Assessing Ranking and Effectiveness of Evolutionary Algorithm Hyperparameters Using Global Sensitivity Analysis Methodologies

Varun Ojha*[1], Jon Timmis[2], and Giuseppe Nicosia[3,4]

[1] *University of Reading, Reading, United Kingdom*
[2] *The University of Sunderland, Sunderland, United Kingdom*
[3] *University of Catania, Catania, Italy*
[4] *University of Cambridge, Cambridge, United Kingdom*

Abstract

We present a comprehensive global sensitivity analysis of two single-objective and two multi-objective state-of-the-art global optimization evolutionary algorithms as an *algorithm configuration problem*. That is, we investigate the *quality of influence* hyperparameters have on the performance of algorithms in terms of their *direct effect* and *interaction effect* with other hyperparameters. Using three sensitivity analysis methods, Morris LHS, Morris, and Sobol, to systematically analyze tunable hyperparameters of covariance matrix adaptation evolutionary strategy, differential evolution, non-dominated sorting genetic algorithm III, and multi-objective evolutionary algorithm based on decomposition, the framework reveals the behaviors of hyperparameters to sampling methods and performance metrics. That is, it answers questions like what hyperparameters influence patterns, how they interact, how much they interact, and how much their direct influence is. Consequently, the ranking of hyperparameters suggests their order of tuning, and the pattern of influence reveals the *stability of the algorithms*.

Keywords: Hyperparameter optimization; evolutionary algorithms; global sensitivity analysis; algorithm design; algorithm stability analysis

1 Introduction

Optimization is at the core of advancement in machine learning and problem-solving. Effective optimization plays a vital role in solving problems, whether a single-objective or multi-objective problem. For example, be it a simple neural network or deep learning, or a simple linear or nonlinear function, optimizing the coefficients (e.g., weights of neural networks) is the most

*Corresponding Author: Varun Ojha, email: v.k.ojha@reading.ac.uk
preprint submitted to *Swarm and Evolutionary Computation*, Elsevier.

crucial aspect, which requires effective optimization algorithms. Evolutionary algorithms (EAs) are global optimization algorithms that iteratively guide a population towards a final population, solving various problems. EAs are widely used because of their agnostic nature to problems being solved (De Jong, 2016). However, their effectiveness relies on hyperparameters like population size and genetic operators (De Jong, 2007). Understanding of hyperparameters sensitivity to an algorithm's performance can be formulated as an *algorithm configuration problem* (ACP) (López-Ibáñez et al., 2016, Iommazzo et al., 2019), where informing optimal hyperparameter selection is essential for solving various tasks: neural networks (Crossley et al., 2013), deep learning (Taylor et al., 2021), and bio-inspired algorithms (Das et al., 2009, Ojha et al., 2014*a*).

Since hyperparameters tuning is crucial in achieving high-quality performance in solving optimization problems, methods such as manual tuning, grid search, and Bayesian search optimization are used. Bergstra and Bengio (2012) have shown the importance of *random search* instead of a *grid search* in sampling hyperparameter values. In addition, *manual tuning* without proper knowledge of hyperparameters can lead to too many trial-and-errors, and grid search and Bayesian search optimization are computationally expensive approaches that are often infeasible for such population-based optimization algorithms. Thus, Bergstra and Bengio (2012) suggest that tuning some hyperparameters is more necessary than the others. Hence, our objective in this research is to assess the ranking and effectiveness of hyperparameters of four well-known EAs: covariance matrix adaptation evolutionary strategy (CMA-ES) (Hansen and Ostermeier, 1996), differential evolution (DE) (Storn and Price, 1997), non-dominated sorting genetic algorithm III (NSGA-III) (Deb and Jain, 2013), and multi-objective evolutionary algorithm based on decomposition (MOEA/D) (Zhang and Li, 2007).

These are well-established algorithms that researchers have investigated and offered different versions to improve their performance. For example, a number of improvements on DE have been provided, including neighborhood-based mutation operator, fitness informed mutation strategy, and scaling factor in mutation for accelerating convergence by Das and Suganthan (2010), Islam et al. (2011), Das et al. (2009), Biswas et al. (2009), Das et al. (2007), and Das et al. (2005). Similarly, an improved step size (mutation) strategy for the CMA-ES algorithm is investigated by Voß et al. (2010), improved decomposition strategy like normal boundary intersection-style Tchebycheff approach, adaptive replacement strategies to assign a new solution to a sub-problem, and adaptive weight vector adjustment strategy for sub-problems, respectively proposed by Zhang et al. (2010), Wang et al. (2015), and Qi et al. (2014) for MOEA/D algorithm. Similarly, for NSGA-III performance enhancement, Cui et al. (2019) designed a selection-and-elimination operator to balance convergence and diversity of the population.

In this work, we develop a framework for comprehensive sensitivity analysis of hyperparameters of these algorithms using global sensitivity analysis methodologies: elementary effects (Morris, 1991) and variance-based sensitivity analysis (Sobol and Kucherenko, 2005). Using these

methodologies, we assess the effectiveness of EA hyperparameters. Such an analysis investigates a model's parameters (or an algorithm's hyperparameters) influence on its output (Iooss and Saltelli, 2016, Brooks et al., 2001), leading to the minimization of the number of critical tunable hyperparameters to improve a model's performance (Conca et al., 2015, Hill et al., 2016).

In our ACP framework, the performance of single-objective EAs was assessed as per *best solution*, while the performance of multi-objective EAs was assessed using three metrics: *generational distance* (Veldhuizen and Lamont, 1998, Veldhuizen, 1999), *inverse generational distance* (Deb and Jain, 2013), and *hyper-volume indicator index* (Zitzler and Thiele, 1998). To evaluate EAs, we use state-of-the-art optimization problems belonging to diverse families: for single-objective optimization, we use a set of 33 problems (Yao et al., 1999, Liang et al., 2013, 2014), and for multi-objective optimization, we use a set of 10 problems (Deb and Jain, 2013).

Our ACP framework assesses each algorithm on three sensitivity analysis methods: Morris Latin Hypercube sampling (Morris, 1991), Morris sampling (Morris, 1991), and Sobol (Sobol and Kucherenko, 2005). For each sample drawn from a hyperparameter search space, we ran each algorithm on 30 independent runs (for some, it was 10 times) and presented results using elementary effects and Sobol indices. These indices inform about (i) the direct effect and (ii) the interaction effect of a hyperparameter with other hyperparameters. Moreover, these two effects form a comparative matrix of low effect to high effect, where the diagonal from low direct and low interaction effects to high direct and high interaction effects shows the *order and ranking* of the hyperparameters. We ran algorithms on a sufficiently large sample set. These experiments were computationally expensive as they, in total, had 19 014 600 000 function evaluations. Computation of these sensitivity analysis indices is expensive, but they are a one-time effort, and once the ranking is determined, results are informative to researchers for further analysis and solving optimization problems. The source code and results are available at `https://github.com/vojha-code/SAofEAs`.

Our results reveal the pattern and behavior of hyperparameters to different sampling methods and matrices used to evaluate the performance of the algorithm. These patterns show how hyperparameters interact with one another or how the influence of one hyperparameter overwhelms the other. Moreover, results reveal how an algorithm is susceptible to its various hyperparameters and sampling methods, highlighting the stability of an algorithm. Consequently, these experiments rank the hyperparameter importance for an algorithm. For example, mutation type was found to have the strongest influence on the performance of DE, and results suggest the high importance of population size followed by the initial step size, crossover probability, and mode of decomposition, respectively in CMA-ES, NSGA-III, and MOEA/D.

Later in Section 2, we present related work. Then, Sections 3, 4, and 5 respectively describe algorithms, methodology, and experiments. The results are discussed in Section 6, followed by conclusions in Section 7. Appendixes A and B offer statistical tests and clustering analysis.

2 Related Work

Hyperparameter tuning is a crucial subject that has continuously been reported in the literature over the past decades (De Jong, 2007). This is because an appropriate hyperparameter setting is challenging since EA hyperparameters exhibit linear and non-linear effects (Lima and Lobo, 2004), meaning that they show various interactions among them (De Jong, 2007, Jansen et al., 2005, Hansen and Ostermeier, 2001). Abundant literature is available on EA hyperparameters tuning (Jansen et al., 2005, Lima and Lobo, 2004, Greco et al., 2019). The majority of which focuses on the *static* or *dynamic* setting of the hyperparameters (Eiben et al., 2007, Kramer, 2010, Iglesias et al., 2007). However, a systematic study of the EA hyperparameters influence is rare (Pinel et al., 2012), and it is largely attributed to the computationally expansive nature of EAs and empirical evaluation requirement for the tuning of their hyperparameters (Maturana et al., 2010). For example, a package *Irace* experimentally evaluates optimal hyperparameters for an optimization algorithm (López-Ibáñez et al., 2016). Therefore, De Jong (2007) posed questions like (i) what EA hyperparameters are useful for improving performance, and (ii) how do changes in a hyperparameter affect the performance of an EA?

Sensitivity analysis answers questions like *how uncertainty in each of the hyperparameters influences the uncertainty in the output of a model* (Saltelli et al., 2004). Hence, sensitivity analysis is useful in answering the questions of De Jong (2007). However, sensitivity analysis is a computationally expansive method since hyperparameters are sampled from a vast hyperparameter search space. Therefore, the sensitivity analysis of EAs has very high computational (time) as well as memory (space) overhead. This has resulted in very few reported works available in the literature, despite its advantages in suggesting a ranking of hyperparameters importance.

The dynamic tuning of hyperparameters requires hyperparameters to adapt during an EA run (Lou et al., 2021), while static tuning informs which hyperparameters to tune before EA run (Kramer, 2010). A systematic approach, like sensitivity analysis, is a static hyperparameter tuning approach. Paul et al. (2011) offered an introductory work on the usage of *local* and *global* sensitivity analysis. However, they used a simple test case, and they mainly performed a sensitivity analysis of EAs from a theoretical perspective. Pinel et al. (2012) performed a comprehensive sensitivity analysis of a parallel asynchronous cellular genetic algorithm on a scheduling problem. They comprehensively evaluated EAs population size, mutation probability, crossover probability, and other cellular genetic algorithm-related hyperparameters using the Fourier amplitude sensitivity test (Fast99) (Saltelli et al., 1999). Pinel et al. (2012) reported a ranking of hyperparameters on scheduling problem instances. On this scheduling problem instance, the crossover probability was ranked first, and in another instance, it was ranked third.

Our work takes an experimental approach to systematically analyze the importance of hyperparameters of state-of-the-art EAs on a testbench of state-of-the-art problems by applying Mor-

ris (Morris, 1991) and Sobol (Sobol and Kucherenko, 2005) sensitivity analysis methodologies. Our methodology comprises both single-objective and multi-objective EAs. Our framework offers a ranking of hyperparameters and insights into their effectiveness on EA performance. Our methodology is an *Algorithm Configuration Problem* (ACP) framework as defined by Iommazzo et al. (2019). This approach is contextually similar to the AutoML approaches (He et al., 2021), where the effort is to find the optimal configuration of algorithms and hyperparameters to solve machine learning tasks by automatic data preparation, feature engineering, hyperparameter optimization, and neural architecture search or even optimization of neural network components such as activation functions (Ojha et al., 2014*b*).

In fact, the ACP scope covers a wider range of methodologies and frameworks that seek to automate the design of algorithm configuration such as AutoMOEA (Bezerra et al., 2015), Auto Weka (Thornton et al., 2013), Auto-sklearn (Feurer et al., 2020), irace (López-Ibáñez et al., 2016), and others for machine learning hyperparameter optimization (Feurer and Hutter, 2019). The goal of these methodologies is to perform hyperparameter optimization and automatic design of new algorithms by assessing components and parameters that offer the best performance on a set of problem instances (Iommazzo et al., 2019, Thornton et al., 2013). The critical issue in such categorization is whether one would consider, for instance, a new evolutionary operator design in an EA framework as a new algorithms design or hyperparameter optimization? In our work, we consider such a scenario as hyperparameter optimization. Thus, our ACP framework seeks to inform the ranking effectiveness of hyperparameters for a set of EAs.

3 Evolutionary Algorithms

EAs are population-based evolution-inspired algorithms. EAs iteratively find solutions to a problem by applying evolutionary operators to candidate solutions. Selection, recombination, and mutation are among evolutionary operators applied to candidate solutions that generate new solutions in each generation. Such a process guides a sequence of generations from an initial population of candidate solutions to a final population. Four different EAs are investigated in this research: two single-objective and two multi-objective algorithms. Each of these EAs has its own version of evolutionary operators. This section briefly describes each of these EAs and their performance measure metrics.

3.1 Single-objective Evolutionary Algorithms

A single-objective optimization (SOO) algorithm (single solution-based or population-based) *minimizes an objective function* (a cost function or a problem) as

$$\begin{aligned} f: \quad & \mathbb{R}^n \to \mathbb{R} \\ & \mathbf{x} \mapsto f(\mathbf{x}), \end{aligned} \tag{1}$$

where $\mathbf{x} \in \mathbb{R}^n$ is a candidate solution (a search point in a solution space X), and we want $f(\mathbf{x})$ to be as minimum as possible. An SOO algorithm converges to a solution $\mathbf{x}^*$ such that $f(\mathbf{x}^*) \leq f(\mathbf{x}),\ \forall \mathbf{x} \in X$. The solution $\mathbf{x}^*$, therefore, is a *global minimum* (global optimum). However, if for $f(\mathbf{x}^*) \leq f(\mathbf{x})$ there exists some $\delta > 0$ such that $|\mathbf{x} - \mathbf{x}^*| \leq \delta$ for any $\mathbf{x} \in X$, then the solution $\mathbf{x}^*$ is a *local minimum* (near-optimum).

We study two population-based single-objective global optimization algorithms: CMA-ES (Hansen and Ostermeier, 1996) and DE (Storn and Price, 1997). The basic steps and operators of CMA-ES and DE are as follows.

3.1.1 Covariance Matrix Adaptation Evolution Strategies (CMA-ES)

CMA-ES is a population-based *evolutionary strategy* optimization algorithm (Hansen and Ostermeier, 1996). CMA-ES algorithm generates new candidate solutions during its search by sampling solutions from a *multivariate normal distribution*, $\mathcal{N}(\mathbf{m}, C)$ uniquely determined by its mean $\mathbf{m} \in \mathbb{R}^n$ and its symmetric positive definite covariance matrix $C \in \mathbb{R}^{n \times n}$. The initial population of λ candidate solutions at generation $g = 0$ are sampled as

$$\mathbf{x}_k^g \sim \mathbf{m}^g + \sigma^g \mathcal{N}(\mathbf{0}, C^g) \quad \text{for } k = 1, \ldots, \lambda, \tag{2}$$

where $\mathcal{N}(\mathbf{0}, C)$ is a multivariate normal distribution with zero mean and covariance matrix $C^g \in \mathbf{I}$, and $\sigma^g \in \mathbb{R}_{>0}$ is an initial step size.

For generation $g = 1, 2, \ldots$, multivariate normal distribution $\mathcal{N}(\mathbf{m}, C^{g+1})$ is generated (updated) with mean $\mathbf{m} \in \mathbb{R}^n$ and covariance matrix $C \in \mathbb{R}^{n \times n}$ updated with scalar factor $\sigma^g \in \mathbb{R}_{>0}$. Selection and recombination operations in CMA-ES are equivalent to computing moving mean m^{g+1}, a weighted average of selected points λ_{ratio} from generation g. Adding a random vector with zero-mean acts as a mutation in CMA-ES during the offspring generation step. The steps size control and covariance matrix adaptation (learning rate α_μ) are additional two necessary steps in a generation of CMA-ES (Hansen and Ostermeier, 1996).

3.1.2 Differential Evolution (DE)

DE is a *gradient-free* EA, originally proposed by Storn and Price (1997). DE iteratively searches for a solution. For an initial population $X = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_\lambda]$ of size λ, DE repeats its steps *selection*, *mutation*, and *recombination* until an optimum solution vector $\mathbf{x}^*$ is obtained, or until a maximum iteration is reached. At each generation $g = 1, 2, \ldots$, DE randomly selects three distinct candidate solutions $\mathbf{x}_{r1}^g$, $\mathbf{x}_{r2}^g$, and $\mathbf{x}_{r3}^g$ from X such that $\mathbf{x}_{r1}^g \neq \mathbf{x}_{r2}^g \neq \mathbf{x}_{r3}^g$. The selection of a base vector $\mathbf{x}_{r1}^g$ plays a crucial in DE.

A mutation operation is performed on a base vector $\mathbf{x}_{r1}^g$ to generate a donor vector $\mathbf{v}^{g+1}$, which is generated using a mutation method $\mathbf{b}_{\text{type}}$, a difference vector $(\mathbf{x}_{r2}^g \mathbf{x}_{r3}^g)$, and acceleration

coefficient β. A mutation method $\mathbf{b}_{\text{type}}$ = "DE/rand/1" or similar mutation is performed as

$$\mathbf{v}^{g+1} = \mathbf{x}^g_{r1} + \beta(\mathbf{x}^g_{r2}\mathbf{x}^g_{r3}). \tag{3}$$

A crossover operation using a crossover method {bin, exp} is performed to generate a trial vector $\mathbf{u}^{g+1}$ which takes its elements from a donor vector $\mathbf{v}^{g+1}$ using a crossover probability P[X]. If the fitness $f(\mathbf{u}^{g+1})$ is better than the target vector $f(\mathbf{x}^{g+1}_t)$, then the trial vector $\mathbf{u}^{g+1}$ replaces the target vector $\mathbf{x}^{g+1}_t$.

3.2 Multi-objective Evolutionary Algorithms

A multi-objective optimization (MOO) algorithm minimizes two or more objective functions *simultaneously* as

$$F(\mathbf{x}) \equiv (f_1(\mathbf{x}), \ldots, f_k(\mathbf{x})), \text{ i.e., } F : \mathbb{R}^n \to \mathbb{R}^k \text{ for } k \geq 2 \tag{4}$$

such that no one objective of the problem can be improved without a simultaneous detriment to at least one of the other objectives. Each $f_l(\mathbf{x}), l = 1, 2, \ldots, k$ is a scalar objective, and MOO optimizes the objective vector $F(\mathbf{x})$ where $\mathbf{x} \in \mathbb{R}^n$ is its feasible solution. More specifically, a MOO algorithm produces a set of non-dominated solutions $\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_{\lambda'}\}$, also known as the Pareto-optimal solutions set (Deb, Pratap, Agarwal and Meyarivan, 2002).

A solution $\mathbf{x}_i$ dominates other solution $\mathbf{x}_j$ if for $j = 1, 2, \ldots, \lambda, i \neq j$, and for all objectives $l = 1, 2, \ldots, k$, $f_l(\mathbf{x}_i) \preccurlyeq f_l(\mathbf{x}_j)$ holds, where $\preccurlyeq$ should be read as "better off." On the contrary, a solution $\mathbf{x}_i$ is non-dominated if, for at least one objective l, $f_l(\mathbf{x}_i) \preccurlyeq f_l(\mathbf{x}_j)$ does not hold. For each $\mathbf{x}_i$, a set of such non-dominated solutions are called a Pareto-optimal set of solutions.

In this paper, we study the population-based multi-objective global optimization algorithms NSGA-III (Deb and Jain, 2013) and MOEA/D (Zhang and Li, 2007) and investigate their algorithmic hyperparameter setting in obtaining a better Pareto-optimal set of solutions.

3.2.1 Non-Dominated Sorting Genetic Algorithm–III (NSGA-III)

NSGA-III is a population-based MOO algorithm (Deb and Jain, 2013). NSGA-III uses fast non-dominated sorting and niching operations to guide an initial population X of size λ candidate solutions through a predefined number of generations to a final population while simultaneously optimizing trade-offs of multiple objectives. In each step of NSGA-III, crossover, mutation, and non-dominated sorting is performed.

The fast non-dominated sorting sorts the λ candidate solutions into several sets (called Fronts) of non-dominated solutions: $F_1, F_2, \ldots, F_s$ such that the Front F_1 contains all the non-dominated candidate solutions of population X. That is, no one solution in F_1 is dominated by any other

solutions. From all the remaining solutions (i.e., except the ones already in F_1), a new Front F_2 that contains all the next non-dominated solutions of X is determined. Similarly, Front F_3 and other Fronts are subsequently obtained using non-dominated sorting. Thus, it is possible to assign a rank to the candidate solutions such that those in the Front F_1 have rank 1, solutions in Front F_2 have rank 2, and so on.

