Top-Two Pareto Front Thompson Sampling

Lennert Saerens[1,2] , **Bram Silue**[1] , **Eleni Litsa**[2] , **Peter Vrancx**[1,2] and **Pieter Libin**[1,3]
[1]Artificial Intelligence Lab, Departement of Computer Science, Vrije Universiteit Brussel, Brussels, Belgium, [2]imec, Leuven, Belgium, [3]Data Science Institute, Interuniversity Institute of Biostatistics and Statistical Bioinformatics, UHasselt, Hasselt, Belgium
Lennert.Saerens@vub.be

Abstract

Identifying Pareto optimal solutions is critical to support multi-objective decision-making. We introduce the first anytime Multi-Objective Multi-Armed Bandit algorithm for the Pareto Set Identification problem, taking a Bayesian approach: *Top-Two Pareto Front Thompson Sampling* (TTPFTS). We benchmark TTPFTS against state-of-the-art fixed-budget Pareto Set Identification algorithms on synthetic environments. Next, we demonstrate its practical utility in a challenging multi-objective molecular discovery setting by efficiently exploring an ultra-large synthesis-on-demand molecular library. Furthermore, we introduce a novel uncertainty quantification metric that estimates our algorithm's confidence in the predicted Pareto set. We demonstrate that this metric effectively proxies true performance, yielding a robust methodology for monitoring learning progress in complex settings. Finally, we complement these empirical findings with a theoretical proof of the algorithm's asymptotic correctness.

1 Introduction

The *Multi-Armed Bandit* (MAB) problem has been widely examined in the literature, primarily as a single-objective stochastic optimization task. In this setting, an agent sequentially draws samples from a collection of K unknown probability distributions, referred to as arms. Two prevalent problem variants are either to maximize the cumulative rewards, commonly framed as *regret minimization* [Auer *et al.*, 2002; Lattimore and Szepesvári, 2020], or to identify the arm with the highest expected value, known as *best-arm identification* [Audibert and Bubeck, 2010; Russo, 2016].

Yet, in many real-world settings, multiple and potentially conflicting objectives must be optimized simultaneously, and no single arm may achieve the optimal score across all criteria [Hayes *et al.*, 2021]. Consequently, the rewards become multi-dimensional and several arms can be Pareto optimal. This gives rise to the *Multi-Objective Multi-Armed Bandit* (MOMAB) framework first proposed by Drugan and Nowé, 2013. A significant body of work focuses on minimizing Pareto regret by prioritizing the selection of optimal arms during the agent's sampling process [Drugan and Nowé, 2013; Yahyaa and Manderick, 2015; Cheng *et al.*, 2024]. We instead focus on the pure-exploration formulation introduced by Zuluaga *et al.*, 2013 and Auer *et al.*, 2016. This problem, known as *Pareto Set Identification* (PSI), aims to identify the set of Pareto optimal arms $\mathcal{P}^\star$ as efficiently and accurately as possible. An arm is included in the Pareto optimal set $\mathcal{P}^\star$ if it is not dominated by any other arm, meaning that no alternative arm has strictly better expected values for all of the objectives. Hence, the arms present in the Pareto set provide optimal trade-offs between objectives. This multi-objective exploration problem of identifying $\mathcal{P}^\star$ arises in numerous applications, including materials discovery [Gopakumar *et al.*, 2018], clinical trials [Zhao *et al.*, 2018], epidemic mitigation policies [Reymond *et al.*, 2024], molecular design [Liu *et al.*, 2025], and drug discovery [Fromer *et al.*, 2024].

Contributions. While significant strides have been made in the fixed-confidence and fixed-budget MOMAB PSI settings, a notable gap remains regarding the *anytime* setting. In this work, we address this limitation by introducing *Top-Two Pareto Front Thompson Sampling* (TTPFTS), the first Bayesian anytime algorithm designed for the MOMAB PSI problem. We first evaluate the algorithm's empirical performance against state-of-the-art fixed-budget *Empirical Gap Elimination* (EGE) algorithms [Kone *et al.*, 2023b] on synthetic benchmarks. Next, we validate the algorithm in a real-world decision-making scenario: exploring synthesis-on-demand molecular libraries to identify molecules with desirable properties. In this setting, we show that our algorithm achieves significant efficiency gains over exhaustive virtual screening, which remains the standard approach despite its high computational cost [Lyu *et al.*, 2019; Klarich *et al.*, 2024], and over a state-of-the-art active learning method [Fromer *et al.*, 2024]. Finally, we introduce a novel uncertainty quantification metric specific to Bayesian MOMAB PSI algorithms. This metric exploits information contained in the posteriors to estimate confidence in current Pareto predictions without ground-truth knowledge, offering a practical and interpretable methodology for guiding decision-making in complex multi-objective settings.

Related work. The seminal works in the MOMAB PSI setting are attributed to Zuluaga *et al.*, 2013 and Auer *et*

al., 2016. The latter proposed a fixed-confidence algorithm that identifies Pareto optimal arms, potentially including sub-optimal arms that become Pareto optimal when increased coordinate-wise by ϵ_1 [Auer *et al.*, 2016]. Concurrently, Zuluaga *et al.* studied a structured variant of fixed-confidence PSI where means are regular functions of arm descriptors. Utilizing Gaussian process modeling, they derived worst-case sample complexity bounds [Zuluaga *et al.*, 2013; Zuluaga *et al.*, 2016]. Building on these foundations, the *Adaptive Pareto Exploration* algorithm was proposed in [Kone *et al.*, 2023a], which allows for various stopping rules to accommodate different relaxations of the PSI problem. More recently, [Crepon *et al.*, 2024] established tight guarantees in the asymptotic regime by iteratively solving an optimization problem characterizing the optimal problem-dependent sample complexity. Also within the fixed-confidence setting, [Kone *et al.*, 2025a] introduced a Bayesian approach relying on posterior distributions to guide sampling and stopping decisions. Addressing the fixed-budget PSI setting, [Kone *et al.*, 2023b] developed the EGE algorithms, which extend the successive-halving and successive-rejects strategies to multi-objective environments. Finally, numerous variants of the standard PSI setting have been explored, including (top) feasible-arm identification [Katz-Samuels and Scott, 2018; Katz-Samuels and Scott, 2019; Kone *et al.*, 2025b] and PSI under partial ordering [Ararat and Tekin, 2021].

2 Setting

In this section, the *Multi-objective Multi-armed Bandit* (MOMAB) *Pareto Set Identification* (PSI) setting is formalized. Throughout this section, for consistency, we adopt the notation used in [Kone *et al.*, 2023b].

A MOMAB PSI setting consists of K arms, each arm corresponding to a reward distribution $\nu_1, \cdots, \nu_K$ over $\mathbb{R}^D$ with respective expectations $\boldsymbol{\theta}_1, \cdots, \boldsymbol{\theta}_K$, where $D > 1$ represents the number of objectives. The bandit instance is defined as $\nu := (\nu_1, \cdots, \nu_K)$. At each time $t = 1, 2, \cdots$ the agent pulls an arm $a_t \in [K]$ and observes an independent sample $\boldsymbol{X}_t \sim \nu_{a_t}$ with $\mathbb{E}[\boldsymbol{X}_t] = \boldsymbol{\theta}_{a_t}$. The goal of an anytime MOMAB PSI algorithm is to estimate the Pareto set after every observation. These are precisely the arms that are not Pareto dominated by any other arm. The following definitions are adopted from [Kone *et al.*, 2023b].

Let $\boldsymbol{\theta}_i$ and $\boldsymbol{\theta}_j$ be the expected reward vectors of arms i and j. Arm i is (Pareto) dominated by j ($\boldsymbol{\theta}_i \preceq \boldsymbol{\theta}_j$) if $\theta_i^d \leq \theta_j^d$ for all objectives $d \in \{1, \ldots, D\}$ and there exists at least one d where $\theta_i^d < \theta_j^d$. It is strongly dominated ($\boldsymbol{\theta}_i \prec \boldsymbol{\theta}_j$) if $\theta_i^d < \theta_j^d$ strictly for all d. The Pareto set, denoted $\mathcal{P}^\star(\nu)$ or simply $\mathcal{P}^\star$, contains all arms not dominated by any other arm in $[K]$. Arms within $\mathcal{P}^\star$ are termed Pareto optimal, while the remaining arms comprise the suboptimal set $\mathcal{S} = [K] \setminus \mathcal{P}^\star$.

While the anytime MOMAB PSI setting requires the algorithm to maintain and update its estimate of the Pareto set after each observation [Bubeck *et al.*, 2009], other formulations of the problem consider stopping conditions that depend on a predefined accuracy or budget constraint. In the *fixed-confidence* MOMAB PSI setting, the learner aims to identify the true Pareto set $\mathcal{P}^\star$ with a probability of at least $1 - \delta$ for some confidence level $\delta \in (0, 1)$, while minimizing the total number of samples collected. Formally, the learner must design a stopping time τ and an estimator $\hat{\mathcal{P}}_\tau^\star$ such that $\mathbb{P}(\hat{\mathcal{P}}_\tau^\star = \mathcal{P}^\star) \geq 1 - \delta$ [Gabillon *et al.*, 2012; Kone *et al.*, 2023a; Crepon *et al.*, 2024; Kone *et al.*, 2025a]. In contrast, the *fixed-budget* MOMAB PSI setting assumes a total sampling budget B is fixed a priori. The learner must then use these B samples to produce an estimator $\hat{\mathcal{P}}_B^\star$ of the Pareto set that maximizes the probability of correct identification $\mathbb{P}(\hat{\mathcal{P}}_B^\star = \mathcal{P}^\star)$ [Gabillon *et al.*, 2012; Kone *et al.*, 2023b]. The anytime formulation, as considered in this work, is more flexible: the algorithm continuously refines its Pareto set estimate $\hat{\mathcal{P}}_t^\star$ after every observation without requiring a predefined stopping rule [Jun and Nowak, 2016]. Finally, our work adopts a *preference-unaware* approach [Osika *et al.*, 2023], aiming to recover the *entire* Pareto front without incorporating decision-maker preferences.

3 Algorithm

In this section, we present the *Top-Two Pareto Front Thompson Sampling* (TTPFTS) algorithm, which, to the best of our knowledge, constitutes the first algorithmic exploration of the anytime MOMAB PSI setting. The proposed TTPFTS algorithm extends the principles of *Top-Two Thompson Sampling* (TTTS) [Russo, 2016] to the multi-objective domain. In single-objective best-arm identification, TTTS selects a candidate best arm and then repeatedly resamples to find a different arm that outperforms the candidate, effectively allocating samples between these two arms. TTPFTS generalizes this mechanism to the MOMAB setting without requiring iterative re-sampling. Without a preference vector to scalarize objectives, there is no single best arm. Instead, there exists a set of Pareto optimal arms representing the trade-off surface. Consequently, TTPFTS generalizes the top-two notion from arms to *Pareto fronts*: it stochastically distributes samples between the current estimated Pareto front, the best set, and the second Pareto front, the challenger set, according to a fixed probability.

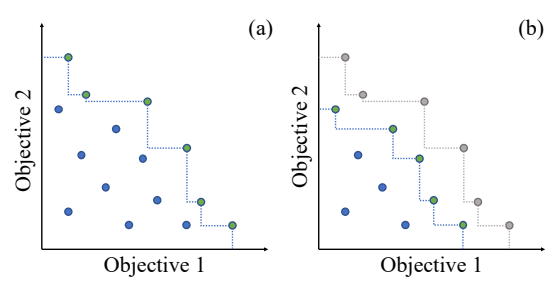


Figure 1: Visualization of the TTPFTS strategy in a bi-objective maximization setting. The samples obtained at time t by sampling from the posterior distributions $\pi(\cdot|\mathcal{H}^{(t-1)})$ are shown as dots. With probability ρ, the optimal arms $\mathcal{F}_1^\star$ are computed and one of them is chosen uniformly at random (a). With probability $1 - \rho$, the optimal arms are removed and one of the optimal arms $\mathcal{F}_2^\star$ of the remaining set is chosen uniformly at random (b).

The detailed procedure is outlined in Algorithm 1. At each time step t, the algorithm begins by sampling from the D-variate posterior distributions $\pi(\cdot \mid \mathcal{H}^{(t-1)})$ for each arm $a \in [K]$. For each arm, this generates an estimate $\tilde{\boldsymbol{\theta}}_a^{(t)}$ of its expected value, as illustrated by the dots in Figure 1. Once these samples are obtained, TTPFTS performs a Bernoulli trial with success probability ρ. The outcome of this trial decides which front to explore. If the trial yields 1, the algorithm identifies the set of Pareto optimal arms $\mathcal{F}_1^\star$ based on the current sampled means. As shown in Figure 1(a), an arm a_t is then selected uniformly at random from this set. This step focuses on verifying arms that are currently believed to be optimal. If the trial yields 0 (probability $1-\rho$), the algorithm removes the optimal set $\mathcal{F}_1^\star$ and considers the subset of remaining arms $\mathcal{A}' = [K] \setminus \mathcal{F}_1^\star$. Subsequently, it computes the Pareto front of this subset, denoted as $\mathcal{F}_2^\star$. We refer to $\mathcal{F}_2^\star$ as the *second front*, visualized in Figure 1(b). An arm a_t is selected uniformly at random from this second front. This step promotes exploration among the most promising suboptimal arms, effectively testing if they have been underestimated or solidifying them as suboptimal. In the event that all arms are mutually non-dominating and the remaining subset $\mathcal{A}'$ is empty, TTPFTS selects an arm uniformly at random from $\mathcal{F}_1^\star$.

Algorithm 1: Top-Two Pareto Front Thompson Sampling (TTPFTS)

Input: MOMAB instance ν with K arms, parameter $\rho \in (0,1)$, prior π, history $\mathcal{H}^{(0)} = \emptyset$
for $t \leftarrow 1$ **to** ∞ **do**
  for $a \leftarrow 1$ **to** K **do**
    Sample $\tilde{\boldsymbol{\theta}}_a^{(t)} \sim \pi(\cdot \mid \mathcal{H}^{(t-1)})$
  Compute the first front:
  $\mathcal{F}_1^\star = \{i \in [K] \mid \nexists j \in [K] : \tilde{\boldsymbol{\theta}}_i^{(t)} \preceq \tilde{\boldsymbol{\theta}}_j^{(t)}\}$
  Sample $b \sim \mathcal{B}(\rho)$
  if $b = 1$ **then**
    Select arm a_t uniformly at random from $\mathcal{F}_1^\star$
  else
    Compute remaining arms $\mathcal{A}' = [K] \setminus \mathcal{F}_1^\star$
    Compute the second front:
    $\mathcal{F}_2^\star = \{i \in \mathcal{A}' \mid \nexists j \in \mathcal{A}' : \tilde{\boldsymbol{\theta}}_i^{(t)} \preceq \tilde{\boldsymbol{\theta}}_j^{(t)}\}$
    Select arm a_t uniformly at random from $\mathcal{F}_2^\star$
  Pull arm a_t and observe reward $\boldsymbol{X}_t \sim \nu_{a_t}$
  Update history $\mathcal{H}^{(t)} \leftarrow \mathcal{H}^{(t-1)} \cup \{(a_t, \boldsymbol{X}_t)\}$
  Compute and recommend $\hat{\mathcal{P}}_t^\star$, the Pareto set of the posterior means

This strategy mirrors the core principle of *Boundary Focused Thompson Sampling* [Libin *et al.*, 2019], which addresses the single-objective top-m problem by targeting the decision boundary between the m-th and $(m+1)$-th arms. By generalizing this concept, TTPFTS focuses exploration on the multi-objective equivalent of that boundary: the frontier between the first and second Pareto fronts. Finally, regardless of which front was selected, the learner pulls the chosen arm a_t, observes the reward vector $\boldsymbol{X}_t \sim \nu_{a_t}$, and updates the history $\mathcal{H}^{(t)}$. As this is an anytime algorithm, the learner maintains a recommendation set $\hat{\mathcal{P}}_t^\star$, which can be queried at any point during the process.

