Customizing Bayesian Optimization Algorithms for Chemical Research

Machine learning (ML) models are undoubtedly paving a way toward intelligent chemistry systems, establishing themselves as a primary research focus in process systems engineering. Among them, Bayesian optimization (BO) — a self-optimization method that expedites the screening of experimental parameters — has particularly excelled in practical applications.

The intersection of ML and chemical engineering presents exciting opportunities for researchers across diverse disciplines. Chemists, the primary end-users, prioritize the convenience and specialization of these algorithms. Thus, it is desirable to package the algorithm as a user-friendly toolbox that is accessible to all users, regardless of their familiarity with the underlying principles and details. Moreover, there is a growing emphasis on integrating more domain knowledge to enhance applicability, rather than using generic statistical ML models. Historically, a lack of familiarity with chemistry has prohibited ML practitioners from integrating this domain knowledge. In the March *AIChE Journal* article, "CAPBO: A Cost-Aware Parallelized Bayesian Optimization Method for Chemical Reaction Optimization," Zhihong Yuan (Tsinghua Univ.) and his coauthors introduced a tailored BO algorithm designed explicitly for chemists.

BO has gained prominence as a black-box optimization approach for addressing challenges in chemical reaction optimization. Implementing BO is relatively straightforward: The model proposes a new experiment point based on existing observations, augments the dataset upon obtaining the experimental result, and repeats the optimization loop. The developed algorithm further considers the characteristics of chemical research from two aspects:

Cost-aware strategies. The original aim of employing BO in reaction optimization was to obtain the highest possible objective function value with the fewest experiments. However, the number of experiments may not be an ideal evaluation. Perhaps experimenters intend to save experimental expenses, but the relationship between these two terms is not linear. For example, a single experiment could require a higher flowrate than several smaller experiments, requiring more reagent consumption and thus more experimental expenses. The concept of experimental costs was introduced as the evaluation in place of the number of experiments, which can refer to total experimental duration, consumption of reagents, total experimental expenses, etc. This means the experimental cost can refer to any quantifiable concern of experimenters and can be selected flexibly. Considering the experimental cost as the number of experiments is only one special case.

Parallelization. Standard BO follows a strictly serial process, where a single reactor is used for continuous, repeated experiments. The total experimental duration for optimization may be prohibitively expensive, requiring an improved asynchronous parallelized strategy. In this way, combining the algorithm with high-throughput screening technology to automatically allocate experimental tasks for multiple reactors enhances screening efficiency.

Additionally, integrating parallelization and cost-aware strategies synergistically addresses the challenges associated with hyperparameter selection, as detailed in the article.

"This is a generalized chemical reaction optimization framework," says Yuan. "To develop a complete specialized ML toolbox for chemistry, this work can be regarded as a solid start, but there is still a long journey ahead." One of the most significant limitations of the current application of ML models is insufficient deep analysis, particularly concerning optimization algorithms. It is necessary to build a more direct bridge between ML and chemistry with a comprehensive understanding of underlying mathematical knowledge and chemical practical experience. Indeed, Yuan underscores the following necessary characteristics for an ML model to be applied in chemical research:

• *Reliability.* Interpretability is crucial for knowledge mining and generalization. The algorithm process is expected to be concise with distinct physical meaning, ideally accompanied by mathematical proofs for convergence.

• *Efficiency.* Multiscale study and validation are recommended to evaluate performance comprehensively. As the lead author, Runzhe Liang (Tsinghua Univ.) designed a series of experiments to explore the behavior of the proposed method, especially the worst-case performance under extreme circumstances. Consistently obtaining optimum performance may be unrealistic, but robustness is vital for adapting to complex chemical systems.

• *Flexibility.* Not every user will be interested in the algorithm's details, so the algorithm should cater to both experts and non-experts. It can be a purely packaged structure, which allows chemists to focus only on input formats, but it should also enable researchers to perform hyperparameter tuning and model secondary development.

These design criteria guide the proposed method, serving as the foundation for developing ML algorithms tailored for chemistry. "ML is not omnipotent, and it should not be overused. We hope that applications of the ML models can truly play an essential role in knowledge discovery and production efficiency improvements. We will strive for it all the time," says Yuan. CEP