Benchmark Tasks for Quality-Diversity Applied to Uncertain Domains

Manon Flageat
manon.flageat18@ic.ac.uk
Imperial College London
London, U.K.

Luca Grillotti
luca.grillotti16@imperial.ac.uk
Imperial College London
London, U.K.

Antoine Cully
a.cully@imperial.ac.uk
Imperial College London
London, U.K.

ABSTRACT

While standard approaches to optimisation focus on producing a single high-performing solution, Quality-Diversity (QD) algorithms allow large diverse collections of such solutions to be found. If QD has proven promising across a large variety of domains, it still struggles when faced with uncertain domains, where quantification of performance and diversity are non-deterministic. Previous work in Uncertain Quality-Diversity (UQD) has proposed methods and metrics designed for such uncertain domains. In this paper, we propose a first set of benchmark tasks to analyse and estimate the performance of UQD algorithms. We identify the key uncertainty properties to easily define UQD benchmark tasks: the uncertainty location, the type of distribution and its parameters. By varying the nature of those key UQD components, we introduce a set of 8 easy-to-implement and lightweight tasks, split into 3 main categories. All our tasks build on the Redundant Arm: a common QD environment that is lightweight and easily replicable. Each one of these tasks highlights one specific limitation that arises when considering UQD domains. With this first benchmark, we hope to facilitate later advances in UQD.

CCS CONCEPTS

• **Computational Methodologies** → **Evolutionary Robotics**.

KEYWORDS

Quality-Diversity optimisation, Uncertain domains, MAP-Elites, Behavioural diversity.

ACM Reference Format:
Manon Flageat, Luca Grillotti, and Antoine Cully. 2023. Benchmark Tasks for Quality-Diversity Applied to Uncertain Domains . In *Genetic and Evolutionary Computation Conference Companion (GECCO '23 Companion), July 15–19, 2023, Lisbon, Portugal.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3583133.3596326

1 INTRODUCTION

Quality-Diversity (QD) optimisation [2, 4] has proven in recent years to be a promising approach across multiple application domains [3, 9, 10]. QD aims to find a collection $\mathcal{A}$ of diverse and high-performing solutions to an optimisation problem. It does so by storing in $\mathcal{A}$ solutions i with high-fitness f_i that induce novel behaviours, quantified using behaviour descriptors d_i. In other words, QD aims to optimise:

$$\max_{\mathcal{A}} \left(\text{QDScore}(\mathcal{A})\right) = \max_{\mathcal{A}} \sum_{i \in \mathcal{A}} f_i \quad \text{w.r.t} \quad \forall i \in \mathcal{A}, d_i \in \text{cell}_i \tag{1}$$

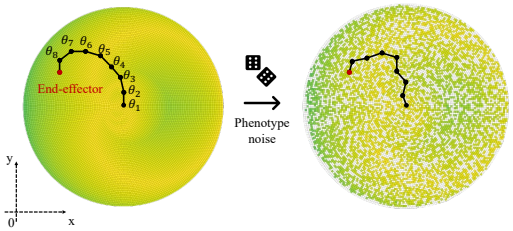


Figure 1: Illustration of the effect of uncertainty in the Redundant Arm QD task; in this case, the uncertainty is applied to the parameters of the arm. The descriptor is given as the end-effector (x, y) position and the fitness as the variance of the joint angles. The background figures the corresponding QD Illusory archive (left) and Corrected archive (right) after multiple reevaluations of each solution.

Despite their encouraging results, multiple works have highlighted the limitations of QD algorithms when considering uncertain domains [6, 7, 11, 12]. A QD domain is seen as uncertain if a solution i can get different fitness f_i and descriptor d_i values from one evaluation to another. In other words, in Uncertain Quality-Diversity (UQD) [7] fitness and descriptor associated with one solution are no more fixed values but distributions: $\tilde{f}_i \sim \mathcal{D}_{f_i}$, $\tilde{d}_i \sim \mathcal{D}_{d_i}$. When applied to such domains, usual QD algorithms such as MAP-Elites [13] tend to keep solutions that have been lucky during evaluation and get outlier fitness or descriptor values. This leads to a loss in the overall performance of their final collection. Contrary to standard QD approaches, UQD algorithms take into consideration the distributional nature of their variables, and adapt the problem formulation accordingly [7]:

