# HypBO: Accelerating Black-Box Scientific Experiments Using Experts' Hypotheses*

**Abdoulatif Cissé**[1,2] , **Xenophon Evangelopoulos**[1,2] , **Sam Carruthers**[1,2] , **Vladimir V. Gusev**[3] and **Andrew I. Cooper**[1,2]

[1]Department of Chemistry, University of Liverpool, England, UK
[2]Leverhulme Research Centre for Functional Materials Design, University of Liverpool, England, UK
[3]Department of Computer Science, University of Liverpool, England, UK
{abdoulatif.cisse, evangx, sgscarru, vladimir.gusev, aicooper}@liverpool.ac.uk

## Abstract

Robotics and automation offer massive acceleration for solving intractable, multivariate scientific problems such as materials discovery, but the available search spaces can be dauntingly large. Bayesian optimization has emerged as a popular sample-efficient optimization engine, thriving in tasks where no analytic form of the target function/property is known. Here, we exploit expert human knowledge in the form of hypotheses to direct Bayesian searches more quickly to promising regions of chemical space. Previous methods have used underlying distributions derived from existing experimental measurements, which is unfeasible for new, unexplored scientific tasks. Also, such distributions cannot capture intricate hypotheses. Our proposed method uses expert human hypotheses to generate improved seed samples. Unpromising seeds are automatically discounted, while promising seeds are used to augment the surrogate model data, thus achieving better-informed sampling. This process continues in a global versus local search fashion, organized in a bilevel optimization framework. We validate the performance of our method on a range of synthetic functions and demonstrate its practical utility on a real chemical design task where the use of expert hypotheses accelerates the search performance significantly.

## 1 Introduction

Bayesian Optimization (BO) is a valuable tool for optimizing experiments in chemistry and materials science, where experiments are costly and time-consuming [Shahriari *et al.*, 2016]. Experimental design methods often involve exhaustive exploration of the parameter space. By contrast, BO offers an efficient framework leveraging Bayesian inference to guide the iterative exploration of the space, ultimately maximizing the target experiment property [Jones *et al.*, 1998]. Formally, BO aims to find the global optimum in the following problem:

$$x^* = \underset{x \in \mathcal{X}}{argmax}\, f(x), \quad (1)$$

where $f : \mathcal{X} \rightarrow \mathbb{R}$ is a continuous function over the $d$-dimensional input space $\mathcal{X} \in \mathbb{R}^d$. Generally, the underlying analytical form of $f(\cdot)$ is unknown, making it a black-box function. The core principle of BO lies in the construction of a probabilistic model, typically a Gaussian Process (GP) [Rasmussen and Williams, 2006], which serves as a *surrogate model* for $f(\cdot)$. This surrogate model is updated iteratively as new experimental data become available, allowing for the refinement of target predictions. The model's uncertainty is quantified, and an *acquisition function* is employed to select the next set of experimental parameters to evaluate, balancing exploration (sampling in unexplored regions) and exploitation (focusing on promising regions).

Injecting domain-specific knowledge into BO to boost optimization performance has gained significant recent attention, especially for scientific tasks [Ramachandran *et al.*, 2020], aiming to tap into previously unused human expertise. In particular, recent studies have used expert knowledge as user-specified priors over possible optima to guide the search toward promising regions [Hvarfner *et al.*, 2022; Li *et al.*, 2020]. While this has shown promising performance in various tasks, it is difficult in many scientific problems to realize external knowledge in the form of a prior distribution. Furthermore, the optimization landscapes of such problems often resemble a needle-in-a-haystack manifold [Siemenn *et al.*, 2023], and inaccurate prior knowledge distributions can introduce negative bias in the problem and quickly degrade performance. More recently, human-in-the-loop approaches have emerged where an interactive optimization framework enables experts to implicitly add chemical or medical knowledge to the problem in the form of feedback on the quality of the samples within the "experimental loop" [Martinelli *et al.*, 2023; Sundin *et al.*, 2018]. However, this knowledge is implicit and sample-specific and can often lead to local optima entrapment. Another category of methods introduces domain-specific knowledge in the form of hard constraints in the problem [Hernández-Lobato *et al.*, 2015], which can, however, over-restrict the search in practice.

In this paper, we propose a novel approach to inject domain knowledge using input from domain experts to direct the

*Supplementary Material (SM) for this work can be found in the extended version of the paper at https://arxiv.org/abs/2308.11787.

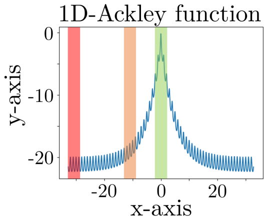

Figure 1: Illustration of three hypothesis locations in the form of confidence regions on the 1D Ackley function. The different colors (red, orange, and green) correspond to different levels of confidence (poor, weak, and good, respectively).

search to more fruitful regions. We specifically represent domain knowledge as human hypotheses or conjectures that are realized as intervals of confidence, i.e., constraints on the parameter space. Figure 1 demonstrates three representative hypothesis regions within the input space of a one-dimensional Ackley function where the input region around zero is clearly the most promising hypothesis. The various hypotheses are realized as GPs restricted to the constrained space, and their utility is being iteratively evaluated, which in turn expands or shrinks the global search space accordingly. Our approach treats the human hypotheses as soft constraints, avoiding to over-restrict the search or getting stuck in local optima. We formulate our approach in a bilevel optimization framework where the lower level evaluates the various hypotheses and the upper level integrates the useful ones in the search. The methodology is further detailed in Section 3.

We test the proposed methodology on a materials design simulation where a set of different chemical hypotheses are injected to guide the search to more fruitful solutions faster. We show that hypotheses with favorable conditions accelerate the search and also improve performance. Interestingly, unfavorable hypotheses do not appear to bias the search negatively in the long run. Extensive synthetic tests further demonstrate the competitive and robust performance of our method.

The remainder of this paper is organized as follows. Section 2 presents recent works about expert knowledge integration in BO, while Section 3 describes the proposed methodology. The robustness and performance of our algorithm are evaluated and discussed in Section 4. Finally, Section 5 summarizes our work and introduces future directions.