4 Theoretical Guarantees

We establish the asymptotic correctness of TTPFTS under standard assumptions: finite arms, consistent posteriors, strict Pareto gaps, and sufficient exploration (see Section D for formal Assumptions 1–4, Lemma 1, and the complete proof).

Theorem 1 (Asymptotic Correctness)**.** *Under Assumptions 1–4 and Lemma 1, the TTPFTS algorithm's estimate of the Pareto set $\hat{\mathcal{P}}_t^\star$ is asymptotically correct:*

$$\lim_{t\to\infty} \mathbb{P}\big(\hat{\mathcal{P}}_t^\star \neq \mathcal{P}^\star\big) = 0.$$

Proof Sketch. By posterior consistency and infinite exploration, the estimated means converge almost surely to the true means. Consequently, the estimated dominance relationships eventually mirror the ground truth: for any suboptimal arm, its true dominator will eventually strongly dominate it in the estimate, ensuring its exclusion. Conversely, for any optimal arm, no other arm will strongly dominate it in the limit, guaranteeing its inclusion. A union bound over the finite set of arms confirms that the probability of any classification error, false positive or false negative, converges to zero. □

5 Empirical Evaluation

Next, we empirically evaluate the performance of the proposed TTPFTS algorithm on the PSI task. Our evaluation is conducted using the eight synthetic MOMAB PSI benchmarking environments introduced by Kone *et al.*, 2023b. These environments are designed to capture a wide range of Pareto front geometries, sub-optimality gap profiles, and objective space dimensionalities commonly encountered in MOMAB problems. For each environment, we use Gaussian reward distributions with diagonal covariance matrix $\sigma^2 I_D$ with variance $\sigma^2 = 0.25$ and the identity matrix I_D of order D. Specifications for each environment are provided in Section A.1.

We compare TTPFTS against both variants of the fixed-budget EGE algorithm proposed in [Kone *et al.*, 2023b]: *Successive Rejects* (EGE-SR) and *Successive Halving* (EGE-SH). These represent the state-of-the-art for the fixed-budget MOMABs. In addition, we include a Uniform Sampling strategy as a baseline. As a Bayesian method, TTPFTS requires specifying a prior distribution over the rewards. To demonstrate robustness, we assume weak prior information about the environment by employing a uniform prior [Honda and Takemura, 2013]. For the likelihood, we assume a Gaussian reward model with unknown mean and unknown variance. This leads to a t-distributed posterior. We set $\rho = 0.5$.

To quantify the quality of the estimated Pareto sets produced by each algorithm, we use the Jaccard metric J [Lu *et al.*, 2019] between the estimated Pareto set $\hat{\mathcal{P}}^\star$ and the ground-truth Pareto set $\mathcal{P}^\star$:

$$J(\mathcal{P}^\star, \hat{\mathcal{P}}^\star) = \frac{|\mathcal{P}^\star \cap \hat{\mathcal{P}}^\star|}{|\mathcal{P}^\star \cup \hat{\mathcal{P}}^\star|} \tag{1}$$

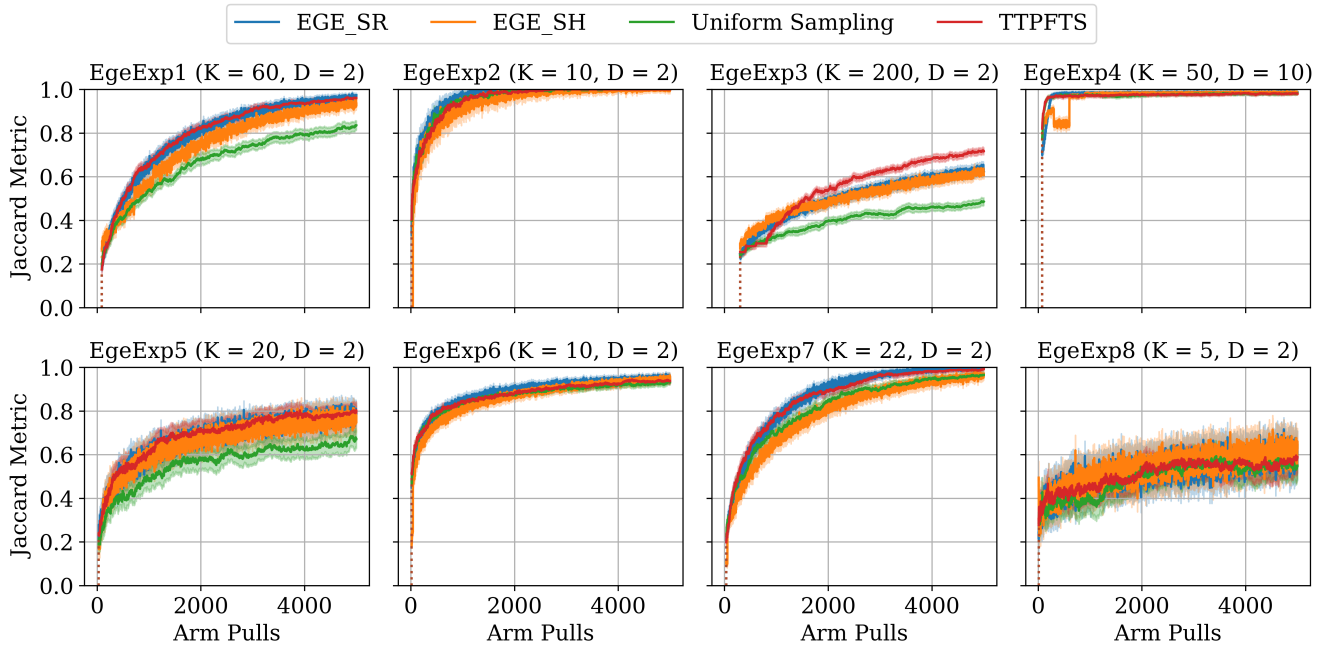


Figure 2: Performance of each of the MOMAB PSI algorithms with respect to the Jaccard metric in each of the eight synthetic benchmarking settings proposed in [Kone *et al.*, 2023b]. We use Gaussian reward distributions with diagonal covariance matrix $\sigma^2 I_D$ with $\sigma^2 = 0.25$. Results are averaged over 100 experimental trajectories, with shaded areas indicating the 95% confidence interval around the mean performance.

For each environment, we report mean performance over 100 independent experimental trajectories, each consisting of 5,000 arm pulls.

Due to the inherently distinct nature of the algorithms under comparison, fixed-budget versus anytime, it is important to clarify how a single experimental trajectory is generated for each method. The fixed-budget EGE algorithms are executed independently for every budget $T \in (0, 5000]$, producing an estimate $\hat{\mathcal{P}}^\star_T$ at the end of each run. In contrast, TTPFTS is executed only once per trajectory and outputs its anytime estimate $\hat{\mathcal{P}}^\star_t$ after each arm pull. Furthermore, it is crucial to recognize that fixed-budget methods are explicitly optimized to minimize the identification error after a predetermined budget T, allowing them to defer exploitation until late in the sampling process. Anytime algorithms, by contrast, must continuously balance exploration and exploitation, as they are required to provide meaningful recommendations at every time step. When examining the results, it is crucial to note that this structural constraint imposes a systematic penalty on anytime methods within fixed-budget comparisons.

Figure 2 reports the evolution of the Jaccard metric starting after K arm pulls, indicated by the dotted line, ensuring that all arms are sampled at least once to yield meaningful Pareto estimates. TTPFTS consistently outperforms the Uniform Sampling baseline and the EGE-SH algorithm across all environments. Furthermore, TTPFTS's scores are competitive with those of the EGE-SR algorithm across all environments. Notably, TTPFTS achieves parity with state-of-the-art algorithms explicitly optimized for fixed horizons, despite operating under the stricter anytime constraint. Most notably, in the EgeExp3 environment, which is characterized by a large number of arms, TTPFTS clearly outperforms both EGE variants. In this challenging setting, TTPFTS converges faster and achieves substantially higher Jaccard scores, suggesting that its posterior-driven allocation is particularly effective in large-scale problems where rigid gap-based arm elimination, as performed by EGE, may be premature. Finally, these results validate the algorithm's effectiveness even under weak prior assumptions. To provide a comprehensive assessment beyond the Jaccard metric, we include Bernoulli and misclassification metrics in Section A.2. These metrics offer complementary insights into the probability of *exact* PSI and the distribution of classification errors, respectively. The complete implementation of all empirical evaluation experiments is available on GitHub[1].

6 Uncertainty Quantification

Up to this stage, we have been using ground-truth dependent metrics to gain insight into the performance of MOMAB PSI algorithms. However, in real-world decision-making settings, the decision maker has no access to the true Pareto front, hence necessitating the PSI process. In this section, we define a measure of TTPFTS's uncertainty about its current estimate $\hat{\mathcal{P}}^\star_t$. The proposed uncertainty measure leverages TTPFTS's Bayesian nature by using the information captured in the posterior distributions of the top-two Pareto fronts.

We quantify uncertainty $\mathcal{U}_\beta$ as the Bhattacharyya overlap [Bhattacharyya, 1943] between the posterior densities of

[1] https://github.com/LennertSaerens/TTPFTS

the arms in the current top-two Pareto fronts [Van Molle *et al.*, 2021]. Intuitively, high overlap indicates that the algorithm is struggling to distinguish optimal arms from suboptimal challengers, i.e. high uncertainty, while limited overlap signals a clear separation, i.e. high confidence. This concept is visualized in Figure 3.

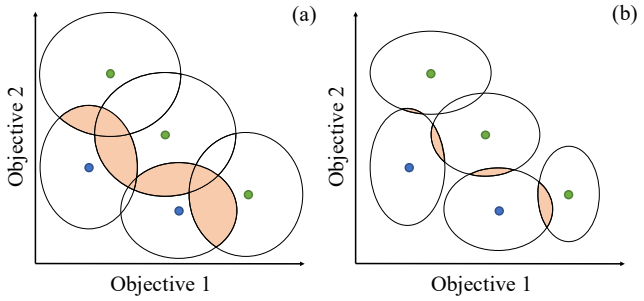


Figure 3: Visualization of proposed uncertainty quantification in the top-two Pareto fronts of a bi-objective maximization setting. The points indicate the mean of the arms' posterior distributions, while the ellipses around the means represent the posterior densities. As TTPFTS samples from the top-two fronts, the posteriors concentrate around the true means, evolving from a situation with high uncertainty, panel (a), to a situation with lower uncertainty, panel (b), indicated by the smaller shaded overlap in the probability density functions.

The Bhattacharyya coefficient β lies in the interval $[0, 1]$, where 1 indicates perfect overlap between distributions. Let $\hat{\mathcal{P}}_1^\star$ denote the estimated first Pareto front and $\hat{\mathcal{P}}_2^\star$ denote the the estimated second Pareto front, based on posterior means. Equation (2) shows the computation of our proposed uncertainty measure, detailed further in Section C.

$$\mathcal{U}_\beta\left(\hat{\mathcal{P}}_1^\star, \hat{\mathcal{P}}_2^\star\right) = \frac{1}{\left|\hat{\mathcal{P}}_1^\star\right|\left|\hat{\mathcal{P}}_2^\star\right|} \sum_{a_i \in \hat{\mathcal{P}}_1^\star} \sum_{a_j \in \hat{\mathcal{P}}_2^\star} \beta(a_i, a_j) \qquad (2)$$

To derive a normalized, environment-independent quantification, we compute the coefficient for every pair of arms between the top-two fronts, summing these values and dividing by the total number of pairwise comparisons. As visualized in Figure 3, this summation may involve double counting in dense regions of the objective space. Consequently, the resulting measure acts as a conservative estimate of total uncertainty. An evaluation justifying our choice of the Bhattacharyya coefficient over alternative information-theoretic metrics, such as KL-divergence and posterior entropy, is detailed in Section C.4.

We first analyze the temporal evolution of this metric across the benchmarks from Section 5. As shown in Figure 4, the uncertainty decreases as the learner gathers evidence. An exponentially decaying trend is evident in most environments. A notable exception is EgeExp8, characterized by a unique optimal arm and geometrically vanishing sub-optimality gaps, which exhibits a linear decline. Notably, these trends align with the performance curves in Figure 2: environments that are harder to learn, exhibiting a slower Jaccard improvement, display a correspondingly slower reduction in uncertainty.

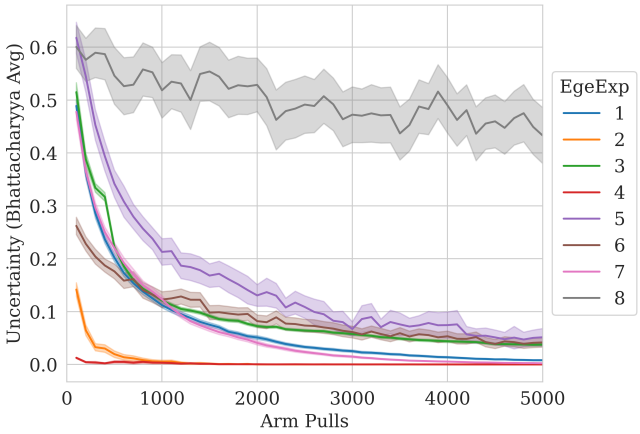


Figure 4: Evolution of the uncertainty of TTPFTS over 5000 arm pulls in each of the eight synthetic benchmarking settings proposed in [Kone *et al.*, 2023b]. Results are averaged over 100 experimental trajectories, with shaded areas indicating the 95% confidence interval around the mean uncertainty.

To formally validate this metric as a proxy for estimation quality, we examine the correlation between the proposed uncertainty measure and the ground-truth Jaccard metric. For each environment, Figure 5 presents boxplots of the Pearson correlation coefficients calculated over 100 independent trajectories, each consisting of 5000 arm pulls.

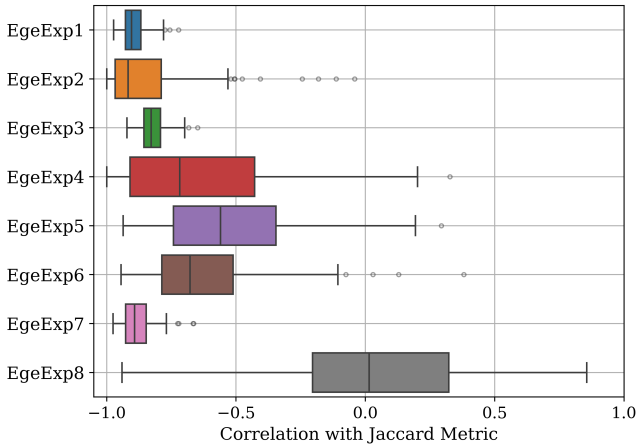


Figure 5: Distribution of Pearson correlation coefficients between the Jaccard metric and the proposed uncertainty quantification over 100 trajectories consisting of 5000 arm pulls. A strong negative correlation indicates that the uncertainty metric is a reliable real-time proxy for identification performance.

