Towards Standardized Evaluation of Feasible Region Identification in Constrained Engineering Design

Ioana Nikova
Siemens Digital Industries Software
Ghent University
Leuven, Belgium
ioana.nikova@siemens.com

Yashesh Dhebar
Siemens Digital Industries Software
East Lansing, United States
yashesh.dhebar@siemens.com

Sebastian Rojas Gonzalez
Ghent University - imec
Ghent, Belgium
sebastian.rojasgonzalez@ugent.be

Tom Dhaene
Ghent University - imec
Ghent, Belgium
tom.dhaene@ugent.be

Ivo Couckuyt
Ghent University - imec
Ghent, Belgium
ivo.couckuyt@ugent.be

***Abstract*—Many benchmarks exist for constrained optimization, which is crucial when it comes to comparing constrained *optimization* algorithms. However, in the field of Feasible Region Identification (FRI), where we are solely interested in discovering and sampling in feasible regions, no standardized way of benchmarking exists. Therefore, we propose a set of constrained problems that cover feasibility properties like number, location, shape, and size of feasible regions, including presence of convex or non-convex regions. Another challenge in FRI is the lack of a standard way to measure FRI performance. In case of model-based FRI methods, the model's performance, e.g., accuracy, is often used to gauge the algorithm's performance. However, we show that this overlooks the fundamental aspect in real world use cases where engineers are interested in both: a good model and a *diverse set* of feasible designs returned by the algorithm. Therefore, we introduce a new performance metric based on Inverted Generational Distance (IGD) that measures the diversity of the acquired feasible points. By providing a collection of benchmark functions and metrics, we aim to contribute towards a more standardized approach for FRI benchmarking. To demonstrate the benchmark collection, we compare five state-of-the-art data-efficient and model-based FRI approaches to show their performance on the proposed benchmark.**

***Index Terms*—feasible region identification, constraint optimization, benchmarking, performance metrics.**

I. Introduction

Rigorous benchmarking is essential for characterizing the performance of black-box learning and optimization [1]. Recent comparative studies [2], [3] demonstrate the practical value of standardized evaluation, using rigorous performance profiles to clearly quantify its impact. However, most of this research has focused on black-box *optimization*. In this work, we bring the same benchmarking perspective to *feasible region identification* (FRI), a setting with different goals and success criteria: rather than searching for optima, the aim is to characterize where constraints are satisfied.

FRI is valuable early in the design process when engineers explore diverse solutions satisfying design requirements. These requirements, represented by design constraints, define one or more feasible regions. Unlike optimization seeking a single (or multiple in case of multi-objective or multi-modal optimization) optimum, FRI methods primarily focus on identifying a set of points, or better yet, regions where the constraints are satisfied. More formally, in a d-dimensional design space, we look for designs $\mathbf{x} \in \mathcal{R}^d : G(\mathbf{x}) = \big(g_1(\mathbf{x}), g_2(\mathbf{x}), \ldots, g_L(\mathbf{x})\big)^\top \leq (t_1, t_2, \ldots, t_L)^\top$, where $g_l : \mathcal{R}^d \to \mathcal{R}$ is the l-th constraint function of the problem and t_l represents its threshold.

FRI commonly uses Bayesian active learning (AL) [4], an iterative and data-efficient sampling approach that uses a Bayesian model, often a Gaussian Process (GP) [5], to learn the expensive constraint functions. An acquisition function uses the prediction of the model as well as uncertainty to make an informed decision about where to sample next, and thus run a time-consuming simulation. Note that this is closely related to Bayesian Optimization [6], [7], where the acquisition function is created specifically for finding the optimum. Similarly, FRI acquisition functions target feasible regions.

In literature, this setting has also been studied under the names *Level Set Estimation (LSE)* [8] and *Constrained Active Search (CAS)* [9]. While the underlying problem setting is identical, the goals and acquisition functions differ. LSE focuses on identifying the boundary between feasible and infeasible designs. We will refer to these acquisition functions as *boundary-based*. In CAS, on the other hand, we focus on identifying a diverse set of solutions within the feasible region, and thus returning more feasible designs to the user. While both aspects are crucial in the broader context captured by FRI, in most real world applications the end user (e.g. a

