Digital Discovery

Accepted Manuscript

This article can be cited before page numbers have been issued, to do this please use: R. A. Patel, M. Li, C. Chang, L. de Lescure, S. Moayedpour, P. Chauvin, A. Cherney, S. Jager and Y. Jangjou, *Digital Discovery*, 2025, DOI: 10.1039/D6DD00052E.

ROYAL SOCIETY OF CHEMISTRY

rsc.li/digitaldiscovery

View Article Online
DOI: 10.1039/D6DD00052E

Distilling and exploiting quantitative insights from Large Language Models for enhanced Bayesian optimization of chemical reactions

Roshan Patel[1], Mingxuan Li[2], Chin-Fei Chang[1], Louis De Lescure[1], Paul Chauvin[3], Alan Cherney[1], Saeed Moayedpour[2], Sven Jager[4], and Yasser Jangjou *[1]

[1]CMC Synthetics Platform, Sanofi, 350 Water St, Cambridge, MA, 02141, USA
[2]Digital R&D, Sanofi, 450 Water St, Cambridge, MA, 02141, USA
[3]Digital R&D, Sanofi, 58-60 avenue de la Grande Armée, Paris, France
[4]Digital R&D, Sanofi, Frankfurt 65929, Germany

March 23, 2026

Abstract

Machine learning and Bayesian optimization (BO) algorithms can significantly accelerate the optimization of chemical reactions. Transfer learning can bolster the effectiveness of BO algorithms in low-data regimes by leveraging pre-existing chemical information or data outside the direct optimization task (i.e., source data). Large Language Models (LLMs) have demonstrated that chemical information present in foundation training data can give them utility for processing chemical data. Furthermore, they can be augmented with and help synthesize potentially multiple modalities of source chemical data germane to the optimization task. In this work, we examine how chemical information from LLMs can be elicited and used for transfer learning to accelerate the BO of reaction conditions to maximize yield. Specifically, we show that a survey-like prompting scheme and preference learning can be used to infer a utility function which models prior chemical information embedded in LLMs over a chemical parameter space; we find that the utility function shows modest correlation to true experimental measurements (yield) over the parameter space despite operating in a zero-shot setting. Furthermore, we show that the utility function can be leveraged to focus BO efforts in promising regions of the parameter space, improving the yield of the initial BO query and enhancing optimization in a majority of the datasets studied. Overall, we view this work as a step towards bridging the gap between the chemistry knowledge embedded in LLMs and the capabilities of principled BO methods to accelerate reaction optimization.

1 Introduction

Machine learning and data-driven approaches can significantly accelerate the optimization of chemical processes [3, 55, 11, 10]. In applications where data is insufficient for comprehensive predictive modeling (e.g., high-throughput screening), Bayesian optimization (BO) algorithms stand out as data-efficient methods to iteratively navigate the chemical and process parameter space to target desired properties from the chemical product [36, 37, 47, 8]. For example, Shields et al. [37] show that BO can work well to identify chemicals (e.g., base, solvent, catalyst ligands) and reaction conditions (e.g., temperature, chemical concentration) to maximize the yields of Buchwald–Hartwig coupling, Suzuki-Miyaura coupling, and direct arylation reactions. We refer readers to a recent review by Guo and Rankovic et al.[17] for comprehensive discussion on successful applications of BO for chemical process development.

Transfer learning can significantly accelerate BO-led workflows by leveraging (source) information or data outside of the direct domain of a given optimization task [4, 40, 13]. For example, source

*Corresponding author: Yasser.Jangjou@sanofi.com

View Article Online
DOI: 10.1039/D6DD00052E

datasets can be used to better inform model development for the domain task [44, 35, 41]. In addition, source data can be used to identify and focus optimization efforts on promising regions of the parameter space through modification of the acquisition function [4, 51, 1, 19, 21, 42, 43]. Overall, though, the application of these and other transfer learning strategies for BO heavily rely on the identification, curation, and numerical encoding of relevant source datasets which are often difficult and laborious to accomplish in practice. Furthermore, qualitative information outside organized datasets (e.g., insights / conclusions found as text in research articles) are typically not leveraged for transfer learning despite representing a large volume of information pertinent for new chemical design tasks.

In recent years, large language models (LLMs) have demonstrated that their ability to model natural language can help perform challenging tasks in disparate chemical domains [18, 45]. For example, with in-context learning, LLMs have been used as regression and classification models to predict chemical properties [23, 31, 24, 22]. In addition, LLMs have shown promise for designing experiments in the context of chemical optimization problems or making meta decisions about the optimization process (e.g., integrating several systems / software needed to execute optimization or deciding a suitable stopping condition) [7, 34, 29, 31, 32, 26]. Given these observations, we posit that LLMs can be queried to transfer pertinent chemical information from source data (e.g., foundation model training data, fine-tuning data) to target Bayesian optimization campaigns and accelerate process development.

In this study, we examine how information from LLMs can be distilled and used to accelerate Bayesian reaction optimization through transfer learning. Specifically, we show that preference learning [14, 9] can be used to infer a utility function over the reaction parameter space from LLM-answered surveys that shows modest correlation to measured reaction yields; promisingly, we accomplish this despite operating in a zero-shot setting with no in-context learning [31, 22, 24, 12] or fine-tuning [22, 26, 20]. Furthermore, we show that when incorporated in the acquisition function, the utility function can be used to focus BO queries in promising regions of the parameter space. We observe that this significantly improves the yield of the initial query to seed BO and enhances optimization in several of the datasets studied. Overall, we view this work as a step towards bridging the gap between the chemistry knowledge embedded in LLMs and the capabilities of principled BO methods to accelerate reaction optimization.

2 Methods

2.1 Datasets

We explore our approach with six chemical reaction datasets compiled by Shields et al.[37, 39, 38] (accessed date: June 2024). Datasets 1-5 correspond to Buchwald-Hartwig (BH) reactions and each contain 792 recorded experiments. Experiments in these datasets are characterized by four reaction parameters: the identity of a specific aryl halide reactant, the palladium precatalyst, the additive, and the base used for the reaction and are labeled by a measured product yield. Dataset 6 corresponds to a direct arylation (DA) reaction and contains 1728 experiments. Experiments in this dataset are characterized by five reaction parameters: the identity of a palladium catalyst ligand, the base, the solvent, temperature, and concentration and are also labeled with a measured product yield. The objective for all datasets is to identify the experiment, characterized by a specific set of reaction parameters, that will give the maximum product yield.

Finally, to evaluate the method's generalization capabilities and rule out data contamination, we utilize three Amide Coupling datasets (AC 1-3, accessed date: Oct. 2025) from a study published in 2025[54, 53]. As these datasets post-date the training cutoff of the LLMs investigated in this study, they serve as a rigorous test of the models' ability to apply chemical reasoning to truly unseen reaction spaces. These datasets involve the optimization of coupling reagents, solvents, and bases for distinct amide bond formation reactions.

2.2 Formulation and implementation of approach

The overall approach taken for each dataset is qualitatively presented in Figure 1. Step 1 of the approach aims at distilling chemical insights from the LLM in the form of a utility function $g(x)$. In step 1a, we formulate a survey in which each question presents two experiments, each characterized by a different set of experimental parameters. In step 1b, we prompt the LLM to answer the survey,

View Article Online
DOI: 10.1039/D6DD00052E

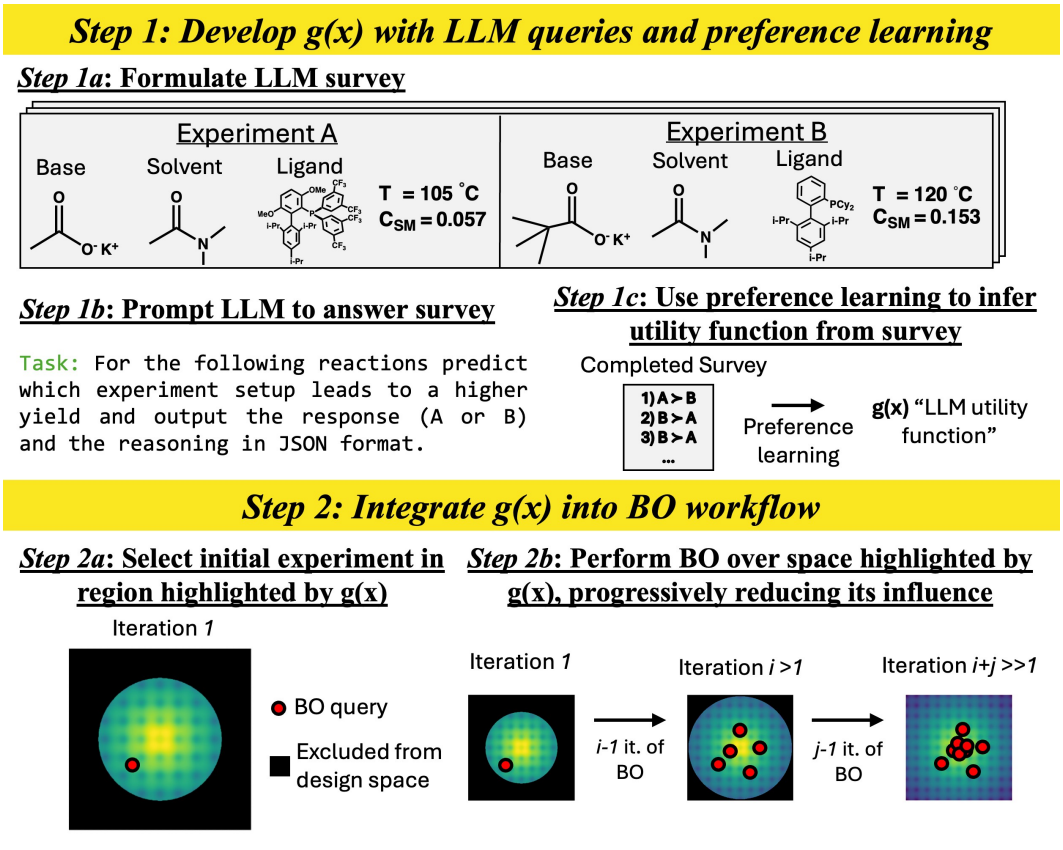


Figure 1: A schematic outlining the major elements of the approach in this study.

