# Towards Self-Driving Labs for Heterogeneous Catalysis: A Bayesian Optimization Approach

Markus GRIMM[1,2] Pierre CHAINAIS[1] Sebastien PAUL[2]

[1]Univ. Lille, CNRS, Centrale Lille, UMR 9189 CRIStAL, F-59000 Lille, France.

[2]Univ. Lille, CNRS, Centrale Lille, Univ. Artois, UMR 8181 - UCCS - Unité de Catalyse et Chimie du Solide, F-59000 Lille, France.

**Résumé –** La catalyse hétérogène implique des interactions des réactifs sur la surface d'un catalyseur solide, qui nécessitent des méthodes d'optimisation pour maximiser le rendement d'une réaction. Nous proposons une approche d'optimisation bayésienne adaptée sur une plateforme d'expérimentation automatisée et robotisée à haut débit pour optimiser le rendement du produit en termes de débit des réactifs et de charge du catalyseur dans une réaction catalysée de manière hétérogène. En 10 expériences, nous déterminons les conditions optimales, à savoir le débit des réactifs et la masse du catalyseur. Nos résultats fournissent un plan d'action pour des laboratoires auto-pilotés accélérés et efficaces dans la catalyse hétérogène.

**Abstract –** Heterogeneous catalysis involves interactions of reactants on the surface of a solid catalyst, which require optimization methods to maximize the yield of a reaction. We propose an adopted Bayesian optimization approach on an automated and robotic high-throughput experimentation platform to optimize the product yield in terms of reactant flow rate and catalyst loading of a heterogeneously catalyzed reaction. Within 10 experiments, we determine optimal conditions, namely reactant flow rate and catalyst mass. Our results provide a roadmap for accelerated and efficient self-driving labs in heterogeneous catalysis.

## 1 Introduction

Heterogeneous catalysis plays a major role in enabling cost-effective production of various chemicals [2]. Generally this involves a solid catalyst that provides a surface with so-called active sites and a non-condensed reactant that reacts on the catalyst surface. Catalysis is a process in which a substance, called a catalyst, accelerates a chemical reaction without being consumed by reducing the energy needed for the reaction to occur. The global market size of this industry was roughly 34 billion USD in 2019 [10]. However, optimizing the reaction conditions to maximize product yield is still mainly an empirical approach that relies on inefficient budget intensive trial and error or design of experiments methods [2].

Closed-loop laboratory systems employing optimization algorithms, such as Bayesian Optimization (BO), have demonstrated success in various fields [3]. These algorithms are particularly useful in optimizing complex processes and selecting the best parameters for specific tasks. In recent years, there has been a growing interest in applying these optimization methods across different domains in chemistry. For example, BO has been effectively used in optimizing physicochemical features and reaction yields [3]. In heterogeneous catalysis, several studies have focused on incorporating sequential optimization methods. BO was used to maximize product yields and improve catalyst properties [6, 5, 8]. These advancements showcase the potential for combining optimization algorithms in various chemistry related fields.

This work results from a collaboration with REALCAT [7]. REALCAT is a state-of-the-art high-throughput experimentation (HTE) robotic-supported platform. High-throughput experimentation enables parallel testing of multiple catalysts and reaction conditions, accelerating catalyst development while reducing the number of experiments and resource consumption to optimize a reaction and its catalyst. REALCAT is specialized in the development of heterogeneous catalysts. It is equipped with advanced high-throughput technologies, including robots with freely movable arms for scheduled automatic parallel synthesis, parallel continuous and batch reactors, and linked analytical devices. Those analytical devices enable online assessment of catalyst performance. REALCAT offers unique capabilities for efficient catalyst development in terms of different parameters. Those parameters can either be continuous or discrete. This results in the possibility to perform several experiments in different conditions in parallel within a programmable set-up. Incorporating signal processing methods into the REALCAT platform, which is recognized as one of the few labs in the world with such extensive resources while belonging to the French "équipements d'excellence" (EQUIPEX) category, is the goal of our work. It aims to further improve the efficiency of heterogeneous catalyst development.

We propose a protocol to optimize the product yield in a heterogeneously catalyzed reaction using the REALCAT platform. Our approach optimizes working conditions, namely the reactant flow rate and catalyst mass. Our protocol consists of the circular integration of REALCAT's high-throughput testing capabilities, gas chromatography (GC), mass spectrometry (MS), and BO. The integration of HTE and BO reduces the number of required experiments significantly to find the optimal working conditions by automatically suggesting a sequence of test conditions. This sequence is guided by a balanced exploration-exploitation strategy. This results in the decrease of optimum search time. Additionally, leveraging faster MS measurements with slower GC measurements enables efficient online catalyst performance monitoring. This reduces measurement time and therefore, lowering the overall experimentation time immensely. Our method represents a significant advancement towards closed-loop automated laboratories for heterogeneous catalysis. It serves as a proof of concept for accelerated lab automation which is crucial for increased efficiency in terms of economic profitability and environmental concerns.

