Digital Discovery

ROYAL SOCIETY OF CHEMISTRY

PAPER

View Article Online
View Journal | View Issue

Cite this: *Digital Discovery*, 2026, **5**, 1132

Can we automate scientific reasoning in closed-loop experiments using large language models?

Abdoulatif Cissé, Max E. Cooper, Mengjia Zhu, Xenophon Evangelopoulos and Andrew I. Cooper *

We present here a detailed study of our hybrid optimisation framework, BORA, which integrates large language model (LLM) reasoning with Bayesian optimisation (BO) for accelerating scientific discovery using closed-loop experiments. We compare five reasoning models (o4-mini, o3, gpt-5-mini, gpt-5, and gemini-2.5-flash) as optimisers for two benchmark problems: a 10-dimensional photocatalytic hydrogen-evolution experiment and a 7-dimensional physics-based pétanque simulation. The results show that LLM/BO hybrids outperform BO-only approaches, particularly in early-stage exploration where the search is warm-started by LLM-driven hypotheses. Among the models tested, o3 delivered the strongest and most consistent optimisation performance after 150 experiments. LLM-only optimisations without the BO component also matched or surpassed hybrid methods in some settings, locating optima with high repeatability. We demonstrate that appending human hypotheses, prior literature, or experimental datasets can improve convergence, and that LLM reasoning can recover in some cases from deliberately misleading prompts. We also explore outlier runs to understand the limitations and failure modes of these methods, as well as considering the energy implications of the LLM queries. The strongest LLM-only performance was observed with a batch size of one, suggesting that experiment-by-experiment machine reasoning is a viable strategy for certain automated scientific optimisation tasks.

Received 23rd November 2025
Accepted 9th February 2026

DOI: 10.1039/d5dd00520e

rsc.li/digitaldiscovery

1. Introduction

There has been much recent interest in self-driving laboratories—that is, laboratories where certain decisions can be made autonomously in closed-loop experiments.[1–31] There is no universal definition of autonomy, but Jensen and colleagues[19] suggested the following: "To avoid ambiguity, we define automation herein as the act of making a process occur without human intervention and autonomy as a paradigm where feedback and adaptive decision making afford the system agency over the manner of its actions". In many ways, autonomy is a natural consequence of succeeding in laboratory automation. If we build robotic workflows that can operate 24/7 for extended periods,[32] then human decision making can become the rate-limiting step. Hence, there are many research problems where some form of autonomous decision making is a prerequisite for unlocking the full potential of laboratory automation.

A popular approach for autonomous closed-looped experiments is Bayesian optimisation (BO) using Gaussian processes.[1,31–40] It uses a probabilistic surrogate and an acquisition function to quantify uncertainty and to balance exploration with exploitation, allowing adaptive, data-efficient selection of the next experiment and convergence to optimal conditions. Bayesian optimisation can also encompass categorical variables[41] and use surrogate models such as random forests[42] and neural networks. BO[1,33–39,41] and multimodal BO[43] are excellent tools for closed-looped experimentation. However, in our experience, it is challenging to build on-the-fly ontologies for the specific research problem in hand. To give one example, we developed an autonomous mobile robotic chemist that used BO to optimise a 10-dimensional photocatalytic hydrogen production reaction.[32] Three of the ten possible components were candidate dye sensitisers. It was found that these dyes all had a negative influence on the catalytic reaction, and hence dye addition was avoided after a few experimental batches. However, the Bayesian algorithm did not learn a general rule that "dye addition is bad" because it had no ontology to do this. Rather, the algorithm correlated, independently, each of these three input variables with the reaction yield. Such uncategorised methods without any semantic ontology become inefficient as the number of input variables increases—if there were 100 candidate dyes, the algorithm would still consider them individually, one by one. Categorical BO methods exist,[41] but it can be challenging to assign neat, unambiguous categories to variables in an experiment. For example, in a chemical reaction, the role of the solvent might be to dissolve the reagents, to act as a heat transfer medium, to yield a nanocrystalline product, or to provide a source of acid to catalyse the reaction—or indeed all four of those things. Likewise, reagents can themselves act as

Materials Innovation Factory and Department of Chemistry, University of Liverpool, Liverpool, L7 3NY, UK. E-mail: aicooper@liverpool.ac.uk

cosolvents, but they would not usually be assigned to the category of solvent. This is just one example of the wider problem of categorisation in chemistry and materials research. At best, the manual creation of bespoke ontologies and categories for each optimisation problem can be unwieldy and time-consuming. At worst, an overly rigid and constrained ontology might reduce the effectiveness of the optimisation.

Optimisation landscapes in science tasks are often non-convex, which presents a further challenge. To better cope with this, intelligent approaches have been developed to use adaptive exploration–exploitation strategies,[44] or enrich the optimisation landscape with domain knowledge by customising the prior distribution.[45] ZoMBI improved efficiency by focusing on local regions assumed to contain the optimum.[46] TuRBO[47] used multiple independent GP surrogate models within identified trust regions and a multi-armed bandit strategy. Incorporating domain knowledge into BO can improve both its efficiency and its performance.[41,48] Some implementations such as ColaBO[49] and HypBO[50] allow researchers to inject their beliefs at the start to guide the optimisation process. However, those methods keep the users' beliefs static and cannot refine them as the optimisation progresses, even if they are wrong. Meanwhile, other human-in-the-loop methods rely on frequent user inputs.[51,52] For robotic experiments that run 24/7 in a closed-loop way,[32] waiting for this human user input could become the rate-limiting step, particularly for fast, small-batch experiments where frequent user input is required.

Large language models (LLMs) have developed rapidly in the last few years, to the extent that they have been compared with human intelligence.[53,54] There have even been claims of "PhD-level" abilities.[55] While these claims are contentious, and we would argue overstated, reasoning models are interesting candidates for closed-loop experiments. Various studies have explored LLMs either to augment or to replace traditional optimisers.[56–63] Methods like LLAMBO[64] and OPRO[65] use LLMs to propose solutions to optimisation problems directly. Hybrid approaches such as BoChemian[66] combine the strengths of LLMs to featurise more traditional optimisation methods. SLLMBO[67] integrates the strengths of LLMs in warm-starting optimisation, and it loops between LLM-based parameter exploitation and Tree-structured Parzen Estimator (TPE)'s exploration capabilities to achieve a balanced exploration–exploitation trade-off. LILO uses an LLM to convert unstructured feedback in the form of natural language into scalar utilities to conduct BO over a numeric search space.[68]

Beyond benchmarking studies, LLMs have been used to guide laboratory experiments and quantum chemistry. ChemCrow[56] is an AI chemistry agent that augmented LLMs with 18 expert-designed chemistry tools to overcome their limitations in chemistry-related problems. ChemCrow planned and executed chemical syntheses autonomously, as demonstrated by the successful creation of an insect repellent and three organocatalysts using the cloud-connected RoboRXN platform, also facilitating human-AI collaboration. Coscientist[58] is an AI system powered (initially) by GPT-4, designed for autonomous chemical research. It integrated LLMs with tools such as internet search, code execution, and experimental automation to autonomously design, plan, and execute experiments, as demonstrated across tasks that included optimising palladium-catalysed cross-coupling reactions and controlling laboratory hardware. Recently, an active learning workflow guided by a LLM was used to select equimolar quinary-cation perovskites from over 850 000 possible cation combinations.[69] In the realm of computational chemistry, El Agente Q is an LLM-based multi-agent system that dynamically generates and executes quantum chemistry workflows from natural language user prompts.[70]

In our own work, we included LLM reasoning within a Bayesian optimisation research advisor (BORA),[71] and showed that this combined method could outperform BO or LLMs used in isolation for problems such as multicomponent photocatalyst optimisation. We built BORA so that it could move dynamically between LLM reasoning and BO, based on a computed "trust score" (Section 2). Specifically, we used the OpenAI gpt-4o-mini model in our first trials.[71] Here, we extend this approach to more recent reasoning models, both in LLM/BO hybrids and as standalone optimisers. We explore in detail the extent to which machine reasoning can be used within closed loop scientific experiments, focusing on both the successes and the failure modes. While we find clear examples where LLMs fall below human intelligence, our results suggest that these tools might be used in tandem with knowledge input from researchers; for example, by guiding experimental searches of large, high-dimensional experimental spaces where the frequency of reasoning input required does not align with human working patterns. We also explore the effect of batch size and the ability of the LLM to use external input in the form of human hypotheses, related background literature, and prior experimental data.

2. Methodology

2.1. Bayesian optimisation research assistant (BORA)

The core methodology used here for hybrid BO/LLM optimisations was described in our earlier publication.[71] Briefly, we created a language-based Bayesian Optimisation Research Assistant (BORA), which is a hybrid framework that enhances BO with the domain knowledge and reasoning capabilities of LLMs. BORA uses an adaptive heuristic policy to determine when to involve the LLM, which can warm-start the optimisation, provide real-time commentary, and generate new hypotheses to guide the search out of local maxima. This dynamic integration allows BORA to blend stochastic inference from BO with contextual insights from the LLM—initially the OpenAI gpt-4o-mini model—leading to faster convergence and more efficient exploration of complex, high-dimensional search spaces in scientific experiments. BORA has three possible actions. It can proceed with a standard vanilla BO (Action a1) to select the next single most promising point (or batch of points) for evaluation. When progress stalls and model uncertainty is high, the LLM can fully intervene (Action a2) to analyse the experiment and to suggest a new set of points based on its own reasoning and hypotheses. As a hybrid option (Action a3), the LLM can instead guide the search by selecting the most promising subset of candidate points that were originally generated

by BO.[71] In the new study here, we evaluated a range of more recent tool-enabled LLMs within this hybrid framework. We also extended it to LLM-only optimisations (Section 3.2); that is, using Action a2 to select all the points.

2.2. Opsight software

All optimisations were conducted using Opsight, a cloud-native platform for hypothesis-driven optimisation that we developed to operationalise our hybrid optimisers such as BORA (https://opsightlab.com/).[72] This software enabled the large-scale benchmarking studies presented here using MS Azure's scalable compute resources (see SI, Section 2, for more details). Opsight has a distributed architecture and detailed logging capabilities, which allowed us to capture the optimisation runs with full reasoning traces. This web-based platform also enables real-time visualisation and analysis of the data (*e.g.*, Fig. 2, below). The LLM is free to choose its own optimisation strategy in each run, but the Opsight software was scripted to produce certain standard sub-sections, which can be seen in any of the 624 optimisation logs supplied in the data repository. These sections were as follows:

- BORA research summary – what the LLM has learned from initial research, including any tool usages (*e.g.*, Google searches for relevant publications).
- A list of hypotheses created by the LLM and experimental points linked to these hypotheses (see *e.g.*, Section 3.2 for examples), along with a brief rationale and a confidence level for each hypothesis.
- A series of summaries throughout the optimisation discussing learning points and next steps (see *e.g.*, Section 3.2).

Once the experimental budget is exhausted, Opsight then produces the following sections:

- Executive summary.
- Optimisation overview (detailing initial hypotheses and major parameter adjustments)
- Results and key insights.
- Recommendations for future experiments.

The LLM was instructed never to suggest variables that are not available in the optimisation (*e.g.*, not to suggest components outside of the 10 available chemicals in the photocatalyst optimisation, see Section 3.1.1). However, for "Recommendations for future experiments", the LLM is unconstrained, and it typically does suggest adjustments, such as new chemical additives, that were not available in the optimisation run.

While each API call to the LLM was stateless by design, memory is persisted within a given optimisation run and kept strictly sandboxed. As such, the learning from one run cannot leak into other runs.

The prompt for the optimisations was presented in the form of an "experiment card".[71] For the photocatalysis optimisation, this prompt was as follows:

"The experiment is about maximizing the rate of hydrogen production from a photocatalyst mixture of different chemicals in water. The mixture is exposed to ultraviolet and visible light to generate hydrogen, and the amount produced is measured using gas chromatography. The goal is to enhance the hydrogen evolution rate (HER) in μmol h^{-1} through modifying the quantity of the chemicals such as adjustments to pH, ionic strength, addition of dyes and surfactants and so on, in the mixture.

The experiment has the following intra-point constraints:

- Total volume: the sum of all parameters excluding P10 should not exceed 5 mL to fit inside the vial.

The goal of this experiment is: maximize hydrogen evolution rate".

We did not provide any explicit contextual information in the experiment card about what a good result was—that is, in the photocatalysis case, we did not state what constitutes a high HER value—but the LLM was free to use background literature as a benchmark, and there were multiple examples where this occurred.

3. Results

3.1. Effect of LLM choice on hybrid LLM/BO optimisation

Our original BORA study used the OpenAI gpt-4o-mini model,[71] chosen because it was one of the most cost-effective models available at that time. Since then, a range of new LLMs have been released, including reasoning models. Here, we evaluated the effect of LLM choice on two different problems: a 10-dimensional (10-D) chemistry problem (photocatalytic hydrogen evolution),[32,71] and a 7-D physics problem (a simulation of the game pétanque).[71] In both cases, we used *in silico* models as the ground truth to allow us to carry out statistically meaningful numbers of repeat optimisations (10–20). The details of these two experiments and models were explained in our earlier publication.[71]

3.1.1. Effect of LLM choice, 10-dimensional photocatalytic H_2 evolution problem. The effect of LLM choice on this hybrid LLM/BO optimisation is illustrated in Fig. 1b–g. Briefly, the design space[32,71] involved 10 variables: a photocatalyst (P10), three dyes, two bases, salt (NaCl), two surfactants, and a sacrificial agent (L-cysteine). Apart from the photocatalyst, P10, where the minimum addition was 1 mg, the other nine variables were allowed to range from zero to a maximum of 5 mL, with the additional constraint that the total combined volume should be less than or equal to 5 mL.

We evaluated five LLMs: four from OpenAI (o4-mini, o3, gpt-5-mini, gpt-5) and one from Google (gemini-2.5-flash). The o3 and gpt-5 models are high-end reasoning models. gemini-2.5-flash is a general reasoning model with a large context window (1 M tokens at the time of writing). gpt-5-mini is a lighter variant of gpt-5 that is cost/speed oriented. Likewise, o4-mini is a leaner, more efficient reasoning model. Unlike the OpenAI gpt-4o-mini model that we used in our first BORA study,[71] these instruction-tuned LLMs are explicitly trained to generate and internally evaluate multi-step reasoning trajectories, using reinforcement learning and/or supervised signals that reward consistency, correctness, and intermediate logical structure rather than only final answers. At inference time, these LLMs natively generate an internal monologue that encourages explicit reasoning, allowing the model to decompose the optimisation problem, reflect on intermediate

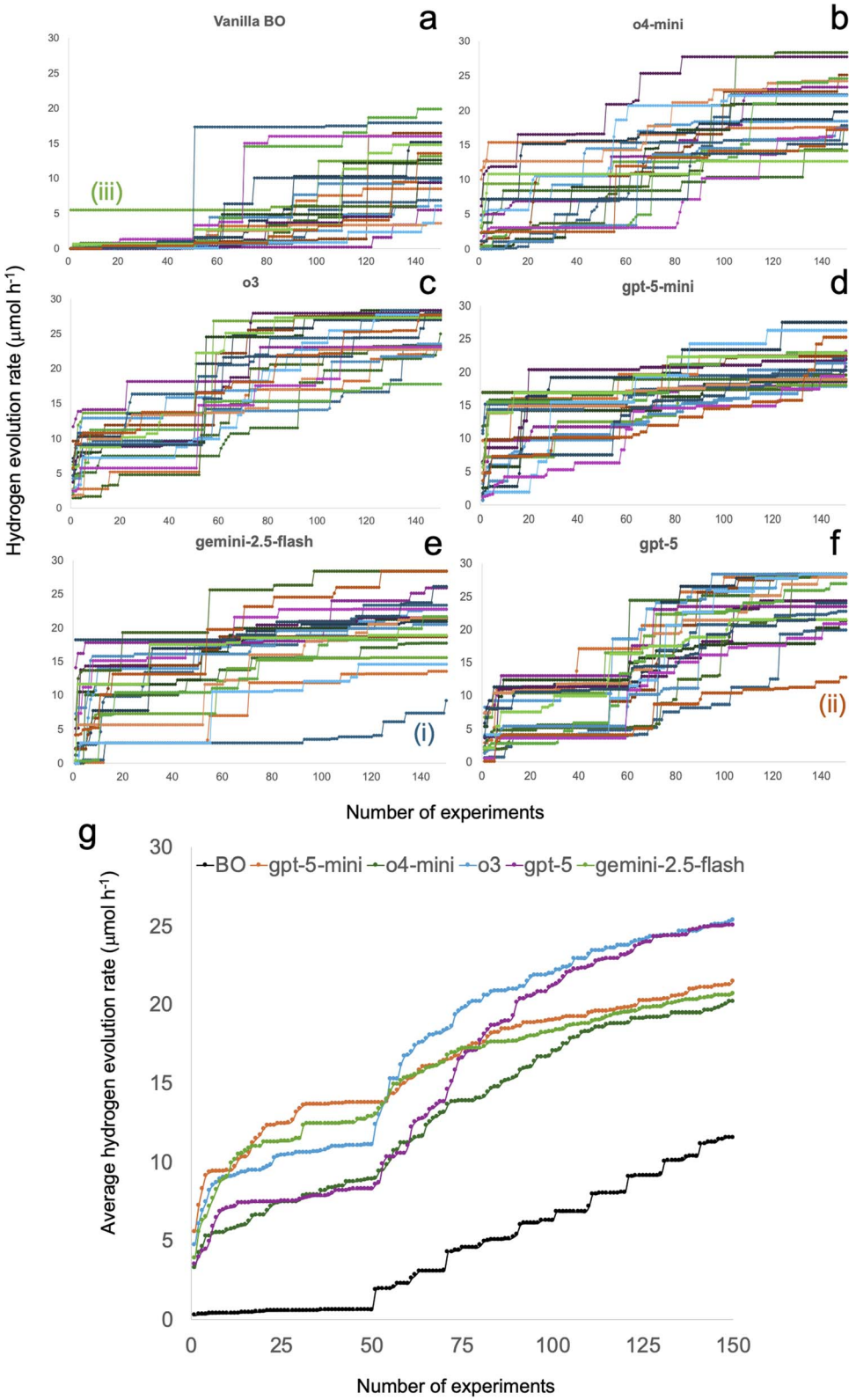


Fig. 1 Effect of LLM choice on hybrid LLM/BO optimisation (BORA) for a 10-dimensional photocatalysis problem. Each run involved 15 batches of 10 experiments (20 repeat optimisations in each case); the first 5 batches used the LLM to warm start the optimisation, after which the hybrid optimiser switched between LLM and BO modes (ref. 71). (a)–(g) show the best hydrogen evolution rate (HER) value found so far. The maximum possible HER in this test is 28.37 μmol h^{-1}. (a) Vanilla BO (ref. 71). (b) BO/o4-mini. (c) BO/o3. (d) BO/gpt-5 mini. (e) BO/gemini-2.5-flash. (f) gpt-5. (g) Average values over 20 repeats for the different models. Outlier runs (i)–(iii) are discussed in the text.

Table 1 Performance of the different LLMs in hybrid LLM/BO optimisers (BORA) after 25 and 150 experiments (units: μmol h^{-1}), averaged over 20 repeat runs. Batch size = 10; 15 batches. LLM reasoning level = medium in all cases. The reasoning levels (low/medium/high) control the LLM's computational effort during response generation. The effects of this are explored in the SI (Fig. S4). The full log files for all 120 of these optimisations are provided in the data repository

Model	Average HER after 25 expt	STD after 25 expt	Average HER after 150 expt	Average STD after 150 expt	# Times reached max
BO	0.6	1.2	11.6	4.4	0
o4-mini	7.5	5.1	20.2	4.4	1
gemini-2.5-flash	11.3	5.1	20.7	4.9	2
gpt-5-mini	**12.5**	4.5	21.5	2.8	0
gpt-5	7.5	3.5	25.1	4.2	**9**
o3	10.5	3.6	**25.3**	2.9	4

hypotheses, and revise proposals before producing an output. One downside of this, as discussed in Section 4, is that these models are significantly more expensive and energy intensive.

