



Delft University of Technology

A Better Multi-Objective GP-GOMEA - But do we Need it?

Harrison, Joe; Alderliesten, Tanja; Bosman, Peter A.N.

DOI

[10.1145/3712255.3734302](https://doi.org/10.1145/3712255.3734302)

Publication date

2025

Document Version

Final published version

Published in

GECCO 2025 Companion - Proceedings of the 2025 Genetic and Evolutionary Computation Conference Companion

Citation (APA)

Harrison, J., Alderliesten, T., & Bosman, P. A. N. (2025). A Better Multi-Objective GP-GOMEA - But do we Need it? In G. Ochoa (Ed.), *GECCO 2025 Companion - Proceedings of the 2025 Genetic and Evolutionary Computation Conference Companion* (pp. 1992-2000). (GECCO 2025 Companion - Proceedings of the 2025 Genetic and Evolutionary Computation Conference Companion). Association for Computing Machinery, Inc. <https://doi.org/10.1145/3712255.3734302>

Important note

To cite this publication, please use the final published version (if applicable).
Please check the document version above.

Copyright

Other than for strictly personal use, it is not permitted to download, forward or distribute the text or part of it, without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license such as Creative Commons.

Takedown policy

Please contact us and provide details if you believe this document breaches copyrights.
We will remove access to the work immediately and investigate your claim.



A Better Multi-Objective GP-GOMEA - But do we Need it?

Joe Harrison 
Joe.Harrison@cw.nl
Centrum Wiskunde & Informatica
Amsterdam, The Netherlands

Tanja Alderliesten 
Leiden University Medical Center
Leiden, The Netherlands

Peter A.N. Bosman 
Centrum Wiskunde & Informatica
Amsterdam, The Netherlands
Delft University of Technology
Delft, The Netherlands

ABSTRACT

In Symbolic Regression (SR), achieving a proper balance between accuracy and interpretability remains a key challenge. The Genetic Programming variant of the Gene-pool Optimal Mixing Evolutionary Algorithm (GP-GOMEA) is of particular interest as it achieves state-of-the-art performance using a template that limits the size of expressions. A recently introduced expansion, modular GP-GOMEA, is capable of decomposing expressions using multiple subexpressions, further increasing chances of interpretability. However, modular GP-GOMEA may create larger expressions, increasing the need to balance size and accuracy. A multi-objective variant of GP-GOMEA exists, which can be used, for instance, to optimize for size and accuracy simultaneously, discovering their trade-off. However, even with enhancements that we propose in this paper to improve the performance of multi-objective modular GP-GOMEA, when optimizing for size and accuracy, the single-objective version in which a multi-objective archive is used only for logging, still consistently finds a better average hypervolume. We consequently analyze when a single-objective approach should be preferred. Additionally, we explore an objective that stimulates re-use in multi-objective modular GP-GOMEA.

CCS CONCEPTS

• Theory of computation → Genetic programming.

KEYWORDS

GOMEA, Symbolic Regression, Genetic Programming, Multi-Objective optimization, Explainable AI, Automatically Defined Functions

ACM Reference Format:

Joe Harrison , Tanja Alderliesten , and Peter A.N. Bosman . 2025. A Better Multi-Objective GP-GOMEA - But do we Need it?. In *Genetic and Evolutionary Computation Conference (GECCO '25 Companion)*, July 14–18, 2025, Malaga, Spain. ACM, New York, NY, USA, 10 pages. <https://doi.org/10.1145/3712255.3734302>

1 INTRODUCTION

Symbolic Regression (SR) is an important eXplainable Artificial Intelligence (XAI) technique, where the goal is to uncover the underlying relationships between input variables and output targets in a given dataset through the construction of symbolic expressions. Symbolic expressions have the potential to be interpretable. As machine learning becomes more integrated into decision-making

processes in critical domains such as healthcare, finance, and criminal justice, the need for interpretable models grows, making SR algorithms an essential tool in the XAI landscape [4, 12, 21].

In XAI, the accuracy of machine learning models is not the only important objective; interpretability is equally important. The interpretability of models is influenced by various factors, including the size of the expression, the number of consecutive compositions, the types of operators employed, and the capabilities of the person interpreting the expression [8, 24]. Unlike traditional regression methods that primarily focus on parameter estimation, SR simultaneously searches for both the structure and parameters of symbolic expressions [9]. Consequently, SR naturally lends itself to Multi-Objective (MO) optimization, where solutions with trade-offs between competing objectives, such as accuracy and expression complexity, can be explicitly searched for.

Recently, it was found that using classic selection based on non-dominated sorting and subsequent variation to generate a new offspring population, as e.g., in NSGA-II, is prone to evolvability degeneration [13]. The population then tends to quickly get flooded with many small solutions, since this objective is easy to optimize, making it more difficult to subsequently find larger expressions.

GP-GOMEA, a state-of-the-art population-based approach to GP [19, 22], differs from classic selectorecombinative EAs in that it employs a technique called optimal mixing, which is more akin to a local search approach. Moreover, the first MO version of GP-GOMEA [20] follows the MO-GOMEA structure in which clustering is used to spread the search bias across the approximated Pareto front. Among the clusters, extreme clusters are identified in which improvements for only one objective are accepted. Optimal mixing combined with clustering essentially mitigates evolvability degeneration. However, different versions of GOMEA exist, for different types of optimization problems, in which clustering is approached differently, e.g., [2, 15]. Particularly, they include a different way to balance the clusters, which could be important when objectives have varying difficulty. We therefore want to study if alternative clustering approaches can have added value in GP-GOMEA.

It is well known that GP tends to bloat, as larger expressions often achieve higher accuracy [17]. Consequently, a single-objective (SO) optimization process, which solely focuses on accuracy, may naturally favor increasingly larger expressions over time, especially when starting with relatively small ones. This could potentially result in a non-dominated front with a higher average hypervolume compared to MO GP-GOMEA, as the latter balances both accuracy and expression size. In this paper, we investigate whether and to what extent this effect occurs in GP-GOMEA.