# 2 Proposed Approach

We propose an accelerated optimization protocol that combines the HTE REALCAT capabilities, MS and GC measurements, and BO for optimizing the product yield of the ODHP reaction in terms of the reactant flow rate and the catalyst mass.

## 2.1 Experimental Setting

The HTE unit on REALCAT consists of 16 continuous batch flow reactors which allow for parallel condition testing of various discrete catalyst masses and allows scheduling experimental settings. Measurement duration is a significant factor for the total experimentation time. Hence, we propose the combination between the MS and GC analytical devices to determine the yield of the reaction in an accelerated manner. MS measurements need 10 seconds to measure the yield of the reaction while GC measurements take 10 minutes. MS measurements need GC measurements for calibration. More GC measurements lead to more accurately calibrated MS measurements. A trade-off between the fast MS measurements and the number of slow GC measurements is necessary.

By integrating BO with the HTE testing unit from REALCAT, we automatically generate a sequence of few as possible conditions to maximize the product yield. Furthermore, by combining MS measurements with GC measurements, we are able to significantly reduce measurement time. To demonstrate our protocol's effectiveness, we introduce a model oxidative-dehydrogenation of propane (ODHP) reaction to validate the successful optimization of working conditions for a heterogeneously catalyzed reaction. This reaction is of significant economic importance in various industries [4]. Thus, this reaction has been extensively researched, leading to the identification of numerous catalysts. Hence, it is an ideal model reaction to showcase our protocol. In this work, we employ the catalyst reported by [9].

The experimental setup involves a circular and iterative process between the HTE unit, a GC, MS, and BO software. In this setup, the BO-suggested condition is manually set in the HTE unit, the yield determined via MS and GC measurements is then passed to the BO software for iterative optimization until the budget is depleted. Additionally, our protocol ensures reproducibility, addressing a common challenge in heterogeneous catalysis.

## 2.2 Bayesian Optimization

In the following we give a brief introduction to Bayesian Optimization (BO) using Gaussian processes (GP) as presented in [1]. BO is used to optimize expansive, black box objective functions like the unknown yield function of the ODHP reaction in a sample-efficient gradientless manner. It is based on a sequential optimization approach and is using a probabilistic surrogate model like GP to optimize black-box objective functions. The GP is characterized by a predictive mean function $\mu(x)$ and a covariance function $\kappa(x, x')$. It defines a distribution over functions and allows optimization within a probabilistic framework. The prior distribution is sequentially updated as new data becomes available to provide a posterior distribution that measures uncertainty about the objective function. The distribution, i.e. the GP of the functions $f$ is given by:

$$f \sim \mathcal{GP}(\mu, \kappa), \tag{1}$$

where $f(x)$ is distributed normally as $\mathcal{N}(\mu(x), \kappa(x, x))$ for all $x \in X$. BO iteratively suggests new points to sample by maximizing an acquisition function $\alpha(x)$ given the state of the current surrogate model $f_{1:t-1}$ that balances exploration and exploitation:

$$x_t = \arg\max_{x \in X} \alpha(x \mid f_{1:t-1}) \tag{2}$$

Common acquisition functions include upper confidence bound (UCB) with a tunable $\lambda$:

$$UCB(x) = \mu(x) + \lambda\sigma(x). \tag{3}$$

The acquisition function selects points based on the surrogate model's posterior distribution, which is updated with each new evaluation of the objective function by taking the Prior $P(f)$ and the likelihood $P(D_{1:t} \mid f)$ of our data given our surrogate model $f$:

$$P(f \mid D_{1:t}) = P(D_{1:t} \mid f)P(f) \tag{4}$$

BO has a finite budget of evaluations, and the goal is to find the optimal point within that budget. Hence, BO is an appropriate sample efficient sequential optimization algorithm that is used in our proposed protocol.

## 2.3 HTE Sequential Optimization Approach

We aim to maximize the product yield of the reaction. The HTE unit presents technical constraints that limit the optimization process. The catalyst masses in the 16 parallel reactors are set at the beginning and cannot be changed during the optimization process. Additionally, the propane flow rate must be the same across all reactors, meaning that the individual propane flow rate cannot be set individually per reactor.