The results reveal a strong negative correlation for the majority of environments. This confirms that a decrease in our uncertainty metric reliably maps to an increase in the accuracy of the Pareto set estimate. Even in environments with higher variance such as EgeExp4 and EgeExp5, the correlation remains predominantly negative. The outlier is EgeExp8, where the correlation is centered near zero with high variance. This decoupling of uncertainty and performance is

consistent with the structural difficulty of EgeExp8 seen in Figure 2, where the algorithm struggles to form a stable estimate. However, for all solvable environments, the proposed metric serves as an effective, ground-truth-independent indication of the uncertainty for the decision-maker.

Finally, we note that this uncertainty quantification creates a powerful synergy with the anytime nature of TTPFTS. Unlike fixed-budget algorithms, where the stopping time is rigid and predetermined, TTPFTS empowers the decision-maker to stop the process dynamically. By monitoring the uncertainty in real-time, decision-makers can decide when the algorithm has reached a low enough uncertainty, where the top-two fronts have sufficiently separated, and terminate the experiment. This transforms TTPFTS from a sampling strategy towards a decision support system, minimizing resource expenditure and maximally informing the decision-maker.

7 Efficiently Screening Molecular Databases

Moving beyond synthetic benchmarking, we apply TTPFTS to the exploration of *ultra-large synthesis-on-demand* molecular libraries [Klarich *et al.*, 2024]. Our goal is to identify molecules that optimally trade off multiple desirable properties by efficiently exploring chemical spaces where exhaustive enumeration is infeasible. This experiment validates TTPFTS's applicability to complex, real-world decision-making tasks.

Virtual screening has become a cornerstone of hit identification, the process of finding promising candidate molecules for drug development. A key innovation in this field is the advent of synthesis-on-demand libraries. These libraries are constructed combinatorially, akin to a Cartesian product: a vast space of product molecules is defined by reacting all valid combinations from smaller, discrete sets of molecular building blocks, called reagents [Grygorenko *et al.*, 2020]. As these libraries have expanded to contain billions of candidates, exhaustive screening has become computationally infeasible. In response, numerous methodologies have been developed to expedite the identification of promising compounds. Notably, [Klarich *et al.*, 2024] introduced a state-of-the-art bandit-based approach where independent Thompson Sampling agents simultaneously select the specific reagents required for each component of the chemical reaction. Their method achieved a $1000\times$ speedup over exhaustive virtual screening, which remains the standard industry approach despite its high computational cost [Lyu *et al.*, 2019; Klarich *et al.*, 2024]. However, their work was limited to optimizing a single molecular property. We demonstrate that TTPFTS naturally extends this efficient screening paradigm to the multi-objective optimization of possibly conflicting molecular properties.

We adopt the experimental framework of [Klarich *et al.*, 2024], utilizing a combinatorial library generated via a three-component reaction. The reagent pools for the three components consist of 376, 500, and 500 unique reagents, respectively, yielding a fully enumerated library of 94 million compounds [Klarich *et al.*, 2024]. In the bandit formulation, this corresponds to three independent agents, i.e. one per component, where the arms represent the available reagents. The reward signal is derived from the evaluation of the final product resulting from the combination of the bandit-selected reagents.

To transition to the MOMAB PSI framework, we introduce two key modifications. First, we expand the optimization landscape by introducing a second objective. In addition to maximizing structural similarity to a query molecule, we simultaneously optimize for the *LogP*, a key molecular property in drug discovery, representing a compound's lipohilicity, which influences properties such as absorption, distribution and membrane permeability [Lipinski *et al.*, 1997; Waring, 2010]. This bi-objective setup mirrors early stage screening in drug discovery, such as for Alzheimer's disease [Gote *et al.*, 2023]. Secondly, we deploy TTPFTS agents for each of the three reagent-selecting bandits instead of the standard Thompson Sampling agents. This algorithmic shift transforms the system from a best-arm identifier into a Pareto-front explorer. Notably, we exclude fixed-budget algorithms like EGE from this multi-agent combinatorial setting due to significant structural incompatibilities. EGE variants rely on rigid phases requiring specific budget allocations across active arms before eliminations can occur. In a synthesis-on-demand loop, where three independent bandits must synchronously propose reagents to form a single valid product, coordinating these phased requirements becomes computationally intractable. Specifically, aligning the distinct active subsets from multiple agents to form valid reaction products introduces a secondary combinatorial optimization problem. TTPFTS avoids this synchronization bottleneck by naturally fitting the system's sequential, single-pull paradigm.

To establish a ground truth, we perform an exhaustive search across the full library of 94 million molecules. This comprehensive screen reveals a true Pareto front consisting of 52 molecules, which are visualized in Figure 6.

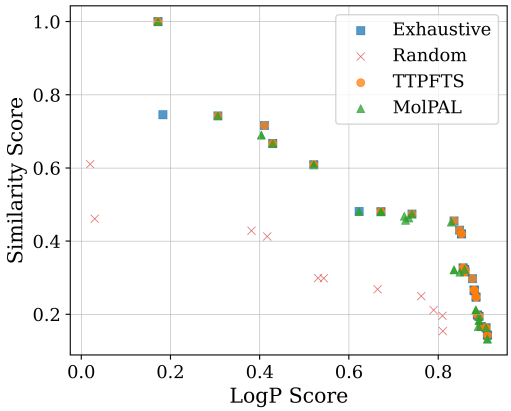


Figure 6: Structural Similarity and LogP scores for the optimal molecules identified in a single run by Random Search, TTPFTS, and MolPAL, compared to the ground-truth Pareto-optimal molecules obtained through exhaustive virtual screening of all 94 million candidates.

Following the protocol in [Klarich *et al.*, 2024], we employ a Random Search strategy as a baseline. This strategy establishes the expected performance of the industry-standard

exhaustive virtual screening approach. We also include the state-of-the-art MolPAL baseline [Fromer *et al.*, 2024] (configuration in Section B.3). All three methods are halted after 50,000 steps. Figure 6 illustrates the outcomes of a representative run. The figure compares the molecules identified by each method to the ground-truth Pareto-optimal set. The results show that, apart from two molecules, the molecules discovered by TTPFTS perfectly coincide with the true Pareto front, despite evaluating only 0.05% of the entire database. In contrast, Random Search fails to recover any of the true Pareto-optimal molecules within the same budget. These findings highlight TTPFTS's ability to efficiently balance exploration and exploitation in a vast chemical space and to rapidly concentrate sampling on promising regions. Additional runs exhibiting similar behavior are provided in Section B.4.

To quantify the average performance over time, Figure 7 reports the Jaccard metric between the estimated and true Pareto sets alongside the evolution of the proposed uncertainty measure. Results are averaged over 100 runs, with shaded regions indicating 95% confidence intervals. The computation of the Jaccard and uncertainty quantification metrics within this multi-agent setting is discussed in Section B.2.

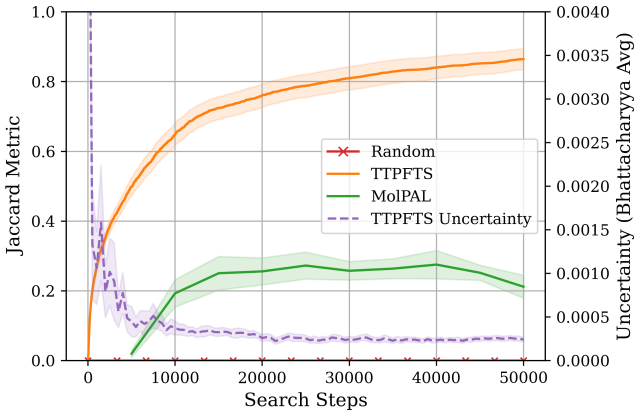


Figure 7: Evolution of the Jaccard metric of the optimal molecules identified by Random Search, TTPFTS, and MolPAL, as well as the proposed uncertainty measure for TTPFTS. Results are averaged over 100 experimental trajectories, with the shaded areas indicating the 95% confidence interval around the mean.

Across all runs, Random Search consistently fails to identify any of the Pareto-optimal molecules. MolPAL reaches a mean Jaccard score of 0.28 at 40,000 oracle calls before declining due to exploitation bias in later iterations. In contrast, TTPFTS reliably obtains scores exceeding 0.8 within approximately 30,000 search steps, confirming the algorithm's ability to rapidly identify optimal trade-offs. Crucially, the figure reveals a strong inverse relationship between the Jaccard score and the uncertainty measure $\mathcal{U}_\beta$. As the estimation quality improves, the uncertainty exhibits a sharp, consistent decline. This validates the metric as a reliable proxy for convergence even in this complex chemical space. This finding highlights a powerful synergy between the anytime nature of TTPFTS and our ground-truth-independent uncertainty quantification. Unlike rigid fixed-budget approaches, this combination provides the drug discovery team with a responsive decision-support system. Drug discovery scientists can monitor the uncertainty in real-time and dynamically determine when the Pareto front is sufficiently resolved to terminate the screening, thereby optimizing the trade-off between computational resources and discovery confidence without requiring prior knowledge of the true Pareto front.

8 Discussion

We introduced *Top-Two Pareto Front Thompson Sampling* (TTPFTS), the first algorithm designed for the anytime MOMAB PSI setting. TTPFTS' anytime nature provides a more flexible alternative to the fixed-budget and fixed-confidence regimes. Our results demonstrate that this flexibility comes at no cost to performance: TTPFTS matches, and in large-action spaces outperforms, state-of-the-art fixed-budget baselines. Crucially, we enable the immediate practical use of this anytime property through a novel uncertainty quantification metric. This provides decision-makers with an interpretable, real-time methodology for monitoring progress which functions without ground-truth knowledge.

Beyond synthetic benchmarks, the combined practical power of these contributions was validated through a challenging multi-objective molecular discovery setting. TTPFTS efficiently navigated the trade-off between conflicting molecular properties, recovering the true Pareto front while exploring less than 0.05% of the search space, and outperforming both the industry-standard approach and the state-of-the-art MolPAL method. Furthermore, the uncertainty metric accurately tracked this progress, effectively acting as a dynamic stopping criterion for the molecular discovery process.

Finally, we acknowledge several limitations, discussed in Section E, which present interesting avenues for future research. First, while Theorem 1 guarantees the asymptotic correctness of TTPFTS, deriving explicit finite-time error bounds remains a non-trivial challenge. Historically, finite-time guarantees for both standard [Thompson, 1933; Agrawal and Goyal, 2013] and Top-Two Thompson Sampling [Russo, 2016; Shang *et al.*, 2019] took years to emerge following their asymptotic proofs. We therefore focus on establishing soundness and immediate practical value through our asymptotic proof, strong empirical performance against state-of-the-art baselines, and a real-time uncertainty metric that tracks convergence. Despite these open challenges discussed in Section E, the successful deployment of TTPFTS in a complex drug discovery application demonstrates its immediate utility and establishes it as a robust advancement in multi-objective decision-making.

Acknowledgments

L.S. is supported by an imec doctoral grant. B.S. acknowledges FWO grant G059423N and OZR-VUB (OZR3863BOF). P.J.K.L. gratefully acknowledges support from the Research council of the Vrije Universiteit Brussel (OZR-VUB via grant number OZR3863BOF), FWO

grant G059423N, FWO grant G0A0S26N, and FWO grant S002626N. This research acknowledges funding from the Flemish Government through the AI Research Program. The computational resources and services used in this work were provided by the VSC (Flemish Supercomputer Center), funded by the Research Foundation Flanders (FWO) and the Flemish Government department EWI. This research is partly funded by the imec.prospect project IMPACT.

References

[Agrawal and Goyal, 2013] Shipra Agrawal and Navin Goyal. Further optimal regret bounds for thompson sampling. In Carlos M. Carvalho and Pradeep Ravikumar, editors, *Proceedings of the Sixteenth International Conference on Artificial Intelligence and Statistics*, volume 31 of *Proceedings of Machine Learning Research*, pages 99–107. PMLR, 2013.

[Ararat and Tekin, 2021] Caugin Ararat and Cem Tekin. Vector optimization with stochastic bandit feedback. In *International Conference on Artificial Intelligence and Statistics*, 2021.

[Audibert and Bubeck, 2010] Jean-Yves Audibert and Sébastien Bubeck. Best arm identification in multi-armed bandits. In *Annual Conference Computational Learning Theory*, 2010.

[Auer *et al.*, 2002] Peter Auer, Nicolò Cesa-Bianchi, and Paul Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine Learning*, 47:235–256, 2002.

[Auer *et al.*, 2016] Peter Auer, Chao-Kai Chiang, Ronald Ortner, and Mădălina M. Drugan. Pareto front identification from stochastic bandit feedback. In *International Conference on Artificial Intelligence and Statistics*, 2016.

[Bhattacharyya, 1943] Anil Kumar Bhattacharyya. On a measure of divergence between two statistical populations defined by their probability distributions. *Bulletin of the Calcutta Mathematical Society*, 35:99–109, 1943.

[Bubeck *et al.*, 2009] Sébastien Bubeck, Rémi Munos, and Gilles Stoltz. Pure exploration in multi-armed bandits problems. In *International Conference on Algorithmic Learning Theory*, 2009.

[Cheng *et al.*, 2024] Ji Cheng, Bo Xue, Jiaxiang Yi, and Qingfu Zhang. Hierarchize pareto dominance in multi-objective stochastic linear bandits. In *AAAI Conference on Artificial Intelligence*, 2024.

[Crepon *et al.*, 2024] élise Crepon, Aurélien Garivier, and Wouter M Koolen. Sequential learning of the Pareto front for multi-objective bandits. In Sanjoy Dasgupta, Stephan Mandt, and Yingzhen Li, editors, *Proceedings of The 27th International Conference on Artificial Intelligence and Statistics*, volume 238 of *Proceedings of Machine Learning Research*, pages 3583–3591. PMLR, 02–04 May 2024.

[Drugan and Nowé, 2013] Mădălina M. Drugan and Ann Nowé. Designing multi-objective multi-armed bandits algorithms: A study. *The 2013 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8, 2013.

[Fromer *et al.*, 2024] Jenna C. Fromer, David E. Graff, and Connor W. Coley. Pareto optimization to accelerate multi-objective virtual screening. *Digital Discovery*, 3:467–481, 2024.

[Gabillon *et al.*, 2012] Victor Gabillon, Mohammad Ghavamzadeh, and Alessandro Lazaric. Best arm identification: A unified approach to fixed budget and fixed confidence. In F. Pereira, C.J. Burges, L. Bottou, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc., 2012.

[Gopakumar *et al.*, 2018] Abhijith M. Gopakumar, Prasanna V. Balachandran, Dezhen Xue, James E. Gubernatis, and Turab Lookman. Multi-objective optimization for materials discovery via adaptive design. *Scientific Reports*, 8(1):3738, 2018.

[Gote *et al.*, 2023] Sharmila Gote, Shankar Thapa, Sonal Dubey, Shachindra L. Nargund, and Mahalakshmi Suresha Biradar. Computational investigation of quinazoline derivatives as keap1 inhibitors for alzheimer's disease. *Informatics in Medicine Unlocked*, 41:101334, 2023.

