compound dominates.

In contrast to MO-CMA-ES, NSGA-II emphasizes the selection of offspring from a union of parent spaces. β parent pairs are selected randomly from the best non-dominated front. This is followed by a selection of λ number of offspring through a crossover mechanism that explores the union space of the each β parent pairs.

The iterative process of NSGA-II involves updating the population with new offspring, followed by another round of non-dominated sorting. Similar to MO-CMA-ES, we iterate this process until a run-time limit or stability in the solution set is reached. The adaptability in the number of parents and offspring generated (β and λ) is also maintained, as described in the MO-CMA-ES section, to ensure diversity and escape from local optima:

The final output of the NSGA-II implementation is a refined set of compounds, representing the most promising solutions according to the defined optimisation and constraint criteria. This process ensures a comprehensive exploration of the solution space, leveraging the strengths of NSGA-II's non-dominated sorting and unique offspring generation methodology.

3.5 Summary

The research methodology refined the application of MOEAs, specifically MO-CMA-ES and NSGA-II, for the sophisticated task of molecular design. It elaborated on the strategic configuration and algorithmic enhancements of these MOEAs to seamlessly integrate with the predictive insights of the *Uni-Mol* model. This tailored approach facilitates a targeted exploration within the molecular space, driven by the objective of identifying compounds that simultaneously maximize beneficial properties and minimize adverse ones. By generating a diverse set of potential solutions through an iterative optimisation process, this methodology underscores a balanced exploration and exploitation strategy. Leveraging accurate molecular property predictions to guide the search, it avoids the pitfalls where mere exploration can lead to decreased computational efficiency (since molecule datasets are extensive), and mere exploitation may result in premature convergence. This balanced approach ensures that the search is both thorough in discovering new possibilities and efficient in converging to optimal solutions.

Furthermore, the methodology introduced novel meta-heuristic strategies —direct correlation and extended search — to innovatively navigate the molecular landscape starting from known seed molecules. These meta-heuristics are designed to expand the search dynamically, either by delving deeper into the vicinity of seed molecules or by broadening the exploration to new regions, thereby enhancing the effectiveness of the MOEAs in discovering optimal molecular configurations. Through these methodological advancements, the research establishes a comprehensive and efficient framework for molecular design, illustrating the potential of customized MOEAs combined with predictive modeling for advanced molecule space exploration.

Chapter 4

Experiments

The Experiments section delineates a comprehensive approach to validate the efficacy and applicability of *Uni-Mol* and EAs in filtering and optimizing molecules with specific properties, primarily focusing on environmental safety and optimum property values in molecules.

Section 4.1 sets up experimenting ground for the application of EAs, specifically MO-CMA-ES and NSGA-II, in optimizing detergent molecules against a set of predefined objectives.

Section 4.2 explains the fine-tuning of *Uni-Mol* using a vast dataset of organic compounds to predict molecular toxicity.

The experiments are conducted on a high-performance system equipped with an AMD Ryzen 9 5950X 16 core CPU running at 4.9GHz with 32GB RAM and Nvidia 3080 Ti 12GB GPU.

4.1 Evolutionary Algorithms

In the experiments chapter, this section aims to address the primary research question by systematically responding to the secondary research questions 2.1 and 2.2 through designing an experimentation setup, as outlined in Table 4.1, that covers both secondary research questions.

Table 4.1: Overview of Experiment Setup

Parameters	**Setup**
MOEAs	MO-CMA-ES or NSGA-II
Meta-heuristics	Direct Correlation or Extended Search
Seed Molecule	CCCCCCC(C)CCCCCCCCCOS(=O)(=O)O
Objectives	Molecule weight, Molecule Complexity, XlogP, Reference Likeness, Fish Toxicity
$\mathcal{A}_{\text{opt}}$	Molecule weight (<), Molecule Complexity (<), XlogP (>), Reference Likeness (90%), Fish Toxicity (remove)
$\mathcal{A}_{\text{constr}}$	Molecule weight (< 500), Molecule Complexity (< 500), XlogP (> 4), Reference Likeness (>70%), Fish Toxicity (remove)
$\mathcal{A}_{\text{constr\_final}}$	Molecule weight (250<.<350), Molecule Complexity (250<.<350), XlogP (5<.<10), Reference Likeness (>70%), Fish Toxicity (remove)
$\hat{T}_0$	70%
$\hat{T}$	90%
Beta	10
Lambda	50

We will run MO-CMA-ES and NSGA-II with both meta heuristic approaches; direct correlation and extended search. The direct correlation approach will initiate with a seed patented detergent molecule, specifically CCCCCCC(C)CCCCCCCCCOS(=O)(=O)O (as illustrated in Figure 4.1). Conversely, the extended search approach will commence with an analysis of five selected compounds that exhibit an 80% similarity to the seed compound, thereby expanding the initial search parameters.

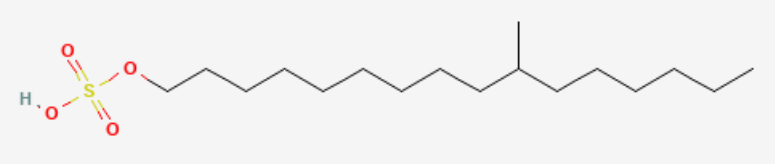


Figure 4.1: 2D structure of *CCCCCCC(C)CCCCCCCCCOS(=O)(=O)O*. This detergent compound functions as a surfactant due to its hydrophobic hydrocarbon chain, which attaches to oils and greases, and a hydrophilic sulfonate group that dissolves in water. This dual nature allows the molecule to form micelles, encapsulating oil particles and effectively removing them when washed away with water. Thus, it effectively breaks down and cleanses oily substances in various cleaning applications.

The optimisation objectives ($\mathcal{A}_{opt}$) for these experiments encompass several critical attributes: molecular weight, molecular complexity, XlogP, reference likeness, and fish toxicity. These attributes have been selected based on their relevance to the efficacy and environmental impact of the detergent molecules. The optimisation attributes are as follows:

- *Reference likeness* targeting an optimal similarity of 90%,
- Minimization of *molecular weight*,
- Minimization of *molecular complexity*,
- Maximization of *XlogP*, and
- Complete elimination of molecules featuring fish toxicity.

Within the experimental framework, setting a reference likeness threshold of 90% represents a strategic balance between maintaining similarity to proven compounds, such as the seed compound, and the introduction of novel structural properties. This methodology embodies a compelling approach to optimisation, where the objective is to fine-tune towards a specific value of likeness. Such a parameterization has a direct impact on the navigational behavior of the optimisation methods, as the search space itself is defined by a measure of similarity akin to this property. This setup is expected to subtly modulate the solution diversity produced by the optimisation methods, constraining their exploratory capacity to some extent. However, this constraint is a deliberate choice, serving as one of the essential factors in managing the trade-offs among various optimisation objectives.

Table 4.2: Solution Step Property Limits

Property	$\mathcal{A}_{constr}$	**Final Step** $\mathcal{A}_{constr}$
Reference Likeness	$\leq 70\%$	$\leq 70\%$
Molecule Weight	≤ 500	$250 \leq \cdot \leq 350$
Molecule Complexity	≤ 500	$250 \leq \cdot \leq 350$
XLogP	≥ 4	$5 \leq \cdot \leq 10$
Fish Toxicity	Remove	Remove

Table 4.3: Properties of the seed compound

Property	**Value**
Reference Likeness	1.00
Molecule Weight	336.54
Molecule Complexity	327.55
XLogP	5.53
Fish Toxicity	0

The minimization of molecular weight and complexity is aimed at reducing production costs, while a higher XlogP value is indicative of improved oil absorption capabilities—a desirable characteristic for detergents. Lastly, the exclusion of compounds with fish toxicity is a crucial consideration for environmental safety. Through these optimized parameters, our study seeks to explore and identify efficient multi-objective compound selections for chemical product design using EAs.

When it comes to optimisation constraints ($\mathcal{A}_{constr}$) and ($\mathcal{A}_{constr\_final}$) they are tuned for our experiments as presented in Table 4.2. These values are chosen according to our seed compound to be able to evaluate resulting solution space according to the seed compound. Its properties are listed in Table 4.3.

Parameters for the experiments can be seen in the Table 4.4. Choice of β and λ values are chosen according to some testings as these values were found to be optimal. For $\hat{T}_0$ 70% is chosen as we want to have high similarity to the seed molecule and for $\hat{T}$ 10% is chosen as we do not really want to methods to get stuck in the gaps of search space. The experiments were run for 50 generations for each method (for extended search meta-heuristic 10 generations for each of the 5 starting compounds) over 20 runs to examine the consistency. Convergence criteria is deliberately excluded to assess comparative method capabilities in escaping local optima and to examine exploratory behaviour. Solution set diversity is empirically assessed using the quantity of non-dominated molecules.

Table 4.4: Parameters for the experiments

Parameter	**Value**
β	10
λ	50
$\hat{T}_0$	70%
$\hat{T}$	90%

The experimental setup, as delineated in Table 4.4, specifies the parameters utilized across the optimisation processes. Notably, the values for β (parents selected) and λ (offsprings selected) are subject to dynamic adjustments during the runs, which is anticipated to minimize their direct impact on the experimental outcomes. This adaptive strategy ensures that the optimisation process remains responsive to the evolving search landscape, thereby enhancing its efficiency and effectiveness and being able to deal with empty gaps in the search space.

The selection of $\hat{T}_0$ at 70% underscores our intent to maintain a strong resemblance to the seed molecule at the outset of the optimisation, facilitating a targeted exploration of the molecular space that builds upon the seed's advantageous properties. Conversely, a higher threshold of 90% for $\hat{T}$ is chosen to prevent the optimisation methods from becoming overly constrained by the immediate vicinity of the search space, thus promoting a broader exploration and avoiding potential stagnation in local optima.

4.1.1 Evaluation of the Methods

In this section of our experiments chapter, we aim to address the primary research question by systematically responding to the secondary research questions 2.1 and 2.2 by generating a series of graphs and plots depicting the solution sets and the processes leading to their formulation. These visual representations will serve as the foundation for our analysis, enabling a comprehensive evaluation of the solution sets derived from the optimisation methods under investigation. By examining these graphs and plots, we intend to extract insights that will directly contribute to answering the secondary research questions. This analytical approach is designed to elucidate the comparative effectiveness and characteristics of the solution spaces generated by the respective optimisation strategies, thereby providing empirical evidence to support our conclusions regarding the optimisation of molecules for product design.

One of the evaluation of the algorithms will be conducted through a line plot graph of the unique molecules explored in each generation, providing a quantitative measure of the algorithms' search efficiency and breadth. This line graph will help us to answer secondary research question 2.1.

Subsequently, the search trends of the algorithms will also be visualized using a MDS plot. This plot will encompass the entire search space, the compounds examined during the search, and the found solution space, thereby offering a graphical representation of the algorithms' behavior within the complex molecular landscape. From the MDS plot, insights into the navigational strategies employed by the algorithms, including their exploration and exploitation tendencies, can be inferred. This visualization facilitates an intuitive understanding of how each algorithm traverses the search space, highlighting patterns of convergence towards optimal solutions or diversification across the space. These visualization plots will help us to answer secondary research question 2.1.

Further, the diversity and quantity of the solution spaces generated by each method will be assessed through the creation of box-plots for the properties of the solution spaces. These box-plots will provide a statistical summary of the distribution of key molecular properties within the solution spaces, enabling a comparative analysis of the diversity and quality of solutions identified by each algorithm. These box-plots will help us to answer secondary research question 2.2.

Runtime for each method will be logged across 50 generations, and to synthesize the findings, all solution spaces generated by the algorithms will be combined, followed by a non-dominated sorting process. This step is designed to identify the most successful compounds—those that are not outperformed across all objectives by any other compound in the combined dataset. The outcome of this non-dominated sorting will serve as a crucial metric for determining which algorithm was most effective in discovering superior molecules, thereby providing a comprehensive evaluation of the algorithms' performance in the context of multi-objective molecular selection for chemical product design. This finding will help us to answer secondary research question 2.2.

4.2 Uni-Mol

In order to effectively pre-train *Uni-Mol*, a large-scale data set composed of organic compounds will be used. The molecular pre-training data set used in our experiments consisted of approximately 19 million compounds, which were sourced from multiple public data sets. To obtain the 3D conformations, a combination of ETKGD and Merck Molecular Force Field optimisation from RDKit tool will be used to generate ten unique conformations for each compound. Additionally, a 2D conformation will be generated for each compound to address rare cases where 3D conformations could not be generated.

