MA-BBOB: A Problem Generator for Black-Box Optimization Using Affine Combinations and Shifts

Diederick Vermetten, Furong Ye, Thomas Bäck, Carola Doerr

MA-BBOB: A Problem Generator for Black-Box Optimization Using Affine Combinations and Shifts

DIEDERICK VERMETTEN, LIACS, Leiden University, The Netherlands
FURONG YE, Institute of Software, Chinese Academy of Science, China
THOMAS BÄCK, LIACS, Leiden University, The Netherlands
CAROLA DOERR, Sorbonne Université, CNRS, LIP6, France

Choosing a set of benchmark problems is often a key component of any empirical evaluation of iterative optimization heuristics. In continuous, single-objective optimization, several sets of problems have become widespread, including the well-established BBOB suite. While this suite is designed to enable rigorous benchmarking, it is also commonly used for testing methods such as algorithm selection, which the suite was never designed around.

We present the MA-BBOB function generator, which uses the BBOB suite as component functions in an affine combination. In this work, we describe the full procedure to create these affine combinations and highlight the trade-offs of several design decisions, specifically the choice to place the optimum uniformly at random in the domain. We then illustrate how this generator can be used to gain more low-level insight into the function landscapes through the use of exploratory landscape analysis.

Finally, we show a potential use-case of MA-BBOB in generating a wide set of training and testing data for algorithm selectors. Using this setup, we show that the basic scheme of using a set of landscape features to predict the best algorithm does not lead to optimal results, and that an algorithm selector trained purely on the BBOB functions generalizes poorly to the affine combinations.

CCS Concepts: • **Theory of computation** → **Random search heuristics**; **Design and analysis of algorithms**; *Bio-inspired optimization.*

Additional Key Words and Phrases: evolutionary computation, black-box optimization, benchmarking, algorithm selection

1 INTRODUCTION

Benchmarking is an essential step in the analysis and comparison of iterative optimization heuristics. By studying an algorithm's behavior on sets of known problems, insights into their working principles can be gained, allowing researchers to improve their algorithms. Because of this interaction between benchmark problems and algorithm design, the availability of benchmark problems inherently influences the kinds of research questions being asked and the ways in which they are answered.

Since designing effective benchmark suites is a challenging undertaking, rigorously designed suites can become widely adopted in the community. Easy access to these established benchmark suites then leads to these suites being used more often, for a wider variety of studies whose aims might not necessarily align with the design choices made in the construction of the benchmark.

Authors' addresses: Diederick Vermetten, LIACS, Leiden University, Leiden, The Netherlands, d.l.vermetten@liacs.leidenuniv.nl; Furong Ye, Institute of Software, Chinese Academy of Science, Beijing, China, f.ye@ios.ac.cn; Thomas Bäck, LIACS, Leiden University, Leiden, The Netherlands, t.h.w.baeck@liacs.leidenuniv.nl; Carola Doerr, Sorbonne Université, CNRS, LIP6, Paris, France, carola.doerr@lip6.fr.

This can lead to biases in the results when algorithms exploit properties of the benchmark, e.g., center-bias in many commonly used problems [Kudela 2022].

In continuous, single-objective optimization, the Black Box Optimization Benchmark (BBOB) [Hansen et al. 2009] is one of the most well-established suites, partly thanks to its integration into the popular COCO framework [Hansen et al. 2021]. The BBOB suite consists of 24 problems, which each represent several high-level landscape characteristics which challenge optimization algorithms in different ways. For each of these problems, instances can be created through a set of transformations, allowing researchers to test several invariances of their algorithm. Because of its popularity, studies into the specifics of the BBOB suite are numerous [Long et al. 2022; Muñoz et al. 2022, 2015], as are studies into the instance landscape which is not covered by BBOB [Long et al. 2023; Muñoz and Smith-Miles 2020; Tian et al. 2020].

One key technique used in the analysis of benchmark functions is Exploratory Landscape Analysis (ELA) [Mersmann et al. 2011]. ELA makes use of a set of randomly sampled points in the domain to calculate sets of low-level landscape features, which in aggregate aim to characterize the high-level properties of the landscape. Studies of ELA on BBOB have identified that it is relatively straightforward to detect differences between all 24 problems [Renau et al. 2021], while different instances of the same problem generally have comparable feature representations (although this is not the case for all functions [Long et al. 2022]).

The ability of ELA to differentiate between BBOB functions suggests that these features would be a useful representation for the algorithm selection problem as well. Indeed, studies on algorithm selection for continuous optimization heuristics often use BBOB as their benchmark suite [Bischl et al. 2012; Kerschke et al. 2019; Kerschke and Trautmann 2019; Kostovska et al. 2022; Lacroix and McCall 2019; Muñoz et al. 2015]. However, the challenge with using BBOB for algorithm selection lies in the evaluation of the results. One method is a leave-one-function-out technique [Nikolikj et al. 2023], which uses 23 functions for training and the remaining one for testing. This approach tends to show poor performance since each problem has been designed to represent different high-level challenges for the optimization algorithm. As such, another technique of cross-validation by splitting function instances is commonly used [Kerschke and Trautmann 2019]. However, this is likely to overfit and overestimate the performance of the selector, since the instances of different problems are inherently very similar. Thus, overfitting to biases of the instance design is an often overlooked risk [Kostovska et al. 2022].

One potential way in which this bias can be reduced is by creating new, larger sets of benchmark problems, or by creating a problem generator (e.g., the W-model in pseudo-Boolean optimization [Weise et al. 2008]). Such problem generators can then create arbitrarily many benchmark functions, to be used in the common train/test or cross-validation mechanisms from the machine learning community [Pikalov and Mironovich 2022].

Extending previous work presented at the GECCO [Vermetten et al. 2023b] and the AutoML [Vermetten et al. 2023c] conferences, we describe in this work the Many-Affine BBOB function generator (MA-BBOB). To construct new functions, we create affine combinations between existing BBOB functions, building on the work of Dietrich and Mersmann [Dietrich and Mersmann 2022]. Our generator is a generalization of their approach, designed to create unbiased combinations of problems where the contribution of the components can be smoothly varied. We highlight the core design choices made in the construction of MA-BBOB in Section 3 and illustrate their impact on the types of problems which can be created.

The parameterization of the MA-BBOB generator allows us to investigate controlled, potentially small differences in functions from both the ELA and algorithm performance perspectives. This is illustrated in Section 4 by investigating the addition of global structure (sphere function) to all other BBOB problems, as well as transitioning between pairs of BBOB problems.

Finally, in Section 5 we make use of a set of 1 000 functions generated with MA-BBOB to illustrate an algorithm selection scenario, and show that the generalization from BBOB to MA-BBOB fails to meet expectations. A comparison of the ELA-based algorithm selection approach with an artificial baseline using the weights of the affine combinations indicates that there is room to improve on the current ELA-based setup, especially when trying to generalize from BBOB to MA-BBOB.

Reproducibility: Links to source code, data, and additional figures are provided in [Vermetten et al. 2024], see Section 3.3 for more details.

2 RELATED WORK

2.1 The BBOB Problem Suite

Within the area of continuous optimization, the BBOB suite stands out as the by far most heavily used one for algorithm comparison and benchmarking. As part of the popular COCO benchmarking framework [Hansen et al. 2021], the original BBOB suite [Hansen et al. 2009] consists of 24 single-objective, noiseless functions. To date, hundreds of different optimization algorithms have been tested on the BBOB suite [Auger and Hansen 2020]. Over the years, the BBOB suite has been integrated into other benchmarking tools – as a whole, as is the case for BBOBtorch [PatrikValkovic 2021] and IOHexperimenter [de Nobel et al. 2021b], or in parts, as is the case for the Nevergrad platform [Rapin and Teytaud 2018].

Of particular note is the fact that the BBOB suite allows for the creation of different instances of each function, by choosing between sets of transformations that are applied to a "base" version. These instances aim to preserve the global function properties while varying factors such as the location of the global optimum. This is done to prevent a biased algorithm from exploiting specific structures of the function generation, e.g., a rotational bias. While the transformations are designed to keep the high-level problem structure intact, the impact of these transformations on the low-level landscape properties is not as straightforward, as some features can vary significantly between instances [Long et al. 2022].