## 2 Related Works

Knowledge distillation has recently been at the center of attention in the BO literature to address issues such as the “cold” start problem where the initial points, usually selected randomly, fail to adequately capture the optimization objective’s landscape. Transfer learning [Niu *et al.*, 2020] has been widely used to extract and use knowledge from previous BO executions to aid in warming up and enhancing optimization [Theckel Joy *et al.*, 2019]. Furthermore, it has also been used effectively in chemical reaction optimization [Hickman *et al.*, 2023] to bias the search space by weighting the current acquisition function with past predictions.

Another approach to improving BO through the incorporation of domain knowledge involves the use of similarities between points in the search space. Gryffin [Häse *et al.*, 2021] uses user-provided physicochemical descriptors to navigate the search space more efficiently by identifying similarities between individual options based on those descriptors. However, when using a large number of descriptors, spurious correlations can occur between descriptors and the optimized objective, leading to irrelevant descriptors being considered important. [Morishita and Kaneko, 2023] suggest using a clustering-based initial sample selection method for optimizing chemical reaction conditions with BO based on a high correlation between molecular descriptors and clustering in chemical space. However, as clustering is based on unsupervised learning, there is a need for expert knowledge to connect it with experimental results; also, not all scientific problems can be codified using molecular descriptors.

Other approaches inject expert prior beliefs as priors to guide the optimization process. [Li *et al.*, 2020] combined prior user beliefs with observed data to compute the posterior distribution via repeated Thompson sampling. This approximates new sampling points using a linear combination of posterior samplings. BOPrO [Souza *et al.*, 2021] uses a prior that is provided by the user and a data-driven model to generate a pseudo-posterior. Similarly, $\pi$BO [Hvarfner *et al.*, 2022] generates a pseudo-posterior by integrating prior beliefs into the acquisition function as a decaying multiplicative factor to improve sampling. Both of these methods, and ColaBO [Hvarfner *et al.*, 2024] which augments the surrogate model with a user-defined prior, are limited to one expert prior, and the use of priors cannot capture intricate knowledge. [Ziatdinov *et al.*, 2022] co-navigates a hypothesis space and the experimental space through a hypothesis learning approach that combines multiple hypotheses as probabilistic models with reinforcement learning. However, major drawbacks of this approach are the difficulty of representing a hypothesis in a probabilistic model from the functional form of the black-box model, the overall computational complexity of the Bayesian inference, and the assumption that only one out of the hypothesis pool is the correct one.

Preference learning can also enrich BO through domain knowledge. [Huang *et al.*, 2022] obtain expert opinions by querying them with pairwise comparisons, thereby approximating the shape of the objective function. [Anjanapura Venkatesh *et al.*, 2022] take a slightly different approach by allowing experts to provide a pair of good and bad points, which are then used to fine-tune the BO’s surrogate model by replacing its current optimal hyperparameters with ones that better align with the expert’s cognitive model. Such approaches are promising but risk biasing the optimizer to mimic the user’s beliefs and result in suboptimal solutions.

Various methods can restrict the search space to regions assumed by the optimizer to contain the optimum [Nguyen *et al.*, 2024]. TuRBO [Eriksson *et al.*, 2019] uses multiple independent GP surrogate models within different trust regions to conduct simultaneous BO runs, and a multi-armed bandit (MAB) strategy [Vermorel and Mohri, 2005] to choose which local optimizations to continue. TREGO [Diouane *et al.*, 2021] proposed alternating between global BO and a trust region-based policy for the local phase when the global BO is failing. Alternatively, LA-MCTS [Wang *et al.*, 2020] proposes the use of Monte Carlo tree search to learn which subregions of the search space are more likely to contain good objective values. The space is then recursively partitioned based on optimization performance. Similarly, ZoMBI [Siemenn *et al.*, 2023] iteratively keeps the best sample points found so far and "zooms" in the sampling search bounds towards the region formed by those samples. While our approach shares similarities with the above in terms of segmenting the search space, ours sets itself apart by integrating human-friendly hypothesis-based constraints that avoid over-restricting the search. This differentiation emphasizes the innovative use of expert knowledge in our approach, particularly in complex scenarios where such insights are crucial. More importantly, we propose a generalized strategy that allows the injections of multiple expert hypotheses, such as those derived from a multi-person research team, where promising seeds from those hypotheses augment BO's surrogate model data to achieve better-informed sampling.

## 3 Methodology

In this section, we propose a novel BO methodology, termed as Hypothesis Bayesian Optimization (HypBO), that uses experts' background knowledge in the form of optimality hypotheses to guide search space exploration more effectively. Let $\{\mathcal{H}\}_{j=1}^{J}$ be a set of manually designed hypotheses w.r.t. promising areas (subspaces) formulated in hyperrectangles and explicitly specified by experts through a system of $p$ equations and $q$ inequalities describing an interval of confidence in the search space:

$$\begin{aligned} Ax &= b, \\ Bx &\leq c, \end{aligned} \tag{2}$$

where $A \in \mathbb{R}^{p\times d}$ and $B \in \mathbb{R}^{q\times d}$ are coefficient matrices, and $b \in \mathbb{R}^p$ and $c \in \mathbb{R}^q$ are solution vectors. This system filters the space $\mathcal{X}$ and forms a solution set, which we refer to as hypothesis subspace $\mathcal{H}_j$. In scientific experiments, experts are accustomed to thinking about parameters and conditions in terms of ranges and relationships; formulating these as mathematical constraints is more intuitive than manually setting up a prior distribution. Details on how to create hypotheses can be found in the SM.

Our goal is to inject practitioners' expertise into the problem at hand by attending to specific regions of the search space based on domain hypotheses, minding at the same time not to over-restrict and negatively bias the search. We, therefore, model the various different hypotheses as *local* GPs acting on a constrained parameter space and using their output samples as *seeds* for the *global* search which in turn is realized by a *global* GP. The utility of the seeds can be measured using any standard acquisition function, and the top-performing seeds are selected and fed through to the global search. This iterative global-versus-local Bayesian search takes effect interchangeably and is organized in a parametric bilevel optimization framework, which in this instance can be solved sequentially as a two-stage decision problem with each level's variables treated as a parameter for the other [Köppe *et al.*, 2010]. The following paragraphs detail and formalize the proposed optimization framework.