We also compared these hybrid optimisations with vanilla BO only,[32,71] without the LLM component (Fig. 1a), which provides no categories or ontology. We used a batch size of one in our earlier study on LLM/BO hybrids[71]—that is single, sequential experiments. Here, we selected a batch size of 10 experiments, initially, to give a total budget of 150 experiments over 15 batches. This batch size was chosen to reflect likely set-ups in chemical and materials research where it is common to automate small batches of experiments to accelerate throughput.[32]

Under these settings, BO alone did not find any catalyst mixtures with significant HER values in the random sampling phase (first 50 experiments) over 20 trial runs (Fig. 1a and g) with just one outlier, labelled (iii) in Fig. 1a, where a composition with HER > 5 μmol h^{-1} was found in the first experiment by chance. This poor initial performance for BO stems from the large size of the experimental space (10-dimensions, 10 discretised levels per variable), in which many combinations give HER values close to zero. We observed better performance with high-dimensional Bayesian optimisation with sparse axis-aligned subspaces (SAAS BO; SI, Fig. S1),[73] but the performance was still less strong than the BO/LLM hybrids. We note that there are many variants of BO, and that performance might be improved further by using different kernels, for example.

The optimisation performance for the different LLMs, averaged over 20 repeat runs, is shown in Fig. 1g and summarised in Table 1. In all cases, the LLM/BO hybrid optimiser, BORA, outperformed BO alone, at least under the settings that we explored here. As we observed previously with the OpenAI gpt-4o-mini model,[71] this was most pronounced in the initial exploration phase (first 50 experiments) where the LLM warm-started the optimisation by creating hypotheses about the chemistry (Fig. 1g). In some cases, the LLM located our earlier publication on this specific catalysis chemistry.[32] While that paper did not directly specify the best catalyst compositions, the LLM was often able to extract general rules (*e.g.*, avoid dyes and surfactants, add bases), which greatly helped the 'warm start', although for reasons that are explained later (Section 3.3.3; Fig. 12e), the fine details of this paper can actually reduce the best HER found after 150 experiments.

On average, the more advanced reasoning models, o3 and gpt-5, reached better solutions—that is, higher hydrogen evolutions rates (HERs)—after 150 experiments (Fig. 1g, and Table 1). The o3 model gave a tighter distribution of results and a lower standard deviation. In the early stages of the optimisation (experiments 1–50), the o3 and gpt-5 models performed relatively less well; indeed, the lighter gpt-5-mini model and gemini-2.5-flash outperformed o3 and gpt-5 in this regime, providing a faster warm start and somewhat higher HERs after the first 25 experiments (Table 1). We do not at present have an explanation for this. In the reasoning phase (experiments 51–150), the o3 and gpt-5 reasoning models were superior, eventually outperforming gpt-5-mini and gemini-2.5-flash (Fig. 1g). The o4-mini model performed least well within the hybrid BORA framework, although it still outperformed BO alone. Overall, the o3/BO hybrid gave the highest average HER value (25.3 μmol h^{-1}) after 150 experiments averaged over 20 runs; the standard deviation was also lower than for BO (2.9 *vs.* 4.4 μmol h^{-1}). The maximum HER value that was possible in these simulated experiments was 28.37 μmol h^{-1}. The o3/BO hybrid found this optimum value in 4/20 runs and it came close to it (within 1 μmol h^{-1}) in a further 5/20 runs. The gpt-5/BO hybrid found the optimum more often (9/20 times; Table 1), but its average HER value was reduced by outlier runs (*e.g.*, run (ii), Fig. 1f). The search performance for the o3/BO and gpt-5/BO hybrids surpassed vanilla BO, where the optimum was not found in any of the 20 runs within this experimental budget (Fig. 1a).

Both the LLM and BO components of the hybrid optimiser are stochastic; that is, the outcomes may be analysed statistically but not predicted exactly, and hence a distribution of performance is to be expected. However, in some cases, we observed outlier runs that gave final HER values that were more than two standard deviations from the average value. For example, one of the gemini-2.5-flash runs, labelled (i) in Fig. 1e, reached a final HER value of just 9.25 μmol h^{-1} after 150 experiments. Our Opsight software creates detailed logs for each optimisation run. Inspection of the Opsight log for run (i) showed that despite creating a reasonable optimisation strategy, the LLM failed to produce correctly formatted initial points after several attempts, and hence random points were used instead, which is the default fallback strategy with the Opsight software should the 'warm start' LLM batch production fail. This initial random sampling over the first 5 batches of 10

experiments explains the slow start in run (i). After the first 5 batches in run (i), the LLM was then able to create hypotheses and to suggest new batches of points, but these hypotheses were not strong enough to fully offset the lack of a warm start in the optimisation. By the end of run (i), the LLM did infer from the simulated experimental data that dyes and surfactants were a negative influence on this reaction, stating the following:

"Absence of dyes and surfactants: notably, the best sample contains no AcidRed871, MethyleneBlue, RhodamineB1, PVP, or SDS. This indicates that these specific dyes and surfactants, at the concentrations explored, are either ineffective or detrimental to HER in this particular P10-based system, especially when combined with sodium disilicate. This discards initial hypotheses that these dyes would act as photosensitizers or that surfactants would universally enhance the process".

These conclusions are broadly correct,[32] but by the time they were reached, the optimiser had exhausted the available budget of 150 experiments.

Another outlier was found in the gpt-5 optimisations, labelled (ii) in Fig. 1f. The final HER reached was 12.77 μmol h^{-1}, which is less than half the average HER value achieved by this model averaged over 20 runs (Table 1). Unlike run (i), above, the LLM did in this case produce initial sets of points based on literature searches and reasoning, creating a quite detailed "mechanistic background" prior to the optimisation (see optimisation log in data repository, gpt-5-med-13 august 2025_Run_20.pdf). The LLM found papers about the effect of NaCl on hydrogen production using related inorganic semiconductor catalysts,[74] as well as a specific mechanistic paper on the polymer photocatalyst, P10.[75] The LLM also found and cited studies on the use of dye sensitizers for photochemical H_2 production.[76,77] Perhaps because of this, the addition of dyes was emphasised in the early experiments.

The LLM produced the following superficially reasonable conjectures before initiating the first experiments:

"I expect the highest HER to emerge from a neutral-to-mildly basic, low-ionic-strength formulation that pairs P10 with a modest amount of a good hole scavenger (L-cysteine) and a visible-light sensitizer (Eosin Y), plus just enough PVP to stabilize the dispersion without passivating interfaces. Mechanistically, P10 benefits from a sacrificial donor to suppress recombination, while Eosin Y extends absorption into the green; together they should create efficient charge separation provided ionic strength and surfactant coverage remain low".

The next two batches of experiments in run (ii) used BO, but this failed to improve upon batches 1–5, whereupon BORA reverted again to LLM sampling. The LLM reasoning now recognised that dyes do not seem to benefit the HER value, and instead investigated whether small amounts of a surfactant, PVP, "improves dispersion at higher P10 without passivating interfaces". In fact, PVP also lowers the HER value, which was later recognised by the LLM. However, at some point the LLM concluded, wrongly, that the addition of base is detrimental to HER, making the following comment at batch 11, effectively sending the search down the wrong (base-free) path:

"Plan: avoid dyes/base/salt/surfactant, fine-scan P10 around 2.6–3.4 mg and tune LCysteine 0.25–0.75".

Subsequent batches did explore base addition, but the poor early choices to emphasise dye sensitisation and to avoid base left insufficient experimental budget to reach higher HER values. Interestingly, the LLM appeared to realise this, since the first point in the "Recommendations for Future Experiments" was as follows:

"Reproducibility and mapping near the optimum: fine-scan P10 4.4–5.0 mg (step 0.2 mg), LCys 1.0–1.5 mL, sodium disilicate 2.5–3.5 mL, and NaCl 0–0.25 mL, enforcing total liquid ≤5 mL".

This is a good set of suggestions, although in fact the optimum value is found with a mixture of two different bases (NaOH plus sodium disilicate), a fact that was missed in run (ii) because of the early decision to avoid addition of base.

Run (i) and run (ii) are negative outliers: in most of the hybrid optimisation runs, the LLM component in BORA warm-

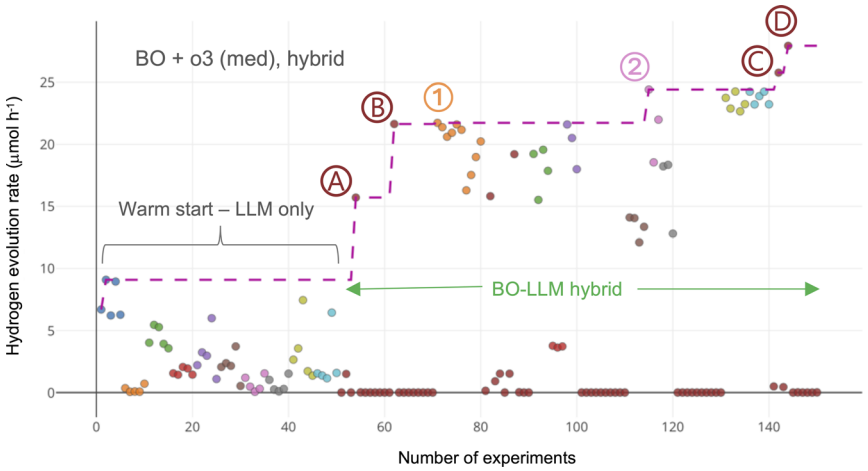


Fig. 2 Single illustrative hybrid BO/o3 run (BORA) showing both LLM points and BO-derived points; the LLM points are colour coded by hypothesis (key not given here; see Fig. 5 for an example). Batch size = 10; 15 batches. Symbols labelled ① and ② denote LLM hypotheses that improve hydrogen evolution rate (HER) directly after the 50-experiment warm start phase. Points labelled Ⓐ–Ⓓ are the points where the BO sampling improved the HER after the LLM warm start in this hybrid adaptive optimisation.

started the optimisation effectively and improved performance over BO alone (Fig. 1a–f).

An illustrative example of a single o3/BO hybrid run is shown in Fig. 2 (see data repository: o3-med-31 july 2025_Run_19.pdf). The first five LLM-derived, hypothesis-led batches provided a 'warm-start' for the optimisation (Fig. 2). In this warm-start phase, the LLM relied on the knowledge contained in its foundation model, plus any information that it gained from literature searches using external tools. In the example run shown here, literature searches within the LLM located a publication relating to the specific polymer photocatalyst, P10,[78] and another study concerning the use of the surfactant SDS in a water splitting experiment[79] where the overall effect of the surfactant was negative.

The first LLM-derived hypothesis, "alkaline high P10 w/ cysteine" (the five initial blue points in Fig. 2) led to a HER value of 9.09 μmol h^{-1} in experiment 2—higher than any values found in the 20 random BO initialisations (Fig. 1a) or in non-random Kennard–Stone sampling of the search space (Fig. S5). Other LLM hypotheses in the first five warm start batches did not improve on this HER value, but upon switching to BO sampling, an HER value of 15.71 μmol h^{-1} was found in experiment 54 (point Ⓐ, Fig. 2) and additional BO sampling led to a further improvement (to 21.65 μmol h^{-1}) in experiment 62 (point Ⓑ). An LLM hypothesis in experiment 71 made a marginal improvement in the HER (to 21.74 μmol h^{-1}; hypothesis = "high-P10/high-cysteine/high-ionic strength"; point ① in Fig. 2). The LLM then hypothesised "Reduced P10 for Better Light Penetration", which led to a more substantial increase to 24.43 μmol h^{-1} (experiment 115, point ②). This was an important change that involved the LLM recognising that adding more of the photocatalyst, P10, could have diminishing returns beyond some optimal value.

Further LLM hypotheses did not improve the HER and the last two jumps in HER (experiments 142 and 144, points Ⓒ and Ⓓ) came from BO sampling. This run illustrates how BO and the LLM component can work in tandem. However, while BO sampling did advance the search profitably (points Ⓐ–Ⓓ, Fig. 2), it also produced many catalyst compositions with HERs that were close to zero. In a real-life experiment, this would add significant time and cost. The proportion of near-zero BO-derived points, which was also observed in the other 19 runs in this trial, prompted us to also study LLM-only optimisations, as discussed in Section 3.2.

Next, we explored the role that tool usage, particularly Google searches, was playing in these hybrid optimisations. Specifically, we instructed the LLM not to carry out literature searches throughout the experiment. This was done by attaching a PDF file with an instruction not to consult the literature (see Fig. S2). Tests with the gpt-5 model showed that the LLM followed this instruction throughout the entire optimisation in only 50% of the 20 runs (see also 3.3.2). In the ten cases where literature searches were performed, in contravention to the instructions, the LLM tended to locate publications that were relevant to the chemistry studied here, especially papers relating to the polymer, P10, as discussed above.[75,78] Nevertheless, the average hybrid optimisation performance was very similar with and without background literature searches (SI, Fig. S3). This tallied with our general perception from reading optimisation logs across multiple experiments. We often found examples where the LLM produced useful initial hypotheses that were derived from related papers discovered in literature searches. Equally, the LLM could also produce less productive hypotheses from papers that were tangentially related. To give a specific example: in several cases, the LLM found publications to support the hypothesis that organic dyes can photosensitise semiconductor photocatalysts, which in turn prompted the model to trial dyes in early experiments. This is not unreasonable because it was our own earlier studies on dye sensitisation[80] that inspired us to include dyes in these experiments in the first place. There is a general literature bias toward positive results; that is, there are multiple studies on the success of dye sensitisation but only one paper (our own 2020 study)[32] that shows that these three particular dyes are ineffective in this specific chemistry. If the LLM were able to locate that paper[32] consistently and to prioritise it as being more relevant, then the optimiser would likely perform more effectively. However, this level of discrimination seems to be beyond the current literature search capabilities of these LLMs. It seems, therefore, that the benefits of literature searching cancel out, at least for this catalysis problem; that is, for each case where a publication aids the optimisation (*e.g.*, "avoid SDS",[79] above), there is a corresponding case, on average, that is unproductive (*e.g.*, "add dye sensitizers").

Taken together, our observations suggest that the improved optimisation performance over BO (Fig. 1g) for this photocatalysis problem stems mostly from well-established chemistry ideas that are embedded in the foundation model (*e.g.*, bases can deprotonate cysteine to make it a better donor), coupled with active reasoning about data as it emerges from the experiment (*e.g.*, learning that dyes, as a class, decrease hydrogen evolution rates). Literature searching by the LLM makes a less obvious net contribution to the optimisation performance for this photocatalyst problem (Fig. S3) using the models tested here. By contrast, as illustrated in Section 3.3.3, the addition of hand-selected literature sources can greatly boost the optimisation performance, especially in the early stages of the search. Overall, we would expect background literature searching by LLMs to be more consistently helpful for clear-cut problems such as our next example, which is based on established, uncontested physics.

3.1.2. Effect of LLM choice, 7-dimensional physics simulation of the game pétanque. We also tested the effect of LLM choice on hybrid BO/LLM optimisations of a 7-D physics-based simulation of the game pétanque[71] that we created ourselves (Fig. 3). There are no papers on specific solutions to this test problem, and our earlier publication[71] used it as a benchmark but did not discuss the optimal solutions that were found. In this case we used a reduced set of models (o4-mini, o3, gpt-5), again comparing with BO alone. As before, 15 batches of 10 experiments were used in each run. A score of 100 here equates to a perfect result—that is, zero distance between the thrown ball and the target ball (the jack). Unlike the photocatalysis case, where there is a single maximum HER value, there are a vast number of valid solutions that can reach this optimal

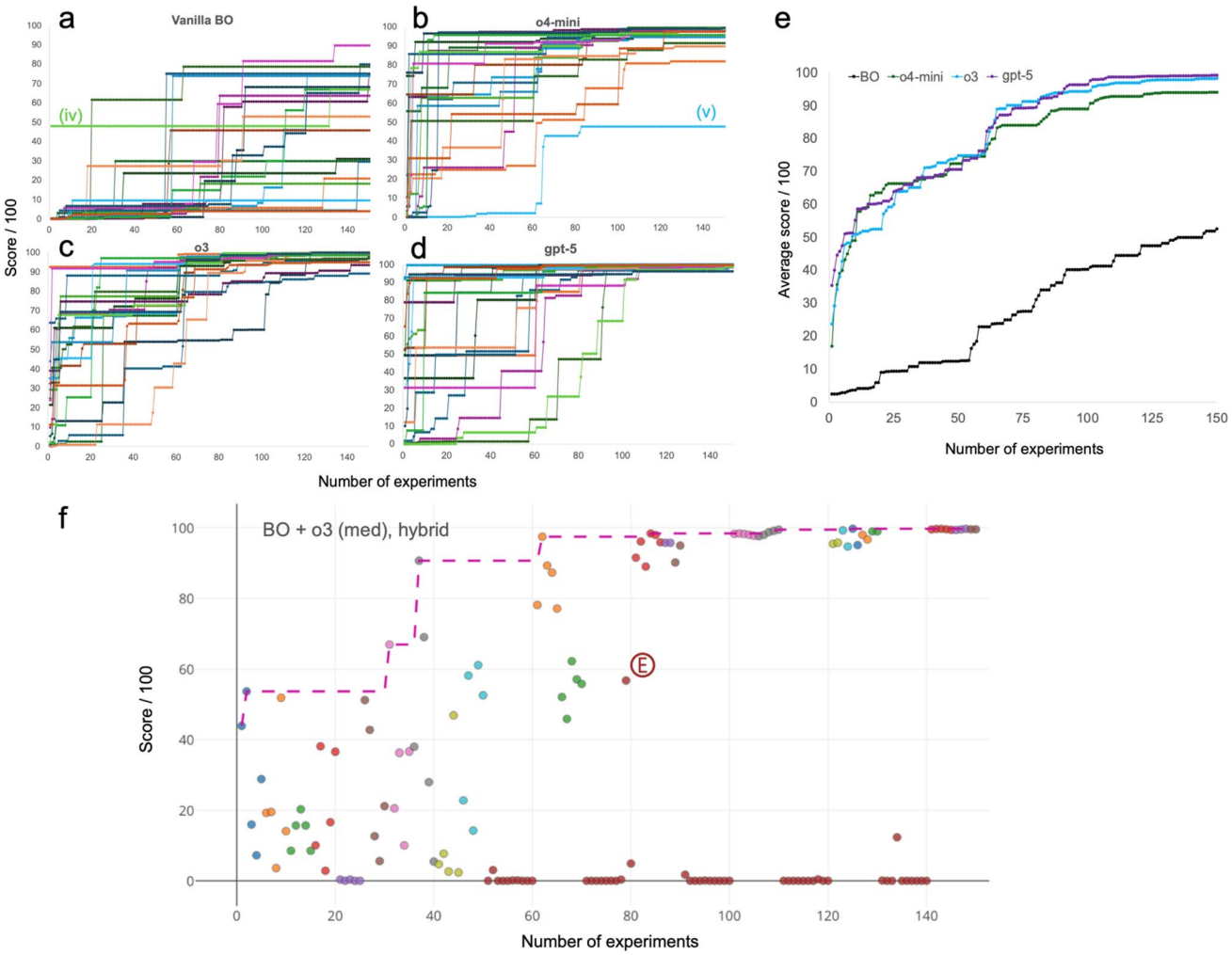


Fig. 3 Effect of LLM choice on LLM/BO hybrid optimisation (BORA) for a 7-dimensional physics simulation of the game pétanque that we built. Each run involved 15 batches of 10 experiments (150 experiments); the first 5 batches used the LLM to warm start the optimisation, after which the hybrid optimiser switched between LLM and BO modes, as defined in our previous study (ref. 71). A total of 20 repeat optimisations was run in each case; all plots show the best value found so far. The maximum possible score in this test is 100, which represents a perfect shot. (a) BO only (ref. 71). (b) BO/o4-mini. (c) BO/o3. (d) gpt-5. (e) Average values over 20 repeats for the different models. (f) Single illustrative hybrid BO/o3 run; point Ⓔ is the best score found by BO sampling after the 50-experiment LLM warm start phase.

score. Nonetheless while the dimensionality of this problem is lower than the photocatalysis example (7-D *versus* 10-D), and there are many valid solutions, it is a challenging task for BO because the optimal solutions are located on a knife-edge. That is, small perturbations in the input variables (*i.e.*, pitch, yaw, velocity, spin, *etc.*) can lead to large deviations from the target. This is apparent from the wide spread of scores achieved by BO alone over 20 repeat runs (Fig. 3a) and the large standard deviation in the final score (Table 2). Also, unlike the photocatalysis case, there are no input variable categories in the pétanque problem that can be dropped, analogous to reasoning that dyes, as a class, are "bad"—all seven input variables matter throughout. The average score achieved by BO under these settings was 52.6 and none of the 20 runs approached the perfect score of 100 (Fig. 3a). Even runs such as (iv), where the initialisation phase starts with a "lucky shot", do not approach the maximum score within the experimental budget.