This instance generation procedure can be seen as one of the reasons for the popularity of the BBOB suite since it allows for the analysis of several invariances in optimization algorithms. Additionally, the availability of instances allows for the evaluation of AutoML approaches, e.g., algorithm selection, by enabling a cross-validation approach across instances. This generally leads to better results than cross-validation on a function-level, since the functions themselves have been designed specifically to contain different challenges to the optimizer, so leaving a function out of the training set generally makes the transfer between train and test challenging [Kostovska et al. 2022; Nikolikj et al. 2023].

2.2 Filling the Instance Space

While using function instances allows the BBOB suite to cover a wider range of problem landscapes than the raw functions alone, there are limits to the types of landscapes which can be created in this way. Instance space analysis is an emerging research direction which investigates the extent to which the existing benchmark functions cover the instance space, and whether they do so evenly with regard to challenges for different optimization algorithms [Smith-Miles and Muñoz 2023]. This analysis allows for studies into different mechanisms for creating new problems, ranging from random function generators [Tian et al. 2020] to genetic programming [Long et al. 2023; Muñoz and Smith-Miles 2020]. In parallel, questions regarding appropriate choice of representative benchmark functions from large sets of problems are gaining more attention, since benchmarking on thousands of functions is still not computationally feasible [Cenikj et al. 2022].

In addition to the function generation methods mentioned above, it has recently been proposed to use affine combinations between pairs of BBOB functions to generate new benchmark functions [Dietrich and Mersmann 2022]. These combinations have been shown to smoothly fill the space of low-level landscape properties, as measured through a low-dimensional projection of a set of ELA features [Mersmann et al. 2011]. These results have shown that even a relatively simple function creation procedure has the potential to give us new insights into the inner workings of benchmark problems and their low-level characteristics.

Similarly, the ability to combine existing BBOB functions would allow for more extensive investigations into the behavior of algorithm selection methods, particularly those based on ELA features [Kerschke et al. 2019], as smooth transitions in the landscape should correspond to smooth transitions in algorithm behavior as well. Our extension of the affine function combination generation mechanism, which is summarized in Section 3, aims to increase the variety of available functions while avoiding issues related to scaling differences between component functions [Vermetten et al. 2023c,b].

3 MANY-AFFINE BBOB

3.1 Pairwise Affine Combinations

To create affine combinations between two BBOB functions, we use a slightly modified version of the procedure proposed in [Dietrich and Mersmann 2022]. Specifically, we define the combination C as follows:

$C(F_{1,I_1}, F_{2,I_2}, \alpha)(x) = 10^X$, with

$$X = \left(\alpha \log_{10} \left(F_{1,I_1}(x) - F_{1,I_1}(O_{1,I_1}) \right) + (1 - \alpha) \log_{10} \left(F_{2,I_2}(x - O_{1,I_1} + O_{2,I_2}) - F_{2,I_2}(O_{2,I_2}) \right) \right)$$

Here, F_1, I_1, F_2, and I_2 are the two base functions and their instance numbers, as defined in BBOB [Hansen et al. 2009]. O_{1,I_1} and O_{2,I_2} represent the location of the optimum of functions F_{1,I_1} and F_{2,I_2} respectively. The transformation to x when evaluating F_{2,I_2} is performed to make sure the location of the optimum is at O_{1,I_1}. To avoid invalid input to the $\log_{10}$ calculation, these values are capped at a value of 10^{-12}. As opposed to the original definition, we subtract the optimal values before aggregating and take a logarithmic mean between the problems. This way, we can use consistent values for α across problems, without having to perform the entropy-based selection performed in [Dietrich and Mersmann 2022]. It has the additional benefit of ensuring the objective value of the optimal solution is always 0, so the comparison of performance across instances and problems is simplified. In Figure 1, we illustrate the change in the landscape for the combination of F21 and F1, for different values of α.

3.2 Combining Multiple BBOB Functions

We extend the pairwise affine combinations from Section 3.1 to create a function generator which uses affine combinations of multiple BBOB functions. In particular, our generator is defined as follows:

$$MA\text{-}BBOB(\vec{W}, \vec{I}, \vec{X}_{\text{opt}})(x) = R^{-1} \Big(\sum_{i=1}^{24} W_i \cdot R_i \big(F_{i,I_i}(x - \vec{X}_{\text{opt}} + O_{i,I_i}) - F_{i,I_i}(O_{i,I_i}) \big) \Big)$$

Here, $\vec{W}$ and $\vec{I}$ are 24-dimensional vectors containing the weight and instance identifiers respectively, and $\vec{X}_{\text{opt}}$ is the location of the optimum, which we generate uniformly at random in the domain $[-5, 5]^d$. Finally, R_i and R^{-1} are rescaling functions, defined in Section 3.2.1. We highlight the motivation behind each of these design choices in the following subsections.
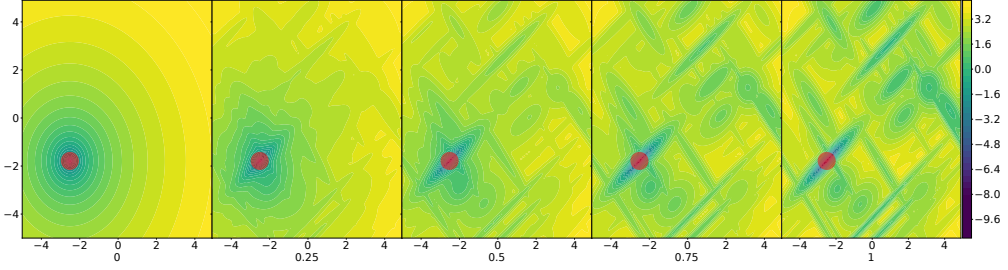
Fig. 1. Evolution of the landscape (log-scaled function-values) of the affine combination between F21 ($\alpha = 1$) and F1 ($\alpha = 0$), instance 1 for both functions, for varying α in 0.25 increments. The red circle highlights the location of the global optimum.

Function ID	1	2	3	4	5	6	7	8	9	10	11	12
Scale Factor	11.0	17.5	12.3	12.6	11.5	15.3	12.1	15.3	15.2	17.4	13.4	20.4
Function ID	13	14	15	16	17	18	19	20	21	22	23	24
Scale Factor	12.9	10.4	12.3	10.3	9.8	10.6	10.0	14.7	10.7	10.8	9.0	12.1

Table 1. Final scale factors used to generate MA-BBOB problems.

3.2.1 Scaling of Function Values. While the geometric weighted average used in Section 3.1 between component functions reduces the impact of small differences in scale, some BBOB problems vary by orders of magnitude, which can still cause one function to dominate the combined landscape. To address this, we add a rescaling function to the MA-BBOB definition, which transforms the log-precision on each component function into approximately $[0, 1]$ before the transformation. This is done by capping the log-precision at -8, adding 8 so the minimum is at 0 and dividing by a *scale factor S_i*. This procedure aims to make the target precision of 10^2 similarly easy to achieve on all component problems. We thus get the following scaling functions:

$$R_i(x) = \frac{\max(\log_{10}(x), -8) + 8}{S_i}$$
$$R^{-1}(x) = 10^{(10 \cdot x) - 8}$$

To determine practical scale factors, we collect a set of 50 000 random samples and evaluate them. We then aggregate the resulting function values (transformed to log-precision) in several ways: min, mean, max, (max + min)/2. In Figure 2, we show the differences between these methods for a selected problem in 2d. Somewhat subjectively, we select the (max + min)/2 scaling as the technique to use for the MA-BBOB generator. To ensure we don't have to repeat this sampling procedure each time we instantiate the problem in a new dimensionality, we investigate the relation between dimensionality and the chosen scale factor calculation. This is visualized in Figure 3, where we see that, with an exception for the smallest dimensions, the values remain quite stable. Because of this, we make use of a static scale factor rather than defining one for each dimension individually. The final factors used are calculated as a rounded median of the values from Figure 3, and shown in Table 1.