selecting which experiment (A or B) it predicts will give the higher yield for each question. In step 1c, we use preference learning to infer a utility function $g(x)$ based on the preferences expressed as choices in the survey. Since we prompt the LLM to prefer experiments with higher predicted yields, we expect $g(x)$ to correlate to yield and thus represent useful, quantitative prior information from the LLM over the set of experiments in a dataset.

Step 2 of the approach aims at leveraging $g(x)$ to expedite BO of reaction parameters. As discussed more in section 2.2.3, we use $g(x)$ to identify promising regions of the parameter space and constrain optimization to these experiments only. Thus in step 2a, we begin optimization by randomly selecting an experiment in the promising set of experiments. In step 2b, we perform BO, gradually removing restrictions on the design space imposed by $g(x)$ as more data is available for surrogate modeling. The pseudo code for the approach is provided at the end of section 2.2.3.

2.2.1 Chemical reaction parameter optimization with Bayesian optimization

Reaction parameter optimization in this work is viewed as a black-box optimization problem:

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

where x is a variable that represents a single candidate experiment, X represents the full set of candidate experiments considered as possible designs, and $f(x_i)$ represents the noiseless output quantity (e.g., yield) that should result from performing experiment x_i; typically, we can access noisy measurements of $f(x_i)$ through experiments: $y_i = f(x_i) + \epsilon$. In our work, we represent experiments x as a concatenation of one-hot-encoded categorical variables (e.g., base, ligand, solvent identity) and continuous variables (e.g., concentration and temperature). Given that obtaining measurements can be time consuming / expensive, the goal is to identify the x^* among all candidate experiments that gives the maximum value of $f(x)$ with as few experiments as possible. Bayesian optimization (BO) is an iterative approach utilizing probabilistic modeling that can be used to solve this class of optimization tasks. At iteration n of BO, we have measured the output of n experiments giving us

View Article Online
DOI: 10.1039/D6DD00052E

dataset $D_n = \{(x_i, y_i)\}_{i=1}^n$. Following, the dataset is used to develop a surrogate model $\hat{f}$ used to predict the output of a given experiment and estimate an uncertainty in that prediction. Gaussian process regression models [33] and Bayesian neural networks [25] are both common surrogate modeling strategies, offering principled methods to estimate a predictive posterior distribution $p(\hat{y}_i \mid D_n, x_i)$. We employ the modeling strategy developed by Shields et al. [37], which leverages GPR with specific priors on kernel parameters suitable for the experiment representation strategy described in this work. All the details of the surrogate model, including kernel specification, prior mean definition and hyperparameter priors, are provided in the supporting information (section 6.3). No deviations from the original formulation were made unless otherwise stated. The surrogate model is used to compute terms in an acquisition function $\alpha(x, D_n)$, the maximizing argument of which is selected as the next best experiment to obtain a measurement for:

$$x^{**} = \underset{x \in X}{\operatorname{argmax}}\, \alpha(x, D_n)$$

The expected improvement (EI) function is a popular choice among studied acquisition functions and has been applied extensively for chemical design [36, 27]:

$$\alpha(x, D_n) = \mathbb{E}_{y \sim p(y|D_n,x)}[\max(y_{\max,n} - y, 0)] \tag{1}$$

where $y_{\max,n}$ is the largest measurement y found in D_n. We use this acquisition as the baseline for our work. Once the maximizing argument x^{**} is found and the measurement y^{**} for the corresponding experiment is taken, x^{**} and y^{**} are added to the dataset ($(x^{**},y^{**}) \cup D_n$) and the next iteration of BO begins. Typically, optimization efforts are concluded when a budget for iterative experimentation has been depleted or improvement upon the largest observed measurement has stagnated for several iterations.

2.2.2 Extracting chemical insights from LLMs

In this section, we discuss our approach to extract quantitative, chemical insights from LLMs in the form of a utility function $g(x)$. We emphasize that LLMs used here as a scalable source of pairwise preference judgments, the proposed BO framework (stated in section 2.2) is agnostic to the choice of preference oracle. First, for each dataset, we formulate a survey consisting of several questions. For each question in a survey, the LLM is presented with two experiments from the dataset (characterized by reaction parameters) and is subsequently prompted to select which of the two would result in a higher yield and to provide its reasoning. An example of a question prompt and the LLM response (specifically by Claude 3.5 Sonnet, version `claude-3-5-sonnet-20240620`) is provided in Figure S3 of the supporting information. To design the survey questions $S_{\text{unanswered}}$ for a given dataset, we created two identical arrays, where each array contains L instances of every experiment in a dataset. Then, we randomly paired elements between each array to form questions, removing repeated questions and questions where the paired experiments were identical. For datasets BH1-BH5 we set L to 10 and for DA L was set to 5 (to keep the total number of questions roughly similar to the BH surveys). This resulted in 7792,7842,7788,7825,7804, and 8610 survey questions designed for the BH1-5 and DA datasets respectively. Overall, we hypothesized that surveys generated using this procedure would facilitate expressing hierarchical preferences over the full set of experiments and subsequent preference learning. In section 4.1, we briefly evaluate the performance of several commonly used foundation LLM models for answering survey questions correctly and select the most accurate model to complete our surveys.

A completed survey is represented as $S_{\text{answered}} = \{(x_{i,j} \succ x_{i,k})\}_{i=1}^m$ where experiment j is preferred over experiment k in question i of the survey with m total questions. Following, we leverage preference learning to infer a utility function $g(x)$ that aligns with LLM predictions made in the survey: namely, $g(x_j)$ should be greater than $g(x_k)$ if $x_j \succ x_k$. Since the LLM was prompted to prefer experiments with higher expected yield based on its chemical reasoning, we expect $g(x)$ to correlate to the true experimental yield measured for experiments in a dataset. We follow the approach of Chu and Gharamani [9], who model $g(x)$ as a Gaussian process and define a function to model the likelihood of observing a preference among pairs of options given their values from the utility function (assumed to contain noise). To tune hyperparameters (e.g., parameters of the kernel), they use the Laplace approximation to define an expression of the posterior density over utility functions conditioned on the data and

View Article Online
DOI: 10.1039/D6DD00052E

optimize it (MAP estimate). For our work, we employ the BoTorch [**BoTorch**] implementation of Chu and Gharamani's approach using the PairwiseGP module. Details of the model implementation are provided in the supporting information (section 6.4). Upon training, we take the mean of the GP posterior conditioned on the survey data as the utility function $g(x)$ and to be a representation of prior chemical knowledge embedded in the LLM over the chemical parameter space.

2.2.3 Leveraging chemical insights from LLMs for enhanced optimization

A common way to incorporate prior-knowledge or information in the BO algorithm is through an adjustment of the acquisition function. For example, Souza et al.[42] and Hvarfner et al. [21]. weight the standard BO acquisition function with a decaying (as a function BO iterations) prior probability function that computes the probability π that experiment x maps to the maximum of f. In doing so, the acquisition function is biased to explore promising regions of the parameter space encoded in $\pi(x)$ in early iterations of BO. Our work follows their weighting framework, computing the modified acquisition function as:

$$\alpha_{\pi,n}(x, D_n, n) = \alpha(x, D_n)\pi(g(x), p(n)) \tag{2}$$

π is computed as a simple indicator function:

$$\pi(g(x), p(n)) = \begin{cases} 1 & \text{if } g(x) \geq P_{p(n)} \\ 0 & \text{if } g(x) < P_{p(n)} \end{cases}$$

where P_p is pth percentile value of the set $G = \{g(x) \mid x \in X\}$. This binary weighting to the acquisition allows optimization to focus on promising regions of the chemical space highlighted by g, without further biasing candidate selection with potentially noisy utility values. Our approach can also be viewed as design space pruning [30, 49, 16], where unpromising portions of the design space are excluded from the set X of candidate experiments. Given that our weighting / pruning approach may adversely impact optimization if g is negatively correlated with f (or by excluding the true maximizing argument of f), we recommend setting percentile $p(n)$ as a decaying function of BO iterations n such that $p \to 0$ as $n \to \infty$. In effect, this relaxes the constraint on the design space imposed by g for candidate selection as more experiments are performed and the surrogate model $\hat{f}$ becomes increasingly reliable. In our work, we select $p(n)$ as a simple 2-step function; Sections 4.2 and 6.1 provide additional details on how parameters for $p(n)$ were selected in our work.