This work is funded by the Flemish Agency on Innovation and Entrepreneurship (VLAIO) under the Baekeland Research Project HBC.2021.0841, and by the Research Foundation Flanders (FWO) under grant #12AZF24N. This work was supported by the Flemish Government under the "Onderzoeksprogramma Artificiële Intelligentie (AI) Vlaanderen".

design engineer) is interested in having a diverse sampling of feasible designs. This necessitates a need to have a reliable metric which can quantify the *sampling-quality* (or coverage) of the FRI method. Common choices for performance metrics are mostly based on the performance of the model. While this is good for measuring how well the boundary has been found, it does not tell us anything about the diversity of the selected feasible points. Therefore, we propose a new performance metric based on Inverted Generational Distance (IGD), which we call Feasible Region IGD (or FR-IGD) with an aim to reliably quantify the *sampling-quality* of FRI methods to sample feasible points.

We provide a diverse benchmark suite[1] tailored to FRI problems, and more specifically, for testing acquisition functions that not only learn the constraints, but also find and sample diversely in all feasible regions.

II. Related Work

A. Performance Metrics for Feasible Region Identification

Various metrics have been used in the literature to evaluate FRI algorithms. These are broadly categorized into two groups: a) model performance metrics, b) dataset metrics. Model performance metrics assess how well a given predictive model of the FRI method can distinguish designs as feasible or infeasible (a binary-classification task). Commonly used metrics are F1-score [4], informedness [10] and Matthew's correlation coefficient (MCC) [11], [12], from which the latter two are better suited for imbalanced datasets, an issue often encountered in FRI. While the previous performance metrics take the predictions into account, the uncertainty is not considered. The Brier score [13], however, does take the probabilistic predictions into account and measures the mean squared difference between predicted probabilities and the actual outcomes.

Good models are crucial for FRI since acquisition depends on them. Yet, engineers often use the acquired set of feasible designs for further steps in the design process. Therefore, it is important to also make sure the acquired data contains enough feasible designs, and ultimately represents the feasible regions well. Previous work has already looked at the proportion of feasible designs or the number of feasible designs in the final dataset [8], [9], [12]. Nevertheless, this does not give any information on the distribution of feasible designs in the design space. Therefore, *coverage-based metrics* are important to consider. With this in mind, the fill-distance and coverage-recall were introduced by Malkomes et al. [9] to measure how well the feasible region is covered by the acquired feasible designs. However, computing fill-distance is NP-hard, requiring feasible region approximation. Another challenge with the fill-distance occurs when we have multiple disjoint feasible regions. In that case, we need to make sure that the largest empty ball the fill-distance is trying to find, is actually situated inside the feasible region (and not between two regions). These limitations motivate the need for a better coverage metric.

Note that algorithms focusing on finding the boundary of the feasible region are not expected to perform well on coverage metrics, as they do not aim to cover the whole feasible region. Often they will perform well on model performance metrics as they learn to distinguish feasible from infeasible quite well by learning the boundary. Nevertheless, coverage-based metrics are important to consider when the goal is to acquire a representative set of feasible designs which is diverse and covers all disjoint feasible regions.

B. Benchmarks for Feasible Region Identification

Most FRI works propose their own test problems to evaluate the performance of their algorithms. There are two types of use cases. First, 2D-3D problems with known feasible regions are used, allowing for easy visualization and interpretation of the results. Second, engineering problems from optimization benchmarks (i.e., analytical functions) or real-world applications (i.e., simulation-based) are used to demonstrate the applicability of proposed algorithms. Engineering problems can have complex feasible regions that, especially in higher dimensions, are difficult to understand. So, there is a lack of interpretable and higher dimensional problems. As we said before, there is no standardized set of benchmark problems for FRI that covers different scenarios in terms of *number* and *shape* of feasible region(s), making it difficult to compare the performance of different algorithms across studies.

In this work, we address two key gaps in the field of FRI, namely a) need for a reliable, computationally efficient and scalable coverage metric, and b) interpretable, controllable[2], scalable and computationally cheap benchmarks. We attempt to address this with our proposed FR-IGD metric and benchmarks.