[Grygorenko *et al.*, 2020] Oleksandr O. Grygorenko, Dmytro S. Radchenko, Igor Dziuba, Alexander Chuprina, Kateryna E. Gubina, and Yurii S. Moroz. Generating multibillion chemical space of readily accessible screening compounds. *iScience*, 23(11):101681, 2020.

[Hayes *et al.*, 2021] Conor F. Hayes, Roxana Ruadulescu, Eugenio Bargiacchi, Johan Kallstrom, Matthew Macfarlane, Mathieu Reymond, Timothy Verstraeten, Luisa M. Zintgraf, Richard Dazeley, Fredrik Heintz, Enda Howley, Athirai Aravazhi Irissappane, Patrick Mannion, Ann Now'e, Gabriel de Oliveira Ramos, Marcello Restelli, Peter Vamplew, and Diederik M. Roijers. A practical guide to multi-objective reinforcement learning and planning. *Autonomous Agents and Multi-Agent Systems*, 36, 2021.

[Honda and Takemura, 2013] Junya Honda and Akimichi Takemura. Optimality of thompson sampling for gaussian bandits depends on priors. In *International Conference on Artificial Intelligence and Statistics*, 2013.

[Jun and Nowak, 2016] Kwang-Sung Jun and Robert Nowak. Anytime exploration for multi-armed bandits using confidence information. In Maria Florina Balcan and Kilian Q. Weinberger, editors, *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 974–982, New York, New York, USA, 20–22 Jun 2016. PMLR.

[Katz-Samuels and Scott, 2018] Julian Katz-Samuels and Clayton D. Scott. Feasible arm identification. In *International Conference on Machine Learning*, 2018.

[Katz-Samuels and Scott, 2019] Julian Katz-Samuels and Clayton D. Scott. Top feasible arm identification. In *International Conference on Artificial Intelligence and Statistics*, 2019.

[Klarich *et al.*, 2024] Kathryn Klarich, Brian Goldman, Trevor Kramer, Patrick Riley, and W. Patrick Walters.

Thompson sampling – an efficient method for searching ultralarge synthesis on demand databases. *Journal of Chemical Information and Modeling*, 64(4):1158–1171, 2024. PMID: 38316125.

[Kone *et al.*, 2023a] Cyrille Kone, Emilie Kaufmann, and Laura Richert. Adaptive algorithms for relaxed pareto set identification. *ArXiv*, abs/2307.00424, 2023.

[Kone *et al.*, 2023b] Cyrille Kone, Emilie Kaufmann, and Laura Richert. Bandit pareto set identification: the fixed budget setting. In *International Conference on Artificial Intelligence and Statistics*, 2023.

[Kone *et al.*, 2025a] Cyrille Kone, Marc Jourdan, and Emilie Kaufmann. Pareto set identification with posterior sampling. In *International Conference on Artificial Intelligence and Statistics*, 2025.

[Kone *et al.*, 2025b] Cyrille Kone, Emilie Kaufmann, and Laura Richert. Constrained pareto set identification with bandit feedback. *ArXiv*, abs/2506.08127, 2025.

[Lattimore and Szepesvári, 2020] Tor Lattimore and Csaba Szepesvári. *Bandit Algorithms*. Cambridge University Press, Cambridge, UK, 2020.

[Libin *et al.*, 2019] Pieter Libin, Timothy Verstraeten, Diederik M Roijers, Wenjia Wang, Kristof Theys, and Ann Nowe. Bayesian anytime m-top exploration. In *2019 IEEE 31st International Conference on Tools with Artificial Intelligence (ICTAI)*, pages 1422–1428, November 2019. ICTAI ; Conference date: 04-11-2019 Through 06-11-2019.

[Lipinski *et al.*, 1997] Christopher A. Lipinski, Franco Lombardo, Beryl W. Dominy, and Paul J. Feeney. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Advanced Drug Delivery Reviews*, 23(1):3–25, 1997. In Vitro Models for Selection of Development Candidates.

[Liu *et al.*, 2025] Yifei Liu, Yiheng Zhu, Jike Wang, Renling Hu, Chao Shen, Wanglin Qu, Gaoang Wang, Qun Su, Yuchen Zhu, Yu Kang, Peichen Pan, Chang-Yu Hsieh, and Tingjun Hou. A multi-objective molecular generation method based on pareto algorithm and monte carlo tree search. *Advanced Science*, 12(20):2410640, 2025.

[Lu *et al.*, 2019] Shiyin Lu, G. Wang, Yao Hu, and Lijun Zhang. Multi-objective generalized linear bandits. In *International Joint Conference on Artificial Intelligence*, 2019.

[Lyu *et al.*, 2019] Jiankun Lyu, Sheng Wang, Trent E. Balius, Isha Singh, Anat Levit, Yurii S. Moroz, Matthew J. O'Meara, Tao Che, Enkhjargal Algaa, Kateryna Tolmachova, et al. Ultra-large library docking for discovering new chemotypes. *Nature*, 566(7743):224–229, 2019.

[Osika *et al.*, 2023] Zuzanna Osika, Jazmin Zatarain Salazar, Diederik M. Roijers, Frans A. Oliehoek, and Pradeep K. Murukannaiah. What lies beyond the pareto front? a survey on decision-support methods for multi-objective optimization. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence (IJCAI-23)*, pages 6741–6749, Macao, SAR, China, 2023. International Joint Conferences on Artificial Intelligence Organization.

[Reymond *et al.*, 2024] Mathieu Reymond, Conor F. Hayes, Lander Willem, Roxana Rădulescu, Steven Abrams, Diederik M. Roijers, Enda Howley, Patrick Mannion, Niel Hens, Ann Nowé, and Pieter Libin. Exploring the pareto front of multi-objective covid-19 mitigation policies using reinforcement learning. *Expert Systems with Applications*, 249:123686, 2024.

[Russo, 2016] Daniel Russo. Simple bayesian algorithms for best arm identification. In Vitaly Feldman, Alexander Rakhlin, and Ohad Shamir, editors, *29th Annual Conference on Learning Theory*, volume 49 of *Proceedings of Machine Learning Research*, pages 1417–1418, Columbia University, New York, New York, USA, 23–26 Jun 2016. PMLR.

[Shang *et al.*, 2019] Xuedong Shang, Rianne de Heide, Emilie Kaufmann, Pierre M'enard, and Michal Valko. Fixed-confidence guarantees for bayesian best-arm identification. In *International Conference on Artificial Intelligence and Statistics*, 2019.

[Thompson, 1933] William R. Thompson. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 25(3-4):285–294, December 1933.

[Van Molle *et al.*, 2021] Pieter Van Molle, Tim Verbelen, Bert Vankeirsbilck, Jonas De Vylder, Bart Diricx, Tom Kimpe, Pieter Simoens, and Bart Dhoedt. Leveraging the bhattacharyya coefficient for uncertainty quantification in deep neural networks. *Neural Comput. Appl.*, 33(16):10259–10275, August 2021.

[Waring, 2010] Michael J Waring. Lipophilicity in drug discovery. *Expert Opinion on Drug Discovery*, 5(3):235–248, Mar 2010.

[Yahyaa and Manderick, 2015] Saba Q. Yahyaa and Bernard Manderick. Thompson sampling for multi-objective multi-armed bandits problem. In *The European Symposium on Artificial Neural Networks*, 2015.

[Zhao *et al.*, 2018] Hui Zhao, Kan Wu, and Edward Huang. Clinical trial supply chain design based on the pareto-optimal trade-off between time and cost. *IISE Transactions*, 50(6):512–524, 2018.

[Zuluaga *et al.*, 2013] Marcela Zuluaga, Guillaume Sergent, Andreas Krause, and Markus Püschel. Active learning for multi-objective optimization. In *International Conference on Machine Learning*, 2013.

[Zuluaga *et al.*, 2016] Marcela Zuluaga, Andreas Krause, and Markus Püschel. e-pal: An active learning approach to the multi-objective optimization problem. *Journal of Machine Learning Research*, 17(104):1–32, 2016.

A Empirical Evaluation

We empirically evaluate the performance of the proposed Top-Two Pareto Front Thompson Sampling (TTPFTS) algorithm on the Pareto Set Identification (PSI) task. Our evaluation is conducted using the eight synthetic Multi-objective Multi-armed Bandit (MOMAB) PSI benchmarking environments introduced in [Kone *et al.*, 2023]. These environments are designed to capture a wide range of structural properties and difficulty levels commonly encountered in MOMAB problems. For each environment, we use Gaussian reward distributions with diagonal covariance matrix $\sigma^2 I_D$ with variance $\sigma^2 = 0.25$.

In this section, we provide detailed specifications for each of the eight synthetic benchmarking environments. **Note:** To ensure completeness and reproducibility, the definitions and parameters listed below are adopted directly from the original specifications provided by [Kone *et al.*, 2023].

A.1 Detailed Synthetic Benchmarking Environment Specifications

EgeExp 1: Arms on a convex Pareto set. This environment features $K = 60$ arms and $D = 2$ objectives. The first 20 parameters $x_1, \dots, x_{20}$ are chosen to be equally spaced in the interval $[0.55, 0.95]$. For arms $i = 1, \dots, 10$, the means are defined as $\boldsymbol{\theta}_i := (x_i^2, 1/(4x_i^2))^\intercal$. The remaining arms $\boldsymbol{\theta}_{11}, \dots, \boldsymbol{\theta}_{60}$ are chosen from the set $\{(x, y) \in [0.1, 0.8]^2 : xy \leq 1/5\}$.

EgeExp 2: Group of sub-optimal arms with unique optimal dominators. Here, $K = 10$ and $D = 2$, with a Pareto set size of $|\mathcal{P}^\star| = 2$. The structure dictates that for each sub-optimal arm i, there is a unique optimal arm j such that $\boldsymbol{\theta}_i \prec \boldsymbol{\theta}_j$. The optimal arms are set as $\boldsymbol{\theta}_1 := (0.4, 0.75)^\intercal$ and $\boldsymbol{\theta}_2 := (0.75, 0.4)^\intercal$. For $i = 1, \dots, 4$, the sub-optimal arms are defined as $\boldsymbol{\theta}_{2i+1} := (0.45 + 0.2^i, 0.35 - 0.2^i)^\intercal$ and $\boldsymbol{\theta}_{2i+2} := (0.10 + 0.20^i, 0.70 - 0.20^i)^\intercal$.

EgeExp 3: Many arms on the unit circle. This setting includes a large number of arms, $K = 200$, with $D = 2$. The instance is generated as an isotropic multivariate normal. Angles $\beta_1, \dots, \beta_{20}$ are evenly spaced in $[\pi/12, \pi/2 - \pi/12]$, while $\beta_{21}, \dots, \beta_{200}$ are evenly spaced in $[\pi/2 + \pi/6, 2\pi - \pi/6]$. The mean vectors are defined as $\boldsymbol{\theta}_i := (\cos(\beta_i), \sin(\beta_i))^\intercal$.

EgeExp 4: High dimension ($D = 10$). This high-dimensional environment sets $K = 50$ and $D = 10$. The mean vectors are generated uniformly: $(\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_{30}) \sim \mathcal{U}([0.2, 0.45]^{10})^{\otimes 30}$ and $(\boldsymbol{\theta}_{31}, \dots, \boldsymbol{\theta}_{50}) \sim \mathcal{U}([0.55, 0.75]^{10})^{\otimes 20}$.

EgeExp 5: 2 clusters of arms. With $K = 20$ arms, this environment samples means from two distinct clusters: $(\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_{10}) \sim \mathcal{U}([0.2, 0.4]^2)^{\otimes 10}$ and $(\boldsymbol{\theta}_{11}, \dots, \boldsymbol{\theta}_{20}) \sim \mathcal{U}([0.5, 0.7]^2)^{\otimes 10}$.

EgeExp 6: All arms are optimal. In this setting, $K = 10$ and $D = 2$, and every arm belongs to the Pareto set. The means are defined such that for any arm i, $\theta_i^1 = 0.75 - 0.65^i$ and $\theta_i^2 = 0.25 + 0.65^i$.

EgeExp 7: All arms have the same sub-optimality gap. This environment uses $K = 22$ and $D = 2$, choosing $\boldsymbol{\Theta}$ such that $\Delta_1 = \dots = \Delta_{22}$. The means are constructed as follows: for $i = 1, \dots, 8$, $\boldsymbol{\theta}_i := (0.3 + c_i, 0.8 - c_i)^\intercal$ where $c_i = (i - 1) * c$; for $i = 9, \dots, 15$, $\boldsymbol{\theta}_i := (0.25 + c_{i-8}, 0.7 - c_{i-8})^\intercal$; and for $i = 16, \dots, 22$, $\boldsymbol{\theta}_i := \boldsymbol{\theta}_{i-7} - (0, -0.05)^\intercal$.

EgeExp 8: Geometric progression with a single optimal arm. This instance features $K = 5$ and $D = 2$, where for any $i \in \{1, \dots, 5\}$, the components are identical: $\theta_i^1 = \theta_i^2 = 0.75 - 0.25^i$. This results in a unique optimal arm.

Figure 8 provides a visual representation of the reward landscapes for the seven bi-objective maximization environments, excluding the high-dimensional EgeExp4. To facilitate reproducibility and encourage further research, we have made the full implementations of these benchmarking environments publicly available. The source code can be accessed in our GitHub repository[2].

A.2 Additional PSI Performance Metrics and Results

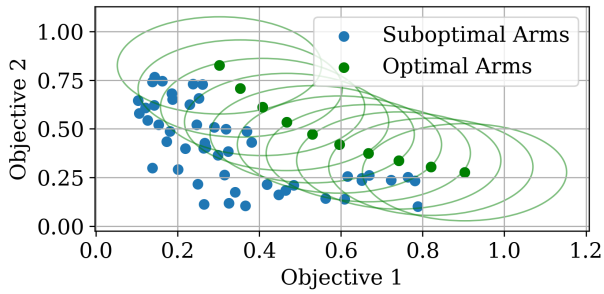
Bernoulli Metric

The first way to evaluate the quality of a Pareto set estimate $\hat{\mathcal{P}}^\star$ with respect to the actual set of Pareto optimal arms $\mathcal{P}^\star$ is to consider the estimation as a *Bernoulli* trial. The trial is considered a success if the recommendation $\hat{\mathcal{P}}^\star$ is exactly equal to the ground-truth set of Pareto optimal arms $\mathcal{P}^\star$. Otherwise, it is considered a failure. A score of 1 is associated with a success, and a score of zero is associated with a failure. These scores for successes and failures can be averaged over multiple runs of the algorithm to obtain the averaged empirical success rate of the algorithm. This yields a proportion that gives a good indication of the algorithm's ability to give perfect recommendations.