### 3.1 Upper Level

In this level, we seek to find the global maximum of $f(\cdot)$ in (1) as in any standard BO task where an acquisition function $\alpha(\cdot)$ is maximized to obtain new candidate samples and evaluate them across its iterations

$$x^* = \underset{x\in\mathcal{X}}{argmax}\ \alpha(x, \mathcal{D}) \tag{3}$$

with $\mathcal{D} = \{x_i, y_i = f(x_i)\}_{i=1}^{n}$ being the observation dataset. To compute $\alpha(\cdot)$, BO relies on constructing a global surrogate model of the underlying function and greatly depends on the initial samples provided as a seed when building this. An appropriate initial sampling has been shown to significantly improve the performance of the search in practice [Morishita and Kaneko, 2023]. At this level, one could use practically any variant of BO, but we have empirically observed that using the LA-MCTS algorithm [Wang *et al.*, 2020] helps the search to focus on promising regions to avoid over-exploring.

### 3.2 Lower Level

The lower level initially uses the subspaces from the given hypotheses to perform a *local* search and yield a set of best-performing seed samples $\{s\}_{t=1}^{T}$, with $T \leq J$, essentially acting as soft constraints on the target objective function $f(\cdot)$. Given that we do not have any analytical information about $f(\cdot)$, we approximate it in the hypothesis subspaces by multiple local GP models $\phi_j \sim \mathcal{N}(\mu_j(x), k_j(x, x'))$ simultaneously, one for each hypothesis subspace $\mathcal{H}_j$. The local models $\phi_j$ are chosen as GP surrogate models for their robustness to noise and uncertainty [Rasmussen and Williams, 2006]. The local search is realized in a MAB fashion where getting a seed $s_t$ translates into selecting the most promising hypothesis via an implicit policy where the hypotheses are the arms before doing a local BO in that hypothesis region. This allows us to evaluate the hypotheses and steer the sampling toward promising regions.

In the initialization phase, before the optimization loop starts, we ensure the hypothesis regions are covered by a specific strategy. For each hypothesis subspace $\mathcal{H}_j$, one random sample is drawn, which provides more informative seeds for the global search. If the number of desired initial samples ($n$) is not exhausted after allocating one sample per hypothesis, additional random points are drawn from the entire search space to enhance diversity and exploration. In the event that the number of hypotheses exceeds $n$, we ensure that at least one additional random point is taken from the whole search space, making $m = \max(1, n - J)$. This strategy guarantees that each hypothesis is represented from the dataset and prevents scenarios where no initial points fall within the hypothesis regions. As the optimization progresses, the hypothesis subspaces $\mathcal{H}_j$ will potentially have more samples, which will update the local models, better evaluating the hypotheses and producing better seeds. The complete bilevel framework is formalized as follows

$$x^* = \underset{x\in\mathcal{X}}{argmax}\ \alpha(x, \mathcal{D} \cup \{(s_t, f(s_t))\}_{t=1}^{T}) \tag{Upper}$$

*s.t*

$$\{s_t\}_{t=1}^{T} \in \underset{s\in\bigcup_{j=1}^{J}\mathcal{H}_j}{argmax}\{\underset{s\in\mathcal{H}_j}{max}\ \alpha_{\phi_j}(s, \{\mathcal{D}\cup\{x^*, f(x^*)\}\}\cap\mathcal{H}_j)\}_{j=1}^{J}. \tag{Lower}$$

Stopping criteria for each level and global convergence are discussed below in Section 3.3.

### 3.3 Convergence Criteria

To maximize the information gained from good (true) hypotheses, we allow the lower level to produce more seed samples until it plateaus. That is, the lower level returns seed samples until these fail to improve upon the best target value:

$$y_{max} + \gamma \geq f(s_i) \text{ for } i = k+1, ..., k + l_{max}, \qquad (4)$$

where $y_{max}$ is the best value found by iteration $i$, $i$ is the current iteration number, $k$ is the iteration number from which the plateauing started, $\gamma \in \mathbb{R}^+$ is the growth step size, and $l_{max} \in \mathbb{N}^+$ dictates after how many consecutive iterations we deem the lower level plateauing.

To mitigate weak and poor (false) hypotheses, we allow the upper level to carry the optimization from the given seeds until it plateaus. It keeps maximizing $\alpha$ until it fails to improve upon the best target value, that is:

$$y_{max} + \gamma \geq f(x_i) \text{ for } i = k+1, ..., k + u_{max}, \qquad (5)$$

where $u_{max} \in \mathbb{N}^+$ dictates after how many consecutive iterations we deem the upper level failed. We set $l_{max} \ll u_{max}$ to direct the search toward the hypotheses' regions if they are helping to improve while still giving the upper level the time to explore the entire search space $\mathcal{X}$.

Parameter $\gamma$ sets how much improvement is considered "significant" for the optimization process. A larger $\gamma$ means that the algorithm requires a larger improvement to consider the level optimization as still progressing. Conversely, a smaller $\gamma$ makes the criterion for progress more strict, as even minor improvements will be considered significant. The optimization steps are detailed in Algorithm 1.

## 4 Experiments

We showcase the effectiveness of our proposed method in optimizing various synthetic functions and real-world problems, such as discovering new materials. We test HypBO's performance and robustness using hypotheses of different qualities, ranging from good to poor. We also compare its performance against other BO algorithms. In Section 4.1, we outline the various experimental settings and comparison methods we used to benchmark our results. Sections 4.2 and 4.3 present the outcomes of an analytical function optimization task and a materials design problem [Burger *et al.*, 2020], respectively.

### 4.1 Experimental Setup

We evaluate HypBO's performance empirically for the following two tasks:

- **Synthetic Functions** We test the precision, convergence speed, and robustness of HypBO using synthetic benchmark functions with various nonconvex landscapes and dimensionalities. The optimization performance is measured using simple and cumulative regrets, and Wilcoxon tests [Rey and Neuhäuser, 2011]. The maximum number of iterations is limited to 100, and the result of 50 repeated trials is reported as the mean value.