$$\max_{\mathcal{A}} \left(\text{QDScore}(\mathcal{A})\right) = \max_{\mathcal{A}} \sum_{i \in \mathcal{A}} \mathbb{E}_{\tilde{f}_i \sim \mathcal{D}_{f_i}} \left[\tilde{f}_i\right] \quad \text{w.r.t} \quad \forall i \in \mathcal{A}, \mathbb{E}_{\tilde{d}_i \sim \mathcal{D}_{d_i}} \left[\tilde{d}_i\right] \in \text{cell}_i \tag{2}$$

Previous work in UQD [6, 7, 11, 12] have proposed solutions tailored for these domains, as well as new metrics that better estimate the performance of QD algorithms when applied to UQD domains. In this paper, we aim to propose a set of benchmark tasks to help

This work is licensed under a Creative Commons Attribution International 4.0 License.
GECCO '23 Companion, July 15–19, 2023, Lisbon, Portugal
© 2023 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-0120-7/23/07.
https://doi.org/10.1145/3583133.3596326

Table 1: Performance-estimation Redundant Arm benchmark tasks.

	Uncertainty location	Distribution	Parameters
Gaussian fitness noise	Fitness after evaluation: $\begin{cases}\tilde{f} = f + \epsilon \\ \tilde{d} = d\end{cases}$	$\epsilon \sim \mathcal{N}(0, \sigma^2)$	$\sigma > 0$
Bimodal fitness noise		$\epsilon \sim \alpha\mathcal{N}(0, \sigma_1^2) + (1-\alpha)\mathcal{N}(-1, \sigma_2^2)$	$\sigma_1, \sigma_2 > 0 \quad \alpha \in (0, 1)$
Small-Gaussian descriptor noise	Descriptor after evaluation: $\begin{cases}\tilde{f} = f \\ \tilde{d} = d + \epsilon\end{cases}$	$\epsilon \sim \mathcal{N}(\mathbf{0}, \sigma^2\boldsymbol{I})$	$\sigma \in (0, 0.02)$
Large-Gaussian descriptor noise		$\epsilon \sim \mathcal{N}(\mathbf{0}, \sigma^2\boldsymbol{I})$	$\sigma > 0.1$
Bimodal descriptor noise		$\epsilon \sim \alpha\mathcal{N}(\mathbf{0}, \sigma_1^2\boldsymbol{I}) + (1-\alpha)\mathcal{N}(\begin{bmatrix}1\\1\end{bmatrix}, \sigma_2^2\boldsymbol{I})$	$\sigma_1, \sigma_2 > 0 \quad \alpha \in (0, 1)$

Table 2: Reproducibility-maximisation Redundant Arm benchmark tasks.

	Uncertainty location	Distribution	Parameters
Gaussian descriptor noise with 2 variance values	Descriptor after evaluation: $\begin{cases}\tilde{f} = 0 \\ \tilde{d} = d + \epsilon\end{cases}$	$\begin{cases}\epsilon \sim \mathcal{N}(\mathbf{0}, \sigma_1^2\boldsymbol{I}) \text{ if } sign(\prod_j \theta^j) \geq 0 \\ \epsilon \sim \mathcal{N}(\mathbf{0}, \sigma_2^2\boldsymbol{I}) \text{ otherwise}\end{cases}$	$\sigma_2 \gg \sigma_1 > 0$
Gaussian descriptor noise with continuous variance values		$\epsilon \sim \mathcal{N}(\mathbf{0}, (\eta * \mathbb{V}(\theta))^2\boldsymbol{I})$	$\eta > 0$

Table 3: Realistic Redundant Arm benchmark tasks.