In the chemistry example above (Section 3.1.1), the catalytic activity was defined by factors at multiple length scales that are not all fully understood. By contrast, the underpinning physics for this pétanque simulation are known, and the background literature is much more homogeneous. In principle, therefore, LLM reasoning should have a good chance of improving the

Table 2 Performance of the different LLMs after 25 and 150 experiments in LLM/BO hybrid optimisation (BORA) for a 7-dimensional physics simulation (max. score = 100), averaged over 20 repeat runs

Model	Average score after 25 expt	STD after 25 expt	Average score after 150 expt	Average STD after 150 expt
BO	9.3	17.2	52.6	26.8
o4-mini	**66.2**	27.6	94.1	11.7
o3	59.9	31.0	98.2	2.8
gpt-5	63.9	35.1	**99.2**	0.9

performance of the optimisation. We found this to be true: unlike BO only, the three hybrid LLM/BO optimisations all approached the perfect score of 100 after 150 experiments (Fig. 3e). As for the photocatalysis problem, the o3/BO hybrid performed well (Fig. 3c), leading to an average final score of 98.2 and a standard deviation over 20 runs of just 2.8 (2.9%, *cf.*, STD = 51% for BO alone). Analysis of the 20 runs performed using the gpt-5/BO hybrid suggested that the optimisations occurred by reasoning, rather than by memorisation, for example, of data relating to this problem that were somehow embedded in the LLM training (Table S3).

For our catalysis example, BO sampling and LLM reasoning contributed in tandem to the optimisation (Fig. 2). By contrast, for this pétanque simulation, the LLM reasoning was dominant. Indeed, in the example o3/BO run shown in Fig. 3f, the BO sampling did not contribute to advancing the optimisation, beyond providing additional points for LLM reasoning, albeit mostly with near-zero and hence unhelpful scores. The best point obtained by BO sampling is labelled Ⓔ in Fig. 3f. All direct gains in score were made by LLM reasoning alone.

Analysis of the Opsight logs showed that the LLMs could produce a logical search strategy, based on known physics, prior to the first five warm-start batches. All the logs are included in the data repository; here is an excerpt from one run (PT_20_run_10_5_15_o3_Run_20.pdf) that used the o3 model to generate the following strategy:

"Maximizing score hinges on delivering a 50 m carry with minimum lateral error while preserving a steep-enough descent to stop near the jack. Classical projectile theory, adjusted for drag and Magnus lift, says that a back-spin-aided arc (pitch $\approx$35–40°, $v_0 \approx$25–30 m s^{-1}) is the natural sweet-spot. Moderate mass (0.3–1 kg) keeps the ball from ballooning yet still lets the Magnus force matter. Spin around the – y axis near 2000 rpm adds 10–20% effective range and softens landing, and only a couple of degrees of yaw are needed to trim cross-range drift. Beyond $\approx$2600 rpm lift gains saturate, and very light balls (<0.1 kg) or very heavy ones (>5 kg) are probably sub-optimal, but exploring those extremes will teach the surrogate model about drag/Magnus scaling. Likewise, contrasting "flat-fast" and "lofty-slow" trajectories maps the ridge of optimal pitch–velocity combinations. The initial 5 batches below therefore concentrate samples in the high-payoff region while peppering in strategic outliers: high-spin, no-spin super-heavy, ultra-light max-spin, opposite yaws, and pitch extremes. These 50 points should give the Bayesian optimizer a well-shaped prior and rapid path to the hundred-point bullseye".

This strategy produced a best score of 69.3 in the first five batches, which was further improved by subsequent LLM hypotheses to a final score of 99.8. As for the example shown in Fig. 3f, BO sampling did not contribute directly to improving the score, despite the LLM's stated intent to produce a "well-shaped prior", and the direct advances in score were made by LLM reasoning. In its final summary, the LLM made the following sensitivity analysis:

"The optimum is extremely sharp: $\pm$0.1 m s^{-1} in velocity or $\pm$0.05° in pitch changes the score by >1 point; $\pm$100 rpm in spin changes it by $\approx$ 6 points".

This observation largely explains the poor performance of BO.

The final scores by the o3 model were consistent after 150 experiments (Fig. 3c). For the o4-mini model, we observed some negative outliers, as noted in the chemistry example (Section 3.1.1). For example, run (v) (Fig. 3b) achieved a final score of 47.2, which was around half the average performance for that model. The LLM did produce hypotheses and suggested points for the first five batches, but these hypotheses lacked detail, were overly general, and they hence proved unproductive; *e.g.*, "Slow, high arcs with moderate spin create long flight for pinpoint landings", "Backspin creates downward Magnus near target to limit overshoot" (for full log, see data repository, PT_20_run_10_5_15_o4-mini_Run_16.pdf). This suggests inherent reasoning limitations in the o4-mini model compared with the more detailed and precise strategies produced by o3, an example of which is reproduced above. The o3/BO hybrid produced no such poorly performing runs (Fig. 3c).

The gpt-5 model performed even better than the o3 model, achieving high consistency for this task after around 100 experiments (Fig. 3d) and a final average score of 99.2 after 150 experiments, close to the maximum score of 100.

For the gpt-5/BO hybrid, we also explored the influence of the reasoning level setting (low, medium, high), both for the photocatalysis and pétanque optimisations. The average performance at these three settings is shown in Fig. S4. For the photocatalysis problem, the low reasoning level setting performed somewhat better (Fig. S4a and Table S1), mainly because there were fewer outliers. For the pétanque problem the differences were marginal (Fig. S4b and Table S2), possibly because this well-defined physics problem involves fewer competing hypotheses.

3.2. Optimisations using LLM reasoning only

3.2.1. Effect of LLM choice for LLM-only reasoning (batch size = 10). The performance of the LLM component in the hybrid LLM/BO optimisations using reasoning models (Section 3.1) prompted us to explore optimisations with the LLM only—that is, without the BO component. Again, we tested this both for the 10-D photocatalysis problem and the 7-D physics problem. The results are shown in Fig. 4, which compares the average performance over 20 runs of the LLM/BO hybrid with pure LLM reasoning for various models using the same batch settings as above (Section 3.1). We note that these LLM-only strategies are also hybrids, of a sort, because the models call on basic data analysis tools (see *e.g.*, Fig. 6).

For the photocatalysis optimisation, we found that the LLM-only method matched or slightly surpassed the performance of the LLM/BO hybrids for the four models tested (Fig. 4a–e), except for the gpt-5 model (Fig. 4c), where the hybrid optimiser performed better. This was mostly because of low-performing outliers that stemmed from the gpt-5 component, as also observed in the gpt-5/BO hybrid optimisations (Table 1; run (ii) in Fig. 1f). As for the LLM/BO hybrid optimisations (Fig. 1g), the o3 model was found to perform most effectively overall when used by itself under these batch settings. Indeed, the o3-only

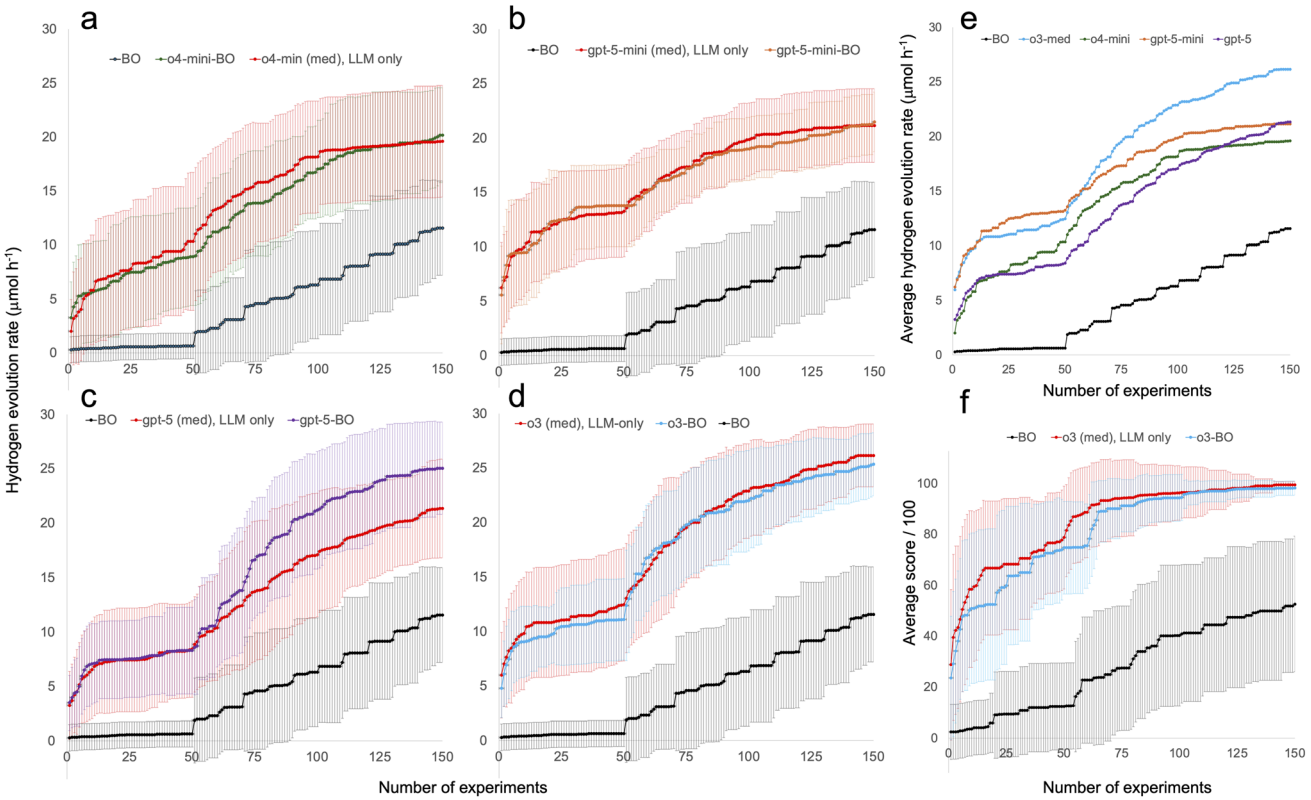


Fig. 4 Comparison of optimisations using LLM reasoning only with BO/LLM hybrid optimisations. BO-only optimisations are also shown for comparison. Each run involved 15 batches of 10 experiments (150 experiments). A total of 20 repeat optimisations was run in each case; the plots show the best value found so far, averaged over the 20 runs. Errors bars are the calculated standard deviations. (a)–(e) Photocatalytic hydrogen production. (f) Physics-based pétanque simulation.

optimisation outperformed all other combinations trialled here for this task (Fig. 4e, and Table 3). The o3-only optimisation also outperformed the o3/BO hybrid marginally, on average (Fig. 4d), but found the optimum twice as frequently (Table 3). We note that LLM-only optimisations for this problem did not outperform LLM/BO hybrids in our previous study using the OpenAI gpt-4o-mini model;[71] indeed, after an initial 'warm start', LLM-only optimisations plateaued at much lower HER values, reflecting the poorer reasoning capabilities of that earlier model.

The rather similar performances of the LLM-only and LLM/BO-hybrid optimisations for this photocatalysis problem (Fig. 4a–d) under these settings can be rationalised by the relatively balanced contribution of the LLM and BO

Table 3 Comparison of optimisation performance for LLM/BO hybrids and LLM-only reasoning (15 batches of 10 experiments = 150 experiments; 5 batches of exploration before active learning; all averages are for 20 repeat runs). Reasoning level = medium. The most effective optimisation results are bolded for both the photocatalysis and pétanque simulations

Photocatalysis expt (20 runs)	Average objective after 150 expt	STD after 150 expt	# Times reached max
o4-mini/BO hybrid	20.2	4.4	0
o4-mini only	19.6	5.2	0
gpt-5-mini/BO hybrid	21.5	2.8	0
gpt-5-mini only	21.2	3.4	1
gpt-5/BO hybrid	25.1	4.2	**9**
gpt-5 only	21.3	4.6	1
o3/BO hybrid	25.3	2.9	4
o3 only	**26.2**	2.9	**9**
BO only	11.6	4.4	0
Pétanque simulation (20 runs)			
o3/BO hybrid	98.2	2.8	0
o3 only	**99.5**	1.2	**0**
BO only	52.6	26.8	0

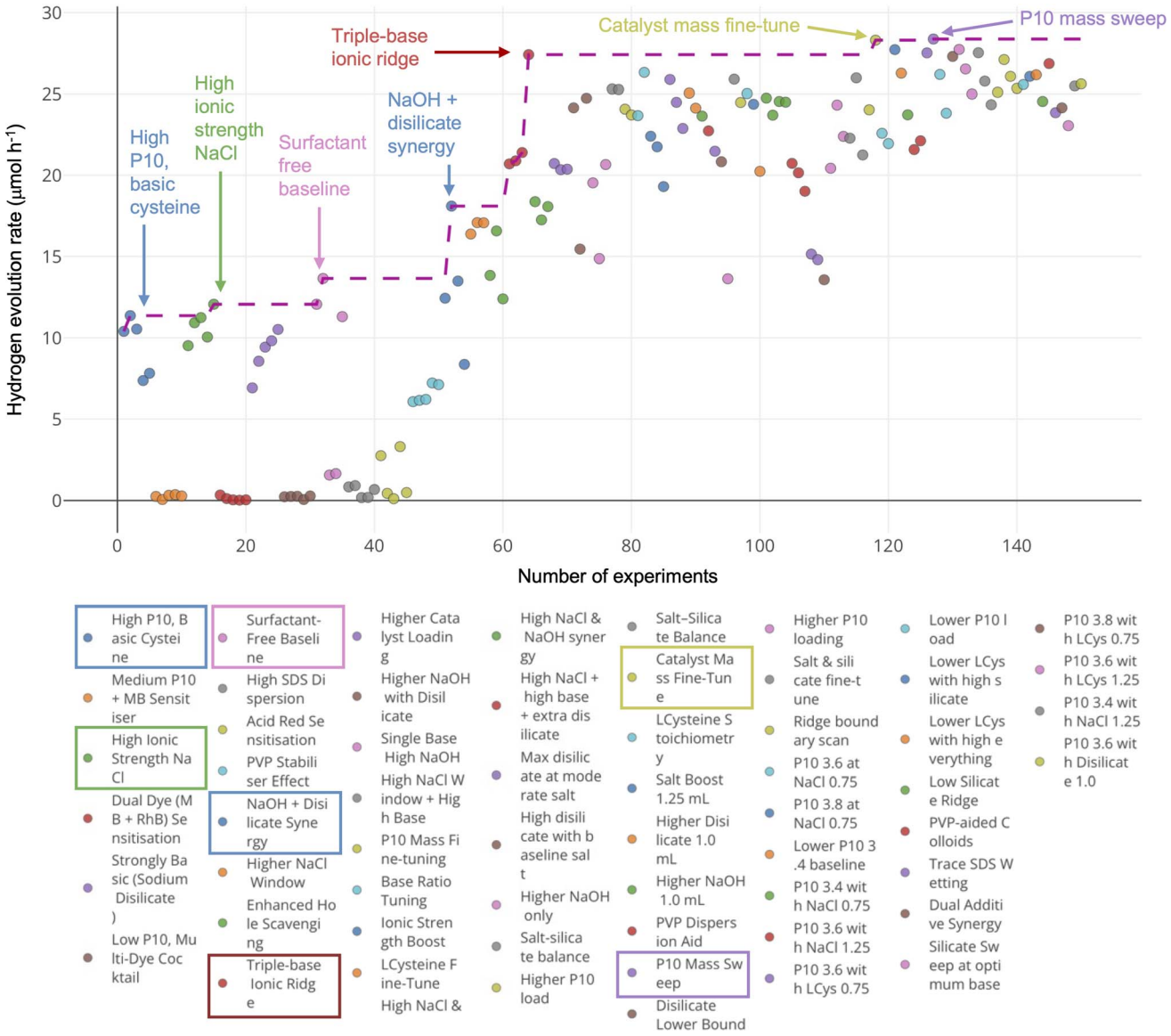


Fig. 5 Example of an LLM-only optimisation for the 10-dimensional photocatalysis problem (o3 model, medium, batch size = 10). There were 5 initial batches (50 experiments) before the active learning phase. The points are colour-coded by LLM hypothesis (see key). There are 59 hypotheses in all over the course of the 150 experiments, and the seven hypotheses that directly advanced the hydrogen evolution rate are highlighted.

components (Section 3.1.1). To take just the single example of the o3/BO run shown in Fig. 2, there were seven jumps in the best HER value found (ignoring experiment 1); three of those jumps derived from LLM hypotheses (experiment 2 plus points ① & ②) while four derived from BO sampling (points Ⓐ–Ⓓ).

The LLM-only optimisations outperformed LLM/hybrid optimisations for the pétanque problem using the o3 model (Fig. 3f); the standard deviation of results after 150 experiments was low in both cases, but the o3-only optimisations gave the tightest distribution of results with a standard deviation of just 1.2% over 20 runs (Table 3). While the average score difference between the LLM-only optimiser (99.5) and the BO/LLM optimiser (98.2) is small in absolute terms, there are chemistry problems, such as maximising the percentage of a pollutant removed from water, where such gains could be significant.

An example of a single run for the o3-only photocatalysis optimisation is shown in Fig. 5, taken from the graphical user interface within our Opsight software. In total, 59 LLM hypotheses were tested over the course of this optimisation and the seven hypotheses that advanced the HER directly are highlighted in Fig. 5. The full optimisation log for this run can be found in the log file, (10 + 5)*10_(o3-medium)_Run_7.pdf, in the data repository. After 5 batches (50 experiments), the LLM calculated variable correlations with HER for the data received so far (Fig. 6) and concluded the following:

"Data so far show a very clear recipe for activity: high P10 (\~4 mg) + ≥1 mL L-cysteine, mildly alkaline (0.5 mL NaOH), and moderate NaCl (≈0.5 mL). Additives that coat the catalyst (SDS, dyes) sharply suppress HER, while extra dispersant (PVP) and high polymer loads beyond 4 mg offer only modest gains.