The dataset underpinning our experiments encompasses 251k molecules, with a mere 2% previously annotated for aquatic toxicity. To enhance training accuracy, our experimental design consolidates acute and chronic toxicity categories. To rigorously assess *Uni-Mol*, we segment the dataset into training, validation, and testing subsets, adhering to an 80-10-10 split ratio, utilizing a scaffold-split strategy. This method, based on the molecular scaffolds, presents more challenge than random splitting and is instrumental in evaluating the model's robustness.

Addressing the imbalance between toxic and non-toxic molecules within the dataset, we implement random oversampling for toxic molecules in the training set. Among the 20 models trained with varied hyperparameters, we identify and report on the performance of the model that exhibited superior efficacy on the test set. Training of these models will aim to achive the highest macro-average accuracy between toxic and non-toxic molecules, serves as the benchmark for our evaluation. This systematic approach ensures a comprehensive and robust evaluation of *Uni-Mol*, positioning it to effectively predict molecular toxicity.

This segment of the experimental investigation contributes to addressing the primary research question, though not representing the ultimate goal of this thesis. Its inclusion introduces an additional layer of complexity to the objectives of our Evolutionary Algorithms (EAs), serving as an enhancement rather than the core focus. Specifically, the aim is to extend beyond the utilization of molecule properties that can be determined through conventional chemical algorithms. Instead, we integrate molecule property prediction models capable of forecasting more intricate properties. This approach is designed to refine and enhance the process of identifying optimal molecules. By leveraging predictive models, we aim to incorporate a broader spectrum of molecular characteristics, thus facilitating a more comprehensive and informed search within the molecular optimisation framework. This addition, while valuable, should be viewed as an auxiliary feature that complements the primary objectives of employing advanced computational methods for molecular optimisation in product design.

4.3 Summary

In this chapter, we presented a detailed experimental setup for evaluating the performance of *Uni-Mol* and evolutionary algorithms in optimizing detergent molecules with specified objectives and constraints. We employed a large-scale dataset for fine-tuning *Uni-Mol*, focusing on molecular toxicity prediction, and used a scaffold-split strategy to enhance model robustness. For evolutionary algorithms, we explored two meta-heuristic approaches using MO-CMA-ES and NSGA-II, targeting optimisation of detergent molecules' molecular weight, complexity, XlogP, reference likeness, and eliminating fish toxicity. The experiments were designed to assess the algorithms' ability to generate diverse, non-dominated molecule solutions efficiently, with a focus on environmental safety and production cost-effectiveness. The methodologies and parameters were meticulously chosen to ensure a comprehensive evaluation of the algorithms' performance in navigating the complex search space for finding optimal compounds.

Chapter 5

Results

The Experiments Results chapter presents the findings obtained from the study's empirical investigations, structured into two primary sections, each focusing on a distinct aspect of the research.

Section 5.1 is dedicated to evaluating the performance of the *Uni-Mol* model, specifically its ability to predict the toxicity of compounds towards fish. Through a detailed Table 5.1, this section showcases the model's predictive accuracy across different epochs, illustrating the significant role of training duration on its performance.

Section 5.2 shifts the focus to the application of EAs in the optimisation process, exploring their efficiency through Direct Correlation and Extended Search heuristics. This section aims to compare the exploration capabilities of different optimisation methods under identical generational constraints. This comparative analysis underscores the impact of algorithmic designs on the methods' ability to explore solution spaces effectively, contributing valuable insights into the strengths and limitations of each optimisation approach.

5.1 Uni-Mol

Table 5.1 presents the performance metrics of the model that achieved the highest Macro Average Accuracy in predicting compounds' toxicity towards fish, as per the test data. The table is structured to display accuracies across several epochs, underlining the significant impact that the number of training epochs has on the model's predictive accuracy. This impact is particularly pronounced due to the imbalance inherent in the

dataset.

Epoch Num	Non-Toxic Accuracy (%)	Toxic Accuracy (%)	Average Accuracy (%)	Macro Average Accuracy (%)
1	55.2	86.3	55.6	70.8
2	**73.2**	**72.0**	**73.2**	**72.6**
3	84.2	58.5	83.9	71.4
4	91.3	38.4	90.6	64.9
5	90.3	42.1	89.7	66.2
6	94.6	28.7	93.7	61.6
7	96.0	25.9	95.1	61.0
8	93.8	32.3	93.0	63.1
9	96.1	22.9	95.1	59.5
10	96.2	23.2	95.2	59.7

Table 5.1: *Uni-Mol* prediction accuracies of the best model (model with highest Macro Average Accuracy) across its training epochs. **Epoch Num** refers to the sequential count of complete passes the model has made over the entire dataset during training. **Non-Toxic Accuracy** quantifies the model's precision in predicting compounds that are known to be non-toxic to fish. **Toxic Accuracy** quantifies the model's precision in predicting compounds that are known to be toxic to fish. **Average Accuracy:** represents the general prediction accuracy of the model across all labels. **Macro Average Accuracy:** mirrors the average accuracy but emphasizes equal consideration of all labels, providing a balanced measure of performance across categories. Bold highlights highest Macro Average Accuracy for all criteria.

The accuracies reported include the model's performance in identifying non-toxic compounds, its accuracy in predicting toxic compounds, the overall accuracy across all categories, and the Macro Average Accuracy. By delineating these accuracies across successive epochs, the table illustrates how the model's predictive capabilities evolve with additional training, providing insights into the dynamics of learning in response to dataset imbalances. This format allows for a nuanced understanding of the temporal development of model accuracy, highlighting the critical role of epoch progression in enhancing the model's ability to discern between toxic and non-toxic compounds accurately.

5.2 Evolutionary Algorithms

In this section, we present the outcomes obtained from our EAs, employing both the Direct Correlation and Extended Search heuristics. These outcomes are presented with line plot graph (Section 5.2.1), MDS plot graphs (Section 5.2.2), box-plot graphs (Section 5.2.3) and finally run-time and solution quality comparisons (Section 5.2.4 and 5.2.5).

Statistical tests, specifically the Friedman test ($p<0.05$), were conducted for each method over its 20 runs to assess the presence of variation between runs. The Friedman test revealed no significant differences across the runs, suggesting that each method's own runs consistently found very similar optimas. This consistency across runs justified our decision to select a random run from each method for diagram plotting. The choice of the Friedman test was further supported by preliminary analyses using the Kolmogorov-Smirnov normality test with Lilliefors correction, which indicated non-parametric distributions of the results (Ghasemi & Zahediasl, 2012; Sheldon et al., 1996).

Statistical analyses were performed to evaluate the differences in molecular properties that are being optimized in experiments. The Kruskal-Wallis test ($p<0.05$) (McKight & Najab, 2010) was employed for each property to assess the presence of significant differences among the methods. Where significant differences were identified, post-hoc Dunn's tests (Dinno, 2015) with Bonferroni correction (Weisstein, 2004) were conducted to determine pairwise differences between methods.

5.2.1 Exploration Line Plot Graph

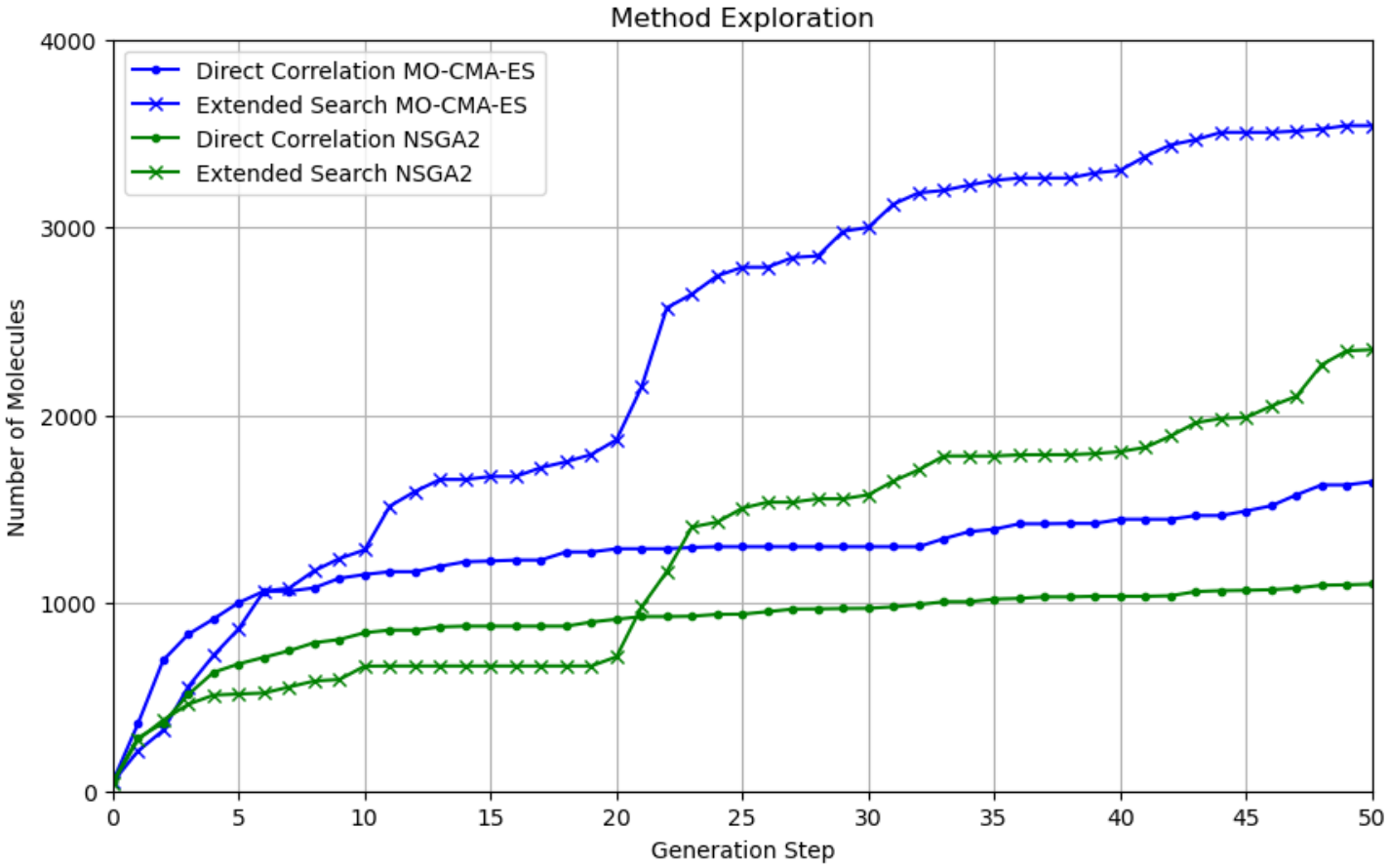


Figure 5.1: Discovered unique molecules per generation given selection by MO-CMA-ES, NSGA-II, direct correlation pruning and extended search pruning chosen from random run over 10 runs for each method.

The line plot diagram Figure 5.1 is designed to facilitate a comparative analysis of the exploration capabilities inherent to different optimisation methods from randomly chosen runs, as Friedman test (p<0.05) indicated there is no significant difference between runs in each method. Given that each method is constrained to operate over an identical number of generations (50 generations), the observed variations in exploration efficiency and strategy can be attributed solely to the architectural nuances of the methods themselves, rather than external factors such as the duration of the run or pre-mature convergence. This controlled setup ensures that any differences in the exploration patterns across methods are indicative of their inherent algorithmic designs and settings.

An initial search space, comprising approximately 700,000 molecules in SMILES format, was sourced from the *PubChem* database, and then refined to 30,000 molecules considering a neighbourhood around seed detergent molecule. This refinement process

utilized the MACCS Tanimoto similarity measure, with a threshold ($\hat{T}_0$) set at 70%.

This Figure 5.1 highlights which method demonstrates a more explorative approach by visualizing the extent to which each method probes various regions of the search space over the course of the optimisation process. Methods that exhibit a greater degree of exploration are expected to sample a broader and potentially more diverse set of solutions, reflecting a proactive search behavior that avoids premature convergence on local optima and seeks out global optima with greater efficiency.

5.2.2 Visualization of Search Behaviour

Figures 5.2, 5.3, 5.4, and 5.5 present search space topology corresponding to various elements: the reference molecule (in the case of direct correlation this is starting seed molecule), starting molecules (in the case of extended search these are starting molecules derived from seed molecule), frontier (final solution set), searched space (all offspring), and search space for all the methods. To facilitate these visualizations, molecules were sampled based on a uniform distribution. Both the search space and the searched space were down-scaled by a factor of five relative to their original volume. This reduction was essential to produce images that are both visually interpretable and informative.