Recently, a modular version of GP-GOMEA was introduced in [6] that allows for the efficient evolution of larger expressions with GP-GOMEA by virtue of re-using subexpressions as functions or input



This work is licensed under a Creative Commons Attribution 4.0 International License. *GECCO '25 Companion*, July 14–18, 2025, Malaga, Spain
© 2025 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-1464-1/2025/07
<https://doi.org/10.1145/3712255.3734302>

features. Re-use of subexpressions can at the same time potentially lower the complexity of expressions because they can be interpreted in a faceted manner, understanding the subexpressions separately and within the whole, without losing much in accuracy. However, a challenge remains in that limited function re-use was observed in the original version of Modular GP-GOMEA, which is single-objective [6]. We hypothesize that the implementation of a special type of parsimony pressure that discounts subexpression re-use in multi-objective optimization may positively influence this aspect, potentially leading to more re-use of expressions.

In this paper, we propose improvements to the MO clustering method for MO Modular GP-GOMEA. We demonstrate that the accuracy-complexity trade-off presents a scenario where SO optimization should be employed. We give insight into when SO or MO Modular GP-GOMEA should be used in light of other objectives in their trade-off with accuracy. Furthermore, we investigate the effects of parsimony pressure on the frequency of functional re-use within the generated expressions.

2 METHODS

This section begins with an explanation of the general functioning of GP-GOMEA, followed by an explanation of Modular GP-GOMEA, and concludes with a detailed description of how these components are integrated into MO Modular GP-GOMEA.

2.1 GP-GOMEA

Symbolic expressions in GP are often modelled as binary-ary trees. Unlike traditional GP [10], which allows trees of varying sizes, GP-GOMEA uses a fixed tree template. This allows individuals in GP-GOMEA to be represented as strings, where each string index bijectively maps to a position in the tree. The mapping is determined using the pre-order traversal of the tree template (see Figure 1).

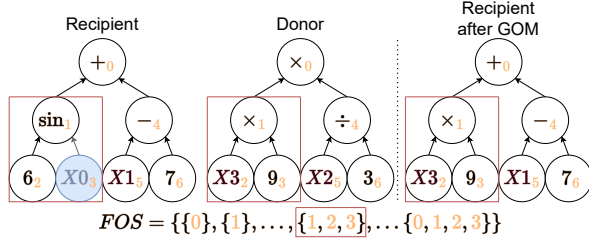


Figure 1: Example of the GOM procedure. Orange subscript numbers denote string indices from the tree's pre-order traversal. Numbers in subsets correspond to tree indices. Tree sections of the recipient and donor matching the subset in the FOS are outlined in red, with introns shaded in blue.

In GP-GOMEA, linkage information is attempted to be leveraged via the identification of dependencies between tree positions. As a surrogate measure of dependence, Mutual Information (MI) between all pairs of positions as measured in the population, is used. The Unweighted Pair Group Method with Arithmetic Mean (UPGMA) [5] is then applied to the MI matrix to hierarchically cluster positions, constructing a Family of Subsets (FOS) in the shape of a tree commonly called the linkage tree. For further details, see [22].

In GP-GOMEA, variation is achieved through Gene-pool Optimal Mixing (GOM). GOM is applied to every individual in each

generation. In GOM, a clone of the individual is created, and for each subset in the FOS, the corresponding genes in the clone are replaced with those of a donor randomly chosen from the population (see Figure 1). If the change swaps meaningful, non-intron, genes, and does not lead to worse fitness, it is accepted. After processing all FOS subsets, the resulting individual is added to the offspring set. Once all individuals are processed, the offspring set replaces the population. As in [23], coefficients are also mutated after GOM, with any mutation being reverted if it worsens fitness.

2.2 Modular GP-GOMEA

In this paper, we use Modular GP-GOMEA [6], a recently introduced multi-tree variant of GP-GOMEA where trees can call other trees as (functional) subexpressions (see Fig. 2). This allows for improving accuracy without affecting efficiency compared to GP-GOMEA with a larger single template. To avoid cyclic calling of subexpressions, only trees i that precede a tree j in the multi-tree can be called. The left and right inputs of subexpression nodes are the input arguments arg_0 and arg_1 , respectively, in the subexpression tree.

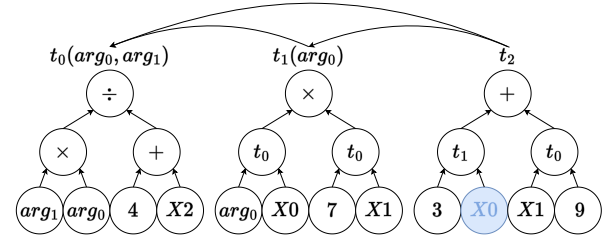


Figure 2: Example of an individual in Modular GP-GOMEA and how trees in its multi-tree representation can call each other. The last tree t_2 is the top-level expression and can call preceding trees t_1 and t_0 , but not vice versa. Nodes X_1 and 9 in t_2 are the input arguments arg_0 and arg_1 in t_0 respectively.

For each tree in the multi-tree, a separate FOS is constructed (e.g., for the example in Figure 2, there is a separate FOS for each of t_0 , t_1 , and t_2). For each tree, except the last (top-level expression) tree, the FOS for that tree contains a subset with all indices in the tree so as to allow full subtrees to be swapped. Before applying GOM, all subtree FOSes are combined into one large FOS and given a subscript to keep track to which tree in the multi-tree each subset pertains, e.g., $FOS = \{\{0, 1, 2\}_0, \{4, 5\}_1, \dots\}$. Upon applying GOM to an individual, the combined FOS is shuffled and a GOM step is performed for each FOS element in turn.

2.3 Multi-Objective GP-GOMEA

A key driver to promote spreading the search bias of an EA along the Pareto approximation front, is to use clustering. The first introduced version of MO-GP-GOMEA [20] follows the first multi-objective version of optimization-based GOMEA [15]. The clustering approach used there, was based on the Balanced k-Leader-Means algorithm (BKLM) introduced in [1]. Two types of clusters are identified: (1) extreme clusters for which Single-Objective (SO) GOM is performed based on their corresponding objective, and (2) middle clusters for which MO GOM is performed. In MO GOM, changes are accepted in case a Pareto improvement is found, when all objectives stay the same, or when a change leads to a solution that is accepted into the

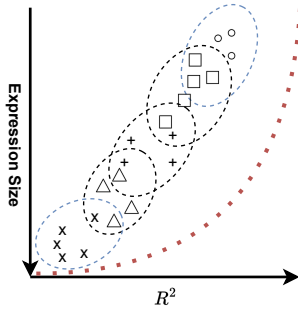


Figure 3: Example of clusters (similar symbols) and corresponding donor clusters (dashed ovals). $\times$ and $\circ$ are extreme clusters for the objectives expression size and R^2 respectively. The blue dashed ovals are the corresponding donor clusters. The red dotted line represents the Pareto front.

elitist archive [14]. In this paper, we use $k = 5$ clusters. Because we furthermore consider two objectives here, we have two extreme and three middle clusters, see Figure 3 for an example.