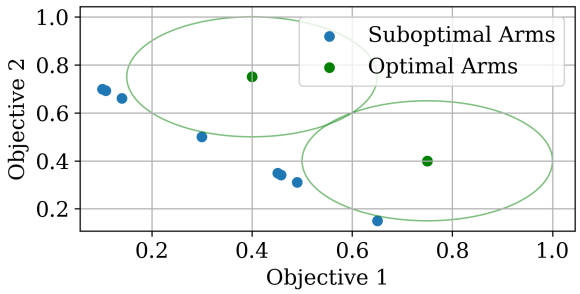
$$\mathcal{B}(\mathcal{P}^\star, \hat{\mathcal{P}}^\star) = \begin{cases} 0 & \text{if } \hat{\mathcal{P}}^\star \neq \mathcal{P}^\star \\ 1 & \text{if } \hat{\mathcal{P}}^\star = \mathcal{P}^\star \end{cases} \tag{3}$$

The Bernoulli metric $\mathcal{B}$ presented in Equation (3) is a very strict metric: only perfect Pareto set estimates $\hat{\mathcal{P}}^\star$ are rewarded. This means that if the estimate contains all but one of the Pareto optimal arms, the Bernoulli metric will be equal to 0. Similarly, if the estimate contains all Pareto optimal arms and some additional suboptimal arm, the Bernoulli metric will also be equal to

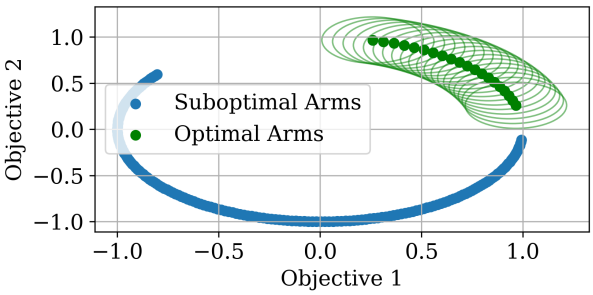
[2] https://github.com/LennertSaerens/TTPFTS



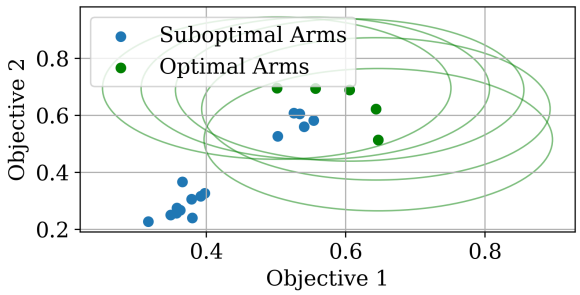
(a) EgeExp1



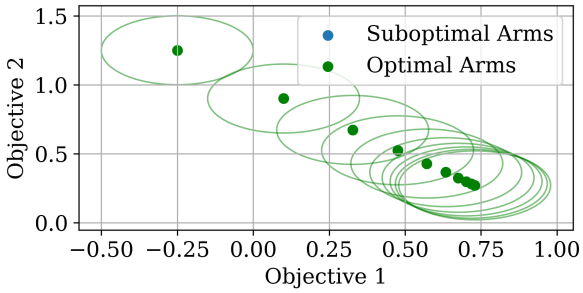
(b) EgeExp2



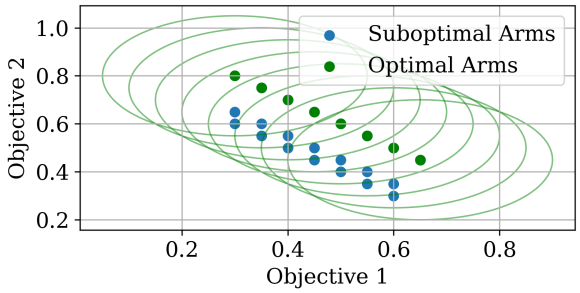
(c) EgeExp3



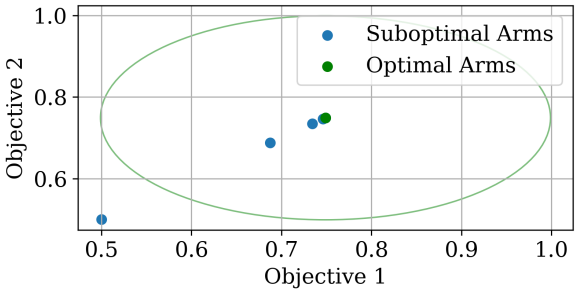
(d) EgeExp5



(e) EgeExp6



(f) EgeExp7



(g) EgeExp8

Figure 8: Visualizations of the bi-objective synthetic maximization benchmarking environments proposed in [Kone *et al.*, 2023]. We use Gaussian reward distributions with diagonal covariance matrix $\sigma^2 I_D$ with $\sigma^2 = 0.25$. For better readability, we visualize σ^2 as an ellipse around the mean only for the optimal arms. EgeExp4 cannot be visualized in the same way due to its high-dimensional nature ($D = 10$).

0. From an intuitive point of view, Pareto set estimates that are far from perfect receive the same score as estimates that are nearly perfect. Consequently, the Bernoulli metric does not provide that much insight into the acquisition process of the Pareto front by the MOMAB PSI algorithm. However, we argue that the Bernoulli metric is a strong indicator of the algorithms ability to identify the Pareto optimal arms, that is particularly useful when the goal is to perfectly identify the Pareto front.

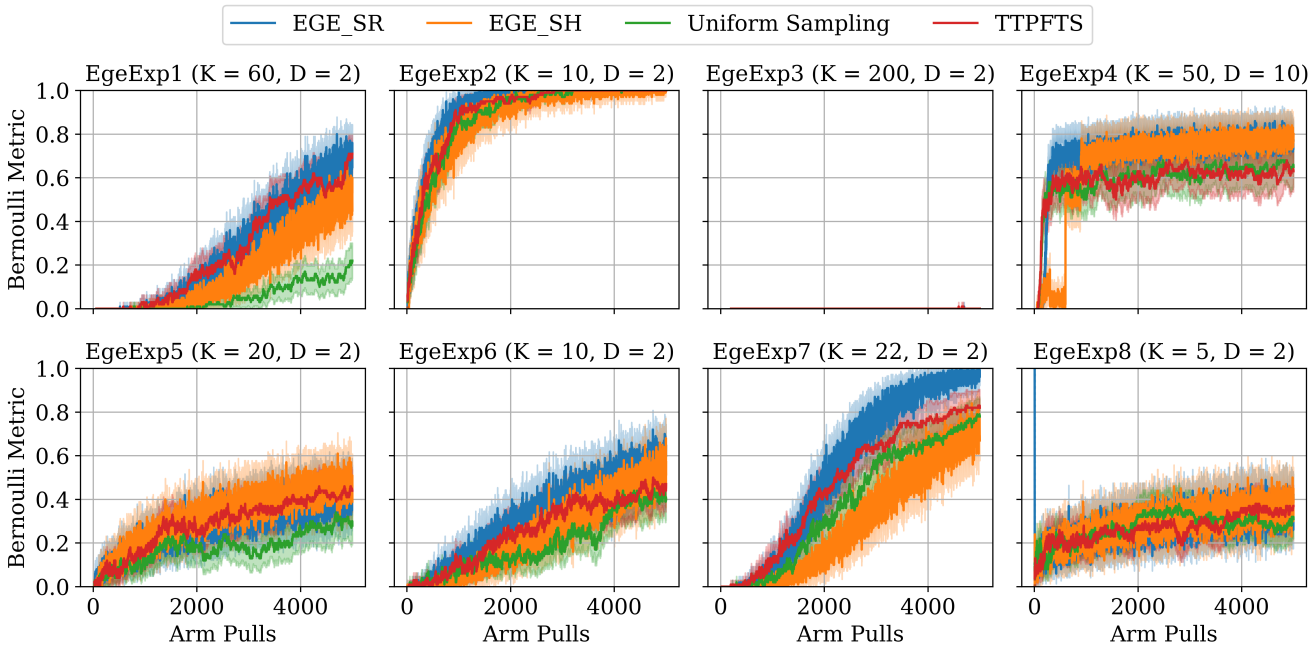


Figure 9: Performance of each of the MOMAB PSI algorithms with respect to the Bernoulli metric in each of the eight synthetic benchmarking settings proposed in [Kone *et al.*, 2023]. We use Gaussian reward distributions with diagonal covariance matrix $\sigma^2 I_D$ with $\sigma^2 = 0.25$. Results are averaged over 100 experimental trajectories, with shaded areas indicating the 95% confidence interval around the mean performance.

Figure 9 illustrates the evolution of the Bernoulli metric across the synthetic benchmarking environments. In general, these results corroborate the findings from the Jaccard metric analysis, confirming that TTPFTS achieves identification rates of similar quality to the state-of-the-art EGE-SR algorithm, despite the inherent challenge of operating without a fixed budget.

A salient observation across almost all environments (notably EgeExp 1, 2, 5, 6, and 8) is the striking contrast in stability between the methods. The EGE algorithm variants exhibit significant performance variability, characterized by jagged learning curves. This volatility stems from the fixed-budget nature of EGE: its performance is highly sensitive to the specific budget T chosen. For certain budgets, EGE may achieve high high quality Pareto set estimates, while for slightly different budgets, performance can drop precipitously. In contrast, TTPFTS demonstrates a much smoother, more monotonic improvement in success rate as it gathers more evidence. This stability highlights a key advantage of the anytime framework: in real-world scenarios where problem hardness is unknown and selecting an optimal budget is difficult, TTPFTS offers a robust and consistent alternative that does not sacrifice performance for flexibility.

The strict nature of the Bernoulli metric is most evident in EgeExp3, which features a large action space ($K = 200$). Here, all algorithms fail to achieve a non-zero success rate within the sampling limit. While this confirms the difficulty of perfectly recovering the exact Pareto set in large-scale settings, it is important to contextualize this with the Jaccard and misidentification results. As previously discussed, TTPFTS achieves the highest Jaccard scores in this environment, indicating that while no method *perfectly* identifies the Pareto set, TTPFTS consistently identifies the highest-quality approximation among all considered methods.

Finally, in the high-dimensional setting of EgeExp4 ($D = 10$), TTPFTS exhibits a more conservative learning profile compared to the EGE algorithm variants regarding the Bernoulli metric. This behavior is likely attributable to the sparsity of the action space in high dimensions, which can make the posterior separation of the top-two fronts less stable. However, this result should be interpreted alongside the Jaccard metric, where TTPFTS remains highly competitive. This divergence suggests that while TTPFTS may occasionally miss a single arm or include a near-optimal contender, resulting in a Bernoulli score of 0, the overall quality of the estimated Pareto set remains high.

Misclassification Metric

Another metric to evaluate the quality of Pareto set estimates $\hat{\mathcal{P}}^\star$ in the context of the MOMAB PSI problem is the *misclassification*. The use of this metrics was proposed in [Kone *et al.*, 2023]. This metric offers a more granular perspective than the

Bernoulli metric and complements the Jaccard metric by quantifying the extent to which individual arms are correctly classified as Pareto (sub)optimal.

Classification	Condition
True Positive (TP)	$a \in \hat{\mathcal{P}}^\star$ and $a \in \mathcal{P}^\star$
False Positive (FP)	$a \in \hat{\mathcal{P}}^\star$ and $a \notin \mathcal{P}^\star$
True Negative (TN)	$a \notin \hat{\mathcal{P}}^\star$ and $a \notin \mathcal{P}^\star$
False Negative (FN)	$a \notin \hat{\mathcal{P}}^\star$ and $a \in \mathcal{P}^\star$

Table 1: Classification of arms based on Pareto set estimate $\hat{\mathcal{P}}^\star$ and ground-truth Pareto front $\mathcal{P}^\star$.

Given a Pareto set estimate, every arm $a \in [K]$ can be classified into one of four categories, based on the comparison between the algorithm's recommendation $\hat{\mathcal{P}}^\star$ and the ground-truth Pareto front $\mathcal{P}^\star$. This is shown in Table 1. Using these classifications, we define the misclassification rate M as the proportion of arms that are incorrectly classified as Pareto (sub)optimal:

$$M(\mathcal{P}^\star, \hat{\mathcal{P}}^\star) = \frac{|\text{FP}| + |\text{FN}|}{K} \tag{4}$$

This metric can be averaged over multiple runs of the algorithm to yield a robust estimate of the average misclassification rate. Unlike the Bernoulli metric, which is binary, the misclassification metric provides a finer-grained signal on how well the algorithm is learning to distinguish optimal from suboptimal arms over time.

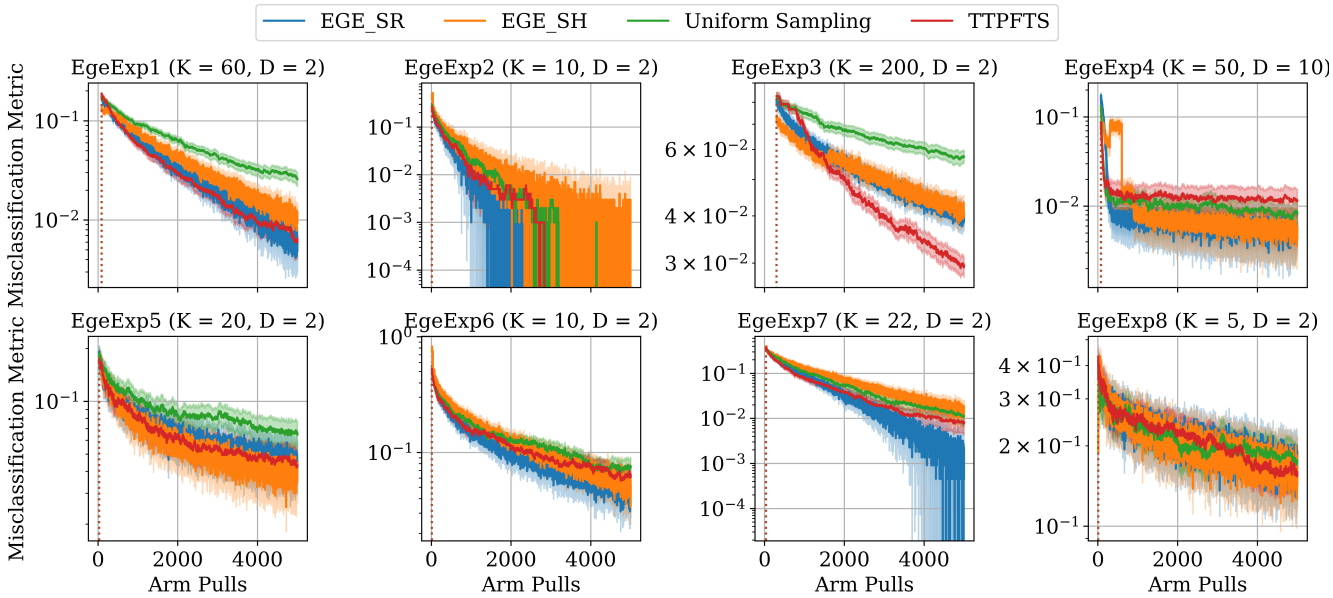


Figure 10: Performance of each of the MOMAB PSI algorithms with respect to the Misclassification metric in each of the eight synthetic benchmarking settings proposed in [Kone *et al.*, 2023]. We use Gaussian reward distributions with diagonal covariance matrix $\sigma^2 I_D$ with $\sigma^2 = 0.25$. Results are averaged over 100 experimental trajectories, with shaded areas indicating the 95% confidence interval around the mean performance.

Figure 10 presents the evolution of the misclassification rate across the eight benchmarking environments. These results strongly corroborate the trends observed in the Jaccard and Bernoulli analyses, confirming that TTPFTS delivers PSI performance comparable to the state-of-the-art EGE-SR algorithm. This parity is particularly significant given that the fixed-budget EGE baselines are optimized to minimize error at specific horizons, whereas TTPFTS operates under the stricter constraint of maintaining a valid anytime estimate.