III. Feasible Region Inverted Generational Distance (FR-IGD) Metric

The Inverted Generational Distance (IGD) [14] is a performance metric commonly used in multi-objective optimization to evaluate the quality of a set of Pareto-optimal solutions. It measures the average distance from a set of reference points (representing the true Pareto front) to the nearest point in the obtained solution set. The IGD metric provides insights into both convergence and diversity of the solution set. A lower IGD value indicates that the solution set is closer to the true Pareto front and better represents its diversity. IGD is calculated as:

$$IGD(S^*, S) = \frac{1}{|S^*|} \sum_{p^* \in S^*} \min_{p \in S} d(p^*, p) \tag{1}$$

where:

- S^* is the set of reference points,
- S is the set of obtained solutions from the algorithm,

[1]Code and supplementary material available at: https://github.com/ioananikova/FRIbenchmark.

[2]By *controllable* we mean that we can modulate the number, shape and location of feasible regions.

- $d(p^*, p)$ is the distance between a reference point p^* and an obtained solution p,
- $|S^*|$ is the number of reference points.

As we are not dealing with multi-objective optimization problems, we adapt the IGD metric to our context of feasible region identification and call it Feasible Region IGD (FR-IGD). Here, the reference set S^* represents a set of points that uniformly cover the true feasible region, while the obtained solution set S consists of the feasible points acquired by the algorithm. The distance $d(p^*, p)$ is calculated using the Euclidean distance in the normalized design space. By using IGD in this context, we can assess how well the identified feasible points represent the true feasible region in terms of both coverage and proximity.

A. Dealing With Empty Feasible Solution Sets

For highly constrained problems, early AL iterations may find no feasible solutions. In such cases, the IGD metric cannot be calculated directly, as there are no feasible points in the obtained solution set S. As such, we propose a modified approach to calculate the FR-IGD when the feasible solution set is empty. Instead of using feasible points, we consider using *infeasible points*. We calculate the distance from each reference point in S^* to the nearest infeasible point in S. However, to ensure that the metric reflects absence of feasible points, we introduce a penalty term. The modified FR-IGD that accounts for empty feasible solution sets is defined as follows:

$$IGD_{FR}(S^*, S) = \begin{cases} IGD(S^*, S_F) & \text{if } S_F \neq \emptyset \\ IGD(S^*, S_{IF}) + \sqrt{d} & \text{if } S_F = \emptyset \end{cases} \tag{2}$$

where S_F and S_{IF} are sets of feasible and infeasible points in the obtained solution set, respectively, and $S = S_F \cup S_{IF}$. The penalty term $\sqrt{d}$ is based on the diagonal length of the normalized design space, i.e., the maximum possible distance between two points in the d-dimensional unit hyper-cube.

B. Reference Sets

As indicated in the previous section, computation of FR-IGD requires a reference set S^*, which uniformly covers the true feasible region. In order to generate such a reference set, we first sample a large number of points in the design space (e.g., 20,000 points) using Sobol sampling and only keep the feasible points. Next, we filter out points that lead to an over-representation of certain areas in the feasible region. This is done by calculating pairwise distances between all feasible points and removing the ones that are closer than a certain threshold to each other. The threshold[3] is determined based on the desired density of points in the reference set. As a result, we obtain one reference set for each problem covering all feasible regions. Thus, the FR-IGD value can reliably measure the *sampling-quality* of FRI algorithm.

[3]In this paper we chose the threshold to be 0.005 in the normalized design space for all problems.

IV. Benchmark Problems

Our benchmark includes lower-dimensional problems (upto 2D) that are easy to visualize and interpret, and higher dimensional problems (upto 7D) that are more representative of real-world engineering applications. An overview of the benchmark problems and their properties is given in Table I. Problems T1 to T11 are created using (known) test functions so that we are certain about the number, location and shape of feasible regions. Problems E1 to E3 are engineering problems commonly used for optimization, adapted to FRI.

Using appropriate parameterization, we can derive variants of problems T6, T9, T10 and T11. By modifying their dimension and thresholds, the size and number of feasible regions can change. In this paper, we focus on the basic versions as listed in Table I.

As indicated in the *Feasible Region Properties* columns, constraint problems in our benchmarking suite have:

- Number of feasible regions (#) varying from 1 to 5,
- Shapes being either convex, non-convex or mix of both (e.g., problem T5) as shown in Figure 1, and
- Feasibility Ratio (ρ) (expressed as a percentage of domain covered by total feasible region and computed using reference set) from as low as 0.11% to as high as 41.67%.

