Multi-Objective Evolution for Chemical Product Design

Bilal Aslan
aslbil001@myuct.ac.za
Department of Computer Science
University of Cape Town
Cape Town, South Africa

Flavio Correa da Silva
fcs@usp.br
Department of Computer Science
University of Sao Paulo
Sao Paulo, Brazil

Geoff Nitschke
gnitschke@cs.uct.za
Department of Computer Science
University of Cape Town
Cape Town, South Africa

ABSTRACT

The design of chemical products requires the optimization of desired properties in molecular structures. Traditional techniques are based on laboratory experimentation and are hindered by the intractable number of alternatives and limited capabilities to identify feasible molecules and either test or infer their properties for optimization. Computational techniques based on deep learning and multi-objective evolutionary optimization have spurred chemical product design, but the definition of appropriate metrics to compare techniques is challenging. We suggest the adoption of two complementary assessments to account for quantitative as well as qualitative features of different techniques, and then test our proposed assessments by comparing two heuristics to build new generations of molecular candidates, termed respectively, *direct correlation* and *extended search*.

CCS CONCEPTS

• **Computing methodologies** → **Co-evolution**.

KEYWORDS

Evolutionary Multi-objective Optimization

ACM Reference Format:
Bilal Aslan, Flavio Correa da Silva, and Geoff Nitschke. 2024. Multi-Objective Evolution for Chemical Product Design. In *Genetic and Evolutionary Computation Conference Companion (GECCO '24 Companion), July 14–18, 2024, Melbourne, Australia.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3583133.3590528

1 INTRODUCTION

Chemical product design requires the optimization of properties in molecules. Traditional optimization techniques are based on laboratory experimentation, which can be expensive and time consuming. In recent years, computational techniques have been successfully employed for generation and selection of molecules given desired properties. These techniques are, in most cases, comparatively faster and more cost effective than their traditional counterparts [14, 18].

The chemical design space is large, with estimated existence of over 10^{200} organic molecules [6]. As a consequence, the development of computational techniques to generate and optimize molecular properties is challenging [17]. Promising results have been obtained with the use of computational chemistry for symbolic representation of molecules and their properties, deep learning for molecular generation and property estimation [12, 23], and evolutionary algorithms for property optimization [4, 7, 8, 10, 11, 16, 19, 20, 22]. Computational techniques employ search procedures to identify solution sets, and search is based on some organization of the chemical design space assuming that: (1) Molecules which are structurally similar present similar property values, and (2) Changes in property values can be controlled by incremental changes in molecular structures. These assumptions are only approximately observed empirically, thus imposing limitations in the accuracy of computational techniques. Moreover, The multifaceted nature of molecular property optimization demands effective trade-off between objectives, and techniques such as the *Multi-objective Covariance Matrix Adaptation Evolution Strategy* (MO-CMA-ES) [13] have been specifically designed for multi-objective optimization (MOO).

Effective heuristics to explore the molecular search space delimit search to a neighbourhood around effective molecules. Suitable selection of effective molecules (*seed molecules*) can reduce the search space to a manageable size, preserve properties of interest across solution candidates, and retain diversity in the constrained search space so that new molecules can still be found. We introduce two heuristics for exploration of the molecular search space starting from seed molecules, coined resp. *direct correlation* and *extended search*. Direct correlation selects a neighbourhood with sufficiently high similarity with respect to seed molecules. The obtained search space is then explored to identify optimal molecules. Extended search initially selects a belt of molecules featuring a specified similarity level around the seed molecules, and then uses this belt to expand the set of seed molecules for direct correlation. Direct correlation is less exploratory than extended search, since search is strongly influenced by the choice of seed molecules, at the cost of reducing solution diversity and innovation in discovered molecules.