Algorithm 1 Pseudo-code for LLM-augmented BO

Input: Parameter space X, Number of BO queries N, Acquisition function a, Percentile function $p(n)$, Experiment instances L
Step 1: Develop utility function $g(x)$ via LLM queries and preference learning
Pair elements of two identical arrays containing L instances of each experiment to formulate the survey $S_{\text{unanswered}} = \{(x_{i,j}, x_{i,k})\}_{i=1}^{m}$, $x_{i,j}, x_{i,k} \in X$
For each question in the survey prompt the LLM to predict which experiment will give a higher yield: $S_{\text{answered}} = \{(x_{i,j} \succ x_{i,k})\}_{i=1}^{m}$
Use preference learning to infer utility function: $g(x) \leftarrow \text{Train}(S_{\text{answered}}, X)$
Step 2: Integrate utility function $g(x)$ into BO workflow
Evaluate $g(x)$ over the parameter space: $G = \{g(x) | x \in X\}$
Randomly select an experiment from the set of promising experiments identified by the utility function: $x_0 \sim \{x | \pi(g(x), p(n=0)) = 1 \text{ and } x \in X\}$
Initialize dataset for BO with this point and its measured label: $D_0 = \{(x_0, y_0)\}$
for n = 1 to $N - 1$ **do**
  Train surrogate model: $\hat{f}(x) \leftarrow \text{Train}(D_n)$
  Obtain candidate by optimizing acquisition function: $x^{**} = \text{argmax}_{x \in X}\, \alpha(x, D_n)\pi(g(x), p(n))$
  Augment dataset with this point and its measured label: $D_n = ((x^{**}, y^{**}) \cup \{(x_i, y_i)\}_{i=0}^{n-1})$
end for
return $x^* = \text{argmax}_{(x_i, y_i) \in D_{N-1}}\, y_i$

View Article Online
DOI: 10.1039/D6DD00052E

3 Related Works

3.1 Transfer learning

In one paradigm of transfer learning, prior information is leveraged to make judicious modifications to the acquisition function to accelerate Bayesian optimization in a target domain [4, 46, 50, 19]. For example, Souza et al.[42] and Hvarfner et al. [21] weight the acquisition function with a prior $\pi(x)$ on the function's maxima, biasing the BO algorithm to focus early optimization efforts on regions of the parameter space with high probability mass. As mentioned, our approach of leveraging information encoded in $g(x)$ follows a similar framework. In their work, however, $\pi(x)$ is typically encoded as a parameterized probability function; choosing the type of function or specific parameter values to match source data or information can be non-trivial. Along with formulating a new acquisition function to incorporate $\pi(x)$, Adachi et al. [1] propose using preference learning to distill insights from human experts and obtain $\pi(x)$. Our own experiments suggested that the quality and quantity of data collected in surveys completed by human experts was not sufficient to apply this method for our domain application. We posited that LLMs offer a promising alternative to human experts: they can answer orders of magnitude more questions in a fraction of the time and could leverage chemical information in source data to accurately answer questions.

3.2 LLM-augmented Bayesian optimization

A few recent works have explored how LLMs can be used to accelerate BO in chemical systems; predominantly, these works have leveraged LLMs to inform the development of the surrogate model. One strategy is to use the LLM as the surrogate model itself. For example, Ramos et al. [31] show that in-context learning and specific prompting strategies (and interpretation of token probabilities) could be used to develop a regressor capable of uncertainty quantification, which they then use for BO. Another approach is to use the LLM to process some description of the chemical system / experiment and produce an embedding from which a surrogate model can be trained to make a prediction. For example, Ranković and Schwaller [32] show that these LLM embeddings are competitive with (and can outperform) those obtained from more sophisticated and domain-informed pre-training procedures. Kristiadi et al. [26] show that the performance of this approach can be further improved when using domain specific and fine-tuned LLMs. Furthermore, they show that parameter-efficient fine-tuning and Bayesian neural networks can offer a principled way to use the LLM as a surrogate model and allow it to further learn informative embeddings of reactions. Similarly, Zeng et al.[52] propose an LLM-enabled multi-task BO framework that uses fine-tuned LLMs to transfer knowledge across tasks via strong initialization points. Aglietti et al.[2] introduce FunBO, a framework that leverages LLM-driven program search to generate new acquisition functions expressed in code. Overall, these are promising developments in leveraging source information in LLMs to expedite BO in a target domain. Our work differs from these approaches in that we separate the modeling of the target information (GPR surrogate model $\hat{f}(x)$) and the source information (LLM-derived utility function $g(x)$); instead the source information is included at the point of defining the acquisition function. Overall, the binary weighting scheme we use to adjust the acquisition function accomplishes a similar purpose to what is presented by Liu et al. [28], who use the LLM to first pre-select which points are considered for initialization and optimization at a given iteration. We suggest that obtaining and exploiting quantitative information present in the utility function can provide finer control for experiment selection strategies.

4 Results and Discussion

4.1 Survey grades and preference learning outcomes

We first evaluated multiple LLMs on their ability to distill chemistry insights based on their performance of survey-style preference questions for each dataset. The LLMs were prompted to answer the surveys designed for each dataset and subsequently each survey was graded for accuracy. A set of 1,000 question pairs was generated for each dataset by randomly pairing distinct, non-identical experiment conditions. Specifically, for each question in a given survey, the question was marked "correct" if the LLM preference (its prediction for whether experiment A or experiment B has the higher yield) aligned with the ground truth; the percentage of questions answered correctly in a given survey is

View Article Online
DOI: 10.1039/D6DD00052E

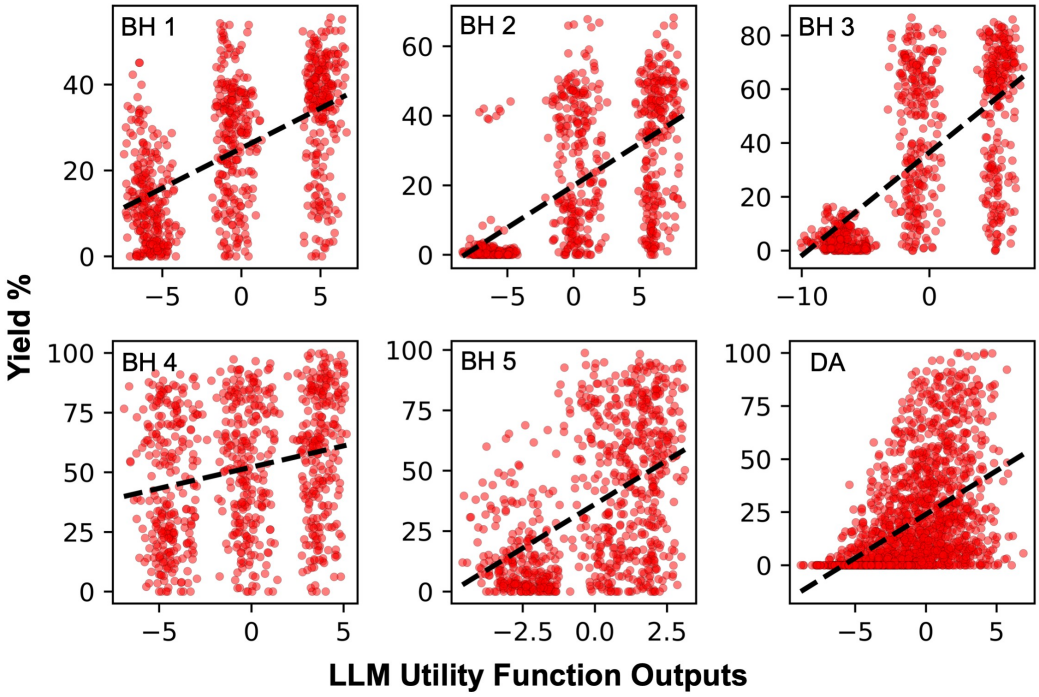


Figure 2: Assessing the correlation between utility function outputs computed for experiments across all datasets and their true measured yield. The Pearson r correlation between utility values and yields are 0.55, 0.63, 0.67, 0.22, 0.49, and 0.48 for datasets BH 1, BH 2, BH 3, BH 4, BH 5, and DA respectively, with a p-value $<$ 1e-10 across all datasets. The least squares regression line between utility values and yields is plotted for each panel in a dotted black line to guide the eye.

defined as the accuracy. The same fixed set of 1,000 question pairs for each dataset was used across all evaluated LLMs (Sonnet-3.5, Sonnet-3, haiku-3 and GPT-4) and the resulting accuracies are shown in Figure S2. Despite a few points drop below 50% on the BH 4 and BH 5 datasets, we observe that the overall accuracies for all LLMs surveys exceed 50% with Sonnet-3.5 consistently outperforming the rest. Based on this result, Sonnet-3.5 was selected and used for all the subsequent works in this study. This suggests that the LLMs could leverage chemical knowledge trained in the foundation model to make informed decisions about which of two experiments would result in a higher yield.