The experimental design for optimization of the catalyst mass will therefore rely on a regular sampling of the presumed interval of masses. Let $M = \{m_1, \ldots, m_{16}\}$ be the discrete set of catalyst masses for the 16 reactors with $i \in I = [1, \ldots 16]$, and let $R = [5.5\text{ml/min}, 50\text{ml/min}]$ represent the continuous range of propane flow rates. We can express the optimization problem as finding the optimal discrete catalyst mass $m_i \in M$ and the continuous propane flow rate $r \in R$ in order to maximize the yield:

$$x^* = \text{argmax}_{\boldsymbol{x} \in \mathcal{X}} F(\boldsymbol{x}) \tag{5}$$

Here, $F : \mathcal{X} \subset \mathbb{R}^2 \to [0, 1]$ is the product yield function with $\boldsymbol{x} \mapsto y(\boldsymbol{x})$ where $y(\boldsymbol{x})$ is the measured yield at a flow $r$ and mass $m_i$ with $\mathcal{X} = M \times R$.

We tackle this problem by optimizing the flow rate on the fly and the mass a posteriori. That is, we define an updated yield function $F' : R \subset \mathbb{R} \to [0, 1]$, with $r \mapsto F'(r) = \max_{m_i \in M} F_{m_i}(r)$. We can optimize $F'$ by formulating a sequential optimization problem by using a surrogate model $f'_{1:t-1}$, which reflects our belief about the underlying updated yield function $F'$. In (6) we present the transformed sequential optimization problem, which we solve on the fly:

$$r_t = \arg\max_{r \in R} f'_{1:t-1}(r) \tag{6}$$

At each step $t \in 1, \ldots, N$, where $N$ represents our budget of flows which we can test, the surrogate model is updated with the results of the most recent experiment. It permits to iteratively refine our estimate of $F'$. After determining the best flow rate $r^*$ based on (6), we can a posteriori determine the best mass. I.e., the optimal mass and flow rate combination, denoted by $\boldsymbol{x}^*$, is determined by choosing the pair $(m_i, r_{1:t})$ that maximizes the measured yields $\boldsymbol{y}_{i,1:t}(\boldsymbol{x})$ for all possible masses $m_i$ and tested flow rates $r_{1:t}$ $\forall i, t$.

In order to solve the optimization problem (5) in a sequential manner as in (6), we propose using Bayesian optimization with the UCB acquisition function to represent the exploration-exploitation trade-off in 1:

**Algorithm 1:** HTE adapted Bayesian Optimization

```
for $t$=$1, \ldots, N$ do
    $r_t$=optimize acquisition function $\alpha(r|f'_{1:t-1})$,
    $\boldsymbol{y}_{i,t} = F(\boldsymbol{x}_t)$, query at flow $r_t$ for $m_i, \forall i$.
    $D_t = D_{1:t-1} \cup (x_t, \max(\boldsymbol{y}_{i,t}))$, update data set,
    $P(f'|D_{1:t}) = P(D_{1:t}|f')P(f')$, update surrogate.
end
$\boldsymbol{x}^* = \text{argmax}_{x \in \{m_i, r_{1:t}\}} \boldsymbol{y}_{i,1:t}(\boldsymbol{x})$, max yield $\forall i, t$
```

In summary, the initial optimization problem is transformed into a sequential optimization problem that takes the process constraints into account by optimizing the reactant flow rate on the fly over a grid of possible masses. The optimal mass is determined a posteriori among tested masses for the optimal flow rate. This approach yields an efficient and adaptive optimization process that respects the constraints of the HTE REALCAT unit while maximizing the product yield of the reaction of interest.

# 3 Experiments

We conducted three sets of experiments utilizing REALCAT's HTE unit in conjunction with the MS and GC. Before the experiments started 14 different catalyst masses were selected as loading in the reactors. Accuracy checks were performed on 2 reactors, which were loaded with inert silica oxide. Therefore, this procedure allows to test as many masses as possible while guarantying accurate measurements. These loadings were used for all subsequent experiments.

In the first experiment set, we focused on determining the optimal number of GC measurements required for accurate transformation of the MS measurements while minimizing the measurement time. We found that using 16 GC measurements resulted in the lowest mean absolute error (MAE) of 0.10%. However, by reducing the number of GC measurements to 8 (one for every second reactor), the MAE only slightly increased to 0.25%. This trade-off between accuracy and measurement time effectively lowers the measurement duration by 50%, while maintaining an acceptable level of accuracy in transforming MS measurements into volume percentages. Further reduction in the number of GC measurements to 5 or 2 led to higher MAE values, indicating a decrease in the accuracy of the transformation. This demonstrates that using one GC measurement for every second reactor is a suitable trade-off between maintaining acceptable accuracy in transforming MS measurements into volume percentages and reducing the measurement time.