The definition of appropriate metrics to compare optimization techniques and heuristics is challenging, since the quantity as well as the quality of candidate design solutions are important: it is clearly important that solution sets comprise optimized properties and, given the accuracy limitations of computational techniques, it is also important that a variety of alternatives are obtained for experimental fine tuning of product design. We suggest the adoption of two complementary assessments to account for quantitative as well as qualitative features of different techniques, and then test

GECCO '24 Companion, July 15–19, 2024, Melbourne, Australia

ACM ISBN 979-8-4007-0120-7/23/07.
https://doi.org/10.1145/3583133.3590528

our proposed assessments by comparing the two proposed heuristics to build new generations of molecular candidates. Quantitative assessment is grounded on the cardinality of solution sets, whereas qualitative assessment is based on mean and variance of property values observed in solution sets. Our results contribute molecular design synthesis guidelines that can be integrated into the methodology of future computational tools for chemical product design.

2 METHODS AND EXPERIMENTS

Our molecular optimization method has been tested for automated domestic detergent synthesis [2, 3], with the optimization attributes:

- *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.

The target similarity threshold of 90% for reference likeness is set to ensure that the selected molecules exhibit a managed deviation from the seed molecules, thereby capitalizing on the advantageous properties of seed molecules and fostering innovation at the same time. The process begins by setting seed molecules and a surrounding molecular space characterized by a minimal 80% similarity with respect to seed molecules. Effective seed molecules are then determined using either direct correlation or extended search. Selection is then conducted iteratively using MOO. Fish toxicity is inferred using a trained model based on *Uni-Mol*, which is a universal 3D molecular representation learning framework [24] grounded on pre-trained 3D structures of 210 million molecules crafted with RDKit and represented using SMILES [21]. To assess fish toxicity, a data set extracted from the publicly available *PubChem* database [9] has been used to tune *Uni-Mol*. Given seed molecules M_0, three hyper-parameters are used to control molecular selection:

(1) similarity threshold T,
(2) parent selection β, and
(3) offspring selection λ.

Higher T preserves similarity (and quality) of candidate solutions with respect to seed molecules, however can lead to additional cycles prior to stabilization; larger β increases breadth in exploration; and larger λ decreases randomness in molecular selection – if $\lambda \geq |\text{solution set}|$, randomness is eliminated. For extended search, an exact similarity value T_0 is also employed to generate *belts* with similarity exactly T_0 with respect to at least one molecule $m \in M_0$, which are used to expand the set of seed molecules. Once candidate molecules are selected, toxic molecules can be eliminated, and a solution set containing only optimal molecules is selected.

We impose constraints to prevent exploration frontiers from veering towards molecules with less desirable properties. These constraints are organized as *in-experiment*, that is, constraints applied to property values between each iteration of optimization procedures, and *post-experiment*, that is, constraints applied to final solution sets only. Post-experiment constraints are more stringent than in-experiment constraints, and are used only at the end of the optimization procedures to avoid premature convergence and, therefore, preserve robust exploratory capabilities. Table 1 presents constraints tuned for our experiments. The final data set (combining all the frontiers found at each step) is then processed to remove any molecule dominated by another in terms of any property.

Our adopted variation of MO-CMA-ES is non-parametric, in that it does not assume any specific prior distribution over the search space. 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 removal of toxic molecules identified using *Uni-Mol*. From these, Pareto optimal solutions are built, 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 offspring are selected using T similarity with the respective parent. Offspring are combined to form M_1 as candidates for the new solution space S_1.

This procedure is repeated to build $S_2, S_3, \ldots$, until a run-time limit or stability criteria is reached in S_N for some 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:

(1) A growth factor $G > 1$ for β and λ
(2) If $\frac{|M_{k+1}|}{|M_k|} < 1$, then β and λ are updated by a factor $\times G$,
(3) And if $\frac{|M_{k+1}|}{|M_k|} > 1$, then they are updated by a factor $times \frac{1}{G}$.