More disjoint feasible regions, non-convex shapes, and low feasibility ratios increase the difficulty. Table I also shows constraint properties in terms of number of constraints (given by L), if the constraints are linear or non-linear, and also if there are redundant constraints that do not actively define the feasible region. A problem may have one constraint but multiple feasible regions *or* multiple constraints defining only one feasible region. The suggested benchmarking suite includes constraint problems with varying *degree* (and *type*) of complexities and can thus serve as a good test-bench for FRI methods.

V. Results and Discussion

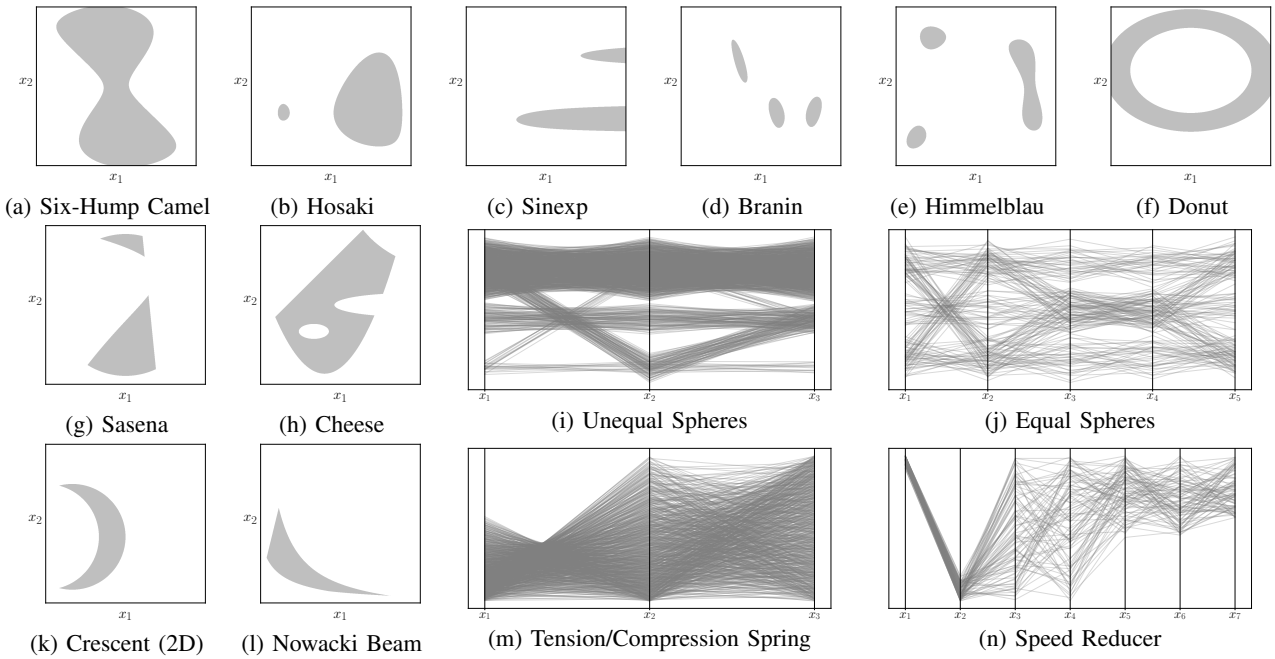
We compare five FRI acquisition functions: *Probability of being at the Boundary and Entropy (PBE)* [10] and *Entropy Feasible (EF)* [4] which are both boundary-based, *Probability of Feasibility and Variance (PoFV)* [21] and *Expected Coverage Improvement (ECI)* [9], both coverage-based, and lastly the scalarization-based *Augmented Tchebysheff (ATCH)* [12] that balances the exploration-exploitation trade-off explicitly. See associated papers for details on the acquisition functions. This selection of acquisition functions will showcase how the performance in terms of model accuracy and dataset diversity varies across different types of acquisition functions.

We ran all five acquisition functions on the benchmark suite using *botorch* [22] and GPs with a Matèrn kernel to model the constraints. Experiments were repeated 10 times with different initial samples according to a Sobol sequence. The initial dataset size and number of AL iterations were chosen depending on the dimensionality and complexity of the problems.

Performance of these FRI methods is compared in terms of the MCC score of the models, the number of feasible points in

TABLE I: Overview of the properties of the benchmark problems and the size of their reference sets. L is number of constraints, ρ is feasibility ratio and $|S^*|$ is the size of the created reference set.