In the second set, we optimized the propane flow rate by implementing the recommendations from the BO software calling these experiments the BO run. We examined 10 different flow rates at a constant temperature of 560 °C, starting with an initial flow rate of 10 ml/min chosen uniformly at random. Due to time constraints, we tested 10 flow rates, as the experiments took one week to perform. To validate the results' reproducibility, we conducted a final set of experiments with evenly spaced propane flow rates. GC measurement was employed for every second reactor for the MS calibration to ensure accuracy during the BO run while for the validation run all reactors were measured via GC.

# 4 Results and Discussion

We optimize the product yield of the ODHP reaction using the REALCAT HTE unit with a budget of only 10 tested pan flow rate conditions. Figure 1 shows all the measurements from the BO-guided run and the validation test set. For both runs, we present the maximum yield of all reactors across all tested conditions. Additionally, we indicate the order of the BO run proposed conditions tested. By the sixth iteration, we achieve a yield that is within 0.5% of the actual optimum yield of 8%, with a propane flow rate that differs by roughly 2 ml/min. Figure 1 shows that, after the complete budget was used, the BO results are in accordance with the validation results. This indicates that the final belief of the underlying objective function was replicated during the BO run.

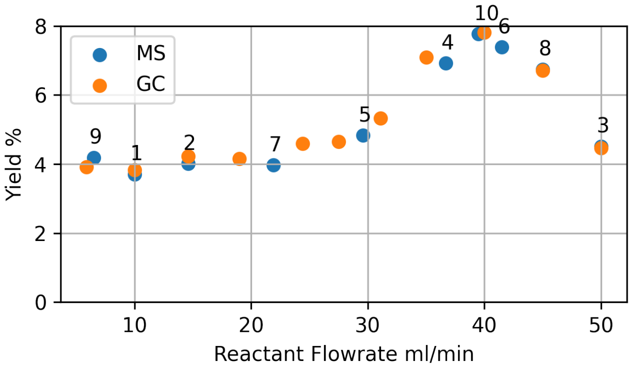

Figure 1 – Comparison BO run and validation run

Figure 2 shows intermediate results of the surrogate model and the acquisition function during the optimization procedure. It illustrates the Gaussian process used to model the underlying yield function with the corresponding acquisition function after 4 propane flow rates tested. The surrogate model slightly resembles a parabolic form, thus indicating that the underlying objective function has a maximum. Additionally, the maximum of the UCB acquisition function suggests a new reactant flow rate. Figure 2 presents the updated state of the mean function of the GP surrogate model at iteration $t = 5$ with the additional tested reactant flow rate that was proposed

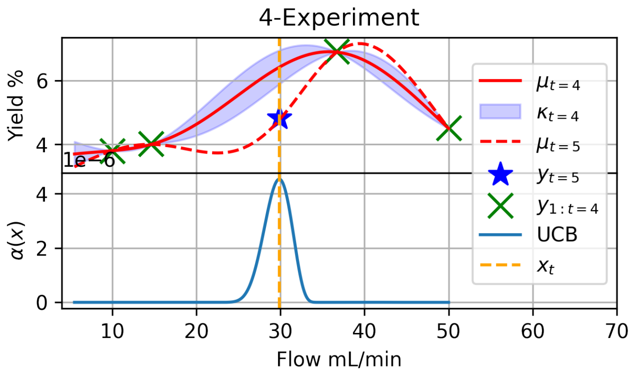

Figure 2 – Evolution of the GP from iteration $t = 4$ to iteration $t = 5$ and the state of UCB at iteration $t = 4$. The solid red line represents the predictive mean function $\mu_{t=4}$ at iteration 4, while the dashed red line shows the updated mean function $\mu_{t=5}$ after querying the suggested flow at the yellow dashed line. The yellow dashed line indicates the maximum of the UCB function in blue. The shaded blue area represents the variance $\kappa_{t=3}$ of the GP at iteration $t = 4$. Green × symbols denote the first four measured MS results, and the 5th measured MS result is illustrated as a blue star.