NSGA-III performs *niching* as its selection operation on non-dominated sorting solutions. Niching takes advantage of a predefined set of reference points placed on a normalized hyperplane of a k-dimensional objective-space (Das and Dennis, 1998), where each individual $\mathbf{x} \in X$ in the population is associated with reference points (Deb and Jain, 2013). The total number of reference points depends on the predefined number of divisions associated with each objective axis. NSGA-III repeats its operations selection, crossover, mutation, and recombination until a maximum iteration or a termination condition is reached. The performance of NSGA-III is measured in terms of the quality of solutions it produces in its iteration and in the final population.

3.2.2 Multi-objective Evolutionary Algorithm based on Decomposition (MOEA/D)

MOEA/D solves a MOO problem by decomposing the MOO problem into many single (scalar) objective sub-problems (Zhang and Li, 2007). Tchebycheff approach (Miettinen, 2012) or normal boundary interaction approach (Das and Dennis, 1998) are typically used approaches for decomposing MOO problem into (say) N scalar sub-problems. A uniform spread of N weight vectors $\{\mathbf{w}_1, \ldots, \mathbf{w}_N\}$ and reference point $\mathbf{z}^* = (z_j^1, \ldots, z_j^k) = \min\{f_i(\mathbf{x}) | \mathbf{x} \in X\}$, for $i = 1, \ldots, k$ is used for computing $j = 1, 2, \ldots, N$ scalar objectives $y^{te}(\mathbf{x}|\mathbf{w}_j)$.

The scalar objective in Tchebycheff decomposition method is $y^{te}(\mathbf{x}|\mathbf{w}_j) = \max_{1 \leq i \leq k}\{\mathbf{w}_j^i | f_i(\mathbf{x}) - \mathbf{z}^*\})$, where weight vector $\mathbf{w}_j = (w_j^1, \ldots, w_j^k)$. The optimal solution of $y^{te}(\mathbf{x}|\mathbf{w}_i)$ for weight vector $\mathbf{w}_i$ should be close to a solution $y^{te}(\mathbf{x}|\mathbf{w}_j)$ for weight vector $\mathbf{w}_j$. Hence, in MOEA/D, a neighborhood of weight vector $\mathbf{w}_i$ is defined with many closest points in $\{\mathbf{w}_1, \ldots, \mathbf{w}_N\}$. The neighborhood may play a vital role in MOEA/D.

Moreover, each objective is optimized as a single (scalar) objective problem. That is, ith objective is optimized such that it minimizes its distance from a reference point on a $k-$objective space. Thus, all decomposed sub-problems move towards the reference point $\mathbf{z}^*$. MOEA/D maintains T closest solution vectors (Neighbor) for each candidate solution in successive steps. In each iteration, MOEA/D generates a new solution by selecting two solution vectors using genetic operators and evaluating them in order to update their neighborhood and the best solution $\mathbf{x}^*$. The details of the MOEA/D algorithm are available in (Zhang and Li, 2007).

3.3 Performance Metrics

3.3.1 Single Objective Metrics

A population-based EA applied to solve a single-objective problem offers the best solution in its final population. The best solution, $\mathbf{x}^*$, is the one that has the lowest $f(\mathbf{x})$ value among all solutions of all generations of a single-objective EA. Hence, the *Best Solution* obtained in fewer generations in a lesser wall clock time measures the quality of a single-objective EA.

3.3.2 Multi-Objective Metrics

Multi-objective EAs applied to a MOO problem typically offer a set of solutions that satisfy trade-offs between the objectives. This set of solutions is non-dominated solutions which are also known as a Pareto-front. A multi-objective EA, therefore, guides a population of candidate solutions from *current Pareto-front* $\mathbf{A}$ toward a *true Pareto-front* $\mathbf{Z}$. In such a setting, three indicators are used to measure and compare the performance of EAs on MOO problems: generational distance, inverse generational distance, and hyper-volume indicator (Fig. 1):

Generational Distance (GD). Generational distance GD_i at an iteration, i, measures the generational distance between *current Pareto-front* and *true Pareto-front* of a multi-objective problem (Veldhuizen and Lamont, 1998, Veldhuizen, 1999). Generational distance GD_i is a measure of error between *current Pareto-front* and *true Pareto-front* as

$$\text{GD}_i \triangleq \frac{\sqrt{\sum_{i=1} d_i^2}}{n}, \tag{5}$$

where d_i^2 is the distance of ith solution in *current Pareto-front* $\mathbf{A}$ with *true Pareto-front* $\mathbf{Z}$ (Veldhuizen, 1999) and GD is typically average of n solutions (Fig. 1). Hence, GD is a minimization metric where a low value indicates a better solution.

Inverse Generational Distance (IGD). Inverse generational point provides combined information on the solutions' diversity and convergence quality. It makes use of a set of target reference points in k-dimensional objective space. Like GD, IGD compares solutions in the *current Pareto-front* $\mathbf{A}$ with *true Pareto-front* $\mathbf{Z}$. However, IGD uses a single reference and computes the average Euclidean distance between all solutions that are nearest to the target reference points (Deb and Jain, 2013) as

$$IGD(\mathbf{A}, \mathbf{Z}) \triangleq \frac{1}{|\mathbf{Z}|} \sum_{i=1}^{|\mathbf{Z}|} \min_{j=1}^{|A|} d(\mathbf{z}_i, \mathbf{a}_j), \tag{6}$$

where $d(\mathbf{z}_i, \mathbf{a}_j) = ||\mathbf{z}_i, \mathbf{a}_j||_2$. Similar to GD, IGD is a measure of error between the *current Pareto-front* and *true Pareto-front*. Hence, lower values of IGD indicate a better solution.

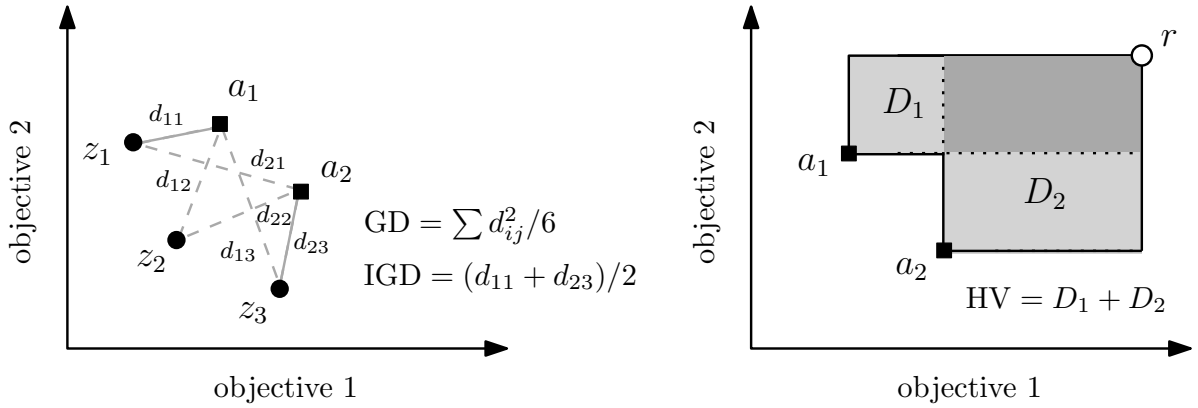

Fig. 1: Example of a 2D objective space and computation of GD, IGD, and HV metrics. Current Pareto front is $\mathbf{A} = \{a_1, a_2\}$ and true Pareto front is $\mathbf{Z} = \{z_1, z_2, z_2\}$, and optimum of two objective is a reference point r. The distance between two points is d_{ij}, and the area framed by a point with the reference point, r, is the area D_i.

Hypervolume Indicator (HV). Hyper-volume indicator, HV measures the dominance of Pareto-front solutions on a geometric space (e.g., area for a 2D objective space) framed by the k-dimensional objective-space with respect to a positive semi-axle r (see Fig. 1). Hence, HV measures the quality Pareto-optimal solutions set (Fonseca et al., 2006), and it is an indicator of the quality of the solutions obtained by two algorithms with respect to the same reference frame. The goal is to maximize the hyper-volume indicator index HV. A greater value indicates that the algorithm's overall performance is better with respect to another algorithm associated with a smaller hyper-volume value. Moreover, the greatest contributing point in a hyper-volume indicator analysis is the point that covers the largest area, and that can be considered as the best solution (Zitzler et al., 2003).

4 Global Sensitivity Analysis

The goal of the *sensitivity analysis* is to study how the uncertainty of a model's output depends on the uncertainty of its inputs (Saltelli, 2002, Saltelli et al., 2008). The *elementary effects* analysis, known as a "Morris method" (Morris, 1991) and *variance-based sensitivity analysis*, known as a "Sobol method" (Sobol and Kucherenko, 2005), are used in this research for the global sensitivity analysis of the hyperparameters of four EAs. This framework of combining sensitivity analysis and EAs is an *algorithms configuration problem* that aims to inform algorithm performance to variations in hyperparameter on problem instances (Iommazzo et al., 2019).

4.1 Elementary Effects

The *elementary effects* (EE) technique, known as a "Morris method" as it was originally introduced by Morris (1991), is an effective way to analyze the effects (sensitivity) of input variables on the outputs of a model or a system. In our case, the Morris method assesses the EE of

the algorithmic hyperparameters on the performances of an EA. This is useful in analyzing the sensitivity of EA hyperparameters as the Morris method determines whether the effects of a hyperparameter on a model's outputs (EA performances on functions) are (a) insignificant and negligible, (b) linearly correlated, or (c) non-linearly correlated or involved in an interaction with other hyperparameters (Saltelli et al., 2008).

We briefly introduce the computation of EE as follows. Let us have $Y = f(\mathbf{X})$ or simply $Y(\mathbf{X})$ be the output of a model $f(\cdot)$ (an algorithm) that takes k hyperparameters $\mathbf{X} = \{X_1, X_2, \ldots, X_k\}$ from a hyperparameter space Ω of the *p-level* grid. Then we compute the *elementary effect* EE_i of ith hyperparameter X_i as

$$\mathrm{EE}_i = \frac{Y(X_1, \ldots, X_{i-1}, X_i + \Delta, \ldots, X_k) - Y(X_1, \ldots, X_{i-1}, X_i, \ldots, X_n)}{\Delta}, \tag{7}$$

where Δ is a value in $\left(\frac{1}{p-1}, \ldots, 1 - \frac{1}{p-1}\right)$ which is an incremental change in the values of hyperparameter X_i when X_i is sampled from *p-level* grid hyperparameter space Ω. In this scenario, for k hyperparameters and p discrete levels, $\Delta = p/2(p-1)$ indicates distance (length) between two levels in the hyperspace Ω along ith axis. The total points in the hyperparameter space Ω, therefore, are $p^{k-1}[p - \Delta(p-1)]$ grid points, which increase exponentially as the number of hyperparameters k increases. However, we use a *one-at-a-time* (OAT) sampling technique for generating r sample points from this space to compute r EEs for each hyperparameter.

In the OAT sampling technique, hyperparameter X_i value is changed from a grid point $X_i^{(j)}$ to the adjacent grid point $X_i^{(j\pm1)}$ by a length of Δ while all other hyperparameters (say $X_{\sim i}$) remain as it is. Then the next hyperparameter X_{i+1} is chosen, whose value is changed while others remain fixed. This way of sampling is a *uniform, non-repeating random walk* through the grid of hyperspace Ω (we call it *Morris* (Morris, 1991)). Another way of sampling points (a set $\mathbf{X}$ of hyperparameters) from the hyperspace Ω is to use the Latin Hypercube Sample (LHS) based Morris method (*Morris LHS*) (Campolongo et al., 2007), which is a stratified sampling approach to cover all region of the hyperspace Ω. Here, we typically select r sample points for each hyperparameter X_i. Hence, both OAT-based Morris LHS and Morris sampling methods give us $r(k+1)$ sample points.

We measure two indices μ_i and σ_i indicate *mean* (central tendency) and *standard deviation* of EE_i of ith hyperparameter X_i. The measure

$$\mu_i = \frac{1}{r}\sum_{j=1}^{r} EE_i^j \tag{8}$$

indicates the overall influence of a hyperparameter X_i where a larger measure of μ_i means a larger *overall* individual ability to influence the outputs of an algorithm. We also measure the

standard deviation σ_i of EE_i as

$$\sigma_i = \sqrt{\frac{1}{r-1}\sum_{j=1}^{r}(EE_i^j - \mu_i)^2}, \tag{9}$$

where a large measure of σ_i indicates that a hyperparameter has high interaction with other hyperparameters. The measure σ is an ensemble influence. That is, if σ_i has a high value, it means that the computed r elementary effects EE_i^r of ith hyperparameter X_i varied a lot because of the variation in the values of other hyperparameters as well. Whereas a low value of σ_i means small differences in the computed r elementary effects EE_i^r of the ith hyperparameter X_i. This indicates that the influence of a hyperparameter on a model's output is independent of the choice of other hyperparameters values. However, to understand the influence of a hyperparameter, both μ and σ measures need to be seen together (see Fig. 2). We normalized the values of μ_i and σ_i between 0 and 1 to effectively show results as per Fig. 2.

4.2 Variance-Based Sensitivity Analysis

The *variance-based sensitivity analysis* is known as the "Sobol method" (Sobol and Kucherenko, 2005), and it shows how much variance of a model's output depends on its inputs. It is an in-depth sensitivity analysis method that uses two sensitivity indices: (a) *first-order effect* S_i to indicate a direct effect of a hyperparameter X_i on a model's output $Y = f(\mathbf{X})$ and (b) *total effect* ST_i to indicate a hyperparameter X_i interaction with its complementary parameters $X_{\sim i}$.

The direct effect S_i, irrespective of the hyperparameter interaction ST_i, indicates that, on average, how much model's variance $V[Y(\mathbf{X})]$ could be reduced if the hyperparameter X_i is fixed to a value. Meaning a low value of S_i shows that the variance of the model's output $Y(\mathbf{X}|X_i = x_i^*)$ does not depend on X_i, and fixing X_i to a value does not have much impact on the model's output, while for a high value of S_i, it strongly does. Indeed, a low value of S_i indicates that ith hyperparameter's influence is negligible. Similarly, the interaction effect or total effect $ST_i = 0$ indicates that the model's output $Y(\mathbf{X}|X_i)$ does not depend on X_i, and it is a non-influential parameter. The large values of interaction effect or total-effect ST_i show proportionally strong interactions between the hyperparameter X_i and its complementary parameter $X_{\sim i}$. The difference $ST_i - S_i \geq 0$, i.e., total interaction influence minus direct influence, shows how much ith hyperparameter is involved in interaction with other hyperparameters. We normalized the values of S_i and ST_i between 0 and 1 for lucid interpretation of their influence (see Fig. 2).

The first-order effect S_i and total effect ST_i of Sobol method are computed as

$$S_i = \frac{V(E(Y|X_i))}{V(Y)} = \frac{y_A \cdot y_{C_i} - f_0^2}{y_A \cdot y_A - f_0^2} = \frac{\frac{1}{N}\sum_{j=1}^{N} y_A^j y_{C_i}^j - f_0^2}{\frac{1}{N}\sum_{j=1}^{N}(y_A^j)^2 - f_0^2} \tag{10}$$

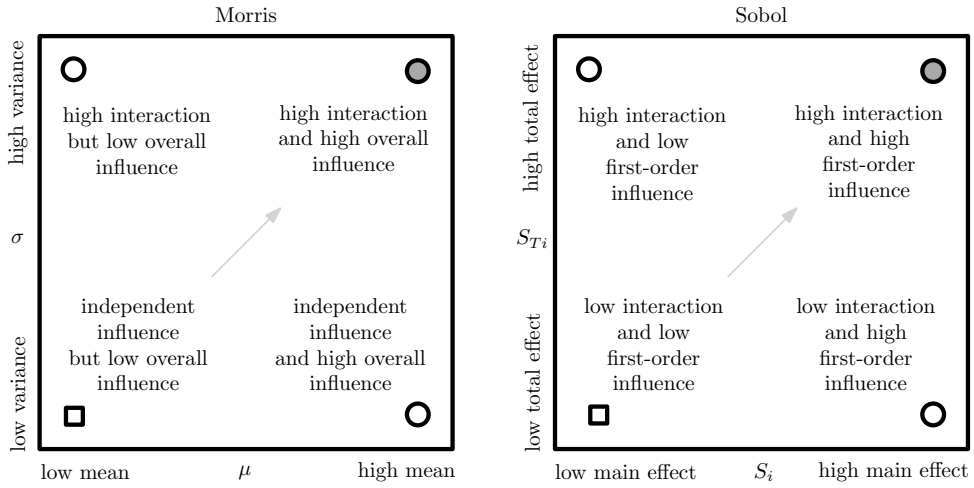

Fig. 2: Morris (*left*) and Sobol (*right*) indices interpretation. Top right corner circle in dark gray is the ideal case where a hyperparameter has high individual influence and high interaction (or total effect). Circles in white at the top left and bottom right corners are cases that have high importance in at least one direction. Bottom left square in white shows the least ideal case where hyperparameters are non-influential, and fixing them at any values within their defined domain will not influence the algorithm's performance. Arrow along the diagonal direction indicates the order of the hyperparameters' importance and influence.

and

$$ST_i = 1 - \frac{V(E(Y|X_{\sim i}))}{V(Y)} = 1 - \frac{y_B \cdot y_{C_i} - f_0^2}{y_A \cdot y_A - f_0^2} = 1 - \frac{\frac{1}{N}\sum_{j=1}^{N} y_B^j y_{C_i}^j - f_0^2}{\frac{1}{N}\sum_{j=1}^{N} (y_A^j)^2 - f_0^2}, \tag{11}$$

where N is the number of random samples, $y_A = f(A)$, $y_B = f(B)$ and $y_{C_i} = f(C_i)$ are model output vectors on sample matrix A, B and C_i respectively; and the estimated mean f_0^2 is

$$f_0^2 = \left(\frac{1}{N} \sum_{j=1}^{N} y_A^j \right)^2 . \tag{12}$$

Matrices $A_{N\times k}$ and $B_{N\times(k-2k)}$ are random sample points (hyperparameter values), and each matrix C_i is formed by taking all columns of matrix B except ith column, which is taken from ith column of matrix A. Such a sampling is similar to OAT sampling, except its rows are not sorted in any specific order, and all elements in a row differ from the other elements in the row.

5 Experiments

Our sensitivity analysis framework has four essential structural components:

1. setup of EAs tunable hyperparameters and optimization problems
2. sampling of hyperparameters from hyperparameter space of respective sensitivity analysis methods for respective algorithms

3. evaluation of EAs on optimization problem (testbench) for all sampled hyperparameter points and for each hyperparameter sample, the evaluation of respective EAs over a number of independent instances to obtain stable results and to observe expected (average-case) performance of algorithms over performance measures

4. computation of Morris and Sobol indices

In the experiment, all EAs start with a population of initial candidate solutions (uniformly randomly drawn from $\mathbb{R}^n$, n being dimensionality of the problem). Other commonalities among EAs are evolutionary operators like "selection," "mutation," and/or "crossover" for generating new (offspring) population and their evaluation. EAs repeat this process for a number of generations until a *termination condition* is met. We set the *termination condition* to be the desired number of *function evaluations*, and we set this to a value of 10 000 for all four algorithms for all problems. The other hyperparameters setting for our experiments were as follows:

5.1 Single-Objective Algorithm Hyperparameters

We analyzed two single-objective EAs over 33 optimization problems: 23 problems from testbench introduced in (Yao et al., 1999), and we took 10 optimization problems regarding shifted problems from CEC2014 (shifted Sphere, Ellipsoid, Ackley, and Griewank; and shifted and rotated Rosenbrock, and Rastrigin) and CEC 2015 (shifted and rotated Weierstrass, Schwefel, Katsuura, HappyCat) (Liang et al., 2006, 2013, 2014). An EA needs to find a single optimal solution for an SOO problem in a few generations at the expanse of some wall-clock time. Hence, the *Best Solution* was used for SOO evaluation. Table 1 lists the hyperparameter tuning space of CMA-ES (Hansen and Ostermeier, 1996) and DE (Storn and Price, 1997) algorithms.

The sensitivity analysis method setup for single-objective optimization was as follows. We used $p = 10$ grid levels to form the hyperparameter space Ω for respective single objective EAs. From this hyperparameter space, we select $r = 50$ sample points for each hyperparameter of CMA-ES and DE in the cases of *Morris LHS* and *Morris* methods (see Equations (8) and (9)). This gave us 300 and 400 sample points in total for CMA-ES and DE algorithms, respectively. The Sobol analysis is $2 + k$ times more expensive than Morris methods since it evaluates hyperparameter matrices A, B, and C_i, $i = 1, 2, \ldots, k$. For Sobol, we use $N = 100$, which gave us 700 and 900 sample points in total for CMA-ES and DE algorithms, respectively.