The final solution set, combining $\mathcal{S}_1, \ldots, \mathcal{S}_N$ then removes dominated molecules. An initial search space of approximately 700, 000 molecules encoded as SMILES strings was sourced from *PubChem*, and then refined to 30, 000 molecules considering neighbourhoods around seed molecules. This refinement process utilized the MACCS Tanimoto similarity measure, with a threshold set at 70%. Experiments gauged MO-CMA-ES effectiveness using either direct correlation or extended search, with respect to solution set diversity and quality. To ensure reproducibility, all code and experimental results have been made publicly available online [1]. The diversity of solutions is assessed considering the cardinality of obtained solution sets, assuming that larger cardinality correlates with higher diversity of alternatives for experimental exploration. The quality of solutions is assessed using box plots of each property independently, considering that optimized mean values (for example, high reference likeness and XLogP are desired properties) and decreased variance for each property correlates with higher quality of solutions as a whole.

3 RESULTS AND DISCUSSION

Experiments were run for 50 generations for each method and for 20 runs per method. Convergence criteria were deliberately excluded to assess comparative method capabilities in escaping local optima and to examine exploratory behaviour. Figure 1 illustrates the average diversity of solutions over the course of an evolutionary run, as measured by the quantity of unique molecules discovered over the evolutionary search process of MO-CMA-ES using direct correlation *versus* extended search for determination of seed molecules.

Table 1: Solution Set Property Constraints

Property	In-experiment	Post-experiment
Molecular Complexity	≤ 500	$250 \leq \cdot \leq 350$
Molecular Weight	≤ 500	$250 \leq \cdot \leq 350$
XLogP	≥ 4	$5 \leq \cdot \leq 10$

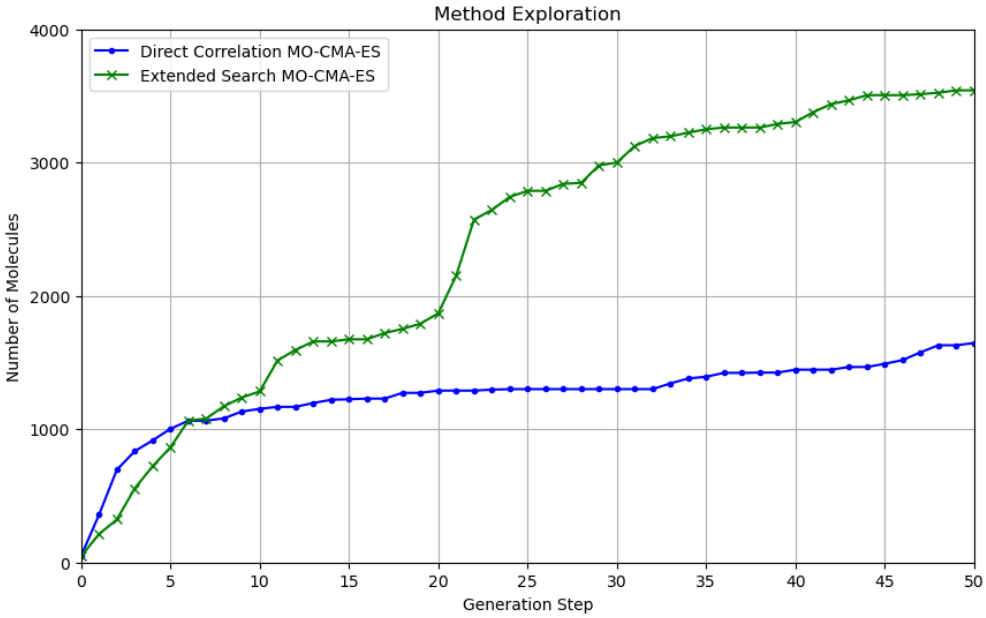

Figure 1: Discovered molecules by MO-CMA-ES using either direct correlation or extended search.

According to this measure, extended search augments solution set diversity in comparison with direct correlation.