**Algorithm 1** Hypothesis Bayesian Optimization (HypBO)

```
Input: Hypotheses $\{\mathcal{H}_j\}_{j=1}^J$, Number of initial samples $n$,
Maximum iteration number $i_{max}$, Improvement growth size
$\gamma$, Number of locally optimal samples $T$ to keep,
convergence parameters $l_{max}$ and $u_{max}$ Output: $y_{max}$
Initialize the dataset $\mathcal{D} = \{\}$;
for each hypothesis subspace $\mathcal{H}_j$ do
    Sample $x_j$ randomly from $\mathcal{H}_j$;
    Evaluate $y_j = f(x_j)$ and add $(x_j, y_j)$ to $\mathcal{D}$;
end for
Randomly sample $m = \max(1, n - J)$ points from the
    entire search space $\mathcal{X}$, evaluate and add them to $\mathcal{D}$;
Set $y_{max}$ as the maximum $y$ value in $\mathcal{D}$ and $i = 0$;
while $i < i_{max}$ do
  Lower Level
    Set attempt without improvement count $l = 0$;
    while $l < l_{max}$ and $i < i_{max}$ do
        for each expert-defined hypothesis $\mathcal{H}_j$ do
            Fit a GP $\phi_j$ within the hypothesis $\mathcal{D} \cap \mathcal{H}_j$;
            Find the best sample $s_j$ maximizing $\alpha_{\phi_j}$;
        end for
        Keep the best samples $\{s_t\}_{t=1}^T$ w.r.t. $\alpha_{\phi_j}$;
        Evaluate the samples $\{s_t\}_{t=1}^T$ and set $y_{t_{max}}$ as
            the maximum of all $y_t = f(s_t)$;
        Increment $l$ if there is no improvement, i.e.
            $y_{t_{max}} \leq y_{max} + \gamma$, else reset $l$ to 0;
        Update the records $\mathcal{D} \leftarrow \mathcal{D} \cup \{(s_t, y_t)\}_{t=1}^T$;
        Update $y_{max}$ as the maximum $y$ value in $\mathcal{D}$;
        $i \leftarrow i + 1$;
    end while
  Upper Level
    Set attempt without improvement count $u = 0$;
    while $u < u_{max}$ and $i < i_{max}$ do
        Fit a GP on the entire search space $\mathcal{D}$;
        Find the best sample $x^*$ maximizing $\alpha$;
        Evaluate $x^*$, $y^* = f(x^*)$;
        Increment $u$ if there is no improvement, i.e.
            $y^* \leq y_{max} + \gamma$, else reset $u$ to 0;
        Update the records $\mathcal{D} \leftarrow \mathcal{D} \cup \{(x^*, y^*)\}$;
        $i \leftarrow i + 1$;
    end while
end while
return The maximum value found, $y_{max}$
```

- **Photocatalytic Hydrogen Production** Here, we replicate the materials design problem addressed in [Burger *et al.*, 2020], aiming to maximize the hydrogen evolution rate (HER) from a mixture of different materials. We follow a more cost-effective approach and emulate that chemistry experiment by interpolating new HER measurements using a GP model trained on existing experimental data points. In section 4.3, we give a comprehensive explanation of this chemistry task. The maximum number of iterations is set to 300, and the mean value of 50 repeated trials is reported.

We further evaluate HypBO against the following baselines whose hyperparameters' values are given in the SM:

- **Random Search (RS)** Random search under uniform distribution over the search space.

- **Trust Region Bayesian Optimization (TuRBO)** with one trust region.
- **Latent Action Monte Carlo Tree Search (LA-MCTS)**
- **LA-MCTS with hypothesis-based initial design (LA-MCTS+)** We modified the previous baseline to initialize it exclusively within the hypothesis subspaces before the regular search in the entire search space.
- **$\pi$BO** This baseline uses expert knowledge throughout optimization. As described in Section 2, it is one of the most competitive methods for priors over optimum. We converted our hypotheses into Gaussian priors centered around the hypothesis subspace center (see SM).

For all experiments, we use preset hyperparameters for HypBO. We set the lower level limit $l_{max}$ to 2, the upper level limit $u_{max}$ to 5, the number of locally optimal samples $T$ to 1, and the growth rate $\gamma$ to 0. This value of $\gamma$ essentially means that as long as any improvement is being made (no matter how small), the level optimization will continue. Note that this could be beneficial in scenarios where even small gains are valuable, but it might also make the optimization process slower or more prone to getting stuck in flat regions where minute fluctuations might appear as improvements. Ablation studies can be found in the SM. Concerning the local GP models for the hypotheses, they have a zero mean and a Matérn ($\nu = 2.5$, $\lambda_i = 1$) kernel with constant scaling. Note that the kernel's hyperparameters are automatically optimized based on the experimental data to fit the models best. All experiments are warm-started with five initial points except for the photocatalytic hydrogen production experiment with mixed hypotheses, whose initial sample count is 10. Reproducibility details are available in the SM and the source code can be found at https://github.com/Ablatif6c/HypBO.

## 4.2 Synthetic Functions

### Hypotheses

A good hypothesis subspace is essentially an interval that contains the optimum, *opt*. By contrast, a weak/poor one does not. The further the weak hypothesis subspace is from the optimum, the worse it is. Here, the "poor" hypothesis is the furthest from the optimum. The hypotheses are hyperrectangles of width $w = 2$ units and centered as follows:

- **Poor hypothesis** at $l_b + w/2$ where $l_b$ is the lower bound of the search space.
- **Weak hypothesis** at $opt - 0.2 * (opt - l_b) - w/2$.
- **Good hypothesis** at $opt$.

We assess HypBO empirically in two different settings. First, we evaluate its performance and robustness against the quality of the hypothesis. Second, we test its ability, when faced with mixed hypotheses simultaneously, to discard the weaker hypotheses and prioritize promising ones.