5.2 Multi-objective Algorithm Hyperparameters

We analyzed multi-objective EAs over a testbench consisting of four families of optimization problems: (i) DTLZ1, DTLZ2, DTLZ3, and DTLZ4 (Deb, Thiele, Laumanns and Zitzler, 2002); (ii) IDTLZ1 and IDTLZ2 (Deb, Thiele, Laumanns and Zitzler, 2002); (iii) CDTLZ2 (Deb and Jain, 2013); and (iv) WFG3, WFG6, and WFG7 (Huband et al., 2006). EAs were evaluated

Table 1: Hyperparameter domain range of CMA-ES (Hansen and Ostermeier, 1996) and DE (Storn and Price, 1997). For both algorithms, the termination condition was $10,000$ function evaluations.

Algo	Params	Domain	Description
CMA-ES	λ	$[10, 1000]$	Population size
	α_μ	$[0, 4]$	Learning rate
	σ_0	$[0.1, 2]$	Initial step size
	$\sigma_{0-scale}$	{False, True}	Re-scaling of σ_0: convergence speed controller
	$\mu\lambda_{\text{ratio}}$	$[0.1, 1]$	Percentage of population's elements usage in co-variance matrix estimation and update
DE	λ	$[10, 1000]$	Population size
	X	{bin, exp}	Crossover methods: Binomial and Exponential
	P[X]	$[0, 1]$	Crossover probability
	$\beta_{\min}$	$[0, 1]$	Minimum Acceleration coefficient
	$\beta_{\max}$	$[0, 2]$	Maximum Acceleration coefficient, $\beta_{\max} = \beta_{\min} + \beta_{\max}$
	$\mathbf{b}_{\text{type}}$	{"best," "target-to-best," "rand-to-best," "rand"}	Base vector selection methods (mutation type or DE algorithm version)
	$\mathbf{b}\lambda_{\text{ratio}}$	$[0.01, 0.5]$	Percentage of base vectors (solution) to be used for difference vectors computation

and analyzed for each listed MOO problem for 3 objectives, and each problem was solved as a 10-dimensional problem. This setting was chosen based on the computation effort required for these MOO problems.

Since the goal of the multi-objective EAs is to obtain a set of solutions where no one objective dominants over the other objectives (Deb, Pratap, Agarwal and Meyarivan, 2002, Zhang and Li, 2007), we use GD (minimization), IGD (minimization), and HV (maximization) as the measures of EA performances (see Section 3.3.2). These metrics result in higher values for large population size λ compared to a small population size λ. Thus, for *population-fair* performance analysis, the metrics were calculated from a union of populations of all generations of EAs and from not only the population of the last generation of the EAs. Moreover, the values were averaged over 30 independent runs for each sampled set of hyperparameters.

NSGA-III and MOEA/D have a few common tunable hyperparameters in addition to their subjective tunable hyperparameters. Table 2 shows the domain setting of these common and subjective tunable hyperparameters of NSGA-III and MOEA/D.

The sensitivity analysis method setup for multi-objective optimization was as follows. We used $p = 10$ grid levels to form the hyperparameter space Ω for respective single objective EAs. From this hyperparameter space, we select $r = 20$ sample points for each hyperparameter of CMA-ES and DE in the cases of *Morris LHS* and *Morris* methods (see Equations (8) and (9)). This gave us 140 and 160 sample points in total for NSGA-III and MOEA/D algorithms, respectively. In the Sobol analysis, we used $N = 30$, and this gave us 240 and 270 sample points in total for NSGA-III and MOEA/D algorithms, respectively, for their matrices A and B from which C_i

Table 2: Hyperparameter domain range of NSGA-III and MOEA/D and their shared (*Common*) hyperparameters domain. For both algorithms, the termination condition was $10,000$ function evaluations.

Algo	Params	Domain	Description
Common	λ	$[10, 1000]$	Population size.
	P[X]	$[0, 1]$	Simulated binary crossover (SBX) probability
	X_{DI}	$[1, 200]$	SBX distribution index
	P[PM]	$[0, 1]$	Polynomial mutation (PM) probability
	PM_{DI}	$[1, 200]$	PM distribution index
NSGA-III	K	$[2, 10]$	Tournament size
	Selection	Tournament	Parents selection for offspring generation
MOEA/D	*Mode*	{"penalty based boundary intersection (PBI)," "Tchebycheff," "Tchebycheff with normalization," "modified Tchebycheff"}	Method for MOO decomposition into many SOO subproblems
	ϵ_N	$[0.05, 0.5]$	Neighbors: percentage of the population considered as neighbors for each sub-problem generation

matrices were created. The number of sampling points in this work is sufficiently large for good sensitivity analysis (Campolongo et al., 2007, Saltelli, 2002, Saltelli et al., 2008).

All algorithms, methods, and sensitivity analysis experiments were performed in MATLAB, and implementations of individual components were taken from MATLAB libraries. We used *safe toolbox* (Pianosi et al., 2015) for the implementation of sensitivity analysis sampling methods, indices calculations, and workflows. Single objective algorithms were implemented using ypea library (Heris, 2019), and the implementation of multi-objective optimization problems and evaluation measure metrics related to optimization algorithms were used from PlatEMO library (Tian et al., 2017). The entire workflow framework was synchronized with the help of inbuilt functions of MATLAB.

The whole experiment was expensive to run since the total number of function evaluations was 19 014 600 000. The breakdown of this function evaluation was as follows (each multiplied by 10 000 concerning termination condition): DE, 858 580; CMA-ES, 720 080; MOEA/D, 171 600; and NSGA-III, 151 200. For DE and CMA-ES, there were 33 objective functions, and each one was run at least 10 times for each combination of hyperparameter settings. Similarly, for MOEA/D and NSGA-III, there were 10 functions, and each was run 30 times for stable results for each set. The hyperparameter sets were sampled in three different ways for all algorithms: Morris LHS, Morris, and Sobol, as mentioned in Sections 5.1 and 5.2. Our implementation of this framework for sensitivity analysis of EAs and results are available in (Ojha et al., 2022).

6 Results and Discussion

The results of sensitivity analysis of each algorithm for their performances on testbench were collected and processed to produce three indicators: (i) sensitivity analysis indices matrix as per Fig. 2, (ii) ordered bar plot arranged from low to high normalized sensitivity analysis total indices values, and (iii) mean score (average performance) of each hyperparameter over select performance measures. Additionally, statistical test and clustering analysis of the results are presented in Sections A and B. This section describes hyperparameter *influence*, *ranking*, and *quality* through these three indicators.

Each sensitivity analysis method varies how they sample hyperparameter sets as they use strategies such as LHS, OAT based uniform random walk, and OAT based uniform sampling. Morris LHS and Sobol use the LHS strategy, which means they stratified the hyperspace to draw samples to cover most of the sample space. Morris uses uniform random walk sampling. In summary, each method may present its own ordering of hyperparameters that influence ranking and interpretation. Hence, we are also interested in the commonality of results among methods.

6.1 Single-Objective EAs

6.1.1 CMA-ES Analysis

CMA-ES results are shown in Figs. 3, 4, and 5, where Fig. 3 is a scatter plot that presents sensitivity analysis indices as per Fig. 2. It shows the tendency of the *quality of influence* a hyperparameter has on CMA-ES performance on all 33 problems in the testbench. For instance, λ, the population size in CMA-ES has a high *overall* influence and high *interaction* influence in all three sensitivity analysis methods. Hence, λ is the most significant hyperparameter of the CMA-ES algorithm, and this must be the first hyperparameter one must select to tune for the performance improvement when CMA-ES is applied to solve a problem.

Population size λ. Population size λ is the most influential factor in CMA-ES algorithms. Both Morris and Sobol methods show a strong overall influence and high interaction for λ. Morris LHS ranked it as a high direct influence but slightly lower interaction influence than covariance matrix size controller $\mu\lambda_{\text{ratio}}$ that has the highest interaction and direct influence in Morris LHS method. Since MOEA/D decomposes problems into several single-objective problems, unsurprisingly, the size of the population and related hyperparameters are the most influential. This corroborates the fact that they offer exploration capabilities to population-based algorithms, which allow them to concurrently search a huge part of the search space. Fig. 4 and Fig. 5 confirm the significance of λ in CMA-ES. Figure 5 also suggests that variation in CMA-ES performance is very high due to this interaction of population size with other hyperparameters as we observe a highly fluctuating performance of CMA-EA for varied λ values.

Covariance matrix size controller $\mu\lambda_{\text{ratio}}$. Hyperparameter $\mu\lambda_{\text{ratio}}$, which controls the percentage of population λ to be used for the covariance matrix estimation and update, has high interaction and direct influence on CMA-ES performance. The $\mu\lambda_{\text{ratio}}$ is the second most influential hyperparameter across all three methods (see Fig. 4). The significance of $\mu\lambda_{\text{ratio}}$ is evident as its values and the choice of λ are closely linked, and the choice of this ratio will increase or decrease the size of the covariance matrix that is at the core of the CMA-ES algorithm functioning. Similar to the performance of λ, $\mu\lambda_{\text{ratio}}$ performance is largely variable for its values (see Fig. 5).

Initial step size σ_0. Fig. 4 confirms the significance of σ_0 (initial step size) influence as this emerged as the next best hyperparameter in Morris and Sobol methods. Morris LHS, which is a stratified sampling method that covers the most hyperspace region, suggests that σ_0 is taken from most regions of its possible values and the CMA-ES performance had varied because of such sampling. However, the scores remain relatively high (see Fig. 5). The performance σ_0 in Morris LHS is also impacted by the fact that for almost half of the time, its re-scaling was switched off by $\sigma_{0-scale}$. Accordingly, $\sigma_{0-scale}$ should have a higher influence on Morris LHS than σ_0, which indeed is the case (see Fig. 4). Examining Fig. 5, we may observe that for range $[1, 2]$ of σ_0 values, CMA-ES mean performances were largely consistent (or above certain high scores). More precisely, a range $[0.8, 1.5)$ σ_0 produces the best performance.

Learning-rate α_μ. Learning-rate α_μ was found to be non-influential. However, since the performance of CMA-ES was consistent for its chosen values across all three methods, the learning-rate α_μ was better than re-scaling $\sigma_{0-scale}$. Moreover, the learning-rate α_μ shows more interaction with other hyperparameters than convergence speed controller $\sigma_{0-scale}$. This is also evident as the gray bars are larger than white bars in Fig. 4 and drop in performance for only a very small range of values around 2 in Fig. 5).

Convergence controller $\sigma_{0-scale}$. CMA-ES convergence controller hyperparameter $\sigma_{0-scale}$, a hyperparameter meant for re-scaling of initial step size σ_0 on and off, is the least influential in both Morris and Sobol methods (see Fig. 3). This result is supported by both Fig. 4 and Fig. 5. However, it is an influential hyperparameter in the sense that it has a very high influence on σ_0, which is the third most influential hyperparameter. From Fig. 5, it is evident that no re-scaling of σ_0 performs better than re-scaled σ_0. This is the reason why for Morris LHS, σ_0 is the least influence hyperparameter.

CMA-ES hyperparameters ranking. In summary, we may provide a *ranking of hyperparameters* for CMA-ES from the most to least influential hyperparameter as λ, $\mu\lambda_{\text{ratio}}$, σ_0, α_μ, and $\sigma_{0-scale}$. One may ignore tuning α_μ and $\sigma_{0-scale}$ completely as setting a sufficiently large function evaluation number would neutralize their importance in the CMA-ES algorithm.

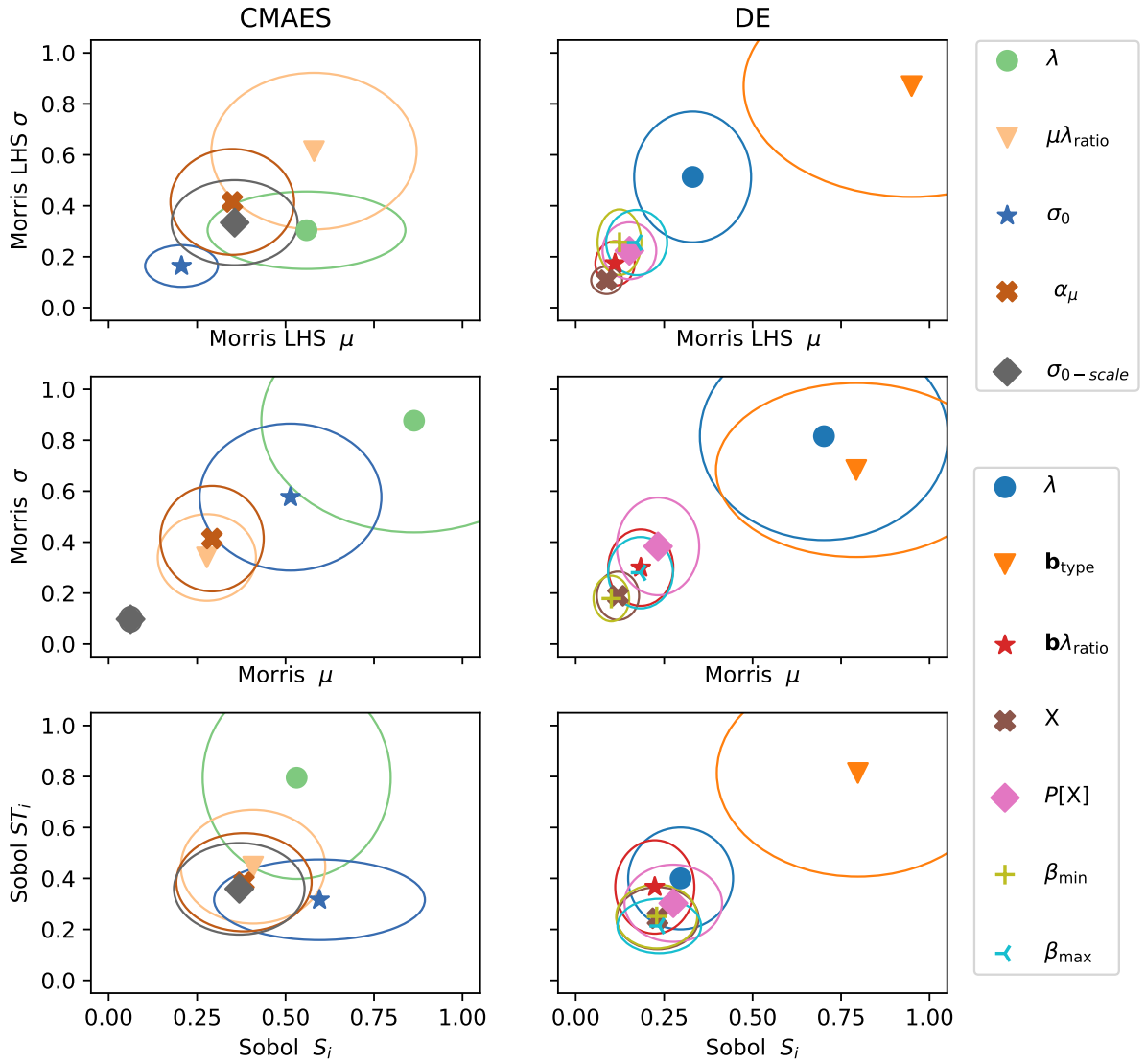

Fig. 3: Single objective algorithms sensitivity analysis. CMA-ES and DE hyperparameters sensitivity analysis are shown in column 1 and column 2, respectively. Rows 1, 2, and 3 respectively indicate Morris LHS, Morris, and Sobol methods. The upper right legend belongs to CMA-ES and the lower right to DE. A symbol and a color represent each hyperparameter. An eclipse centered at a hyperparameter is the span of the standard deviation of the influence along with direct and interaction influences. A larger width of the eclipse of a hyperparameter in the x-axis direction means more variation in direct dominance of that hyperparameter, and a larger height in the y-axis direction means its variation in total (or interaction) influence. In each plot, further apart a hyperparameter in the diagonal direction from the origin $(0, 0)$ is, higher is its importance to the algorithm. CMA-ES hyperparameter λ, $\mu\lambda_{\text{ratio}}$, σ_0, α_μ, and $\sigma_{0-scale}$ respectively are population size, percentage of the population for covariance matrix, initial step size, learning rate, and convergence speed controller. DE hyperparameters λ, $\mathbf{b}_{\text{type}}$, $\mathbf{b}\lambda_{\text{ratio}}$, X, P[X], $\beta_{\min}$, and $\beta_{\max}$ respectively are population size, base vector selection type (mutation type), percentage of population for base vector selection, crossover type, crossover probability, minimum acceleration coefficient, and maximum acceleration coefficient. The hyperparameters definitions and domains are given in Table 1. Supporting statistical tests (Kalpić et al., 2011) between direct and interaction effects and clustering analysis are provided in Sections A and B.

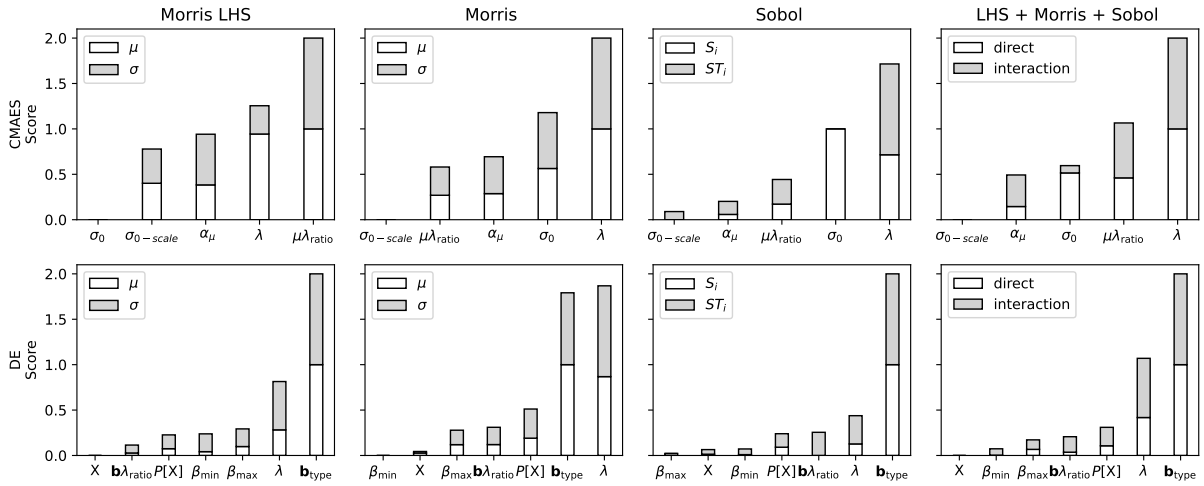

Fig. 4: Ordering (small to larger) of the sum of sensitivity analysis indices of single objective algorithms. CMA-ES (row 1) and DE (row 2) algorithms hyperparameters performance across all problems (functions). Columns 1, 2, and 3 respectively show performance evaluated using Morris LHS, Morris, and Sobol methods. The white color portion of a bar is the direct influence normalized value in $[0, 1]$ and gray color portion is interaction (total) influence value in $[0, 1]$. Larger height bar implies a higher influence. The hyperparameters name, definition and domain are given in Table 1.