To create the clusters, first, the solution with the best objective value for a randomly chosen objective is chosen. Then, the leaders of the remaining $k - 1$ clusters are chosen from the remaining population by iteratively picking the individual with the largest Euclidean distance in normalized objective space to the other leaders [15]. Then, while there remain unpicked individuals, the clusters are iterated over in random order and the individual that is closest to the cluster leader is assigned to that cluster. Once every individual is assigned to a cluster, k -means clustering is performed. Afterward, the cluster centers are used to first construct donor clusters by choosing the $\frac{2n}{k}$ individuals closest to each cluster center, where n is the population size. These clusters all have equal size to ensure control over the numbers of individuals from which linkage is learned. Indeed, for each donor cluster, a separate FOS is learned for each tree in the multi-tree representation (i.e. $5 \times 4 = 20$ FOSes in the case of this paper). In GOMEA, each individual must undergo GOM, but the clustering procedure thus far may have left individuals unassigned or assigned to multiple clusters. Individuals without an assigned cluster are assigned to a random cluster. For individuals with multiple assigned clusters, ties are broken randomly. The clusters for which their center represents the best objective value for one of the objectives, becomes the extreme cluster for that respective objective.

We take inspiration from [14] and use a method similar to Adaptive Grid Discretisation to keep the number of individuals in the elitist archive at a manageable size. The objective space is divided into grid cells. Individuals are only admitted to the elitist archive if they are non-dominated and occupy an empty grid cell or if they dominate the individual currently residing in the same grid cell. We evenly space the grid with 100 steps between the minimum and maximum value of the elitist archive of the previous generation. We opted for an adaptive method because it is unknown beforehand how large an individual will grow and size is one of our objectives.

Algorithm 2 Combined description of various clustering methods in MO (Modular) GP-GOMEA. The original clustering method is defined by all steps in black and olive. The BKRR clustering method introduced in this paper is defined by all steps in black and purple. The BKMRR clustering method introduced in this paper is defined by all steps in black, blue, and purple.

- 1: Normalise objective values for all individuals.
- 2: **for each objective in random order do**
- 3: Sort population in objective.
- 4: Assign top $\frac{n}{k}$ individuals to an extreme (SO) cluster.
- 5: Remove these individuals from the pool of individuals.
- 6: **end for**
- 7: Set k to $k - 2$.
- 8: Choose a random objective o .
- 9: Initialize first leader as individual with best value for o .
- 10: **while** fewer than k leaders selected **do**
- 11: Compute distances from remaining candidates to leaders.
- 12: Select candidate with maximum distance as new leader.
- 13: **end while**
- 14: Assign individuals to nearest leader to form initial clusters.
- 15: Perform k -means clustering.
- 16: Undo all cluster assignments, but keep cluster centers.
- 17: **while** Individuals not assigned to a cluster **do**
- 18: Loop over k clusters in random order and assign to each cluster the unassigned individual that is closest to the cluster center, without updating the cluster center.
- 19: **end while**
- 20: For each cluster center, create donor cluster with closest $\frac{2n}{k}$ individuals.
- 21: Identify clusters as extreme or middle clusters.
- 22: Identify clusters as extreme or middle clusters.
- 23: Create final clusters from donor clusters by assigning unassigned individuals to random clusters and breaking ties for individuals with multiple cluster assignments randomly.

2.4 MO GP-GOMEA vs SO GP-GOMEA with an MO archive

We distinguish between MO and SO GP-GOMEA (respectively referred to as MO and SO in tables). While the sole objective in SO GP-GOMEA is R^2 (maximization), we also maintain an MO elitist archive. This MO elitist archive only tracks non-dominated solutions when considering size besides R^2 , i.e., the MO archive does not influence the evolutionary process in SO optimization in any way.

3 EXPERIMENTS AND RESULTS

We conduct experiments to address six research questions. All are studied in the context of the modular version of GP-GOMEA, even though we omit the term modular in most of the remainder of this paper. The experiments for each question may build on the findings of prior ones. Consequently, we present the results immediately following the explanation of the setup for each research question:

- (1) Is clustering necessary for MO GP-GOMEA?
- (2) How does MO GP-GOMEA compare to SO GP-GOMEA?
- (3) What improvements can be derived for MO GP-GOMEA from SO GP-GOMEA?

- (4) Does the removal of duplicate solutions affect performance?
- (5) Can functional re-use be promoted in MO GP-GOMEA?
- (6) How does MO GP-GOMEA perform with other objectives?

3.1 General Setup

We run experiments on 5 real-world datasets (see Table 1). Each experiment is run on a separate core of an AMD EPYC ROME 7282 and is terminated, either when a time budget of 3 hours is reached, or when the MO elitist archive (which is maintained in all experiments) does not change for 100 consecutive generations. The average hypervolume [25] over all repetitions is used as a measure of comparison. To allow for easy comparison between datasets we use the coefficient of determination $R^2 = 1 - \frac{\text{var}(x)}{\text{MSE}(x)}$ as a measure of accuracy. For the size of the expression, we expand all subexpression nodes and sum up its non-intron nodes. The hypervolume is computed by normalising the objectives to a range of 0 to 1, achieved by subtracting the minimum objective value and dividing by the maximum objective value after the subtraction step, observed across all experiments, with the exception of R^2 which is set to have a fixed minimum and maximum of 0 and 1 respectively. The reference point is set to 0 and 1 for the R^2 and size objective respectively. Statistical testing is done using the Wilcoxon signed rank test with Bonferroni-Holm correction (significant results indicated in bold).

Dataset	Samples	Features	Mean	Variance
Airfoil	1503	5	124.8	6.9
Bike Daily	731	11	4504.3	1935.9
Concrete	1030	8	35.8	16.7
Dow Chemical	1066	57	3.0	0.1
Tower	4999	25	342.1	87.8

Table 1: Real-world datasets used in our experiments. Two columns that are a linear combination of the target were removed from the Bike Daily dataset.