	Uncertainty location	Distribution	Parameters
Gaussian Phenotype J noise	Dimension J of the phenotype: $\tilde{\theta} = \begin{pmatrix}\theta^1 & \cdots & \theta^{J-1} & \theta^J + \epsilon & \theta^{J+1} & \cdots & \theta^N\end{pmatrix}$	$\epsilon \sim \mathcal{N}(0, \sigma^2)$	$\sigma > 0$

future developments of UQD algorithms. As first contribution, we propose a more systematic way to define UQD benchmark tasks for a given environment, as the ensemble of (1) an uncertainty location, (2) an uncertainty distribution type and (3) its corresponding parameters. These properties also constitute initial tools to analyse inherent uncertainty in more realistic UQD tasks, allowing to better select tailored approaches. As second contribution, we define, based on this definition, 8 new tasks, split into 3 main categories. These tasks aim to highlight some of the main limitations that arise in UQD settings. All our tasks are based on the Redundant Arm task, a simple-to-implement and lightweight QD task that has been widely studied across the existing literature [3, 8]. To facilitate the use of our benchmark, we make our code available at URL https://github.com/adaptive-intelligent-robotics/QDax .

2 PROPERTIES OF UQD TASKS

One can define multiple UQD benchmark tasks for a given environment by varying the properties of the uncertainty induced in the environment. We identified 3 main uncertainty properties: the uncertainty location, the type of distribution, and the parameters of this distribution. In this section, we analyse those properties and explain how they can be used to define UQD benchmark tasks. They can also be used as tools to analyse more realistic UQD tasks.

2.1 Uncertainty Location

As explained above, in UQD tasks, fitnesses and descriptors are no more fixed values but distributions. These distributions can arise from uncertainty at different levels within the environment. The clear definition of this uncertainty location is critical to define benchmark tasks and study UQD tasks. Previous work in uncertain evolutionary algorithms proposed a classification that separates prior noise, applied to the phenotype before the fitness evaluation, and posterior noise, applied to the fitness after the evaluation [14]. In the case of UQD, we propose to decompose further the prior noise category to encompass dynamic and applied problems as considered in QD. In simple QD environments such as Rastrigin or the Redundant Arm, one could consider adding (1) noise on the fitness and descriptor reading, or (2) noise on the phenotype (i.e. the parameters in Rastrigin and the joint angles in the Arm), which corresponds directly to the prior and posterior noise categories. However, in more complex settings, such as robot control environments [3], one could consider adding (1) noise on the fitness and descriptor reading, (2) noise on the phenotype, i.e. the controller parameters, (3) noise on the initial position of the robot, and (4) stochasticity in the dynamics of the environment. Similar locations could be considered in closed-loop QD-RL environments [1, 8].

2.2 Distribution Type

Another key property of UQD tasks is the distribution of the noise or stochasticity. Most previous work in UQD used Gaussian or Uniform distributions. However, more complex distributions such as skewed, multi-modal, or non-continuous distributions can lead to interesting properties as will be discussed later in this paper. Depending on the uncertainty location within the environment, the structure of this distribution will then be shaped and modified to give the final fitness and descriptor distributions, leading to complex uncertainty structures.

2.3 Distribution Parameters

Varying the parameters of a distribution may also have a significant impact on the performance of UQD algorithms. For example, gaussian descriptor distributions with a very low variance may have a negligible impact on an algorithm's results; indeed, if an individual's descriptor has a low variance, then most re-evaluations of that descriptor will land in the same behaviour niche (i.e. MAP-Elites cell). Conversely, a really large fitness variance can make the comparison between two fitness values almost non-meaningful and thus prevent convergence.

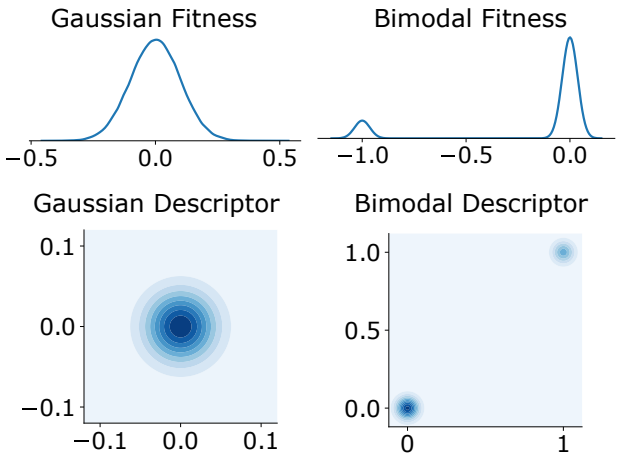


Figure 2: Distributions in the Performance-estimation tasks.