3.2.2 Instance Creation. Another design choice we made was to place the optimum of the combined function uniformly at random in the domain ($[-5, 5]^d$). This differs from the earlier versions used
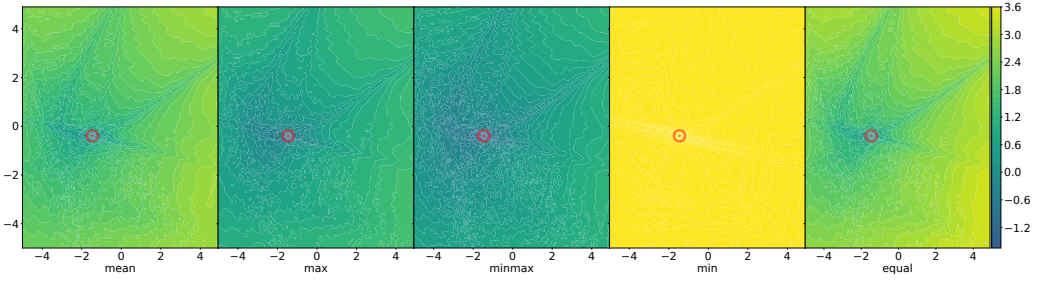
Fig. 2. Log-scaled fitness values of an example of a single many-affine function with 5 different ways of scaling. The first 4 are taking the mean, max, $(\text{max} + \text{min})/2$ and min of 50 000 random samples ($\mathcal{U}([-5, 5]^d)$) to create the scale factor, while the fifth ('equal') option does not make use of this scaling.
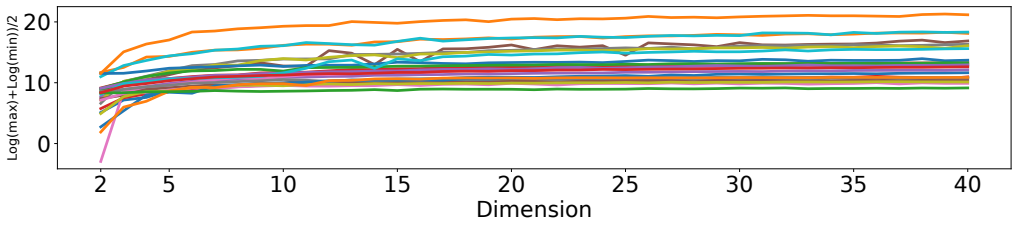

Fig. 3. Evolution of the log-scaled $(\text{max} + \text{min})/2$ scaling factor, relative to the problem dimension. The values are based on 50 000 samples. Each line corresponds to one of the 24 BBOB functions.
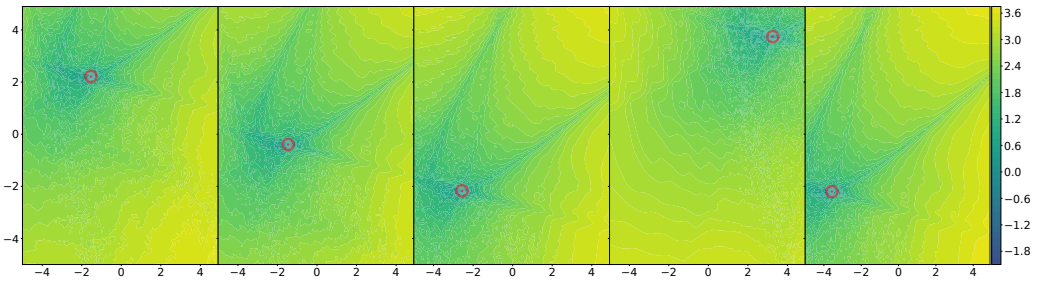

Fig. 4. Log-scaled fitness values of an example of a single many-affine function with changed location of optimum.

for pairwise combinations of BBOB functions [Dietrich and Mersmann 2022; Vermetten et al. 2023b], where the optimum of one of the component functions was re-used. However, the biases in the original BBOB instance generation procedure would then be transferred into the combinations as well [Long et al. 2022]. Since our function generator does not have to guarantee the preservation of global function properties, we take the risk of moving parts of the regions of interest outside the domain to have a less biased location of the global optimum. Figure 4 shows how a $2d$-function changes when moving the optimum location.

3.2.3 Sampling Random Functions. To allow for the usage of MA-BBOB as a function generator, we need to create a default setting to generate useful weight-vectors. This could be done uniformly at random (given a normalization step). However, in this way, the weight for every component is likely to be non-zero, so most functions contribute to the final combination, erasing the possibility of generating unimodal problems since some multimodality will always be included from some of the multimodal component functions.

To address this issue, we adapt the sampling technique to combine fewer component functions on average. Our approach is based on a threshold value to determine which functions contribute to the problem. The procedure for generating weights is thus as follows: (1) Generate initial weights uniformly at random, (2) adapt the threshold to be the minimum of the user-specified threshold and the third-highest weight, (3) this threshold is subtracted from the weights, all negative values are set to 0. The second step is to ensure that at least two problems always contribute to the new problem. We decide to set the default value at $T = 0.85$, such that on average 3.6 (i.e., 15% of 24) problems will have a non-zero weight.

3.3 Availability and Reproducibility

The MA-BBOB generator as described here is made available directly in the IOHexperimenter package [de Nobel et al. 2021b] as part of the IOHprofiler project [Doerr et al. 2018]. This enables us to use any of the built-in logging and tracking options of IOHexperimenter. In particular, it allows us to store the performance data into a file format which can be directly processed into IOHanalyzer [Wang et al. 2022] for post-processing.

The experiments and corresponding analyses described in this paper (including some additional versions for presented figures) can be reproduced by following the steps outlined in our Zenodo repository [Vermetten et al. 2024], which also contains the performance data and the ELA-features for all analyzed functions. This repository documents the steps required to generate, analyze and visualize the data described in this paper using an interactive notebook as well as several guided scripts.

4 PAIRWISE AFFINE COMBINATIONS

For the first analysis of the MA-BBOB functions, we limit ourselves to the combination of pairs of functions. This allows a more low-level investigation into the transition of both landscape features and algorithm performance.

4.1 Setup

For the algorithm performance-based analysis, we make use of a portfolio of five algorithms. Of these, three are accessed through the Nevergrad framework [Rapin and Teytaud 2018]:

- Differential Evolution (DE) [Storn and Price 1997]
- Constrained Optimization by Linear Approximation (Cobyla) [Powell 1994]
- Diagonal Covariance Matrix Adaptation Evolution Strategy (dCMA-ES) [Hansen and Ostermeier 2001]

The remaining algorithms are two modular algorithm families: modular Differential Evolution [Vermetten et al. 2023a] (modde) and modular CMA-ES [de Nobel et al. 2021a] (modcma). All algorithms, including the modular ones, use default parameter settings. Each run we perform has a budget of $2\,000d$, where d is the dimensionality of the problem. We perform 50 independent runs per function. For the pairwise function combinations, we stick to the terminology introduced in Section 3.1 for easier comparison with the previous results in [Vermetten et al. 2023b].

In the remainder of this section, we set $I_2 = 1$. As such, when discussing the *instance* of a pairwise affine combination $C(F_1, I_1, F_2, I_2, \alpha)$, we are referring to I_1. Note that we also introduce the uniform sampling of the optima for these experiments, following the description in Section 3.2.2. For our performance measure, we make use of the normalized area over the convergence curve (AOCC), to be maximized. The AOCC is an anytime performance measure, which is equivalent to the area under the cumulative distribution curve (AUC) given infinite targets for the construction of the ECDF [López-Ibáñez et al. 2024]. This measure is thus slightly more precise than the AUC, and can easily be computed online. To remain consistent with the performance measures used in our previous work, and analysis of results on BBOB in general, we use 10^2 and 10^{-8} as the bounds for our function values, and perform a log-scaling before calculating the AOCC. We thus calculate the normalized AOCC of a single run as follows:

$$AOCC(\vec{y}) = \frac{1}{B} \sum_{i=1}^{B} 1 - \frac{\min(\max((\log_{10}(y_i), -8), 2) + 8}{10}$$

where $\vec{y}$ is the sequence of best-so-far function values reached, B is the budget of the run. To obtain the AOCC over multiple runs, we simply take the average.

For the landscape analysis, we make use of the pFlacco [Prager and Trautmann 2024] package to calculate the ELA features. We use a sample size of $1\,000d$ points, sampled using a Sobol' sequence. We note this large sample size is used to remove some of the inherent variability in the ELA features, even though practical applications usually rely on much smaller budgets. To be consistent with our previous work [Vermetten et al. 2023c], we don't include features which require additional samples and remove all features which lead to NaN-values or remain static for all functions, resulting in a set of 44 features.