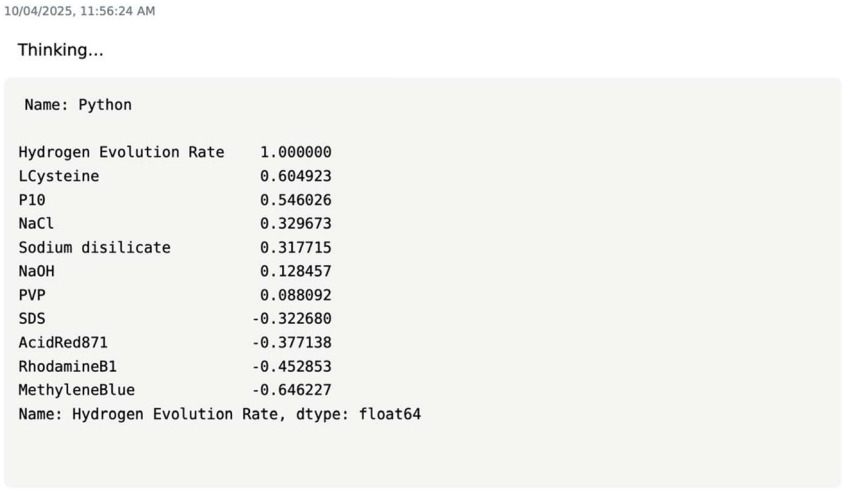


Fig. 6 Variable correlations calculated by the LLM after batch 5 (first 50 experiments). Note that this was not scripted but performed spontaneously by the LLM (o3).

Correlation analysis ranks variables as follows: L-cysteine (+0.60) > P10 (+0.55) >NaCl (+0.33) ≈ sodium-disilicate (+0.32) > NaOH (+0.13); strongly negative: methylene blue, RhB, acid red, SDS. This suggests remaining upside lies in fine-tuning the alkaline environment and ionic strength rather than introducing new organics. Two avenues look promising: (i) supplement NaOH with small sodium-disilicate doses, which may raise pH and supply silicate anions that interact favourably with the polymer surface; (ii) explore whether HER keeps climbing with NaCl up to ≈1–1.5 mL when sufficient base is present. A third variable is the exact cysteine dose—earlier 1.25 mL gave low activity under sub-optimal conditions, but in the now-optimised matrix it may increase hole-scavenging further. The ten proposed points systematically probe these three factors while keeping the proven core unchanged. All satisfy the 5 mL limit and maintain dye- and surfactant-free conditions, thereby targeting incremental yet potentially significant performance gains above the current 13.7 μmol h^{-1} record".

The LLM computed the correlations in Fig. 6 autonomously by invoking the code interpreter tool within Opsight, executing Python code in an isolated sandbox environment to calculate correlation coefficients between each input variable and the HER target across all accumulated data. Based on this reasoning, the LLM produced the following hypothesis and four sets of suggested points:

"NaOH + disilicate synergy

Rationale: test whether combining conventional base and silicate boosts pH and interfacial chemistry.

Confidence: medium.

Points: {'AcidRed871': 0.0, 'LCysteine': 1.0, 'MethyleneBlue': 0.0, 'NaCl': 0.5, 'NaOH': 0.25, 'P10': 4.0, 'PVP': 0.0, 'RhodamineB1': 0.0, 'SDS': 0.0, 'Sodium disilicate': 0.5}leaner {'AcidRed871': 0.0, 'LCysteine': 1.0, 'MethyleneBlue': 0.0, 'NaCl': 0.5, 'NaOH': 0.5, 'P10': 4.0, 'PVP': 0.0, 'RhodamineB1': 0.0, 'SDS': 0.0, 'Sodium disilicate': 0.5} {'AcidRed871': 0.0, 'LCysteine': 1.0, 'MethyleneBlue': 0.0, 'NaCl': 0.5, 'NaOH': 0.25, 'P10': 4.0, 'PVP': 0.0, 'RhodamineB1': 0.0, 'SDS': 0.0, 'Sodium disilicate': 0.75} {'AcidRed871': 0.0, 'LCysteine': 1.0, 'MethyleneBlue': 0.0, 'NaCl': 0.5, 'NaOH': 0.0, 'P10': 4.0, 'PVP': 0.0, 'RhodamineB1': 0.0, 'SDS': 0.0, 'Sodium disilicate': 0.75}".

As shown in Fig. 5, this "NaOH + disilicate synergy" hypothesis advanced the HER. A more substantial improvement in HER was then made in batch 7; again, the LLM recalculated variable correlations (not shown) and provided the following analysis:

"HER has climbed from <14 to 18 μmol h^{-1} after revealing a clear synergy between (i) moderate-high base (0.5 mL NaOH), (ii) supplemental silicate base (0.5 mL), and (iii) raised ionic strength (≥1 mL NaCl). The correlation matrix confirms LCysteine and P10 loadings remain the strongest single drivers, while every organic dye and SDS sharply quenches activity. The current best recipe still consumes only 2.5 mL of the 5 mL liquid budget, leaving room to push pH and ionic strength further or to increase the hole-scavenger concentration. Literature on polymer photocatalysts suggests HER often peaks at pH ≈ 11–12 and at 1–2 M salt; likewise, thiol donors can keep boosting rates up to \∼0.3 M (≈1.5 mL here). Therefore the next batch (i) maps a 2-D ridge of NaCl (1.0–1.25 mL) *vs.* silicate (0.5–0.75 mL) under constant NaOH 0.5–0.75 mL, (ii) tests whether LCysteine 1.25–1.5 mL provides additional hole scavenging when bases are optimised, and (iii) checks if a slight mass increase of P10 (4.2–4.4 mg) yields extra photon harvesting without excessive scattering. All points retain the proven dye-/surfactant-free matrix and stay below 5 mL total volume. This focused exploration should clarify which of the three levers—ionic strength, donor amount, or catalyst loading—offers the next jump beyond 18 μmol h^{-1}".

Here, the LLM generalised about the negative effect of dyes. It also reasoned about how much of the 5 mL "liquid budget" was consumed by the beneficial components. The resulting hypothesis, "Triple-Base Ionic Ridge", improved the HER significantly to 27.4 μmol h^{-1} (Fig. 4). The rationale for this hypothesis was to "Fine-map NaCl × silicate synergy while holding NaOH high".

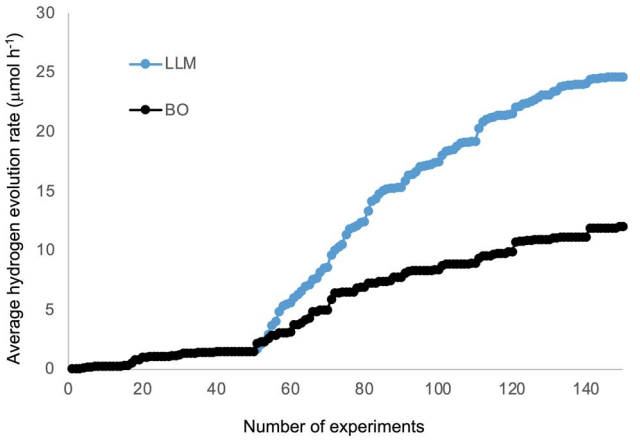


Fig. 7 Comparison of LLM-only (o3, med) and BO-only optimisations for the 10-dimensional photocatalysis problem, initialised with the same 50 points (Kennard–Stone algorithm). Each run involved 15 batches of 10 experiments (150 experiments). A total of 20 repeat optimisations was run in each case; these plots show the average of those 20 runs (see Fig. S5 for plots of all runs).

The LLM reaches the optimum HER (28.37 μmol h^{-1}) in experiment 127 with the following hypothesis:

"P10 mass sweep

Rationale: fine-tune catalyst loading around 3.8–4.0 mg under best chemical environment.

Confidence: high.

Points: {'AcidRed871': 0.0, 'LCysteine': 1.0, 'MethyleneBlue': 0.0, 'NaCl': 1.0, 'NaOH': 0.75, 'P10': 3.6, 'PVP': 0.0, 'RhodamineB1': 0.0, 'SDS': 0.0, 'Sodium disilicate': 0.75} {'AcidRed871': 0.0, 'LCysteine': 1.0, 'MethyleneBlue': 0.0, 'NaCl': 1.0, 'NaOH': 0.75, 'P10': 4.0, 'PVP': 0.0, 'RhodamineB1': 0.0, 'SDS': 0.0, 'Sodium disilicate': 0.75}".

We also tested the relative performance of the LLM-only (o3, medium reasoning) and BO optimisations when both were initialised with an identical set of samples (Fig. 7 and S5). The initial samples were selected using the Kennard–Stone algorithm to ensure maximum dissimilarity among the conditions, thereby covering the search space broadly and minimising bias toward any particular region. The same optimisation setup was used in each case; that is, the campaign comprised 15 batches of 10 experiments (total 150), with the first 5 batches (50 experiments) constituting the initialisation stage. Because the methods share identical samples during initialisation, any metric computed before 50 experiments reflects the seed set, rather than algorithm choice. Consistent with this, at 25 experiments (Table 4), both methods necessarily exhibit the same performance.

After the first 50 experiments, the two methods show clear divergence (Fig. 7). The LLM-only approach using o3 achieves an HER of 24.6 μmol h^{-1} (STD = 3.4) averaged over 20 runs and reaches the optimum value in five of these runs. BO attains an average of 12.0 μmol h^{-1} (STD = 4.6) and never reaches the optimum. The superior final performance of the LLM-based optimiser with the o3 LLM can therefore be attributed to its post-initialisation acquisition strategy.

We used a budget of 150 experiments for the tests described so far. Within this budget, BO could not reach the optimum HER (28.37 μmol h^{-1}) for the photocatalyst problem (Table 1). The hybrid LLM/BO optimisers (Table 1) and LLM-only optimisers (Table 3) did reach the optimum HER, but not consistently within 150 experiments. We also explored LLM-only optimisations with larger experimental budgets. Our tests here were limited, but we found optimisation runs that plateaued at sub-optimal HER values within a budget of 150 experiments could nonetheless reach the maximum HER within 500 experiments. For example, the two runs illustrated in Fig. S6 and S7 both plateaued at HER < 25 μmol h^{-1} after around 100 experiments, but subsequent LLM hypotheses broke out of these plateaus, eventually to reach the optimum HER. As discussed in the SI (Section 1), our hybrid LLM/BO optimiser, BORA,[71] inherits asymptotic convergence from its BO component. The BO component is absent in the LLM-only optimisations, but other recent studies suggest that certain implementations of LLM optimisers can be characterised by a finite-state Markov chain model, thus also providing a theoretical guarantee on the convergence to the global optimum.[81] It is unclear whether this applies to the LLMs that we explored here.

3.2.2. Effect of batch size on LLM-only reasoning (o3 model). Batch size also had a pronounced effect on the optimisation performance, both for the o3 LLM optimiser and for BO (Fig. 8, and Table 5). In this case, we considered only the 10-D photocatalysis problem. For this problem, the LLM search and BO both performed better with small batches (Fig. 8a), but the LLM search outperformed uncategorised BO for all batch sizes tested. The influence of batch size for the photocatalysis problem (Fig. 8a) supports our findings in Section 3.1.1. Overall, the ability of the LLM to reason actively about experimental data and to create associated ontologies seems to be stronger than its foundational chemistry knowledge or its ability to search and prioritise literature (*e.g.*, Fig. S3). As such, it is advantageous to use smaller batches (more frequent reasoning) and a shorter initialisation phase (50, 25 and 15 experiments for batch size = 10, 5, and 3, respectively) with LLM-only searches. This was manifested in the inflection observed after the initialisation

Table 4 Comparison of average performance for LLM-only (o3, med) and BO-only optimisations for the 10-dimensional photocatalysis problem, initialised with the same 50 points (Kennard–Stone algorithm). A total of 20 repeat optimisations was run in each case

Model	Average HER after 25 expt	STD after 25 expt	Average HER after 150 expt	Average STD after 150 expt	# Times reached max
BO	1.1	1.4	12.0	4.6	0
o3	1.1	1.4	**24.6**	3.4	**5**

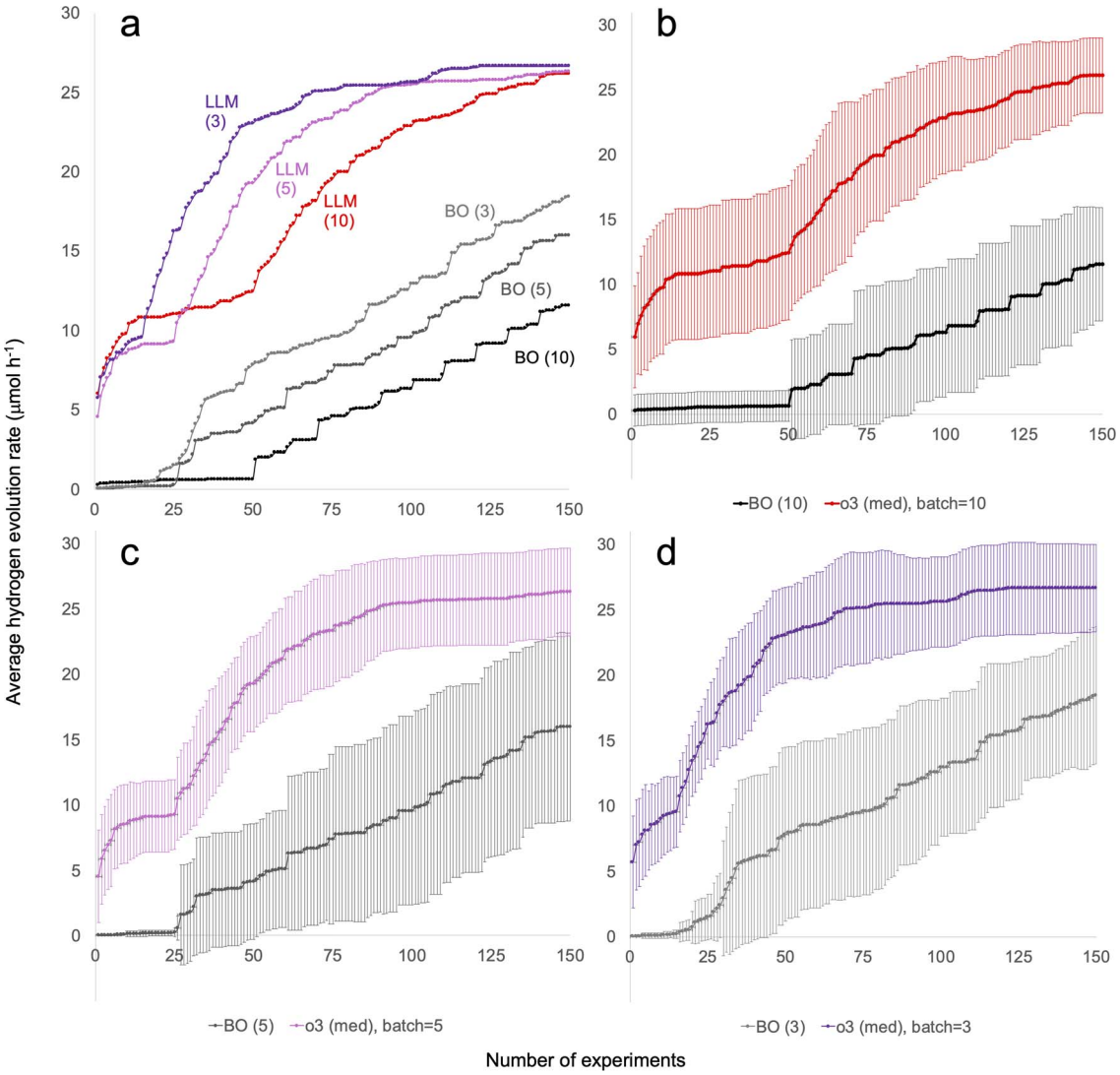


Fig. 8 Effect of batch size on LLM-only reasoning (o3, medium) and BO-only sampling for the 10-dimensional photocatalysis problem. (a) Best HER found, averaged over 20 repeat runs (batch size = 3, 5 and 10; three batches in each case before active learning commenced). (b)–(d) Comparisons of LLM-only optimisations and BO for the different batch sizes showing calculated standard deviations.

phase for each batch size. The early plateaus observed in the initialisation phases for a batch size of 5 and 10 experiments (Fig. 8a) suggest that the LLM might be exhausting its productive hypotheses within a budget of 25 and 50 experiments, respectively, based on foundation model knowledge and literature searches alone. Unlike BO, which is more problem-agnostic, the ideal length of this initialisation phase, prior to adaptive sampling, will depend on the problem, the problem's representation in the foundation model, and the quantity, accessibility, and uniformity of any associated background literature.

3.3. Adding additional data or instructions using LLM/BO hybrid optimisers

3.3.1. Adding human hypotheses. Previously, we developed HypBO, which is a method for incorporating human-in-the-loop hypotheses into Bayesian optimisation,[50] as tested using the same 10-D photocatalysis model explored here. In HypBO, the human hypotheses were added as numerical constraints at the start of the optimisation, as implemented in our Opsight software using graphical sliders. It was not possible in our first implementation of HypBO to add hypotheses during the optimisation—they could only be added at the start—nor was it possible to add complex hypotheses in the form of free text. The use of LLM reasoning now allows us to add hypotheses at any point during the optimisation in the form of free text. In Opsight, this is done by attaching a PDF file or a CSV data file. We tested this here for the 10-D photocatalysis problem, which we understand well, by attaching 'bad', 'mixed', and 'good' hypotheses prior to the optimisation. Here we returned to the hybrid LLM/BO optimiser set-up (BORA) using the o3 model and the same batch settings that we explored in Section 3.1.1.

The bad hypotheses are reproduced in Fig. 9; the mixed and good hypotheses are included in the SI (Fig. S8). Hypotheses 1–5 in Fig. 9 are the reverse of what we knew to be true, while

Table 5 Performance LLM-only and BO-only optimisations after 25 and 150 experiments at different batch sizes (units: μmol h^{-1}). Averages are all for 20 repeat runs. Five initial batches in each case prior to active learning phase

Batch size	Model	Average HER after 25 expt	STD after 25 expt	Average HER after 150 expt	Average STD after 150 expt	# Times reached max
10	o3/BO hybrid	10.5	3.6	25.3	2.9	4
10	o3 only	11.0	5.1	26.2	2.9	9
10	BO only	0.6	1.2	11.6	4.4	0
5	o3 only	9.3	2.6	26.3	3.3	12
5	BO only	0.2	0.3	16.6	5.5	0
3	**o3 only**	**16.2**	5.0	**26.6**	2.5	**13**
3	BO only	3.3	4.8	18.4	4.7	0

hypothesis 6 is neutral. The good hypotheses were essentially the opposite of these (Fig. S8), while the mixed batch contained equal numbers of good and bad hypotheses. In each case, we used the hybrid o3/BO optimiser. Fig. 1c shows the comparable unprompted optimisations without these additional hypotheses.

The average performance over 20 runs is summarised in Fig. 10d and Table 6. Both the bad and the mixed hypotheses reduced the performance of the optimisation with respect to the unprompted (no hypotheses) o3/BO case. This shows that the hybrid optimiser can adopt hypotheses added in a free text form, albeit here poor ones. By contrast, the good hypotheses lead to an increase in average performance over the unprompted o3/BO hybrid, particularly in the early stages of the optimisation (Fig. 10d). The good hypotheses led to higher HER values, on average, but they did not cause the optimiser to locate the maximum value (28.37 μmol h^{-1}) more frequently. Indeed, the success rate for that task was halved (Table 6). Upon close examination, we noticed a detail in the so-called "good" prompt that might explain this anomaly. Specifically, the third "good" hypothesis was that "The amount of P10 should be high (>4 mg)" (Fig. S8). In fact, the maximum HER was observed when P10 is equal to 4 mg. We therefore believe that the appended prompt was directing the optimiser into a good but not optimal range for the P10 variable, supported by the observation that the LLM tended to suggest points where P10 > 4 mg throughout the 150 experiments (*e.g.*, Fig. S9c). In some cases, the BO component could break the optimiser out of this regime (Fig. S10–S12) but often this did not occur within the experimental budget, leading to high but not maximal HER values. We therefore believe that this bias in the P10 variable—also present in the mixed hypotheses (Fig. S8)—accounts for the apparent paradox that the runs with the bad hypotheses located the maximum HER more often than the good or mixed ones (Table 6). The learning here is that a sloppily crafted prompt might be worse than no prompt at all, and this in turn suggests that explicit prompts should perhaps favour general heuristics over specific numerical prescriptions.