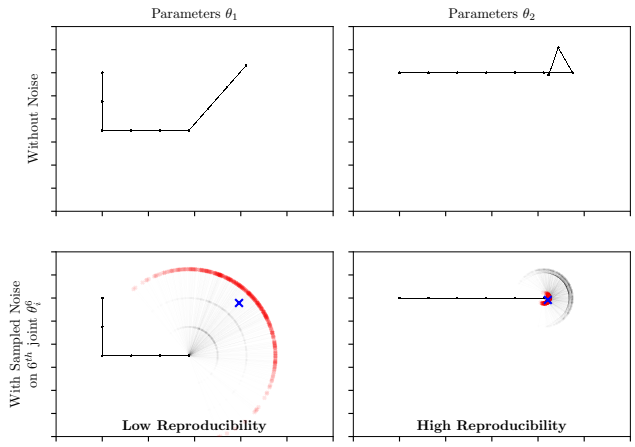


Figure 3: Illustration of the effect of noise applied on the 6^{th} joint of 2 distinct arm configurations θ_1 (left) and θ_2 (right). The red crosses indicate the sampled descriptors, and the blue cross their average. The descriptor reproducibility achieved by θ_2 is way higher than the one achieved by θ_1.

3 REDUNDANT ARM UQD BENCHMARK

3.1 Redundant Arm

The Redundant Arm [2–4], illustrated in Figure 1, is one of the most commonly used QD benchmark task. It aims to learn how to achieve every reachable point with a planar N-DoF robotic Arm:

- **Controller:** real-valued angles of the N-DoF: $\theta_i = (\theta_i^j)_{1 \leq j \leq N}$.
- **Descriptor:** defined as the (x, y) position of the end-effector, after applying the control values in the joints.
- **Fitness:** in this paper, we consider two versions of this task:
  (1) fitness 0 for all individuals.
  (2) fitness as negative variance of the N joint angles $(\theta_i^j)_{1 \leq j \leq N}$: $f_i = -\mathbb{V}(\theta_i) = -\frac{1}{N} \sum_{j=1}^{N} (\theta_i^j - \overline{\theta_i})^2$.

3.2 UQD Benchmark Tasks

In the following, we propose a set of benchmark UQD tasks in the Redundant Arm environment, chosen to illustrate some of the common challenges in UQD. As highlighted by Flageat and Cully [7], UQD approaches usually tackle two main problems:

- **Performance estimation:** correctly estimating the expected fitness $\mathrm{E}_{f_i \sim \mathcal{D}_f}[f_i]$ and expected descriptor $\mathrm{E}_{d_i \sim \mathcal{D}_d}[d_i]$ of solutions from their evaluations.
- **Reproducibility maximisation:** encouraging solutions that are more likely to get similar fitness and descriptor values from one evaluation to another. This corresponds to maintaining solutions with lower variance of fitness $\mathrm{V}_{f_i \sim \mathcal{D}_f}[f_i]$ and variance of each descriptor dimension $\mathrm{V}_{d_i^j \sim \mathcal{D}_{d^j}}\left[d_i^j\right]$. This issue only arises in tasks where solutions can have different variances.

In Section 4, we present 5 benchmark tasks designed to assess the Performance estimation ability of UQD algorithms; and in Section 5, we introduce 2 other tasks to assess their ability to select reproducible solutions. We also propose in Section 6 a more realistic task that combines both problems. For each section, we give some preliminary results obtained by MAP-Elites and MAP-Elites-sampling (MAP-Elites-smpl) with 30 samples [7, 12] with 100×100 grid, two approaches used as baselines in existing UQD works [6, 7, 11, 12].