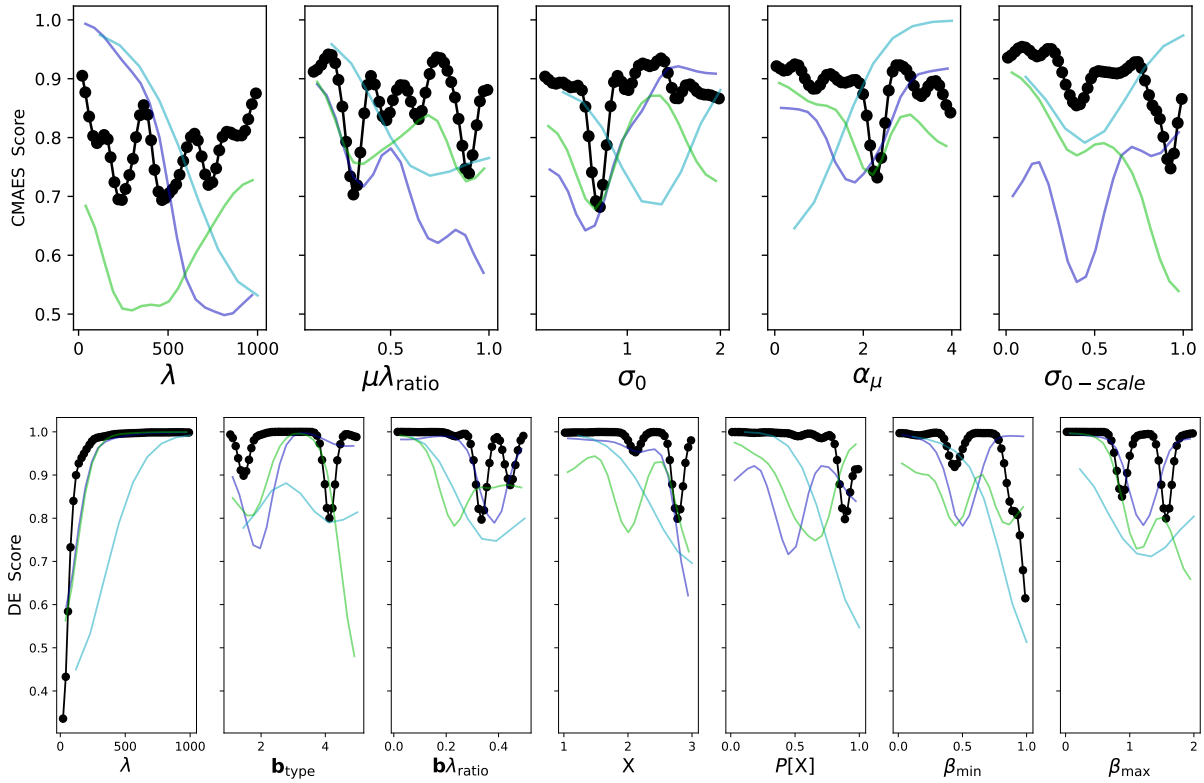

Fig. 5: CMA-ES and DE algorithms average performance on 10 runs (for 72 cases, 30 runs) of each hyperparameter set. CMA-ES and DE algorithms had 2740 and 2.980 hyperparameter sample sets evaluated in total (black dots) by Morris LHS (blue lines), Morris (cyan lines), Sobol (green lines) methods. Each dot in a subplot is a mean performance of a bin of total samples. Along x-axis there are 50 bins from lower to higher values which are plotted against each hyperparameter normalized score filtered (using Gaussian filter with sigma 2) in the y-axis.

6.1.2 DE Analysis

DE versions $\mathbf{b}_{\text{type}}$. DE results are shown in Figs. 3, 4, and 5, where Fig. 3 shows the quality of influence each hyperparameter has on DE performance over 33 problems. As per the result of the three sensitivity analysis methods, it is clear that the type of DE base vector selection method (mutation type) $\mathbf{b}_{\text{type}}$ (DE version) is by far the most significant hyperparameter. Examining Fig. 5, we observe that the type of DE mutation strategy *best* and *rand* have lower scores, and they fluctuate. In comparison, the base vector selection methods *target-to-best* and *best-to-best* performed more consistently with high scores. Therefore, the average performance of DE over the testbench was highly sensitive to the selection of $\mathbf{b}_{\text{type}}$. We also observe that $\mathbf{b}_{\text{type}}$ in Fig. 3 remains at $(1, 1)$ corner of the plot, meaning it had both a high *direct effect* and high *interaction effect*.

Population size λ. Overall population size λ is the second most influential hyperparameter in DE (cf. Figs. 3 and 4). Comparatively, it had produced better scores for larger population size than the smaller population size (see Fig. 5). However, the size of the population of DE was a distant second influential hyperparameter. This indicates that for apart a very small population size (less than 200), DE's performance was invariable when population size was increased from 200 to 1000 (Fig. 5). This was when the number of function evaluations was the same for each population size, i.e., the number of function evaluations was 10 000 for each population size.

Crossover type X **and crossover probability** P[X]**.** The crossover related hyperparameters are the type of crossover X and the probability of crossover P[X]. Between these, P[X] plays a vital role in DE's performance, and the type of DE $\mathbf{b}_{\text{type}}$ was the least influential (cf. Figs. 3 and 4). For crossover-type *binomial* offered better scores than the crossover-type *exponential* (see Fig. 5). The crossover probability P[X] has its usage only for binomial crossover. Hence, it was an influential hyperparameter in this setting.

Base vectors selection pool $\mathbf{b}\lambda_{\text{ratio}}$. The hyperparameter $\mathbf{b}\lambda_{\text{ratio}}$ defines the percentage of the population used for the selection of base vectors for DE. We found that $\mathbf{b}\lambda_{\text{ratio}}$ has a negligible influence on the performance of DE (cf. Figs. 3 and 4).

Acceleration coefficients $\beta_{\min}$ and $\beta_{\max}$. Similar to $\mathbf{b}\lambda_{\text{ratio}}$, acceleration coefficients hyperparameters $\beta_{\min}$ and $\beta_{\max}$ are least significant in DE. However, the setting of an appropriate range is vital for DE performance, as we observed in Fig. 5. This is evident because $\beta_{\min}$ and $\beta_{\max}$ acquire a relatively moderate influence in Morris LHS methods (see Fig. 4). Since the Morris LHS method uses a stratified sampling approach, it forced the selection of $\beta_{\min}$ and $\beta_{\max}$ values across their whole range and the performance of DE is impacted negatively by the higher values of $\beta_{\min}$ and $\beta_{\max}$. Examining Fig. 5, we observed that $\beta_{\min}$ scores for values in $[0.0, 0.5]$ performed consistently with better scores than the values in $(0.5, 1.0]$. However, Morris and Sobol had a uniform distribution and show that the influence of this hyperparameter is

non-influential; therefore, setting these values somewhere in $[0.0, 0.5]$ will suffice, and one may not need to exhaustively tune this hyperparameter.

Similarly, $\beta_{\max}$ was found sensitive to its range selection. Fig. 5 offers us the ways to investigate which range had a positive influence and which had a negative. We observe that the lower values had higher scores than the larger values of $\beta_{\max}$ (see Fig. 5). Investigating closely, we found that scores in $[0.2, 0.9]$ are by far better than the scores for other values. This means tuning $\beta_{\max}$ values within range $[0.2, 0.9]$ for a problem is an appropriate strategy.

DE hyperparameters ranking. In summary, we *rank* DE hyperparameters from the most significant to least significant as $\mathbf{b}_{\text{type}}$, λ, P[X], $\mathbf{b}\lambda_{\text{ratio}}$, $\beta_{\max}$, $\beta_{\min}$, and X. That one would safely use DE with binomial crossover and set appropriate values (discussed above) of $\beta_{\max}$, $\beta_{\min}$.

6.1.3 Remarks on SOO hyperparameter rankings and algorithms

We evaluated two single objective optimization algorithms and presented rankings of their hyperparameter influence based on a combined assessment of three sensitivity analysis methods. Each method, as mentioned, has its own way of drawing samples from the hyperparameter space and thus has produced its own ranking. However, the results reveal some obvious signs of influence based on direct and interaction effects.

Section A provides rich information on statistical tests among hyperparameters that one can thoroughly examine to reach the presented ranking and may find more information should one is interested in studying specific hyperparameters. For instance, the interaction effect of population size in CMA-ES is more significant than its direct effect (see Section A), which confirms the analysis presented in Fig. 5. Additionally, clustering analysis of hyperparameters and objective function provides information behaviors of the algorithm on the class of problems they solve (see Section B). For example, the type of mutation in DE forms a distinct cluster of its performance characteristics, and all other hyperparameters are grouped together in one cluster (see Section B).

As a consequence of this analysis and results presented in Section 6.1, it is clear that *DE is a more stable algorithm than CMA-ES*. See variation in scores of the hyperparameters of CMA-ES algorithm compared to DE's hyperparameters in Fig. 5 and high interaction among CMA-ES's hyperparameters. In contrast, DE has a clear ranking of hyperparameters. Additionally, during the experiments, CMA-ES failed to solve some class of problems for some combination of hyperparameters (see results in Ojha et al. (2022)).

6.2 Multi-objective EAs

6.2.1 NSGA-III Analysis

Population size λ. Results of NSGA-III are presented in Figs. 6, 7, and 8. In Fig. 6, we present results of three measures GD, IGD, and HV; see columns in Fig. 6, and along rows in Fig. 6, we present Morris LHS, Morris, and Sobol sensitivity methods. For NSGA-III, we clearly observe that the population size λ is a significant hyperparameter, and the probability of crossover is the second most significant hyperparameter. Population size influence has approximately equal high direct influence and high interaction influence. That is, although population size is the most significant hyperparameter, NSGA-III performance varied because of the variation of the other hyperparameters as well (see NSGA-III has a monotonous line for λ in Fig. 8 that indicates a more liner influence on NSGA-III). This fact was found true across all methods and all measures as the eclipse of its influence centered around coroner $(1, 1)$ in Fig. 6, and the white and gray bars have comparable lengths in Fig. 8.

An examination of scores of the population size shows that population size does not fluctuate much for HV metric after a certain population size, but for GD and IGD metrics, the scores keep increasing for increasing population size (see Fig. 8). However, this is monotonous, and one would expect such performance for GD and IGD metrics. The probability of crossover shows more fluctuations of all three metrics. Therefore, the variations in the performance of NSGA-III after a sufficiently large population size (in this case, 200) come from the variations of other hyperparameters, including crossover probability.

Crossover and mutation hyperparameters. The probability of crossover $P[\mathrm{X}]$ shows a more linear relationship between its values and NSGA-III performance measures GD, IGD, and HV. For increasing values of crossover probability, we see decreasing GD and IGD scores (signs of better performance) and increasing scores of HV for some values (see Fig. 8). A crossover rate around 0.6 lead to better solutions along the problem's objective dimensions, i.e., increasing scores of HV and lower scores of GD and IGD. This fact is supported by the strong direct and interaction influence of crossover $P[\mathrm{X}]$ for IGD and HV metrics and relatively direct influence on GD. Sobol method on $P[\mathrm{X}]$ does show a very strong total influence compared to direct influence on all metrics. In summary, the $P[\mathrm{X}]$ performance has a behavior of monotonous increase and is one of the most influential hyperparameters in NSGA-III.

For crossover related hyperparameter crossover distribution indices X_{DI}, the performance remains consistent and largely non-influential (cf. Figs 6 and 8) as only for a certain range of its value (a small range around 100), it shows a spike in performance of NSGA-III. Similarly, mutation distribution index $\mathrm{PM}_{\mathrm{DI}}$, the performance of NSGA-III is better for a certain range (around $100 - -150$ or low values of $\mathrm{PM}_{\mathrm{DI}}$, see Fig. 8). For both X_{DI} and $\mathrm{PM}_{\mathrm{DI}}$, this phenomenon occurred roughly around a value of 100 of these indices, which aligned with the range

for these hyperparameters suggested in (Deb and Deb, 2014, Deb et al., 1995).

Similar to the probability of crossover P[X], the probability of mutation P[PM] shows a sudden change in performance around a value of 0.6, but in a complementary direction (see a drop in HV and spike in GD and IGD metric in Fig. 8). The direct and interaction influence of mutation related hyperparameters P[PM] and $\mathrm{PM_{DI}}$ is low for NSGA-III (cf. Figs. 6 and 7).

Tournament size K. Tournament size K, probability of polynomial mutation P[PM], polynomial mutation distribution index $\mathrm{PM_{DI}}$ and simulated binary crossover (SBX) distribution index X_{DI} have comparable significance. However, they differ for different methods and metrics. Among these hyperparameters, tournament size K clearly shows a high influence on NSGA-III performance. Tournament size K shows more interaction influence than direct influence, except for the HV metric of the Sobol method. The high score of K in Fig. 8 with clear fluctuation is the evidence of its interaction with other hyperparameters, but the scores (especially in GD and IGD scores) shows an upward trend, indicating it has comparatively less influence on guiding population towards true Pareto-front than hyperparameters P[PM], $\mathrm{PM_{DI}}$ and X_{DI}. We may also confirm that the lower value of K is more influential than its higher values.

NSGA-III hyperparameters ranking. Considering the hyperparameters performance influence, we rank them from most influential to least influential hyperparameters as λ, P[X], X_{DI}, K, P[PM], and $\mathrm{PM_{DI}}$. Here, λ is effective up to a certain population size and then λ saturates. The tuning of crossover P[X] linearly influence NSGA-III, and X_{DI}, P[PM], and $\mathrm{PM_{DI}}$ require setting a fixed value, but their influence fluctuates, i.e., they are affected by the setting of values of other hyperparameters a lot.

6.2.2 MOEA/D Analysis

Population size λ. MOEA/D results are shown in Figs. 9, 10, and 11. In Fig. 9, the results of three sensitivity analysis methods for three metrics of MOEA/D performance are presented. Unlike NSGA-III results, population size λ is not a clear most significant hyperparameter for MOEA/D multi-objective algorithm. Rather, MOEA/D's hyperparameters $Mode$, the MOO decomposition method, is also among the influential hyperparameters. Morris LHS method shows that the $Mode$ is the most significant hyperparameter overall on three metrics. Fig. 11 also confirms this fact as for the population size values, the GD, IGD, and HV metrics show a strong relation.

For example, the HV metric in Fig. 11 shows a linear trend, but it has clear fluctuations in scores. This is because population size has high interaction with other hyperparameters, and tuning population size alone cannot compensate for the role of the other hyperparameters in the performance of MOEA/D on GD metric. However, for the IGD metric, population size improves the performance of MOEA/D. This shows a highly fluctuating behavior of population

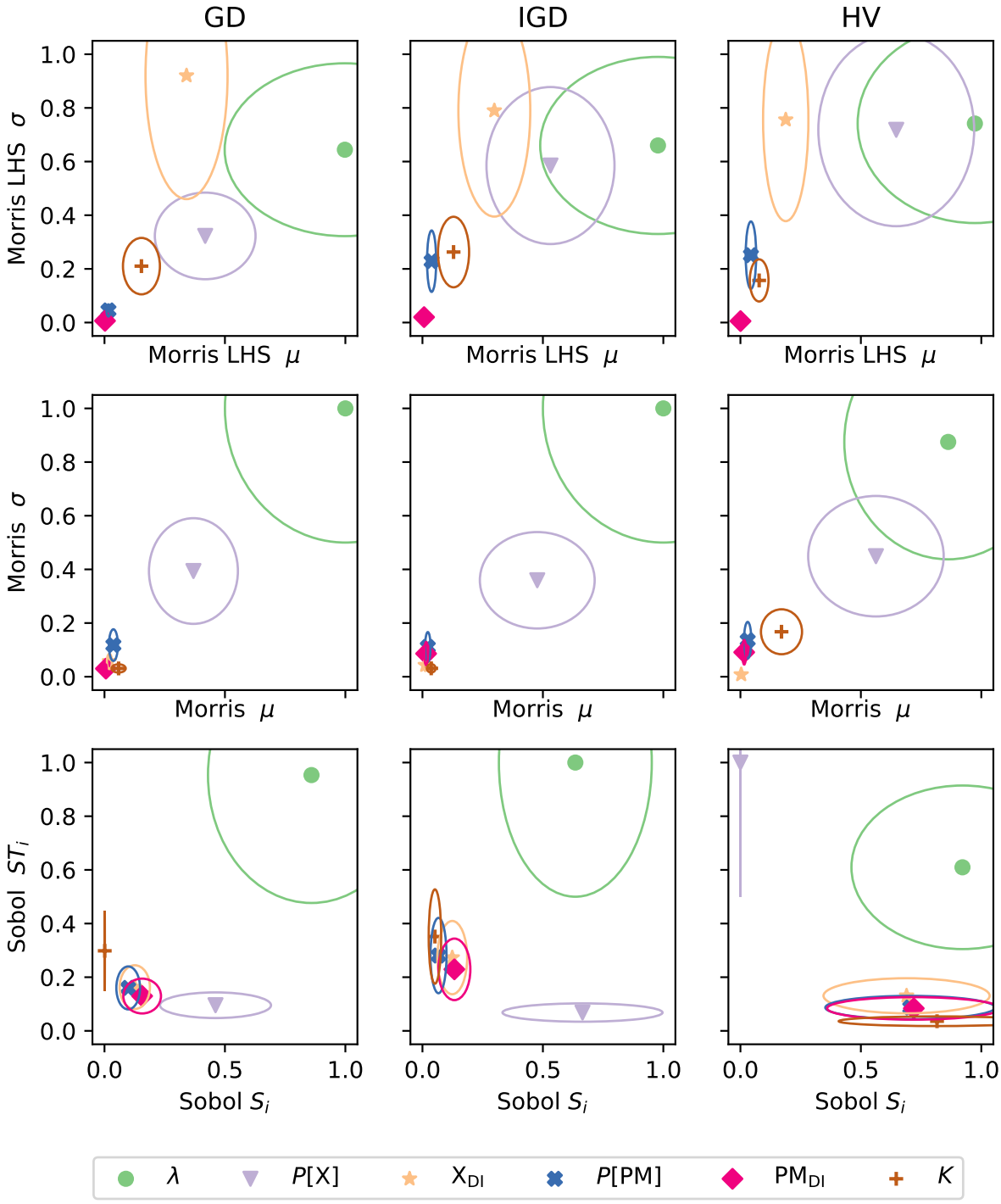

Fig. 6: NSGA-III hyperparameters sensitivity analysis. Columns 1, 2, and 3 respectively indicate performance metrics GD, IGD, and HV. Rows 1, 2 and 3 respectively indicate Morris LHS, Morris, and Sobol methods. Legends of hyperparameter are shown at the bottom. Each hyperparameter is represented by a symbol and a color. An eclipse centered at a hyperparameter is the standard deviation of its influence and direction of its influence. Further apart a hyperparameter in the diagonal direction from the origin $(0, 0)$ is, the higher is its importance to the algorithm. A larger width of the eclipse of a hyperparameter in the x-axis direction means more variation in direct influence of a hyperparameter, and a larger height in the y-axis direction means variation in total (or interaction) influence. Supporting statistical tests and clustering analysis are provided in Sections A and B.

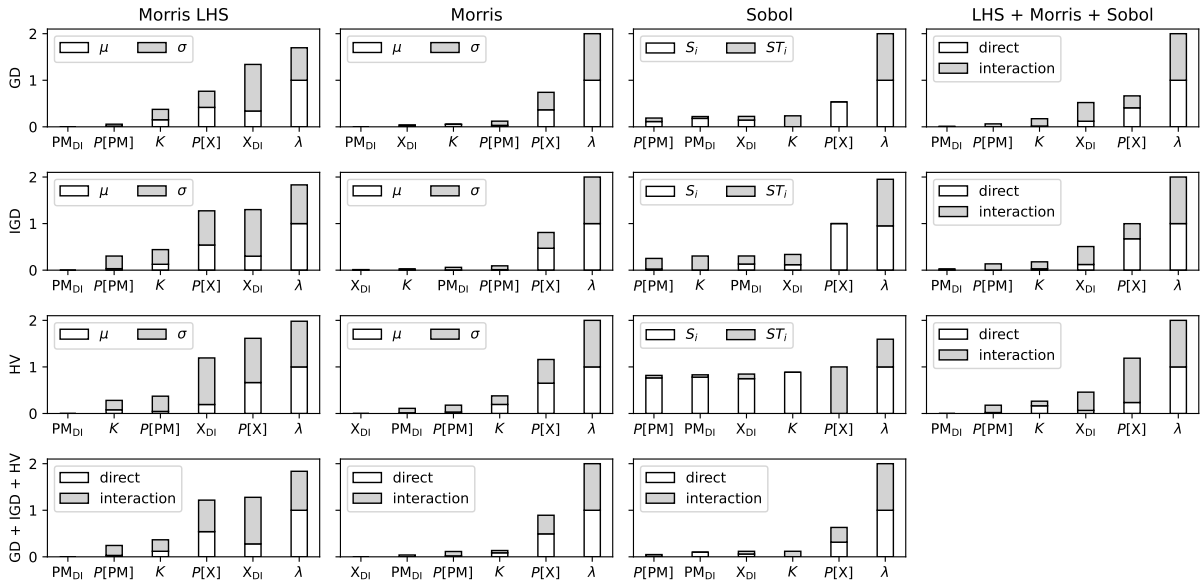

Fig. 7: NSGA-III algorithm's hyperparameters performance across all problems (functions). Rows 1, 2, and 3 respectively show performance evaluated using GD, IGD, and HV metrics. Columns 1, 2, and 3 respectively indicate Morris LHS, Morris, and Sobol methods. The white color portion of a bar is direct influence normalized value in $[0, 1]$ and gray color portion is interaction (total) influence value in $[0, 1]$. A larger height bar implies a higher influence, and hyperparameters in each subplot are arranged from low to high influence.