Problem			Constraint Properties			Feasible Region Properties			FR-IGD
id	Name	d	L	(Non-)linear?	Redundant	#	# Convex/Non-convex	$\rho(\%)$	$\lvert S^*\rvert$
T1	Six-Hump Camel [15]	2	1	non-linear	0	1	0/1	41.67	5908
T2	Hosaki [4], [16]	2	1	non-linear	0	2	2/0	20.23	2857
T3	Sinexp	2	1	non-linear	0	2	2/0	10.59	1509
T4	Branin [4], [16]	2	1	non-linear	0	3	3/0	4.13	600
T5	Himmelblau [17]	2	1	non-linear	0	3	2/1	9.42	1345
T6	Donut	2	2	non-linear	0	1	0/1	32.45	4614
T7	Sasena [15]	2	3	mixed	0	2	0/2	13.61	1934
T8	Cheese [18]	2	6	mixed	1	1	0/1	32.51	4609
T9	Unequal Spheres	3	1	non-linear	0	5	5/0	4.24	847
T10	Equal Spheres	5	1	non-linear	0	5	5/0	0.50	100
T11	Crescent [19]	6	2	non-linear	0	1	0/1	0.36	356(*)
E1	Nowacki Beam [16]	2	6	non-linear	1	1	0/1	8.62	1226
E2	Tension/Compression Spring [12]	3	3	mixed	0	1	0/1	33.54	6708
E3	Speed Reducer [20]	7	11	mixed	4	1	0/1	0.11	105(*)



(a) Six-Hump Camel (b) Hosaki (c) Sinexp (d) Branin (e) Himmelblau (f) Donut

(g) Sasena (h) Cheese (i) Unequal Spheres (j) Equal Spheres

(k) Crescent (2D) (l) Nowacki Beam (m) Tension/Compression Spring (n) Speed Reducer

Fig. 1: Feasible regions (grey) of the benchmark problems. For 2D problems the full design space is shown, higher dimensional problems are visualized using parallel plots and feasible designs of the reference sets.

the dataset at each iteration (i.e., $|S_F|$), and our proposed FR-IGD metric. Reference set for FR-IGD calculations is created based on a Sobol sequence of 20,000[4] samples. The size of FR-IGD reference sets (after filtering the feasible designs) is provided in the last column of Table I. The full Sobol sequence is used as reference set for the MCC score.

Results in Table II indicate that the ranking order of FRI methods in terms of final FR-IGD values on problems T1, T3, T6, T7, T8, T11 and E3, is identical. A visual analysis for problem T8[5] is provided in Figure 2 and it clearly highlights the correlation between FR-IGD values and the sampling-quality of the obtained set of feasible points. While the MCC fails to show significant differences between the FRI methods, the $|S_F|$ and FR-IGD clearly signal that boundary-based methods (PBE and EF) sample less feasible designs, and hence have higher FR-IGD values. In contrast, PoFV and ECI sampled more *diversely scattered* feasible points and thus have low FR-IGD values, with PoFV having lower FR-IGD value since it has more uniform coverage, which further cements the reliability of FR-IGD metric in capturing the quality of

[4]We use 100,000 samples for the T11 and E3 problem as they have an extremely low feasibility ratio and a higher input dimensionality. This is shown with (*) in Table I.

[5]Figures for other problems are provided in the supplementary material.

TABLE II: Performance results at the final step given by the median. The arrow indicates if higher (↑) or lower (↓) values are better. Best value is put in **bold**. Statistically similar results, according to the Wilcoxon signed-rank test, are underlined.