4 PERFORMANCE ESTIMATION TASKS

In this section, we propose benchmark tasks that assess the ability of UQD approaches to estimate performance. We consider two cases: tasks where only the fitness is a distribution and tasks where only the descriptor is a distribution. One can easily combine these cases to get complex tasks with noise on both. These tasks all build on the Redundant Arm with non-zero fitness from Section 3.1. We introduce each task below, give corresponding equations in Table. 1 and Figure 2, and some results in Figures 4 and 5.

4.1 Gaussian Fitness Noise

Gaussian noise on the fitness is the most commonly used noise distribution in previous UQD works. Here, the variance parameter σ^2 for this distribution can take almost any value. The results shows the benefit of sampling-based approaches in this case (Fig. 4).

4.2 Bimodal Fitness Noise

One of the particularity of Gaussian distributions, as proposed in the previous section, is that sampling is always beneficial. Here, we propose a task where too low number of samples can actually be detrimental. We define a bimodal distribution with a distant low-probability second mode. This second mode is really detrimental for low number of samples, leading to an interesting phenomenon where MAP-Elites outperforms MAP-Elites-sampling (Fig. 4).

4.3 Small-Gaussian Descriptor Noise

Gaussian noise on the descriptor is the most commonly used noise distribution in previous UQD works. The value of the variance σ^2 for this distribution is critical. Here, we consider small values that lead individuals to stay with high probability within the same behavioural niche (i.e. MAP-Elites cell). The results show the benefit of sampling-based approaches in this case (Fig. 4).

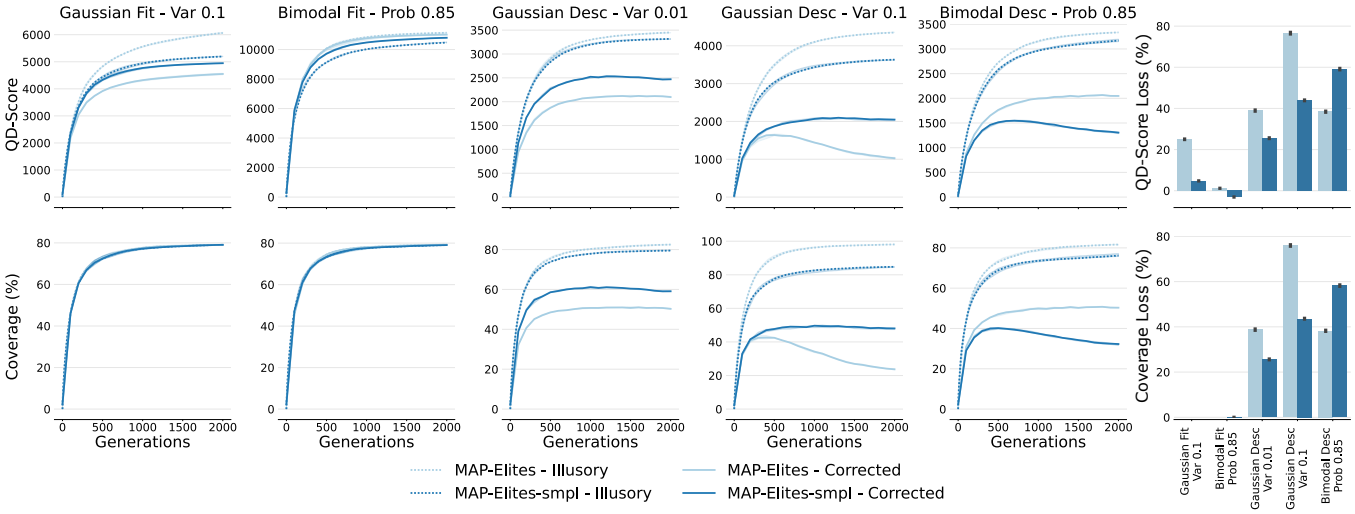


Figure 4: Results on the Performance-estimation benchmark tasks. Each column corresponds to one of the tasks from Table 1. For each baseline, we give the Illusory (left - dashed line) and Corrected (left - plain line) QD-Score and Coverage [7, 11]. Solid lines are the median across 5 replications, and shaded area the quartiles. We also give the QD-Score loss (right - top) and Coverage loss (right - bottom) from Flageat et al. [5], that quantify the estimation capability of each baseline.