4.2 Adding Global Structure

For the first set of experiments, we make use of affine combinations where we combine each function with F1: the sphere model (as the F_2 function in the combination). As can be seen in Figure 1, adding a sphere model to another function creates an additional global structure that can guide the optimization toward the global optimum. As such, these kinds of combinations might allow us to investigate the influence of an added global structure on the performance of optimization algorithms. While to some extent this can already be investigated by comparing results on the function groups of the original BBOB with different levels of global structure, the affine function combinations allow for a much more fine-grained investigation. For these experiments, we will vary the function between the sphere ($\alpha = 0$) and the target function ($\alpha = 1$) by varying α in steps of 0.05.

In Figure 5a, we can see that the performance of CMA-ES does indeed seem to move smoothly between the sphere and the function with which it is combined. It is however interesting to note the differences in speed at which this transition occurs. For example, while the final performance on functions 3 and 10 seems similar, the transition speed differs significantly. These differences in transition speeds might in part be influenced by local differences in scaling, rather than the global corrections applied in Section 3.2.1, but could also indicate a difference in the impact of adding global structure to the function.

We can perform a similar analysis on other optimization algorithms. In Figures 5b and 5c, we show the same heatmap as Figure 5a, but for Differential Evolution and Cobyla, respectively. It is clear from these heatmaps that the performance of DE is more variable than that of CMA-ES, while Cobyla's performance drops off much more quickly. The overall trendlines for DE do seem to be somewhat similar to those seen for diagonal CMA-ES: the transition points between high and
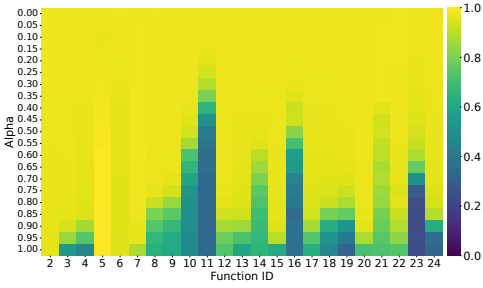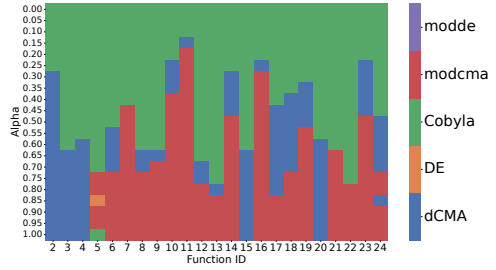
(a) AOCC values for Diagonal CMA-ES.
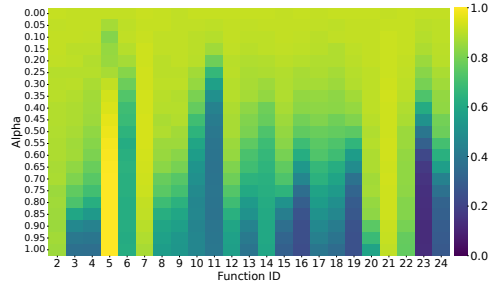
(b) AOCC values for Differential Evolution.
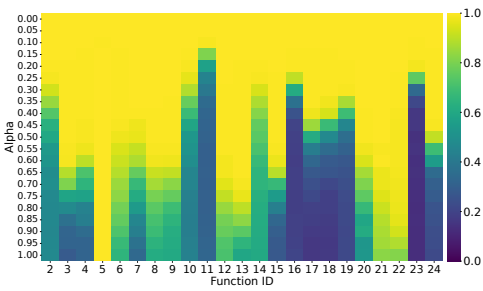
(c) AOCC values for Cobyla.

(d) Best performing algorithm from the portfolio, based on AOCC.

Fig. 5. Normalized area under the ECDF curve of three selected algorithms (a–c) and best-ranking algorithm from the full portfolio (d) for each combination of the BBOB-function (x-axis) with a sphere model, for given values of α (y-axis) AOCC is calculated after 10 000 function evaluations, based on 50 runs on 50 instances (and location of optimum). Note that $\alpha = 0$ corresponds to the sphere function.

low AOCC in Figure 5b are comparable to those seen in Figure 5a. There are however still some differences in behavior, especially relative to Cobyla. These differences then lead to the question of whether there exist transition points in ranking between algorithms as well. Specifically, if one algorithm performs well for $\alpha = 0$ but gets overtaken as $\alpha \rightarrow 1$, exploring this change in ranking would give further insight into the relative strengths and weaknesses of the considered algorithms.

To study the impact on the relative ranking of algorithms, we make use of the full portfolio of 5 algorithms and rank them based on AOCC on each affine function combination. We then visualize the top ranking algorithm on each setting in Figure 5d. From this figure, we can see that Cobyla deals well with the sphere model, managing to outperform the other algorithms when the weighting of the sphere is relatively high. Then, after a certain threshold, the CMA-ES variants consistently outperform the rest of the portfolio, with dCMA taking over when Cobyla is no longer preferred. However, as α increases further, and the influence of the sphere model diminishes, an interesting pattern seems to occur. For several problems, there is a second transition point to modcma, indicating that the differences in default parameterizations between the used libraries have a large impact on the algorithms' behavior. Based on some preliminary experiments, we identified that one significant factor could be related to the initial stepsize, which is smaller for dCMA, and thus might lead to it becoming more easily stuck in local optima when the global structure is not as strong.
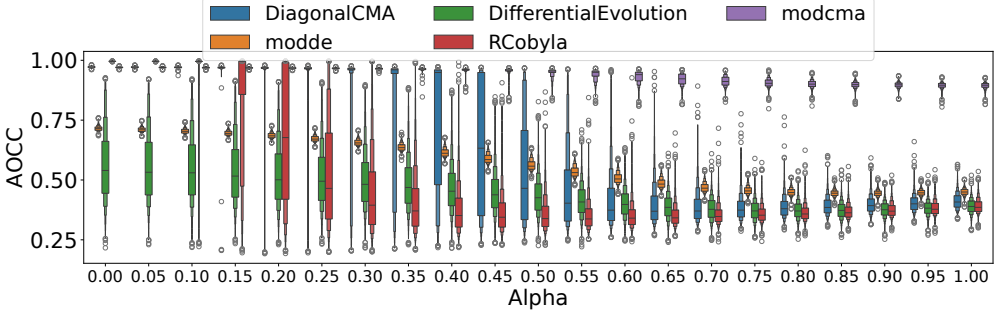
Fig. 6. Distribution of AOCC values for 5 algorithms on the affine combinations between F10 ($\alpha = 1$) and F1 ($\alpha = 0$), for selected values of α.

In order to better understand what the transitions in algorithm ranking look like, we can zoom in on one of the functions and plot the distribution of AOCC for all values of α. This is done in Figure 6, where we look at the combination between F10 and the sphere model. In this figure, we observe that Cobyla is very effective at optimizing the sphere and the combinations with low α. However, when α increases, Cobyla quickly starts to fail, while, for example, DiagonalCMA still manages to solve most instances at $\alpha = 0.25$ with similar AOCC. As α increases further, the modcma's performance remains stable, showing only a minor drop in performance relative to the one seen in dCMA.

4.3 Impact on ELA Features

In addition to the performance perspective, we can also look at what happens to the landscape feature of the BBOB functions as we add increasingly more influence from the sphere function. Since we measure 44 different ELA features, our analysis of the impact is rather more high-level than the algorithm performance viewpoint, as we first aim to capture the overall stability of the features for increasing α values (from 0 to 1 with steps of 0.05). This is measured as the sum of absolute differences in feature mean for consecutive α's, which is plotted in Figure 7. From this figure, we can see that the mean of most features remains quite stable, with a few notable exceptions. In particular, functions 16 and 23 show many feature changes from the sphere, which matches observations from e.g. [Renau et al. 2019; Škvorc et al. 2020].

In addition to the differences between functions, it is also clear to see that features don't all behave in the same way. Since Figure 7 only shows absolute changes in mean, the deviation between instances might also play a role. To analyze this in some more detail, we select a single function (F10) and look at the evolution of both the feature mean and its standard deviation in Figure 8. From this figure, we can see that the mean of each feature seems to transition rather smoothly between the two component functions, in a similar way to the performance plots in Section 4.2. However, we should note that the standard deviation of many features is relatively large for each α value.