A key advantage of the anytime framework is evident in the stability of the TTPFTS learning curves. Consistent with the Bernoulli results, the EGE variants display marked variability, particularly in EgeExp2, where the misclassification rate fluctuates significantly depending on the budget. In contrast, TTPFTS exhibits a smoother, more monotonic reduction in misclassification errors as it gathers evidence. This stability ensures reliable performance without the need for a priori knowledge of the optimal budget, a crucial feature for real-world deployment.

In terms of specific environments, EgeExp3 ($K = 200$) again highlights the efficacy of TTPFTS in large-scale settings. Here, TTPFTS achieves a distinctly lower misclassification rate compared to both EGE variants, reinforcing the conclusion from the Jaccard analysis that its posterior-driven exploration is well-suited for larger action spaces. Conversely, in the high-dimensional setting of EgeExp4 ($D = 10$), TTPFTS shows a slightly higher misclassification rate than the EGE algorithms. However, when viewed in conjunction with the high Jaccard scores reported in the main text, this suggests that while TTPFTS may misclassify a small number of arms near the decision boundary, the overall quality and utility of its recommended Pareto set remain high.

B Molecular Discovery Setting

B.1 Formalization of the Multi-Agent Nature

In the main text, we demonstrate the application of TTPFTS in exploring an ultra-large synthesis-on-demand molecular library. Unlike the synthetic benchmarks, which are evaluated in a standard single-agent framework, this chemical space is generated combinatorially, akin to a Cartesian product: a vast space of product molecules is defined by reacting all valid combinations from smaller, discrete sets of molecular building blocks, called reagents [Grygorenko *et al.*, 2020]. To clarify the mechanics of this experiment, we formalize this domain as a cooperative multi-agent MOMAB PSI problem.

Let M denote the number of reaction components. We deploy an independent TTPFTS agent for each component $c \in \{1, \dots, M\}$. Each agent c faces a distinct action space $\mathcal{A}_c$, where each arm $a_c \in \mathcal{A}_c$ corresponds to a unique chemical reagent available for that specific component.

At each time step t, every agent c simultaneously and independently selects a reagent $a_{c,t} \in \mathcal{A}_c$ according to its own TTPFTS policy. These selections form a complete combinatorial action tuple, $\boldsymbol{a}_t = (a_{1,t}, \dots, a_{M,t}) \in \mathcal{A}_1 \times \cdots \times \mathcal{A}_M$, which defines a fully specified product molecule. The properties of this assembled molecule are then evaluated, yielding a D-dimensional deterministic reward $\boldsymbol{X}_t = f(\boldsymbol{a}_t) \in \mathbb{R}^D$.

Since in the main text we are considering a three-component reaction, in our experiment $M = 3$. Furthermore, in our case, the number of objectives $D = 2$, representing Structural Similarity and LogP.

Stochasticity and Non-Stationarity. A defining feature of this formulation is the source of the reward stochasticity. From the perspective of a complete combinatorial action tuple, $\boldsymbol{a}_t$, the evaluation function $f(\boldsymbol{a}_t)$ is deterministic. However, from the localized perspective of a single agent c, the reward vector observed after pulling arm $a_{c,t}$ is stochastic, as this observed reward $\boldsymbol{X}_t$ depends on the concurrent selections made by the *other* agents, denoted as $\boldsymbol{a}_{-c,t}$.

Consequently, the reward observed by agent c for pulling reagent a_c is an evaluation of $f(a_c, \boldsymbol{a}_{-c,t})$. Because the other agents are learning from online interactions, continuously updating their posterior distributions and shifting their sampling between their respective top-two Pareto fronts, the distribution of $\boldsymbol{a}_{-c,t}$ changes over time. Thus, the expected marginal utility of a reagent a_c is not fixed; the reward distribution perceived by agent c is inherently *non-stationary*.

B.2 Pareto Front Extraction and Metric Computation.

To evaluate the algorithm's search efficiency over time, we compute the Jaccard similarity between the estimated Pareto front and the true Pareto front at each evaluation step. During the search process, every evaluated molecule and its corresponding multi-objective score are logged sequentially, preserving the exact chronological exploration order. In post-processing, we reconstruct the algorithm's estimated Pareto set at any given time step t by iteratively applying non-dominated sorting to the cumulative collection of molecules evaluated up to that point. As each new evaluation is incorporated, any previously retained molecule that becomes strictly dominated is pruned from the active set. This rolling Pareto front represents the exact non-dominated subset of the agent's search history at step t. The learning curve is then generated by calculating the Jaccard metric between this step-wise estimated set and the ground-truth Pareto optimal set.

To compute the uncertainty quantification (UQ) metric in the combinatorial molecular discovery setting, we extend the Bhattacharyya coefficient to evaluate the joint product space. Because the library consists of 94 million possible products derived from three reaction components, explicitly tracking posteriors for all combinations is computationally prohibitive. Instead, we leverage the independence of the agents: for any given product, its 6-dimensional joint posterior (representing two objectives across three components) is constructed by concatenating the independent 2-dimensional Gaussian mean and variance vectors of its constituent reagents. We construct these 6-dimensional product posteriors and immediately apply non-dominated sorting to retain only those products belonging to the local top-two Pareto fronts, discarding the vast majority of strictly dominated combinations. After aggregating these candidates, we compute the global first and second Pareto fronts for the joint space. The final UQ metric is then calculated as the average Bhattacharyya coefficient across all pairs of products between the first and second fronts.

B.3 MolPAL Baseline Configuration

The surrogate model and hyperparameters for the MolPAL baseline were selected through a systematic two-phase tuning process conducted on a reservoir sample of 200,000 molecules drawn from the 94-million quinazoline library. In the first phase, four tree-based ensemble architectures (Random Forest, Extra Trees, Gradient Boosted Regression, and Histogram-based Gradient Boosting) were compared using an offline evaluation protocol that mirrors the active learning training schedule, training each model on incremental subsets of 5,000 to 50,000 molecules and measuring R^2, RMSE, and wall-clock time

on a held-out set for both objectives. Histogram-based Gradient Boosting achieved the highest predictive accuracy (similarity $R^2 = 0.960$; LogP $R^2 = 0.667$ at 50k training points) while being 5–18× faster to train (35s vs. 200–625s), a critical advantage given that the surrogate is retrained from scratch each iteration and inference over 94M molecules requires 1.5h per cycle. In the second phase, a focused grid search over 36 Histogram-based Gradient Boosting configurations yielded the final setting (learning rate = 0.1, 500 iterations, unlimited depth, 63 leaf nodes), raising surrogate quality to $R^2 = 0.973$ for similarity and $R^2 = 0.721$ for LogP score.

Molecules were represented as Morgan fingerprints (2048-bit, radius 2) [Rogers and Hahn, 2010]. For acquisition, we employed non-dominated sorting, which ranks candidates front-by-front without scalarization. Because this strategy relies only on mean predictions, it is fully compatible with Histogram-based Gradient Boosting. Neural network surrogates were not pursued, as gradient boosting methods match or exceed Neural Network performance on fingerprint-based inputs while remaining more computationally efficient [Zhang *et al.*, 2019].

B.4 Additional Results

In the main text, we presented the results of a single representative experimental run to illustrate the efficacy of TTPFTS in exploring the ultra-large synthesis-on-demand molecular library. To demonstrate the robustness and consistency of our approach, and to verify that the reported performance is not the result of a favorable random seed, we provide visualizations of ten additional independent experimental runs in this section.

Figures 11 and 12 display the Pareto fronts identified by TTPFTS and Random Search compared to the ground-truth Pareto front, obtained via exhaustive virtual screening, for ten additional independent trajectories. These qualitative results align strongly with the quantitative Jaccard metric results reported in the main text.

Across all additional runs, TTPFTS consistently recovers the vast majority of the ground-truth Pareto optimal molecules. Notably, the algorithm achieves perfect identification of the entire Pareto front in Runs 3, 6, and 7, capturing every optimal trade-off between structural similarity and LogP. In Runs 4, 5, and 9, the algorithm misses only a very small number (one or two) of the 52 optimal molecules. This high success rate across diverse random seeds confirms that the performance illustrated in the main text is representative of the algorithm's stable behavior.

In the remaining runs where TTPFTS does not fully recover the exact ground-truth set (e.g., Runs 2, 8, and 10), the algorithm exhibits a graceful approximation behavior. The molecules that are identified in place of the missing optimal solutions lie in the immediate vicinity of the true Pareto front in the objective space. This indicates that even in slightly sub-optimal trajectories, TTPFTS successfully focuses its sampling on the most promising regions of the chemical space. Consequently, the resulting estimated Pareto set remains highly valuable for decision-making, providing drug discovery scientists with candidate molecules that are nearly indistinguishable from the optimal set in terms of their property profiles.

Consistent with the findings in the main text, the Random Search baseline fails to identify a single ground-truth Pareto optimal molecule across all ten additional runs. This stark contrast highlights the necessity of intelligent, feedback-driven exploration strategies like TTPFTS when navigating ultra-large combinatorial libraries where exhaustive virtual screening is computationally infeasible.

C Uncertainty Quantification Metric

In this section, we provide the detailed mathematical formulation used to calculate the uncertainty quantification metric $\mathcal{U}_\beta$ introduced in the main text. The metric is defined as the average Bhattacharyya coefficient (overlap) [Bhattacharyya, 1943] between the posterior distributions of arms in the estimated first Pareto front $\mathcal{F}_1^\star$ and the estimated second Pareto front $\mathcal{F}_2^\star$.

C.1 Posterior Assumptions

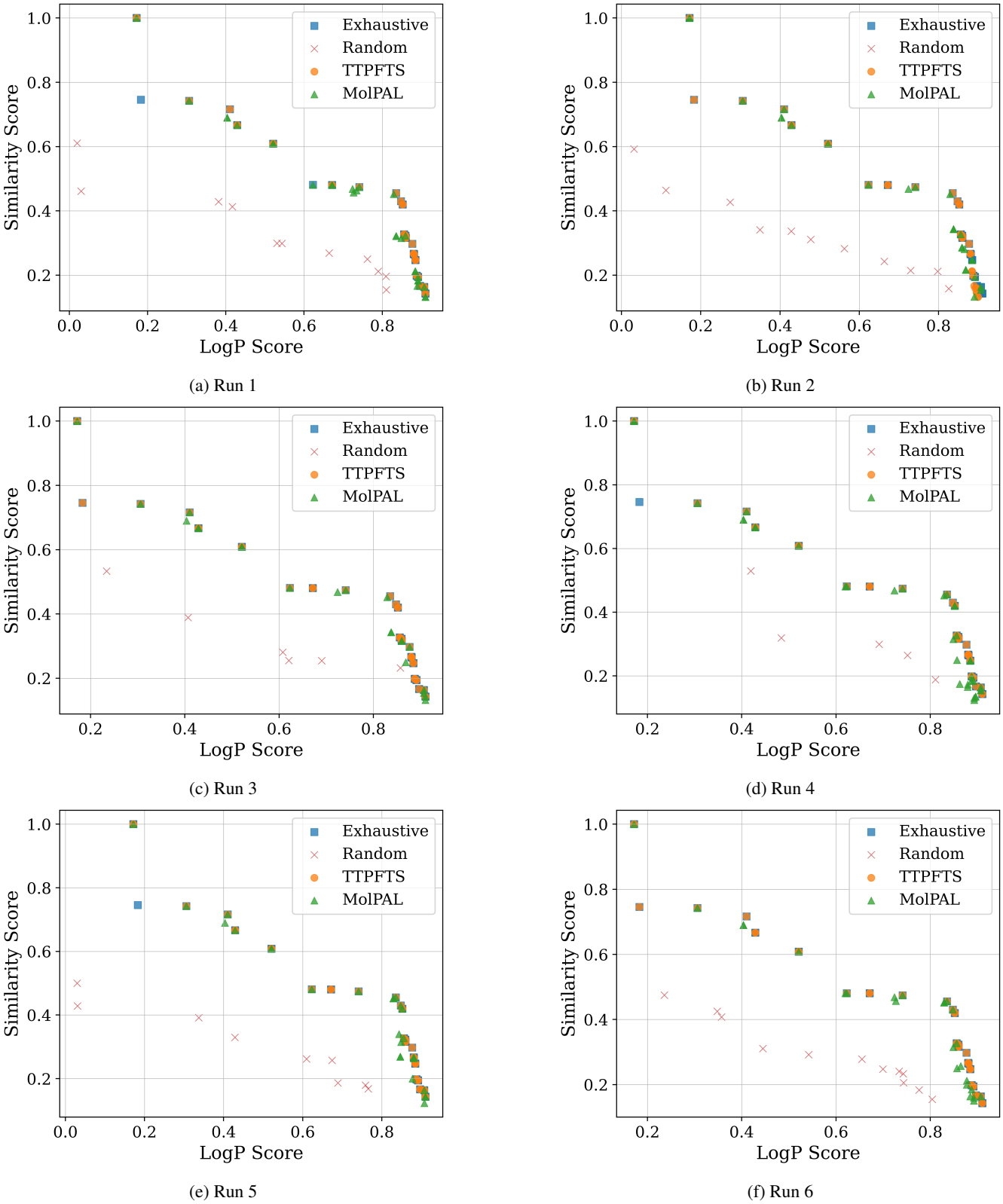
Let the bandit instance ν consist of K arms and D objectives. At any time step t, the algorithm maintains a posterior distribution over the mean reward vector for each arm $a \in [K]$, conditioned on the history $\mathcal{H}^{(t)}$. In the main text, for brevity, we refer to the overlap between arms simply as $\beta(a_i, a_j)$. Here, we rigorously define this quantity based on the underlying posterior distributions. We model the posterior belief for any arm a as a multivariate Gaussian distribution:

$$p_a(\mathbf{x} \mid \mathcal{H}^{(t)}) = \mathcal{N}(\mathbf{x}; \boldsymbol{\mu}_a, \boldsymbol{\Sigma}_a), \tag{5}$$

where $\boldsymbol{\mu}_a \in \mathbb{R}^D$ is the posterior mean vector and $\boldsymbol{\Sigma}_a \in \mathbb{R}^{D\times D}$ is the posterior covariance matrix. In our experiments, objectives are modeled independently, resulting in diagonal covariance matrices $\boldsymbol{\Sigma}_a = \mathrm{diag}(\sigma_{a,1}^2, \ldots, \sigma_{a,D}^2)$, though the derivation below holds for general covariance matrices [Fukunaga, 1990].