The MDS visualization plots use MACCS Tanimoto similarity serving as the metric for distance measurement. In these visualizations, the axes, denoted as *Similarity Distance*, inversely correlates with the Tanimoto similarities between molecules. 0.1 distance indicates 10% similarity distance. The Grid circles also expand with 10% similarity.

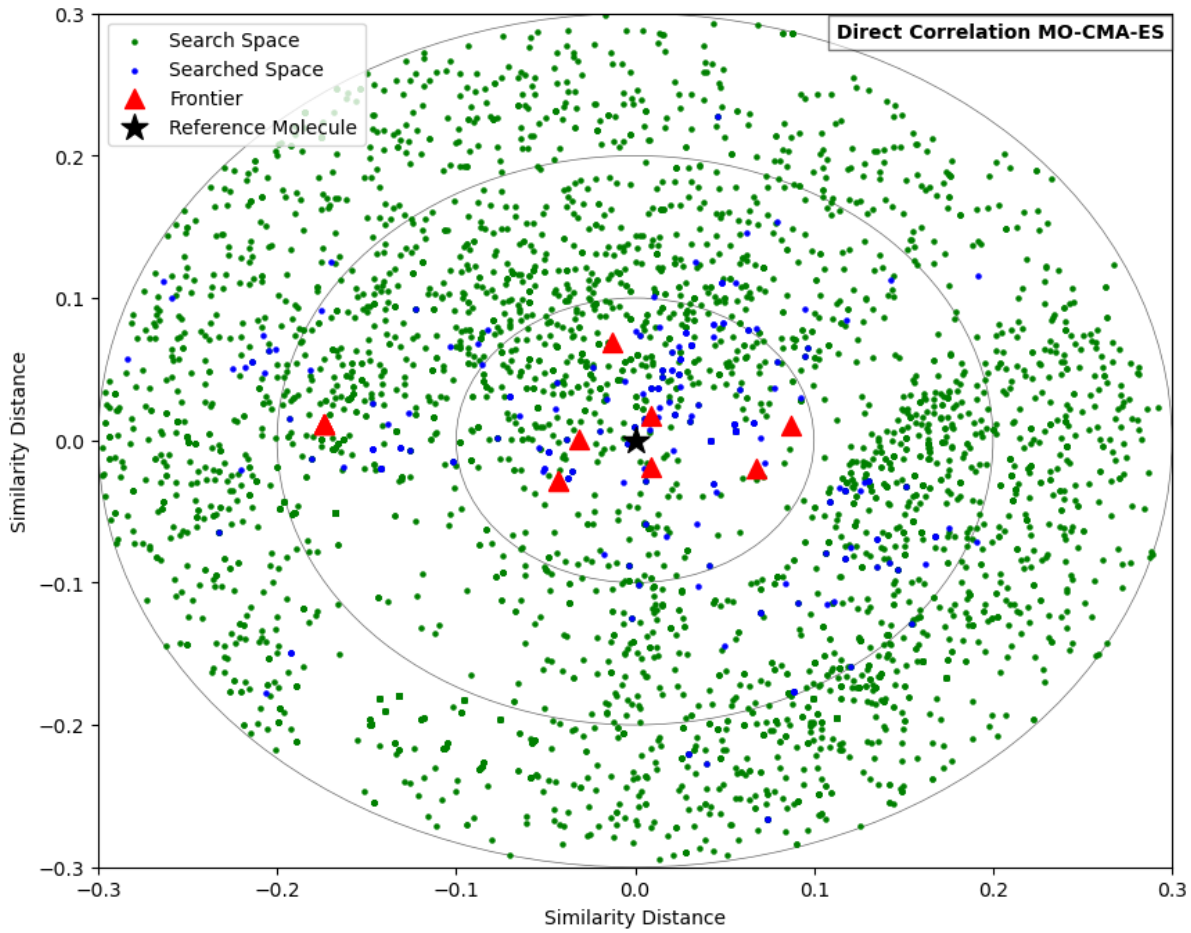


Figure 5.2: This figure depicts an MDS visualization of the MO-CMA-ES algorithm's exploration strategy using the Direct Correlation meta-heuristic, chosen from random run over 10 runs. It highlights three key areas: unexplored compounds, explored molecules, and the optimally selected solution set, with the seed detergent molecule as the reference point for initiation. The visualization employs a similarity distance metric, where a 0.1 unit indicates a 10% difference in similarity, and uses expanding grid circles to represent this variance, effectively illustrating the algorithm's explored part of the molecular search space.

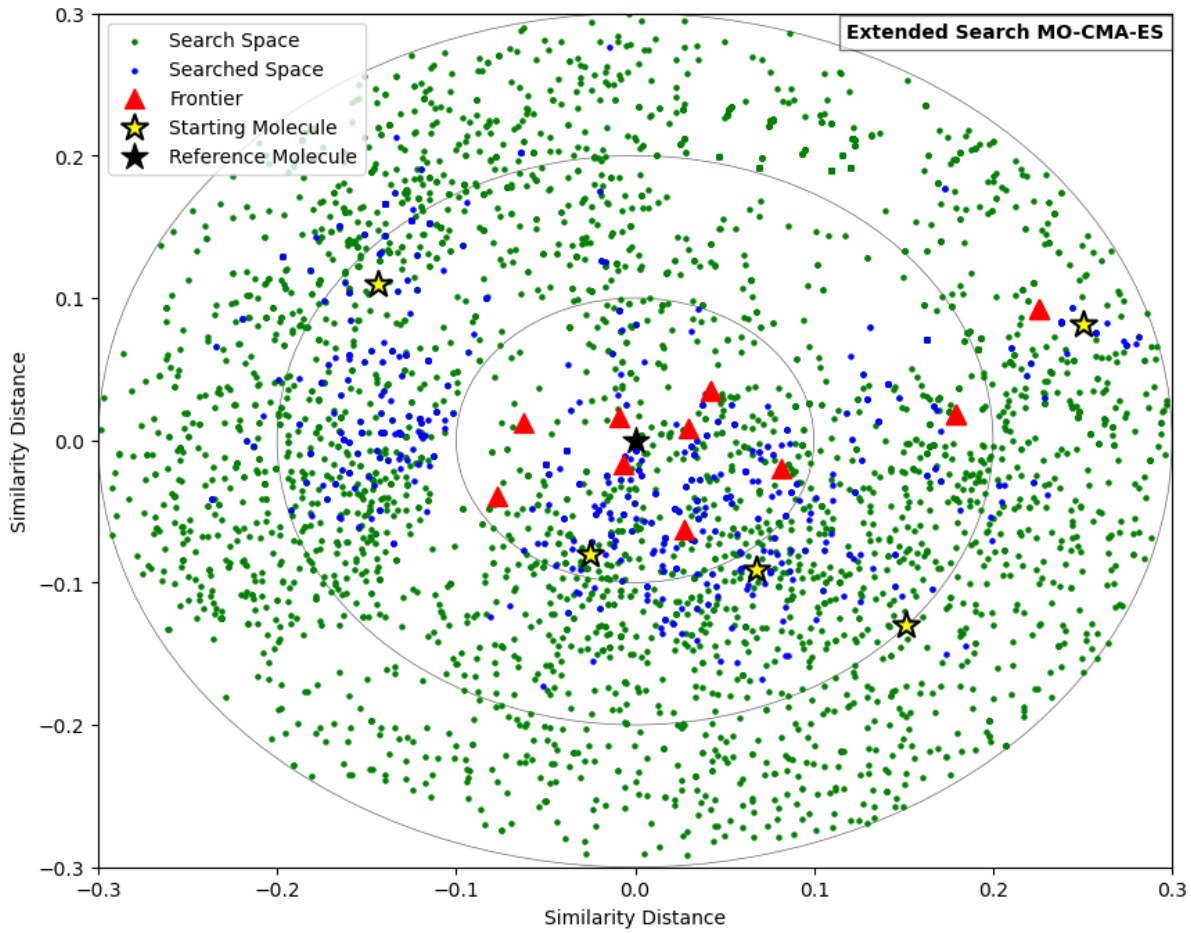


Figure 5.3: This figure depicts an MDS visualization of the MO-CMA-ES algorithm's exploration strategy using the Extended Search meta-heuristic, chosen from random run over 10 runs. It highlights three key areas: unexplored compounds, explored molecules, and the optimally selected solution set, with the seed detergent molecule as the reference point for initiation. The visualization employs a similarity distance metric, where a 0.1 unit indicates a 10% difference in similarity, and uses expanding grid circles to represent this variance, effectively illustrating the algorithm's explored part of the molecular search space.

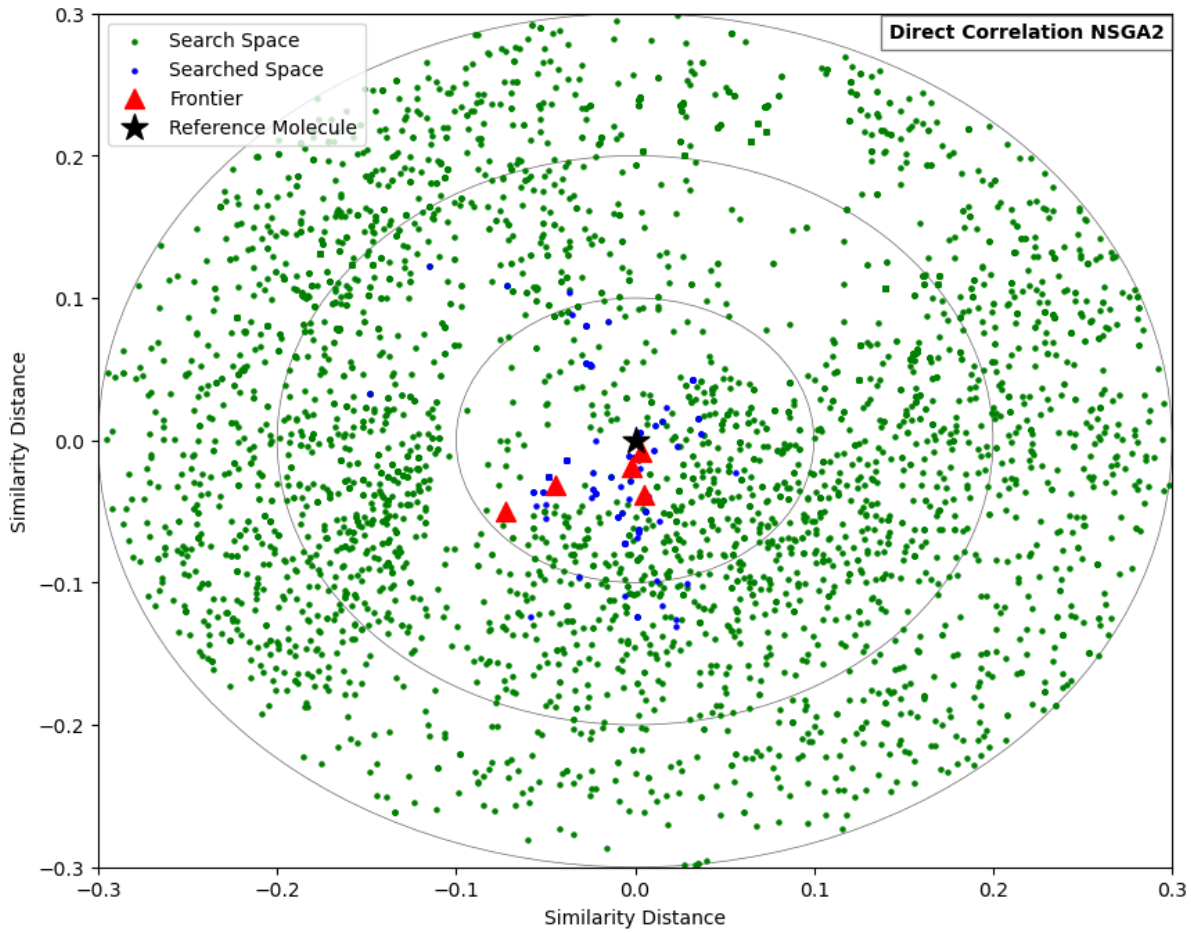


Figure 5.4: This figure depicts an MDS visualization of the NSGA-II algorithm's exploration strategy using the Direct Correlation meta-heuristic, chosen from random run over 10 runs. It highlights three key areas: unexplored compounds, explored molecules, and the optimally selected solution set, with the seed detergent molecule as the reference point for initiation. The visualization employs a similarity distance metric, where a 0.1 unit indicates a 10% difference in similarity, and uses expanding grid circles to represent this variance, effectively illustrating the algorithm's explored part of the molecular search space.