	MCC ↑					$\lvert S_F \rvert$ ↑					FR-IGD ↓				
id	PoFV	ECI	ATCH	PBE	EF	PoFV	ECI	ATCH	PBE	EF	PoFV	ECI	ATCH	PBE	EF
T1	**0.99**	0.98	**0.99**	**0.99**	0.98	**34**	29	21	23	25	**0.048**	0.056	0.061	0.076	0.073
T2	0.98	0.98	0.98	**0.99**	0.98	**25**	15	20	14	22	**0.039**	0.063	0.053	0.067	0.066
T3	**1.00**	0.99	**1.00**	**1.00**	**1.00**	**28**	12	18	10	16	**0.038**	0.046	0.049	0.102	0.067
T4	**0.98**	0.94	0.97	0.97	0.81	18	4	14	4	**21**	**0.026**	0.093	0.030	0.062	0.101
T5	0.88	0.86	0.87	**0.89**	0.79	16	8	19	12	**26**	**0.036**	0.062	0.068	0.046	0.211
T6	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**36**	30	20	18	18	**0.045**	0.050	0.067	0.086	0.082
T7	**1.00**	**1.00**	**1.00**	**1.00**	0.99	**32**	18	15	16	4	**0.031**	0.044	0.048	0.065	0.146
T8	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**35**	28	17	25	14	**0.045**	0.051	0.065	0.069	0.100
T9	0.94	0.98	0.97	**0.99**	0.93	**104**	45	48	28	54	0.073	**0.059**	0.061	0.099	0.102
T10	0.88	0.68	0.93	**0.97**	0.86	**203**	198	74	54	93	0.117	0.280	**0.109**	0.123	0.162
T11	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	268	**300**	94	96	88	0.126	**0.114**	0.156	0.161	0.162
E1	**0.95**	0.94	0.88	**0.95**	0.91	**23**	12	18	20	9	**0.030**	0.045	0.045	0.032	0.070
E2	0.96	0.92	0.90	0.96	**0.99**	56	**61**	38	54	22	**0.114**	0.120	0.140	0.115	0.200
E3	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	33	**65**	32	4	0	0.254	**0.223**	0.262	0.417	3.235

sampling. We notice that this ranking-order of FRI methods is maintained for most problems having one feasible region.

Considering problems such as T4, which have multiple feasible regions, we can see that not all acquisition functions manage to consistently find all distinct regions (Figure 3). Although EF has highest number of feasible designs ($|S_F|$) in the final dataset for T4 (Table II), its MCC and FR-IGD are worse than other methods. We can visually see in Figure 3 that despite EF having high $|S_F|$ value (21), it misses to discover the rightmost feasible region. This highlights the shortcoming of $|S_F|$ as a metric to reliably gauge the quality of sampling. Besides that, ECI and PBE have the lowest $|S_F|$ and FR-IGD score closer to EF. ECI and PBE showcase good MCC scores, since they are able to sample around the feasible regions, but note that they have poor sampling quality (Figure 3), again highlighting that MCC values can be deceiving. This poor sampling quality is rightfully reflected with a bad FR-IGD value, indicating that FR-IGD metric is capable to capture the information about the sampling quality even for constrained problems with *disconnected* feasible regions.

VI. Conclusion

We present a set of test functions and engineering cases as benchmark problems for feasible region identification. The problems vary in dimensionality, constraints, and most importantly, in their feasible region characteristics. We introduce a new performance metric, FR-IGD, specifically designed for measuring the diversity of the sampled feasible designs. The conducted experiments show that model metrics like MCC do not capture any information on the sampling. Moreover, the number of feasible designs $|S_F|$ can also give a biased view on the diversity, while FR-IGD more clearly shows the relationship between the sampling and the diversity.

In our future work, we would like to investigate how FR-IGD can be used to monitor sampling within individual feasible regions for problems with multiple disconnected feasible regions, thereby giving us an indication regarding how many feasible regions were discovered by a given FRI method and what was the quality of sampling in each of them, individually. Currently, the FR-IGD metric is dependent on the reference set, which limits its usage to cases where the ground truth is known or can be exhaustively sampled. Secondly, reliability of FR-IGD is dictated by the quality of reference set S^*. We attempt to attenuate this limitation by having a high number of reference points, however further investigation is necessary to develop a robust metric. In current work, FR-IGD is successfully able to capture and quantify the information about the quality of points sampled by FRI methods. Therefore, it can serve as a guideline oracle metric to derive a metric which does not rely on a reference set. As the field grows further, we would also like to extend our benchmarking suite to engineering problems beyond 7D. The proposed work thus paves the way for many research directions in the field of FRI.

References

[1] M. L. Santoni, E. Raponi, R. D. Leone, and C. Doerr, "Comparison of high-dimensional bayesian optimization algorithms on bbob," *ACM Transactions on Evolutionary Learning*, vol. 4, no. 3, pp. 1–33, 2024.

[2] J. de Nobel, F. Ye, D. Vermetten, H. Wang, C. Doerr, and T. Bäck, "Iohexperimenter: Benchmarking platform for iterative optimization heuristics," *Evolutionary Computation*, vol. 32, no. 3, pp. 205–210, 2024.