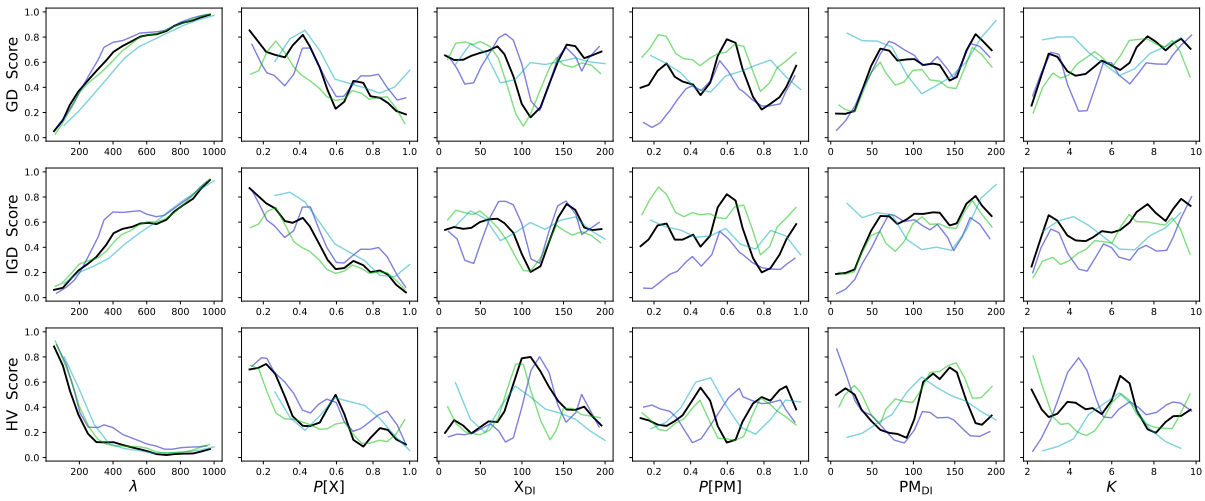

Fig. 8: NSGA-III algorithm average performance on 30 runs of each set of hyperparameters. NSGA-III hyperparameter (x-axis) against the mean metric value (y-axis). Rows 1, 2, and 3 respectively are GD, IGD, and HV metrics. The scores are normalized between 0 and 1 and smooth out using a Gaussian 1D filter with sigma 0.99. The y-axis is GD, IGD, and HV metrics values normalized between a score 0 and 1, where 0 is the best score for GD and IGD, and 1 is the best score for HV. A total of 520 samples were evaluated for the NSGA-III algorithm jointly by Morris LHS (blue lines), Morris (cyan lines), Sobol (green lines) methods. The hyperparameter values are arranged in 20 bins (lower values to higher values) across the x-axis. Each line in each plot connects the mean values of 20 bins of such samples.

size in MOEA/D for varied metrics, i.e., MOEA/D performance has a non-linear relation with the population size. This means population size is rather highly involved with interaction with other hyperparameters as the variation in other hyperparameters also influences the performance of MOEA/D.

MOO decomposition type $Mode$**.** The next set of hyperparameters that we observe as highly influential is $Mode$, as it shows high interaction and high overall influence in Morris LHS, Morris, and Sobol for GD and HV metrics. HV metric for Sobol placed the hyperparameters on the direct influence to high total influence diagonal (see Fig. 9), which suggests that the hyperparameters either have a good high interaction or good overall influence. Hence, the sum of these, presented in Fig. 10, differs only marginally. Sobol rank $Mode$ is second in the GD metric as both high interaction and high overall influence and third in HV metric as it has a high direct influence.

Examining the performance of $Mode$ in Fig. 11, we confirm that the type of MOO decomposition "Tchebycheff with normalization" had the best performance followed by "penalty based boundary intersection (PBI)" and "Tchebycheff" has significantly poor performance and "modified Tchebycheff," decomposition mode had the worse scores among MOO decomposition methods. MOEA/D hyperparameter ϵ_N refers to the number of neighbors for selecting the percentage of the population for sub-problems selection MOEA/D has an equivalent influence as the probability of mutation distribution index $\mathrm{PM_{DI}}$. However, ϵ_N value less than 0.2 show a sharp improvement in MOEA/D performance.

Crossover and mutation hyperparameters. Genetic operator related hyperparameters $P[\mathrm{X}]$, X_{DI}, $P[\mathrm{PM}]$ and $\mathrm{PM_{DI}}$ show varied significance on different metrics on different sensitivity methods. For example, the probability of mutation distribution index $\mathrm{PM_{DI}}$ has a high influence on HV metrics (pink diamond and eclipse in Fig. 9) and high total influence on HV metrics in the Sobol method. The probability of mutation $P[\mathrm{PM}]$ is second to $\mathrm{PM_{DI}}$ in total influence on HV as per the Sobol method. This suggests that *mutation has a high influence in diversifying the population* in MOEA/D, helping it produces a better Pareto-front. We also observe that $P[\mathrm{PM}]$ and $\mathrm{PM_{DI}}$ have mirror image like performance (see Fig. 11), which suggests that values of $P[\mathrm{PM}]$ around 0.8 and higher values of $\mathrm{PM_{DI}}$ are more effective in MOEA/D performance. The probability of crossover $P[\mathrm{X}]$ has competing performance in the MOEA/D, and it is similar to performances of mutation related hyperparameters. That is, unlike NSGA-III, the probability of crossover does not outshine among the crossover and mutation related hyperparameters.

MOEA/D hyperparameters ranking. In summary, the *ranking of hyperparameters* of MOEA/D from the most influential to least influential hyperparameters is λ, $Mode$, $\mathrm{PM_{DI}}$, $P[\mathrm{PM}]$, $P[\mathrm{X}]$, ϵ_N, and X_{DI}.

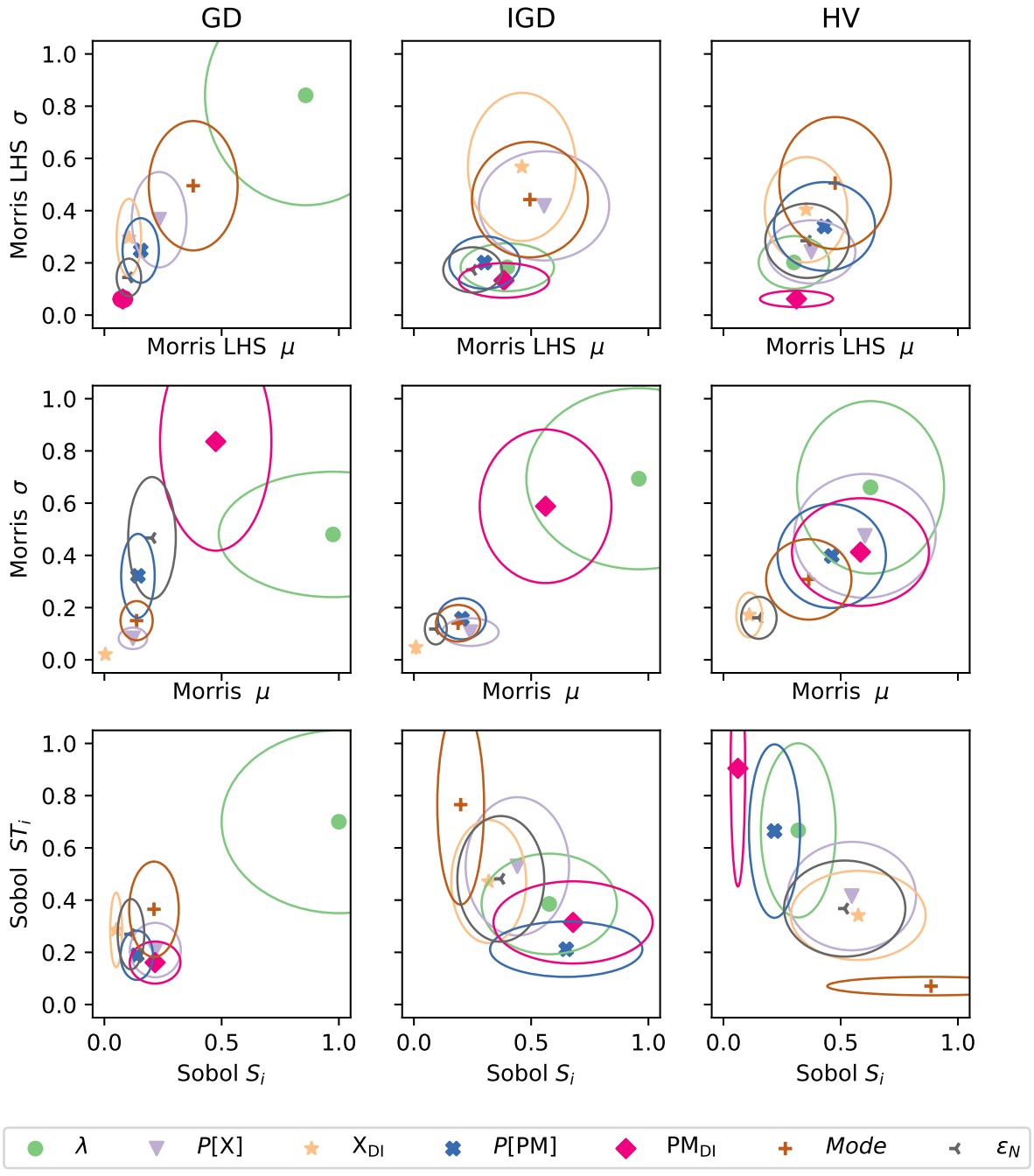

Fig. 9: MOEA/D hyperparameters sensitivity analysis. Columns 1, 2, and 3 respectively indicate performance metrics GD, IGD, and HV. Rows 1, 2 and 3, respectively indicate Morris LHS, Morris, and Sobol methods. Legends of hyperparameter are shown at the bottom. Each hyperparameter is represented by a symbol and a color. An eclipse centered at a hyperparameter is the standard deviation of its influence and direction of its influence. Further apart a hyperparameter in the diagonal direction from the origin $(0, 0)$ is, the higher is its importance to the algorithm. A larger width of the eclipse of a hyperparameter in the x-axis direction means more variation in direct influence of a hyperparameter, and a larger height in the y-axis direction means variation in total (or interaction) influence. Supporting statistical tests and clustering analysis are provided in Sections A and B.

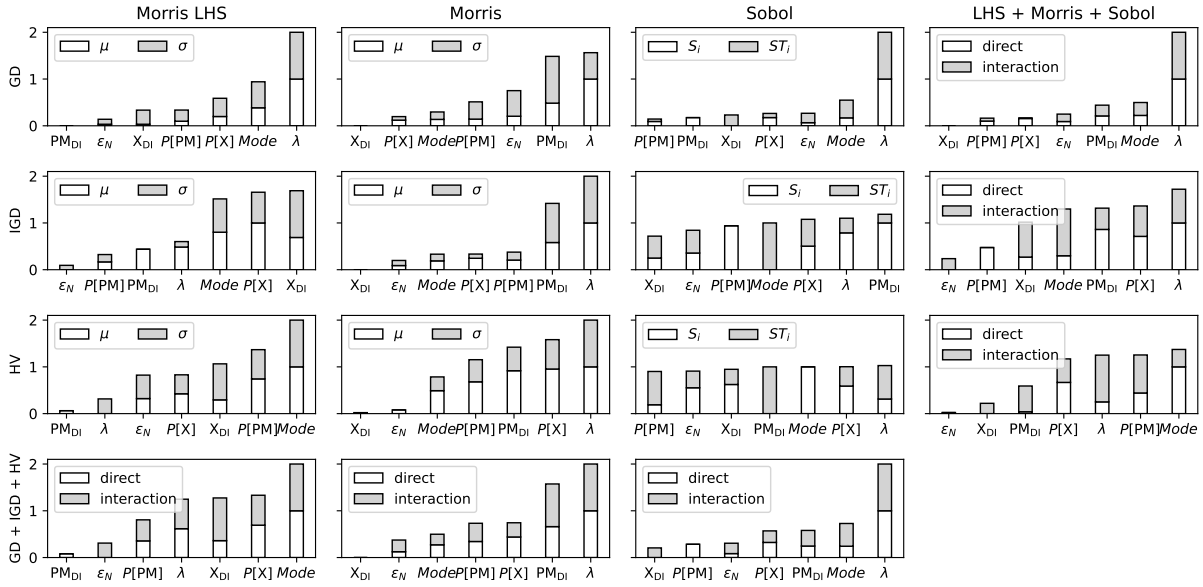

Fig. 10: MOEA/D algorithm's hyperparameters performance across all problems (functions). Rows 1, 2, and 3 respectively show performance evaluated using GD, IGD, and HV metrics. Columns 1, 2, and 3 respectively indicate metric Morris LHS, Morris, and Sobol methods. The white color portion of a bar is direct influence normalized value in $[0, 1]$ and gray color portion is interaction (total) influence value in $[0, 1]$. A larger height bar implies a higher influence, and hyperparameters in each subplot are arranged from low to high influence.

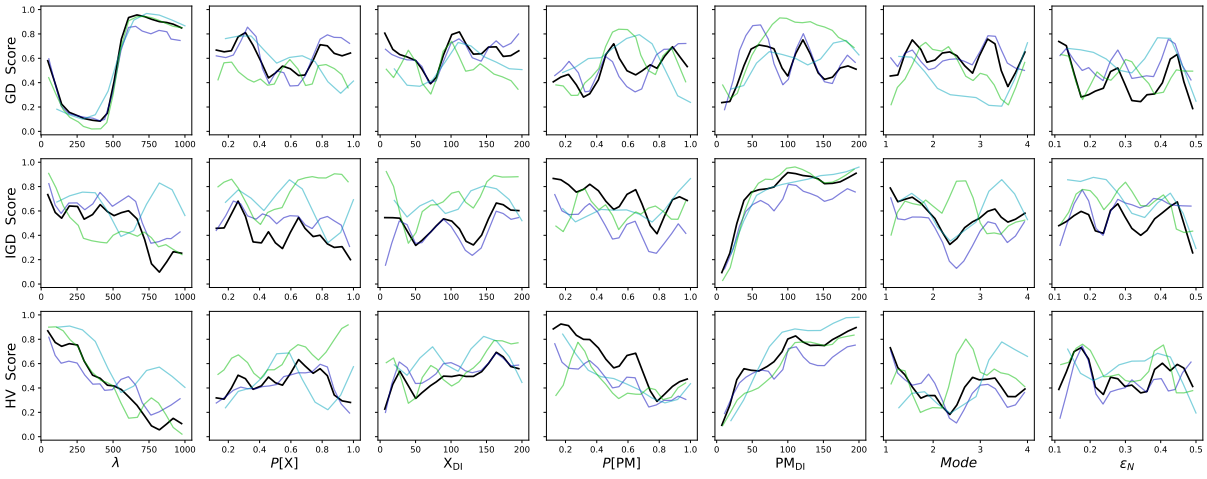

Fig. 11: MOEA/D algorithm average performance on 30 runs of each set of hyperparameters. MOEA/D hyperparameter (x-axis) against the mean metric value (y-axis). Rows 1, 2, and 3 respectively are GD, IGD, and HV metrics. The scores are normalized between 0 and 1 and smooth out using a Gaussian 1D filter with sigma 0.99. The y-axis is GD, IGD, and HV metrics values normalized between a score 0 and 1, where 0 is the best score for GD and IGD, and 1 is the best score for HV. A total of 590 samples were evaluated for the MOEA/D algorithm jointly by Morris LHS (blue lines), Morris (cyan lines), Sobol (green lines) methods. The hyperparameter values are arranged in 20 bins (lower values to higher values) across the x-axis. Each line in each plot connects the mean values of 20 bins of such samples.

6.2.3 Remarks on MOO hyperparameter rankings and algorithms

Providing ranking to hyperparameters for MOO is more challenging than SOO since it uses three distinct sensitivity analysis methods and uses three distinct performance metrics. However, we look for potential agreement between these distinct measures. We observe that the population size λ clearly emerged as the most influential hyperparameter in all three analyses and metrics for NSGA-III and probability of crossover the second most influential. These two hyperparameters significantly dominate all other hyperparameters in NSGA-III. Whereas for MOEA/D, the population size λ dominates only for the GD metric and for Morris analysis. For HV and IGD metric and Morris LHS and Sobol analysis, *Mode* and mutation probability are dominant factors. Unlike NSGA-III, there is no clear, significantly dominant hyperparameter in MOEA/D. Therefore, considering hyperparameters' strong variability and dependency on the type of hyperparameter sampling methods and type of performance metrics, we may confirm that NSGA-III is a more stable algorithm than MOEA/D.

7 Conclusions

We present a framework for systematic and methodological analysis of the effectiveness of the evolutionary algorithm hyperparameters. This analysis results in (i) identifying the pattern of influence each hyperparameter has on the algorithm, (ii) recommending rankings of hyperparameters influence, and (iii) analyzing the stability of algorithms related to hyperparameter sampling and performance metrics. We apply our methodology to state-of-the-art evolutionary algorithms: two single-objective algorithms and two multi-objective algorithms. The single-objective algorithms used are covariance matrix adaptation evolutionary strategy (CMA-ES), differential evolution (DE), and multi-objective algorithms used are non-dominated sorting genetic algorithm III (NSGA-III), and multi-objective evolutionary algorithm based on decomposition (MOEA/D). Our methodology involves two global sensitivity analysis methods, Morris and Sobol. This methodology is computationally heavy, but it produces widely usable and effective recommendations on hyperparameters ranking, being the order in which one can tune EA hyperparameters to achieve high performance. For example, the initial step size, base vector selection type (mutation), probability of crossover, and mode multi-objective problem decomposition were among the most influential hyperparameters of CMA-ES, DE, NSGA-III, and MOEA/D algorithms, respectively. The results show how the hyperparameters interact with one another when they are sampled differently, and different performance measures are used.

References

Bergstra, J. and Bengio, Y. (2012), 'Random search for hyper-parameter optimization', *Journal of Machine Learning Research* **13**, 281–305.

Bezerra, L. C., López-Ibánez, M. and Stützle, T. (2015), Comparing decomposition-based and automatically component-wise designed multi-objective evolutionary algorithms, *in* 'International Conference on Evolutionary Multi-Criterion Optimization', Springer, pp. 396–410.

Biswas, A., Das, S., Abraham, A. and Dasgupta, S. (2009), 'Design of fractional-order PIλDμ controllers with an improved differential evolution', *Engineering Applications of Artificial Intelligence* **22**(2), 343–350.

Brooks, R., Semenov, M. and Jamieson, P. (2001), 'Simplifying Sirius: sensitivity analysis and development of a meta-model for wheat yield prediction', *European Journal of Agronomy* **14**, 43–60.

Campolongo, F., Saltelli, A. and Cariboni, J. (2007), 'An effective screening design for sensitivity analysis of large models', *Environmental Modelling & Software* **22**, 1509–1518.

Conca, P., Stracquadanio, G. and Nicosia, G. (2015), Automatic tuning of algorithms through sensitivity minimization, *in* P. Pardalos, M. Pavone, G. M. Farinella and V. Cutello, eds, 'Machine Learning, Optimization, and Big Data', Springer, pp. 14–25.

Crossley, M., Nisbet, A. and Amos, M. (2013), Quantifying the impact of parameter tuning on nature-inspired algorithms, *in* 'The 12th European Conference on Artificial Life', MIT Press, pp. 925–932.

Cui, Z., Chang, Y., Zhang, J., Cai, X. and Zhang, W. (2019), 'Improved NSGA-III with selection-and-elimination operator', *Swarm and Evolutionary Computation* **49**, 23–33.

Das, I. and Dennis, J. E. (1998), 'Normal-boundary intersection: A new method for generating the pareto surface in nonlinear multicriteria optimization problems', *SIAM Journal on Optimization* **8**(3), 631–657.

Das, S., Abraham, A., Chakraborty, U. K. and Konar, A. (2009), 'Differential evolution using a neighborhood-based mutation operator', *IEEE Transactions on Evolutionary Computation* **13**(3), 526–553.

Das, S., Abraham, A. and Konar, A. (2007), 'Automatic clustering using an improved differential evolution algorithm', *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans* **38**(1), 218–237.

Das, S., Konar, A. and Chakraborty, U. K. (2005), Two improved differential evolution schemes for faster global search, *in* 'Proceedings of the 7th Annual Conference on Genetic and Evolutionary Computation', pp. 991–998.

Das, S. and Suganthan, P. N. (2010), 'Differential evolution: A survey of the state-of-the-art', *IEEE Transactions on Evolutionary Computation* **15**(1), 4–31.

De Jong, K. (2007), 'Parameter setting in EAs: a 30 year perspective', *Studies in Computational Intelligence (SCI)* **54**, 1–18.

De Jong, K. (2016), Evolutionary computation: a unified approach, *in* 'Proceedings of the 2016 on Genetic and Evolutionary Computation Conference Companion', pp. 185–199.