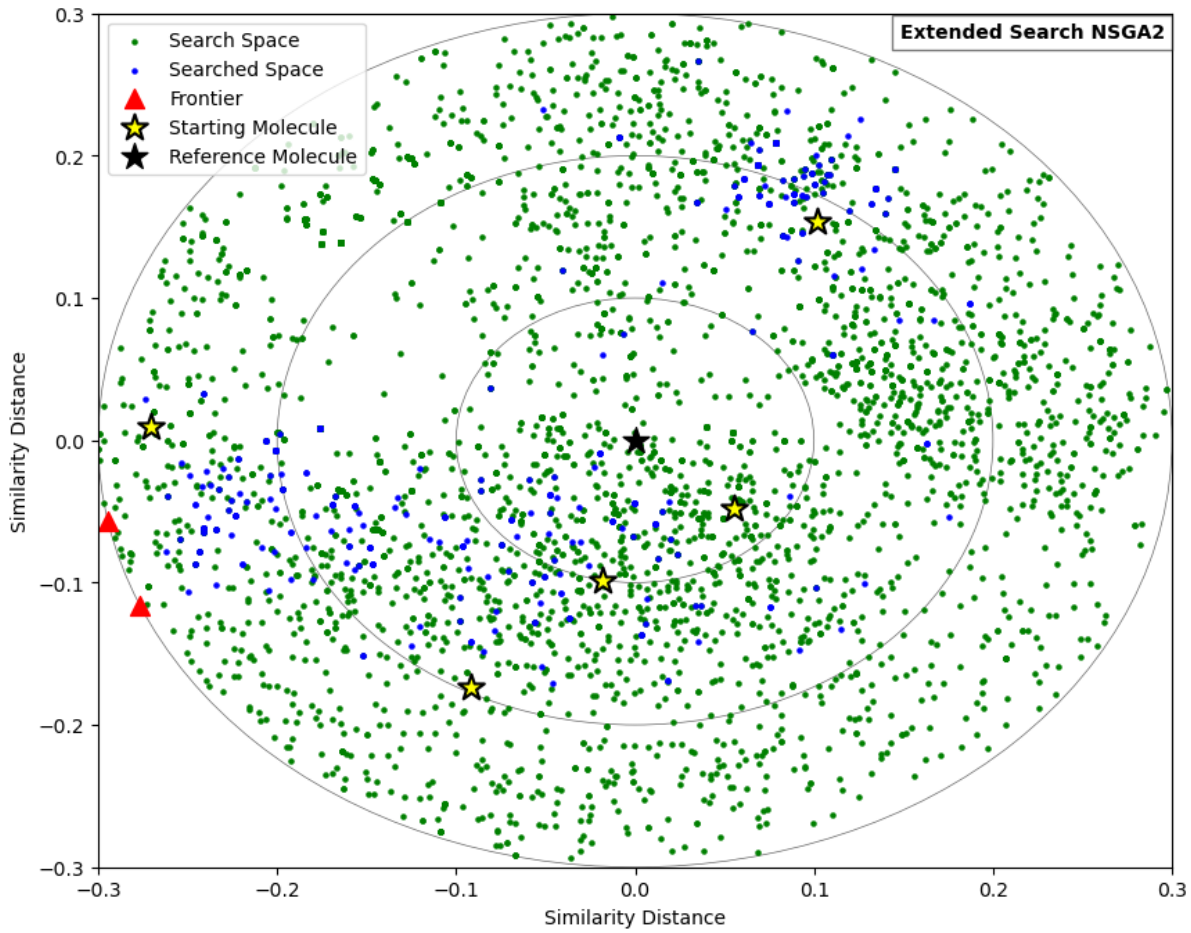


Figure 5.5: This figure depicts an MDS visualization of the NSGA-II algorithm's exploration strategy using the Extended Search meta-heuristic, chosen from random run over 10 runs. It highlights three key areas: unexplored compounds, explored molecules, and the optimally selected solution set, with the seed detergent molecule as the reference point for initiation. The visualization employs a similarity distance metric, where a 0.1 unit indicates a 10% difference in similarity, and uses expanding grid circles to represent this variance, effectively illustrating the algorithm's explored part of the molecular search space.

5.2.3 Box-Plot Diagrams of Solution Spaces

Figures 5.6, 5.7, 5.8, and 5.9 display box plots that compare the quality of the final solution sets with respect to various optimized variables: molecular complexity, molecular weight, XlogP, and reference likeness. The data for these variables was normalized within a range of 0 to 1, based on the maximum and minimum values derived from the Final Step $\mathcal{A}_{constr}$ as outlined in table 4.2. Reference likeness was already normalized, given its inherent representation as a percentage value. Random run was chosen from the 10 runs for each method to plot these box-plots.

The properties of the seed compound are prominently featured in the box-plots to enable a detailed comparison with the properties of the solution set. This approach allows for an analytical assessment of how the characteristics of the compounds identified by the EAs align with the benchmark properties of the seed detergent. Through this comparative analysis, we can discern the extent to which the solution set mirrors the desired attributes of the seed molecule, highlighting the efficacy and precision of the search and optimisation process.

Statistical analyses employing the Kruskal-Wallis test ($p<0.05$) followed by post-hoc Dunn's test with Bonferroni correction reveal no significant differences in the performance of MOO approaches regarding the optimisation outcomes. In contrast, notable distinctions were identified through box-plot between MO-CMA-ES, NSGA-II direct correlation and NSGA-II with extended search concerning molecular weight (as illustrated in Figure 5.6), molecular complexity (Figure 5.7), and reference likeness (Figure 5.9). These findings suggest a more stable and better efficacy of MO-CMA-ES methods in optimizing these molecular properties. Conversely, for the property XlogP, the box-plot reveal no significant differences across the examined methods, indicating a comparable performance in optimizing this specific criterion.

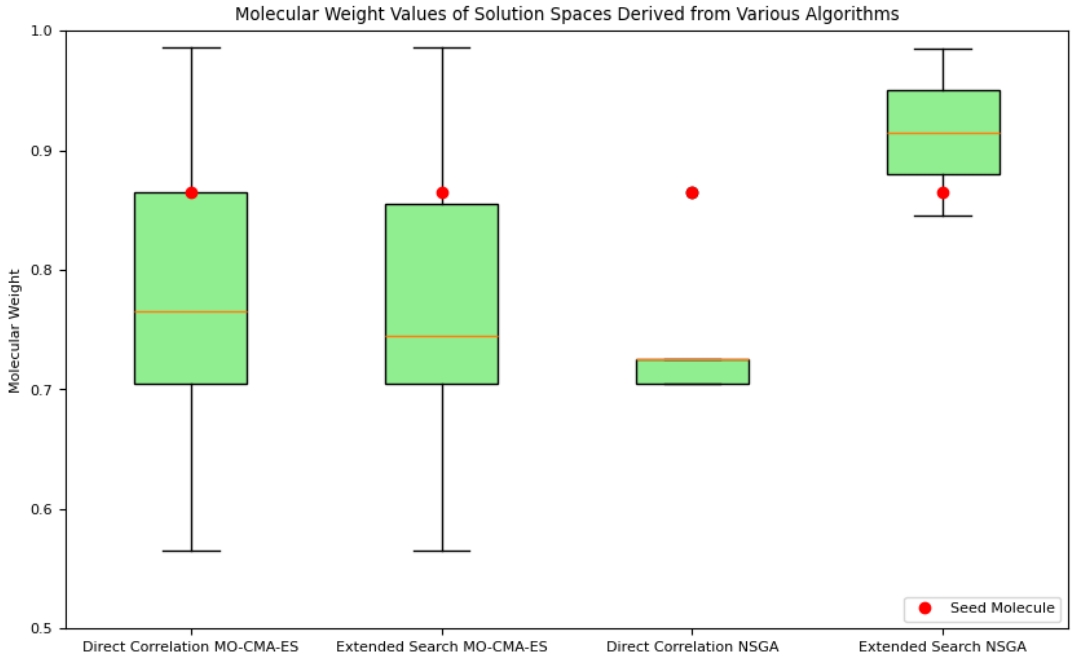


Figure 5.6: Box-plot Comparisons of Molecule Weight Properties of Compounds in Solution Spaces Derived from Various Methods (chosen from random run over 10 runs). The weight of the seed molecule is highlighted with a red dot to facilitate direct comparison. The properties of the molecules have been normalized to fall within the range [0,1].

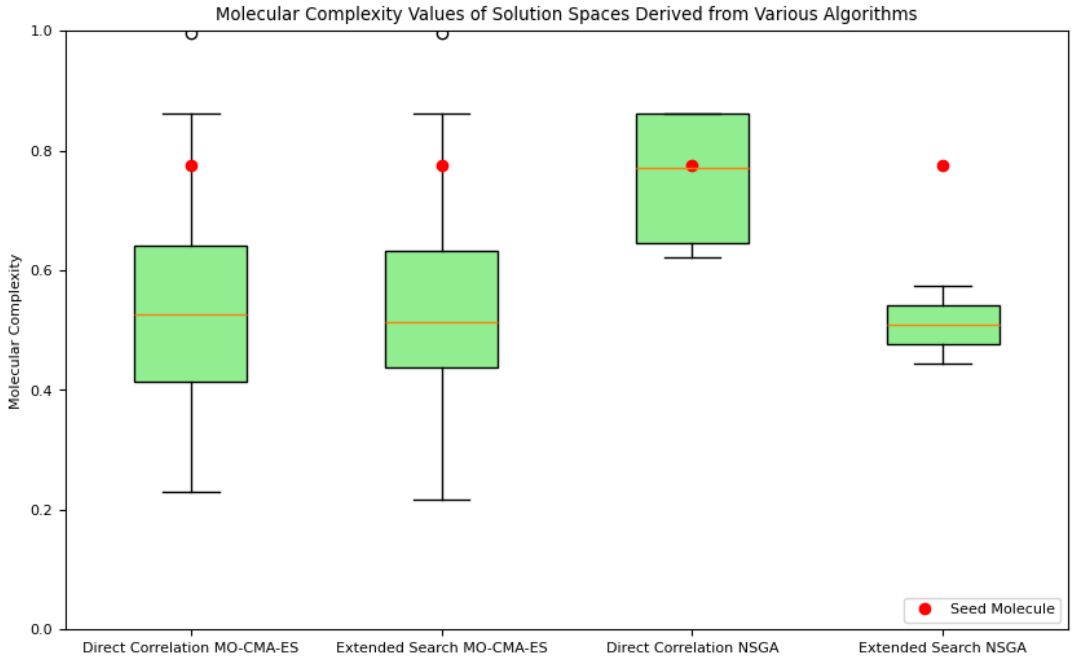


Figure 5.7: Box-plot Comparisons of Molecule Complexity Properties of Compounds in Solution Spaces Derived from Various Methods (chosen from random run over 10 runs). The complexity of the seed molecule is highlighted with a red dot to facilitate direct comparison. The properties of the molecules have been normalized to fall within the range [0,1].

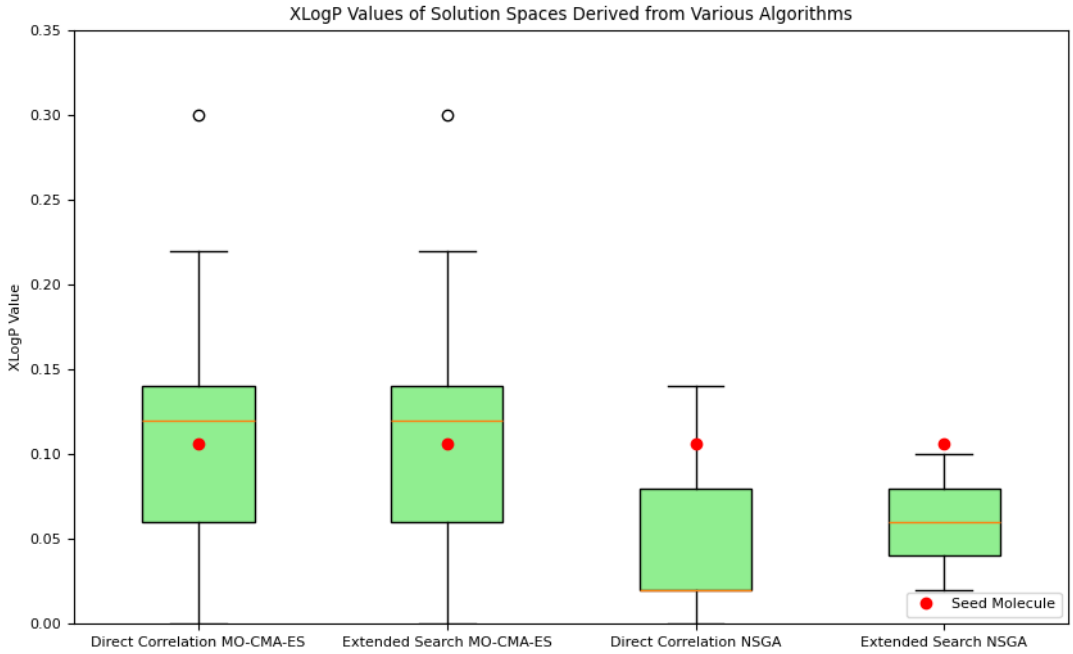


Figure 5.8: Box-plot Comparisons of Molecule XlogP Properties of Compounds in Solution Spaces Derived from Various Methods (chosen from random run over 10 runs). The XlogP of the seed molecule is highlighted with a red dot to facilitate direct comparison. The properties of the molecules have been normalized to fall within the range [0,1].