[3] D. Vermetten, J. Rook, O. L. Preuß, J. de Nobel, C. Doerr, M. López-Ibañez, H. Trautmann, and T. Bäck, "Mo-iohinspector: Anytime benchmarking of multi-objective algorithms using iohprofiler," in *International Conference on Evolutionary Multi-Criterion Optimization*. Springer, 2025, pp. 242–256.

[4] N. Knudde, I. Couckuyt, K. Shintani, and T. Dhaene, "Active learning for feasible region discovery," in *2019 18th IEEE International Conference On Machine Learning And Applications (ICMLA)*. IEEE, 2019, pp. 567–572.

[5] C. E. Rasmussen and C. K. Williams, *Gaussian Processes for Machine Learning*. MIT Press, 2005.

[6] P. I. Frazier, "A tutorial on bayesian optimization," *arXiv preprint arXiv:1807.02811*, 2018.

[7] I. Couckuyt, S. Rojas Gonzalez, and J. Branke, "Bayesian optimization: tutorial," in *Proceedings of the genetic and evolutionary computation conference companion*, 2022, pp. 843–863.

[8] D. Azzimonti, D. Ginsbourger, C. Chevalier, J. Bect, and Y. Richet, "Adaptive design of experiments for conservative estimation of excursion sets," *Technometrics*, vol. 63, no. 1, pp. 13–26, 2021.

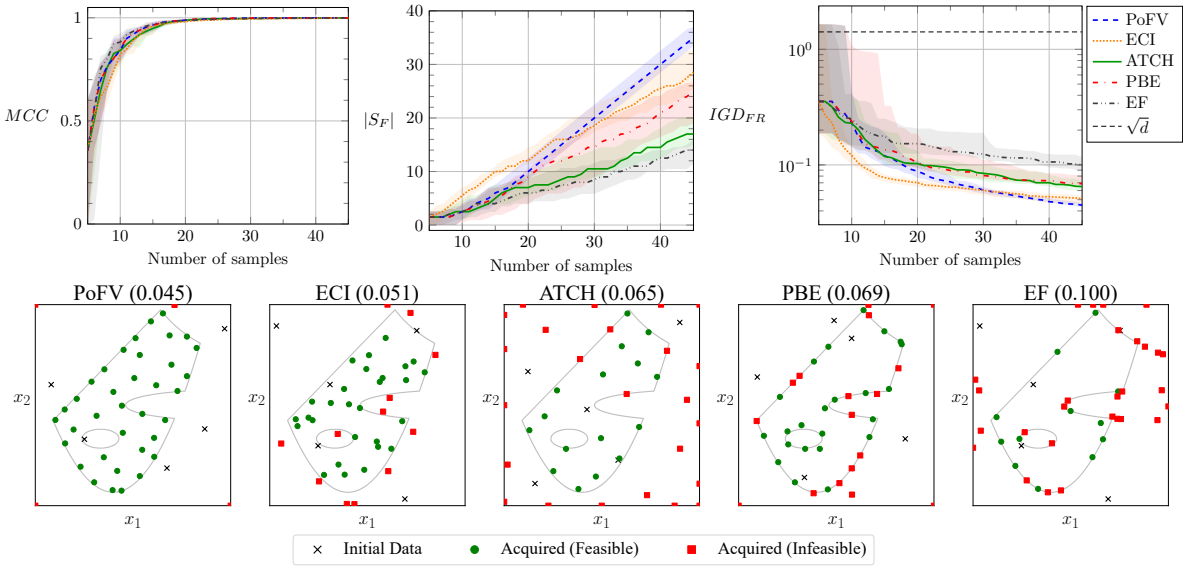


Fig. 2: [Top] Performance of the different acquisition functions on the T8 problem: (left) MCC performance where higher is better, (middle) number of feasible designs $|S_F|$ in the dataset, (right) proposed FR-IGD metric performance where lower is better. The median and the 5th-95th percentile interval are shown. [Bottom] Acquired samples of the run with the median final FR-IGD performance (as given in the title) is shown.