If we examine all 20 runs (Fig. 10a–c), rather than just the averages (Fig. 10d), we again identify some outliers. In the bad hypotheses case, run (vi) (Fig. 10a) managed to find a relatively high HER value early in the search and went on to locate a final HER value of 27.94 μmol h^{-1}, close to the maximum possible value. Examination of the optimisation log (PC_20_run_10_5_15_o3_prompt_1_Run_4.pdf) showed that the LLM did initially emphasise dye addition in batches 1–5, as suggested by 'bad' hypothesis 1, but after batch 6, the LLM made the following comment:

"Analysis of the first 60 experiments shows that dyes and large additive volumes suppress HER, while a moderately basic, lightly salted medium with a small amount of L-cysteine produced a dramatic jump to 6.47 μmol h^{-1} (P10 2.2 mg, NaOH 0.5 mL, NaCl 0.25 mL, sodium-disilicate 0.5 mL, L-cys 0.25 mL)".

Team Ideas 1

Our research team has discussed this problem just now and we've got the following thoughts.

1. Addition of dyes will increase hydrogen production
2. Addition of surfactants will increase hydrogen production
3. The amount of P10 added should be low (< 1 mg)
4. The amount of cysteine added should be low to zero
5. Bases (NaOH or sodium disilicate) will be bad for hydrogen production
6. We don't know what to think about NaCl.

Our confidence in these hypotheses is medium.

Existing constraints such as keeping the total reaction volume equal to 5 mL must still be obeyed.

Fig. 9 Bad hypotheses for the 10-D photocatalysis optimisation. The good and mixed hypotheses are included in the SI (Fig. S8).

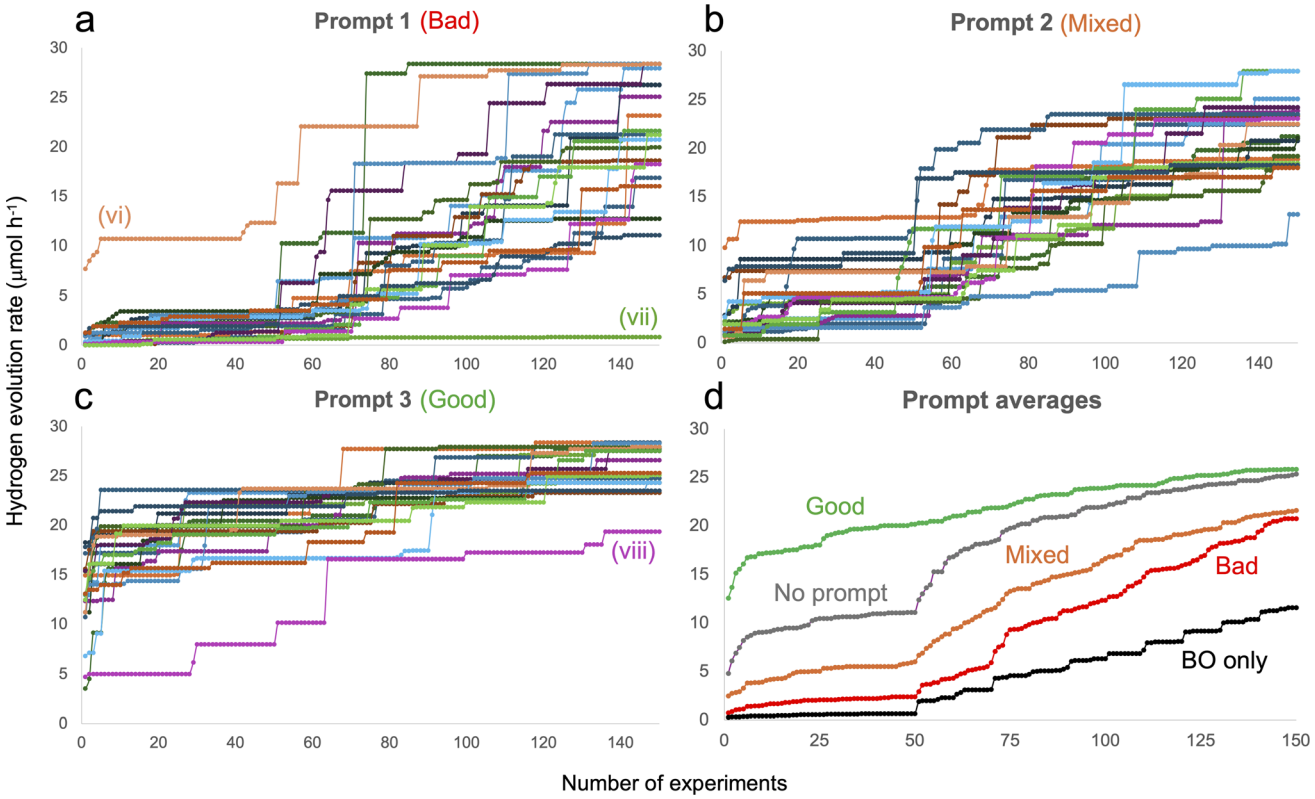


Fig. 10 (a)–(c) Effect of bad, mixed, and good hypothesis prompts on o3/BO hybrid optimisations (20 runs in each case; batch size = 10; 5 initialisation batches prior to adaptive sampling). (d) Average performance over 20 runs compared with o3/BO case (all 20 runs for the "no prompt" case are shown in Fig. 1c).

Effectively, the LLM has used experimental data to quickly reason its way out of the initial bad suggestion of dye addition. At the end of the optimisation, the LLM summarised its "Major Parameter Adjustments" showing how it had moved away from all the initial bad hypotheses based on experimental data:

"Major parameter adjustments

1. P10 progressively increased from 1.4 mg → 4.0 mg; >4 mg showed diminishing returns.

2. Dyes and surfactants virtually eliminated after iteration 3 owing to consistent performance collapse.

3. Base strategy shifted from NaOH-only to a 1 : 1 NaOH : sodium-disilicate buffer (\~0.75 mL each).

4. L-Cysteine incrementally raised from 0.25 mL to 1.0 mL, maximising sacrificial donor efficiency without optical shading.

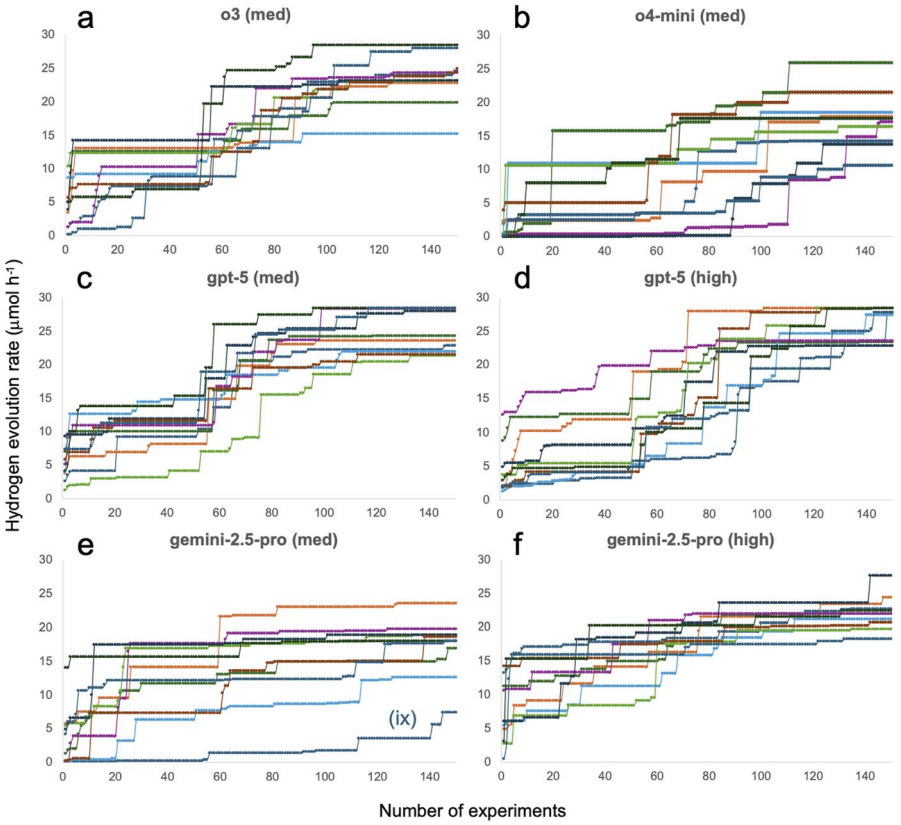
5. NaCl dialled in around 1.25 mL; both lower and higher values reduced HER".

Most other runs also reasoned away from the original bad hypotheses to reach high HER values (Fig. 10a). Indeed, 4/20 runs reached the optimum value of 28.37 μmol h^{-1}. However, apart from run (vi), the injection of bad hypotheses impaired the search over the first 60 or so experiments (*cf.*, Fig. 1c and 10a) and on average, the optimisation did not fully recover within the budget of 150 experiments (Fig. 10d).

The final average performance was also reduced by run (vii), where the log (PC_20_run_10_5_15_o3_prompt_1_Run_6.pdf) showed that the optimiser did not abandon the initial "add dyes" hypothesis, as illustrated in Fig. S13, nor did intermittent BO sampling manage to break the optimiser out of this regime. This was further illustrated by the fifth of six recommendations for future experiments made at the end of run (vii) by the LLM, which remained wedded obstinately to the concept of dye addition:

Table 6 Comparison of performance for o3/BO hybrid optimisations (BORA) with no prompt, bad prompt (Fig. 9), mixed prompt (Fig. S8), and good prompt (Fig. S8). Batch size = 10; 15 batches

Model	Average HER after 25 expt	STD after 25 expt	Average HER after 150 expt	Average STD after 150 expt	# Times reached max
o3 (no prompt)	10.5	3.6	25.3	2.9	**4**
o3 (Prompt 1, bad)	2.0	2.3	20.8	7.0	**4**
o3 (Prompt 2, mixed)	5.0	3.2	21.6	3.6	0
o3 (Prompt 3, good)	**18.0**	1.2	**25.8**	4.4	2



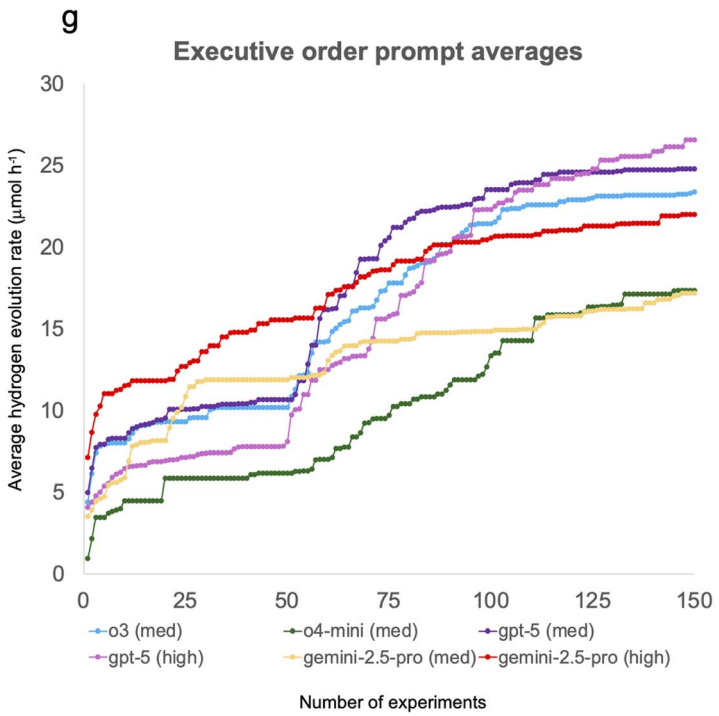


Fig. 11 (a)–(f) Effect of a beneficial executive order, "do not add dyes", on o3/LLM hybrid optimisations (10 runs in each case; batch size = 10; 15 batches). (g) Average performance over 10 runs compared the executive order. Note that the o4-mini and o3 models failed to follow the order in most cases, but the o3 model performed well despite this.

Table 7 Comparison of performance for o3/LLM hybrid optimisations (BORA) with the executive order, "do not add dyes" (10 repeat runs in each case). Order followed = number of runs that follow the order to the end of the optimisation in the LLM-derived batches of points

Model	Reasoning level	Average HER after 25 expt	STD after 25 expt	Average HER after 150 expt	Average STD after 150 expt	Order followed	# Times reached max
gemini-2.5-pro	Medium	10.9	6.0	17.2	4.4	9/10	0/10
o4-mini	Medium	5.9	5.2	17.4	4.2	0/10	0/10
o3	Medium	9.3	4.0	23.4	3.8	2/10	1/10
gemini-2.5-pro	High	**12.7**	3.5	22.0	2.6	10/10	0/10
gpt-5	Medium	10.1	3.1	24.8	3.1	10/10	2/10
gpt-5	**High**	7.1	4.6	**26.6**	2.4	**10/10**	**4/10**

"5. Sensitizer Stability Study. Quantify acid red 871 degradation during runs; if significant, test more robust organic dyes (*e.g.*, Eosin Y) at 0.75–1.0 mL".

Run (vii) is an outlier: overall the data demonstrate that the hybrid o3/BO optimiser recovered, at least partially, from the bad hypotheses in 19/20 of the test runs by applying reasoning to the experimental data (Fig. 10a). The final average performance is still significantly better than uncategorised BO (Fig. 10d), despite us suggesting the worst starting points.

There is also an outlier in the good hypotheses test in run (viii) (Fig. 10c; PC_20_run_10_5_15_o3_prompt_3_Run_17). Investigation of the optimisation log showed that this was a software bug; the Opsight software failed to retrieve the list of "good" hypotheses, and hence the optimisation proceeded without them, which explains the lower performance of that run.

As demonstrated in Fig. S14–S16, it is also possible to add hypotheses during the optimisation, rather than at the start. The example given is a somewhat artificial one, where we switched bad hypotheses for good ones (Fig. S14) during the optimisation (Fig. S15) and these new ideas are then assimilated by the LLM (Fig. S16). The real-life use case for this would be to inject human hypotheses that are prompted by the evolving experimental data: for example, to add a chemistry interpretation that was missed by the LLM, or to correct a misunderstanding made by the model.

3.3.2. Adding executive orders: "do not add dyes". In Section 3.3.1, we added hypotheses, phrased as team suggestions, along with a specified confidence level. Here, we tested the ability of the hybrid LLM/BO optimiser to follow direct instructions, or rules, again for the 10-D photocatalysis problem. This applied only to the LLM component of the optimiser; the BO sampling was unconstrained by these instructions. We attached an "executive order" at the start of the optimisation that read as follows:

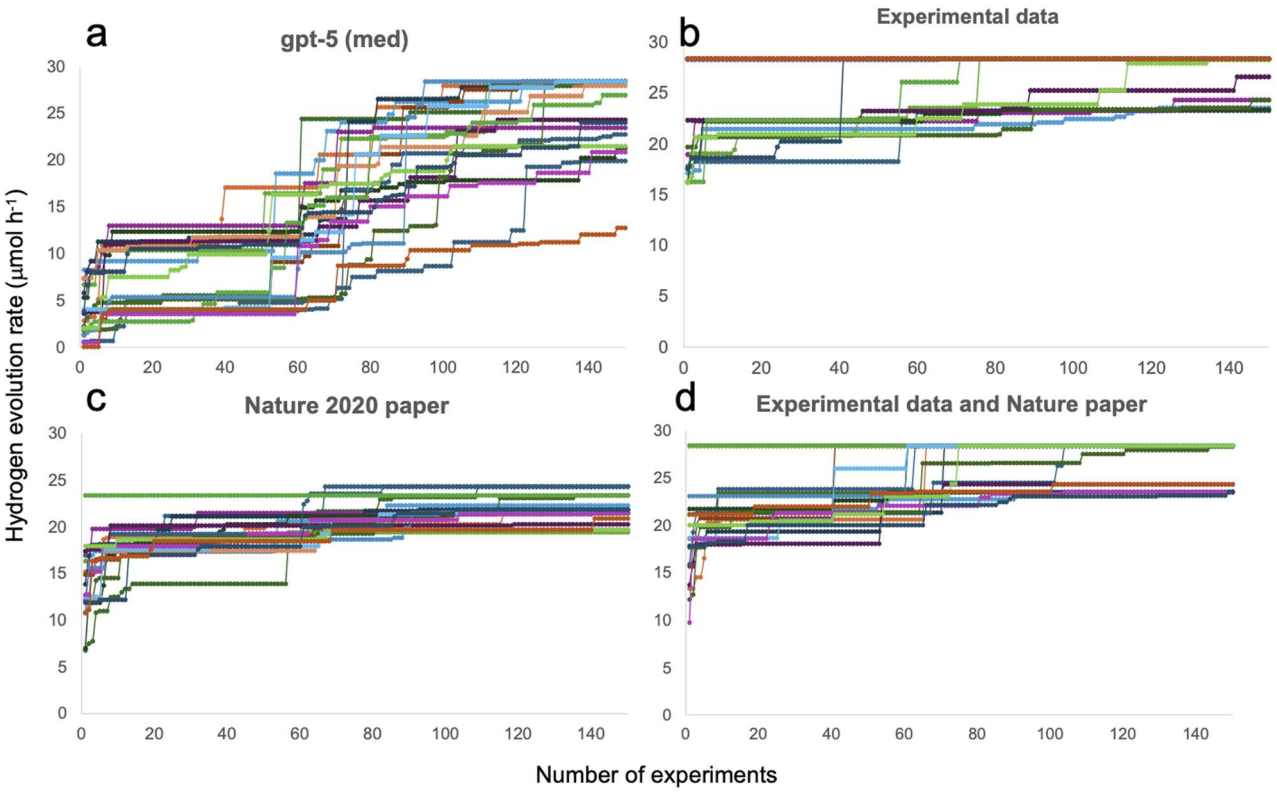
"Do not add any dyes throughout this experiment. These will have a negative effect. This is an 'executive order', overriding any other ideas that might come from literature searches, *etc.* The Bayesian optimisation module is unconstrained so will likely suggest batches that contain dyes as components; irrespective of this, do not suggest any addition of dyes in batches that you suggest, nor pick any BO points where dyes are included. This instruction is for the entire experiment".

The results of this test are shown in Fig. 11. Each optimisation was run 10 times; the average performance for the different models is shown in Fig. 11g. Three of the models tested were found to obey the "no dye" directive throughout the whole optimisation in 10/10 runs (Table 7). This boosted the early optimisation performance: for example, for the gpt-5(med)/BO hybrid (Fig. 11a), the average HER after 25 experiments was 10.1 μmol h^{-1}, *versus* 7.5 μmol h^{-1} (Table 1) for the optimisations without the executive order. By contrast, the gemini-2.5-pro(med) LLM followed the directive in 9/10 cases, while the o4-mini(med) and o3(med) models mostly failed to follow the directive to the end of the optimisation run (Table 7). These tests suggest that the injection of hard constraints or rules into LLM-based optimisations is more effective with more recent reasoning models. For that reason, the gpt-5 model performed best in this hybrid optimisation setting with the "do not add dyes" order (Table 7).