We applied Sonnet-3.5 to a completed surveys (full length instead of 1,000-question, sample surveys) together with preference learning to infer the utility function. Figure S3 gives an example of the typical reasoning provided by the LLM in answering survey questions; we observe that decisions are made from relatively simple chemical reasonings (e.g., polarity of the solvent, strength of base, stereochemistry of ligand). Overall though, we find that this is enough to achieve survey accuracies above 50% (one-tailed binomial test statistically significant for all surveys, $p < 0.01$), suggesting that despite the simplicity, the chemical information embedded in the LLM is pertinent enough to help make (on average) informed decisions.

Following, for a given dataset, we use the LLM-completed survey and preference modeling to infer the utility function; we compare its output for experiments to their true measured yields. Overall, we observe a positive correlation between utility function outputs and true experimental yields for each dataset (Figure 2), indicating that preference modeling could be used to infer a utility function that aligned with the chemically informed, LLM-completed surveys. Importantly, the outputs are not on the same scale as measured yield given that they only encode the utility of a given experiment and not the yield directly. We compared our approach to directly asking the LLM to predict the yield from descriptions of the reaction parameters and for the given reaction (i.e., zero-shot regression), which we observe gives output values that do not correlate positively to yield (Figure S4). Overall, this suggests that the LLM survey + preference modeling approach presented herein is a promising way to distill quantitative insights from LLMs in the zero-shot setting.

View Article Online
DOI: 10.1039/D6DD00052E

Interestingly, for several of the datasets we observe distinct clusters where the utility function gives similar output values for different experiments (Figure 2), likely reflecting the prior observation that (for datasets that show clustering) the LLM is largely leveraging simple chemical reasoning (i.e., based on 1 or 2 features) to rank one experiment over another. For datasets BH 1-4 we observe three distinct clusters, dataset BH 5 has two loosely defined clusters, and dataset DA shows no clustering. Notably, while the mean yield of experiments increases with the mean utility value of each cluster (giving rise to the overall correlation), the yield of experiments within a given cluster correlate relatively poorly to corresponding preference model outputs. We posit that for experiments within a cluster, the LLM was not able to apply sound chemical reasoning to predict why one experiment should result in a higher yield than another resulting in random predictions in surveys, and manifesting as overfit noise in the preference model. This observation motivated the formulation of the approach detailed in section 2.2.3, where we essentially attempt to restrict BO to query experiments found in clusters with the highest mean preference value and forgo the precise value due to the apparent noise within the cluster. We suspect that future workflows may benefit from identifying questions in the survey where the LLM is uncertain (e.g., "hallucinating") in its response (e.g., through repeated questioning) and removing uncertain responses from the preference model training data.

4.2 BO experiments

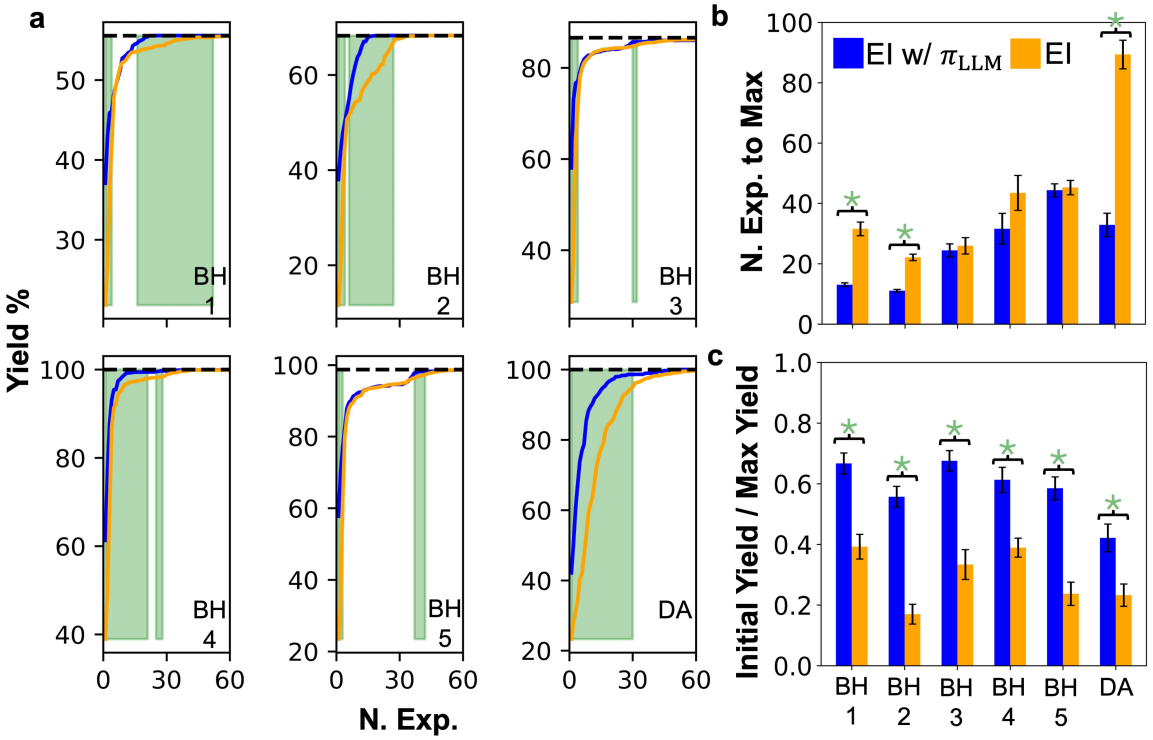


Figure 3: Comparison of BO reaction parameter optimization using the expected improvement acquisition function versus the LLM-preference-guided expected improvement acquisition. Panel **a** plots the best measured yield as a function of the number of experiments performed for each dataset in BO campaigns using a given acquisition function. Each line represents the mean value at a given number of experiments across $n = 50$ randomly seeded campaigns; across all lines, the standard errors are small, and the shaded regions closely track the mean. Panel **b** shows the mean number of experiments required to observe the maximum yield for a given dataset and acquisition function, along with the standard error from $n = 50$ trials. Panel **c** shows the average yield observed in the initial experiment selected during BO, again with standard error from $n = 50$ trials. All values are normalized by the maximum observed yield for each dataset. A two-tailed Welch's t-test is used to assess the significance ($p < 0.01$) of differences in mean metrics between the two acquisition functions across all panels. No marker denotes no significant difference, green markers indicate significant improvement of our method over the baseline, and red markers indicate significant underperformance relative to the baseline.

View Article Online
DOI: 10.1039/D6DD00052E

Next, we aimed to compare the performance of the expected improvement (EI) acquisition function (eq. 1) to the LLM utility function–modified EI acquisition function (LLM-EI) (eq. 2) for Bayesian reaction parameter optimization in the datasets included in this study. We perform 50 independent optimization runs for each dataset and acquisition function. Optimization runs using the EI acquisition function are initialized by randomly selecting a single experiment from the given dataset. Optimization runs for the LLM-EI acquisition function are seeded by randomly selecting an experiment from the set $\{\, x \mid x \in X \ \wedge \ \pi(g(x), p(n=0)) = 1 \,\}$ within a given dataset. For each run, we track the maximum yield observed at a given iteration of BO.

Before performing our comparison, however, the precise functional form of $p(n)$ required specification. In general, we expect that the functional form of $p(n)$ best suited for an optimization task will depend, in part, on topological features of the true property surface (e.g., modality in $y(x)$) and on the quality of the optimization prior $g(x)$ (i.e., its correlation with the true property surface $y(x)$). Since neither of these are known *a priori*, we sought to develop a form of $p(n)$ that performs well across several datasets (BH1–5) empirically and to subsequently evaluate its performance on additional reaction optimization datasets not used during parameter tuning (DA, AC1–3). Section 6.1 provides additional details of the procedure used to develop $p(n)$ and specifies the functional form used for all optimization results presented henceforth.

Overall, we observe that the LLM-EI acquisition function either significantly outperforms or performs comparably to the EI acquisition function, with no statistically significant differences, across all measured metrics for BH1-5, which were used to tune $p(n)$, as well as DA, which was not. Specifically, Figure 3a shows that for BH1, BH2, BH4, and DA, LLM-EI often achieves a higher average maximum yield at a given number of experiments in the optimization campaign compared to EI, whereas datasets BH3 and BH5 exhibit comparable performance between the two acquisition functions. Furthermore, Figure 3 shows that the mean number of experiments required to identify reaction parameters yielding the maximum outcome decreases significantly when using LLM-EI compared to EI, from 32 to 13 (59% decrease), 22 to 11 (50% decrease), and 89 to 33 (63% decrease) for datasets BH1, BH2, and DA, respectively. Due to convergence difficulties in dataset BH4, in which near-optimal yields (>99% of the maximum) are reached early but additional experiments are required to locate the absolute maximum (potentially due to GP noise or a multimodal objective landscape), the mean number of experiments needed to reach the maximum yield shows no statistically significant difference between the two acquisition functions. As a result, the mean number of experiments required to reach the maximum yield shows no statistically significant difference between the two acquisition functions. However, Figure S5 shows that the mean number of experiments required to identify reaction parameters achieving 99% of the maximum attainable yield decreases significantly from 19 to 9 (53%) for dataset BH4; LLM-EI is similarly advantageous for identifying 99% of the maximum yield for BH1, BH2, and DA. For BH3 and BH5, the mean number of experiments required to identify either 99% or 100% of the maximum yield is similar between the two acquisition functions and does not show statistically significant differences. Additionally, Figure 3c shows that Bayesian optimization seeding experiments selected using the utility function yield, on average, significantly higher outcomes than those selected at random across all datasets. This may be advantageous in applications where moderate yields or property values are sufficient to advance development, rather than requiring near-maximum values. Overall, these results suggest that utility functions inferred from LLM-completed surveys can help identify promising regions of chemical space and improve the efficiency of Bayesian optimization. Furthermore, they demonstrate that the optimized $p(n)$ performs well across BH1-5 and generalizes effectively to DA (which was not included in the optimization process) despite differences between the datasets, such as the quality of $g(x)$ shown in Figure 2.