C.2 Bhattacharyya Distance and Coefficient

To quantify the separability between two probability distributions p_i and p_j, we utilize the *Bhattacharyya distance* $D_B(p_i, p_j)$. For two multivariate Gaussian distributions $p_i = \mathcal{N}(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$ and $p_j = \mathcal{N}(\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)$, the Bhattacharyya distance is derived in [Fukunaga, 1990] as:



(a) Run 1

(b) Run 2

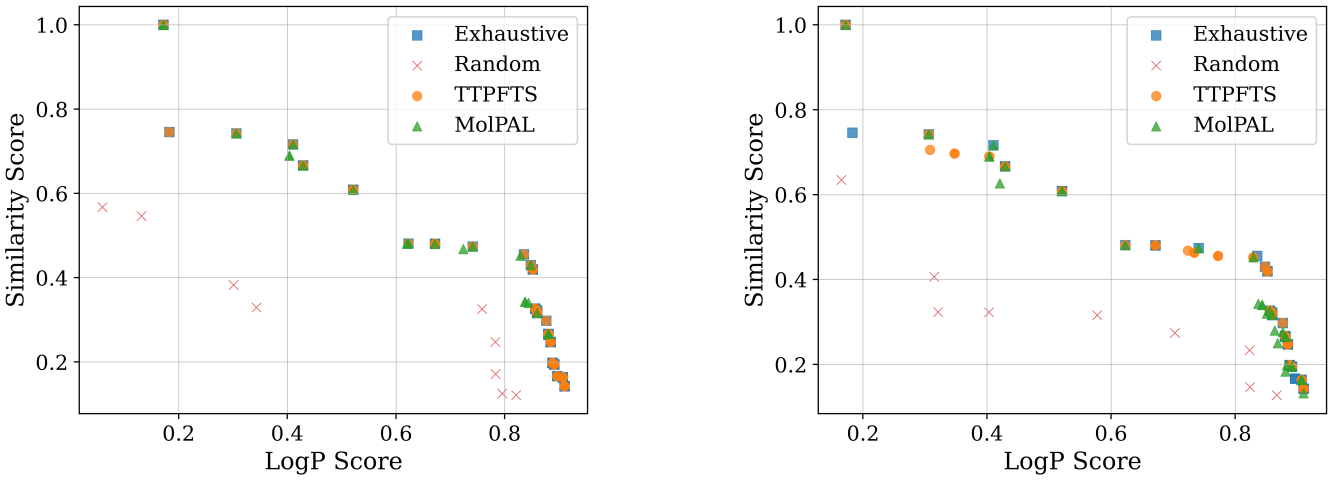
(c) Run 3

(d) Run 4

(e) Run 5

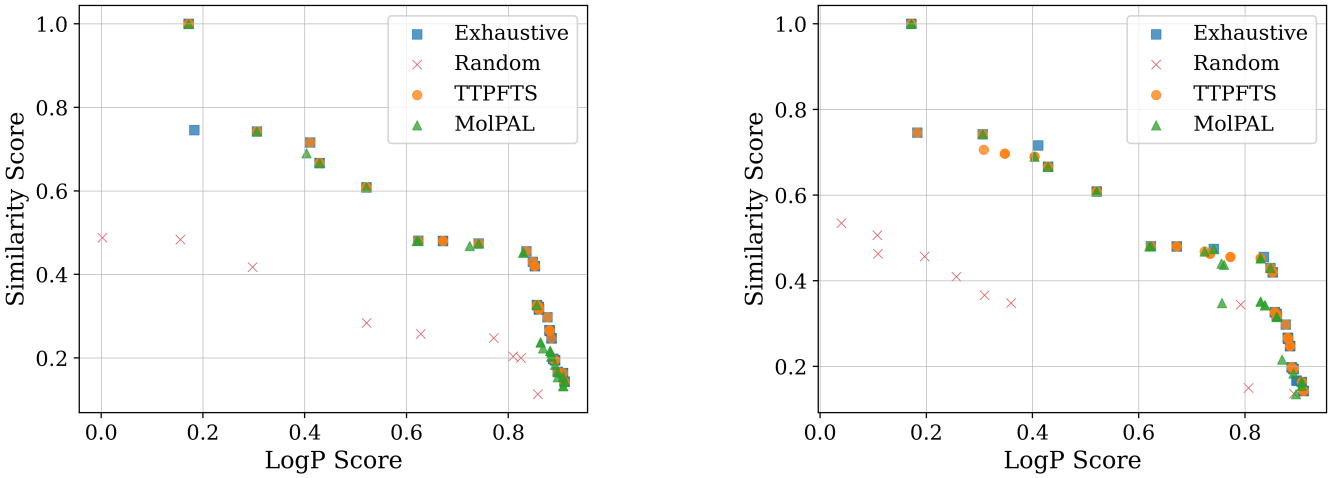
(f) Run 6

Figure 11: Pareto optimal molecules identified by the Random Search baseline, MolPAL, and TTPFTS, compared to the ground-truth Pareto set of optimal molecules obtained through exhaustive virtual screening of the synthesis-on-demand molecular library. (Runs 1–6)



(a) Run 7

(b) Run 8



(c) Run 9

(d) Run 10

Figure 12: Pareto optimal molecules identified by the Random Search baseline, MolPAL and TTPFTS, compared to the ground-truth Pareto set of optimal molecules obtained through exhaustive virtual screening of the synthesis-on-demand molecular library. (Runs 7–10)

$$D_B(p_i, p_j) = \underbrace{\frac{1}{8}(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)}_{\text{Mahalanobis term}} + \underbrace{\frac{1}{2}\ln\left(\frac{\det(\boldsymbol{\Sigma})}{\sqrt{\det(\boldsymbol{\Sigma}_i)\det(\boldsymbol{\Sigma}_j)}}\right)}_{\text{Determinant term}}, \tag{6}$$

where $\boldsymbol{\Sigma}$ represents the average covariance matrix of the two distributions:

$$\boldsymbol{\Sigma} = \frac{\boldsymbol{\Sigma}_i + \boldsymbol{\Sigma}_j}{2}. \tag{7}$$

The *Bhattacharyya coefficient* β between two arms a_i and a_j, which represents the geometric overlap between their underlying posterior distributions p_i and p_j, is derived directly from the Bhattacharyya distance:

$$\beta(a_i, a_j) \triangleq \exp\left(-D_B(p_i, p_j)\right). \tag{8}$$

The coefficient ranges from 0 to 1, where 0 implies no overlap, i.e. perfect separation, and 1 implies identical distributions, indicating maximum uncertainty regarding the arms' ordering.

C.3 Aggregate Uncertainty Measure

The total uncertainty metric $\mathcal{U}_\beta$ is computed by averaging the Bhattacharyya coefficients over all pairwise combinations of arms between the first and second Pareto fronts.

Let $\hat{\mathcal{P}}_1^\star$ denote the set of arms in the estimated first Pareto front, based on posterior means, and $\hat{\mathcal{P}}_2^\star$ denote the set of arms in the estimated second Pareto front. The uncertainty measure is calculated as:

$$\mathcal{U}_\beta\left(\hat{\mathcal{P}}_1^\star, \hat{\mathcal{P}}_2^\star\right) = \frac{1}{|\hat{\mathcal{P}}_1^\star||\hat{\mathcal{P}}_2^\star|} \sum_{a_i \in \hat{\mathcal{P}}_1^\star} \sum_{a_j \in \hat{\mathcal{P}}_2^\star} \beta(a_i, a_j) \tag{9}$$

This aggregation captures the average overlap between the currently believed best set, i.e. the first front, and the challenger set, i.e. the second front. A high value indicates that the posterior masses of the optimal and suboptimal arms are heavily intertwined, requiring further exploration to resolve the decision boundary. Conversely, a low value indicates that the probability masses have separated, signifying low uncertainty regarding the current estimate of the Pareto set. The source code containing the full implementation of the proposed uncertainty quantification metric can be accessed in our GitHub repository[3].

C.4 Comparison of Uncertainty Quantification Metrics

In the main text, we introduce an uncertainty quantification (UQ) metric based on the Bhattacharyya coefficient to serve as a ground-truth-independent proxy for algorithmic convergence. To contextualize and justify this design choice, this section presents a comparative analysis of our proposed metric against two standard information-theoretic alternatives: Symmetric Kullback-Leibler (KL) divergence and Posterior (Differential) Entropy.

All three metrics evaluate the posterior distributions of the arms estimated to be on the top-two Pareto fronts (denoted as $\mathcal{F}_1^\star$ and $\mathcal{F}_2^\star$). The evaluation was conducted across the eight synthetic benchmarking environments (EgeExp1–8). For each environment, we executed 100 independent runs of the TTPFTS algorithm (using an uninformative t-distributed prior) for 5,000 arm pulls, saving and evaluating the posterior states at 100-step intervals.

- **Bhattacharyya Coefficient (BC):** Our proposed metric computes the average pairwise overlap between distributions of arms in $\mathcal{F}_1^\star$ and arms in $\mathcal{F}_2^\star$. It is bounded in $[0, 1]$, where 0 indicates no overlap and 1 indicates identical distributions. [Bhattacharyya, 1943]
- **Symmetric KL-Divergence:** This metric computes the average pairwise symmetric KL-divergence, $\frac{1}{2}\left(D_{\mathrm{KL}}(p \parallel q) + D_{\mathrm{KL}}(q \parallel p)\right)$, between the same $\mathcal{F}_1^\star \times \mathcal{F}_2^\star$ arm pairs. It is unbounded and grows as the posteriors separate. [Kullback and Leibler, 1951]
- **Posterior Entropy:** This metric calculates the average differential entropy over all arms in the union $\mathcal{F}_1^\star \cup \mathcal{F}_2^\star$. Unlike BC and KL-divergence, it measures the overall posterior uncertainty rather than the pairwise separation between the top-two fronts. [MacKay, 1992]

To determine how well each metric acts as a proxy for actual performance, we calculated the Pearson correlation coefficient (r) between the metric's time series and the ground-truth Jaccard similarity across the experimental runs. A reliable UQ metric should exhibit a strong, consistent correlation with the Jaccard score across diverse environments. The results of this analysis are visualized in Figure 13.

The Bhattacharyya coefficient proves to be the strongest and most consistent tracker of convergence. It achieves the highest median absolute correlation ($r = -0.80$) and maintains the expected negative correlation, meaning that as overlap decreases,

[3] https://github.com/LennertSaerens/TTPFTS

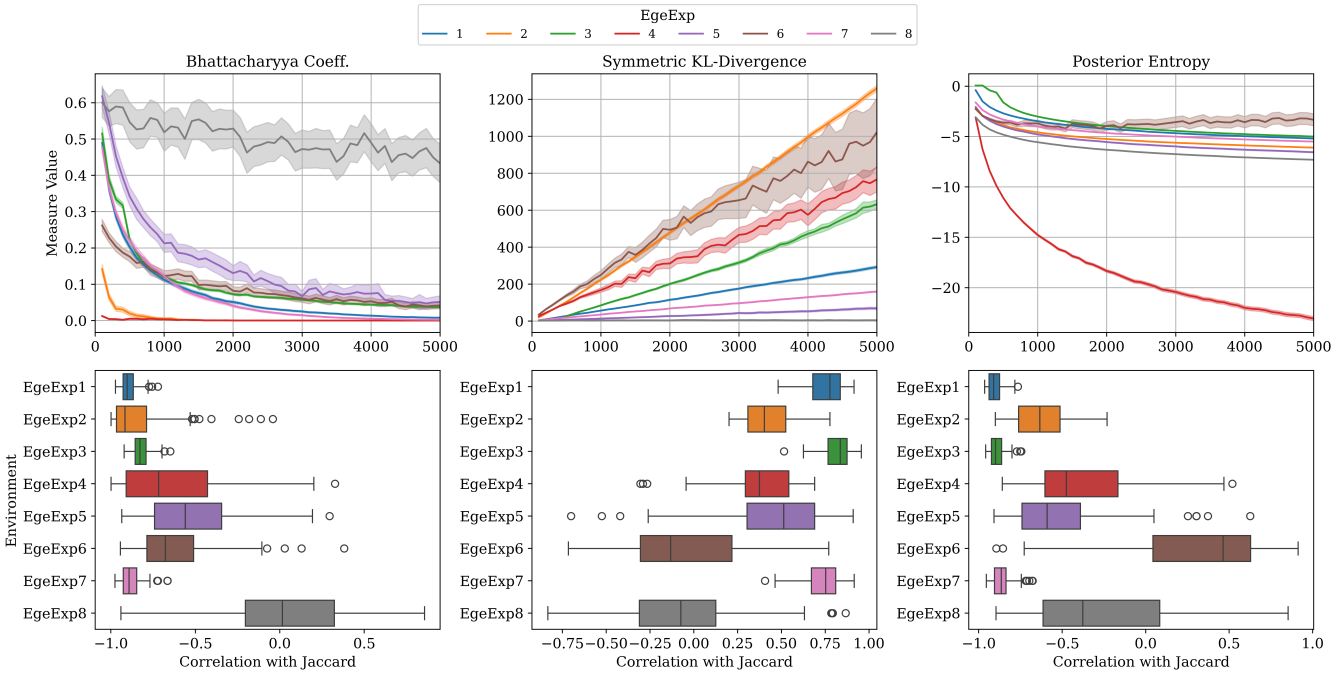


Figure 13: Comparison of uncertainty quantification metrics. **Top row:** Evolution of the Bhattacharyya Coefficient, Symmetric KL-Divergence, and Posterior Entropy over 5,000 arm pulls (mean $\pm$ 95% CI over 100 runs). **Bottom row:** Distribution of per-run Pearson correlation coefficients between each metric and the ground-truth Jaccard similarity, grouped by environment.

identification accuracy increases, across seven of the eight environments. In contrast, Symmetric KL-Divergence is the weakest overall predictor (median $r = +0.55$). Posterior Entropy demonstrates a strong median correlation ($r = -0.70$) but exhibits high variance across different environments (standard deviation of 0.48).

The limitations of Posterior Entropy and KL-Divergence become apparent in specific edge cases, most notably in the Ege-Exp6 environment, where all arms are Pareto-optimal. Because entropy measures absolute posterior spread rather than the separation between fronts, it naturally decreases as the algorithm pulls arms and variances shrink, even if the fronts have not been accurately separated. This structural flaw leads to a problematic sign flip in EgeExp6 ($r = +0.46$). KL-Divergence similarly struggles in this environment and in EgeExp8, yielding near-zero or incorrect-sign correlations.

D Theoretical Guarantees

D.1 Assumptions

Assumption 1 (Finite arms)**.** *K is finite.*

Assumption 2 (Reward & model regularity)**.** *For each arm $a \in [K]$, the rewards $\boldsymbol{X} \sim \nu_a$ are i.i.d. with mean vector $\boldsymbol{\theta}_a$. The reward distributions belong to a model class for which Bayesian posterior consistency holds (e.g., priors with full support over the parameter space). Specifically, as the number of times arm a is selected $N_a \to \infty$, the posterior mean $\hat{\boldsymbol{\mu}}_a$ converges to the true mean $\boldsymbol{\theta}_a$ almost surely (and thus in probability).*

Assumption 3 (Strict Pareto gaps)**.** *For every suboptimal arm $s \in \mathcal{S}$, there exists at least one optimal arm $a^* \in \mathcal{P}^\star$ such that a^* strongly dominates s (denoted $\boldsymbol{\theta}_{a^*} \succ \boldsymbol{\theta}_s$). That is, $\theta_{a^*,d} > \theta_{s,d}$ for all objectives d. We assume a positive margin of separation sufficient to distinguish this domination.*

Assumption 4 (Distinct optimal arms)**.** *Every optimal arm in the true Pareto set possesses a strictly unique expected reward vector. Formally, for any two distinct optimal arms $i, j \in \mathcal{P}^\star$ where $i \neq j$, we assume $\boldsymbol{\theta}_i \neq \boldsymbol{\theta}_j$.*

Lemma 1 (Infinite Exploration)**.** *Given that the TTPFTS algorithm utilizes a prior with full support over $\mathbb{R}^D$ (e.g., Gaussian) and a selection parameter $\rho \in (0, 1)$, every arm $k \in [K]$ is pulled infinitely often almost surely. That is,*

$$\mathbb{P}\left(\lim_{t\to\infty} N_k^{(t)} = \infty\right) = 1.$$

Proof. We proceed by contradiction. Assume there exists an arm k that is pulled only a finite number of times. This implies there exists a finite time step T_0 such that for all $t > T_0$, arm k is never selected. Consequently, the posterior distribution for

arm k, denoted $p_k^{(t)}$, becomes static (frozen) for all $t > T_0$. Since the frozen posterior has full support, finite samples cannot reduce the posterior variance to zero.