We initialize the population using the half-and-half method (50% grow, 50% full). In the grow method, terminal nodes (input features or coefficients) are sampled with a 50% probability. Moreover, the probability of sampling a coefficient is then also 50%. Unused string indices in the fixed-sized tree templates are filled with introns sampled randomly from the operator set. For unary operators the leftmost child node is used (see also Figure 1).

Linear scaling (LS) terms are recalculated at every generation for each expression to ensure optimal scaling during the evolutionary process, following the approach proposed by [7].

In Table 2 we provide an overview of the general settings that we used in all of the experiments in this paper.

3.2 Is clustering necessary for MO GP-GOMEA?

In [13] it was shown that with expression size as secondary objective, a classic selection and variation approach to MO GP such as NSGA-II [3] is prone to evolvability degeneration because generating smaller expressions is a much easier objective than improving the accuracy. However, given the more local search nature of GOMEA where small changes are made and immediately tested for improvement, it is not directly clear whether the additional

Parameter	Setting
Population size	4096
Tree height	4
# Multi-trees	4
Coefficient sampling	$\sim U(\text{min}_{\text{target}}, \text{max}_{\text{target}})$
Probability sampling coefficient	50%
Function set	$+, -, *, /$, sin, cos, log, $\sqrt{\cdot}$, subexpression, arg_0 , arg_1
# Repetitions	10
Stopping criterion	3 hours
# Clusters	5

Table 2: General experiment settings used in every experiment (unless indicated otherwise).

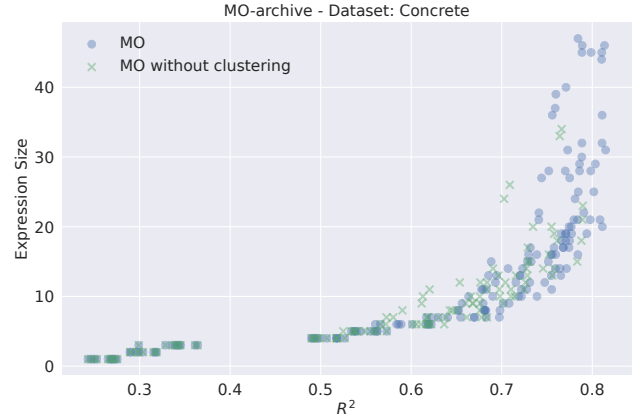


Figure 4: Individuals in the MO elitist archives in all 10 experiment repetitions.

layer of complexity that comes with clustering is required. We therefore first perform an experiment where we compare MO Modular GP-GOMEA using the clustering approach of the first MO GP-GOMEA [20] to not using clustering.

Dataset	MO with clustering	MO without clustering
Air	0.696 ± 0.046	0.668 ± 0.056
Bike	0.839 ± 0.017	0.809 ± 0.019
Concrete	0.794 ± 0.011	0.741 ± 0.032
Dowchemical	0.709 ± 0.036	0.636 ± 0.056
Tower	0.827 ± 0.021	0.746 ± 0.045

Table 3: Comparison of average hypervolume ($\pm$ standard deviation) obtained with MO Modular GP-GOMEA with and without clustering. Significant results are indicated in bold.

The results in Table 3 show that without clustering the average hypervolume is significantly smaller in 4/5 datasets. In Figure 4 we see that using clustering, especially larger and more accurate solutions can be found within the time budget. We therefore conclude that, while GOMEA does not suffer from evolvability degeneration as does NSGA-II, adding clustering to its procedure is beneficial.

3.3 How does MO GP-GOMEA compare to SO GP-GOMEA?

Clustering and performing linkage learning for each cluster separately adds complexity in MO GP-GOMEA. A natural question is to ask whether the added complexity leads to improvements. To this

end, we now consider the single-objective version of GP-GOMEA using as sole objective maximizing accuracy (R^2).

Dataset	MO with clustering	SO
Air	0.696 ± 0.046	0.820 ± 0.016
Bike	0.839 ± 0.017	0.893 ± 0.007
Concrete	0.794 ± 0.011	0.855 ± 0.005
Dowchemical	0.709 ± 0.036	0.853 ± 0.010
Tower	0.827 ± 0.021	0.899 ± 0.004

Table 4: Comparison of average hypervolume ($\pm$ standard deviation) between MO optimization and SO optimization. Significant results are indicated in bold typeface.

In Table 4 we see that SO GP-GOMEA significantly outperforms MO GP-GOMEA on all datasets within the limits of our experiments. A larger average hypervolume is achieved with SO GP-GOMEA due to it being able to divert all its search efforts towards finding expressions with high R^2 . Because the initial population has very small solutions, this method incidentally finds increasingly larger expressions that lie on the non-dominated front, even though it is not explicitly minimizing the size objective. Moreover, MO GP-GOMEA continually uses fruitless search efforts to minimize the expression size further, while the smallest expression size has already been found. The individuals with the highest R^2 often have large tree structures that could benefit the most from optimising their expression size, but these individuals end up in the extreme cluster optimising only accuracy. We therefore conclude that for the bi-objective optimization of accuracy and expression size, an SO approach is preferable, but we also find that the clustering approach used so far has issues, which may be improved upon.

3.4 What improvements can be derived for MO GP-GOMEA from SO GP-GOMEA?

In each of the previous experiments, the same population size (4096) was used for both the MO and SO experiments. MO GP-GOMEA however divides its population over five clusters. Hence, the extreme cluster with the accuracy objective is smaller than the total population size used in the SO approach. As population size is often a crucially important factor in what can be achieved with an EA, the population size for each EA should always be tuned individually for the fairest comparison. Therefore, we also perform an experiment in which the MO approach has a population 5 times larger (making each cluster equally large as the population in SO GP-GOMEA).

Cluster sizes can become skewed, as many individuals in the population end up in a cluster with other individuals that have an expression size of 1 (see Figure 5). This goes against the idea of spreading the search bias equally across the approximation front. To address this, we propose *Balanced K-leader-means Round Robin* (BKRR) clustering and *Balanced K-m-leader-means Round Robin* (BKmRR) where m is the number of objectives, which in this paper is always 2, but the approach is generic for any m . The approaches take inspiration both from the clustering approach in the MO real-valued GOMEA [2] and the original BKLM clustering method [1].