The outlier run (ix) in Fig. 11e can be explained because the LLM failed to generate initial points, and hence it reverted to random points instead (PC_10_run_10_5_15_gemini-2.5-pro-med_executive-order_Run_1.pdf). This run also concluded that sodium disilicate was an inhibitor for hydrogen evolution, one of few cases across all tests in this study where the LLM drew an obviously incorrect conclusion at the end of the 150 experiments.

3.3.3. Adding literature and experimental data. Here we tested the addition of relevant literature and experimental data related to the photocatalysis problem at the start of the optimisation, which is analogous to the injection of hypotheses in Section 3.3.1. We chose the gpt-5 model here because it seemed to follow appended instructions more consistently (Table 7).

Our Opsight software allows the attachment of multiple literature sources but for simplicity of testing and interpretation, we used a single paper here—our 2020 publication[32] on the use of a mobile robotic chemist for this photocatalysis problem. In that study, the best experimental HER found was 21.05 μmol h^{-1}; lower than the 28.37 μmol h^{-1} maximum in the Gaussian Process Regressor (GPR) oracle model used for the tests here. This is because the GPR model was built using additional experimental data collected after the 2020 publication, where higher HER values were found.[71] The appended data source for these tests was the csv file containing the results of the real-life experiments used to build the GPR model in our first BORA publication[71] (1027 laboratory experiments in all). That is, we provided the LLM



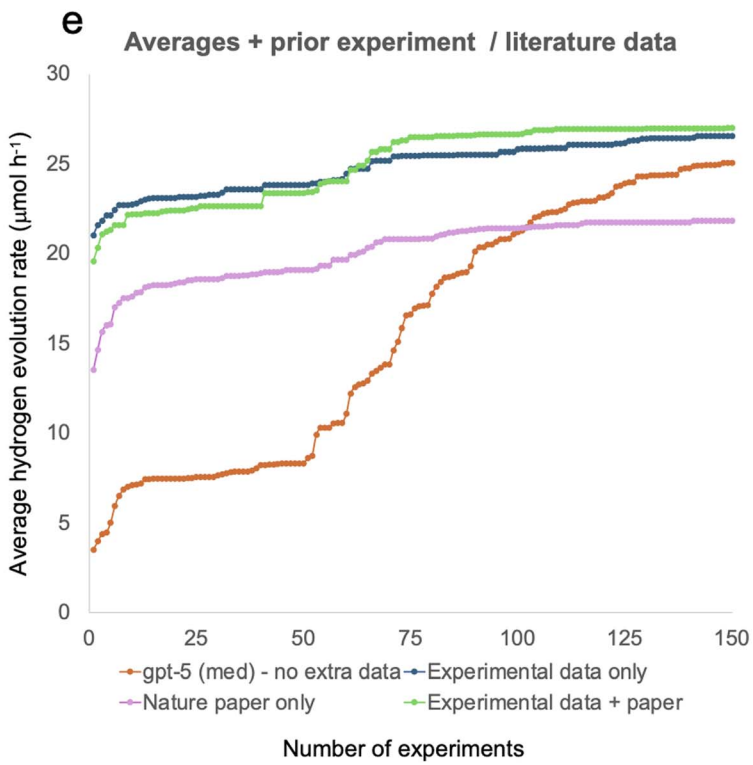


Fig. 12 (a) gpt-5/BO hybrid optimisation (BORA) without any additional prompt. (b) gpt-5/BO hybrid optimisation (BORA) with addition of experimental dataset from ref. 71. (c) gpt-5/BO hybrid optimisation (BORA) with addition of 2020 *Nature* publication (ref. 32). (d) gpt-5/BO hybrid optimisation with addition of both experimental dataset from ref. 71 and the publication (ref. 32). (e) Overlay plot showing average best HER value found for the four scenarios tested (20 runs in each case).

Table 8 Comparison of average performance for gpt-5/BO hybrid optimisations with and without the addition of SI data (20 repeat runs in each case). Batch size = 10; 15 batches

Model	Average HER after 25 expt	STD after 25 expt	Average HER after 150 expt	Average STD after 150 expt	# Times reached max
gpt-5 (med)	7.5	3.5	25.1	4.2	9
+Experimental data	**23.2**	3.3	26.6	2.7	11
+*Nature* 2020	18.5	1.9	21.8	1.5	0
+Data and paper	22.5	3.3	**27.0**	2.2	**13**

with the experimental data used to build the GPR model, whereas the *in silico* experiments in these tests were derived from the GPR oracle model that was fitted to these laboratory data.

As shown in Fig. 12b, the addition of prior experimental data has a marked effect on the optimisation. In total, 11 of the 20 runs reached the optimal HER of 28.37 μmol h^{-1}. Eight of the runs went straight to this maximum value in the first experiment. In other cases, the LLM did not manage to identify the best-performing sample from the csv file and failed to initialise the search with that optimal point. This is a clear case where the LLM fell short of human intelligence; most researchers would repeat the best previous experiment as the first trial. While this paper was under review, OpenAI released the models gpt-5.1 and gpt-5.2. We have not benchmarked these new models for all the tests in this study, but we were intrigued by claims that they are better at file handling. Preliminary testing corroborated this: with the attachment of the same experimental data, a gpt-5.2/BO hybrid reached the optimum HER in 20/20 runs (Fig. S17 and Table S4) compared to the 11/20 success rate observed for gpt-5 (Table 8). Our Opsight software also allows initialisation of runs directly with predefined experimental datapoints (*e.g.*, Fig. 7), rather than providing spreadsheet attachments: in that case, the suggested points will always be explored.

The effect of appending the PDF file of our 2020 paper on photocatalysis discovery using a mobile robotic chemist[32] is shown in Fig. 12c and e. That publication did not explicitly state the best experimental catalyst composition found, nor did it tabulate experimental data, although these were supplied in separate SI, which was not attached in these tests. The supplied paper did, however, discuss the general positive and negative influences of the various classes of components, as well as the specific observation of an apparent synergy between the two bases, NaOH and sodium disilicate.

Adding this paper warm-started the optimisation, and the average best HER value found after 25 experiments was more than double that found without the paper (Table 8). Analysis of the relevant log files showed that overall, the LLM did a good job of summarising the key learning points from the attached paper. This accounts for the pronounced warm start with respect to the unprompted search (*cf.*, Fig. 12a and c). For example, based on learning from the paper, the addition of dyes and surfactants was mostly avoided from the outset. This early performance gain suggests that the fine detail of the knowledge contained in this publication[32] was not embedded strongly in the gpt-5 foundation model. However, despite this warm start, the optimisation performance after 150 experiments was in fact lower than we observed without addition of the paper (Fig. 12e, and Table 8). At first this seems counterintuitive: the earlier discovery of more effective catalyst compositions (Fig. 12c) should allow the LLM to carry out finer levels of reasoning in later batches, as well as providing better seeds for subsequent BO sampling in the hybrid optimiser. It is impossible to account unambiguously for this anomaly because of the black-box nature of LLMs. Also, even though Opsight is scripted to explain its reasoning, the details are buried within 40 lengthy logs across two sets of 20 optimisation runs. However, after scrutinising these logs and the associated data, we speculate that the difference in performance might lie in the role of salt (NaCl). Notably, our 2020 paper contained the following sentence: "NaCl had a small positive effect, but less so than the four selected components, and it was therefore deselected".[32] The paper stated further that "Increased ionic strength is beneficial for hydrogen production (NaCl addition), but not as beneficial as increasing the pH (NaOH/sodium disilicate addition), which also increases the ionic strength". The best catalyst composition found in that study (HER = 21.05 μmol h^{-1}) contained no NaCl.[32] The statements quoted from this paper convey that the elimination of NaCl will increase the HER, and this idea was understood consistently by the LLM prior to the optimisation, for example:

PC_20_Run_10_5_15_gpt-5-med-Nature_paper_Run_1.pdf: "Slightly positive: NaCl (ionic strength) early, then deprioritized relative to base effects".

PC_20_Run_10_5_15_gpt-5-med-Nature_paper_Run_2.pdf: "NaCl: small positive effect (ionic strength) but inferior to bases; ultimately deselected [Nature 2020]"

PC_20_Run_10_5_15_gpt-5-med-Nature_paper_Run_3.pdf: "NaCl had a small positive effect but was inferior to the base additives and was deselected [Nature 2020]".

PC_20_Run_10_5_15_gpt-5-med-Nature_paper_Run_4.pdf: "NaCl: small positive *via* ionic strength, but far less than pH elevation; ultimately deprioritized [Nature 2020]".

By contrast, in the combined experimental dataset used to build the GPR oracle model,[71] which post-dated our 2020 publication,[32] six out of the ten most active catalyst formulations contained finite levels of NaCl. This included the most active catalyst formulation in that combined dataset (NaCl = 1 mL). We therefore suggest that attaching the 2020 publication led the hybrid optimiser away from NaCl addition, which we only later discovered to be beneficial. This is substantiated by the data: for the unprompted gpt-5/BO hybrid optimisation (Fig. 12a), the average NaCl addition across the 20 best-

performing catalyst compositions found was 0.96 mL. For the publication-prompted experiment (Fig. 12c), the average NaCl addition was 0.34 mL (NaCl = zero in 6/20 cases). The plot in Fig. S18 illustrates that the NaCl addition for the best-performing experiment across the 20 prompted runs was lower than in the unprompted runs in 18/20 cases. We did not inform the LLM about the differences between the datasets in our earlier paper[32] and the larger combined dataset that was used to build the underpinning GPR model.[71] As such, the optimiser did not necessarily behave wrongly in de-emphasizing NaCl addition, based on the information that we provided, because that was indeed the conclusion of the appended 2020 publication.[32] This mirrors the "good" team hypotheses (Section 3.3.1 and Fig. S9–S12) where a broadly helpful set of starting hypotheses provided an initial warm start, but then a persistent misleading hypothesis ("P10 > 4 mg", "do not add NaCl") compromised performance over unprompted active learning in the longer term. These observations suggest that LLMs might be insufficiently explorative and too trusting of prior literature, and that this is not fully compensated by using BO hybrids, as here, within the experimental budget that we explored.

Fig. 12d and e illustrates the effect of adding both the experimental data (csv file) and the 2020 publication (PDF file) together. This led to the highest average HER after 150 experiments (27.0 μmol h^{-1}; Table 8). The optimal HER (28.37 μmol h^{-1}) was reached in 13 of the 20 test runs. Evidently, the addition of this experimental dataset, which superseded the 2020 paper, cancelled out any "no salt" bias that was introduced by appending the publication alone. The average NaCl addition across the 20 runs shown in Fig. 12d was 0.98 mL, which is similar to the average obtained when we attached only the experimental data (0.96 mL), and close to the optimal value in the appended experimental dataset (NaCl = 1 mL in the best experiment). Again, the newer model, gpt-5.2, performed better still with such file handling tasks (Fig. S17, and Table S4).

3.4. Maximising performance of the photocatalysis optimisation

Here, we took the learning from Sections 3.1–3.3 to maximise the optimisation performance for the 10-D photocatalysis experiment. Specifically, we chose the o3 LLM, which was most effective in both hybrid optimisations (Section 3.1.1; Table 1) and in LLM-only optimisations (Section 2; Table 3) in the absence of any additional data or human hypotheses. We also chose the smallest possible batch size (one experiment), extrapolating from the trials in Section 3.2 (Fig. 7, and Table 5). The results are shown in Fig. 13 (and Fig. S19) and summarised in Table 9. We used an initialisation phase of 30 experiments, prior to adaptive sampling; that is, three times the problem dimensionality.

The unprompted LLM-only optimisation under these settings outperformed, ultimately, even the most prompted hybrid optimisation tests with a batch size of 10 where we added both experimental data and the publication that is most directly relevant to this problem[32] (*cf.*, Tables 8 and 9). The average best

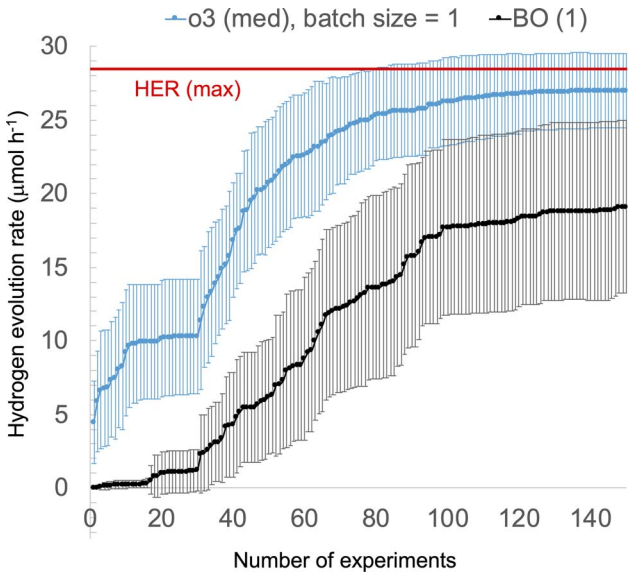


Fig. 13 LLM-only (o3, med) and BO-only optimisation for the 10-D photocatalysis problem with a batch size of one. Initialisation phase = first 30 experiments. HER (max) = optimal HER (28.37 μmol h^{-1}). Error bars are the standard deviation across 20 repeat runs. All 20 runs for the LLM-only optimisation are plotted in Fig. S22a.

HER value after 150 experiments was 27.0 μmol h^{-1}, close to the optimum value. The optimiser located the optimum HER (28.37 μmol h^{-1}) in 15 out of 20 runs. This performance was also superior to that observed for o3/BO and o3/SAAS BO hybrid optimisations with a batch size of one (Table 8, and Fig. S19).

When we appended prior experimental data, as in Section 3.3.3, then the LLM-only optimiser was able to find the optimum conditions in 100% of 20 trial runs (Table S5, Fig. S20, and S21) to give an average best HER value of 28.37 μmol h^{-1} (STD = 0) after 150 experiments. However, when using o3, the LLM went straight to the optimal HER in the first experiment in just 3/20 runs (Fig. S20), whereas a researcher would most likely replicate the best known reaction conditions in the first experiment, as a baseline, 100% of the time. As discussed above, the gpt-5.2 model, released while this paper was under review, is superior for this file handling task, going directly to the optimum HER in experiment one in 18/20 trials using a hybrid gpt-5.2/BO optimiser and a batch size of 10 (Fig. S18).

Fig. 14 shows the optimisation performance for an example run selected from the 20 LLM-only runs with a batch size of one. This example was not the most rapidly converging among the 20 runs; others reached the optimum HER in fewer experiments (*e.g.*, Run 1, Fig. S22, reached the optimum in 43 experiments). We chose this example (Run 18 in Fig. S22) to illustrate a reasoning chain that extended over the entire budget of 150 experiments. In total, the LLM made 149 separate hypotheses over the course of the optimisation. In all but one case (experiments 101 & 102), there was a single hypothesis per experimental point suggested. The details of all hypotheses can be found in the optimisation log file in the data repository (o3-med-(30 + 120)*1_28 october 2025, 10_33 am-2_Run_18.pdf). The hypotheses and their rationales are also provided as

Table 9 Comparison of average performance for hybrid (BORA) and LLM-only optimisations with batch size = 1 (20 repeat runs in each case; 30 initial experiments before active learning phase)

Model	Average HER after 25 expt	STD after 25 expt	Average HER after 150 expt	Average STD after 150 expt	# Times reached max
BO-only	1.1	1.4	19.1	5.9	0
SAAS BO-only	0.4	0.4	9.8	6.7	0
o3(med)/BO hybrid	**10.4**	4.8	24.9	2.7	2
o3/SAAS BO hybrid	**10.4**	3.2	24.0	2.4	3
o3 (med)-only	10.3	3.9	**27.0**	2.5	**15**

a spreadsheet (Log_file_run_18_o3-med-batch_1.xlsx; also Fig. S23). Fig. 13b summarises the 22 hypotheses that advanced the HER directly. The LLM summaries of the optimisation progress are provided in Fig. S24–S27.

As in all LLM-only and LLM-hybrid optimisations (other than the tool-disabled runs in Section 3.1.1, Fig. S3), the optimisation commenced with a literature search. In this example, the LLM located our 2020 publication on this chemistry[32] *via* a Google search that accessed SCISPACE,[82,83] thus retrieving this paper, which is behind a paywall, from an open-access university repository. Perhaps influenced by this publication, the LLM made little exploration of the unprofitable dye sub-space throughout the optimisation (Fig. S28), thus largely discounting three of the ten experimental variables from the outset.

After literature searching, the LLM proceeded to produce and test hypotheses. The rationale for the first hypothesis, "Baseline" (experiment 1) was as follows:

"Literature-inspired starting point: balanced donor, base, dispersant".

The next two hypotheses explored higher NaOH concentration (experiment 2) and higher L-cysteine concentration (experiment 3), both of which improved the HER, as did addition of NaCl (experiment 8), which was absent in experiment 1. In experiment 9, the following hypothesis was explored:

"PVP stabiliser: test non-ionic steric stabiliser instead of SDS".

This improved the HER further, not in fact because PVP improves the HER *per se*, but because it is less detrimental than SDS, which had been kept constant (0.25 mL) in experiments 1–8. The LLM then explored combining both SDS and PVP ("SDS + PVP: synergistic dual surfactant system"; confidence = low), which decreased the HER, as did the hypotheses of "High SDS" (experiment 18) and "very high SDS". Presumably inspired by these negative observations, the next hypothesis that improved the HER was "No SDS" (experiment 20). After experiment 20, SDS was largely ignored and was only returned to four times in the next 130 experiments (Fig. S29a). For example. SDS was reintroduced (unsuccessfully) in experiment 69 ("Optimum + SDS 0.25"), based on the following rationale:

"Small SDS may improve catalyst wetting and photogenerated H_2 bubble release without excessive site blocking. Added volume keeps ionic strength similar".

The next hypothesis to improve the HER was "No SDS + High NaCl" (experiment 31). The rationale was "Combine best baseline with 1 mL NaCl to test additive ionic-strength boost without surfactant". Up until experiment 31, NaCl had been largely ignored (Fig. S29b). As noted in Section 3.3.3 (and Fig. S18), this might have been influenced by our 2020 publication,[32] which

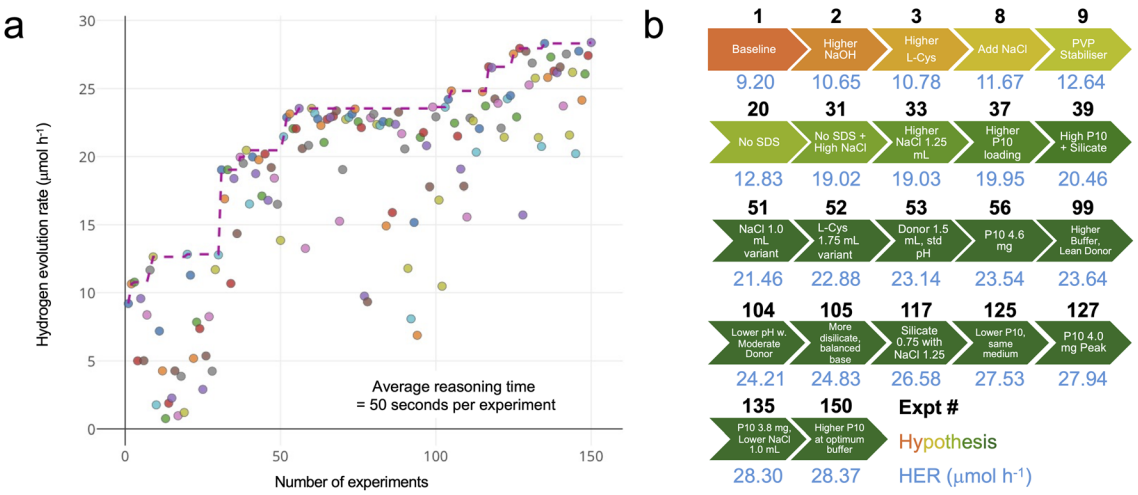


Fig. 14 (a) Example of an LLM-only optimisation for the 10-dimensional photocatalysis problem (o3 model, medium reasoning level, batch size = 1). There were 30 experiments before the active learning phase. The points are colour-coded by LLM hypothesis (see Fig. S23 for full list of the 149 hypotheses tested; also Log_file_run_18_o3-med-batch_1.xlsx in the data repository). (b) Hypotheses that increase the HER; experiment number in black text, HER value in blue text.

was identified at the start of the search. We note that using a batch size of one (*cf.*, batch = 10 in Section 3.3.3) provides more opportunities—that is, more hypotheses—for the LLM to reason away from such initial biases.