4.4 Large-Gaussian Descriptor Noise

In this second version, we choose the standard deviation σ to be sufficiently high to lead solutions to change behavioural niche with high probability. This task makes the correct estimation of descriptor highly complex. For instance, with a MAP-Elites grid of size 100×100 and a value of σ higher than 0.1, most descriptor estimates with fewer than 100 samples will not be placed in the appropriate cell. The results highlight the limitation of MAP-Elites and MAP-Elites-sampling in this case (Fig. 4 and 5).

4.5 Bimodal Descriptor Noise

The Bimodal distribution is the equivalent of the Bimodal Fitness Noise from Section 4.2 for the descriptor. In this task, the noise takes the form of a bimodal distribution in 2 dimensions with a distant low-probability second mode (see Fig. 2). This distant second mode makes it difficult to estimate expected descriptors based on a low amount of samples, which is why MAP-Elites-sampling achieves the worst results in this task (Fig. 4 and 5).

5 REPRODUCIBILITY MAXIMISATION TASKS

In this section, we propose 2 benchmark tasks that assess the ability of UQD approaches to keep solutions with high reproducibility. To simplify the problem, we consider only the reproducibility in term of descriptor, and we only focus on the Redundant Arm with zero-fitness from Section 3.1. We introduce each task below, give corresponding equations in Table. 2, and some results in Figures 6 and 7. To quantify the performance of approaches, we display the Reproducibility-Score as defined by Flageat and Cully [7]. Additionally, inspired by Grillotti et al. [11], we add as baseline MAP-Elites-sampling-Reproducibility (MAP-Elites-smpl-Reproducibility), which directly optimises for the reproducibility of solutions.

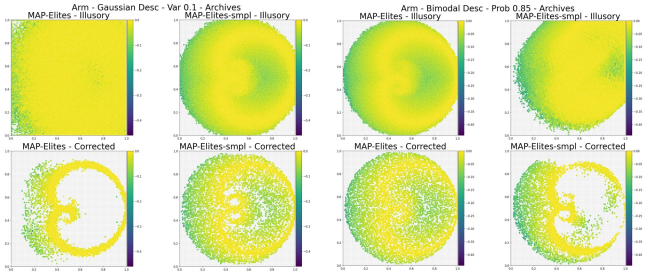


Figure 5: Illusory (top) and Corrected (bottom) archives for the two Performance-estimation descriptor noise tasks.

5.1 Gaussian Descriptor Noise with Two Possible Variance Values

We first consider a simple task where there are two distinct possible values of variance for the descriptor: one small σ_1^2 and one large σ_2^2. Thus, solutions ending in the same cell can be either highly reproducible or poorly reproducible. In this task, we attribute variance based on phenotype: we give low variance σ_1^2 to solutions with positive product of joint angles, and high variance σ_2^2 to the others. Ideally, as the fitness is the same (zero) for all solutions, UQD approaches are expected to prefer low-variance solutions to high-variance ones. However, the results show that none of the two considered approaches manages to select reproducible solutions across the descriptor space, unlike MAP-Elites-sampling-Reproducibility [11] (Fig. 6 and 7). UQD algorithms require more complex mechanisms to be able to effectively solve this task [7, 11, 12].

5.2 Gaussian Descriptor Noise with Continuous Variance Values

We propose a more complex version of the previous task, with the aim to refine the study of UQD algorithms in reproducibility-maximisation tasks. In this task, the variance can take any values,