Deb, K., Agrawal, R. B. et al. (1995), 'Simulated binary crossover for continuous search space', *Complex Systems* **9**(2), 115–148.

Deb, K. and Deb, D. (2014), 'Analysing mutation schemes for real-parameter genetic algorithms', *International Journal of Artificial Intelligence and Soft Computing* **4**(1), 1–28.

Deb, K. and Jain, H. (2013), 'An evolutionary many-objective optimization algorithm using reference-point-based nondominated sorting approach, Part I: Solving problems with box constraints', *IEEE Transactions on Evolutionary Computation* **18**(4), 577–601.

Deb, K., Pratap, A., Agarwal, S. and Meyarivan, T. (2002), 'A fast and elitist multiobjective genetic algorithm: NSGA-II', *IEEE Transactions on Evolutionary Computation* **6**(2), 182–197.

Deb, K., Thiele, L., Laumanns, M. and Zitzler, E. (2002), Scalable multi-objective optimization test problems, *in* 'Proceedings of 2002 IEEE Congress on Evolutionary Computation', Vol. 1, pp. 825–830.

Eiben, A. E., Michalewicz, Z., Schoenauer, M. and Smith, J. E. (2007), Parameter control in evolutionary algorithms, *in* 'Parameter setting in evolutionary algorithms', Springer, pp. 19–46.

Feurer, M., Eggensperger, K., Falkner, S., Lindauer, M. and Hutter, F. (2020), 'Auto-sklearn 2.0: Hands-free AutoML via meta-learning', *arXiv:2007.04074* .

Feurer, M. and Hutter, F. (2019), Hyperparameter optimization, *in* 'Automated Machine Learning', Springer, Cham, pp. 3–33.

Fonseca, C. M., Paquete, L. and López-Ibánez, M. (2006), An improved dimension-sweep algorithm for the hypervolume indicator, *in* 'IEEE International Conference on Evolutionary Computation', IEEE, pp. 1157–1163.

Greco, A., Riccio, S. D., Timmis, J. and Nicosia, G. (2019), Assessing algorithm parameter importance using global sensitivity analysis, *in* I. Kotsireas, P. Pardalos, K. E. Parsopoulos, D. Souravlias and A. Tsokas, eds, 'Analysis of Experimental Algorithms', Springer, pp. 392–407.

Hansen, N. and Ostermeier, A. (1996), Adapting arbitrary normal mutation distributions in evolution strategies: the covariance matrix adaptation, *in* 'Proceedings of IEEE International Conference on Evolutionary Computation', pp. 312–317.

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

He, X., Zhao, K. and Chu, X. (2021), 'AutoML: A survey of the state-of-the-art', *Knowledge-Based Systems* **212**, 106622.

Heris, S. (2019), 'YPEA: Yarpiz evolutionary algorithms'. `https://github.com/smkalami/ypea`. Accessed on 22 September 2021.

Hill, M. C., Kavetski, D., Clark, M., Ye, M., Arabi, M., Lu, D., Foglia, L. and Mehl, S. (2016), 'Practical use of computationally frugal model analysis methods', *Groundwater* **54**(2), 159–170.

Huband, S., Hingston, P., Barone, L. and While, L. (2006), 'A review of multiobjective test problems and a scalable test problem toolkit', *IEEE Transactions on Evolutionary Computation* **10**(5), 477–506.

Iglesias, P. L., Mora, D., Martinez, F. J. and Fuertes, V. S. (2007), 'Study of sensitivity of the parameters of a genetic algorithm for design of water distribution networks', *Journal of Urban and Environmental Engineering* **1**(2), 61–69.

Iommazzo, G., d'Ambrosio, C., Frangioni, A. and Liberti, L. (2019), Algorithmic configuration by learning and optimization, *in* 'Cologne-Twente Workshop on Graphs and Combinatorial Optimization'.

Iooss, B. and Saltelli, A. (2016), Introduction to sensitivity analysis, *in* R. Ghanem, D. Higdon and H. Owhadi, eds, 'Handbook of Uncertainty Quantification', Springer, pp. 1–20.

Islam, S. M., Das, S., Ghosh, S., Roy, S. and Suganthan, P. N. (2011), 'An adaptive differential evolution algorithm with novel mutation and crossover strategies for global numerical optimization', *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)* **42**(2), 482–500.

Jansen, T., Jong, K. A. D. and Wegener, I. (2005), 'On the choice of the offspring population size in evolutionary algorithms', *Evolutionary Computation* **13**(4), 413–440.

Kalpić, D., Hlupić, N. and Lovrić, M. (2011), *Student's t-Tests*, Springer, pp. 1559–1563.

Kramer, O. (2010), 'Evolutionary self-adaptation: a survey of operators and strategy parameters', *Evolutionary Intelligence* **3**(2), 51–65.

Liang, J. J., Baskar, S., Suganthan, P. N. and Qin, A. K. (2006), 'Performance evaluation of multiagent genetic algorithm', *Natural Computing* **5**(1), 83–96.

Liang, J. J., Qu, B. Y. and Suganthan, P. N. (2013), Problem definitions and evaluation criteria for the CEC 2014 special session and competition on single objective real-parameter numerical optimization, Technical report, Zhengzhou University, Zhengzhou China and Nanyang Technological University, Singapore.

Liang, J., Qu, B., Suganthan, P. and Chen, Q. (2014), Problem definitions and evaluation criteria for the CEC 2015 competition on learning-based real-parameter single objective optimization, Technical report, Zhengzhou University, Zhengzhou China and Nanyang Technological University, Singapore.

Lima, C. F. and Lobo, F. G. (2004), Parameter-less optimization with the extended compact genetic algorithm and iterated local search, *in* 'Genetic and Evolutionary Computation Conference', Springer, pp. 1328–1339.

Lloyd, S. P. (1982), 'Least squares quantization in PCM', *IEEE Transactions on Information Theory* **28**, 129–137.

López-Ibáñez, M., Dubois-Lacoste, J., Cáceres, L. P., Birattari, M. and Stützle, T. (2016), 'The irace package: Iterated racing for automatic algorithm configuration', *Operations Research Perspectives* **3**, 43–58.

Lou, Y., Yuen, S. Y. and Chen, G. (2021), 'Non-revisiting stochastic search revisited: Results, perspectives, and future directions', *Swarm and Evolutionary Computation* **61**, 100828.

Maturana, J., Lardeux, F. and Saubion, F. (2010), 'Autonomous operator management for evolutionary algorithms', *Journal of Heuristics* **16**(6), 881–909.

Miettinen, K. (2012), *Nonlinear multiobjective optimization*, Vol. 12, Springer.

Morris, M. D. (1991), 'Factorial sampling plans for preliminary computational experiments', *Technometrics* **33**(2), 161–174.

Ojha, V. K., Abraham, A. and Snášel, V. (2014*a*), ACO for continuous function optimization: A performance analysis, *in* '2014 14th International Conference on Intelligent Systems Design and Applications', IEEE, pp. 145–150.

Ojha, V. K., Abraham, A. and Snášel, V. (2014*b*), Simultaneous optimization of neural network weights and active nodes using metaheuristics, *in* '2014 14th International Conference on Hybrid Intelligent Systems', IEEE, pp. 248–253.

Ojha, V., Timmis, J. and Nicosia, G. (2022), 'Sensitivity analysis evolutionary algorithms'. `https://github.com/vojha-code/SAofEAs`. Accessed on 10 February 2022.

Paul, G., Müller, C. L. and Sbalzarini, I. F. (2011), Sensitivity analysis from evolutionary algorithm search paths, *in* 'EVOLVE - A bridge between Probability, Set Oriented Numerics and Evolutionary Computation', Studies in Computational Intelligence, Springer.

Pianosi, F., Sarrazin, F. and Wagener, T. (2015), 'An effective screening design for sensitivity analysis of large models', *Environmental Modelling & Software* **70**, 80–85.

Pinel, F., Danoy, G. and Bouvry, P. (2012), Evolutionary algorithm parameter tuning with sensitivity analysis, *in* P. Bouvry, M. A. Kłopotek, F. Leprévost, M. Marciniak, A. Mykowiecka and H. Rybiński, eds, 'Security and Intelligent Information Systems', Springer, pp. 204–216.

Qi, Y., Ma, X., Liu, F., Jiao, L., Sun, J. and Wu, J. (2014), 'MOEA/D with adaptive weight adjustment', *Evolutionary Computation* **22**(2), 231–264.

Rousseeuw, P. J. (1987), 'Silhouettes: A graphical aid to the interpretation and validation of cluster analysis', *Journal of Computational and Applied Mathematics* **20**, 53–65.

Saltelli, A. (2002), 'Sensitivity analysis for importance assessment', *Risk Analysis* **22**(3), 579–590.

Saltelli, A., Ratto, M., Andres, T., Campolongo, F., Cariboni, J., Gatelli, D., Saisana, M. and Tarantola, S. (2008), *Global Sensitivity Analysis: The primer*, John Wiley & Sons.

Saltelli, A., Tarantola, S., Campolongo, F. and Ratto, M. (2004), *Sensitivity analysis in practice: A guide to assessing scientific models*, Vol. 1, John Wiley & Sons.

Saltelli, A., Tarantola, S. and Chan, K.-S. (1999), 'A quantitative model-independent method for global sensitivity analysis of model output', *Technometrics* **41**(1), 39–56.

Sobol, I. M. and Kucherenko, S. (2005), 'Global sensitivity indices for nonlinear mathematical models, review', *Wilmott Magazine* **2005**, 56–61.

Storn, R. and Price, K. (1997), 'Differential evolution - a simple and efficient heuristic for global optimization over continuous spaces', *Journal of Global Optimization* **11**, 341–359.

Taylor, R., Ojha, V., Martino, I. and Nicosia, G. (2021), Sensitivity analysis for deep learning: ranking hyper-parameter influence, *in* '2021 IEEE 33rd International Conference on Tools with Artificial Intelligence (ICTAI)', IEEE, pp. 512–516.

Thornton, C., Hutter, F., Hoos, H. H. and Leyton-Brown, K. (2013), Auto-WEKA: combined selection and hyperparameter optimization of classification algorithms, *in* 'Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining', pp. 847–855.

Tian, Y., Cheng, R., Zhang, X. and Jin, Y. (2017), 'PlatEMO: A MATLAB platform for evolutionary multi-objective optimization', *IEEE Computational Intelligence Magazine* **12**(4), 73–87.

Veldhuizen, D. A. V. (1999), Multiobjective Evolutionary Algorithms: Classifications, Analyses, and New Innovations, PhD thesis, School of Engineering, Air Force Institute of Technology, Wright-Patterson AFB, Ohio.

Veldhuizen, D. A. V. and Lamont, G. B. (1998), Evolutionary computation and convergence to a pareto front, *in* 'Late Breaking Papers at the 1998 Genetic Programming Conference', Stanford University, pp. 221–228.

Voß, T., Hansen, N. and Igel, C. (2010), Improved step size adaptation for the MO-CMA-ES, *in* 'Proceedings of the 12th Annual Conference on Genetic and Evolutionary Computation', ACM, pp. 487–494.

Wang, Z., Zhang, Q., Zhou, A., Gong, M. and Jiao, L. (2015), 'Adaptive replacement strategies for MOEA/D', *IEEE Transactions on Cybernetics* **46**(2), 474–486.

Yao, X., Liu, Y. and Lin, G. (1999), 'Evolutionary programming made faster', *IEEE Transactions on Evolutionary Computation* **3**(2), 82–102.

Zhang, Q. and Li, H. (2007), 'MOEA/D: A multiobjective evolutionary algorithm based on decomposition', *IEEE Transactions on Evolutionary Computation* **11**(6), 712–731.

Zhang, Q., Li, H., Maringer, D. and Tsang, E. (2010), MOEA/D with NBI-style Tchebycheff approach for portfolio management, *in* 'IEEE Congress on Evolutionary Computation', IEEE, pp. 1–8.

Zitzler, E. and Thiele, L. (1998), Multiobjective optimization using evolutionary algorithms—a comparative case study, *in* 'International Conference on Parallel Problem Solving from Nature', Springer, pp. 292–301.

Zitzler, E., Thiele, L., Laumanns, M., Fonseca, C. M. and Da Fonseca, V. G. (2003), 'Performance assessment of multiobjective optimizers: An analysis and review', *IEEE Transactions on Evolutionary Computation* **7**(2), 117–132.

A Statistical Tests of Hyperparameters Influence

We present a pairwise statistical test (Kalpić et al., 2011) of hyperparameters in groups of direct effect and interaction effect of each algorithm. Tables A1 A2, A3, A4, A5, A6, A7, and A8 are statistical independent two-sample t-test results, where symbol 't' indicates t-statistics and 'p' is the p-value. The samples for this t-test are direct effect and interaction (or total) effect values of each hyperparameter for the functions evaluated in this study. For single objective testbench, the sample size is 33, and for multi-objective testbench, the sample size is 10. In the t-test, one may choose a value of 0.05 or less to indicate the t-test is significant. The magnitude of t-statistics indicates the significance of the mean and the sign (negative or positive) indicates the direction that is which sample is more significant than the other.

For example, in Table A1, t-test between pair CMA-ES parameters λ (first column) and $\mu\lambda_{\text{ratio}}$ (first row) indicate no significance of λ, although it indicates that the mean λ is better than the mean of $\mu\lambda_{\text{ratio}}$. This is because the p-value does not confirm its significance. Table A1 interestingly shows that the interaction effect of λ is much higher than the direct effect of λ, which is confirmed by its statistical significance (see row 9 and 10 and column 1 of Table A1 where mean of λ indicating interaction is better (t-stat is -2.14) since it has the p-value 0.04). Moreover, the interaction effect of λ is significantly better than all other hyperparameters. In summary, Tables A1 A2, A3, A4, A5, A6, A7, and A8 have arranged such a way that one needs to read *row* to find that the t-stat is significant only if it is negative and corresponding p-value less than 0.05. Contrarily, if one reads *column* direction, t-stat is significant only if it is positive and the corresponding p-value is less than 0.05.

In the case of single-objective optimization algorithms, results are shown in Tables A1 and A2 that support results of the ranking of CMA-ES and DE hyperparameters. For instance, in Table A1, the high interaction of λ is found to be more statistically significant than that of all other hyperparameters, including its own direct influence (read row-wise). This is followed by the strong direct influence of σ_0 (read column-wise). For DE, $\mathbf{b}_{\text{type}}$ has both strong (and statistically significant) direct and interaction effects on the DE performance compared to all other hyperparameters (read column-wise in Table A2). The hyperparameter λ has higher statistically significant interaction influence on DE performance compared to hyperparameters $\mathbf{b}\lambda_{\text{ratio}}$, X, β_{min}, and β_{max} (read row-wise).

For multi-objective optimization, NSGA-III hyperparameters λ and P[X] are observed to have a higher influence on NSGA-III performance than other hyperparameters (read columns in GD, IGD and HV metrics in Tables A3, A4, and A5). Similarly, for MOEAD, hyperparameter λ and mutation related hyperparameters P[PM] and PM_{DI} observed to have a higher statistical significance of their performance (see Tables A6, A7, and A8).

Table A1: CMA-ES Best Solution Metric T-Test Statistics

			Direct Effect					Interaction Effect			
	Param		λ	$\mu\lambda_{\text{ratio}}$	σ_0	α_μ	$\sigma_{0-scale}$	λ	$\mu\lambda_{\text{ratio}}$	σ_0	α_μ
Direct Effect	$\mu\lambda_{\text{ratio}}$	t	1.05								
		p	0.30								
	σ_0	t	-0.50	-1.72							
		p	0.62	0.09							
	α_μ	t	1.30	0.28	2.01						
		p	0.20	0.78	0.05						
	$\sigma_{0-scale}$	t	1.38	0.41	2.06	0.14					
		p	0.18	0.68	0.04	0.89					
Interaction Effect	λ	t	-2.14	-3.80	-1.72	-4.20	-4.15				
		p	0.04	0.00	0.09	0.00	0.00				
	$\mu\lambda_{\text{ratio}}$	t	0.72	-0.40	1.36	-0.70	-0.80	3.41			
		p	0.47	0.69	0.18	0.49	0.43	0.00			
	σ_0	t	1.67	0.85	2.29	0.63	0.48	4.16	1.19		
		p	0.10	0.40	0.03	0.53	0.63	0.00	0.24		
	α_μ	t	1.25	0.25	1.93	-0.03	-0.17	4.04	0.65	-0.64	
		p	0.22	0.81	0.06	0.98	0.87	0.00	0.52	0.53	
	$\sigma_{0-scale}$	t	1.56	0.57	2.32	0.28	0.11	4.64	1.01	-0.42	0.30
		p	0.13	0.57	0.02	0.78	0.91	0.00	0.32	0.67	0.76

Table A2: DE Best Solution Metric T-Test Statistics

			Direct Effect							Interaction Effect					
	Param		λ	$\mathbf{b}_{\text{type}}$	$\mathbf{b}\lambda_{\text{ratio}}$	X	P[X]	β_{min}	β_{max}	λ	$\mathbf{b}_{\text{type}}$	$\mathbf{b}\lambda_{\text{ratio}}$	X	P[X]	β_{min}
Direct Effect	$\mathbf{b}_{\text{type}}$	t	-5.99												
		p	0.00												
	$\mathbf{b}\lambda_{\text{ratio}}$	t	0.85	7.05											
		p	0.40	0.00											
	X	t	0.81	7.50	-0.10										
		p	0.42	0.00	0.92										
	P[X]	t	0.27	7.13	-0.70	-0.65									
		p	0.79	0.00	0.49	0.52									
	β_{min}	t	0.85	7.54	-0.07	0.04	0.68								
		p	0.40	0.00	0.94	0.97	0.50								
	β_{max}	t	0.75	7.39	-0.16	-0.07	0.57	-0.10							
		p	0.46	0.00	0.87	0.95	0.57	0.92							
Interaction Effect	λ	t	-1.19	4.79	-2.08	-2.13	-1.62	-2.17	-2.06						
		p	0.24	0.00	0.04	0.04	0.11	0.03	0.04						
	$\mathbf{b}_{\text{type}}$	t	-6.34	-0.21	-7.44	-7.96	-7.61	-8.00	-7.85	-5.12					
		p	0.00	0.83	0.00	0.00	0.00	0.00	0.00	0.00					
	$\mathbf{b}\lambda_{\text{ratio}}$	t	-0.76	4.90	-1.60	-1.60	-1.11	-1.63	-1.54	0.37	5.21				
		p	0.45	0.00	0.12	0.11	0.27	0.11	0.13	0.71	0.00				
	X	t	0.61	6.79	-0.25	-0.16	0.42	-0.19	-0.10	1.84	7.19	1.37			
		p	0.54	0.00	0.81	0.87	0.67	0.85	0.92	0.07	0.00	0.18			
	P[X]	t	-0.07	6.31	-0.98	-0.96	-0.37	-0.99	-0.89	1.19	6.71	0.74	-0.73		
		p	0.94	0.00	0.33	0.34	0.71	0.33	0.38	0.24	0.00	0.46	0.47		
	β_{min}	t	0.59	7.46	-0.36	-0.28	0.38	-0.32	-0.21	1.94	7.94	1.41	-0.09	0.72	
		p	0.56	0.00	0.72	0.78	0.71	0.75	0.83	0.06	0.00	0.16	0.93	0.48	
	β_{max}	t	1.10	8.26	0.14	0.27	0.98	0.24	0.34	2.51	8.80	1.92	0.42	1.28	0.59
		p	0.27	0.00	0.89	0.79	0.33	0.81	0.73	0.01	0.00	0.06	0.68	0.20	0.56

Table A3: NSGA-III GD Metric T-Test Statistics

			Direct Effect					Interaction Effect					
	Param		λ	P[X]	X_{DI}	P[PM]	PM_{DI}	K	λ	P[X]	X_{DI}	P[PM]	PM_{DI}
Direct Effect	P[X]	t	2.62										
		p	0.02										
	X_{DI}	t	7.98	2.53									
		p	0.00	0.02									
	P[PM]	t	8.87	2.83	0.64								
		p	0.00	0.01	0.53								
	PM_{DI}	t	7.61	2.29	-0.56	-1.32							
		p	0.00	0.03	0.58	0.20							
	K	t	10.27	3.63	3.26	5.07	3.95						
		p	0.00	0.00	0.00	0.00	0.00						
Interaction Effect	λ	t	-0.98	-3.64	-13.71	-16.94	-13.06	-20.40					
		p	0.34	0.00	0.00	0.00	0.00	0.00					
	P[X]	t	8.29	2.76	0.55	0.07	1.10	-2.44	14.15				
		p	0.00	0.01	0.59	0.95	0.29	0.03	0.00				
	X_{DI}	t	6.21	2.02	-0.44	-0.84	-0.08	-2.17	8.97	-0.80			
		p	0.00	0.06	0.66	0.41	0.94	0.04	0.00	0.43			
	P[PM]	t	6.14	2.02	-0.40	-0.77	-0.04	-2.06	8.78	-0.75	0.03		
		p	0.00	0.06	0.70	0.45	0.97	0.05	0.00	0.47	0.98		
	PM_{DI}	t	6.55	2.26	-0.05	-0.41	0.32	-1.75	9.45	-0.41	0.32	0.28	
		p	0.00	0.04	0.96	0.68	0.75	0.10	0.00	0.69	0.75	0.78	
	K	t	4.78	1.08	-1.90	-2.37	-1.56	-3.61	6.92	-2.23	-1.22	-1.22	-1.53
		p	0.00	0.30	0.07	0.03	0.14	0.00	0.00	0.04	0.24	0.24	0.14