Let $\mathcal{A}_\infty = [K] \setminus \{k\}$ be the set of arms that are pulled infinitely often. As $t \to \infty$, by the consistency of the posterior (Assumption 2), the posterior mass for any active arm $j \in \mathcal{A}_\infty$ concentrates around its true mean $\boldsymbol{\theta}_j$. Therefore, there exists a vector $\mathbf{C} \in \mathbb{R}^D$ (strictly larger than all $\boldsymbol{\theta}_j$) and a time $T_1 > T_0$ such that for all $t > T_1$, the samples from all active arms are bounded by $\mathbf{C}$ with high probability. Specifically, for any $\delta > 0$:

$$\mathbb{P}(\forall j \in \mathcal{A}_\infty : \tilde{\boldsymbol{\theta}}_j^{(t)} < \mathbf{C} \mid \mathcal{H}^{(t-1)}) \geq 1 - \delta. \tag{10}$$

Consider the frozen arm k. Since its posterior has full support on $\mathbb{R}^D$, there exists a strictly positive constant probability $\epsilon > 0$ that a sample $\tilde{\boldsymbol{\theta}}_k^{(t)}$ drawn from this frozen distribution exceeds the bounding vector $\mathbf{C}$ in all objectives simultaneously:

$$\epsilon = \mathbb{P}(\tilde{\boldsymbol{\theta}}_k^{(t)} > \mathbf{C} \mid \mathcal{H}^{(t-1)}) > 0. \tag{11}$$

Crucially, because the posterior for arm k is fixed for $t > T_0$, ϵ is constant and does not vanish.

Now, consider the event D_t at time $t > T_1$ where arm k samples a value larger than $\mathbf{C}$ while all active arms sample values smaller than $\mathbf{C}$. The conditional probability of this event is bounded below:

$$\mathbb{P}(D_t \mid \mathcal{H}^{(t-1)}) \geq \epsilon(1 - \delta). \tag{12}$$

If event D_t occurs, arm k strongly dominates all other sampled arms and is therefore included in the first Pareto front $\mathcal{F}_1^\star$. The algorithm selects an arm uniformly from this first front with probability ρ. Thus, the probability of pulling arm k at time t is lower-bounded by:

$$\mathbb{P}(a_t = k \mid \mathcal{H}^{(t-1)}) \geq \frac{\rho}{|\mathcal{F}_1^\star|} \cdot \epsilon(1 - \delta) \geq \frac{\rho}{K}\epsilon(1 - \delta) := \gamma > 0. \tag{13}$$

Since the conditional probabilities are bounded below by a positive constant γ for all $t > T_1$, the sum of these probabilities diverges:

$$\sum_{t=T_1}^{\infty} \mathbb{P}(a_t = k \mid \mathcal{H}^{(t-1)}) = \infty. \tag{14}$$

By the conditional second Borel-Cantelli lemma (Lévy's extension), the divergence of this sum implies that arm k is selected infinitely often almost surely. This contradicts the assumption that arm k is never pulled after T_0. Therefore, every arm must be pulled infinitely often. □

D.2 Proof of Asymptotic Correctness

We now provide the full proof for Theorem 1:

Proof. Fix an arbitrary suboptimal arm $s \in \mathcal{S}$. By Assumption 3, there exists a dominating arm $a^* \in \mathcal{P}^\star$ such that $\boldsymbol{\theta}_{a^*} \succ \boldsymbol{\theta}_s$. Under Lemma 1, all arms are pulled infinitely often, implying $N_{a^*}^{(t)} \to \infty$ and $N_s^{(t)} \to \infty$ almost surely as $t \to \infty$.

By Assumption 2, the posterior means concentrate around their true values. Thus,

$$\hat{\boldsymbol{\mu}}_{a^*}^{(t)} \to \boldsymbol{\theta}_{a^*} \quad \text{and} \quad \hat{\boldsymbol{\mu}}_s^{(t)} \to \boldsymbol{\theta}_s$$

almost surely. Since $\boldsymbol{\theta}_{a^*} \succ \boldsymbol{\theta}_s$ with a strictly positive margin, for sufficiently large t, the ordering is preserved by the estimates, and we have $\hat{\boldsymbol{\mu}}_{a^*}^{(t)} \succ \hat{\boldsymbol{\mu}}_s^{(t)}$ almost surely. Consequently, the probability of the estimated means failing to reflect this dominance decays to zero:

$$\lim_{t\to\infty} \mathbb{P}\big(\hat{\boldsymbol{\mu}}_{a^*}^{(t)} \not\succ \hat{\boldsymbol{\mu}}_s^{(t)}\big) = 0.$$

Recall that for a suboptimal arm s to be included in the estimated Pareto set $\hat{\mathcal{P}}_t^\star$, it must not be dominated by *any* other arm in the current estimation. Therefore, if $s \in \hat{\mathcal{P}}_t^\star$, it specifically implies that s is not dominated by a^*:

$$\{s \in \hat{\mathcal{P}}_t^\star\} \subseteq \{\hat{\boldsymbol{\mu}}_{a^*}^{(t)} \not\succ \hat{\boldsymbol{\mu}}_s^{(t)}\}.$$

Taking probabilities and limits, we obtain:

$$\mathbb{P}\big(s \in \hat{\mathcal{P}}_t^\star\big) \leq \mathbb{P}\big(\hat{\boldsymbol{\mu}}_{a^*}^{(t)} \not\succ \hat{\boldsymbol{\mu}}_s^{(t)}\big) \xrightarrow{t\to\infty} 0.$$

Since the set of suboptimal arms $\mathcal{S}$ is finite (Assumption 1), we apply the union bound:

$$\mathbb{P}\big(\hat{\mathcal{P}}_t^\star \cap \mathcal{S} \neq \varnothing\big) = \mathbb{P}\left(\bigcup_{s\in\mathcal{S}} \{s \in \hat{\mathcal{P}}_t^\star\}\right) \leq \sum_{s\in\mathcal{S}} \mathbb{P}\big(s \in \hat{\mathcal{P}}_t^\star\big).$$

As $t \to \infty$, each term in the finite sum tends to zero, concluding that:

$$\lim_{t\to\infty} \mathbb{P}\big(\hat{\mathcal{P}}_t^\star \cap \mathcal{S} \neq \varnothing\big) = 0.$$

While the previous argument establishes *Soundness* (no suboptimal arms are included), asymptotic correctness also requires *Completeness* (all optimal arms are included). We proceed similarly by contradiction. Consider an arbitrary optimal arm $a^* \in \mathcal{P}^\star$. For a^* to be excluded from the estimate $\hat{\mathcal{P}}^\star$, there must exist some other arm $j \in [K]$ that strongly dominates it in the current estimation, i.e., $\hat{\boldsymbol{\mu}}_{a^*}^{(t)} \prec \hat{\boldsymbol{\mu}}_j^{(t)}$. By the definition of Pareto optimality, in the ground-truth, no arm j strongly dominates a^* (i.e., $\boldsymbol{\theta}_j \prec \boldsymbol{\theta}_{a^*}$). Furthermore, by Assumption 4, no other arm shares the exact expected reward vector of a^*. Under Assumption 2 and Lemma 1, as $t \to \infty$, the posterior means converge almost surely to the true means: $\hat{\boldsymbol{\mu}}_j^{(t)} \to \boldsymbol{\theta}_j$ and $\hat{\boldsymbol{\mu}}_{a^*}^{(t)} \to \boldsymbol{\theta}_{a^*}$. Consequently, the event that an estimated mean vector $\hat{\boldsymbol{\mu}}_j^{(t)}$ erroneously dominates $\hat{\boldsymbol{\mu}}_{a^*}^{(t)}$ becomes impossible in the limit. Taking the union bound over the finite set of optimal arms $\mathcal{P}^\star$, the probability of excluding any optimal arm vanishes:

$$\lim_{t\to\infty} \mathbb{P}\big(\exists a^* \in \mathcal{P}^\star : a^* \notin \hat{\mathcal{P}}_t^\star\big) = 0.$$

Combining this with the result for suboptimal arms, we conclude that the estimator is asymptotically consistent:

$$\lim_{t\to\infty} \mathbb{P}(\hat{\mathcal{P}}_t^\star \neq \mathcal{P}^\star) = 0.$$

□

While Theorem 1 establishes asymptotic consistency, deriving explicit finite-time rates of the error probability requires characterizing the diffusion of the posterior mass for top-two sampling in the multi-objective setting. In the single-objective domain, analyses by Russo [Russo, 2016] and subsequent works [Shang *et al.*, 2019] suggest that the error probability for top-two strategies decays exponentially with the number of samples, scaled by the problem complexity. We hypothesize that a similar exponential decay holds for TTPFTS, governed by the multi-objective Pareto gaps defined in Assumption 3.

E Limitations

We acknowledge several theoretical, empirical, and computational limitations that provide avenues for future research.

Theoretical Guarantees and Finite-Time Bounds. While Theorem 1 establishes the asymptotic correctness of TTPFTS, deriving explicit finite-time error bounds remains an important direction for future work. Historically, finite-time guarantees for both standard Thompson Sampling [Thompson, 1933; Agrawal and Goyal, 2013] and Top-Two Thompson Sampling [Russo, 2016; Shang *et al.*, 2019] took years to emerge following their initial asymptotic proofs. In this paper, we focus on establishing foundational theoretical soundness alongside immediate practical utility. We validate this utility through strong empirical performance against state-of-the-art baselines, the introduction of a real-time uncertainty metric, and a deployment in a complex drug discovery application.

Robustness to Prior Selection. The theoretical guarantee of infinite exploration (Lemma 1), which is necessary for asymptotic correctness, relies heavily on the algorithm utilizing a prior with full support over an unbounded space. If a bounded prior, such as a Beta distribution, were employed, the algorithm might permanently ignore temporarily "frozen" arms if their bounded support fails to overlap with the empirical optimal arms. Consequently, users must exercise caution; applying TTPFTS with strictly bounded priors in domains with bounded rewards may compromise the algorithm's asymptotic correctness.

Structural Dilution of the Uncertainty Metric. The proposed uncertainty quantification metric aggregates the Bhattacharyya coefficients by averaging them over all pairwise combinations between the top-two Pareto fronts. In the multi-objective space, arms that are distant from one another naturally exhibit near-zero overlap. Hence, as the sizes of the Pareto fronts increase, the number of localized (overlapping) pairs grows approximately linearly, whereas the total number of pairs (the denominator) grows quadratically. This structural dynamic inherently shrinks the aggregate metric as the Pareto front widens. Consequently, this dilution effect could mask localized areas of high uncertainty, which adds the need for nuance to the claim that the metric is environment-independent. Future iterations of this metric might explore maximum overlap or localized aggregation methods to provide a more robust and uniform stopping criterion.

Empirical Hardness and Factors Influencing Performance. Empirically, the performance of TTPFTS degrades in environments characterized by geometrically vanishing sub-optimality gaps, as evidenced by the results on the EgeExp8 benchmark. In such difficult settings, distinguishing the single optimal arm from near-optimal challengers requires a significantly larger sampling budget, causing the algorithm to struggle in forming a stable, high-confidence estimate early in the process.

Computational Scalability. Finally, the computational efficiency of TTPFTS scales differently than simple fixed-budget methods. The algorithm requires continuous non-dominated sorting to update the first and second Pareto fronts, as well as posterior sampling for every arm at each time step. While this computational overhead is manageable for the settings explored in this paper (including the implicit 94-million combinatorial space), the cost of exact non-dominated sorting scales poorly as the number of arms K or the number of objectives D becomes extremely large. Deploying TTPFTS in environments with massive, unstructured action spaces may require approximate sorting mechanisms or more aggressive arm-pruning heuristics.

References

[Agrawal and Goyal, 2013] Shipra Agrawal and Navin Goyal. Further optimal regret bounds for thompson sampling. In Carlos M. Carvalho and Pradeep Ravikumar, editors, *Proceedings of the Sixteenth International Conference on Artificial Intelligence and Statistics*, volume 31 of *Proceedings of Machine Learning Research*, pages 99–107. PMLR, 2013.

[Bhattacharyya, 1943] Anil Kumar Bhattacharyya. On a measure of divergence between two statistical populations defined by their probability distributions. *Bulletin of the Calcutta Mathematical Society*, 35:99–109, 1943.

[Fukunaga, 1990] Keinosuke Fukunaga. *Introduction to Statistical Pattern Recognition*. Academic Press, San Diego, CA, 2 edition, 1990.

[Grygorenko *et al.*, 2020] Oleksandr O. Grygorenko, Dmytro S. Radchenko, Igor Dziuba, Alexander Chuprina, Kateryna E. Gubina, and Yurii S. Moroz. Generating multibillion chemical space of readily accessible screening compounds. *iScience*, 23(11):101681, 2020.

[Kone *et al.*, 2023] Cyrille Kone, Emilie Kaufmann, and Laura Richert. Bandit pareto set identification: the fixed budget setting. In *International Conference on Artificial Intelligence and Statistics*, 2023.

[Kullback and Leibler, 1951] Solomon Kullback and Richard A. Leibler. On information and sufficiency. *The Annals of Mathematical Statistics*, 22(1):79–86, 1951.

[MacKay, 1992] David J. C. MacKay. Information-based objective functions for active data selection. *Neural Computation*, 4(4):590–604, 1992.

[Rogers and Hahn, 2010] David Rogers and Mathew Hahn. Extended-connectivity fingerprints. *Journal of Chemical Information and Modeling*, 50(5):742–754, 2010.

[Russo, 2016] Daniel Russo. Simple bayesian algorithms for best arm identification. In Vitaly Feldman, Alexander Rakhlin, and Ohad Shamir, editors, *29th Annual Conference on Learning Theory*, volume 49 of *Proceedings of Machine Learning Research*, pages 1417–1418, Columbia University, New York, New York, USA, 23–26 Jun 2016. PMLR.

[Shang *et al.*, 2019] Xuedong Shang, Rianne de Heide, Emilie Kaufmann, Pierre M'enard, and Michal Valko. Fixed-confidence guarantees for bayesian best-arm identification. In *International Conference on Artificial Intelligence and Statistics*, 2019.

[Thompson, 1933] William R. Thompson. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 25(3-4):285–294, December 1933.

[Zhang *et al.*, 2019] Jian Zhang, Damian Mucs, Ulf Norinder, and Filip Svensson. Lightgbm: An effective and scalable algorithm for prediction of chemical toxicity–application to the tox21 and mutagenicity data sets. *Journal of Chemical Information and Modeling*, 59(10):4150–4158, 2019.