4.3 Validation on Newer Datasets: Amide Coupling

A critical concern in using LLMs for scientific tasks is data contamination—the possibility that the model performs well simply because it has seen the optimization landscape in its training data. To address this and further validate the elements of our methodology, we applied the LLMBO workflow to three Amide Coupling datasets (AC 1-3) published in 2025[54], which are temporally disjoint from the LLM's training data.

The results for these unseen datasets are summarized in Figure 4. We observe that the qualitative chemical knowledge embedded in the LLM successfully transfers to these new reaction spaces. In Panel **a**, the LLM-EI acquisition function (blue) finds high-yielding conditions significantly faster than

View Article Online
DOI: 10.1039/D6DD00052E

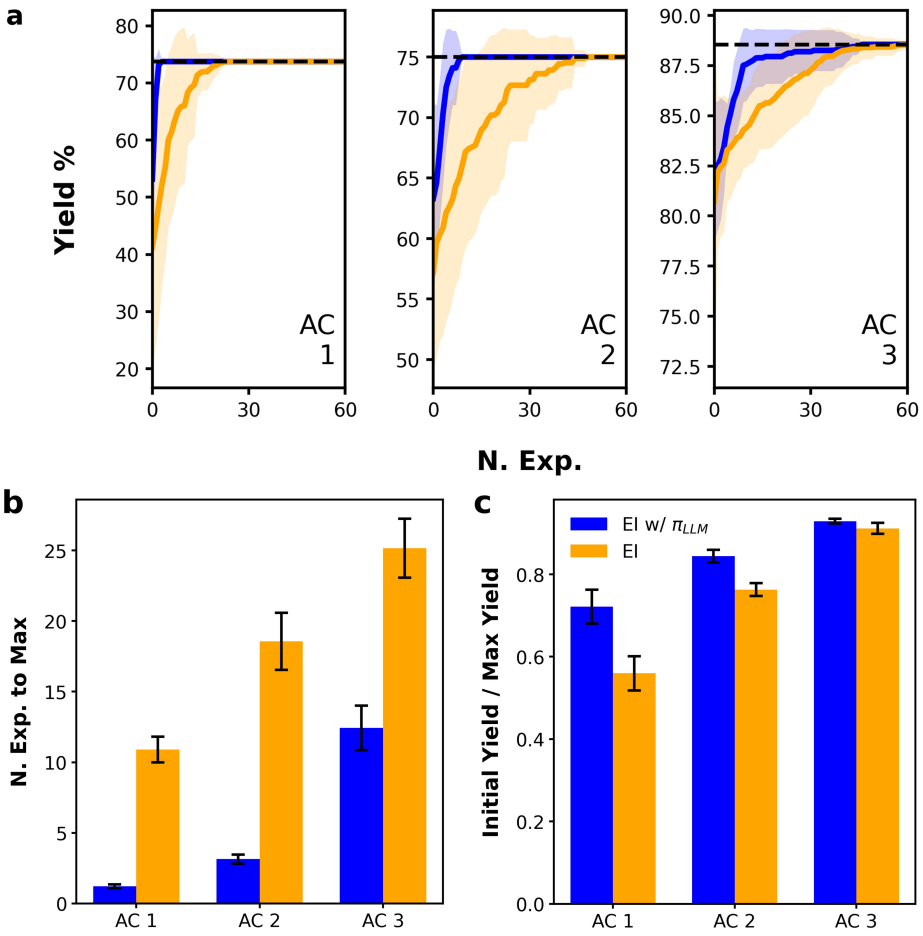


Figure 4: Performance on "unseen" Amide Coupling datasets (AC 1-3). **a** Optimization traces showing the best yield observed versus the number of experiments performed; shaded regions represent the standard deviation ($n = 50$). The blue line (LLM-EI) demonstrates faster convergence compared to the baseline (EI, orange). **b** The mean number of experiments required to reach the maximum yield; error bars represent the standard error of the mean (SEM). LLM-EI significantly reduces the experimental burden across all three datasets. **c** The ratio of the yield obtained in the initial experiment to the maximum possible yield. The LLM-guided initialization consistently selects starting conditions with higher yields than random selection.

standard EI (orange). This efficiency is quantified in Panel **b**, where the mean number of experiments required to reach the maximum yield is drastically reduced—most notably in AC 1 and AC 2, where the experimental budget is cut by more than half.

Furthermore, Panel **c** illustrates that even in this unseen domain, the LLM's "zero-shot" preferences allow for a superior initialization strategy. The initial experiments selected by the LLM consistently achieve 70-90% of the maximum possible yield, compared to significantly lower starting points for random selection. This confirms that the performance gains observed in the Buchwald-Hartwig and Direct Arylation datasets are driven by genuine chemical reasoning rather than rote memorization of literature data. Furthermore, we suggest these results provide strong evidence that the various methodological choices made in this study (e.g.,design of survey questions, selection of a foundational LLM, specification of $p(n)$) can generalize well to other reaction optimization tasks.

5 Discussion and Conclusion

In this study, we presented an approach to distill and use quantitative insights from LLMs to accelerate Bayesian reaction optimization with transfer learning. Specifically, we prompted the LLM to complete surveys in which each question of a survey asks the LLM to predict which of two experiments is expected

View Article Online
DOI: 10.1039/D6DD00052E

to provide the higher yield. We find that the LLM typically employs simple chemical logic to make predictions which led to (on average) correct predictions in surveys. Following, for each dataset, we used preference learning to infer a utility function $g(x)$ which quantitatively models LLM preferences expressed in a survey. We found that the outputs of utility functions show modest correlation to the true yield measured for experiments in a given dataset; thus we interpret $g(x)$ as an expression of prior information provided by the LLM over the chemical parameter space. Lastly, we show that the outputs of $g(x)$ can be used to focus BO queries on promising regions of the parameter space, leading to significantly enhanced optimization in several of the datasets examined and higher experimental yields for initial BO queries.

Moving forward we anticipate several avenues of investigation to improve the performance of the method presented in this work. In the first line of investigation, we posit that working to maximize the correlation between $g(x)$ and $f(x)$ for a given optimization task would enable the pruning algorithm to better focus optimization efforts on promising regions of the design space and further accelerate discovery. We imagine several areas of improvement in our algorithm. First, we suspect that fine-tuning the LLM with domain-specific literature or using in-context learning (possibly identified via document retrieval systems) could be used to refine the information used to answer survey questions, improving the chemical knowledge encoded in completed surveys. In addition, we suspect that parameters surrounding the survey itself can be further optimized to better encode the chemical knowledge / reasoning of the LLM. For example, it may be advantageous to explore alternative formulations of the survey (e.g., ranking several experiments at the same time) and corresponding preference modeling strategies. In addition, it may be possible in some capacity to remove survey questions (e.g., repeated queries, specific prompting) where the LLM is very uncertain about a response, which would remove noisy responses from the dataset used to infer the preference model. Lastly, in any application, it would be important to explore sensitive to the precise wording used to elicit responses from the LLM.

In another line of work, we posit that it may be advantageous to estimate the quality of $g(x)$, for example, by using the first few labeled experiments obtained during an optimization campaign to validate the LLM's reasoning. As a primary benefit, this could allow the user to avoid failure modes of the method, i.e., when $g(x)$ is negative because the LLM consistently expresses incorrect chemical reasoning in survey responses, which we do not observe but could in principle occur. In such situations, it may be necessary to remove the influence of $g(x)$ in optimization efforts by reverting to the baseline acquisition function. In other cases, additional information about $g(x)$ could enable informed modifications to the $p(n)$ used for pruning in this work, which was designed to operate across a wide range of $g(x)$ values and other factors that likely affect the success of pruning (e.g., modality of $f(x)$, roughness of $f(x)$). For example, in cases where $g(x)$ is estimated to be close to 1, it may be advantageous to adapt $p(n)$ such that it more aggressively prunes experiments at smaller n. Overall, however, we suspect that such informed modifications to $p(n)$ will require identifying which aspects of $f(x)$ most strongly influence improvements, as well as characterizing how improvement depends on the combined effects of the quality of $g(x)$, the characteristics of $f(x)$, and the choice of $p(n)$.