### Optimization With a Single Hypothesis

Figures 2 and 3 show that HypBO benefits from informative hypotheses and can also recover from weak ones. The method improves the search performance dramatically over RS, TuRBO, and LA-MCTS for a good hypothesis. The seeds from that hypothesis aid in recognizing the promising subspace and focusing efforts there, resulting in a faster location of the optimum, akin to cost-effective optimizations [Tay *et al.*, 2023]. A similar behavior is observed in both LA-MCTS+ and $\pi$BO, but they notably outperform HypBO on smaller and simpler landscapes like Branin$_{\text{d2}}$ and Sphere$_{\text{d2}}$, as their initial sampling is entirely hypothesis-focused. Concerning the weak hypothesis, HypBO converges toward the optimum faster than LA-MCTS, TuRBO, and RS. In fact, the seeds coming from the weak hypothesis, by outperforming the existing samples in the dataset, direct HypBO towards the hypothesis' surroundings, too, which are more promising. As would be expected, poor hypotheses lead to a slower search in the early stages, but HypBO displays desired robustness by recovering from the poor seeds to approximately equal regret as LA-MCTS and TuRBO. However, it is interesting to note that HypBO with a poor hypothesis outperforms LA-MCTS and TuRBO in high-dimensional functions due to its more diverse initial sampling strategy. Its initial sampling strategy combines one sample from the poor hypothesis with others from across the search space, leading to a more comprehensive understanding of the overall landscape. As shown in Figure 3, this approach efficacy seems to increase with the search space complexity, i.e., its dimensionality, making the advantage of diverse sampling more pronounced. Both LA-MCTS+ and $\pi$BO lag behind HypBO in these two last hypothesis scenarios as their initial sampling, being entirely made of samples from the weak (respectively poor) hypothesis region, is not diverse enough and $\pi$BO's trade-off decay hyperparameter $\beta$ keeps it unnecessarily longer in that weak (respectively poor) region as shown in Figure 3. This highlights HypBO's

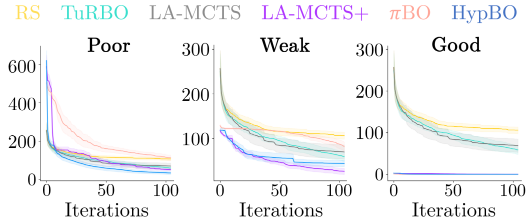

Figure 2: Comparison of RS, TuRBO, LA-MCTS, LA-MCTS+, $\pi$BO, and HypBO on Levy$_{\text{d20}}$ for various hypothesis qualities. Solid lines show mean values, while shaded areas represent standard error.

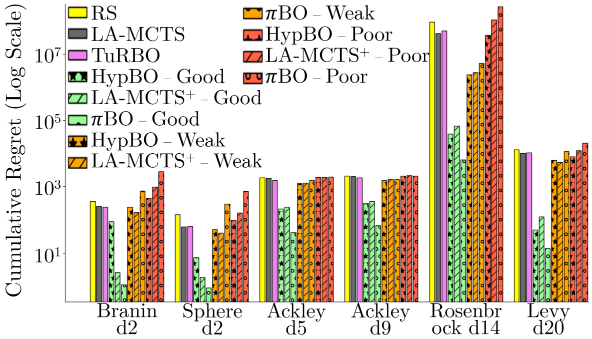

Figure 3: Cumulative regret on functions with various dimensions.

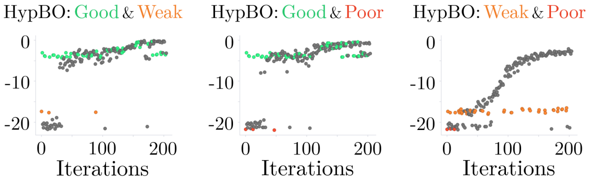

Figure 4: HypBO on the 9D-Ackley function with three different mixtures of hypotheses of various qualities for 200 iterations. Colored sample points came from the hypotheses, i.e., the lower level, while the grey ones came from the upper level.

ability to exploit the explicit and implicit information the hypothesis provides faster and more intelligently. Moreover, we conducted Wilcoxon signed-rank tests [Rey and Neuhäuser, 2011] (at a 95% confidence level with Bonferroni correction [Bonferroni, 1936]) and examined the mean and median cumulative regrets. The tests reveal that HypBO performs significantly better than RS, LA-MCTS, and TuRBO. Although the p-values for comparisons with LA-MCTS+ and $\pi$BO were not statistically significant, they were low for $\pi$BO ($p = 0.06$ and $0.15$ for weak and poor hypotheses), indicating $\pi$BO's weaker performance. Moreover, HypBO showed lower median and mean regrets compared to LA-MCTS+ and $\pi$BO variants for weak and poor hypotheses, leading us to conclude that HypBO generally outperforms both of them.

### Optimization With Mixed Hypotheses

We use three different binary combinations of hypotheses of varying quality to test HypBO's ability to uncover and prioritize promising hypotheses, and to discard bad ones from a pool of hypotheses. HypBO takes seeds from all the given hypotheses, which it uses to update its beliefs about each hypothesis via the MAB procedure. As the optimization progresses, it has a better representation of the hypotheses and can abandon the weaker one of the pair and select seeds from the more promising hypothesis. As shown in Figure 4, for the Ackley$_{d9}$ function, this approach allows HypBO to deselect the weaker hypothesis early on. It keeps the remaining stronger hypothesis, which it uses to expedite the search as described in the previous subsection. Figure 5 shows that these findings are consistent when applied to a variety of synthetic functions of higher dimensions. For lower dimensions, HypBO with mixed hypotheses has approximately equal regret performance to TuRBO, LA-MCTS, and LA-MCTS+ as the search space is smaller, and it becomes easier to capture the underlying behavior of the objective function. For higher dimensions, even in the case of combined weak and poor hypotheses, HypBO outperforms the other methods, demonstrating robustness when faced with multiple items of inaccurate knowledge and the ability to use these for better sampling.