4.4 Impact of Optimum Location and the Instance

As can be seen from the relatively large variance in both ELA features and algorithm performance, the instance and location of the optimum can have a major impact on both the landscape and the corresponding algorithm behaviour. In particular, the way in which we defined an instance in our
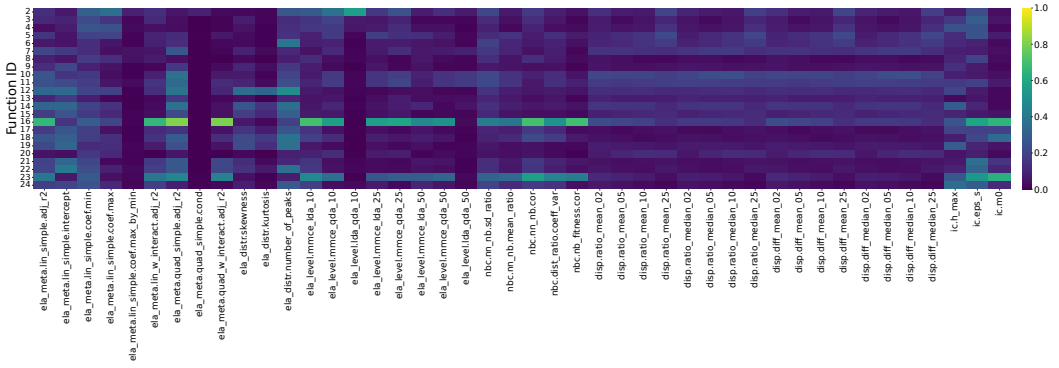
Fig. 7. Total changes in each ELA feature when transitioning from the sphere to the function indicated in the row. Each cell represents the sum of differences in mean between pairs of consecutive values of α, so high values indicate a large total change in mean, while low values indicate features which remain stable throughout the transition.
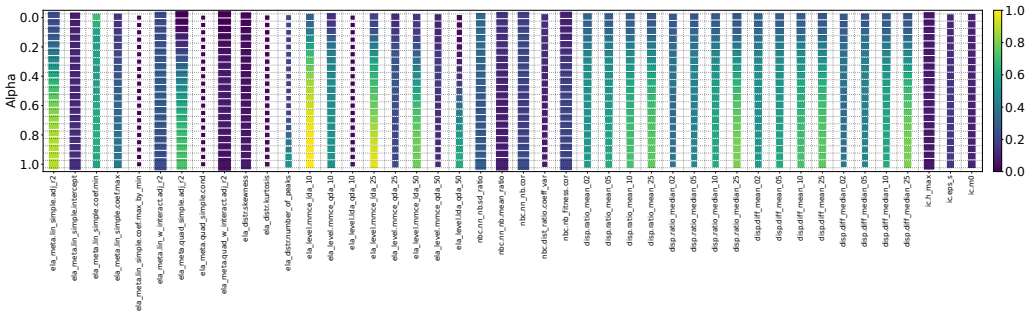

Fig. 8. Evolution of ELA features with changing between Sphere ($\alpha = 0$) and F10 ($\alpha = 1$). The color indicates the mean of the feature over the 50 instances (lighter = larger), while the size indicates the variance (larger = higher variance).

setup is not necessarily equivalent to the common interpretation, e.g., from BBOB. Since we allow the optimum of a component function to be moved anywhere in the domain, this can lead to large parts of the original function no longer being reachable. This is the main reason why some BBOB functions have very restricted distributions for their optima, which can be seen by analyzing the overall distribution of optima across all BBOB functions, visualized in Figure 9 and previously observed in e.g.[Long et al. 2022].

To analyse how much the algorithms in our portfolio are influenced by the choice of instance and location of optimum, we determine the relative impact of different instances for each function (F_1 and α value). This is done by first averaging the performance across all 50 runs on each instance. By dividing this by the total variance present across all runs on all instances of that function, we obtain a relative measure of 'stability' across instances, which is visualized in Figure 10. This figure shows that some algorithms are inherently more impacted by the instance/location of the optimum (modde), while, for example, for Cobyla, the variance increases with increasing α, which suggests that it is very stable on the sphere problem, but becomes much more impacted by variations in the landscape when more non-sphere influence is added.
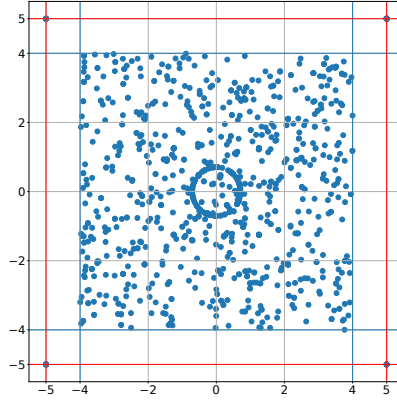
Fig. 9. Location of optima of the 24 2d BBOB functions (1000 random instances). The red lines mark the commonly used box-constraints of $[-5, 5]^d$.
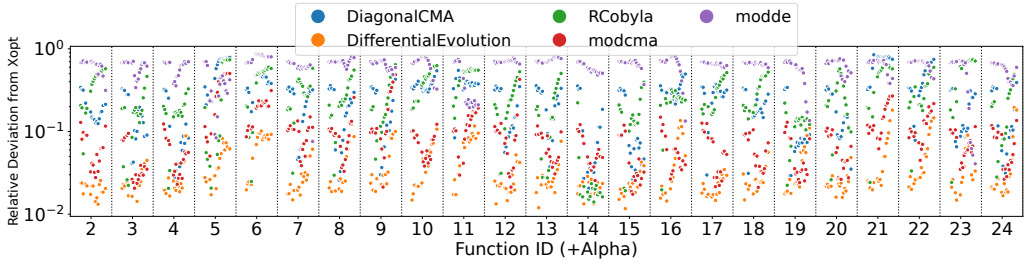

Fig. 10. Relative deviation in AOCC caused by the change of the global optimum for all combinations of BBOB functions with the sphere model (calculated as deviation per instance divided by deviation across all instances). X-axis indicates changing function ID, and within each function ID the transition goes from $\alpha = 0$ to $\alpha = 1$ (left to right).

To further analyse the impact of this increased flexibility in terms of function generation, we perform an experiment involving two versions of the original BBOB functions. The first is the data from Section 4.2 where $\alpha \in \{0, 1\}$, which corresponds to data for 50 instances of each function, with each of them having its optimum moved to a random point in the domain. The second set of instances are created by taking the same instances of the BBOB functions, but not shifting their optimum (the rescaling from Section 3.2.1 is still applied). This allows us to compare the influence of moving the optimum on the landscape of the resulting problem. In Figure 11, we compare the distribution of all features on these two versions of BBOB function 23. Note that versions of this figure for other functions are available in our repository [Vermetten et al. 2024].

Finally, we can take an aggregated view of the features, and project the 44 dimensional space into two dimensions using PCA on the original BBOB versions. Then, we can plot the moved versions of the function into the same space, and observe the differences. The result of this projection is shown in Figure 12, where we see that many functions are moved much closer to the center of the projected space. This suggests that some of the 'unique' feature combinations present in the original BBOB functions are being lost when moving their optimum. This happens because large
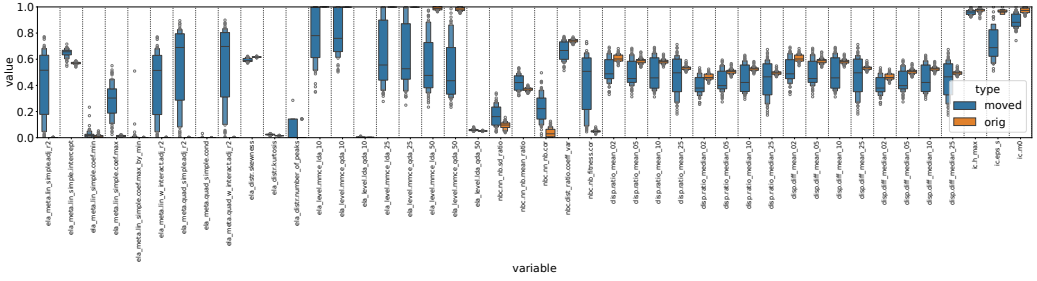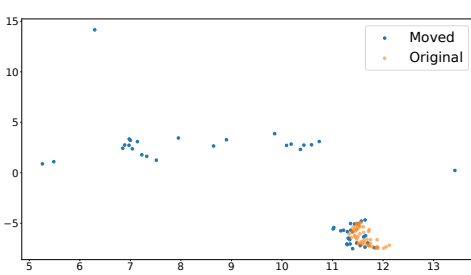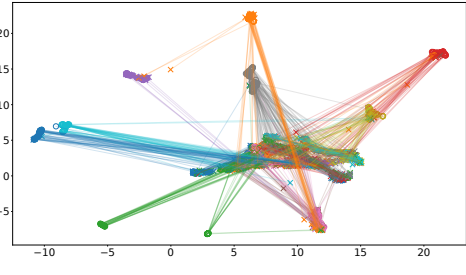
Fig. 11. Distribution of normalized ELA features for the BBOB instance creation procedure and the same instance moved to have an optimum location uniformly in the domain, for F23. (other functions in the bbob vs moved folder)

(a) Projection of the moved and original version of F18 (50 instances).