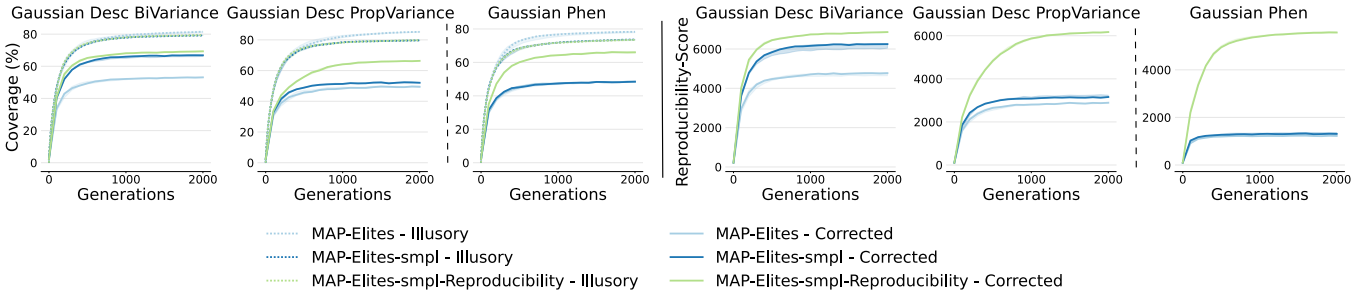


Figure 6: Results on the Reproducibility-maximisation and Realistic-combination benchmark tasks. Each column corresponds to one of the task defined in Table 2 or 3. For each baseline, we give the Illusory Coverage (dashed line - left) and Corrected Coverage (plain line - left). We also display the Reproducible-Score from Flageat and Cully [7] that quantifies the ability of baselines to maintain reproducible solutions. It is computed based on the normalised variance σ^2_{norm} as the sum over the Corrected archive of $1 - \sigma^2_{norm}$. Solid lines are the median across 5 replications, and shaded area the quartiles.

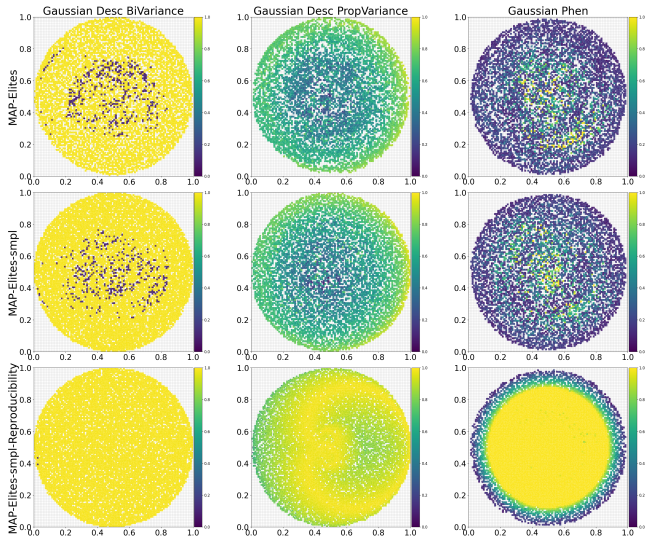


Figure 7: Reproducibility archives [7] for benchmark tasks of Reproducibility-maximisation and Realistic-combination.

continuously, within a range. Here, we attribute variance proportionally to the variance of the joint angles. Ideally, as the fitness is the same (zero) for all solutions, UQD approaches are expected to select the most reproducible solutions for each cell. However, as for the previous task, results in Figure 6 show that neither MAP-Elites nor MAP-Elites-sampling favour reproducible solutions.

6 REALISTIC COMBINATION TASK

In addition to the targeted tasks proposed in the previous sections, we propose here one more realistic task that combines both the issues of performance estimation and reproducibility maximisation: the Gaussian Phenotype J noise. We give the corresponding equations in Table 3 and some results in Figures 6 and 7.

6.1 Gaussian Phenotype J Noise

In this task, we propose to change the uncertainty location to consider the phenotype (i.e. the joint angles). Noise is applied on only one single pre-defined joint angle, and all solutions have zero-fitness. UQD algorithms are expected to maximise the reproducibility of the solutions while also providing accurate estimates of their expected descriptors. In this task, the best solutions are fully reproducible, i.e. have a descriptor variance of 0, as depicted on Figure 3. The results show that MAP-Elites and MAP-Elites-sampling do not manage to find these solutions (Fig. 6 and 7).