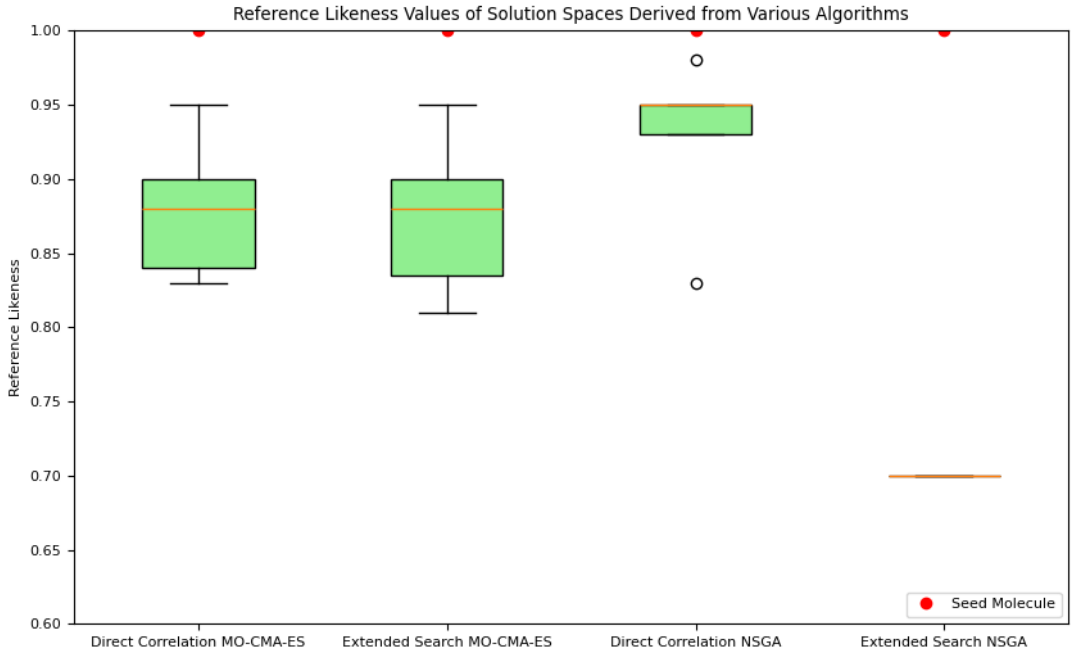


Figure 5.9: Box-plot Comparisons of Reference Likeness of Compounds in Solution Spaces Derived from Various Methods (chosen from random run over 10 runs). The reference likeness of the seed molecule is highlighted with a red dot which is 100%. The values are percentage and fall within the range [0,1].

5.2.4 Runtime of the Methods

EA	MH	SS Size	Time (min)
MO-CMA-ES	Direct Correlation	13	15
MO-CMA-ES	Extended Search	15	13
NSGA-II	Direct Correlation	6	30
NSGA-II	Extended Search	2	55

Table 5.2: Comparison of Solution Sets and Computational Time. EA: Evolutionary Algorithm, MH: Meta-heuristic, SS Size: Solution Set Size, Time: Time required in minutes for 50 generations.

Table 5.2 presents runtime for each method experimented. MO-CMA-ES and direct correlation produced a solution set comprising 13 molecules within a time frame of 15 minutes. When extended search pruning was employed, MO-CMA-ES generated a slightly larger solution set of 15 molecules, but remarkably, in a reduced duration of 13 minutes. In contrast, NSGA-II and direct correlation yielded a solution set of only 6 molecules after 30 minutes. Furthermore, when extended search was used, the solution set was limited to only two molecules, and the algorithm required the significantly longer period of 55 minutes.

5.2.5 Quality of Solution Sets

EA	MH	Contributed Molecules
MO-CMA-ES	Direct Correlation	12
MO-CMA-ES	Extended Search	15
NSGA-II	Direct Correlation	1
NSGA-II	Extended Search	0

Table 5.3: Contributions to the Collective Solution Set. EA: Evolutionary Algorithm, MH: Meta-heuristic, Contributed Molecules: Number of molecules contributed by each method to the collective set of 17 non-dominated molecules.

Table 5.3 presents when we consider the combined solution sets generated by all methods, we observed the following. Upon aggregating all the non-dominated molecules, the collective set comprised 17 molecules. Given this set, MO-CMA-ES with direct correlation, contributed 12 molecules, while MO-CMA-ES with extended search contributed with 15 molecules. However, NSGA-II with direct correlation contributed with only one molecule and NSGA-II with extended search did not contribute any molecules.

5.3 Summary

In this chapter, we systematically presented the experimental results, focusing on the predictive accuracy of a molecule toxicity model then the exploration and optimisation capabilities of EAs, and the comparative quality of their solution sets. Key findings include the significant impact of training epochs on model accuracy, the superior exploration and solution diversity offered by the MO-CMA-ES algorithm, especially when combined with extended search meta-heuristics, and the comprehensive evaluation of solution sets across various molecular properties. The results underscored the effectiveness of the employed computational methods in navigating complex molecular spaces, offering insights into their applicability for optimizing molecule design for specific properties.

Chapter 6

Discussion

This chapter provides an in-depth analysis and interpretation of the findings in relation to the research questions posed in this study. It comprises two main sections, each addressing critical aspects of chemical space exploration and compound optimisation using EAs and the *Uni-Mol* model.

Section 6.1 addresses the primary research question by detailing the integration of the *Uni-Mol* model with EAs. It highlights the balance between training duration and accuracy in predicting compound toxicity, identifying the limitations of the SMILES format and suggesting improvements for model precision in future research.

Section 6.2 directly addresses secondary research questions 2.1 and 2.2, evaluating the exploration dynamics and solution quality generated by MO-CMA-ES and NSGA-II in the realm of chemical space exploration. This section underscores the exceptional performance of MO-CMA-ES in navigating complex molecular landscapes, showcasing its ability to produce diverse and highly optimized solution sets. It highlights the instrumental role of strategic exploration methodologies in augmenting the effectiveness of EAs for advanced compound identification, thereby facilitating a more nuanced and efficient approach to chemical compound discovery.

6.1 Predicting Fish Toxicity with *Uni-Mol*

This study advances the field of chemical space exploration for optimal compound identification by integrating *Uni-Mol* with the EAs, addressing the primary research question with a sophisticated approach. The experimental observations reveal a strong trade-off between the training duration (epochs) and the model's proficiency in distinguishing toxic from non-toxic compounds, as detailed in Section 5.1, Table 5.1. Notably, enhanced accuracy in recognizing toxic molecules minimizes false positives, while improved detection of non-toxic molecules reduces false negatives. An increase in training epochs consistently amplifies the accuracy for non-toxic molecule identification, simultaneously diminishing toxic molecule detection accuracy, a phenomenon likely attributed to the data set's skewed distribution of non-toxic labels.

In selecting a model for our EAs, we prioritized the one exhibiting the optimal balance of toxic and non-toxic detection accuracies (macro average accuracy), as emphasized in bold within Section 5.1, Table 5.1. This chosen model demonstrates an efficacy on par with, yet distinct advantages over, existing methodologies (Chen & et al., 2018; Pu et al., 2019; Schneider, 2018; Schneider & Fechner, 2005) due to its avoidance of the traditional, resource-intensive optimisation processes.

Despite the promising accuracy of our premier model (referenced in Section 5.1, Table 5.1), it is crucial to acknowledge its limitations in fully replacing laboratory testing. Nevertheless, it provides valuable preliminary insights into the fish toxicity of compounds, facilitating the early-stage chemical analysis. The ambition of this thesis extends beyond mere predictive accuracy; the model, even with its current limitations, is instrumental in guiding EAs towards the discovery of optimal compounds.

To further refine the model's accuracy, the utilization of the SMILES format for generating three-dimensional structures is identified as a potential limitation. The SMILES format, despite its ability to represent molecular structures, fails to capture the complexity of toxicological responses, as evidenced by molecules with identical SMILES representations exhibiting divergent toxicological profiles. Advancing the model's predictive accuracy necessitates exploring alternative input modalities capable of encapsulating the nuanced spatial characteristics of molecules, thereby enhancing the model's reliability and utility in the optimisation process.

6.2 Optimizing Compound Discovery in Chemical Space Using Evolutionary Algorithms

In Chapter 5, we presented findings pertinent to the secondary research questions 2.1 and 2.2. This section delves into a comprehensive discussion and analysis of the algorithms' exploration strategies (Section 6.2.1) as well as the diversity and quality of solutions generated (Section 6.2.2).

6.2.1 Analyzing the Exploratory Dynamics of Evolutionary Algorithms in Chemical Space

Section 5.2.1, illustrated by Figure 5.1, we delve into the comparative analysis of the unique molecules identified throughout the evolutionary search processes employed by MO-CMA-ES and NSGA-II. A particular focus is given to the impact of utilizing direct correlation versus extended search strategies. The implementation of extended search notably enhanced the diversity of the solution sets, a phenomenon that was especially pronounced when applied in conjunction with MO-CMA-ES. Relative to NSGA-II, MO-CMA-ES exhibited a consistent superiority in generating a broader array of solution sets, affirming its efficacy in maintaining elevated levels of diversity throughout the evolutionary search, while being a lot more computationally efficient (Section 5.2.4, Table 5.2). These observations reinforce the adeptness of MO-CMA-ES in navigating complex solution spaces, underscoring its value in the context of molecular optimisation.

Further insights into the search behaviors of these methods are presented in Section 5.2.2, through Figures 5.2, 5.3, 5.4, and 5.5. These visualizations facilitate a nuanced comparison of the exploratory efficiency and effectiveness of MO-CMA-ES versus NSGA-II and their meta-heuristics. Highlighted within these figures are the noticeable gaps within the solution space, representing regions that pose significant challenges to both methods. These gaps often funnel the exploration efforts towards local optima, presenting obstacles to the attainment of global solutions. This analysis sheds light on the inherent complexities of the solution space and the adaptive responses of the EAs in their quest to identify optimal molecular configurations.

Section 5.2.2, as depicted in Figure 5.2, showcases the MDS visualization for MO-CMA-ES employing direct correlation. This visualization reveals a tendency of the method to remain within close proximity to the initial seed compound throughout the search process. This phenomenon is attributed to the objective of achieving a 90% likeness to the reference molecule, which confines the search primarily within the smallest grid circle indicative of this threshold. Nevertheless, the algorithm occasionally ventures further afield in pursuit of exploration, discovering non-dominated compound that maintain roughly 80% similarity to the reference molecule. This behavior indicates a balance between explorative actions and the avoidance of becoming ensnared in local optima.

In contrast, Section 5.2.2 Figure 5.3 elucidates the MDS visualization for MO-CMA-ES with extended search, illustrating a markedly more explorative approach than its direct correlation counterpart. The initiation of search from molecules distanced from the reference molecule facilitates a broader exploration of the search space, with a noticeable trend of movement towards the reference molecule driven by the likeness objective. This strategy diverges from the outward movement observed in the direct correlation setting, highlighting the dynamic exploration capabilities of MO-CMA-ES when augmented with extended search criteria.

Section 5.2.2 Figure 5.4 examines the MDS visualization for NSGA-II utilizing direct correlation. This method appears to struggle in navigating the search space, failing to identify non-dominated compounds even within close proximity to the seed. This limitation suggests a constrained exploratory scope, potentially hindering the method's ability to uncover viable molecular configurations.

Lastly, Section 5.2.2 Figure 5.5 provides the MDS visualization for NSGA-II with extended search, demonstrating a superior exploration of the search space relative to its direct correlation iteration. Despite this expanded exploratory effort, the method falls short in identifying compounds that satisfy the constraints ($\mathcal{A}constr$) and ($\mathcal{A}constr\_final$), indicating challenges in balancing exploration with the achievement of specified optimisation criteria.