Here, the Wilcoxon tests with Bonferroni correction show no significant difference between HypBO and LA-MCTS+, albeit much lower p-values for the Good & Poor and Good & Weak scenarios ($p = 0.15$). Along with an examination of the mean and median regrets where HypBO has lower values, we conclude that HypBO outperforms LA-MCTS+. These statistical tests also show that HypBO greatly outperformed LA-MCTS, TuRBO, and RS ($p \leq \alpha_{adjusted}$) with much lower mean and median regrets.

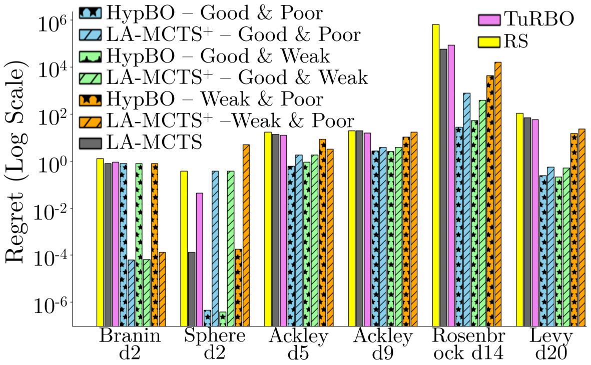

Figure 5: Regret on synthetic functions with mixed hypotheses.

## 4.3 Photocatalytic Hydrogen Production Optimization

We test HypBO on a real materials design problem where we seek an optimal composition of ten materials to maximize hydrogen production via photocatalysis [Wang *et al.*, 2019]. Due to the combinatorially large search space (98,423,325 possible combinations), [Burger *et al.*, 2020] used an autonomous mobile robotic chemist along with a discretized Bayesian optimizer (DBO), which can discretize the input space, to search for the optimal combination of materials.

We recast this experimental problem as a more cost-effective multivariable simulation; that is, we mapped out the chemical space by interpolating available experimental observations using a Gaussian process regression (GPR). Specifically, this GPR model has a zero mean, a Matérn ($\nu = 2.5$) kernel with constant scaling and homoscedastic noise; each variable lengthscale $\lambda_i$ is initialized as its discretization step. We fitted this model against a total "ground truth" dataset of 1119 experimental observations supplied by the authors of [Burger *et al.*, 2020]. While the interpolated model is only approximate, close inspection suggested that it is broadly representative of the known real chemical space and sufficiently accurate to draw safe conclusions here. Our main goal is to test whether we can capture and inject experts' knowledge and intuition towards a better-informed and faster search. For a fair comparison, in place of TuRBO, LA-MCTS, and LA-MCTS+, we use the same DBO developed by [Burger *et al.*, 2020] for experimental photocatalytic hydrogen production, capable of discretization, as a baseline.

### Retrospective Application of Knowledge

First, we used HypBO to fold in, retrospectively, knowledge of the underlying chemistry that was not captured in the [Burger *et al.*, 2020] study to investigate whether injecting such hypotheses might improve performance. We explored three separate cases, outlined briefly below with more detailed explanations in the SM:

- **What They Knew** In 2019, there was extra chemical knowledge available prior to the [Burger *et al.*, 2020] study that could not have been injected using DBO; here, it is injected, retrospectively, using HypBO.

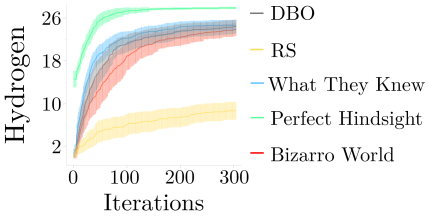

Figure 6: Retrospective application of hypotheses derived from [Burger *et al.*, 2020] using HypBO, compared to the no hypothesis run using DBO and RS. Shaded area is the standard deviation.

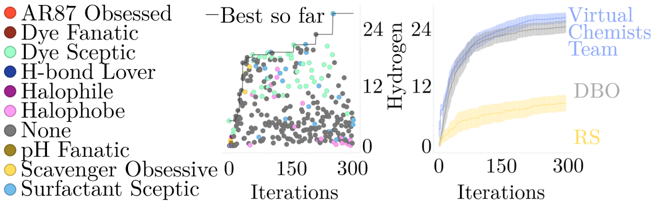

Figure 7: HypBO pruning the hypotheses and selecting seeds from the most promising ones to boost the optimization. Left: Scatter plot of the HypBO optimization with all nine virtual hypotheses; color denotes followed hypothesis, if any. Right: Best value obtained so far by HypBO using all nine hypotheses compared to DBO and RS.

- **Perfect Hindsight** limits the search to within the optimal subspace based on *post facto* knowledge of the outcomes of all 1119 robotic experiments.
- **Bizarro World** purposefully focuses the search within the worst areas of the chemical space in all dimensions.

As shown in Figure 6, HypBO with ‘What They Knew’ boosts performance somewhat in the early stages of the search, and overall it improves upon DBO, thus validating the benefits of considering expert hypotheses in real-world problems. For example, one can posit that any or all of the three dye components (MB, AR87, RB) might be beneficial but that high values would be counterproductive, based on chemical reasoning. We captured this in ‘What They Knew’ by lowering the dyes’ upper bounds (MB $\leq$ 0.5mL, AR87 $\leq$ 1mL, RB $\leq$ 0.5mL). The somewhat modest boost given by ‘What They Knew’ (Figure 6) can be explained by the partial knowledge available in 2019; indeed, some of ‘What They Knew’ was, in fact, wrong. For example, as reported in [Burger *et al.*, 2020], all three dyes were strongly negative at all concentrations. We have not captured this post-experiment knowledge here; rather, ‘What They Knew’ captures the knowledge that was available to this team in 2019, building on their initial formulation of hypotheses, prior to any robotic experiments.

Unsurprisingly, ‘Perfect Hindsight’ leads to a much faster optimization. By contrast, although the artificially bad case of ‘Bizarro World’ does lead to a slower search than DBO, the effects are greatly mitigated because HypBO can abandon unproductive hypotheses.