7 CONCLUSION AND DISCUSSION

This paper presents the first set of Uncertain Quality-Diversity benchmark tasks. Each of these tasks aims to highlight one limitation of existing QD algorithms in uncertain domains and points toward interesting directions of improvement. They are all defined for the Redundant Arm, a commonly used QD environment. In total, we propose 8 tasks that can be split into 3 main categories: Performance-estimation, Reproducibility-maximisation and Realistic tasks. The first two correspond to the two main issues that arise when considering Uncertain domains, as highlighted in previous work. We hope that this paper will constitute a meaningful first benchmark for later advances in UQD and that it will lead to more similar tasks to assess QD algorithms in Uncertain domains.

REFERENCES

[1] Felix Chalumeau, Thomas Pierrot, Valentin Macé, Arthur Flajolet, Karim Beguir, Antoine Cully, and Nicolas Perrin-Gilbert. 2022. Assessing Quality-Diversity Neuro-Evolution Algorithms Performance in Hard Exploration Problems. *arXiv preprint arXiv:2211.13742* (2022).

[2] Konstantinos Chatzilygeroudis, Antoine Cully, Vassilis Vassiliades, and Jean-Baptiste Mouret. 2021. Quality-Diversity Optimization: a novel branch of stochastic optimization. In *Black Box Optimization, Machine Learning, and No-Free Lunch Theorems*. Springer, 109–135.

[3] Antoine Cully, Jeff Clune, Danesh Tarapore, and Jean-Baptiste Mouret. 2015. Robots that can adapt like animals. *Nature* 521, 7553 (2015), 503–507.

[4] Antoine Cully and Yiannis Demiris. 2017. Quality and diversity optimization: A unifying modular framework. *IEEE Transactions on Evolutionary Computation* 22, 2 (2017), 245–259.

[5] Manon Flageat, Felix Chalumeau, and Antoine Cully. 2023. Empirical analysis of PGA-MAP-Elites for Neuroevolution in Uncertain Domains. *ACM Transactions on Evolutionary Learning* 3, 1 (2023), 1–32.

[6] Manon Flageat and Antoine Cully. 2020. Fast and stable MAP-Elites in noisy domains using deep grids. In *ALIFE 2020: The 2020 Conference on Artificial Life*. MIT Press, 273–282.

[7] Manon Flageat and Antoine Cully. 2023. Uncertain Quality-Diversity: Evaluation methodology and new methods for Quality-Diversity in Uncertain Domains. *arXiv preprint arXiv:2302.00463* (2023).

[8] Manon Flageat, Bryan Lim, Luca Grillotti, Maxime Allard, Simón C Smith, and Antoine Cully. 2022. Benchmarking Quality-Diversity Algorithms on Neuroevolution for Reinforcement Learning. *arXiv preprint arXiv:2211.02193* (2022).

[9] Matthew C Fontaine, Ruilin Liu, Ahmed Khalifa, Jignesh Modi, Julian Togelius, Amy K Hoover, and Stefanos Nikolaidis. 2021. Illuminating mario scenes in the latent space of a generative adversarial network. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 35. 5922–5930.

[10] Adam Gaier, Alexander Asteroth, and Jean-Baptiste Mouret. 2017. Aerodynamic design exploration through surrogate-assisted illumination. In *18th AIAA/ISSMO multidisciplinary analysis and optimization conference*. 3330.

[11] Luca Grillotti, Manon Flageat, Bryan Lim, and Antoine Cully. 2023. Don't Bet on Luck Alone: Enhancing Behavioral Reproducibility of Quality-Diversity Solutions in Uncertain Domains. *arXiv preprint arXiv:2304.03672* (2023).

[12] Niels Justesen, Sebastian Risi, and Jean-Baptiste Mouret. 2019. Map-elites for noisy domains by adaptive sampling. In *Proceedings of the Genetic and Evolutionary Computation Conference Companion*. 121–122.

[13] Jean-Baptiste Mouret and Jeff Clune. 2015. Illuminating search spaces by mapping elites. *arXiv preprint arXiv:1504.04909* (2015).

[14] Frank Neumann, Mojgan Pourhassan, and Vahid Roostapour. 2020. Analysis of evolutionary algorithms in dynamic and stochastic environments. *Theory of evolutionary computation: recent developments in discrete optimization* (2020), 323–357.