These observations collectively underscore the distinct behaviors of MO-CMA-ES and NSGA-II under different meta-heuristics, emphasizing the critical role of initial conditions and objective settings in influencing the exploration dynamics and solution quality within complex chemical spaces.

The extended search implementation of MO-CMA-ES underscores its enhanced ability to navigate the search space more effectively, discovering a broader and more diverse set of non-dominated molecules. This is contrasted with the narrower scope of exploration and the found solution sets of NSGA-II, which tend to exhibit greater similarity to the seed molecules and are less varied in their distribution across the molecular property objectives.

The empirical findings suggest that the adoption of NSGA-II might lead to a more constrained and potentially less innovative exploration process when juxtaposed with the capabilities of MO-CMA-ES. This is further evidenced by the solution sets generated by NSGA-II, which not only closely resemble the seed molecules in terms of chemical structure but also are quantitatively smaller, indicating a potential limitation in the algorithm's ability to foster innovation within the chemical space.

An intriguing aspect of the study's findings is the observation that the spatial distribution of feasible molecules, as depicted in Section 5.2.2, Figures 5.2, and 5.3, reveals areas of sparsity within the search space. This characteristic poses a challenge especially to NSGA-II's strategy, which inherently seeks to diversify solutions but is constrained by the algorithm's similarity threshold for including solutions in the solution sets.

The contrast between conservative and more assertive exploration strategies, such as those exemplified by direct correlation versus extended search, respectively, illuminates the inherent differences in the MOO capabilities of MO-CMA-ES and NSGA-II. Specifically, the application of extended search strategies has been shown to accentuate these differences, facilitating the fine-tuning of hyper-parameters to optimize the exploration of search spaces and the identification of viable solution sets.

Empirical evidence from the study highlights MO-CMA-ES's superior performance in search-space exploration over NSGA-II, across various exploration strategies. This performance advantage is attributed to MO-CMA-ES's more liberal criteria for non-dominated selection and its less conservative approach in offspring selection, enabling a broader exploration of the search space and yielding a more diverse and populated set of solutions.

Furthermore, the findings suggest that the offspring selection mechanism of MO-CMA-ES plays a pivotal role in its effectiveness, surpassing the impact of offspring selection diversity. This mechanism facilitates continuous exploration and prevents the premature convergence seen in NSGA-II, thereby underscoring the robustness of MO-CMA-ES in maintaining an ongoing discovery process of new non-dominated molecules, as supported by the results presented in figure 5.1. These insights affirm the critical influence of offspring selection diversity on the performance of MOO algorithms in navigating and populating complex solution spaces.

In addressing the secondary research question 2.1 concerning the comparative performance of MO-CMA-ES and NSGA-II in the context of navigating the complex molecular space for identifying molecules with optimal properties, the empirical evidence and analytical observations presented in this study elucidate distinct advantages and limitations inherent to each algorithm. MO-CMA-ES demonstrates superior capability in exploring and exploiting the molecular space, facilitated by its adaptive strategy that effectively balances exploration and exploitation through a dynamic adjustment of its meta-heuristic approaches. This adaptability enables MO-CMA-ES to discover a broader and more diverse array of non-dominated solutions, signifying its effectiveness in identifying molecules with optimal properties amidst the vast and complex molecular landscape.

Conversely, NSGA-II, while being very exploitative in its approach, exhibits limitations in its exploratory scope, often resulting in solution sets that are less varied and more closely aligned with the initial seed molecules. This tendency, coupled with a proneness towards premature convergence in certain contexts, underscores a potential constraint in NSGA-II's ability to innovate and identify novel compounds within the chemical space.

The comparison thus reveals that MO-CMA-ES's exploratory dynamics, characterized by a more flexible and adaptive mechanism for navigating the search space, render it more efficacious than NSGA-II in the context of this study's objectives. MO-CMA-ES's proficiency in generating diverse and expansive solution sets, coupled with its robustness against premature convergence, positions it as the more favorable algorithm for the task of identifying molecules with optimal properties within the complex molecular space.

6.2.2 Assessing the Solution Quality of Evolutionary Algorithms in Compound Optimisation

The assessment of solution diversity is predicated on the cardinality of the solution sets, with the underlying assumption that a greater cardinality indicates a richer diversity of alternatives for experimental scrutiny. Solution quality is evaluated through independent box plot analyses of each property, where optimized mean values and reduced variance for each property are indicative of superior solution quality.

Section 5.2.3, Figures 5.6, 5.7, 5.8, and 5.9, box plots elucidate the final solution sets' quality as determined by the EAs across various optimized variables, including molecular complexity, weight, XlogP, and reference likeness.

In the context of minimizing molecular weight to reduce potential detergent compound production costs (Cheng et al., 2020), Section 5.2.3, Figure 5.6 showcases that MO-CMA-ES yielded a diverse array of solutions with generally lower molecular weights compared to the initial seed detergent molecule. The results from both of it's meta-heuristics were somewhat similar, indicating no significant difference. NSGA-II's solution sets, especially those from the extended search, were markedly heavier (box-plot analysis), diverging significantly from both MO-CMA-ES's solutions and its own solutions under direct correlation, which exhibited lesser diversity.

Similarly, for molecular complexity, aimed to be minimized to lower production costs, MO-CMA-ES consistently identified compounds with reduced complexity relative to the seed, across both heuristic strategies. In contrast, NSGA-II's performance varied significantly (box-plot analysis) between strategies, with its direct correlation approach yielding compounds close in complexity to the seed molecule, whereas its extended search identified compounds with notably lower molecular complexity. MO-CMA-ES found molecules significantly lower (box-plot analysis) in terms of molecular complexity compared to NSGA-II.

Regarding XlogP, aimed to be maximized to enhance oil solubility, MO-CMA-ES identified molecules with a broader range of XlogP values, including those higher than the seed molecule's. Conversely, NSGA-II was less successful in this regard, often failing to find molecules surpassing the seed molecule's XlogP. Yet, there was no significant difference between methods.

In terms of similarity to the seed molecule, while MO-CMA-ES effectively identified compounds around the 90% similarity mark, its extended search demonstrated a capacity for greater exploration, identifying less similar molecules, aligning with the goal of balancing property trade-offs. NSGA-II, however, showed a stark difference (box-plot analysis) between its search strategies, with direct correlation finding highly similar molecules and extended search yielding compounds at the lower threshold of 70% similarity, deviating from the optimisation objectives. Solution sets produced by MO-CMA-ES was significantly closer (box-plot analysis) to 90% than solution sets from NSGA-II.

MO-CMA-ES exhibited a consistent ability to discover compounds with lower molecule complexity, molecule weight and higher XlogP values compared to NSGA-II, regardless of the heuristic applied. This suggests an overall superiority of MO-CMA-ES in optimizing for compound complexity, weight, and solubility properties. Moreover, MO-CMA-ES's performance in maintaining reference likeness across strategies contrasts with NSGA-II's variable outcomes, further underscoring the former's stability and the latter's sensitivity to search strategy.

These findings, supported by the comprehensive comparison in Section 5.2.5, Table 5.3, where MO-CMA-ES's solutions outperformed those of NSGA-II across all properties, highlight the differential exploration efficacy of these MOO methods in discovering optimized molecules. MO-CMA-ES's adaptability and exploratory depth enable it to uncover a wider array of optimized solutions, thereby proving its efficacy over NSGA-II in the realm of molecular optimisation within complex chemical spaces.

This analysis addresses secondary research question 2.2, focusing on the comparative performance of MO-CMA-ES and NSGA-II in the context of identifying molecules with optimal properties. The findings reveal that MO-CMA-ES, regardless of the specific meta-heuristic applied, significantly surpasses NSGA-II in its efficacy to discover molecules that excel across various desired attributes. Notably, MO-CMA-ES demonstrates superior capability in identifying molecules that are not only of lower weight and complexity—which implies a potential reduction in production costs—but also exhibit higher XlogP values, enhancing oil solubility. These enhancements open up new possibilities for discovering detergents that are more optimal than the seed reference detergent, offering avenues for innovation in detergent formulation that are cheaper to produce and more oil soluble. Additionally, this approach successfully maintains close reference likeness to the predetermined criteria. These results underscore the robustness and efficiency of MO-CMA-ES in optimizing molecular properties, thereby offering a compelling answer to the posed research question and illustrating its practical advantages in molecular design optimisation.

6.3 Summary

In this chapter, the study's results were analyzed and interpreted in light of the research questions, focusing on the EAs for chemical space exploration and compound optimisation and innovative integration of the *Uni-Mol* model. Two central sections form the crux of the discussion, each providing critical insights into the study's findings.

The section 6.1 directly tackled the primary research question, illustrating how the integration of *Uni-Mol* with EAs enhanced the prediction accuracy for toxic versus non-toxic compounds. This analysis underlines a significant trade-off between training duration and model accuracy, shedding light on the limitations inherent to the SMILES format and suggesting pathways for refining model precision.

In addressing the secondary research questions, Optimizing Compound Discovery with EAs evaluates the exploratory dynamics and solution quality produced by MO-CMA-ES and NSGA-II. This section 6.2 highlights MO-CMA-ES's superior capability in generating diverse and optimized solutions, emphasizing the pivotal role of exploration strategies in bolstering the efficiency of compound discovery processes.

The findings underscored the MO-CMA-ES algorithm's adeptness in navigating the complex molecular landscape, affirming its effectiveness in identifying a broader, more diverse array of optimized solutions compared to NSGA-II. This comparison elucidates the nuanced strategies and dynamics at play in EAs exploration of chemical spaces, offering valuable perspectives on algorithm selection and the strategic deployment of meta-heuristics for compound optimisation.

Overall, this chapter not only dissected the intricacies of molecular optimisation and toxicity prediction but also paves the way for future research directions aimed at enhancing the accuracy, efficiency, and applicability of computational models and EAs in the realm of chemical product design.

Chapter 7

Conclusions

The primary objective of this thesis is to assess the exploration and efficacy of MOEAs, specifically MO-CMA-ES and NSGA-II, in the optimisation of molecules within a complex landscape, augmented by molecular property prediction models. This investigation aimed to enhance the automated molecular discovery process through a comprehensive comparative analysis of these leading MOEAs, supplemented by the introduction of novel meta-heuristic approaches, Direct Correlation and Extended Search. The contributions of this research offer significant advancements in the methodology of molecule design and optimisation.

Firstly, the comparative analysis between MO-CMA-ES and NSGA-II, enhanced by the innovative meta-heuristics, illuminated the adaptability, efficiency, and scalability of MOEAs in navigating the intricate molecular spaces. This detailed examination provided a framework for selecting the most suitable MOEAs for specific molecular optimisation tasks. The empirical evidence and detailed analysis underpinning this framework allow for a strategic selection of computational strategies, optimizing the likelihood of success in various molecular design contexts.

Furthermore, the introduction of the Direct Correlation and Extended Search meta-heuristics represents a strategic advancement in the exploration and optimisation of the molecular search space. These meta-heuristics, focusing on outward and inward exploration respectively, offer a nuanced approach to navigating the search space, thereby enhancing the effectiveness of automated molecular optimisation applications. The empirical evaluations of these approaches have yielded critical insights, contributing significantly to the field's understanding and development.

The interdisciplinary integration of algorithmic objectives with the predictive capabilities of Molecular Property Prediction Models stands as a final contribution of this thesis. This approach infuses the design process with predictive insights. Such integration facilitates the achievement of more targeted and sophisticated molecular designs, marking a transformative step forward in the precision and innovation of molecular product design.

The insights gleaned from this research suggest that MO-CMA-ES, particularly when augmented with the Extended Search heuristic, is a superior choice for addressing complex MOO problems in molecular discovery, specifically over NSGA-II. This algorithm, supported by the novel meta-heuristics, has demonstrated the potential to generate a diverse and viable set of molecules with desirable properties, thus establishing MO-CMA-ES as significantly better in molecule optimisation.

In conclusion, the contributions of this thesis significantly advance the automated molecular discovery field, enhancing the understanding of MOEAs in complex optimisation scenarios and paving the way for future research at the intersection of computational chemistry and MOEAs. Through a strategic framework for algorithm selection, novel meta-heuristic approaches, and the integration of predictive modeling, positions MO-CMA-ES as a choice for MOO in the ongoing advancement of chemical product design and automated molecular discovery.

7.1 Future Directions

To extend the contributions of this thesis, several avenues for future work are proposed:

1. **Integration of Generative Models:** The development and implementation of molecule generative models capable of interpolating and extrapolating within the search space could uncover novel molecules with optimal properties. Fine-tuning these models with the discovered optimal molecules could further enhance their effectiveness.
2. **Advancements in Molecular Representation:** Exploring beyond the limitations of SMILES representations could offer more accurate and diverse molecular characterizations. Alternative representations may better capture molecular similarities and differences, facilitating improved optimisation processes.