Figure 2 displays box plots that compare the average quality of the final solution sets with respect to various optimized variables: *molecular complexity*, *molecular weight*, *XLogP*, and *reference likeness*. In these diagrams, data is normalized within a range of 0 to 1, based on the maximum and minimum values obtained after application of constraints as defined in table 1, and quality is measured according to fluctuations of mean values of properties towards desired goals (which can be either maximization or minimization of values) and to reduction of variance across solution sets (corresponding to reduction in height of the green rectangles). In this particular experiment, no significant fluctuation in quality was observed with changes between direct correlation and extended search, suggesting that, in this case, extended search can improve diversity of solution sets while preserving the quality of solutions with respect to optimization of properties. In other scenarios, it can happen that the quality of solutions is also affected by the choice of heuristics to guide evolutionary search. In these cases, quality and diversity must be balanced, according to priorities determined for each particular (molecular property optimization) problem.

4 CONCLUSION

The study conducted comparisons between two heuristics to guide search in MOO for chemical product design (direct correlation and extended search). Empirical findings indicate that extended search can improve the diversity of solution sets without altering the quality of obtained solutions. Our key contribution is the design of appropriate metrics to compare different optimization strategies for innovative product design. Specifically, we propose the combination of two perspectives, focusing respectively on diversity and quality of solution sets, measured objectively by the cardinality of solution sets and by statistical measures for each property of interest as observed in each obtained solution set. Experimental results also indicated that the chemical design space is highly unevenly distributed in terms of similarity across molecules and corresponding observable property values, leading to entrapment by local optima, particularly in MOO scenarios. Ongoing work is enhancing dynamic parameter optimization and fine-tuning parameter interactions (such as β and λ), and enriching datasets with virtually generated molecules (for example, using Generative Adversarial Networks) and evolutionary transfer learning [5, 15] to avoid local optima and produce diverse solutions.