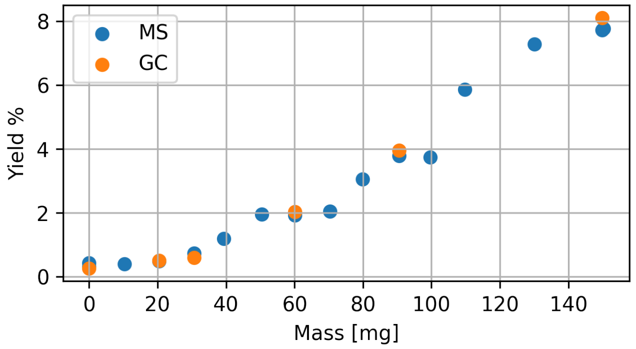

Figure 3 – Reactor loadings and propylene yield of MS and GC measurements at optimal propane flow rate $r = 39.5$ml/min of the BO run.

by maximizing UCB at iteration $t = 4$. The new tested flow updates the belief of the underlying yield function and thus updates the surrogates model mean function from $\mu_4$ to $\mu_5$.

The Bayesian optimization procedure is generally successful in updating our belief of the underlying yield function at each iteration. The UCB acquisition function guides the selection of the next tested pan flow rate condition. Besides determining the optimal flow the optimal catalyst mass of 150 mg was successfully determined.

Figure 3 illustrates the measured yields of all reactors at the optimal propane flow rate $r = 39.5$ml/min. A steady increase of the yield almost linearly in dependence of the catalyst mass can be observed. At a mass around 0 a Propylene yield of approximately 0% is measured while the highest yield is measured at the maximum loading of 150 mg. Since the number of active sites linearly correlates with catalyst mass and influence the yield of a catalyst, it is reasonable to see that the highest mass, i.e. highest number of active sites has the highest yield.

Our results demonstrate the effectiveness of the proposed Bayesian optimization approach for optimizing propene yield of the ODHP reaction in dependence of propane flow rate and catalyst loading via a semi automized experimental setting using the HTE unit in combination with the MS and GC.

## 5 Conclusion

We present a semi-automated Bayesian optimization protocol on a robotic high-throughput experimentation platform, efficiently identifying optimal working conditions for a heterogeneously catalyzed reaction. The protocol combines GC-MS measurements to reduce measurement time and uses a minimum number of experiments to find the optimum working conditions based on the integration of BO into HTE. We successfully optimize catalyst loading and reactant flow rate using UCB acquisition function.

Our results demonstrate the effectiveness of this protocol in optimizing the product yield for the ODHP reaction using the HTE facilities from REALCAT. Future work will focus on adapting sequential optimization approaches to fully exploit Gaussian process properties, enabling exploration of a broader range of conditions, such as temperature. Additionally, once an API becomes available for the HTE unit, we aim to fully automate the optimization process.

**Acknowledgement:** This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie grant agreement number 847568. This work is supported by the ANR project Chaire IA Sherlock ANR-20-CHIA-0031-01.

## References

[1] Eric Brochu, Vlad M. Cora, and Nando de Freitas. A tutorial on bayesian optimization of expensive cost functions, with application to active user modeling and hierarchical reinforcement learning. *ArXiv*, abs/1012.2599, 2010.

[2] Avelino et al. Corma. Heterogeneous catalysis: Understanding for designing, and designing for applications. *Angewandte Chemie International Edition*, 55(21):6112–6113, 2016.

[3] Christensen Melodie et al. Data-science driven autonomous process optimization. *Communications Chemistry*, 4(1):112, 2021.

[4] Gambo Yahya et al. Catalyst design and tuning for oxidative dehydrogenation of propane – A review. *Applied Catalysis A: General*, 609(July 2020):117914, 2021.

[5] Junya Ohyama et al. Bayesian-optimization-based improvement of cu-cha catalysts for direct partial oxidation of ch4. *Journal of Physical Chemistry C*, 126:19660–19666, 2022.

[6] Keisuke Takahashi et al. The rise of catalyst informatics: Towards catalyst genomics. *ChemCatChem*, 11:1146–1152, 2019.

[7] Sébastien Paul et al. Realcat : A new platform to bring catalysis to the lightspeed to cite this version : Hal id : hal-01772468. *Oil Gas Sci. Technol. - Rev. IFP Energies nouvelles*, 2018.

[8] Wenjin Yan et al. Bayesian migration of gaussian process regression for rapid process modeling and optimization. *Chemical Engineering Journal*, 166:1095–1103, 2011.

[9] Zhou Hang et al. Isolated boron in zeolite for oxidative dehydrogenation of propane. *Science*, 372(6537):76–80, 2021.

[10] Xijun et al. Hu. Heterogeneous catalysis: Enabling a sustainable future. *Frontiers in Catalysis*, 1:667675, 2021.