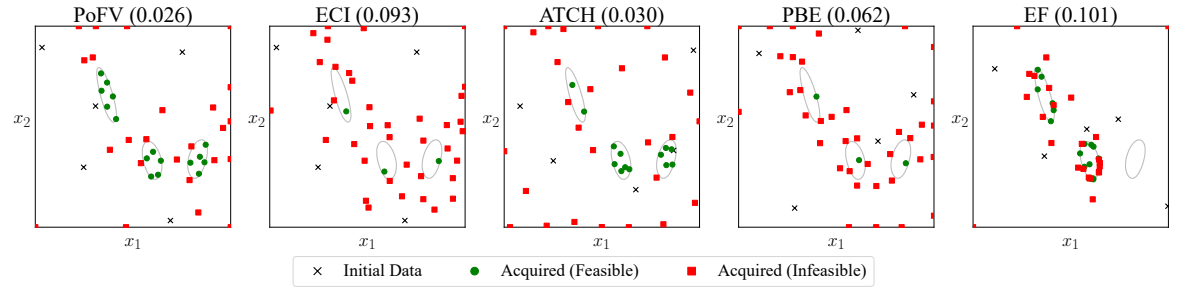


Fig. 3: Acquired samples according to the different acquisition functions on the T4 problem. The run with the median final FR-IGD performance (as given in the title) is shown.

[9] G. Malkomes, B. Cheng, E. H. Lee, and M. Mccourt, "Beyond the pareto efficient frontier: Constraint active search for multiobjective experimental design," in *International Conference on Machine Learning*. PMLR, 2021, pp. 7423–7434.

[10] A. Rahat and M. Wood, "On bayesian search for the feasible space under computationally expensive constraints," in *International conference on machine learning, optimization, and data science*. Springer, 2020, pp. 529–540.

[11] I. Nikova, T. Dhaene, and I. Couckuyt, "Cost-aware active learning for feasible region identification," in *Proceedings of the Companion Conference on Genetic and Evolutionary Computation*, 2023, pp. 2286–2288.

[12] I. Nikova, S. Rojas Gonzalez, T. Dhaene, and I. Couckuyt, "Trade-offs in bayesian active learning for feasible region identification," *Journal of Intelligent Manufacturing*, pp. 1–24, 2025.

[13] B. Letham, P. Guan, C. Tymms, E. Bakshy, and M. Shvartsman, "Look-ahead acquisition functions for bernoulli level set estimation," in *International Conference on Artificial Intelligence and Statistics*. PMLR, 2022, pp. 8493–8513.

[14] C. A. Coello Coello and M. Reyes Sierra, "A study of the parallelization of a coevolutionary multi-objective evolutionary algorithm," in *Mexican international conference on artificial intelligence*. Springer, 2004, pp. 688–697.

[15] Z. Wang and M. Ierapetritou, "A novel feasibility analysis method for black-box processes using a radial basis function adaptive sampling approach," *AIChE Journal*, vol. 63, no. 2, pp. 532–550, 2017.

[16] W. Chen and M. Fuge, "Active expansion sampling for learning feasible domains in an unbounded input space," *Structural and Multidisciplinary Optimization*, vol. 57, no. 3, pp. 925–945, 2018.

[17] D. M. Himmelblau *et al.*, *Applied nonlinear programming*. McGraw-Hill, 2018.

[18] V. S. K. Adi, R. Laxmidewi, and C.-T. Chang, "An effective computation strategy for assessing operational flexibility of high-dimensional systems with complicated feasible regions," *Chemical Engineering Science*, vol. 147, pp. 137–149, 2016.

[19] K. Deb, "An efficient constraint handling method for genetic algorithms," *Computer methods in applied mechanics and engineering*, vol. 186, no. 2-4, pp. 311–338, 2000.

[20] T. Ray, "Golinski's speed reducer problem revisited," *AIAA journal*, vol. 41, no. 3, pp. 556–558, 2003.

[21] A. Kaintura, K. Foss, I. Couckuyt, T. Dhaene, O. Zografos, A. Vaysset, and B. Sorée, "Machine learning for fast characterization of magnetic logic devices," in *2018 IEEE electrical design of advanced packaging and systems symposium (EDAPS)*. Ieee, 2018, pp. 1–3.

[22] M. Balandat, B. Karrer, D. R. Jiang, S. Daulton, B. Letham, A. G. Wilson, and E. Bakshy, "BoTorch: A Framework for Efficient Monte-Carlo Bayesian Optimization," in *Advances in Neural Information Processing Systems 33*, 2020. [Online]. Available: http://arxiv.org/abs/1910.06403