In BKRR, we initially proceed as in the clustering approach previously used in MO GP-GOMEA. After the k-means clustering step however, we do not create donor clusters with the closest $\frac{2n}{k}$

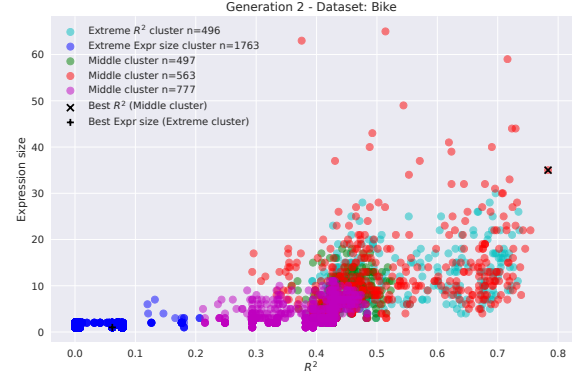


Figure 5: Two issues in MO GP-GOMEA with the original clustering approach: the clusters are unbalanced (see the large cluster size of the expression size cluster that occurs after only 2 generations) and the solution with the best R^2 value (marked $\times$) is assigned to a middle cluster instead of the appropriate extreme cluster.

individuals and then subsequently assign potentially unassigned individuals to random clusters and break ties randomly for individuals assigned to multiple clusters. Instead, we perform multiple assignment rounds in which we iterate over the remaining clusters, shuffling the cluster order each round. Each time a cluster is considered, we assign the closest individual in the remaining unassigned population to that cluster, in a round-robin fashion, ending up with almost exactly $\frac{n}{k}$ solutions in each cluster (deviations by 1 are possible based on population size and cluster ordering).

In BKmRR, we first iterate over the objectives in random order and select, for each objective, the top $\frac{n}{k}$ unassigned individuals with respect to that objective, and assign them to the extreme (SO) cluster corresponding to that objective (see Algorithm 1). The remaining unassigned individuals are then subjected to BKRR using the remaining $k - 2$ clusters (i.e., excluding the individuals in the extreme clusters). The round-robin style assignment combined with the initial splitting off of extreme clusters based on their objective values for singular objectives now leads to no longer requiring donor clusters and obtaining a clustering with equal size clusters where extreme clusters are truly representative of the best solutions in that objective. For an example, see Figure 6.

Dataset	MO with BKRR	MO with BKmRR
Air	0.789 ± 0.016	0.814 ± 0.016
Bike	0.880 ± 0.007	0.889 ± 0.006
Concrete	0.841 ± 0.004	0.847 ± 0.009
Dowchemical	0.809 ± 0.018	0.845 ± 0.007
Tower	0.882 ± 0.008	0.900 ± 0.005

Table 5: Comparison of the use of the two new clustering methods in MO GP-GOMEA in terms of average hypervolume ($\pm$ standard deviation). Significant results are in bold.

From the results in Tables 4 and 5 we observe that both versions of MO GP-GOMEA that use the new clustering methods perform better than when the original clustering method is used. Moreover, the use of BKmRR statistically significantly outperforms the use of BKRR in all of the datasets.

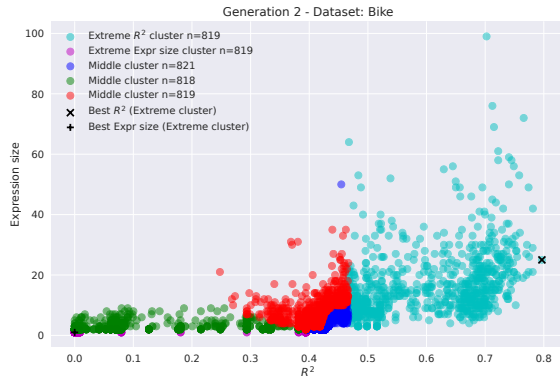


Figure 6: Example of a balanced clustering from our proposed BKmRR clustering method. Each cluster has approximately the same number of individuals. The best solution in each objective is assigned to an appropriate SO cluster. The R^2 SO cluster has a visible R^2 cut-off around 0.48.

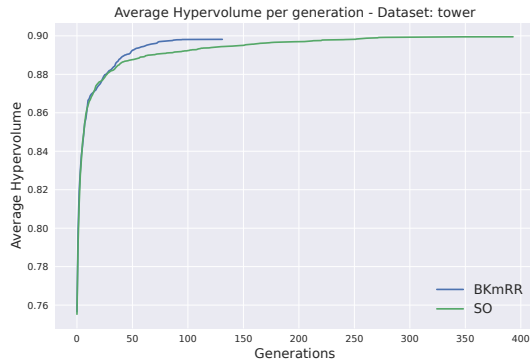


Figure 7: Comparison of average hypervolume versus generations for MO GP-GOMEA with BKmRR and SO GP-GOMEA.

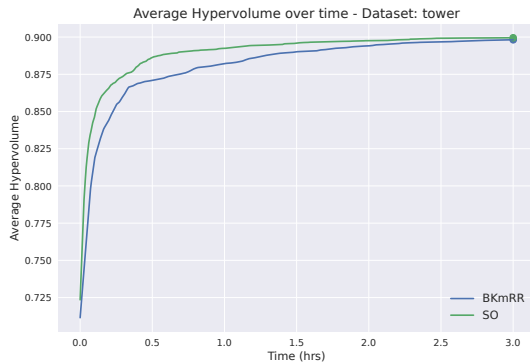


Figure 8: Comparison of average hypervolume versus time for MO GP-GOMEA with BKmRR and SO GP-GOMEA.

Effectively, the BKmRR method has one extreme SO cluster that performs exactly the same task as in SO GP-GOMEA. Yet, even if the population size in MO GP-GOMEA is set so that the cluster size is the same as the population size in SO GP-GOMEA, MO GP-GOMEA is outperformed by SO GP-GOMEA in all datasets (but

only significantly in the Bike and Concrete datasets). Key to understanding this difference in performance is comparing the average hypervolume measured per generation to the average hypervolume measured over time. In Figure 7 we observe that per generation, MO GP-GOMEA with BKmRR closely matches SO GP-GOMEA, but, when looking at a time-based comparison in Figure 8 we see that SO GP-GOMEA obtains a larger hypervolume faster. MO GP-GOMEA needs a larger population (#clusters times larger) to construct larger expressions and also bears the overhead of constructing multiple FOSes and clustering each generation.