6 Supporting Information

6.1 Selection and Parameterization of the Step Function $p(n)$

For our work, we constructed p(n) as a step function parameterized by three values:

$$p(n) = \begin{cases} v_1 & \text{if } n \leq c_1 \\ v_2 & \text{if } c_1 < n \leq 40 \\ 0 & \end{cases}$$

To optimize these three parameters, we used 300 iterations of the Tree-Structured Parzen Estimator [6] implemented by the Hyperopt package [5] over the parameter space: $v_1 \in [70, 95]$, $v_2 \in [0, 70]$, $c_1 \in [1, 39]$ and select the parameter set that gives the best performance among datasets BH1-5 relative to the baseline EI. Specifically, the minimization objective was computed using the following formula:

View Article Online
DOI: 10.1039/D6DD00052E

$$a(v_1, v_2, c_1) = \sum_{i=1}^{5} (\bar{n}_{\text{max,EI-LLM,BH}i}(v_1, v_2, c_1) - \bar{n}_{\text{max,EI,BH}i}) - 50 \cdot \mathbb{I}\left[\forall i,\ \bar{n}_{\text{max,EI-LLM,BH}i} < \bar{n}_{\text{max,EI,BH}i}\right] \tag{3}$$

where $\bar{n}_{\text{max,EI-LLM,BH}i}(v_1, v_2, c_1)$ is the number of experiments needed to observe the maximum yield (averaged over 50 trials) using the EI w/LLM acquisition function for dataset BHi, $\bar{n}_{\text{max,EI,BH}i}$ is the corresponding quantity obtained using the standard EI acquisition function, and $\mathbb{I}[\cdot]$ denotes the indicator function. The indicator term applies a fixed bonus of -50 whenever the EI w/LLM acquisition function requires fewer experiments than standard EI across all five datasets used for parameter optimization(BH1–BH5). This bonus term was introduced to explicitly favor parameter settings that consistently improved optimization performance across all datasets to emphasize robustness in the designed $p(n)$.

Optimization yields parameters $v_1 = 85\%$, $v_2 = 15\%$, and $c_1 = 30$ giving the $p(n)$ shown in Figure S1. We use this $p(n)$ for all BO trials employing the EI w/LLM acquisition function in Figure 3 and Figure 4. The good performance of LLM-EI on the DA dataset shown in Figure 3 and the three additional AC 1-3 dataset in Figure 4 (not used for parameter optimization) serves as validation of the $p(n)$ optimized in our study.

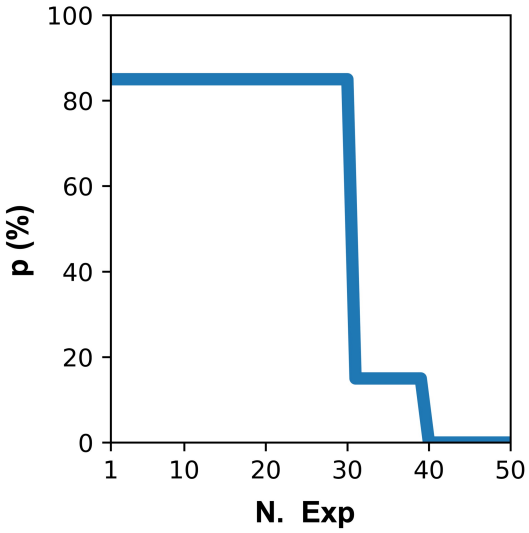


Figure S1: Optimized $p(n)$ used for the LLM-EI acquisition function. The plot shows the cutoff percentile p% (y-axis) for outputs of g(x) as a function of iterations of Bayesian optimization (x-axis). For example, according to this function and the definition of $\pi(x)$ presented in section 2.2.3, the first 30 experiments are selected from the subset experiments with preference values above the 85 percentile of g(x) for $x \in X$.

View Article Online
DOI: 10.1039/D6DD00052E

6.2 Supplementary Figures

To understand the performance of LLMs in reasoning the relation of chemical reaction conditions and yield, four LLMs (Sonnet-3.5, Sonnet-3, haiku-3 and GPT-4) were applied in the yield preference survey. Each datasets has 1,000 question pairs and the accuracy for each LLM is shown in Figure S2. Accuracy of a full-length survey for each dataset using Sonnet-3.5 is also presented (purple dot line).

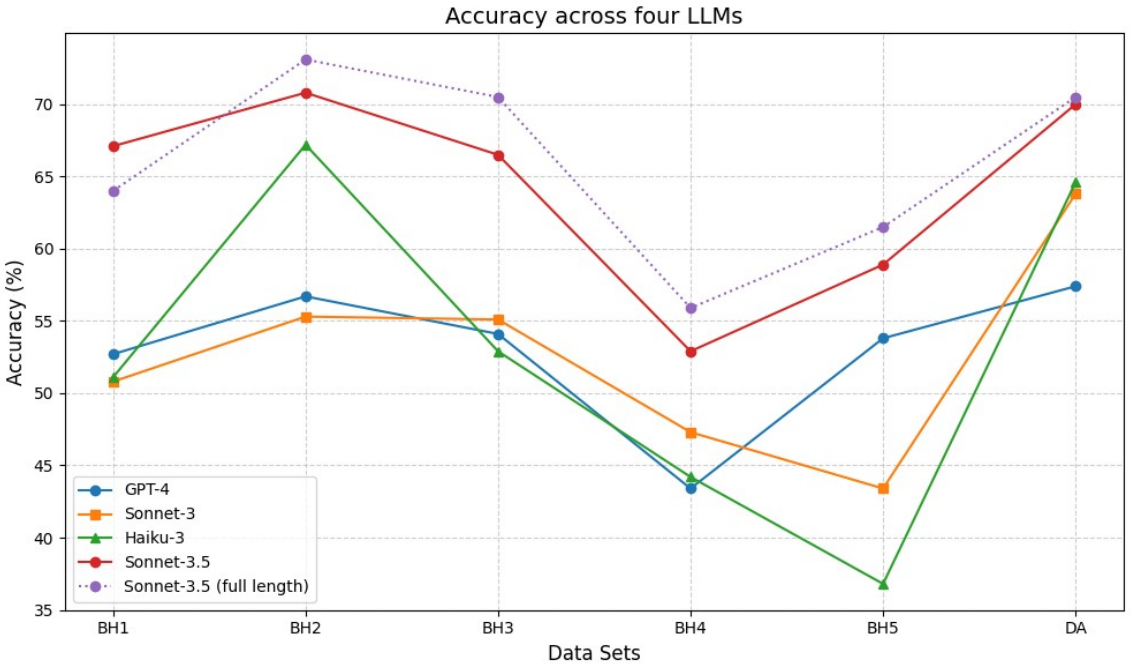
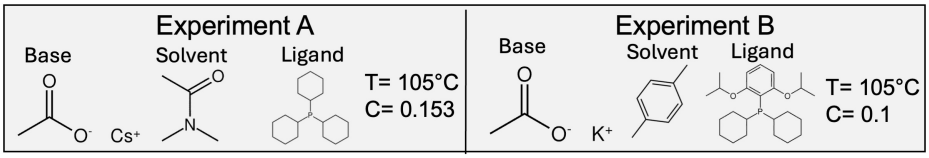


Figure S2: Comparison of LLMs performance in chemical reaction yield reasoning.

View Article Online
DOI: 10.1039/D6DD00052E

Figure S3 shows an example of typical reasoning used by the LLM to answer survey questions. This example is a randomly selected question from the survey generated for the DA dataset.



Example Prompt
"Task: For the following reaction, predict which experiment setup leads to a higher yield and output the response (A or B) along with reasoning in JSON format. Task reaction: "Direct Arylation" For any reaction, an additive, a base, and a palladium-ligand are present.

Setup A:
Base SMILES: O=C([O-])C.[Cs+]
Solvent SMILES: CC(N(C)C)=O
Ligand SMILES: P(C1CCCCC1)(C2CCCCC2)C3CCCCC3
Temperature: 105°C
Concentration: 0.153 M

Setup B:
Base SMILES: O=C([O-])C.[K+]
Solvent SMILES: CC1=CC=C(C)C=C1
Ligand SMILES: CC(OC1=C(P(C2CCCCC2)C3CCCCC3)C(OC(C)C)=CC=C1)C
Temperature: 105°C
Concentration: 0.1 M

Output: Reaction setup (A or B) with higher yield and the reasoning in a JSON object with 'Setup' and 'reasoning' keys."

Example Response
"Setup A is likely to lead to a higher yield for the given reaction due to the following reasons:
1- Base: Cesium carbonate (Cs2CO3) in Setup A is generally a stronger base than potassium carbonate (K2CO3) in Setup B. Stronger bases are often more effective in promoting cross-coupling reactions.
2- Ligand: The ligand in Setup A (triphenylphosphine) is a common and effective ligand for many cross-coupling reactions. It's simpler structure may provide better stability and catalytic activity compared to the more complex ligand in Setup B.
3- Solvent: Dimethylacetamide (DMAc) in Setup A is a polar aprotic solvent that is often used in cross-coupling reactions. It can effectively dissolve both organic and inorganic compounds, which is beneficial for this type of reaction. Toluene in Setup B is less polar and may not be as effective.
4- Concentration: Setup A has a higher concentration (0.153 M) compared to Setup B (0.1 M), which could lead to faster reaction rates and potentially higher yields.
5- Temperature: Both setups have the same temperature, so this factor does not differentiate between them.
6- Overall, the combination of a stronger base, effective ligand, suitable solvent, and higher concentration in Setup A is more likely to result in a higher yield for this cross-coupling reaction."