### Searching the Chemistry Space with Mixed Hypotheses

We test HypBO’s ability to exploit good hypotheses and discard bad ones in a more realistic setting by creating a team of nine ‘virtual chemists’, each with a virtual hypothesis based on plausible chemical reasoning detailed in the SM. The combined “knowledge” of this virtual team was then used to redo the simulated experiment for [Burger *et al.*, 2020] in tandem with HypBO. The virtual team was designed to emulate the diverse and sometimes contradictory views of a real research team tackling a new problem. For example, some pairs of hypotheses (e.g., ‘Halophile’ / ‘Halophobe’) are in direct contradiction. Based on retrospective knowledge, ‘Dye Sceptic’ and ‘Surfactant Sceptic’ might be expected empirically to be the strongest hypotheses, while ‘Dye Fanatic’ is probably the weakest. We applied all nine virtual hypotheses simultaneously and used the same oracle model described in the section above. For the purposes of these initial tests, all virtual hypotheses were considered of equal weighting. Figure 7 illustrates the power of including human insights throughout the optimization. All hypotheses were selected initially, but as the optimization progressed, HypBO filtered out bad hypotheses and prioritized the most promising ones, improving the search compared to DBO and RS. While the least profitable hypotheses, such as ‘AR87 Obsessed’, ‘Dye Fanatic’, and ‘Halophobe’, were deselected early on, HypBO does not discard them completely. For example, ‘Halophobe’ was selected at times when its Expected Improvement was greater than that of others that were available for evaluation. This captures the importance of re-evaluating hypotheses in the face of new data. Likewise, certain hypotheses are used while they are profitable and then discarded when they become delimiting; this can be observed for ‘Scavenger Obssessive’, where some scavenger is indeed required, but not too much. The modest search improvement of the virtual chemist team over DBO (Figure 7, right) is somewhat arbitrary because we purposefully built this virtual team to be mediocre, with both “good” (informative) and “bad” (uformative or misleading) hypotheses in near equal numbers.

## 5 Conclusions

To fully exploit the opportunities in laboratory robotics and automation, we need optimization methods that work in tandem with teams of human scientists. So far, BO has not fully leveraged the experience and hunches of experimenters. We harness that knowledge here by allowing them to inject their hypotheses about which parts of the input space will yield the best performance. We propose a BO variant, HypBO, that achieves this in a bi-level framework by recursively pruning and turning the hypotheses into seeds that augment sampling as a springboard for global optimization. HypBO stands out in its ability to concurrently handle and evaluate multiple expert-formulated hypotheses. It can also use weak hypotheses to converge faster than cases with no hypotheses and recover from poor ones. This highlights the power of human-computer interaction, re-imagining the role of humans in autonomous scientific discovery. Future work includes initializing the hypotheses with weights based on the experimenter’s profile or confidence estimation. These weights might be particularly valuable in quick-starting hypothesis selection in large, diverse research teams, where expertise levels and domain specializations can vary quite widely.

## Acknowledgments

The authors acknowledge financial support from the Leverhulme Trust via the Leverhulme Research Centre for Functional Materials Design. AIC thanks the Royal Society for a Research Professorship (RSRP\S2\232003).

## References

[Anjanapura Venkatesh *et al.*, 2022] Arun Kumar Anjanapura Venkatesh, Santu Rana, Alistair Shilton, and Svetha Venkatesh. Human-ai collaborative bayesian optimisation. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 16233–16245. Curran Associates, Inc., 2022.

[Bonferroni, 1936] Carlo E. Bonferroni. *Teoria statistica delle classi e calcolo delle probabilità*. Pubblicazioni del R. Istituto superiore di scienze economiche e commerciali di Firenze. Seeber, 1936.

[Burger *et al.*, 2020] Benjamin Burger, Phillip M. Maffettone, Vladimir V. Gusev, Catherine M. Aitchison, Yang Bai, Xiao yan Wang, Xiaobo Li, Ben M. Alston, Buyin Li, Rob Clowes, Nicola Rankin, Brandon Harris, Reiner Sebastian Sprick, and Andrew I. Cooper. A mobile robotic chemist. *Nature*, 583:237–241, 2020.

[Diouane *et al.*, 2021] Youssef Diouane, Victor Picheny, Rodolophe Le Riche, and Alexandre Scotto Di Perrotolo. Trego: a trust-region framework for efficient global optimization. *Journal of Global Optimization*, 86:1–23, 2021.

[Eriksson *et al.*, 2019] David Eriksson, Michael Pearce, Jacob Gardner, Ryan D Turner, and Matthias Poloczek. Scalable global optimization via local bayesian optimization. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

[Hernández-Lobato *et al.*, 2015] José Miguel Hernández-Lobato, Michael Gelbart, Matthew Hoffman, Ryan Adams, and Zoubin Ghahramani. Predictive entropy search for bayesian optimization with unknown constraints. In *International Conference on Machine Learning*, pages 1699–1707. PMLR, 2015.

[Hickman *et al.*, 2023] Riley J. Hickman, Jurgis Ruža, Hermann Tribukait, Loïc M. Roch, and Alberto García-Durán. Equipping data-driven experiment planning for self-driving laboratories with semantic memory: case studies of transfer learning in chemical reaction optimization. *Reaction Chemistry & Engineering*, 2023.

[Huang *et al.*, 2022] Daolang Huang, Louis Filstroff, Petrus Mikkola, Runkai Zheng, and Samuel Kaski. Bayesian optimization augmented with actively elicited expert knowledge. arXiv:2208.08742, 2022.

[Hvarfner *et al.*, 2022] Carl Hvarfner, Danny Stoll, Artur Souza, Luigi Nardi, Marius Lindauer, and Frank Hutter. $\pi$BO: Augmenting acquisition functions with user beliefs for bayesian optimization. In *10th International Conference on Learning Representations, ICLR'22*, pages 1–30, April 2022.

[Hvarfner *et al.*, 2024] Carl Hvarfner, Frank Hutter, and Luigi Nardi. A general framework for user-guided bayesian optimization. In *The Twelfth International Conference on Learning Representations*, 2024.