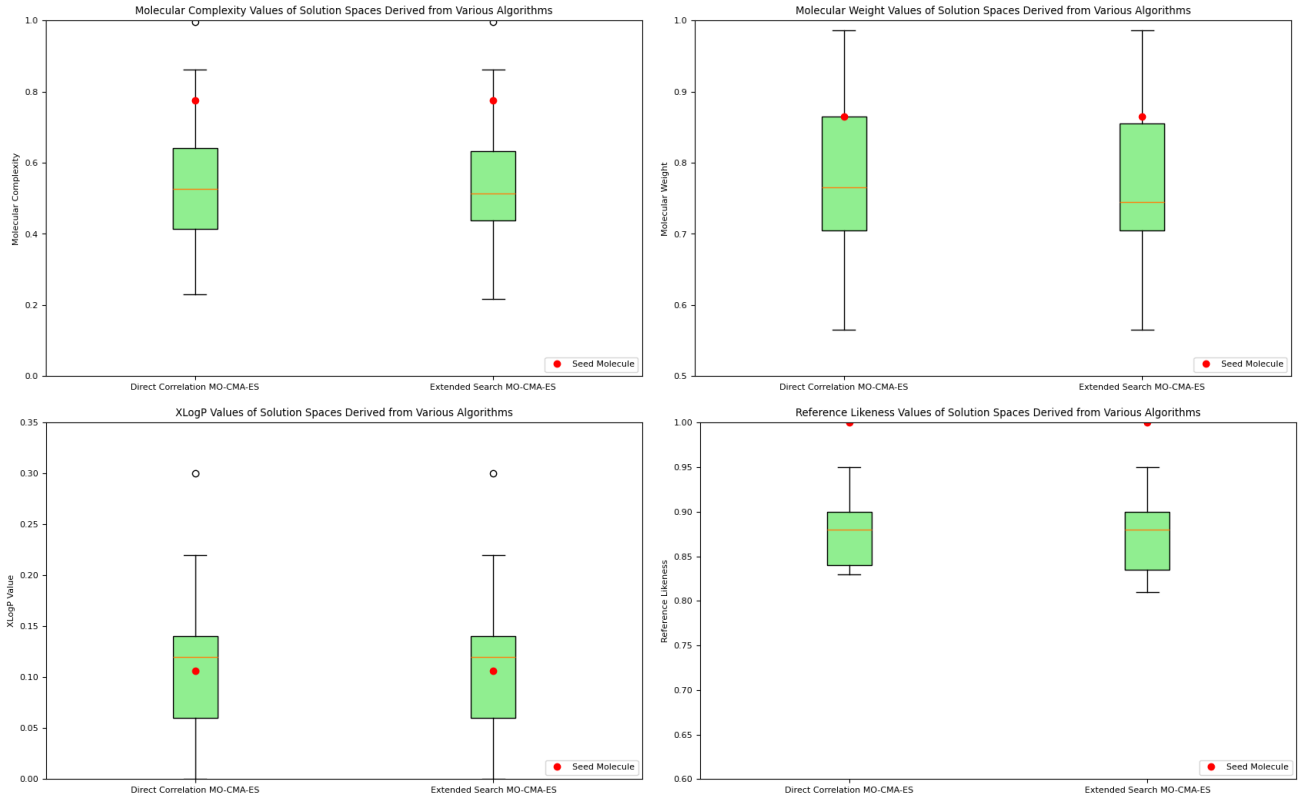

Figure 2: Box plot comparisons of compound properties. Upper Left: Molecular Complexity; Upper Right: Molecular Weight; Lower Left: XLogP; Lower Right: Reference Likeness.

REFERENCES

[1] Anon. 2024. Anonymous Repository. *https://anonymous.4open.science/r/Molecule-Selection-with-MOO-6CCC* (2024).

[2] B. Aslan, F. da Silva, and G. Nitschke. 2023. A Computational Method to Support Chemical Product Design Based on Multi-objective Optimisation and Graph Transformers. In *Proceedings of the Conference on Artificial Life*. MIT Press, Sapporo, Japan.

[3] B. Aslan, F. da Silva, and G. Nitschke. 2023. Multi-objective Evolution for Automated Chemistry. In *Proceedings of the IEEE Symposium Series on Computational Intelligence*. IEEE Press, Mexico City, Mexico, 152–157.

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

[5] S. Didi and G. Nitschke. 2016. Hybridizing Novelty Search for Transfer Learning. In *Proceedings of the IEEE Symposium Series on Computational Intelligence*. IEEE Press, Athens, Greece, 2620–2628.

[6] T. Fink, H. Bruggesser, and J. Reymond. 2005. Virtual Exploration of the Small-molecule Chemical Universe Below 160 Daltons. *Angewandte Chemie* 44 (2005), 1504–1508.

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

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

[9] S. Kim et al. 2023. PubChem 2023 Update. *Nucleic Acids Research* 51, 1 (2023), 1373–1380.

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

[11] J. Leguy et al. 2020. EVOMOL: A Flexible and Interpretable Evolutionary Algorithm for Unbiased de novo Molecular Generation. *Journal of Cheminformatics* 12, 55 (2020), 1–19.

[12] J. Lim et al. 2020. Scaffold-based Molecular Design with a Graph Generative Model. *Chemical Science* 11, 4 (2020), 1153–1164.

[13] I. Loshchilov and F. Hutter. 2016. CMA-ES for Hyperparameter Optimization of Deep Neural Networks. *arXiv preprint arXiv:1604.07269* (2016).

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

[15] G. Nitschke and S. Didi. 2017. Evolutionary Policy Transfer and Search Methods for Boosting Behavior Quality: Robocup Keep-Away Case Study. *Frontiers in Robotics and AI* 4, 1 (2017).

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

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

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

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

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

[21] D. Weininger. 1988. SMILES, A Chemical Language and Information System. *Journal of Chemical Information and Computer Sciences* 28, 1 (1988), 31–36.

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

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

[24] G. Zhou et al. 2023. Uni-Mol: A Universal 3D molecular Representation Learning Framework. *ChemRxiv* 10.26434/chemrxiv-2022-jjm0j (2023).