Even higher NaCl concentration improved the HER marginally (experiment 33). A larger increase was made by the hypothesis "Higher P10 loading" (experiment 37), and after experiment 37, higher P10 values (>4 mg) were explored for the next 88 experiments (Fig. S29c). The next increase was the hypothesis "High P10 + Silicate" (experiment 39); until that point, sodium disilicate had not been prioritised, but after experiment 39, it was included in all but seven of the next 111 experiments (the points on the *x*-axis in Fig. S29d).

At experiment 51 (Fig. S29b and e), the LLM decided to "Test if slightly lower ionic strength synergises with silicate buffer at high P10" (NaCl 1.0 mL variant), which led to an increase in HER, as did the decision (experiment 52, Fig. S29e) to "Probe finer donor optimum between 1.5 and 2.0 mL" (L-Cys 1.75 mL variant).

The rationale to reduce the L-cysteine concentration (experiment 53) even further, from 1.75 mL to 1.5 mL (Fig. S28e) was that:

"HER peaked when L-cysteine was reduced; 1.5 mL may sit at the true maximum of the inverted-U while maintaining optimum pH".

Thus, experiment 53 appeared to learn from experiment 52.

Experiment 56 was informed by the hypothesis that "A slight decrease in P10 may alleviate light-scattering and aggregation seen at 4.8–5 mg, potentially boosting effective active area". Whether or not those two chemically plausible hypotheses were the precise cause, it did increase the HER from 23.14 to 25.54 μmol h^{-1}.

There then followed a series of experiments (57–98) where the LLM hypotheses did not improve the HER, resulting in the plateau shown in Fig. 14a. The full hypothesis list is given in the data repository (Log_file_run_18_o3-med-batch_1.csv): in précis, the LLM mostly explored the variables P10 (fine tuning, Fig. S28c), NaCl (larger changes, centred around 1.25 mL, Fig. S28b), L-cysteine (mostly reductions from 1.5 mL, Fig. S29e), and to a lesser extent, sodium disilicate (Fig. S29d) and NaOH (Fig. S29f). There were also a small number of unsuccessful attempts to reintroduce variables that had been discounted previously, such as the hypotheses "Optimum + AcidRed87" (experiment 77) and "More Base + PVP Dispersion" (experiment 84).

The next HER increases were achieved by lowering the L-cysteine donor further (experiments 99 and 104) and then replacing some of the L-cysteine with sodium disilicate (experiment 105; Fig. S29e), based on the hypothesis that:

"Replacing some L-cysteine with sodium disilicate increases alkalinity and silicate surface interactions that may enhance proton reduction while keeping scavenger present".

Further fine-tuning of the disilicate/salt balance (experiment 117) and the P10 amount (experiments 125 & 127, again with the hypothesis of reducing light scattering) improved the HER further to 27.94 μmol h^{-1}, close to the optimal value. This experiment prompted the following summary (time stamp 12 : 16 : 23 in the log file):

"We now have a clear "sweet-spot" formulation: P10 ≈ 4 mg, NaCl 1.25 mL, NaOH 0.75 mL, L-Cys 1.0 mL, Na-disilicate 0.75 mL and no other additives. Over the last 10 iterations this recipe out-performed all others and delivered the current best HER of 27.94 μmol h^{-1}.

Analysis of 128 datapoints shows:

• HER rises steeply when P10 is increased from 3 → 4 mg but falls again beyond 4.2 mg.

• 0.75 mL Na-disilicate offers the best buffer action; 0.5 or 1.0 mL are inferior.

• HER maximises at NaOH 0.75 mL; 0.5 mL is often too acidic and 1.0 mL or higher suppresses activity.

• Ionic strength: 1.25 mL NaCl is optimal; 1.0 mL is slightly worse and 1.5 mL drops HER significantly unless P10 is lowered.

• Donor amount: 1.0 mL L-Cys consistently outperforms 0.75 or 1.25 mL in this window.

Two avenues remain unexplored: (i) fine P10 tuning below 4.0 mg, and (ii) interaction of the 4 mg catalyst with a marginally higher ionic strength (NaCl 1.5 mL). These could uncover an incremental gain without breaching the 5 mL volume cap".

After 132 batches, the LLM summary was revised to the following (time stamp 12 : 20 : 57):

"Across 132 batches, HER climbed from ~9 to 27.9 μmol h^{-1}. All top-performing formulations share: (i) SDS/PVP/dyes = 0, (ii) NaCl ≈ 1.25 mL, (iii) LCys ≈ 1 mL, (iv) NaOH ≈ 0.75 mL, (v) sodium-disilicate 0.75 mL and P10 between 3.8–4.2 mg. Simple correlations confirm this: NaCl, P10 and silicate are positively correlated with HER, while SDS, dyes and excess donor/base depress activity. Trends suggest the system benefits from moderate ionic strength and buffer capacity but is sensitive to excessive alkalinity. We have never tested the present optimum with a slightly milder pH (NaOH 0.50 mL) – correlation data suggest HER might increase as NaOH is reduced. Likewise, raising silicate from 0.75 → 1.00 mL (keeping total base constant by lowering NaOH) could enhance buffering and surface adsorption without exceeding the 5 mL volume. Both new points stay in the sweet spot of NaCl = 1.25 mL and LCys = 1 mL and may nudge HER past 28 μmol h^{-1}.

Key insights saved: (1) avoid surfactants/dyes; (2) optimal NaCl ≈ 1.25 mL; (3) HER peaks at P10 ≈ 4 mg; (4) lowering NaOH while keeping silicate 0.75–1.0 mL is unexplored and promising".

At experiment 135, the LLM hypothesised the following:

"P10 3.8 mg, Lower NaCl 1.0 mL: data show 3.8 mg P10 with NaCl 1.25 mL → 27.72 μmol h^{-1}; lowering NaCl to 1.0 mL gave 26.53 μmol h^{-1} at 4.4 mg P10. Expect synergistic boost at 3.8 mg by reducing ionic strength slightly".

A small reduction in NaCl (by 0.25 mL; Fig. 14b) and reduction of P10 from 4.0 mg to 3.8 mg found a new maximum HER of 28.3 μmol h^{-1}. The final increase to the optimal value, 28.37 μmol h^{-1} (experiment 150), resulted from increasing the P10 amount back to 4.0 mg at the new "champion salt/base/buffer conditions".

At the end of the optimisation, the LLM also provided cogent although not overly surprising suggestions for future experiments (Fig. S30).

4. Ethics and responsible research using LLMs

The use of LLMs in R&D raises some questions of ethics and responsible research. For example, a recent study discussed the ethics of using LLMs in the chemical sciences,[84] citing risks that included "misinformation harms" (*i.e.*, LLMs outputting erroneous or misleading information), "environmental and socio-economic harms" (*e.g.*, energy and water consumption by LLMs), and "human–computer interaction harms" (*e.g.*, anthropomorphising LLMs, leading to overreliance or unsafe use). In some scenarios, LLMs have been found to exhibit "agentic misalignment" to achieve their goals.[85] They can also fail to distinguish beliefs from facts.[86] We note that scientific research involves a mixture of both facts and beliefs—that is, theories and hypotheses—but it is, of course, desirable to differentiate between these. There is also evidence that frontier models are capable of "in-context scheming",[87] and pursuing misaligned goals, although the specific LLMs evaluated in that study were different to those investigated here. Another risk is access inequity: we have focused on proprietary models where the larger models were found to perform better, but these are also the most expensive commercial tools. Our Opsight software is not tied to a specific LLM, and it could be configured to use free, open-source models, but these may not be the best-performing options available.

To some extent, these risks prompted the detailed *in silico* testing strategy here. It is expensive to carry out statistically meaningful numbers of repeat optimisations for wet-lab experiments, and it is improbable, for example, that outlier (ii) (Fig. 1f) would have been found in a single experimental optimisation run. There is a strong case, therefore, for robust *in silico* testing of new LLM-based methods before deploying them in the real world.

Notwithstanding these genuine risks, there are also counterarguments for the potential societal benefits of LLM reasoning in scientific research. The societal pay-off from material discoveries enabled by LLMs might, in the long-term, offset negative aspects such as energy usage. Moreover, if LLMs can guide us to good solutions in fewer experiments, then we might use less energy and fewer chemicals, which could yield both safety and environmental benefits.

A comprehensive ethics discussion is beyond the scope of this study, but we did attempt a high-level analysis of the environmental impact of using LLM queries (Table 10). It is challenging to make precise estimates here, so these values should be taken as a rough guide only. A recent analysis concluded that a "pessimistic" estimate of the energy cost of a long gpt-4o query was 40.4 wh (Table 10). Our own estimates for the LLM queries in this study using the gpt-5 model with LLM reasoning only (Section 3.2) were much higher than this: around 1045 wh per run, on average. Our analysis involved five repeat photocatalysis optimisations (150 experiments each) and considered a batch size of 10 to reflect a batched experiment. A smaller batch size one would lead to more LLM queries and a higher energy consumption. We considered both input tokens and output tokens, as well as tool usage. The average number of input tokens, including agentic output (web searches, python scripts) was 1.7×10^6. The average number of output tokens (reasoning and chat output) was 2.54×10^5. The average time to complete 150 virtual experiments was 4508 seconds.

Table 10 Estimated energy usage of LLM queries and common laboratory and domestic equipment

Device/Task	Approx. energy cost/day (wh)
Simple Google search (x1)	0.3
LLM query (gpt-4o)[a,c]	40.5
LLM query (gpt-5) (our estimate)[b,c]	1045
Heaters/stirrers[d]	666
NMR instrument[d]	15 418
Single laboratory fume hood[d]	26 904
−80 °C freezer[d]	28 800
Domestic microwave oven[e]	250
Electric car[f]	13 400

[a] Estimate taken from https://epoch.ai/gradient-updates/how-much-energy-does-chatgpt-use, assuming "pessimistic" estimate for a single gpt-4o query and maximum query length. [b] Our own estimates based on averages over 5 real LLM-only queries in this study (photocatalysis problem) using the 'worst case' gpt-5 model (batch size = 10). [c] For the LLM estimates, each optimisation is associated with a single (complex) query; as such, the daily energy cost will depend on the length of the experiment. These energy estimates are for a single query; for an *N*-day experiment, the numbers should be divided by *N* to give the daily estimate. [d] Estimates taken from https://mygreenlab.org/wp-content/uploads/2025/07/ie_-_energy_consumption_of_univeristy_laboratories_-_s-labs.pdf (published in 2011). Energy use was given per annum; the numbers in this table are normalised per day. [e] 1000 W oven, 15 min usage per day. [f] Taken from https://www.energysage.com/electricity/house-watts/how-many-watts-does-an-electric-car-charger-use/.

This LLM energy usage is substantial: to put it in context, we estimate that the energy used by one of these complex queries is on the same order as running a laboratory heater/stirrer plate for one day or using a domestic microwave for one hour (Table 10). However, this is low compared to some other common laboratory equipment. For example, a standard laboratory fume hood consumes about 25 times as much energy every day, whether it is being used or not.

The benchmarking exercises in this study have artificially high energy costs because we ran 10–20 repeat optimisations for virtualised experiments that are effectively instantaneous. In a more realistic closed-loop laboratory experiment, a single LLM query might support a search that runs over multiple days. For example, in our original mobile robotic chemist study,[32] 688 experiments were carried out over 8 days. If we assume roughly linear scaling of the LLM energy costs from the 150-experiment benchmarks discussed here, again with an experimental batch size of ten, then that would equate to 4.8 kwh of energy usage over 8 days, or 600 wh per day, on average. This would constitute a small fraction of the total daily energy use in that workflow—around 2% of the energy consumed by a single fume hood associated with the experiments, without considering the additional energy used by the mobile robot, analytical instruments, and photolysis lamps, *etc.*

Thinking longer term, much laboratory infrastructure, such as fume hoods, and many analytical instruments are left powered on permanently, irrespective of whether they are being used, either for safety reasons or because of operational stability. Hence, closed-loop experiments with LLM or hybrid LLM/BO guidance could exploit what is otherwise dead time for these laboratory assets, thus increasing energy efficiency. If we can reach good experimental solutions in fewer experiments, as suggested by the examples in Sections 3.1–3.4, then there are further opportunities to save energy and to reduce chemical and solvent usage. We also note that a single researcher commuting to a laboratory by electric car might consume 20 times as much energy daily as our daily LLM energy estimates for the hypothetical 8-day experiment example provided above (Table 10). As such, while the energy consumption of LLM queries cannot be dismissed, it is unclear that more traditional ways of doing research will have a lower overall environmental footprint.

5. Conclusions

In this study we show that reasoning models can be used either as standalone optimisers or as hybrids with BO, building on our previous studies.[71] Used as standalone optimisers, these reasoning models match or even surpass the performance of BO or LLM/BO hybrids for two problems: a 10-D photocatalysis problem and a 7-D physics problem (Fig. 4 and S18).

The key insights from the study are as follows:

Insight 1: while the inner workings of the LLMs used here are opaque, the Opsight software package is configured to explain its reasoning and hypotheses as it proceeds (*e.g.*, Fig. 5, and 14).

Insight 2: it is straightforward to inject additional knowledge or information into optimisations, either as natural language human hypotheses or instructions (3.3.1, 3.3.2), prior experimental data (3.3.3), or scientific publications (3.3.3), or some mixture of these formats. This can be done at the start of the experiments or during the optimisation (Fig. S15); for example, to add a new hypothesis in response to the evolving data.

Insight 3: even when deliberately bad starting hypotheses are added (3.3.1), the LLM can reason away from them in most cases, sometimes even reach the optimal solution within the experimental budget (Fig. 10a, and Table 6), albeit at the cost of more experiments.

Insight 4: LLMs are stochastic, and some optimisation runs are outliers. Like humans, LLMs can occasionally become fixated on unprofitable hypotheses (*e.g.*, run (vii) in Fig. 10a). It is important to carry out a statistically meaningful number of test runs when developing these platforms. We found outliers of just one or two anomalous runs within a batch of 20 repeat optimisations (*e.g.*, Fig. 1e, f, 3b, 10a, c, and 11e).

Insight 5: not all the LLMs tested were able to carry out direct user instructions consistently, although the recent gemini-2.5-pro and gpt-5 models did succeed in this task (Table 7).

Insight 6: several LLMs had difficulties in handling appended datasets: for example, in identifying the best experimental conditions within an appended spreadsheet of previous experimental results (Fig. 12b, S20 and Table 8). Recent models such as gpt-5.2 appear to do much better with such file handling tasks (Fig. S17).

Insight 7: the ability of the LLMs to find and exploit scientific literature also had limitations. The LLMs located papers that were broadly relevant to the problem in hand, but the prioritisation of this information was less strong. Consequently, it was unclear that these literature searches were always making a net positive contribution, at least for the photocatalysis problem studied here (Fig. S3). This is perhaps unsurprising for questions with no clear consensus: for example, there are papers showing that dyes are good sensitisers for photocatalysis and other papers that show the reverse. It is hard for an LLM—or indeed a human researcher—to make more nuanced reasoning without additional physicochemical information or calling up specific calculations.[70]

Insight 8: LLMs and LLM/BO hybrids are agentic frameworks, but they have less agency than a researcher. Unless an explicit experimental budget or termination condition is set by the user, these optimisers will continue to explore the chemical space unprofitably, in a way that a human researcher would not (Fig. S31). Conversely, there may be scenarios where automated machine reasoners, like BO, could locate better solutions in high-dimensional chemical space long after a human researcher might have given up (Fig. S6, and S7).

Insight 9: prompts must be crafted precisely, since even small ambiguities can affect the optimisation performance strongly, as discussed for the "good" hypotheses case in Section 3.3.1. It would be possible to input spoken hypotheses,[88] but the evident sensitivity to small details suggests that carefully crafted written prompts might be more effective.

Insight 10: while the energy used by these LLM optimisers is substantial, and rising as the models become more complex, this energy consumption is still small compared to the energy costs of physical experiments (Section 4).

Insight 11: the most effective approach was to use an LLM-only optimiser (Fig. 13). This has important ramifications because it suggests that future optimisers may not require components such as BO, at least for some problem classes.

Insight 12: iterative decisions with a batch size of one experiment[89,90] was the most effective strategy (Fig. 13), but this is also repetitious and poorly aligned with human working patterns. The introduction of machine reasoning is an attractive strategy to augment more periodic and deeper human insights.

We have focused our tests here on noise-free ground truth models of a 10-D chemistry problem and a 7-D physics problem. This allowed in-depth testing with multiple repeats in a way that would be impossible with real experiments. The extension of this approach to functional materials discovery problems that also include significant experimental noise will be the next step in our research programme.

Author contributions

The Opsight software was written by A. C. M. E. C. and A. I. C. carried out the optimisations, apart from those in Fig. 7/S5, which were performed by M. Z. The experiments were interpreted by M. E. C. and A. I. C. The asymptotic convergence proof

was written by X. E, A. C. and M. Z. The figures were made by M. E. C. and A. I. C. The paper was written by A. I. C., with input from coauthors. We thank Dr Stirling Baird for comments on the work, and specifically for suggesting the SAAS BO experiments.

Conflicts of interest

All authors declare no competing interests.

Data availability

All datasets (optimisation logs, experimental data, ground truth model and data used to build it) are openly available at https://github.com/Ablatif6c/llm-closed-loop-experiments and at https://doi.org/10.5281/zenodo.18632872. In total, the data repository contains 624 log files for optimisations involving LLMs, split by section as follows (Section 3.1.1–120 log files; Section 3.1.2–80 log files; Section 3.2–160 log files; Section 3.3.1–60 log files; Section 3.3.2–60 log files; Section 3.3.3–60 log files; Section 3.4–80 log files; SI – 4 log files).

Supplementary information (SI): BORA convergence proof, Opsight software details, plus 31 SI figures and 5 SI tables. See DOI: https://doi.org/10.1039/d5dd00520e.

Acknowledgements

The authors acknowledge financial support from the Leverhulme Trust *via* the Leverhulme Research Centre for Functional Materials Design. The authors also acknowledge the AI for Chemistry: AIchemy hub for funding (EPSRC grant EP/Y028775/1 and EP/Y028759/1). A. I. C thanks the Royal Society for a Research Professorship (RSRP\S2\232003).