3. **Conservation of Molecular Substructures:** Incorporating an objective to conserve specific molecular substructures during the optimisation process with MOEAs could address the requirements of certain chemical products that necessitate particular substructures. This approach would ensure that essential molecular features are retained in the optimized molecules.

These proposed directions aim to not only address the limitations identified through this research but also to push the boundaries of what is currently possible in automated molecular discovery and optimisation. By pursuing these avenues, future research can continue to enhance the efficiency and applicability of computational methods in chemical product design, contributing to the discovery of innovative molecules with significant commercial impacts.

Bibliography

Alves, V. M., Muratov, E., Fourches, D., Strickland, J., Kleinstreuer, N., Andrade, C. H., & Tropsha, A. (2015). Predicting chemically-induced skin reactions. part i: Qsar models of skin sensitization and their application to identify potentially hazardous compounds. *Toxicology and applied pharmacology*, *284*(2), 262–272.

Bajusz, D., Rácz, A., & Héberger, K. (2015). Why is tanimoto index an appropriate choice for fingerprint-based similarity calculations? *Journal of cheminformatics*, *7*(1), 1–13.

Bender, A., & Cortes-Ciriano, I. (2021). Artificial Intelligence in Drug Discovery: What Is Realistic, What Are Illusions? *Drug Discovery Today*, *26*(1), 1040.

Bengio, Y., Courville, A., & Vincent, P. (2013). Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, *35*(8), 1798–1828.

Bento, A. P., Hersey, A., Félix, E., Landrum, G., Gaulton, A., Atkinson, F., Bellis, L. J., De Veij, M., & Leach, A. R. (2020). An open source chemical structure curation pipeline using rdkit. *Journal of Cheminformatics*, *12*, 1–16.

Bilodeau, C., Jin, W., Jaakkola, T., Barzilay, R., & Jensen, K. F. (2022a). Generative models for molecular discovery: Recent advances and challenges. *Wiley Interdisciplinary Reviews: Computational Molecular Science*, *12*(5), e1608.

Bilodeau, C., Jin, W., Jaakkola, T., Barzilay, R., & Jensen, K. F. (2022b). Generative models for molecular discovery: Recent advances and challenges. *Wiley Interdisciplinary Reviews: Computational Molecular Science*, *12*(5), e1608.

Brabazon, A., & O'Neill, M. (2006). *Biologically inspired algorithms for financial modelling*. Springer Science & Business Media.

Brown, N., & et al. (2004). A Graph-based Genetic Algorithm and its Application to the Multiobjective Evolution of Median Molecules. *Journal of Chemical Information and Modeling*, *44*(1), 1079–1087.

Brown, N. (2009). Chemoinformatics—an introduction for computer scientists. *ACM Computing Surveys (CSUR)*, *41*(2), 1–38.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. (2020). Language models are few-shot learners. *Advances in neural information processing systems*, *33*, 1877–1901.

Butler, K. T., Davies, D. W., Cartwright, H., Isayev, O., & Walsh, A. (2018). Machine learning for molecular and materials science. *Nature*, *559*(7715), 547–555.

Chen, R., & et al. (2018). Machine Learning for Drug-Target Interaction Prediction. *Molecules*, *23*(9), 2208.

Cheng, K. C., Khoo, Z. S., Lo, N. W., Tan, W. J., & Chemmangattuvalappil, N. G. (2020). Design and performance optimisation of detergent product containing binary mixture of anionic-nonionic surfactants. *Heliyon*, *6*(5).

Coello, C. C. (2006). Evolutionary multi-objective optimization: A historical view of the field. *IEEE computational intelligence magazine*, *1*(1), 28–36.

Coello, C. A. C. (2007). *Evolutionary algorithms for solving multi-objective problems*. Springer.

Crum-Brown, A., & Fraser, T. (1865). The connection of chemical constitution and physiological action. *Trans R Soc Edinb*, *25*(1968-1969), 257.

Curtarolo, S., & et al. (2013). The High-throughput Highway to Computational Materials Design. *Nature Materials*, *12*(1), 191–201.

De Cao, N., & Kipf, T. (2018). Molgan: An implicit generative model for small molecular graphs. *arXiv preprint arXiv:1805.11973*.

Deb, K., Pratap, A., Agarwal, S., & Meyarivan, T. (2002). A fast and elitist multiobjective genetic algorithm: Nsga-ii. *IEEE transactions on evolutionary computation*, *6*(2), 182–197.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Dinno, A. (2015). Nonparametric pairwise multiple comparisons in independent groups using dunn's test. *The Stata Journal*, *15*(1), 292–300.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Douguet, D., Thoreau, E., & Grassy, G. (2000). A genetic algorithm for the automated generation of small organic molecules: Drug design using an evolutionary algorithm. *Journal of computer-aided molecular design*, *14*, 449–466.

Duvenaud, D. K., Maclaurin, D., Iparraguirre, J., Bombarell, R., Hirzel, T., Aspuru-Guzik, A., & Adams, R. P. (2015). Convolutional networks on graphs for learning molecular fingerprints. *Advances in neural information processing systems*, *28*.

Eiben, A. E., & Smith, J. E. (2015). *Introduction to evolutionary computing*. Springer.

Emadi, G., Rahmani, A. M., & Shahhoseini, H. (2017). Task scheduling algorithm using covariance matrix adaptation evolution strategy (cma-es) in cloud computing. *Journal of Advances in Computer Engineering and Technology*, *3*(3), 135–144.

Fang, X., Liu, L., Lei, J., He, D., Zhang, S., Zhou, J., Wang, F., Wu, H., & Wang, H. (2022). Geometry-enhanced molecular representation learning for property prediction. *Nature Machine Intelligence*, *4*(2), 127–134.

Fasel, U., Keidel, D., Molinari, G., & Ermanni, P. (2017). Aerostructural optimization of a morphing wing for airborne wind energy applications. *Smart Materials and Structures*, *26*(9), 095043.

Feinberg, E. N., Sur, D., Wu, Z., Husic, B. E., Mai, H., Li, Y., Sun, S., Yang, J., Ramsundar, B., & Pande, V. S. (2018). Potentialnet for molecular property prediction. *ACS central science*, *4*(11), 1520–1530.

Ghasemi, A., & Zahediasl, S. (2012). Normality tests for statistical analysis: A guide for non-statisticians. *International journal of endocrinology and metabolism*, *10*(2), 486.

Gillet, V. J., Khatib, W., Willett, P., Fleming, P. J., & Green, D. V. (2002). Combinatorial library design using a multiobjective genetic algorithm. *Journal of chemical information and computer sciences*, *42*(2), 375–385.

Gómez-Bombarelli, R., & et al. (2018). Automatic Chemical Design using a Data-driven Continuous Representation of Molecules. *ACS Central Science*, *4*(2), 268–276.

Gómez-Bombarelli, R., Wei, J. N., Duvenaud, D., Hernández-Lobato, J. M., Sánchez-Lengeling, B., Sheberla, D., Aguilera-Iparraguirre, J., Hirzel, T. D., Adams, R. P., & Aspuru-Guzik, A. (2018). Automatic chemical design using a data-driven continuous representation of molecules. *ACS central science*, *4*(2), 268–276.

Guimaraes, G. L., Sanchez-Lengeling, B., Outeiral, C., Farias, P. L. C., & Aspuru-Guzik, A. (2017). Objective-reinforced generative adversarial networks (organ) for sequence generation models. *arXiv preprint arXiv:1705.10843*.

Gunantara, N. (2018). A review of multi-objective optimization: Methods and its applications. *Cogent Engineering*, *5*(1), 1502242.

Hamilton, W. L., Ying, R., & Leskovec, J. (2017). Representation learning on graphs: Methods and applications. *arXiv preprint arXiv:1709.05584*.

Handsel, J., Matthews, B., Knight, N. J., & Coles, S. J. (2021). Translating the inchi: Adapting neural machine translation to predict iupac names from a chemical identifier. *Journal of cheminformatics*, *13*(1), 1–11.

Hansch, C., & Fujita, T. (1964). P-σ-π analysis. a method for the correlation of biological activity and chemical structure. *Journal of the American Chemical Society*, *86*(8), 1616–1626.

Hansen, N., Müller, S. D., & Koumoutsakos, P. (2003). Reducing the time complexity of the derandomized evolution strategy with covariance matrix adaptation (cma-es). *Evolutionary computation*, *11*(1), 1–18.

Hansen, N., & Ostermeier, A. (2001). Completely derandomized self-adaptation in evolution strategies. *Evolutionary computation*, *9*(2), 159–195.

He, W., Qiao, P.-L., Zhou, Z.-J., Hu, G.-Y., Feng, Z.-C., & Wei, H. (2018). A new belief-rule-based method for fault diagnosis of wireless sensor network. *IEEE Access*, *6*, 9404–9419.

Heller, S. R., McNaught, A., Pletnev, I., Stein, S., & Tchekhovskoi, D. (2015). Inchi, the iupac international chemical identifier. *Journal of cheminformatics*, *7*(1), 1–34.

Hu, W., Liu, B., Gomes, J., Zitnik, M., Liang, P., Pande, V., & Leskovec, J. (2019). Strategies for pre-training graph neural networks. *arXiv preprint arXiv:1905.12265*.

Hughes, J. P., Rees, S., Kalindjian, S. B., & Philpott, K. L. (2011). Principles of early drug discovery. *British journal of pharmacology*, *162*(6), 1239–1249.

Igel, C., Hansen, N., & Roth, S. (2007). Covariance matrix adaptation for multi-objective optimization. *Evolutionary computation*, *15*(1), 1–28.

Igel, C., Suttorp, T., & Hansen, N. (2006). A computational efficient covariance matrix update and a (1+ 1)-cma for evolution strategies. *Proceedings of the 8th annual conference on Genetic and evolutionary computation*, 453–460.

Jensen, J. (2019). A Graph-based Genetic Algorithm and Generative Model/Monte Carlo Tree Search for the Exploration of Chemical Space. *Chemical Science*, *10*(12), 3567–3572.

Jimoh, A. A., & Lin, J. (2019). Biosurfactant: A new frontier for greener technology and environmental sustainability. *Ecotoxicology and Environmental safety*, *184*, 109607.

Jin, W., Barzilay, R., & Jaakkola, T. (2020). Hierarchical generation of molecular graphs using structural motifs. *International conference on machine learning*, 4839–4848.

Keith, J., & et al. (2021). Combining Machine Learning and Computational Chemistry for Predictive Insights Into Chemical Systems. *Chemical Reviews*, *121*, 9816–9872.

Kim, S., Chen, J., Cheng, T., Gindulyte, A., He, J., He, S., Li, Q., Shoemaker, B. A., Thiessen, P. A., Yu, B., et al. (2019a). Pubchem 2019 update: Improved access to chemical data. *Nucleic acids research*, *47*(D1), D1102–D1109.

Kim, S., Chen, J., Cheng, T., Gindulyte, A., He, J., He, S., Li, Q., Shoemaker, B. A., Thiessen, P. A., Yu, B., et al. (2019b). Pubchem 2019 update: Improved access to chemical data. *Nucleic acids research*, *47*(D1), D1102–D1109.

Krenn, M., Häse, F., Nigam, A., Friederich, P., & Aspuru-Guzik, A. (2019). Selfies: A robust representation of semantically constrained graphs with an example application in chemistry. *arXiv preprint arXiv:1905.13741*, *1*(3).

Kusner, M. J., Paige, B., & Hernández-Lobato, J. M. (2017). Grammar variational autoencoder. *International conference on machine learning*, 1945–1954.

Kuwahara, H., & Gao, X. (2021). Analysis of the effects of related fingerprints on molecular similarity using an eigenvalue entropy approach. *Journal of Cheminformatics*, *13*, 1–12.

Kwon, Y., & et al. (2021). Evolutionary Design of Molecules based on Deep Learning and a Genetic Algorithm. *Nature Scientific Reports*, *11*(17304), 4–6.

Kwon, Y., & Lee, J. (2021). Molfinder: An evolutionary algorithm for the global optimization of molecular properties and the extensive exploration of chemical space using smiles. *Journal of cheminformatics*, *13*, 1–14.

Landrum, G., et al. (2013). Rdkit: A software suite for cheminformatics, computational chemistry, and predictive modeling. *Greg Landrum*.