(b) Projection of all BBOB functions connecting the original (circles) and moved (crosses) versions of each function.

Fig. 12. UMAP projection trained on original bbob, then used to plot both the moved and original functions in 2D.

parts of the function are moved outside of the domain, and replaced by parts which were originally located outside the bounds. For some functions, these components are exponentially increasing, leading to a large part of the space which is dominated by these artifacts, which is represented in the ELA-features.

In addition to moving the location of the optima, we also perform a rescaling of the BBOB functions before using them as components. Since it has been shown that the ELA-features used in this work, with the notable exception of the linear slope intercept, are invariant to scale of the function [Prager and Trautmann 2023], we do not provide a detailed analysis of this aspect of the function transformation.

4.5 Pairwise Combinations

While combining functions with a sphere model can be viewed as adding global structure to a problem, combinations between other functions can provide interesting insights into the transition points between different types of problems. To illustrate the kinds of insights that can be gained from these combinations, we select a subset of 5 functions and collect performance data on each combination with the same 21 α values (with both orderings of the function). We show the performance in terms of normalized AOCC of diagonal CMA-ES on these function combinations in
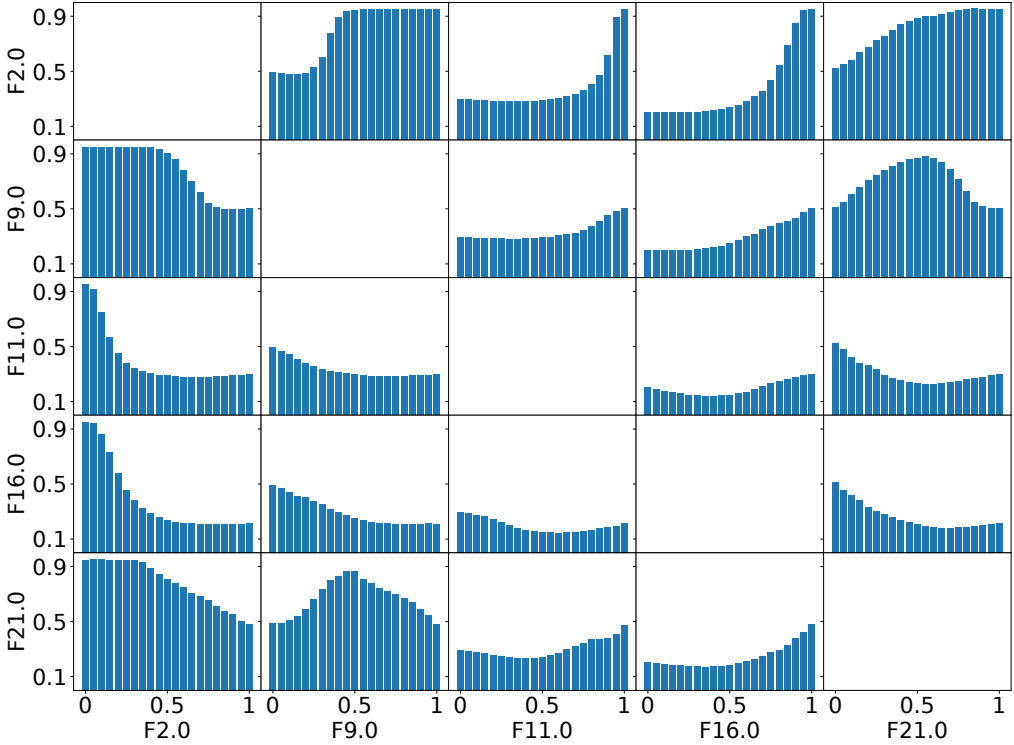
Fig. 13. Normalized area over the convergence curve for Diagonal CMA-ES on each of the affine combinations between the selected BBOB problems. Each facet corresponds to the combination of the row and column function, with the x-axis indicating the used α. AOCC values are calculated based on 50 runs on 25 instances, with a budget of 10 000 function evaluations.

Figure 13. Note that for $\alpha = 1$, we are using the function specified in the column label, while for $\alpha = 0$ we have the function specified in the row label.

When comparing this Figure to its equivalent from the GECCO paper [Vermetten et al. 2023b], it is important to note the fact that Figure 13 is almost fully symmetric around the diagonal, which was not the case in the GECCO paper. Even though we might expect (F_1, F_2, α) to be similar to $(F_2, F_1, 1 - \alpha)$, this was not the case when the location of the global optimum was selected as the optimum of one of the component functions, as different BBOB functions can have significantly different distributions of potential global optima [Long et al. 2022]. This is in large part the reason why we enabled the MA-BBOB generator to sample the optimum uniformly at random in the domain. While Section 4.4 showed that this can potentially move interesting parts of some component functions outside the domain, we view this as a worthwhile tradeoff to achieve fully unbiased global optima.

From Figure 13, we can also see that the transition of performance between the two extreme α values is mostly smooth. While there are some rather quick changes, e.g., for the transition between F2 and F11, these seem to be the exception rather than the rule. Particularly interesting are the settings where the performance of affine combinations between two functions proves to be much

easier or harder than the functions which are being combined. For example, this is the case for the combinations of F21 and F9.

5   COMBINING MULTIPLE FUNCTIONS: TESTING GENERALIZABILITY

For our final set of experiments, we make use of a set of 1000 functions generated using the setup described in Section 3.2.3. This data is taken directly from the Zenodo repository [Vermetten et al. 2023d], and contains both ELA and performance data (for the same set of algorithms described in Section 4.1, but using the original AUC measure instead of the AOCC). In [Vermetten et al. 2023c] we analyzed this data to understand the MA-BBOB instance generation procedure, with the goal of generating a wide set of benchmark problems on which algorithm selection and other automated machine learning techniques can be tested.

In this experiment, we take the perspective of algorithm selection and train a random forest model to predict the best algorithm to use for each function, based either on the ELA features of the problem or the weights of the component functions. We can then compare the loss in terms of AUC relative to the virtual best solver (VBS) for both of these models, in different training contexts. We can either use the common cross-validation setup, or attempt to test for generalization ability based only on the original BBOB functions. In Figure 14a we show the cumulative loss for the cross-validation setup on the 5-dimensional functions. From this, we can see that the ELA-based selector performs worse than the one based on the weights. This confirms the previous observation that the ELA features might not be sufficiently representative to accurately represent the problems in a way which is relevant for ranking optimization algorithms.

In order to better estimate how much the structure of the ELA features helps the prediction, we can add in a naive baseline. This is created by shuffling the labels (best ranked algorithm) of all samples before training. This shuffled model is in essence just a selector based on the frequency of labels in the training data, and the difference between this version and the original ELA-based selector shows how much the structure of the ELA-features helps improve the predictions. The results for the cross-validation setup in 2D are shown in Figure 14b, where we see that the benefit over most of the individual algorithms is inherent to the selected portfolio, since the shuffled model outperforms all algorithms except modcma. This suggests that our algorithm portfolio is severely unbalanced.

When looking at the generalization task, this imbalance is exacerbated further, since for MA-BBOB the modcma is ranked first on an even larger fraction of functions than on BBOB, as shown in Figure 15. In combination with the added challenge of transferring to a new suite, this leads to our algorithm selection models being outperformed by the modcma, which is the Single Best Solver (SBS) in this case, as illustrated in Figure 16. While the algorithm portfolio is partly responsible for this shortcoming, the generalization ability does not significantly improve when removing the modcma from our portfolio. This suggests that training on the original BBOB instances does not sufficiently represent the challenges faced in the MA-BBOB suite. An important aspect of the challenge of this transfer is the location of the optima, as discussed in Section 4.4.