References

1 G. Tom, *et al.*, Self-driving laboratories for chemistry and materials science, *Chem. Rev.*, 2024, **124**, 9633–9732, DOI: 10.1021/acs.chemrev.4c00055.
2 C. Scheurer and K. Reuter, Role of the human-in-the-loop in emerging self-driving laboratories for heterogeneous catalysis, *Nat. Catal.*, 2025, **8**, 13–19, DOI: 10.1038/s41929-024-01275-5.
3 A. Low, J. Cheng, K. Hippalgaonkar and L. Ng, Self-driving laboratories: translating materials science from laboratory to factory, *ACS Omega*, 2025, **7**, 29902–29908, DOI: 10.1021/acsomega.5c02197.
4 R. Canty, *et al.*, Science acceleration and accessibility with self-driving labs, *Nat. Commun.*, 2025, **16**, 11, DOI: 10.1038/s41467-025-59231-1.
5 J. Li, C. Ding, D. Liu, L. Chen and J. Jiang, Autonomous laboratories in China: an embodied intelligence-driven platform to accelerate chemical discovery, *Digital Discovery*, 2025, **4**, 1672–1684, DOI: 10.1039/D5DD00072F.
6 B. Zhang, Z. Zhu, H. Li, J. Cao and J. Jiang, Revolutionizing chemistry and material innovation: an iterative theoretical-experimental paradigm leveraged by robotic AI chemists, *CCS Chem.*, 2025, **7**, 345–360, DOI: 10.31635/ccschem.024.202404860.
7 X. Zhang, *et al.*, Material intelligence by the convergence of artificial intelligence and robotic platforms, *Nexus*, 2025, 100083, DOI: 10.1016/j.ynexs.2025.100083.
8 Y. Su, *et al.*, Automation and machine learning augmented by large language models in a catalysis study, *Chem. Sci.*, 2024, **15**, 12200–12233, DOI: 10.1039/d3sc07012c.
9 S. Lo, *et al.*, Review of low-cost self-driving laboratories in chemistry and materials science: the "frugal twin" concept, *Digital Discovery*, 2024, **3**, 28, DOI: 10.1039/d3dd00223c.
10 H. Hysmith, *et al.*, The future of self-driving laboratories: from human in the loop interactive AI to gamification, *Digital Discovery*, 2024, **3**, 621–636, DOI: 10.1039/d4dd00040d.
11 R. Canty and M. Abolhasani, Reproducibility in automated chemistry laboratories using computer science abstractions, *Nat. Synth.*, 2024, **3**, 1327–1339, DOI: 10.1038/s44160-024-00649-8.
12 S. Bräse, Digital chemistry: navigating the confluence of computation and experimentation – definition, status quo, and future perspective, *Digital Discovery*, 2024, **3**, 1923–1932, DOI: 10.1039/d4dd00130c.
13 O. Bayley, E. Savino, A. Slattery and T. Noël, Autonomous chemistry: Navigating self-driving labs in chemical and material sciences, *Matter*, 2024, **7**, 2382–2398, DOI: 10.1016/j.matt.2024.06.003.
14 J. Zhang, J. Hauch and C. Brabec, Toward self-driven autonomous material and device acceleration platforms (AMADAP) for emerging photovoltaics technologies, *Acc. Chem. Res.*, 2024, **57**, 1434–1445, DOI: 10.1021/acs.accounts.4c00095.
15 Y. Xie, K. Sattari, C. Zhang and J. Lin, Toward autonomous laboratories: Convergence of artificial intelligence and experimental automation, *Prog. Mater. Sci.*, 2023, **132**, 44, DOI: 10.1016/j.pmatsci.2022.101043.
16 R. Hickman, P. Bannigan, Z. Bao, A. Aspuru-Guzik and C. Allen, Self-driving laboratories: A paradigm shift in nanomedicine development, *Matter*, 2023, **6**, 1071–1081, DOI: 10.1016/j.matt.2023.02.007.
17 M. Abolhasani and E. Kumacheva, The rise of self-driving labs in chemical and materials sciences, *Nat. Synth.*, 2023, **2**, 483–492, DOI: 10.1038/s44160-022-00231-0.
18 B. MacLeod, F. Parlane and C. Berlinguette, How to build an effective self-driving laboratory, *MRS Bull.*, 2023, **48**, 173–178, DOI: 10.1557/s43577-023-00476-w.
19 R. Canty, B. Koscher, M. Mcdonald and K. Jensen, Integrating autonomy into automated research platforms, *Digital Discovery*, 2023, **2**, 1259–1268, DOI: 10.1039/d3dd00135k.
20 M. Seifrid, *et al.*, Autonomous chemical experiments: Challenges and perspectives on establishing a self-driving lab, *Acc. Chem. Res.*, 2022, **55**, 2454–2466, DOI: 10.1021/acs.accounts.2c00220.
21 B. MacLeod, F. Parlane, A. Brown, J. Hein and C. Berlinguette, Flexible automation accelerates materials

discovery, *Nat. Mater.*, 2022, **21**, 722–726, DOI: 10.1038/s41563-021-01156-3.
22 N. Szymanski, *et al.*, Toward autonomous design and synthesis of novel inorganic materials, *Mater. Horiz.*, 2021, **8**, 2169–2198, DOI: 10.1039/d1mh00495f.
23 M. Soldatov, *et al.*, Self-driving laboratories for development of new functional materials and optimizing known reactions, *Nanomaterials*, 2021, **11**, 16, DOI: 10.3390/nano11030619.
24 H. Stein and J. Gregoire, Progress and prospects for accelerating materials science with automated and autonomous workflows, *Chem. Sci.*, 2019, **10**, 9640–9649, DOI: 10.1039/c9sc03766g.
25 D. Tabor, *et al.*, Accelerating the discovery of materials for clean energy in the era of smart automation, *Nat. Rev. Mater.*, 2018, **3**, 5–20, DOI: 10.1038/s41578-018-0005-z.
26 R. King, *et al.*, Functional genomic hypothesis generation and experimentation by a robot scientist, *Nature*, 2004, **427**, 247–252, DOI: 10.1038/nature02236.
27 R. King, *et al.*, The automation of science, *Science*, 2009, **324**, 85–89, DOI: 10.1126/science.1165620.
28 P. Nikolaev, *et al.*, Autonomy in materials research: a case study in carbon nanotube growth, *npj Comput. Mater.*, 2016, **2**, 6, DOI: 10.1038/npjcompumats.2016.31.
29 V. Dragone, V. Sans, A. Henson, J. Granda and L. Cronin, An autonomous organic reaction search engine for chemical reactivity, *Nat. Commun.*, 2017, **8**, 8, DOI: 10.1038/ncomms15733.
30 B. MacLeod, *et al.*, Self-driving laboratory for accelerated discovery of thin-film materials, *Sci. Adv.*, 2020, **6**, 8, DOI: 10.1126/sciadv.aaz8867.
31 C. Tamura, *et al.*, Autonomous organic synthesis for redox flow batteries via flexible batch Bayesian optimization, *Digital Discovery*, 2025, **4**, 2737–2751, DOI: 10.1039/D5DD00017C.
32 B. Burger, *et al.*, A mobile robotic chemist, *Nature*, 2020, **583**, 237–241, DOI: 10.1038/s41586-020-2442-2.
33 F. Häse, L. Roch, C. Kreisbeck and A. Aspuru-Guzik, Phoenics: A Bayesian optimizer for chemistry, *ACS Cent. Sci.*, 2018, **4**, 1134–1145, DOI: 10.1021/acscentsci.8b00307.
34 R. J. Hickman, M. Aldeghi, F. Häse and A. Aspuru-Guzik, Bayesian optimization with known experimental and design constraints for chemistry applications, *Digital Discovery*, 2022, **1**, 732–744, DOI: 10.1039/d2dd00028h.
35 A. M. Mroz, P. N. Toka, E. A. D. Chanona and K. E. Jelfs, Web-BO: towards increased accessibility of Bayesian optimisation (BO) for chemistry, *Faraday Discuss.*, 2025, **256**, 221–234, DOI: 10.1039/d4fd00109e.
36 S. G. Baird, A. R. Falkowski & T. D. Sparks, Honegumi: An interface for accelerating the adoption of Bayesian optimization in the experimental sciences, *arXiv*, 2025, arXiv:2502.06815, DOI: 10.48550/arXiv.2502.06815, https://ui.adsabs.harvard.edu/abs/2025arXiv250206815B.
37 R. J. Hickman, G. Tom, Y. H. Zou, M. Aldeghi and A. Aspuru-Guzik, Anubis: Bayesian optimization with unknown feasibility constraints for scientific experimentation, *Digital Discovery*, 2025, **4**, 2104–2122, DOI: 10.1039/d5dd00018a.
38 R. J. Hickman, *et al.*, Atlas: a brain for self-driving laboratories, *Digital Discovery*, 2025, **4**, 1006–1029, DOI: 10.1039/d4dd00115j.
39 Y. Jin and P. Kumar, Bayesian optimisation for efficient material discovery: a mini review, *Nanoscale*, 2023, **15**, 10975–10984, DOI: 10.1039/d2nr07147a.
40 B. Ranković, R.-R. Griffiths, H. B. Moss and P. Schwaller, Bayesian optimisation for additive screening and yield improvements – beyond one-hot encoding, *Digital Discovery*, 2024, **3**, 654–666, DOI: 10.1039/D3DD00096F.
41 F. Häse, M. Aldeghi, R. Hickman, L. Roch and A. Aspuru-Guzik, Gryffin: An algorithm for Bayesian optimization of categorical variables informed by expert knowledge, *Appl. Phys. Rev.*, 2021, **8**, 16, DOI: 10.1063/5.0048164.
42 M. Lindauer, *et al.*, SMAC3: A versatile Bayesian optimization package for hyperparameter optimization, *J. Mach. Learn. Res.*, 2022, **23**, 9.
43 M. A. McDonald, *et al.*, Bayesian optimization over multiple experimental fidelities accelerates automated discovery of drug molecules, *ACS Cent. Sci.*, 2025, **11**, 346–356, DOI: 10.1021/acscentsci.4c01991.
44 A. S. Raihan, H. Khosravi, S. Das and I. Ahmed, Accelerating material discovery with a threshold-driven hybrid acquisition policy-based Bayesian optimization, *Manuf. Lett.*, 2024, **41**, 1300–1311, DOI: 10.1016/j.mfglet.2024.09.157.
45 A. Ramachandran, S. Gupta, S. Rana, C. Li and S. Venkatesh, Incorporating expert prior in Bayesian optimisation via space warping, *Knowl.-Based Syst.*, 2020, **195**, 105663, DOI: 10.1016/j.knosys.2020.105663.
46 A. E. Siemenn, Z. K. Ren, Q. X. Li and T. Buonassisi, Fast Bayesian optimization of Needle-in-a-Haystack problems using zooming memory-based initialization (ZoMBI), *npj Comput. Mater.*, 2023, **9**, 79, DOI: 10.1038/s41524-023-01048-x.
47 D. Eriksson, M. Pearce, J. R. Gardner, R. Turner and M. Poloczek in *Advances in Neural Information Processing Systems 32 (NIPS 2019)*, 2019, vol. 32.
48 M. Adachi *et al.* in *International Conference on Artificial Intelligence and Statistics*, 2024, p. 238.
49 C. Hvarfner, F. Hutter and L. Nardi, A General framework for user-guided Bayesian optimization, *12th International Conference on Learning Representations*, ICLR 2024 2-s2.0-85197024019, 2024.
50 A. Cissé, X. Evangelopoulos, S. Carruthers, V. V. Gusev and A. I. Cooper, HypBO: Accelerating black-box scientific experiments using experts' hypotheses, *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence (IJCAI-24)*, 2024, pp. 3881–3889, https://www.ijcai.org/proceedings/2024/0429.pdf.
51 T. Savage and E. A. D. Chanona, Human-algorithm collaborative Bayesian optimization for engineering systems, *Comput. Chem. Eng.*, 2024, **189**, 108810, DOI: 10.1016/j.compchemeng.2024.108810.
52 A. Biswas, *et al.*, A dynamic Bayesian optimized active recommender system for curiosity-driven partially Human-

in-the-loop automated experiments, *npj Comput. Mater.*, 2024, **10**, 29, DOI: 10.1038/s41524-023-01191-5.
53 A. Ananthaswamy, How close is AI to human-level intelligence?, *Nature*, 2024, **636**, 22–25, DOI: 10.1038/d41586-024-03905-1.
54 C. Criddle, M. Murgia and M. Heikkilä, in *Financial Times*, 2025.
55 L. Jamali and L. McMahon, OpenAI claims GPT-5 model boosts ChatGPT to 'PhD level', 2025, https://www.bbc.co.uk/news/articles/cy5prvgw0r1o.
56 A. Bran, *et al.*, Augmenting large language models with chemistry tools, *Nat. Mach. Intell.*, 2024, **6**, 13, DOI: 10.1038/s42256-024-00832-8.
57 T. Nguyen and A. Grover, *LICO: Large language models for in-context molecular optimization*, 2024, https://arxiv.org/abs/2406.18851v2.
58 D. Boiko, R. Macknight, B. Kline and G. Gomes, Autonomous chemical research with large language models, *Nature*, 2023, **624**, 570, DOI: 10.1038/s41586-023-06792-0.
59 R. Gupta, J. Hartford and B. Liu, LLMs for Bayesian optimization in scientific domains: Are we there yet?, *arXiv*, 2025, arXiv:2509.21403, DOI: 10.48550/arXiv.2509.21403, https://ui.adsabs.harvard.edu/abs/2025arXiv250921403G.
60 M. Akke *et al.*, Bayesian optimization for biochemical discovery with LLMs, *ChemRxiv*, 2025, DOI: 10.26434/chemrxiv-2025-w1wsh.
61 Z. Yang *et al.*, Reasoning BO: Enhancing Bayesian optimization with long-context reasoning power of LLMs, *arXiv*, 2025, arXiv:2505.12833, DOI: 10.48550/arXiv.2505.12833, https://ui.adsabs.harvard.edu/abs/2025arXiv250512833Y.
62 D. Han *et al.*, ChemBOMAS: Accelerated BO in chemistry with LLM-enhanced multi-agent system, *arXiv*, 2025, arXiv:2509.08736, DOI: 10.48550/arXiv.2509.08736, https://ui.adsabs.harvard.edu/abs/2025arXiv250908736H.
63 B. Ranković, R.-R. Griffiths and P. Schwaller, Large language models as uncertainty-calibrated optimizers for experimental discovery, *arXiv*, 2025, arXiv:2504.06265, DOI: 10.48550/arXiv.2504.06265, https://ui.adsabs.harvard.edu/abs/2025arXiv250406265R.
64 T. Liu, N. Astorga, N. Seedat and M. v. d. Schaar, Large language models to enhance Bayesian optimization, *arXiv*, 2024, ArXiv abs/2402.03921, DOI: 10.48550/arXiv.2402.03921.
65 C. Yang *et al.*, Large language models as optimizers, *arXiv*, 2023, ArXiv abs/2309.03409, DOI: 10.48550/arXiv.abs/2309.03409.
66 B. Ranković and P. Schwaller, BoChemian: Large language model embeddings for Bayesian optimization of chemical reactions, *NeurIPS 2023 Workshop on Adaptive Experimental Design and Active Learning in the Real World*, 2023.
67 K. Mahammadli and S. Ertekin, Sequential large language model-based hyper-parameter optimization, *arXiv*, 2024, arXiv:2410.20302, DOI: 10.48550/arXiv.2410.20302, https://ui.adsabs.harvard.edu/abs/2024arXiv241020302M.
68 K. Kobalczyk *et al.*, LILO: Bayesian optimization with interactive natural language feedback, *arXiv*, 2025, arXiv:2510.17671, DOI: 10.48550/arXiv.2510.17671, https://ui.adsabs.harvard.edu/abs/2025arXiv251017671K.
69 K. Ji *et al.*, A closed-loop AI framework for hypothesis-driven and interpretable materials design, *arXiv*, 2025, arXiv:2509.18604, DOI: 10.48550/arXiv.2509.18604, https://ui.adsabs.harvard.edu/abs/2025arXiv250918604J.
70 Y. H. Zou, *et al.*, El Agente: An autonomous agent for quantum chemistry, *Matter*, 2025, **8**, 102263, DOI: 10.1016/j.matt.2025.102263.
71 A. Cissé, X. Evangelopoulos, V. V. Gusev and A. I. Cooper, Language-Based Bayesian Optimization Research Assistant (BORA), *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence (IJCAI-25)*, 2025, pp. 4967–4975, https://www.ijcai.org/proceedings/2025/0553.pdf.
72 Opsight uses external paid frontier models and we do not have a budget available to make this fully open access, but limited access can be provided for non-commercial research purposes upon request.
73 D. Eriksson and M. Jankowiak in *Uncertainty in Artificial Intelligence*, 2021, vol. 161, pp. 493–503.
74 Y. X. Li, *et al.*, Effects of electrolyte NaCl on photocatalytic hydrogen evolution in the presence of electron donors over Pt/TiO2, *J. Mol. Catal. A:Chem.*, 2011, **341**, 71–76, DOI: 10.1016/j.molcata.2011.03.026.
75 M. Sachs, *et al.*, Understanding structure-activity relationships in linear polymer photocatalysts for hydrogen evolution, *Nat. Commun.*, 2018, **9**, 4968, DOI: 10.1038/s41467-018-07420-6.
76 P. Chowdhury, G. Malekshoar and A. K. Ray, Dye-sensitized photocatalytic water splitting and sacrificial hydrogen generation: current status and future prospects, *Inorganics*, 2017, **5**, 34, DOI: 10.3390/inorganics5020034.
77 C. J. Wu, *et al.*, Further studies of photodegradation and photocatalytic hydrogen production over Nafion-coated Pt/P25 sensitized by rhodamine B, *Int. J. Hydrogen Energy*, 2020, **45**, 22700–22710, DOI: 10.1016/j.ijhydene.2020.06.098.
78 Y. Bai, *et al.*, Photocatalytic overall water splitting under visible light enabled by a particulate conjugated polymer loaded with palladium and iridium, *Angew. Chem., Int. Ed.*, 2022, **61**, e202201299, DOI: 10.1002/anie.202201299.
79 A. Dolan, *et al.*, Surfactant effects on hydrogen evolution by small-molecule nonfullerene acceptor nanoparticles, *ACS Appl. Nano Mater.*, 2022, **5**, 12154–12164, DOI: 10.1021/acsanm.2c02350.
80 X. Y. Wang, *et al.*, Sulfone-containing covalent organic frameworks for photocatalytic hydrogen evolution from water, *Nat. Chem.*, 2018, **10**, 1180–1189, DOI: 10.1038/s41557-018-0141-5.
81 H. Lee, W. Zhou, M. Debbah and I. Lee, On the convergence of large language model optimizer for black-box network management, *IEEE Trans. Commun.*, 2025, **73**, 11385–11402, DOI: 10.1109/TCOMM.2025.3592598.
82 https://scispace.com.

83 https://scispace.com/pdf/a-mobile-robotic-chemist-204cvrj0qk.pdf.
84 E. W. C. Spotte-Smith, Considering the ethics of large machine learning models in the chemical sciences, *Mach. Learn.: Sci. Technol.*, 2025, 035007, DOI: 10.1088/2632-2153/adec3c.
85 A. Lynch *et al.*, Agentic misalignment: How LLMs could be insider threats, *arXiv*, 2025, arXiv:2510.05179, DOI: 10.48550/arXiv.2510.05179, https://ui.adsabs.harvard.edu/abs/2025arXiv251005179L.
86 M. Suzgun, *et al.*, Language models cannot reliably distinguish belief from knowledge and fact, *Nat. Mach. Intell.*, 2025, **7**, 1780–1790, DOI: 10.1038/s42256-025-01113-8.
87 A. Meinke *et al.*, Frontier models are capable of in-context scheming, *arXiv*, 2024, arXiv:2412.04984, DOI: 10.48550/arXiv.2412.04984, https://ui.adsabs.harvard.edu/abs/2024arXiv241204984M.
88 K. Darvish, *et al.*, ORGANA: A robotic assistant for automated chemistry experimentation and characterization, *Matter*, 2025, **8**, 20, DOI: 10.1016/j.matt.2024.10.015.
89 X. Li, *et al.*, Sequential closed-loop Bayesian optimization as a guide for organic molecular metallophotocatalyst formulation discovery, *Nat. Chem.*, 2024, **16**, 15, DOI: 10.1038/s41557-024-01546-5.
90 A physically encoded Bayesian assistant for the optimization of multicomponent reactions, *Nat. Chem.*, 2024, **16**, 1225–1226, DOI: 10.1038/s41557-024-01547-4.