Figure S3: Example of LLM reasoning in answering survey

View Article Online
DOI: 10.1039/D6DD00052E

Figure S4 provides a comparison between two methods aiming to extract quantitative insights from the LLM. Panel a shows the result of zero-shot LLM regression for the DA dataset, i.e., the LLM was now described the conditions of a single experiment (in the same way that was done for answering the survey) but instead prompted to provide the predicted yield instead of a preference. Panel b shows the utility function $g(x)$ outputs. The utility function outputs show a modest correlation to the true experimentally measured yield, whereas the LLM-predicted yields do not.

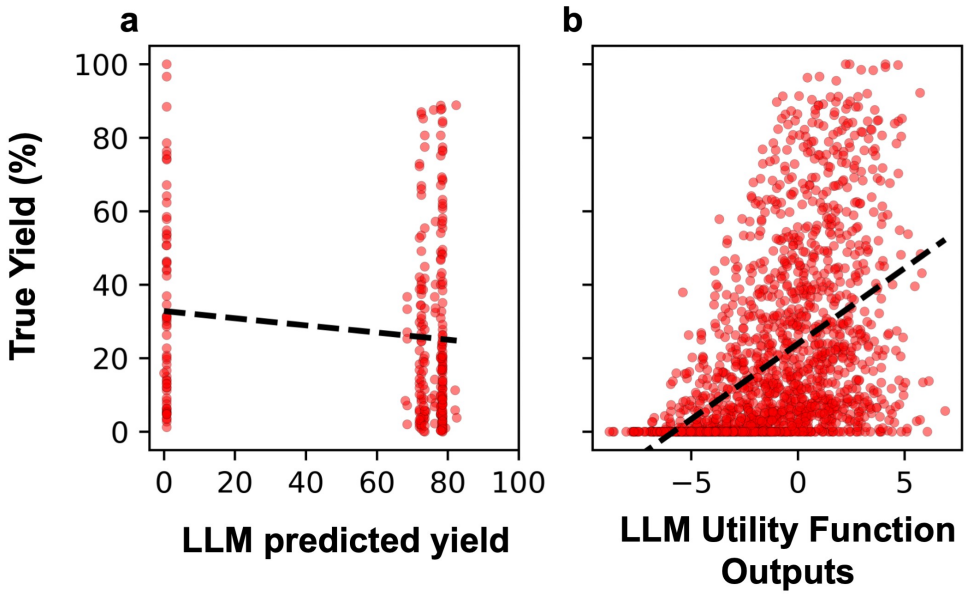


Figure S4: Comparison between simple zero-shot regression (panel a) and preference learning on survey (answered by zero-shot LLM) for the DA dataset (panel b)

View Article Online
DOI: 10.1039/D6DD00052E

Figure S5 shows the mean (and standard error from n=50 trials) number of experiments needed to run to observe 99% of the maximum yield for a given dataset and acquisition function. The x-axis shows dataset names ranging from Buchwald-Hartwig (BH) 1 to 5 and Direct Arylation (DA) (left to right), and the y-axis shows the number of experiment to reach 99% maximum.

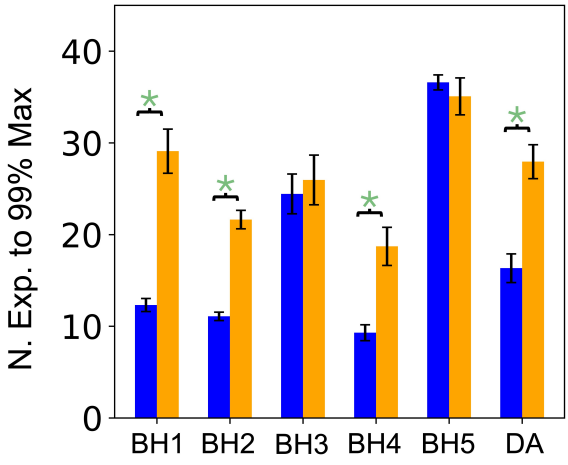


Figure S5: Comparing the average number of experiments needed to identify experimental conditions that give 99% of the maximum observed yield using the LLM-EI (blue) and EI (yellow) acquisition functions for datasets BH1-BH5 and DA. Refer to the caption of Figure 3 for additional details regarding significance testing.

6.3 Gaussian Process Surrogate Model Developed by Shields et al.[37]

The Gaussian process surrogate model used in this work follows the formulation introduced by Shields et al.[37]. The reaction outcomes were standardized to zero mean and unit variance and modeled by a constant mean Gaussian process with a Matérn kernel of shape parameter $v = 5/2$. The kernel function is given by:

$$k_{Matérn52}(r) = \alpha(1 + \frac{\sqrt{5}r}{l} + \frac{5r^2}{3l^2}) \exp(-\frac{\sqrt{5}r}{l}) \tag{4}$$

where α is an output scale parameter, $r = \sqrt{\sum_{i=1}^{n}(\mathbf{x}_{1i} - \mathbf{x}_{2i})^2}$ is the distance between two input points ($\mathbf{x}_1$ and $\mathbf{x}_2$), and l is the characteristic length scale parameter.

In addition, automatic relevance determination [48] was employed by learning a separate length scale for each input dimension. Gamma priors favoring longer length scales were placed on the length-scale parameters to regularize the model in the low-data regime. Output variance and observation noise variance were learned during training. All kernel and likelihood hyperparameters were optimized by maximizing the log marginal likelihood using stochastic gradient-based optimization, as implemented in GPyTorch[15]

6.4 Pairwise Gaussian Process Preference Model

The Preference learning was performed using a Pairwise Gaussian process (PairwiseGP) model provided in the BoTorch[**BoTorch**] library, which is an approach introduced by Chu and Ghahramani[9]. A latent function $g(x)$ is assumed, and observed preferences are modeled as comparisons between pairs of inputs, which are the reaction conditions preference generated by the LLMs survey. Specifically, when one reaction condition x_j is preferred than another x_k, the model assumes that the latent value of the preferred condition is larger $g(x_j) > g(x_k)$.

In this present work, the kernel is set as the default Scaled RBF kernel defined in PairwiseGP, and the model employs the `PairwiseLikelihood`, $\frac{g(x_j)-g(x_k)}{\sqrt{2}\sigma}$, where the σ is implicitly set to 1 as the function is scaled by the implemented kernel. Posterior inference over the latent function was carried out using a Laplace approximation, as implemented via `PairwiseLaplaceMarginalLogLikelihood` in BoTorch. This approximation enables training of the preference model by optimizing a marginal log likelihood objective that accounts for uncertainty in pairwise comparison data.

View Article Online
DOI: 10.1039/D6DD00052E

References

[1] M. Adachi et al. "Looping in the Human Collaborative and Explainable Bayesian Optimization". In: *arXiv preprint arXiv:2310.17273* (2023).

[2] V. Aglietti et al. "Funbo: Discovering acquisition functions for bayesian optimization with funsearch". In: *arXiv preprint arXiv:2406.04824* (2024).

[3] M. Aldeghi et al. "Golem: an algorithm for robust experiment and process optimization". In: *Chemical Science* 12.44 (2021), pp. 14792–14807.

[4] T. Bai et al. "Transfer learning for Bayesian optimization: A survey". In: *arXiv preprint arXiv:2302.05927* (2023).

[5] J. Bergstra, D. Yamins, and D. Cox. "Making a Science of Model Search: Hyperparameter Optimization in Hundreds of Dimensions for Vision Architectures". In: *Proceedings of the 30th International Conference on Machine Learning*. Ed. by S. Dasgupta and D. McAllester. Vol. 28. Proceedings of Machine Learning Research 1. Atlanta, Georgia, USA: PMLR, 17–19 Jun 2013, pp. 115–123.

[6] J. Bergstra et al. "Algorithms for Hyper-Parameter Optimization". In: *Advances in Neural Information Processing Systems*. Ed. by J. Shawe-Taylor et al. Vol. 24. Curran Associates, Inc., 2011.

[7] K. Chen et al. "Chemist-X: Large language model-empowered agent for reaction condition recommendation in chemical synthesis". In: *arXiv preprint arXiv:2311.10776* (2023).

[8] M. Christensen et al. "Data-science driven autonomous process optimization". en. In: *Communications Chemistry* 4.1 (Aug. 2021), p. 112.

[9] W. Chu and Z. Ghahramani. "Preference learning with Gaussian processes". In: *Proceedings of the 22nd international conference on Machine learning* (2005).

[10] C. W. Coley, W. H. Green, and K. F. Jensen. "Machine Learning in Computer-Aided Synthesis Planning". In: *Accounts of Chemical Research* 51.5 (2018). PMID: 29715002, pp. 1281–1289.

[11] A. F. De Almeida, R. Moreira, and T. Rodrigues. "Synthetic organic chemistry driven by artificial intelligence". In: *Nature Reviews Chemistry* 3.10 (2019), pp. 589–604.

[12] Q. Dong et al. "A Survey on In-context Learning". In: *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*. Ed. by Y. Al-Onaizan, M. Bansal, and Y.-N. Chen. Miami, Florida, USA: Association for Computational Linguistics, Nov. 2024, pp. 1107–1128.