Le, T., & Winkler, D. (2016). Discovery and Optimization of Materials using Evolutionary Approaches. *Chemical Reviews*, *116*(1), 6107–6132.

Leguy, J., & al. (2009). EVOMOL: A Flexible and Interpretable Evolutionary Algorithm for Unbiased de novo Molecular Generation. *Journal of Cheminformatics*, *12*(55), 1–19.

Li, P., Wang, J., Qiao, Y., Chen, H., Yu, Y., Yao, X., Gao, P., Xie, G., & Song, S. (2021). An effective self-supervised framework for learning expressive molecular global representations to drug discovery. *Briefings in Bioinformatics*, *22*(6), bbab109.

Lim, J., Hwang, S.-Y., Moon, S., Kim, S., & Kim, W. Y. (2020). Scaffold-based molecular design with a graph generative model. *Chemical science*, *11*(4), 1153–1164.

Loshchilov, I., & Hutter, F. (2016). Cma-es for hyperparameter optimization of deep neural networks. *arXiv preprint arXiv:1604.07269*.

Lwin, K., Qu, R., & Kendall, G. (2014). A learning-guided multi-objective evolutionary algorithm for constrained portfolio optimization. *Applied Soft Computing*, *24*, 757–772.

Mayr, A., Klambauer, G., Unterthiner, T., & Hochreiter, S. (2016). Deeptox: Toxicity prediction using deep learning. *Frontiers in Environmental Science*, *3*, 80.

McKight, P. E., & Najab, J. (2010). Kruskal-wallis test. *The corsini encyclopedia of psychology*, 1–1.

Mitchell, J. (2014). Machine learning methods in chemoinformatics. *Wiley Interdisciplinary Reviews: Computational Molecular Science*, *4*(5), 468–481.

Mohammed, R., Rawashdeh, J., & Abdullah, M. (2020). Machine learning with oversampling and undersampling techniques: Overview study and experimental results. *2020 11th international conference on information and communication systems (ICICS)*, 243–248.

Montavon, G., Hansen, K., Fazli, S., Rupp, M., Biegler, F., Ziehe, A., Tkatchenko, A., Lilienfeld, A., & Müller, K.-R. (2012). Learning invariant representations

of molecules for atomization energy prediction. *Advances in neural information processing systems*, *25*.

Namasivayam, V., & Bajorath, J. (2012). Multiobjective particle swarm optimization: Automated identification of structure–activity relationship-informative compounds with favorable physicochemical property distributions. *Journal of chemical information and modeling*, *52*(11), 2848–2855.

Ng, K., & Gani, R. (2019). Chemical Product Design: Advances in and Proposed Directions for Research and Teaching. *Computers & Chemical Engineering*, *126*, 147–156.

Nicolaou, C. A., Brown, N., & Pattichis, C. S. (2007). Molecular optimization using computational multi-objective methods. *Current Opinion in Drug Discovery and Development*, *10*(3), 316.

Pal, S. K., Bandyopadhyay, S., & Ray, S. S. (2006). Evolutionary computation in bioinformatics: A review. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, *36*(5), 601–615.

Paul, D. (2021). Artificial Intelligence in Drug Discovery and Development. *Drug Discovery Today*, *26*(1), 80–93.

Polishchuk, P. (2017). Interpretation of quantitative structure–activity relationship models: Past, present, and future. *Journal of Chemical Information and Modeling*, *57*(11), 2618–2639.

Pu, L., Naderi, M., Liu, T., Wu, H., Mukhopadhyay, S., & Brylinski, M. (2019). eToxPred: A Machine Learning-based Approach to Estimate the Toxicity of Drug Candidates. *BMC Pharmacology and Toxicology*, *20*(2), 1–15.

Pyzer-Knapp, E., & et al. (2015). What is High-throughput Virtual screening? A Perspective from Organic Materials Discovery. *Annual Review of Materials Research*, *45*(1), 195–216.

Radford, A., Narasimhan, K., Salimans, T., Sutskever, I., et al. (2018). Improving language understanding by generative pre-training.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., et al. (2019). Language models are unsupervised multitask learners. *OpenAI blog*, *1*(8), 9.

Rahimi, I., Gandomi, A. H., Nikoo, M. R., & Chen, F. (2023). A comparative study on evolutionary multi-objective algorithms for next release problem. *Applied Soft Computing*, 110472.

Reed, P. M., Hadka, D., Herman, J. D., Kasprzyk, J. R., & Kollat, J. B. (2013). Evolutionary multiobjective optimization in water resources: The past, present, and future. *Advances in water resources*, *51*, 438–456.

Reymond, J. (2015). The Chemical Space Project. *Accounts of Chemical Research*, *48*, 722–730.

Riniker, S., & Landrum, G. A. (2015). Better informed distance geometry: Using what we know to improve conformation generation. *Journal of chemical information and modeling*, *55*(12), 2562–2574.

Rong, Y., Bian, Y., Xu, T., Xie, W., Wei, Y., Huang, W., & Huang, J. (2020). Self-supervised graph transformer on large-scale molecular data. *Advances in Neural Information Processing Systems*, *33*, 12559–12571.

Rosen, M. J., & Kunjappu, J. T. (2012). *Surfactants and interfacial phenomena*. John Wiley & Sons.

Russom, C. L., Bradbury, S. P., Broderius, S. J., Hammermeister, D. E., & Drummond, R. A. (1997). Predicting modes of toxic action from chemical structure: Acute toxicity in the fathead minnow (pimephales promelas). *Environmental Toxicology and Chemistry: An International Journal*, *16*(5), 948–967.

Samanta, B., & et al. (2019). NeVAE: A Deep Generative Model for Molecular Graphs. *Proceedings of the AAAI Conference on Artificial Intelligence*, 1110–1117.

Schneider, G. (2018). Automating Drug Discovery. *Nature Reviews Drug Discovery*, *17*(2), 97–113.

Schneider, G., & Fechner, U. (2005). Computer-based de novo Design of Drug-like Molecules. *Nature Reviews Drug Discovery*, *4*(1), 649–663.

Shekar, C., & Shivakumar, M. (2019). Multi-objective wind farm layout optimization using evolutionary computations. *Int. J. of Adv. in Appl. Sci. Vol*, *8*(4), 293–306.

Sheldon, M. R., Fillyaw, M. J., & Thompson, W. D. (1996). The use and interpretation of the friedman test in the analysis of ordinal-scale data in repeated measures designs. *Physiotherapy Research International*, *1*(4), 221–228.

Tadros, T. F. (2006). *Applied surfactants: Principles and applications*. John Wiley & Sons.

Tan, U., Rabaste, O., Adnet, C., & Ovarlez, J.-P. (2019). On the eclipsing phenomenon with phase codes. *2019 International Radar Conference (RADAR)*, 1–5.

Tkatchenko, A. (2020). Machine Learning for Chemical Discovery. *Nature Communications*, *11*, 4125.

Varela, D., & Santos, J. (2022). Niching Methods Integrated with a Differential Evolution Memetic Algorithm for Protein Structure Prediction. *Swarm and Evolutionary Computation*, *71*(1), 101062.

Vo-Duy, T., Duong-Gia, D., Ho-Huu, V., Vu-Do, H. C., & Nguyen-Thoi, T. (2017). Multi-objective optimization of laminated composite beam structures using nsga-ii algorithm. *Composite Structures*, *168*, 498–509.

Walters, W. P., & Barzilay, R. (2020). Applications of deep learning in molecule generation and molecular property prediction. *Accounts of chemical research*, *54*(2), 263–270.

Wang, S., Guo, Y., Wang, Y., Sun, H., & Huang, J. (2019). Smiles-bert: Large scale unsupervised pre-training for molecular property prediction. *Proceedings of the 10th ACM international conference on bioinformatics, computational biology and health informatics*, 429–436.

Wang, Y., Wang, J., Cao, Z., & Barati Farimani, A. (2022). Molecular contrastive learning of representations via graph neural networks. *Nature Machine Intelligence*, *4*(3), 279–287.

Weininger, D. (1988a). Smiles, a chemical language and information system. 1. introduction to methodology and encoding rules. *Journal of chemical information and computer sciences*, *28*(1), 31–36.

Weininger, D. (1988b). Smiles, a chemical language and information system. 1. introduction to methodology and encoding rules. *Journal of chemical information and computer sciences*, *28*(1), 31–36.

Weisstein, E. W. (2004). Bonferroni correction. *https://mathworld. wolfram. com/*.

Willett, P. (2006). Similarity-based virtual screening using 2d fingerprints. *Drug discovery today*, *11*(23-24), 1046–1053.

Winder, C., Azzi, R., & Wagner, D. (2005). The development of the globally harmonized system (ghs) of classification and labelling of hazardous chemicals. *Journal of hazardous materials*, *125*(1-3), 29–44.

Winter, R., Montanari, F., Noé, F., & Clevert, D.-A. (2019a). Learning continuous and data-driven molecular descriptors by translating equivalent chemical representations. *Chemical science*, *10*(6), 1692–1701.

Winter, R., Montanari, F., Steffen, A., Briem, H., Noé, F., & Clevert, D.-A. (2019b). Efficient multi-objective molecular optimization in a continuous latent space. *Chemical science*, *10*(34), 8016–8024.

Winter, R., Montanari, F., Steffen, A., Briem, H., Noé, F., & Clevert, D.-A. (2019c). Efficient multi-objective molecular optimization in a continuous latent space. *Chemical science*, *10*(34), 8016–8024.

Wu, Z., Ramsundar, B., Feinberg, E. N., Gomes, J., Geniesse, C., Pappu, A. S., Leswing, K., & Pande, V. (2018). Moleculenet: A benchmark for molecular machine learning. *Chemical science*, *9*(2), 513–530.

Xu, Z., Wang, S., Zhu, F., & Huang, J. (2017). Seq2seq fingerprint: An unsupervised deep molecular embedding for drug discovery. *Proceedings of the 8th ACM international conference on bioinformatics, computational biology, and health informatics*, 285–294.

Yang, M., Tao, B., Chen, C., Jia, W., Sun, S., Zhang, T., & Wang, X. (2019). Machine learning models based on molecular fingerprints and an extreme gradient boosting method lead to the discovery of jak2 inhibitors. *Journal of Chemical Information and Modeling*, *59*(12), 5002–5012.

Yangxin, Y., Jin, Z., & Bayly, A. E. (2008). Development of surfactants and builders in detergent formulations. *Chinese Journal of Chemical Engineering*, *16*(4), 517–527.

Ying, C., Cai, T., Luo, S., Zheng, S., Ke, G., He, D., Shen, Y., & Liu, T.-Y. (2021). Do transformers really perform badly for graph representation? *Advances in Neural Information Processing Systems*, *34*, 28877–28888.

Yoshikawa, N., & et al. (2018). Population-based de novo Molecule Generation using Grammatical Evolution. *Chemistry Letters*, *47*(11), 1431–1434.

Yuan, Q., & et al. (2020). Molecular Generation Targeting Desired Electronic Properties via Deep Generative Models. *Nanoscale*, *12*(12), 6744–6758.

Zang, Q., Mansouri, K., Williams, A. J., Judson, R. S., Allen, D. G., Casey, W. M., & Kleinstreuer, N. C. (2017). In silico prediction of physicochemical properties of environmental chemicals using molecular fingerprints and machine learning. *Journal of chemical information and modeling*, *57*(1), 36–49.

Zhang, D., Yin, J., Zhu, X., & Zhang, C. (2018). Network representation learning: A survey. *IEEE transactions on Big Data*, *6*(1), 3–28.

Zhang, R., Waibel, C., & Wortmann, T. (2020). Aerodynamic shape optimization for high-rise conceptual design. *Proceedings of the eCAADe*.

Zhou, A., Qu, B.-Y., Li, H., Zhao, S.-Z., Suganthan, P. N., & Zhang, Q. (2011). Multiobjective evolutionary algorithms: A survey of the state of the art. *Swarm and evolutionary computation*, *1*(1), 32–49.

Zhou, G., Gao, Z., Ding, Q., Zheng, H., Xu, H., Wei, Z., Zhang, L., & Ke, G. (2022). Uni-mol: A universal 3d molecular representation learning framework.

Zhou, G., Gao, Z., Ding, Q., Zheng, H., Xu, H., Wei, Z., Zhang, L., & Ke, G. (2023). Uni-mol: A universal 3d molecular representation learning framework.

Zitzler, E., Laumanns, M., & Thiele, L. (2001). Spea2: Improving the strength pareto evolutionary algorithm. *TIK report*, *103*.

Zitzler, E., & Thiele, L. (1998). Multiobjective optimization using evolutionary algorithms—a comparative case study. *International conference on parallel problem solving from nature*, 292–301.