Table A4: NSGA-III IGD Metric T-Test Statistics

			Direct Effect					Interaction Effect					
	Param		λ	P[X]	X_{DI}	P[PM]	PM_{DI}	K	λ	P[X]	X_{DI}	P[PM]	PM_{DI}
Direct Effect	P[X]	t	-0.16										
		p	0.87										
	X_{DI}	t	3.86	3.92									
		p	0.00	0.00									
	P[PM]	t	4.39	4.42	1.34								
		p	0.00	0.00	0.20								
	PM_{DI}	t	3.86	3.92	-0.20	-1.87							
		p	0.00	0.00	0.85	0.08							
	K	t	4.36	4.40	1.32	0.31	1.65						
		p	0.00	0.00	0.20	0.76	0.12						
Interaction Effect	λ	t	-2.87	-2.52	-24.39	-40.08	-32.61	-23.14					
		p	0.01	0.02	0.00	0.00	0.00	0.00					
	P[X]	t	4.28	4.32	1.09	-0.04	1.43	-0.30	25.81				
		p	0.00	0.00	0.29	0.97	0.17	0.77	0.00				
	X_{DI}	t	2.32	2.43	-1.54	-2.22	-1.49	-2.23	8.06	-2.11			
		p	0.03	0.03	0.14	0.04	0.15	0.04	0.00	0.05			
	P[PM]	t	2.16	2.27	-1.43	-2.01	-1.38	-2.05	6.94	-1.93	-0.06		
		p	0.04	0.04	0.17	0.06	0.18	0.05	0.00	0.07	0.96		
	PM_{DI}	t	2.53	2.64	-1.02	-1.63	-0.96	-1.68	7.94	-1.55	0.33	0.36	
		p	0.02	0.02	0.32	0.12	0.35	0.11	0.00	0.14	0.74	0.72	
	K	t	1.74	1.87	-2.12	-2.74	-2.09	-2.74	6.40	-2.63	-0.58	-0.49	-0.87
		p	0.10	0.08	0.05	0.01	0.05	0.01	0.00	0.02	0.57	0.63	0.40

Table A5: NSGA-III HV Metric T-Test Statistics

			Direct Effect						Interaction Effect					
	Param		λ	P[X]	X_{DI}	P[PM]	PM_{DI}	K	λ	P[X]	X_{DI}	P[PM]	PM_{DI}	
Direct Effect	P[X]	t	19.13											
		p	0.00											
	X_{DI}	t	2.60	-9.16										
		p	0.02	0.00										
	P[PM]	t	2.29	-8.67	-0.14									
		p	0.04	0.00	0.89									
	PM_{DI}	t	2.17	-9.02	-0.27	-0.13								
		p	0.05	0.00	0.79	0.90								
	K	t	1.02	-8.86	-1.06	-0.90	-0.79							
		p	0.33	0.00	0.31	0.38	0.44							
Interaction Effect	λ	t	2.43	-5.12	0.57	0.66	0.77	1.37						
		p	0.03	0.00	0.58	0.52	0.46	0.20						
	P[X]	t	-1.62	–	-4.12	-3.63	-3.52	-2.00	-3.28					
		p	0.13	0.00	0.00	0.00	0.00	0.07	0.01					
	X_{DI}	t	15.88	-10.34	7.32	6.98	7.29	7.37	4.00	68.71				
		p	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00				
	P[PM]	t	14.17	-2.57	7.30	7.02	7.30	7.43	4.22	26.94	1.21			
		p	0.00	0.02	0.00	0.00	0.00	0.00	0.00	0.00	0.25			
	PM_{DI}	t	16.08	-4.21	7.78	7.42	7.74	7.77	4.36	46.06	1.99	0.08		
		p	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.07	0.94		
	K	t	15.30	-1.12	7.99	7.65	7.95	8.00	4.65	29.99	2.75	1.09	1.27	
		p	0.00	0.29	0.00	0.00	0.00	0.00	0.00	0.00	0.02	0.30	0.23	

Table A6: MOEA/D GD Metric T-Test Statistics

			Direct Effect							Interaction Effect						
	Param		λ	P[X]	X_{DI}	P[PM]	PM_{DI}	$Mode$	ϵ_N	λ	P[X]	X_{DI}	P[PM]	PM_{DI}	$Mode$
Direct Effect	P[X]	t	9.23												
		p	0.00												
	X_{DI}	t	50.53	1.92											
		p	0.00	0.07											
	P[PM]	t	25.00	0.87	-2.24										
		p	0.00	0.40	0.04										
	PM_{DI}	t	12.55	0.01	-2.55	-1.10									
		p	0.00	0.99	0.02	0.29									
	$Mode$	t	11.84	0.05	-2.33	-0.98	0.05								
		p	0.00	0.96	0.03	0.34	0.96								
	ϵ_N	t	19.16	1.08	-1.26	0.43	1.33	1.21							
		p	0.00	0.30	0.22	0.67	0.20	0.24							
Interaction Effect	λ	t	1.96	-2.76	-4.22	-3.59	-2.93	-2.93	-3.68						
		p	0.07	0.01	0.00	0.00	0.01	0.01	0.00						
	P[X]	t	9.63	0.07	-1.88	-0.79	0.08	0.03	-1.01	2.83					
		p	0.00	0.94	0.08	0.44	0.94	0.98	0.33	0.01					
	X_{DI}	t	8.66	-0.58	-2.78	-1.65	-0.67	-0.70	-1.82	2.39	-0.66				
		p	0.00	0.57	0.01	0.12	0.51	0.49	0.09	0.03	0.52				
	P[PM]	t	8.11	0.22	-1.36	-0.48	0.23	0.19	-0.68	2.80	0.15	0.75			
		p	0.00	0.83	0.19	0.64	0.82	0.85	0.50	0.01	0.88	0.46			
	PM_{DI}	t	9.52	0.46	-1.23	-0.24	0.51	0.46	-0.48	3.06	0.40	1.03	0.21		
		p	0.00	0.65	0.24	0.81	0.61	0.65	0.64	0.01	0.70	0.31	0.84		
	$Mode$	t	4.70	-0.93	-2.31	-1.63	-1.00	-1.02	-1.76	1.64	-0.99	-0.50	-1.05	-1.27	
		p	0.00	0.37	0.03	0.12	0.33	0.32	0.09	0.12	0.34	0.62	0.31	0.22	
	ϵ_N	t	9.21	-0.45	-2.69	-1.52	-0.52	-0.56	-1.70	2.50	-0.53	0.14	-0.63	-0.91	0.61
		p	0.00	0.66	0.02	0.15	0.61	0.58	0.11	0.02	0.60	0.89	0.54	0.37	0.55

Table A7: MOEA/D IGD Metric T-Test Statistics

			Direct Effect						Interaction Effect					
	Param		λ	P[X]	X_{DI}	P[PM]	PM_{DI}	*Mode*	ϵ_N	λ	P[X]	X_{DI}	P[PM]	PM_{DI}	*Mode*
Direct Effect	P[X]	t	0.77												
		p	0.45												
	X_{DI}	t	1.64	1.02											
		p	0.12	0.32											
	P[PM]	t	-0.43	-1.54	-3.02										
		p	0.68	0.14	0.01										
	PM_{DI}	t	-0.59	-1.71	-3.15	-0.22									
		p	0.56	0.10	0.01	0.83									
	Mode	t	2.27	1.85	1.15	3.69	3.80								
		p	0.04	0.08	0.26	0.00	0.00								
	ϵ_N	t	1.27	0.56	-0.53	2.40	2.56	-1.54							
		p	0.22	0.58	0.60	0.03	0.02	0.14							
Interaction Effect	λ	t	0.92	0.31	-0.41	1.51	1.64	-1.08	-0.09						
		p	0.37	0.76	0.69	0.15	0.12	0.29	0.93						
	P[X]	t	0.26	-0.57	-1.57	0.81	0.98	-2.28	-1.14	-0.75					
		p	0.80	0.58	0.13	0.43	0.34	0.03	0.27	0.46					
	X_{DI}	t	0.65	-0.24	-1.57	1.54	1.72	-2.45	-0.96	-0.51	0.42				
		p	0.52	0.81	0.13	0.14	0.10	0.02	0.35	0.62	0.68				
	P[PM]	t	2.19	1.75	1.03	3.57	3.68	-0.11	1.42	1.01	2.19	2.33			
		p	0.04	0.10	0.32	0.00	0.00	0.92	0.17	0.33	0.04	0.03			
	PM_{DI}	t	1.63	1.03	0.05	2.94	3.07	-1.06	0.55	0.42	1.56	1.54	-0.94		
		p	0.12	0.32	0.96	0.01	0.01	0.30	0.59	0.68	0.14	0.14	0.36		
	Mode	t	-1.08	-2.31	-3.85	-0.88	-0.64	-4.43	-3.23	-2.12	-1.54	-2.41	-4.31	-3.75	
		p	0.29	0.03	0.00	0.39	0.53	0.00	0.00	0.05	0.14	0.03	0.00	0.00	
	ϵ_N	t	0.58	-0.31	-1.58	1.39	1.57	-2.43	-1.01	-0.56	0.33	-0.09	-2.31	-1.55	2.24
		p	0.57	0.76	0.13	0.18	0.13	0.03	0.33	0.58	0.74	0.93	0.03	0.14	0.04

Table A8: MOEA/D HV Metric T-Test Statistics

			Direct Effect						Interaction Effect					
	Param		λ	P[X]	X_{DI}	P[PM]	PM_{DI}	*Mode*	ϵ_N	λ	P[X]	X_{DI}	P[PM]	PM_{DI}	*Mode*
Direct Effect	P[X]	t	-1.27												
		p	0.23												
	X_{DI}	t	-1.38	-0.15											
		p	0.19	0.88											
	P[PM]	t	0.71	2.32	2.41										
		p	0.49	0.04	0.03										
	PM_{DI}	t	1.82	3.44	3.47	1.77									
		p	0.09	0.00	0.00	0.10									
	Mode	t	-2.67	-1.59	-1.43	-3.68	-4.56								
		p	0.02	0.13	0.17	0.00	0.00								
	ϵ_N	t	-1.12	0.17	0.32	-2.19	-3.34	1.76							
		p	0.28	0.87	0.75	0.05	0.00	0.10							
Interaction Effect	λ	t	-1.68	-0.58	-0.44	-2.57	-3.47	0.93	-0.74						
		p	0.12	0.57	0.67	0.02	0.00	0.37	0.47						
	P[X]	t	-0.60	0.82	0.96	-1.70	-3.07	2.40	0.65	1.32					
		p	0.56	0.42	0.35	0.11	0.01	0.03	0.52	0.21					
	X_{DI}	t	-0.14	1.28	1.40	-1.05	-2.38	2.76	1.12	1.70	0.53				
		p	0.89	0.22	0.18	0.31	0.03	0.02	0.28	0.11	0.60				
	P[PM]	t	-2.00	-0.68	-0.50	-3.36	-4.56	1.07	-0.87	0.01	-1.64	-2.10			
		p	0.07	0.51	0.62	0.00	0.00	0.30	0.40	0.99	0.12	0.05			
	PM_{DI}	t	-2.95	-1.80	-1.62	-4.17	-5.14	-0.08	-1.99	-1.06	-2.71	-3.10	-1.25		
		p	0.01	0.09	0.13	0.00	0.00	0.93	0.07	0.31	0.02	0.01	0.23		
	Mode	t	1.70	3.28	3.32	1.55	-0.10	4.43	3.17	3.35	2.86	2.21	4.35	4.97	
		p	0.11	0.01	0.01	0.14	0.92	0.00	0.01	0.00	0.01	0.04	0.00	0.00	
	ϵ_N	t	-0.31	1.16	1.28	-1.37	-2.81	2.70	0.99	1.61	0.36	-0.20	2.02	3.04	-2.60
		p	0.76	0.27	0.22	0.19	0.01	0.02	0.34	0.13	0.72	0.85	0.06	0.01	0.02

B Clustering Analysis of Hyperparameters and Testbench

We present a clustering analysis of hyperparameters performance. For this analysis, we take direct and interaction (total) influence values together in a matrix form and apply the k-means algorithm (Lloyd, 1982). The number of clusters was automatically chosen based on silhouette scores (Rousseeuw, 1987) as per formula: K = max{2, argmax(silhouette scores)}. This formula selects a value of K equal to 2 or more, for which K-means clustering produces maximum silhouette score. We analyze algorithms performance for hyperparameters and functions in the testbench for each sampling and metric used.

We hypothesized that the hyperparameters that have a similar influence on the performance of algorithms tend to cluster together, and similarly, the functions that have similar characteristics may cluster together. Moreover, we may find the number of clusters and the trajectory of clustering based on silhouette score also indicating differences in characteristics as a silhouette score defines how well the clusters are separated from each other and how good is the proximity of data points within a cluster. Figures B1, B2, and B3 are hyperparameters clustering. Whereas Figs. B4, B5, and B6 are functions clusters. The x and y axes are two principal components of the matrix formed by direct and interaction (total) influence values in these plots.

Hyperparameters performance characteristics. The hyperparameters clustering for single-objective algorithms in Fig. B1 shows that hyperparameter $\mathbf{b}_{\text{type}}$ of DE has clear, distinct influence characterization, whereas all other hyperparameters tend to cluster in a similar group. For CMA-ES, different sampling methods tend to cluster hyperparameters differently, where hyperparameter λ shows more distinct behavior than others (see Fig. B1). We may also observe the line of silhouette score that represents how optimal clustering number varies as the number of clusters gradually tend to go to the maximum number of data points characterizing variation in the similarity and dissimilarity between hyperparameters due to different sampling methods and metrics.

Multi-objective clustering results for MOEA/D in Fig. B3 shows that hyperparameter influence vary a lot (mostly on HV metric and less on IGD metric), and only a few clusters can be seen, such as λ and the hyperparameters that are related also appear to have high proximity such as $\{P[\text{X}]$ and $\text{X}_{\text{DI}}\}$ and $\{P[\text{PM}]$ and $\text{PM}_{\text{DI}}\}$. For NSGA-III, λ on GD metric has a separate identity (see Fig. B2). However, there are 2-3 groups of hyperparameters.

Functions performance characteristics. When analyzing functions clustering characteristics, we may observe that (although difficult to read each individual number due to overlapping points) each hyperparameters sampling method has different influences on the performance of the algorithms. Not only that, different algorithms for different metrics also produce a variety of clustering of functions. However, one may be able to group functions into a few clusters, as evident from Figs B4, B6, and B5.

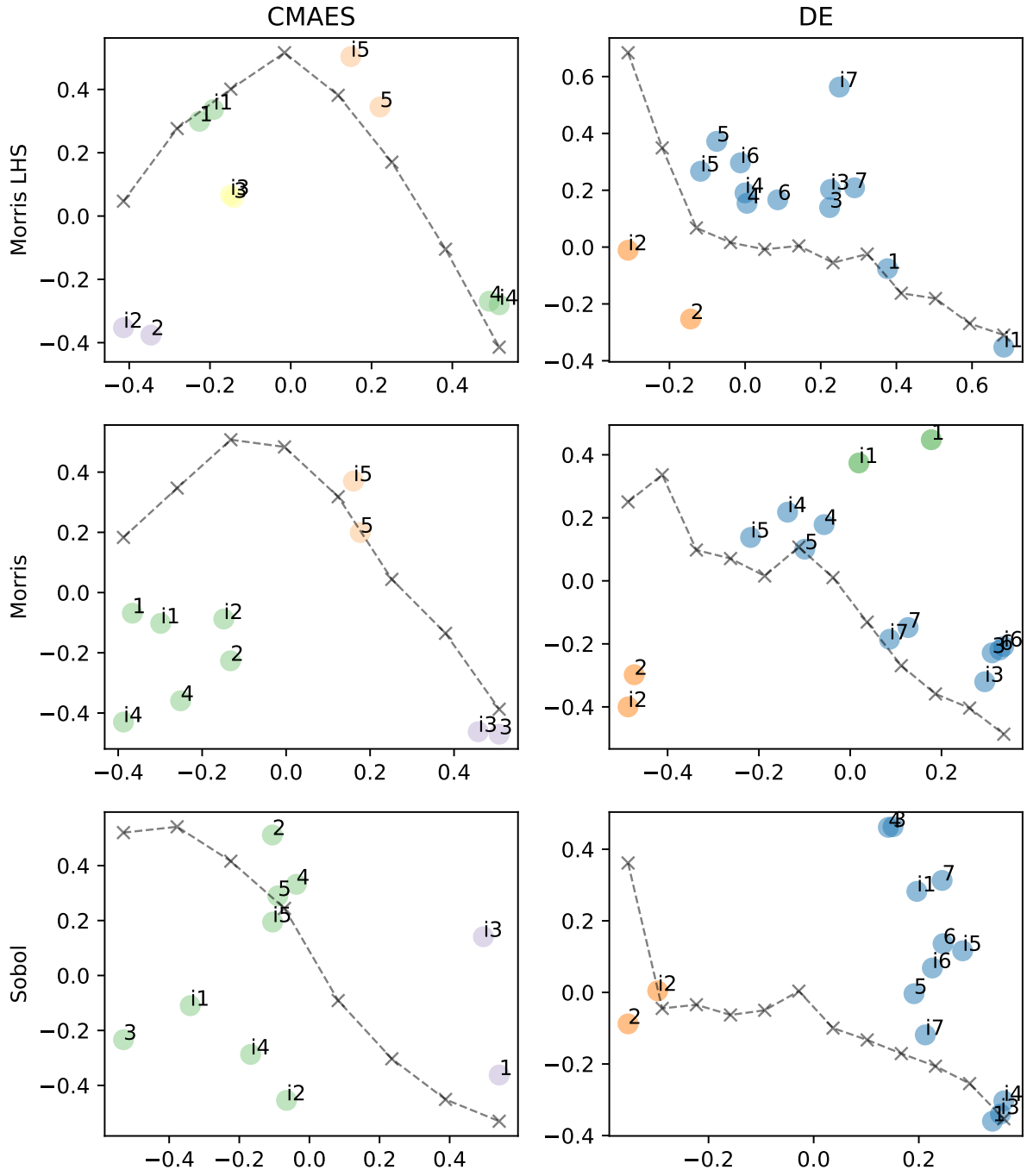

Fig. B1: DE and CMA-ES algorithm hyperparameters K-means influence clustering. K is chosen automatically by using the formula K = max{2, argmax(silhouette scores)}, which is the cross at the highest peak of the dotted line. CMA-ES hyperparameters numbered from 1 to 5 are λ, $\mu\lambda_{\text{ratio}}$, σ_0, and α_μ, $\sigma_{0-scale}$, respectively and DE hyperparameters numbered from 1 to 7 are λ, $\mathbf{b}_{\text{type}}$, $\mathbf{b}\lambda_{\text{ratio}}$, X, P[X], β_{min}, and β_{max}, respectively. A hyperparameter number appearing twice indicates the values of direct and interaction effect. A letter 'i' as a prefix to a number in the plot indicates interaction/total effect, and the of 'i' indicates direct effect. Different colors represent different clusters.