3.5 Does the removal of duplicate solutions affect performance?

In Figure 6 we see that the extreme cluster with the expression size objective is balanced in the number of individuals (i.e., it has a similar number of individuals as other clusters). However, it mostly contains duplicate expressions with a single coefficient or input feature (i.e., the size is 1, the minimum value for the size objective). These individuals will repeatedly end up in the same extreme cluster and MO GP-GOMEA will pointlessly keep trying to minimize the expression size even further. This search effort can better be spent elsewhere. One way to achieve this, is to mutate all individuals with an expression size of one, but this method does not work for other objectives where a similar problem may occur. To combat this problem in a more general fashion, each generation, we mutate all expressions that have duplicate fitness for each objective until an active node is mutated similar to work done in [16]. Symbolic expressions in GP are often modelled as binary-unary trees.

Dataset	MO with BKmRR mutated	SO mutated
Air	0.819 $\pm$ 0.016	0.839 $\pm$ 0.008
Bike	0.892 $\pm$ 0.005	0.902 $\pm$ 0.006
Concrete	0.850 $\pm$ 0.008	0.876 $\pm$ 0.007
Dowchemical	0.847 $\pm$ 0.005	0.863 $\pm$ 0.008
Tower	0.897 $\pm$ 0.00	0.905 $\pm$ 0.006

Table 6: Comparison in terms of average hypervolume ($\pm$ standard deviation) of SO GP-GOMEA and MO GP-GOMEA with BKmRR clustering, both including mutation of individuals with duplicate fitness. Significant results are in bold.

Comparing the results in Tables 4 and 6, we observe that the average hypervolume of both MO GP-GOMEA with BKmRR and SO GP-GOMEA can be improved by mutating individuals with duplicate fitness. However, SO GP-GOMEA significantly outperforms MO GP-GOMEA in all datasets. Because there are clusters with different objectives, and individuals can switch between clusters per generation, the population in MO GP-GOMEA is less likely to converge. Conversely, SO GP-GOMEA, especially with a small population size, tends to quickly (prematurely) converge. It therefore benefits from our proposed mutations that avoid convergence.

3.6 Can functional re-use be promoted in MO GP-GOMEA?

While the modular version of GP-GOMEA that we use in this paper is capable of creating decomposed expressions, which may enhance

interpretability of the final expression, the use of accuracy and expression size alone as objectives do not actively promote re-use of subexpressions. To address this issue, we propose a new definition of the expression size objective in which we subtract all duplicate non-leaf nodes from the total expression size, encouraging re-use. We only subtract all duplicate non-leaf nodes because leaf nodes can occur multiple times without subexpression nodes and we do not want to encourage multiple uses of leaf nodes. We refer to this new objective as the de-duplicated size objective. We use this objective only during the runtime of the algorithm, while the hypervolume is calculated using the normal size objective for comparison.

Whether the expressions underlying the real-world datasets from Table 1 contain any subexpression re-use is unknown. We therefore perform experiments in this section with five synthetic datasets from [6] (see Appendix B in the supplementary) that are specially constructed to re-use subexpressions and have a known ground-truth. No LS terms are used or calculated in this specific set of experiments.

Dataset	MO with BKmRR mutated	MO with BKmRR de-duplicated mutated
Synthetic 1	0.962 $\pm$ 0.001	0.963 $\pm$ 0.001
Synthetic 2	0.982 $\pm$ 0.000	0.982 $\pm$ 0.000
Synthetic 3	0.973 $\pm$ 0.002	0.975 $\pm$ 0.001
Synthetic 4	0.978 $\pm$ 0.000	0.981 $\pm$ 0.001
Synthetic 5	0.953 $\pm$ 0.01	0.974 $\pm$ 0.007

Table 7: Comparison in terms of average hypervolume ($\pm$ standard deviation) of MO GP-GOMEA with BKmRR with a normal size objective and the proposed de-duplicated size objective. Significant results are in bold.

From Table 7 we observe that promoting re-use via the use of de-duplicated size as an objective, leads to a significantly larger average hypervolume in 4/5 synthetic datasets.

Dataset	Avg. Diff. #subexpr. used	Avg. Diff. #subexpr. re-used	Avg. Diff. #subexpr. re-used as function
Synthetic 1	0.00	0.40	0.10
Synthetic 2	-0.60	-0.40	-0.20
Synthetic 3	-1.40	-1.30	0.00
Synthetic 4	0.20	-0.10	0.3
Synthetic 5	0.00	0.00	0.00

Table 8: The average difference in the number of (re-)used subexpressions between the normal size objective and the de-duplicated size objective (i.e., the de-duplicated count minus the normal count). A positive number indicates more reuse when using the de-duplicated size objective.

We define subexpression use as the number of times a tree representing that subexpression is called within the expanded tree representation. Re-used subexpressions are those that are invoked more than once. Functional re-use specifically refers to re-used subexpressions that are called multiple times and contain at least one non-intron argument node. For a clearer illustration, we refer to the example figures in Appendix D in the supplementary. In Table 8,

the de-duplicated size objective increases hypervolume despite leading to fewer used and re-used subexpressions in 4/5 datasets. This suggests that without it, more incorrect subexpressions are selected, lowering the achieved hypervolume.

3.7 How does MO GP-GOMEA perform with other objectives?

While simultaneously optimizing for both expression size and accuracy is inherently a multi-objective problem, SO GP-GOMEA with an MO archive consistently outperforms MO GP-GOMEA in terms of average hypervolume. This discrepancy can be attributed, in part, to the initialization process. GP-GOMEA initializes using the half-and-half method. The grow part of this method can cause issues because there is a high probability of sampling a small tree. Typically, multi-objective optimization with an EA starts with a set of individuals of which the objective values are far away from the Pareto front, in all objectives. To achieve this for MO GP, this would require that for the expression size, the algorithm starts with a population of only full, large trees. However, as shown in Figure 6, within just two generations, the smallest possible expression size combined with optimal accuracy for that size, is already discovered, whereas most individuals remain far from achieving the maximum attainable accuracy for their respective size.

Mindful of the importance of the non-dominated front at initialization time, in this section we analyse various types of objectives in combination with the accuracy objective to get insights regarding whether MO GP-GOMEA or SO GP-GOMEA with an MO archive performs better if other objectives than expression size are used.