[13] M. Feurer et al. "Practical transfer learning for bayesian optimization". In: *arXiv preprint arXiv:1802.02219* (2018).

[14] J. Fürnkranz and E. Hüllermeier. "Preference Learning: An Introduction". In: *Preference Learning*. Ed. by J. Fürnkranz and E. Hüllermeier. Berlin, Heidelberg: Springer Berlin Heidelberg, 2011, pp. 1–17. ISBN: 978-3-642-14125-6.

[15] J. Gardner et al. "Gpytorch: Blackbox matrix-matrix gaussian process inference with gpu acceleration". In: *Advances in neural information processing systems* 31 (2018).

[16] D. E. Graff et al. "Self-Focusing Virtual Screening with Active Design Space Pruning". In: *Journal of Chemical Information and Modeling* 62.16 (2022). PMID: 35938299, pp. 3854–3862.

[17] J. Guo, B. Ranković, and P. Schwaller. "Bayesian optimization for chemical reactions". en. In: *Chimia (Aarau)* 77.1-2 (Feb. 2023), pp. 31–38.

[18] T. Guo et al. "What can large language models do in chemistry? a comprehensive benchmark on eight tasks". In: *Advances in Neural Information Processing Systems* 36 (2023), pp. 59662–59688.

[19] R. J. Hickman et al. "Equipping data-driven experiment planning for Self-driving Laboratories with semantic memory: case studies of transfer learning in chemical reaction optimization". In: *Reaction Chemistry & Engineering* 8.9 (2023), pp. 2284–2296.

[20] J. Howard and S. Ruder. "Universal Language Model Fine-tuning for Text Classification". In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Ed. by I. Gurevych and Y. Miyao. Melbourne, Australia: Association for Computational Linguistics, July 2018, pp. 328–339.

View Article Online
DOI: 10.1039/D6DD00052E

[21] C. Hvarfner et al. "πBO: Augmenting Acquisition Functions with User Beliefs for Bayesian Optimization". In: *arXiv preprint arXiv:2204.11051* (2022).

[22] K. M. Jablonka et al. "Is GPT all you need for low-data discovery in chemistry?" In: *ChemRxiv* (2023).

[23] K. M. Jablonka et al. "Leveraging large language models for predictive chemistry". en. In: *Nat. Mach. Intell.* 6.2 (Feb. 2024), pp. 161–169.

[24] R. Jacobs et al. "Regression with large language models for materials and molecular property prediction". In: *arXiv preprint arXiv:2409.06080* (2024).

[25] L. V. Jospin et al. "Hands-On Bayesian Neural Networks—A Tutorial for Deep Learning Users". In: *IEEE Computational Intelligence Magazine* 17.2 (May 2022), pp. 29–48. ISSN: 1556-6048.

[26] A. Kristiadi et al. "A sober look at LLMs for material discovery: Are they actually good for Bayesian optimization over molecules?" In: *arXiv preprint arXiv:2402.05015* (2024).

[27] Q. Liang et al. "Benchmarking the performance of Bayesian optimization across multiple experimental materials science domains". en. In: *Npj Comput. Mater.* 7.1 (Nov. 2021).

[28] T. Liu et al. "Large language models to enhance bayesian optimization". In: *arXiv preprint arXiv:2402.03921* (2024).

[29] A. M Bran et al. "Augmenting large language models with chemistry tools". en. In: *Nat. Mach. Intell.* 6.5 (May 2024), pp. 525–535.

[30] V. Nguyen et al. "Filtering Bayesian optimization approach in weakly specified search space". en. In: *Knowl. Inf. Syst.* 60.1 (July 2019), pp. 385–413.

[31] M. C. Ramos et al. "Bayesian optimization of catalysts with in-context learning". In: *arXiv preprint arXiv:2304.05341* (2023).

[32] B. Ranković and P. Schwaller. "BoChemian: Large language model embeddings for Bayesian optimization of chemical reactions". In: *NeurIPS 2023 Workshop on Adaptive Experimental Design and Active Learning in the Real World*. 2023.

[33] C. E. Rasmussen. "Gaussian Processes in Machine Learning". In: *Advanced Lectures on Machine Learning: ML Summer Schools 2003, Canberra, Australia, February 2 - 14, 2003, Tübingen, Germany, August 4 - 16, 2003, Revised Lectures*. Ed. by O. Bousquet, U. von Luxburg, and G. Rätsch. Berlin, Heidelberg: Springer Berlin Heidelberg, 2004, pp. 63–71. ISBN: 978-3-540-28650-9.

[34] Y. Ruan et al. "An automatic end-to-end chemical synthesis development platform powered by large language models". en. In: *Nat. Commun.* 15.1 (Nov. 2024), p. 10160.

[35] P. Schwaller et al. "Molecular Transformer: A Model for Uncertainty-Calibrated Chemical Reaction Prediction". In: *ACS Central Science* 5.9 (2019). PMID: 31572784, pp. 1572–1583.

[36] B. Shahriari et al. "Taking the Human Out of the Loop: A Review of Bayesian Optimization". In: *Proceedings of the IEEE* 104.1 (2016), pp. 148–175.

[37] B. J. Shields et al. "Bayesian reaction optimization as a tool for chemical synthesis". en. In: *Nature* 590.7844 (Feb. 2021), pp. 89–96.

[38] B. J. Shields et al. *Buchwald-Hartwig Dataset*. https://github.com/b-shields/edbo/blob/master/experiments/data/aryl_amination/experiment_index.csv [Accessed: 2024-06-01]. 2021.

[39] B. J. Shields et al. *Direct-Arylation Dataset*. https://github.com/b-shields/edbo/blob/master/experiments/data/direct_arylation/experiment_index.csv [Accessed: 2024-06-01]. 2021.

[40] E. Shim et al. "Machine Learning Strategies for Reaction Development: Toward the Low-Data Limit". In: *Journal of Chemical Information and Modeling* 63.12 (2023). PMID: 37312524, pp. 3659–3668.

[41] S. Singh and R. B. Sunoj. "A transfer learning protocol for chemical catalysis using a recurrent neural network adapted from natural language processing". en. In: *Digital Discovery* 1.3 (2022), pp. 303–312.

View Article Online
DOI: 10.1039/D6DD00052E

[42] A. Souza et al. "Bayesian optimization with a prior for the optimum". In: *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer. 2021, pp. 265–296.

[43] K. Swersky, J. Snoek, and R. P. Adams. "Multi-task bayesian optimization". In: *Advances in neural information processing systems* 26 (2013).

[44] P. Tighineanu et al. "Transfer learning with gaussian processes for bayesian optimization". In: *International conference on artificial intelligence and statistics*. PMLR. 2022, pp. 6152–6181.

[45] J. Van Herck et al. "Assessment of fine-tuned large language models for real-world chemistry and material science applications". en. In: *Chem. Sci.* 16.2 (Jan. 2025), pp. 670–684.

[46] M. Volpp et al. "Meta-learning acquisition functions for transfer learning in bayesian optimization". In: *arXiv preprint arXiv:1904.02642* (2019).

[47] K. Wang and A. W. Dowling. "Bayesian optimization for chemical products and functional materials". In: *Current Opinion in Chemical Engineering* 36 (2022), p. 100728. ISSN: 2211-3398.

[48] C. K. Williams and C. E. Rasmussen. *Gaussian processes for machine learning*. Vol. 2. 3. MIT press Cambridge, MA, 2006.

[49] M. Wistuba, N. Schilling, and L. Schmidt-Thieme. "Hyperparameter search space pruning – A new component for sequential model-based hyperparameter optimization". In: *Machine Learning and Knowledge Discovery in Databases*. Lecture notes in computer science. Cham: Springer International Publishing, 2015, pp. 104–119.

[50] M. Wistuba, N. Schilling, and L. Schmidt-Thieme. "Scalable gaussian process-based transfer surrogates for hyperparameter optimization". In: *Machine Learning* 107.1 (2018), pp. 43–78.

[51] W. Xu et al. "Principled Bayesian optimization in collaboration with human experts". In: *Advances in Neural Information Processing Systems* 37 (2024), pp. 104091–104137.

[52] Y. Zeng et al. "Large Scale Multi-Task Bayesian Optimization with Large Language Models". In: *arXiv preprint arXiv:2503.08131* (2025).

[53] C. Zhang et al. *Amide Coupling Datasets*. https://github.com/aichemeco/amide_coupling/blob/main/data/all_HTE_with_condition.csv [Accessed: 2025-10-01]. 2025.

[54] C. Zhang et al. "Intermediate knowledge enhanced the performance of the amide coupling yield prediction model". In: *Chemical Science* (2025).

[55] Z. Zhou, X. Li, and R. N. Zare. "Optimizing Chemical Reactions with Deep Reinforcement Learning". In: *ACS Central Science* 3.12 (2017). PMID: 29296675, pp. 1337–1344.

View Article Online
DOI: 10.1039/D6DD00052E

This study was carried out using publicly available data from Shields et al. (Nature, 2021, 590, 89-96) at (https://doi.org/10.1038/s41586-021-03213-y) and Zhang et al. (Chem. Sci., 2025,16, 11809-11822) at (https://doi.org/10.1039/D5SC03364K). The code for this article is available on GitHub. The most recent version of the code and datasets are available at 10.5281/zenodo.18882258.