6   CONCLUSIONS AND FUTURE WORK

We propose MA-BBOB as a way to generate new benchmark problems by creating affine combinations of rescaled and translated BBOB functions. Our approach generalizes earlier work by Dietrich and Mersmann who investigated combinations of pairs of BBOB functions and the transitions between them [Dietrich and Mersmann 2022]. The construction of MA-BBOB still allows for detailed studies into the relation between landscape features and algorithm performance, but at the same time, it has the potential to be used for testing algorithm selection techniques for optimization algorithms.
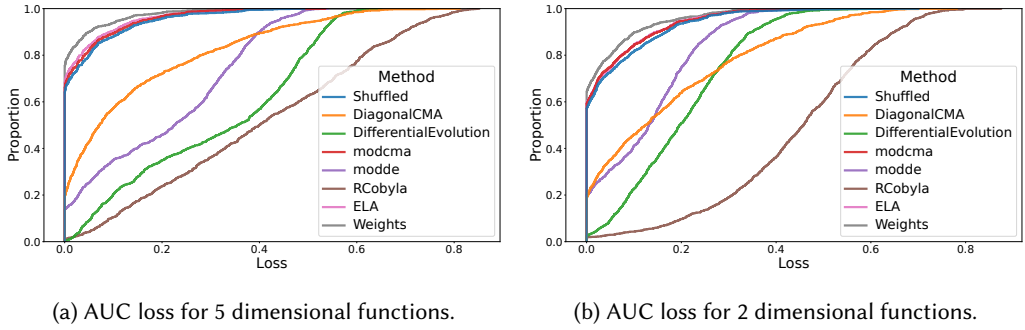
(a) AUC loss for 5 dimensional functions.

(b) AUC loss for 2 dimensional functions.

Fig. 14. Cumulative loss (AUC) for different models: cross-validation (mixture of BBOB + MA-BBOB generated combinations) based on weights and ELA, and each of the single-algorithm models.
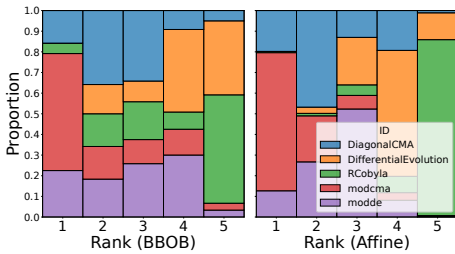

Fig. 15. Distribution of ranks based on per-function AUC after 10 000 evaluations.
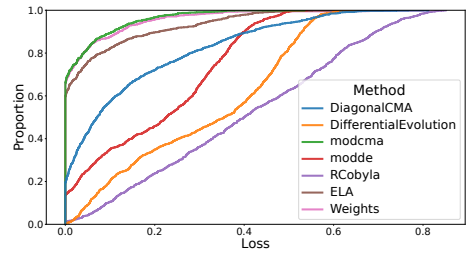
Fig. 16. Cumulative loss (AUC) on the 5 dimensional MA-BBOB problems for models trained on the BBOB functions, and each of the single-algorithm models.

While MA-BBOB does not bias the location of the optimum to any part of the domain, this comes at a potential cost in terms of instance creation. Specifically, we have shown that using this procedure, the landscape features of the original BBOB features can be drastically changed. Because of this, we shouldn't consider two MA-BBOB functions as instances of the same problem if they share the same weights, but view them as different problems instead.

When analyzing the changes in both performance and landscape of combinations of two functions, we observe relatively smooth transitions between the components. There is however a large variance in both of these metrics. It is additionally worthwhile to note that combinations between two functions might become easier or harder for a given optimization algorithm to solve. Further studies into these transitions might help reveal some closer relation between the performance on the components and on their affine combination, and link this with the changes in landscape characteristics.

The selection of 'representative' instance collections still remains to be done. Another important step for future work is to test the generalization ability of AutoML systems that are trained on MA-BBOB functions and tested on numerical black-box optimization problems that do not originate from the BBOB family. In this context, our basic Random Forest-based algorithm selector indicates that the ELA features might not be as suitable for this generalization task as expected, motivating further research on feature engineering for single-objective continuous black-box optimization.

ACKNOWLEDGMENTS

REFERENCES

Anne Auger and Nikolaus Hansen. 2020. A SIGEVO Impact Award for a Paper Arising from the COCO Platform: A Summary and Beyond. https://evolution.sigevo.org/issues/HTML/sigevolution-13-4/home.html. Issue 3.

Bernd Bischl, Olaf Mersmann, Heike Trautmann, and Mike Preuß. 2012. Algorithm selection based on exploratory landscape analysis and cost-sensitive learning. In *Proceedings of the 14th annual conference on Genetic and evolutionary computation*. 313–320.

Gjorgjina Cenikj, Ryan Dieter Lang, Andries Petrus Engelbrecht, Carola Doerr, Peter Korošec, and Tome Eftimov. 2022. SELECTOR: selecting a representative benchmark suite for reproducible statistical comparison. In *Proc. of Genetic and Evolutionary Computation Conference (GECCO)*. 620–629.

Jacob de Nobel, Diederick Vermetten, Hao Wang, Carola Doerr, and Thomas Bäck. 2021a. Tuning as a means of assessing the benefits of new ideas in interplay with existing algorithmic modules. In *Proc. of Genetic and Evolutionary Computation Conference (GECCO'21, Companion material)*. ACM, 1375–1384. https://doi.org/10.1145/3449726.3463167

Jacob de Nobel, Furong Ye, Diederick Vermetten, Hao Wang, Carola Doerr, and Thomas Bäck. 2021b. IOHexperimenter: Benchmarking Platform for Iterative Optimization Heuristics. *CoRR* abs/2111.04077 (2021). arXiv:2111.04077 https://arxiv.org/abs/2111.04077

Konstantin Dietrich and Olaf Mersmann. 2022. Increasing the Diversity of Benchmark Function Sets Through Affine Recombination. In *Proc. of Parallel Problem Solving from Nature (PPSN'22) (LNCS, Vol. 13398)*, Günter Rudolph, Anna V. Kononova, Hernán E. Aguirre, Pascal Kerschke, Gabriela Ochoa, and Tea Tusar (Eds.). Springer, 590–602. https://doi.org/10.1007/978-3-031-14714-2_41

Carola Doerr, Hao Wang, Furong Ye, Sander van Rijn, and Thomas Bäck. 2018. IOHprofiler: A Benchmarking and Profiling Tool for Iterative Optimization Heuristics. *CoRR* abs/1810.05281 (2018). arXiv:1810.05281 http://arxiv.org/abs/1810.05281

Nikolaus Hansen, Anne Auger, Raymond Ros, Olaf Mersmann, Tea Tušar, and Dimo Brockhoff. 2021. COCO: A platform for comparing continuous optimizers in a black-box setting. *Optim. Methods Softw.* 36, 1 (2021), 114–144.

Nikolaus Hansen, Steffen Finck, Raymond Ros, and Anne Auger. 2009. *Real-Parameter Black-Box Optimization Benchmarking 2009: Noiseless Functions Definitions*. Technical Report RR-6829. INRIA. https://hal.inria.fr/inria-00362633/document

Nikolaus Hansen and Andreas Ostermeier. 2001. Completely Derandomized Self-Adaptation in Evolution Strategies. *Evolutionary Computation* 9, 2 (2001), 159–195. https://doi.org/10.1162/106365601750190398

Pascal Kerschke, Holger H. Hoos, Frank Neumann, and Heike Trautmann. 2019. Automated Algorithm Selection: Survey and Perspectives. *Evol. Comput.* 27, 1 (2019), 3–45. https://doi.org/10.1162/evco_a_00242

P. Kerschke and H. Trautmann. 2019. Automated Algorithm Selection on Continuous Black-Box Problems by Combining Exploratory Landscape Analysis and Machine Learning. *Evolutionary Computation* 27, 1 (2019), 99–127. https://doi.org/10.1162/evco_a_00236

Ana Kostovska, Anja Jankovic, Diederick Vermetten, Jacob de Nobel, Hao Wang, Tome Eftimov, and Carola Doerr. 2022. Per-run Algorithm Selection with Warm-starting using Trajectory-based Features. In *Proc. of Parallel Problem Solving from Nature (PPSN'22) (LNCS, Vol. 13398)*. Springer, 46–60. https://doi.org/10.1007/978-3-031-14714-2_4 Free version available at https://arxiv.org/abs/2204.09483.