3.7.1 Maximum error. The average error (i.e., the standard accuracy objective that we use) and the maximum error are correlated. As the average error is minimized, the maximum error is often reduced as well. In Table 9, we observe that SO GP-GOMEA still outperforms MO GP-GOMEA with this objective instead of the size objective. While SO GP-GOMEA focuses solely on optimizing the standard accuracy objective, it also minimizes the maximum error, effectively optimizing both objectives simultaneously.

Dataset	MO with BKmRR mutated	SO mutated
Air	0.796 $\pm$ 0.010	0.817 $\pm$ 0.029
Bike	0.888 $\pm$ 0.009	0.900 $\pm$ 0.008
Concrete	0.844 $\pm$ 0.007	0.865 $\pm$ 0.015
Dowchemical	0.796 $\pm$ 0.022	0.809 $\pm$ 0.029
Tower	0.849 $\pm$ 0.012	0.849 $\pm$ 0.012

Table 9: Comparison in terms of average hypervolume ($\pm$ standard deviation) between MO GP-GOMEA with BKmRR clustering and mutation and SO GP-GOMEA with mutation using maximum error as a second objective. Significant results are in bold.

3.7.2 Different complexity measure. In [8], a novel complexity measure is introduced which gives each operator in the tree a complexity score and sums them up to obtain a complexity measure for the entire tree. In a similar fashion, we give each operator its own score (see Appendix A in the supplementary for scores). In Table 10 we

see that SO GP-GOMEA significantly outperforms MO GP-GOMEA in 5/5 datasets. The new complexity metric and the size objective are correlated in the sense that when the size objective is minimized the complexity metric is also minimized.

Dataset	MO with BKmRR mutated	SO mutated
Air	0.820 ± 0.015	0.843 ± 0.013
Bike	0.895 ± 0.007	0.905 ± 0.008
Concrete	0.858 ± 0.007	0.873 ± 0.009
Dowchemical	0.847 ± 0.011	0.868 ± 0.010
Tower	0.898 ± 0.004	0.909 ± 0.005

Table 10: Comparison between MO and SO with different complexity measure as second objective in terms of average hypervolume. Significant results are in bold typeface.

3.7.3 Regularization objective with LS. Another means of performing regularization other than minimizing expression size, is to regularize the LS scaling (b) and offset (a) terms. The motivation for doing so is that while LS can greatly increase accuracy, it also can make an expression less interpretable because to extremely large numbers for the offset and/or scaling may be obtained. We therefore consider minimizing $\log(1 + a^2 + (b - 1)^2)$ as an objective instead of expression size. In Table 11 we see again that SO GP-GOMEA always outperforms MO GP-GOMEA using this objective instead of the expression size objective. This time the correlation with the R^2 is not as strong, but SO GP-GOMEA within the allotted time budget is able to find a large variety of different LS terms and is able to outperform MO GP-GOMEA in terms of hypervolume by finding more accurate solutions.

Dataset	MO with BKmRR mutated	SO mutated
Air	0.823 ± 0.011	0.841 ± 0.010
Bike	0.901 ± 0.006	0.908 ± 0.005
Concrete	0.861 ± 0.005	0.882 ± 0.009
Dowchemical	0.856 ± 0.008	0.874 ± 0.011
Tower	0.892 ± 0.004	0.909 ± 0.004

Table 11: Comparison in terms of average hypervolume between MO GP-GOMEA and SO GP-GOMEA with the MSE of LS terms as second objective. Significant results are in bold.

3.7.4 Number of cosine operators. The previous subsections shows that it is hard to find objectives that do not correlate with accuracy or that result in the single-objective EA starting out close to the non-dominated front. As a final, perhaps somewhat contrived case, we consider maximizing the number of cosine operators. This objective is certainly not correlated with accuracy, nor is it likely to have good values for this objective upon initialization. In Table 12, we observe that for this trade-off, MO GP-GOMEA achieves a significantly higher hypervolume than SO GP-GOMEA with an MO archive. As shown in Appendix C in the supplementary, the number of cosine operators is low at initialization and remains low in SO GP-GOMEA, as it is not actively optimized throughout evolution.

Dataset	MO with BKmRR mutated	SO mutated
Air	0.616 ± 0.011	0.446 ± 0.159
Bike	0.843 ± 0.007	0.325 ± 0.107
Concrete	0.746 ± 0.018	0.468 ± 0.201
Dowchemical	0.705 ± 0.011	0.232 ± 0.060
Tower	0.770 ± 0.016	0.255 ± 0.044

Table 12: Comparison in terms of average hypervolume between MO GP-GOMEA and SO GP-GOMEA with the number of cosines as second objective. Significant results are in bold.

4 DISCUSSION AND CONCLUSIONS

In this paper, we introduce and improve MO Modular GP-GOMEA by replacing the clustering method that was previously introduced for the non-modular MO GP-GOMEA. Specifically, we employ a similar strategy as in RV-GOMEA and first determine extreme clusters that excel in individual objectives before clustering the rest of the population in objective space. We find that our new clustering methods improve the performance of MO Modular GP-GOMEA. Furthermore, we tackled the issue of clusters containing individuals with duplicate fitness by introducing a mutation strategy.

While these improvements enhanced MO Modular GP-GOMEA, SO Modular GP-GOMEA with an MO archive generally outperformed it in terms of average hypervolume across most datasets within a time budget of three hours when optimizing for expression size and expression accuracy. The main reason is that SO Modular GP-GOMEA effectively finds many individuals that lie on the non-dominated front even though it focuses exclusively on optimizing accuracy. Due to the half-and-half initialization strategy used in this paper, individuals of the smallest size are already found at initialization and diverting search effort towards further minimizing the size of individuals is a waste of effort with a small time budget. The superiority of SO GP-GOMEA was also observed with the regularization of LS terms as a secondary objective, but not for the number of cosines second objective. The reason here is that for the former case, upon initialization already solutions are found with good values for the objective not being optimized on, whereas this is not so for the latter case. With a longer time budget, however, MO optimization can surpass SO in identifying solutions with comparable accuracy and reduced size. Additionally, MO performs better than SO when the secondary objective is uncorrelated with accuracy or individuals do not have objective values near the non-dominated front at initialization.

Initialising the individuals differently, e.g., uniformly in terms of size as done in [18] or [11], may potentially positively influence the performance of MO GP-GOMEA, but this requires further research.