[Häse *et al.*, 2021] Florian Häse, Matteo Aldeghi, Riley J. Hickman, Loïc M. Roch, and Alán Aspuru-Guzik. Gryffin: An algorithm for Bayesian optimization of categorical variables informed by expert knowledge. *Applied Physics Reviews*, 8(3):031406, 07 2021.

[Jones *et al.*, 1998] Donald R. Jones, Matthias Schonlau, and William J. Welch. Efficient global optimization of expensive black-box functions. *Journal of Global Optimization*, 13(4):455–492, 1998.

[Köppe *et al.*, 2010] Matthias Köppe, Maurice Queyranne, and Christopher Thomas Ryan. Parametric integer programming algorithm for bilevel mixed integer programs. *Journal of Optimization Theory and Applications*, 146(1):137–150, 2010.

[Li *et al.*, 2020] Cheng Li, Sunil Gupta, Santu Rana, Vu Nguyen, Antonio Robles-Kelly, and Svetha Venkatesh. Incorporating expert prior knowledge into experimental design via posterior sampling. arxiv:2002.11256, 2020.

[Martinelli *et al.*, 2023] Julien Martinelli, Yasmine Nahal, Duong Lê, Ola Engkvist, and Samuel Kaski. Leveraging expert feedback to align proxy and ground truth rewards in goal-oriented molecular generation. In *NeurIPS 2023 Workshop on New Frontiers of AI for Drug Discovery and Development*, 2023.

[Morishita and Kaneko, 2023] Toshiharu Morishita and Hiromasa Kaneko. Initial sample selection in bayesian optimization for combinatorial optimization of chemical compounds. *ACS Omega*, 8(2):2001–2009, 2023.

[Nguyen *et al.*, 2024] Quoc Phong Nguyen, Wan Theng Ruth Chew, Le Song, Bryan Kian Hsiang Low, and Patrick Jaillet. Optimistic bayesian optimization with unknown constraints. In *The Twelfth International Conference on Learning Representations*, 2024.

[Niu *et al.*, 2020] Shuteng Niu, Yongxin Liu, Jian Wang, and Houbing Song. A decade survey of transfer learning (2010–2020). *IEEE Transactions on Artificial Intelligence*, 1(2):151–166, 2020.

[Ramachandran *et al.*, 2020] Anil Ramachandran, Sunil Gupta, Santu Rana, Cheng Li, and Svetha Venkatesh. Incorporating expert prior in bayesian optimisation via space warping. *Knowledge-Based Systems*, 195:105663, 2020.

[Rasmussen and Williams, 2006] Carl Edward Rasmussen and Christopher K. I. Williams. *Gaussian processes for machine learning*. Adaptive computation and machine learning. MIT Press, 2006.

[Rey and Neuhäuser, 2011] Denise Rey and Markus Neuhäuser. *Wilcoxon-signed-rank test*, pages 1658–1659. Springer Berlin Heidelberg, Berlin, Heidelberg, 2011.

[Shahriari *et al.*, 2016] Bobak Shahriari, Kevin Swersky, Ziyu Wang, Ryan P. Adams, and Nando de Freitas. Taking the human out of the loop: A review of bayesian optimization. *Proceedings of the IEEE*, 104(1), 2016.

[Siemenn *et al.*, 2023] Alexander E. Siemenn, Zekun Ren, Qianxiao Li, and Tonio Buonassisi. Fast bayesian optimization of needle-in-a-haystack problems using zooming memory-based initialization (zombi). *npj Computational Materials*, 9(1), May 2023.

[Souza *et al.*, 2021] Artur Souza, Luigi Nardi, Leonardo B. Oliveira, Kunle Olukotun, Marius Lindauer, and Frank Hutter. Bayesian optimization with a prior for the optimum. In Nuria Oliver, Fernando Pérez-Cruz, Stefan Kramer, Jesse Read, and Jose A. Lozano, editors, *Machine Learning and Knowledge Discovery in Databases. Research Track*, pages 265–296, Cham, 2021. Springer International Publishing.

[Sundin *et al.*, 2018] Iiris Sundin, Tomi Peltola, Luana Micallef, Homayun Afrabandpey, Marta Soare, Muntasir Mamun Majumder, Pedram Daee, Chen He, Baris Serim, Aki Havulinna, Caroline Heckman, Giulio Jacucci, Pekka Marttinen, and Samuel Kaski. Improving genomics-based predictions for precision medicine through active elicitation of expert knowledge. *Bioinformatics (Oxford, England)*, 34:i395–i403, 07 2018.

[Tay *et al.*, 2023] Sebastian Tay, Chuan Sheng Foo, Daisuke Urano, Richalynn Leong, and Bryan Kian Hsiang Low. Bayesian optimization with cost-varying variable subsets. In A. Oh, T. Neumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 3008–3031. Curran Associates, Inc., 2023.

[Theckel Joy *et al.*, 2019] Tinu Theckel Joy, Santu Rana, Sunil Gupta, and Svetha Venkatesh. A flexible transfer learning framework for bayesian optimization with convergence guarantee. *Expert Systems with Applications*, 115:656–672, 2019.

[Vermorel and Mohri, 2005] Joannès Vermorel and Mehryar Mohri. Multi-armed bandit algorithms and empirical evaluation. In *Proceedings of the 16th European Conference on Machine Learning*, ECML'05, page 437–448, Berlin, Heidelberg, 2005. Springer-Verlag.

[Wang *et al.*, 2019] Zheng Wang, Can Li, and Kazunari Domen. Recent developments in heterogeneous photocatalysts for solar-driven overall water splitting. *Chemical Society Reviews*, 48:2109–2125, 2019.

[Wang *et al.*, 2020] Linnan Wang, Rodrigo Fonseca, and Yuandong Tian. Learning search space partition for black-box optimization using monte carlo tree search. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 19511–19522. Curran Associates, Inc., 2020.

[Ziatdinov *et al.*, 2022] Maxim A. Ziatdinov, Yongtao Liu, Anna N. Morozovska, Eugene A. Eliseev, Xiaohang Zhang, Ichiro Takeuchi, and Sergei V. Kalinin. Hypothesis learning in automated experiment: application to combinatorial materials libraries. *Advanced Materials*, 34(20), April 2022.