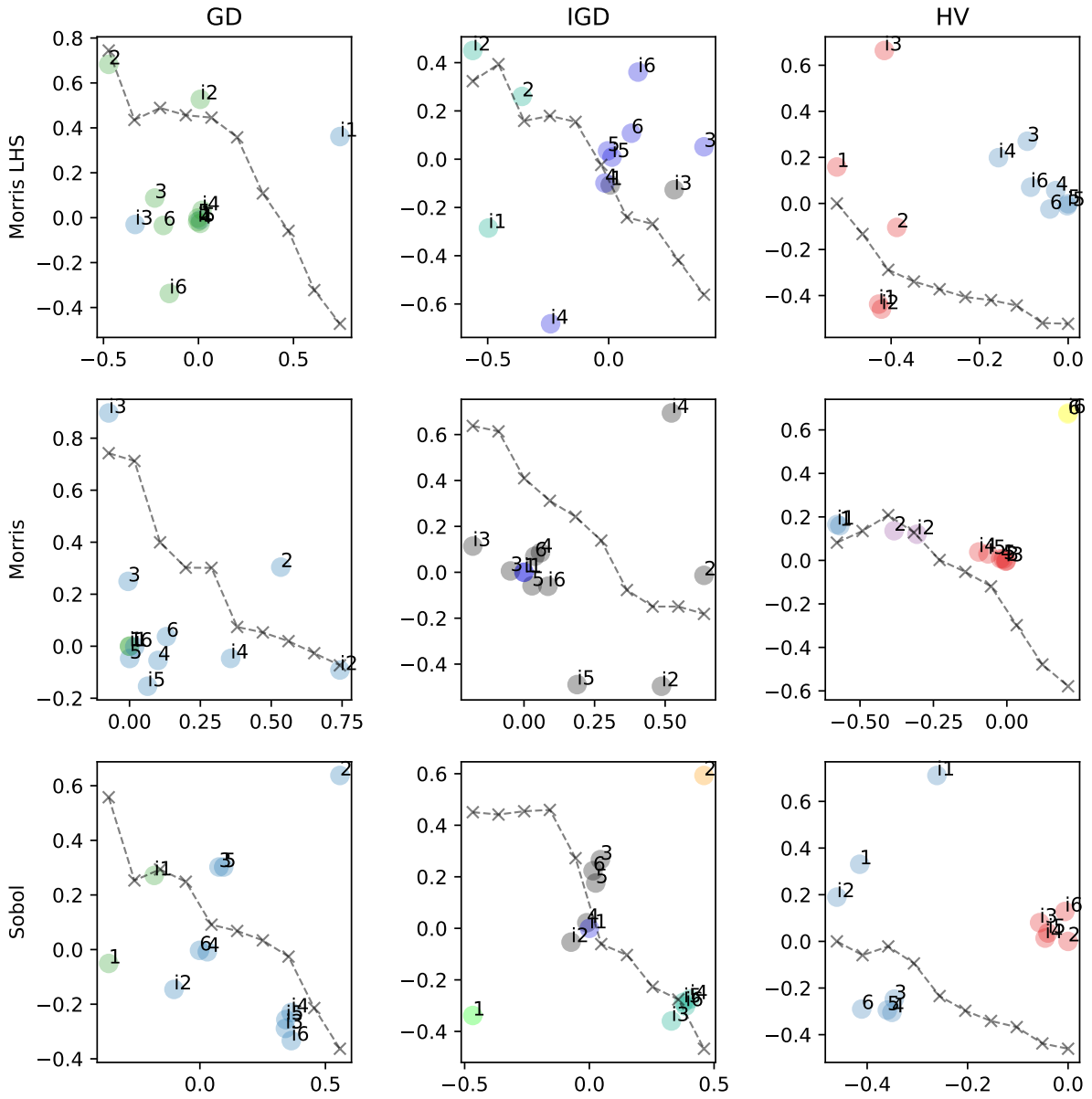

Fig. B2: NSGA-III algorithm hyperparameters K-means influence clustering. K is chosen automatically by using the formula K = max{2, argmax(silhouette scores)}, which is the cross at the highest peak of the dotted line. NSGA-III hyperparameters numbered from 1 to 6 are λ, P[X], $\mathrm{X_{DI}}$, P[PM], $\mathrm{PM_{DI}}$, and K, respectively. A hyperparameter number appearing twice indicates the values of direct and interaction effect. A letter 'i' as a prefix to a number in the plot indicates interaction/total effect, and the of 'i' indicates direct effect. Different colors represent different clusters.

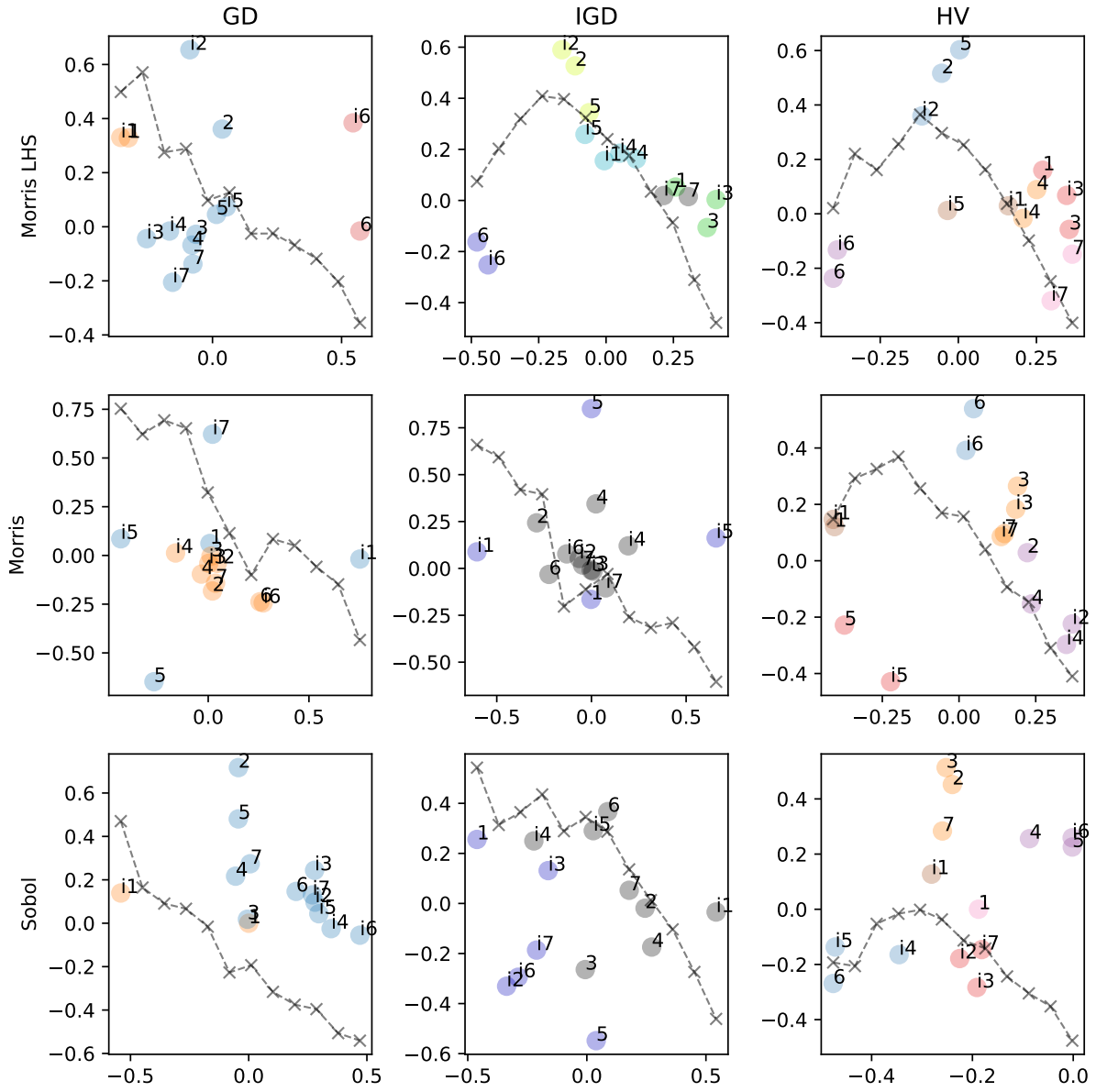

Fig. B3: MOEA/D algorithm hyperparameters K-means influence clustering. K is chosen automatically by using the formula K = max{2, argmax(silhouette scores)}, which is the cross at the highest peak of the dotted line. MOEA/D hyperparameters numbered from 1 to 7 are λ, P[X], X_{DI}, P[PM], $\mathrm{PM}_{\mathrm{DI}}$, $Mode$, and ϵ_N, respectively. A hyperparameter number appearing twice indicates the values of direct and interaction effect. A letter 'i' as a prefix to a number in the plot indicates interaction/total effect, and the of 'i' indicates direct effect. Different colors represent different clusters.

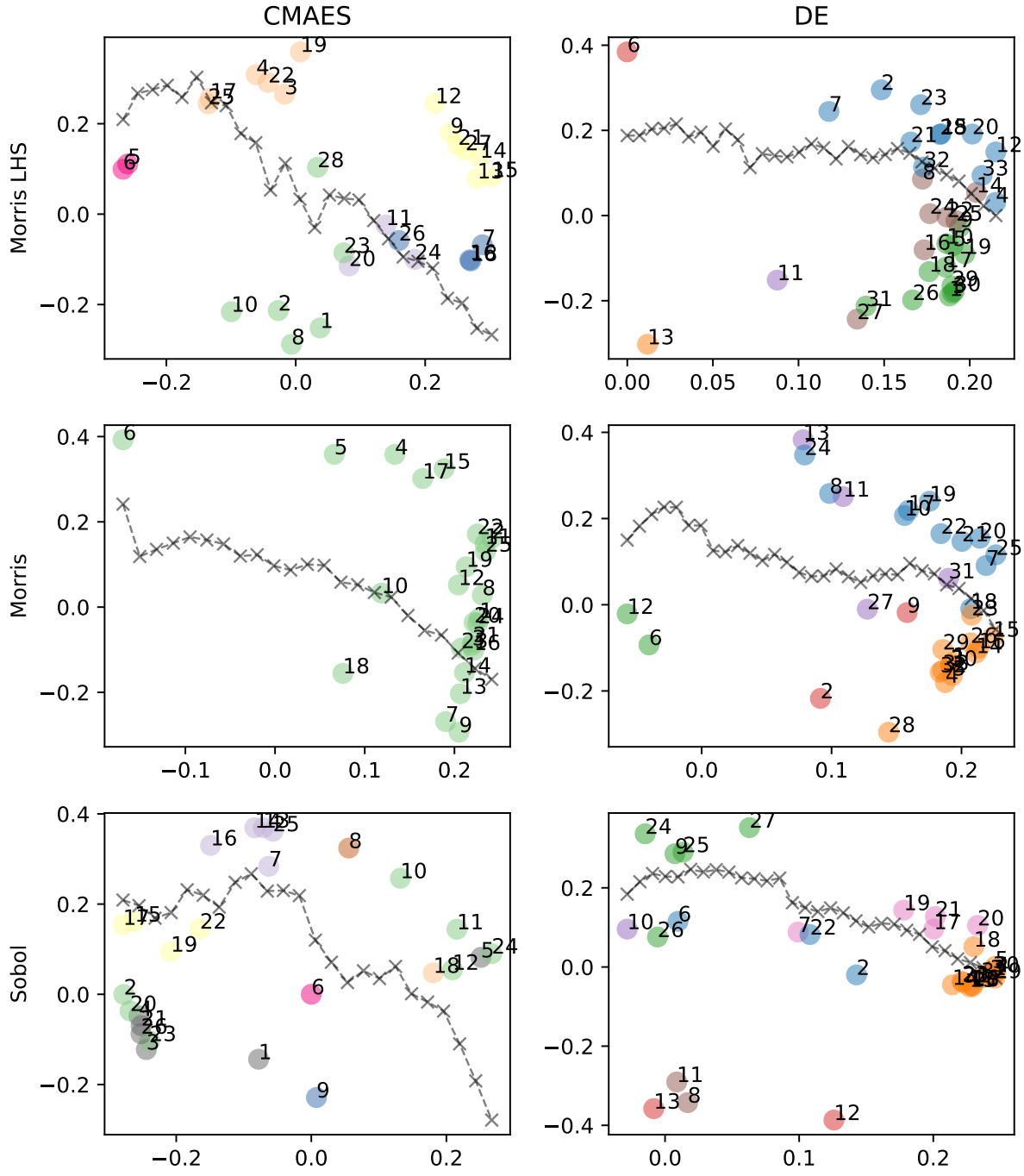

Fig. B4: DE and CMA-ES function clustering based on hyperparameters influence on them. The numbers 1 to 23 are of the function in (Yao et al., 1999), 24 to 27 are shifted functions (Liang et al., 2013) Sphere, Ellipsoid, Ackley, and Griewank; and 28 to 33 are shifted and rotated functions (Liang et al., 2013, 2014) Rosenbrock, and Rastrigin, Weierstrass, Schwefel, Katsuura, and HappyCat. Different colors represent different clusters.

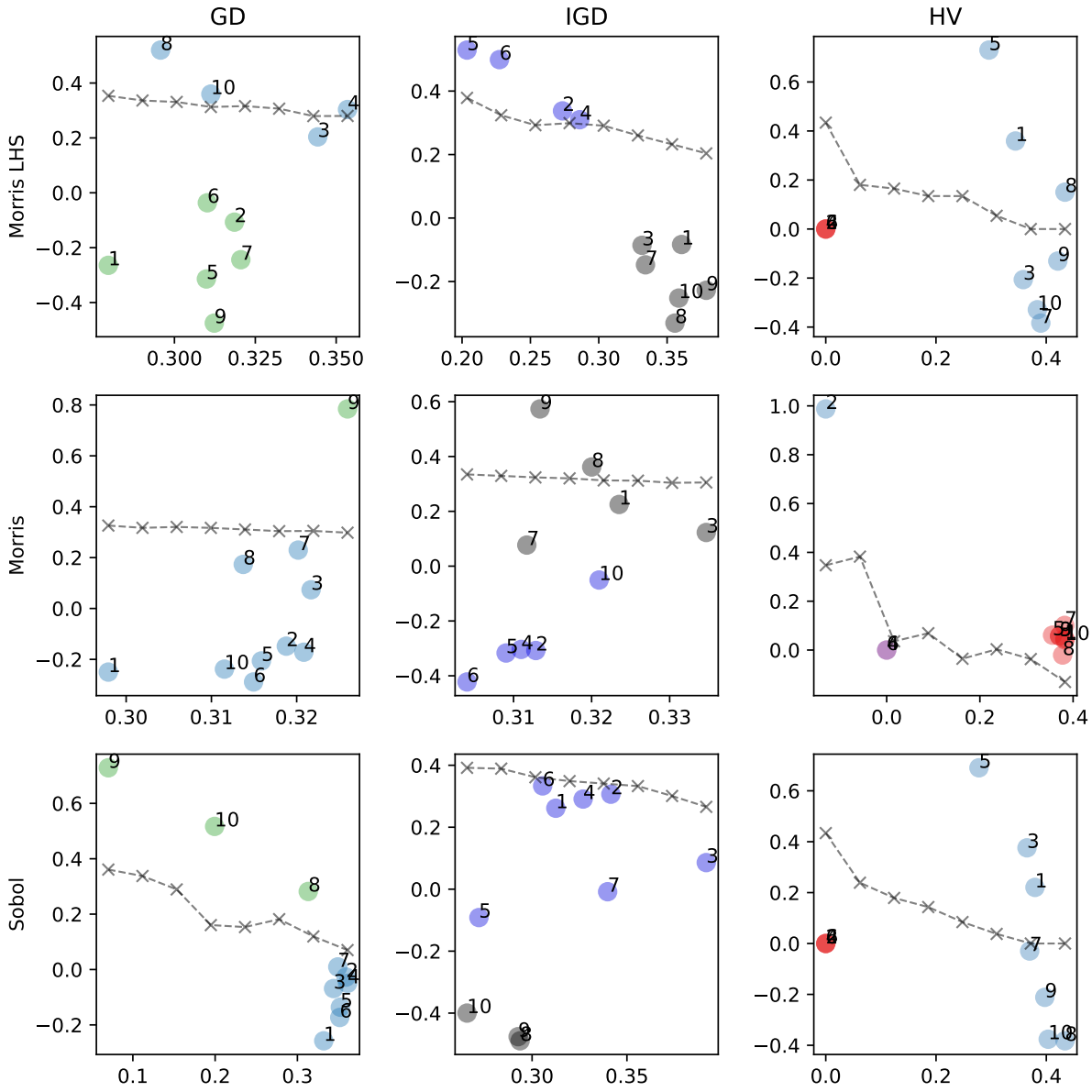

Fig. B5: NSGA-III performance on multi-objective functions. Clustering based on the influence of the hyper-parameters on the functions. The numbers 1 to 10 represent optimization problems CDTLZ2, DTLZ1, DTLZ2, DTLZ3, DTLZ4, IDTLZ1, IDTLZ2, WFG3, WFG6, and WFG7, respectively. Different colors represent different clusters.

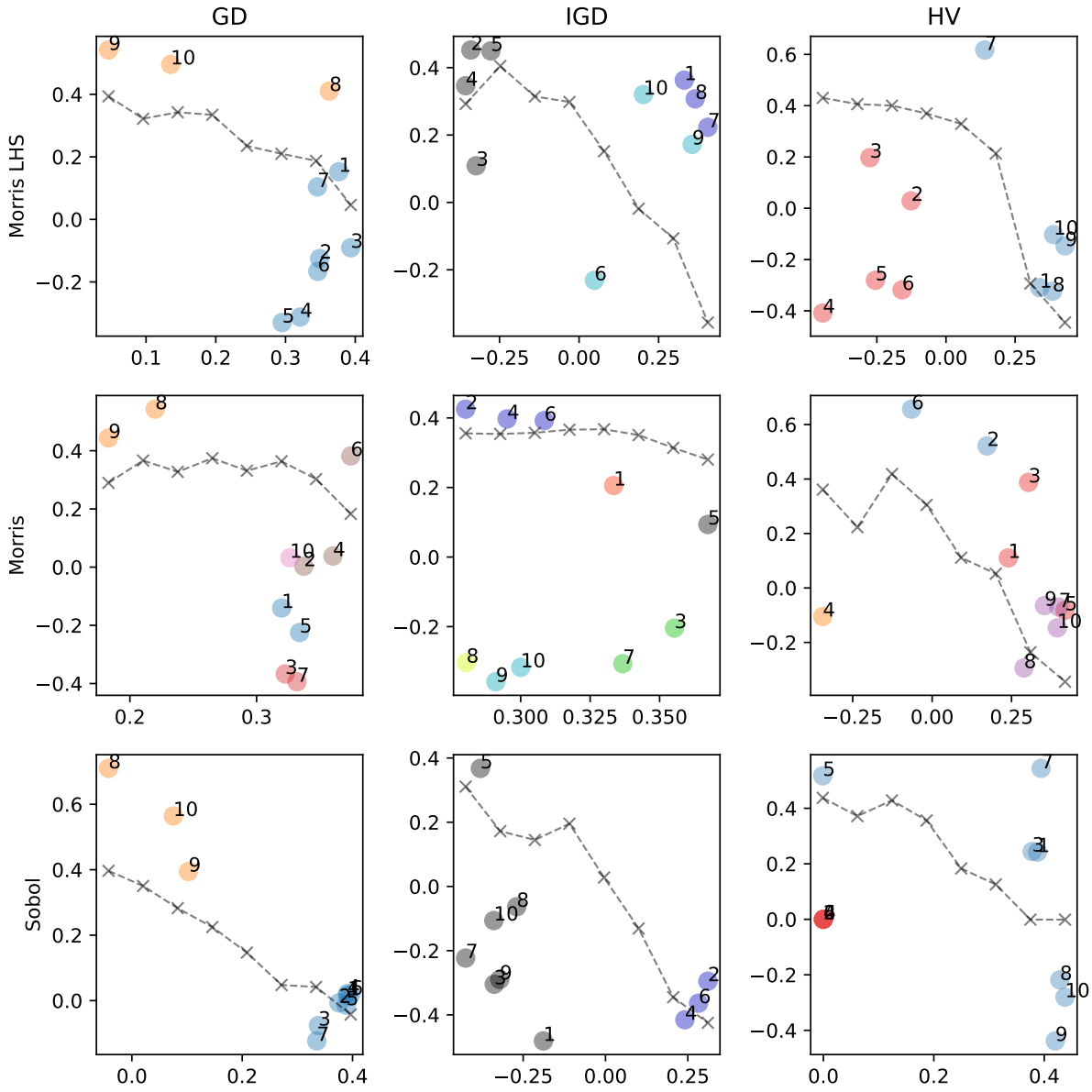

Fig. B6: MOEA/D performance on multi-objective functions. Clustering based on the influence of the hyperparameters on the functions. The numbers are 1 to 10 represent optimization problems CDTLZ2, DTLZ1, DTLZ2, DTLZ3, DTLZ4, IDTLZ1, IDTLZ2, WFG3, WFG6, and WFG7, respectively. Different colors represent different clusters.