Based on our findings, we recommend using SO optimization with an MO elitist archive in GP, particularly when using GP-GOMEA, as a first step for most scenarios.

REFERENCES

- [1] Peter AN Bosman. 2010. The anticipated mean shift and cluster registration in mixture-based EDAs for multi-objective optimization. In *Proceedings of the 12th annual conference on Genetic and evolutionary computation*. 351–358.
- [2] Anton Bouter, Ngoc Hoang Luong, Cees Witteveen, Tanja Alderliesten, and Peter A. N. Bosman. [n. d.]. The multi-objective real-valued gene-pool optimal mixing evolutionary algorithm. In *Proceedings of the Genetic and Evolutionary*

- Computation Conference* (New York, NY, USA, 2017-07). ACM, 537–544. <https://doi.org/10.1145/3071178.3071274>
- [3] Kalyanmoy Deb, Amrit Pratap, Sameer Agarwal, and TAMT Meyarivan. 2002. A fast and elitist multiobjective genetic algorithm: NSGA-II. *IEEE transactions on evolutionary computation* 6, 2 (2002), 182–197.
 - [4] Finale Doshi-Velez and Been Kim. 2017. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608* (2017).
 - [5] Ilan Gronau and Shlomo Moran. 2007. Optimal implementations of UPGMA and other common clustering algorithms. *Information processing letters* 104, 6 (2007), 205–210.
 - [6] Joe Harrison, Peter A.N. Bosman, and Tanja Alderliesten. 2025. Thinking Outside the Template with Modular GP-GOMEA. *arXiv preprint arXiv:2505.01262* (2025).
 - [7] Maarten Keijzer. 2003. Improving symbolic regression with interval arithmetic and linear scaling. In *European Conference on Genetic Programming*. Springer, 70–82.
 - [8] Michael Kommenda, Andreas Beham, Michael Affenzeller, and Gabriel Kronberger. 2015. Complexity measures for multi-objective symbolic regression. In *Computer Aided Systems Theory—EUROCAST 2015: 15th International Conference, Las Palmas de Gran Canaria, Spain, February 8–13, 2015, Revised Selected Papers 15*. Springer, 409–416.
 - [9] Michael Kommenda, Bogdan Burlacu, Gabriel Kronberger, and Michael Affenzeller. 2020. Parameter identification for symbolic regression using nonlinear least squares. *Genetic Programming and Evolvable Machines* 21, 3 (2020), 471–501.
 - [10] John R Koza. 1992. *Genetic Programming: On the Programming of Computers by Means of Natural Selection*. Vol. 1. MIT press Cambridge, MA, USA.
 - [11] Guillaume Lample and François Charton. 2019. Deep learning for symbolic mathematics. *arXiv preprint arXiv:1912.01412* (2019).
 - [12] Zachary C Lipton. 2018. The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queue* 16, 3 (2018), 31–57.
 - [13] Dazhuang Liu, Marco Virgolin, Tanja Alderliesten, and Peter AN Bosman. 2022. Evolvability degeneration in multi-objective genetic programming for symbolic regression. In *Proceedings of the Genetic and Evolutionary Computation Conference*. 973–981.
 - [14] Hoang N Luong and Peter AN Bosman. 2012. Elitist archiving for multi-objective evolutionary algorithms: To adapt or not to adapt. In *Parallel Problem Solving from Nature-PPSN XII: 12th International Conference, Taormina, Italy, September 1–5, 2012, Proceedings, Part II 12*. Springer, 72–81.
 - [15] Ngoc Hoang Luong, Han La Poutré, and Peter AN Bosman. 2014. Multi-objective gene-pool optimal mixing evolutionary algorithms. In *Proceedings of the 2014 Annual Conference on Genetic and Evolutionary Computation*. 357–364.
 - [16] Julian F Miller and Stephen L Smith. 2006. Redundancy and computational efficiency in cartesian genetic programming. *IEEE Transactions on evolutionary computation* 10, 2 (2006), 167–174.
 - [17] Michael O'Neill. 2009. Riccardo Poli, William B. Langdon, Nicholas F. McPhee: A Field Guide to Genetic Programming. Lulu. com, 2008, 250 pp, ISBN 978-1-4092-0073-4.
 - [18] Pablo Ramos Criado, D Barrios Rolanía, Daniel Manrique, and Emilio Serrano. 2020. Grammatically uniform population initialization for grammar-guided genetic programming. *Soft Computing* 24, 15 (2020), 11265–11282.
 - [19] Thalea Schlender, Mafalda Malafaia, Tanja Alderliesten, and Peter Bosman. 2024. Improving the efficiency of GP-GOMEA for higher-arity operators. In *Proceedings of the Genetic and Evolutionary Computation Conference*. 971–979.
 - [20] EMC Sijben, Tanja Alderliesten, and Peter AN Bosman. 2022. Multi-modal multi-objective model-based genetic programming to find multiple diverse high-quality models. In *Proceedings of the Genetic and Evolutionary Computation Conference*. 440–448.
 - [21] Bas HM Van der Velden, Hugo J Kuijff, Kenneth GA Gilhuijs, and Max A Viergever. 2022. Explainable artificial intelligence (XAI) in deep learning-based medical image analysis. *Medical Image Analysis* 79 (2022), 102470.
 - [22] Marco Virgolin, Tanja Alderliesten, Cees Witteveen, and Peter AN Bosman. 2021. Improving model-based genetic programming for symbolic regression of small expressions. *Evolutionary computation* 29, 2 (2021), 211–237.
 - [23] Marco Virgolin and Peter AN Bosman. 2022. Coefficient mutation in the gene-pool optimal mixing evolutionary algorithm for symbolic regression. In *Proceedings of the Genetic and Evolutionary Computation Conference Companion*. 2289–2297.
 - [24] Marco Virgolin, Andrea De Lorenzo, Eric Medvet, and Francesca Randone. 2020. Learning a formula of interpretability to learn interpretable formulas. In *Parallel Problem Solving from Nature-PPSN XVI: 16th International Conference, PPSN 2020, Leiden, The Netherlands, September 5–9, 2020, Proceedings, Part II 16*. Springer, 79–93.
 - [25] Eckart Zitzler and Lothar Thiele. 1999. Multiobjective evolutionary algorithms: a comparative case study and the strength Pareto approach. *IEEE transactions on Evolutionary Computation* 3, 4 (1999), 257–271.