Jakub Kudela. 2022. A critical problem in benchmarking and analysis of evolutionary computation methods. *Nature Machine Intelligence* 4, 12 (2022), 1238–1245.

Benjamin Lacroix and John McCall. 2019. Limitations of Benchmark Sets and Landscape Features for Algorithm Selection and Performance Prediction. In *Proc. of Genetic and Evolutionary Computation (GECCO'19)* (Prague, Czech Republic). ACM, New York, NY, USA, 261–262. https://doi.org/10.1145/3319619.3322051

Fu Xing Long, Diederick Vermetten, Anna V. Kononova, Roman Kalkreuth, Kaifeng Yang, Thomas Bäck, and Niki van Stein. 2023. Challenges of ELA-Guided Function Evolution Using Genetic Programming. In *Proceedings of the 15th International Joint Conference on Computational Intelligence, IJCCI 2023, Rome, Italy, November 13-15, 2023*, Niki van Stein, Francesco Marcelloni, H. K. Lam, Marie Cottrell, and Joaquim Filipe (Eds.). SCITEPRESS, 119–130. https://doi.org/10.5220/0012206200003595

Fu Xing Long, Diederick Vermetten, Bas van Stein, and Anna V. Kononova. 2022. BBOB Instance Analysis: Landscape Properties and Algorithm Performance across Problem Instances. *CoRR* abs/2211.16318 (2022). https://doi.org/10.48550/arXiv.2211.16318 arXiv:2211.16318

Manuel López-Ibáñez, Diederick Vermetten, Johann Dréo, and Carola Doerr. 2024. Using the Empirical Attainment Function for Analyzing Single-objective Black-box Optimization Algorithms. *CoRR* abs/2404.02031 (2024). https://doi.org/10.48550/ARXIV.2404.02031 arXiv:2404.02031

Olaf Mersmann, Bernd Bischl, Heike Trautmann, Mike Preuss, Claus Weihs, and Günter Rudolph. 2011. Exploratory landscape analysis. In *Proc. of Genetic and Evolutionary Computation (GECCO'11)*. ACM, 829–836.

Mario Andrés Muñoz, Michael Kirley, and Kate Smith-Miles. 2022. Analyzing randomness effects on the reliability of exploratory landscape analysis. *Natural Computing* 21, 2 (2022), 131–154.

Mario A. Muñoz and Kate Smith-Miles. 2020. Generating New Space-Filling Test Instances for Continuous Black-Box Optimization. *Evol. Comput.* 28, 3 (2020), 379–404. https://doi.org/10.1162/evco_a_00262

Mario A Muñoz, Yuan Sun, Michael Kirley, and Saman K Halgamuge. 2015. Algorithm selection for black-box continuous optimization problems: A survey on methods and challenges. *Information Sciences* 317 (2015), 224–245.

Ana Nikolikj, Carola Doerr, and Tome Eftimov. 2023. RF+ clust for Leave-One-Problem-Out Performance Prediction. In *Proc. of Applications of Evolutionary Computation (Evo Applications'23)*. Springer, 285–301.

PatrikValkovic. 2021. BBOBtorch. https://github.com/PatrikValkovic/BBOBtorch.

Maxim Pikalov and Vladimir Mironovich. 2022. Parameter tuning for the (1+(λ, λ)) genetic algorithm using landscape analysis and machine learning. In *International Conference on the Applications of Evolutionary Computation (Part of EvoStar)*. Springer, 704–720.

Michael JD Powell. 1994. *A direct search optimization method that models the objective and constraint functions by linear interpolation.* Springer.

Raphael Patrick Prager and Heike Trautmann. 2023. Nullifying the inherent bias of non-invariant exploratory landscape analysis features. In *International Conference on the Applications of Evolutionary Computation (Part of EvoStar)*. Springer, 411–425.

Raphael Patrick Prager and Heike Trautmann. 2024. Pflacco: Feature-based landscape analysis of continuous and constrained optimization problems in Python. *Evolutionary Computation* (2024), 1–6.

Jérémy Rapin and Olivier Teytaud. 2018. Nevergrad - A gradient-free optimization platform. https://GitHub.com/FacebookResearch/Nevergrad.

Quentin Renau, Johann Dreo, Carola Doerr, and Benjamin Doerr. 2019. Expressiveness and Robustness of Landscape Features. In *Proc. of Genetic and Evolutionary Computation (GECCO'19)* (Prague, Czech Republic). ACM, 2048–2051. https://doi.org/10.1145/3319619.3326913

Quentin Renau, Johann Dréo, Carola Doerr, and Benjamin Doerr. 2021. Towards explainable exploratory landscape analysis: extreme feature selection for classifying BBOB functions. In *Applications of Evolutionary Computation: 24th International Conference, EvoApplications 2021, Held as Part of EvoStar 2021, Virtual Event, April 7–9, 2021, Proceedings 24*. Springer, 17–33.

Urban Škvorc, Tome Eftimov, and Peter Korošec. 2020. Understanding the problem space in single-objective numerical optimization using exploratory landscape analysis. *Applied Soft Computing* 90 (2020), 106138.

Kate Smith-Miles and Mario Andrés Muñoz. 2023. Instance space analysis for algorithm testing: Methodology and software tools. *Comput. Surveys* 55, 12 (2023), 1–31.

Rainer Storn and Kenneth Price. 1997. Differential evolution-a simple and efficient heuristic for global optimization over continuous spaces. *Journal of global optimization* 11, 4 (1997), 341.

Ye Tian, Shichen Peng, Xingyi Zhang, Tobias Rodemann, Kay Chen Tan, and Yaochu Jin. 2020. A recommender system for metaheuristic algorithms for continuous optimization based on deep recurrent neural networks. *IEEE transactions on artificial intelligence* 1, 1 (2020), 5–18.

Diederick Vermetten, Fabio Caraffini, Anna V Kononova, and Thomas Bäck. 2023a. Modular Differential Evolution. In *Proc. of Genetic and Evolutionary Computation (GECCO'23)*. ACM. https://doi.org/10.1145/3583131.3590417 To appear. Code available at https://github.com/Dvermetten/ModDE.

Diederick Vermetten, Furong Ye, Thomas Bäck, and Carola Doerr. 2023c. MA-BBOB: Many-Affine Combinations of BBOB Functions for Evaluating AutoML Approaches in Noiseless Numerical Black-Box Optimization Contexts. In *International Conference on Automated Machine Learning, 12-15 November 2023, Hasso Plattner Institute, Potsdam, Germany (Proceedings of Machine Learning Research, Vol. 224)*, Aleksandra Faust, Roman Garnett, Colin White, Frank Hutter, and Jacob R. Gardner (Eds.). PMLR, 7/1–14. https://proceedings.mlr.press/v224/vermetten23a.html

Diederick Vermetten, Furong Ye, and Carola Doerr. 2023b. Using Affine Combinations of BBOB Problems for Performance Assessment. In *Proc. of Genetic and Evolutionary Computation Conference (GECCO'23)*, Vol. abs/2303.04573. ACM. https://doi.org/10.1145/3583131.3590412

Diederick Vermetten, Furong Ye, Carola Doerr, and Thomas Back. 2023d. Many-Affine BBOB Function Combinations autoML - Data and Figures. https://doi.org/10.5281/zenodo.7826036

Diederick Vermetten, Furong Ye, Carola Doerr, and Thomas Bäck. 2024. MA-BBOB - Reproducibility and Additional Data. https://doi.org/10.5281/zenodo.11547124

Hao Wang, Diederick Vermetten, Furong Ye, Carola Doerr, and Thomas Bäck. 2022. IOHanalyzer: Detailed Performance Analysis for Iterative Optimization Heuristic. *ACM Trans. Evol. Learn. Optim.* 2, 1 (2022), 3:1–3:29. https://doi.org/10.1145/3510426 IOHanalyzer is available at CRAN, on GitHub, and as web-based GUI, see https://iohprofiler.github.io/

IOHanalyzer/ for links.

Thomas Weise, Stefan Niemczyk, Hendrik Skubch, Roland Reichle, and Kurt Geihs. 2008. A tunable model for multi-objective, epistatic, rugged, and neutral fitness landscapes. In *Proceedings of the 10th annual conference on Genetic and evolutionary computation*. 795–802.