A Bayesian Framework Coupling Categorical and Continuous Variables for Accelerated Catalyst Discovery

Yi Zhang[1†], Changquan Zhao[2†], Yulian He[2*], Cheng Hua[2*]

[1]Columbia University.
[2*]Shanghai Jiao Tong University.

*Corresponding author(s). E-mail(s): yulian.he@sjtu.edu.cn; cheng.hua@sjtu.edu.cn;
Contributing authors: yz5195@columbia.edu; chomaster@sjtu.edu.cn;
†These authors contributed equally to this work.

Abstract

Identifying optimal catalyst compositions and reaction conditions is central to catalyst discovery, but remains a formidable challenge due to the vast multidimensional design spaces encompassing both continuous and categorical variables. In this work, we present a Bayesian optimization framework for accelerated catalyst discovery by jointly optimizing discrete and continuous experimental variables in a single model. Our approach introduces a unified Gaussian Process surrogate with a novel spectral mixture kernel combining Gaussian and Cauchy components. By leveraging Bochner's theorem, this kernel captures both smooth local trends and abrupt, non-smooth patterns in the continuous parameter space, which in turn reduces the modeling burden on the discrete variable side. Discrete choices, such as catalyst compositions and supports, are efficiently navigated via a trust-region strategy based on Hamming distance, allowing early broad exploration followed by focused refinement. In the virtual experimentations of important catalytic processes including oxidative coupling of methane and selective catalytic reduction, where both discrete and continuous parameters were involved in the optimization process, the proposed algorithm significantly outperforms state-of-the-art approaches with high robustness, identifying top catalyst recipes and reaction conditions within only several tens of iterations without any prior knowledge. Notably, the combinatorial optimization process was achieved by aggressive initial explorations in the discrete catalyst composition spaces to quickly identify and converge to the optimal catalyst choices, followed by continuous conditions optimization near the optimal regimes. The

proposed methodology should be of high generalisability to accelerate materials discovery in multidimensional experimental design spaces with minimal experimental costs.

Keywords: Bayesian Optimization, Spectral Mixture Kernel

1 Introduction

Catalyst design and optimization represents a central challenge in the development of catalytic processes. While whether in catalyst compositions, synthesis parameters, or reaction condition optimization, researchers often face a vast design space comprising thousands to millions of possible candidates [1–4]. An exhaustive search in such a high-dimensional space via either the typical intuition-based or model-based practices or even the advanced high-throughput experimentations/screening [5–7] is essentially impractical and consuming. In reality where experimental budgets are limited, chemists often seek to sample the vast space using factorial design methods such as orthogonal experiments, such that the dimension can be simplified by systematically deconvoluting factorial effects and interactions. While these methods demand pre-defined experimental matrices that typically derived from literature precedence and chemical intuitions, rendering them infeasible for global optimization. Most importantly, they become inefficient and inflexible in high-dimensional or nonlinear design systems, as is often the case for a typical catalyst optimization task, where co-optimization of categorical and continuous parameters will be involved. For example, in the oxidative coupling of methane (OCM), a reaction that directly upgrades methane to valuable C-C coupled products like ethylene, one must consider a huge combination of discrete catalyst compositional variables (e.g. metal, support, promoter choices) alongside continuous reaction conditional variables (e.g. temperature, flow rate, gas concentrations) for optimal product yields.

Alternatively, data-driven optimization approaches have emerged as potential solutions for accelerating catalyst discovery in complex design spaces. Without the need for deriving an explicit mechanistic model, data-driven methods excel in inferring causality and correlation based on existing data in an efficient manner. Specifically in experimental design, Bayesian Optimization (BO) stands out as an efficient probabilistic global optimization method as often only small sample sizes are available [8–15]. In brief, BO adopts an adapative framework by iteratively building a probabilistic surrogate model of the objectives (e.g., product yields, selectivities) and suggesting the most informative experiment to perform next. By balancing exploration of new regions against exploitation of known good candidates, BO can often find high-performing solutions in only tens of trials, orders of magnitude fewer than unguided searches. This approach has been successfully applied to various chemical and materials optimization tasks, particularly suitable for autonomous experimentations [8, 11].

Despite its success, applying BO to mixed-variable optimization problems involving both discrete and continuous parameters remains a formidable challenge in catalyst discovery. Classical BO methods usually assume a continuous, smooth search space;

the presence of categorical choices breaks this smoothness and renders gradient-based search for optima difficult [16]. In essence, when discrete parameters are involved, the input space becomes non-differentiable and more complex to model. Several researchers have addressed this issue by extending or modifying BO. For example, an early attempt that could handle both variable types was made by Bergstra et al. (2011) [17] using evolutionary strategies (CMA-ES) for continuous parameters and an Estimation of Distribution Algorithm (EDA) for discrete ones, coupled with a tree-structured Parzen estimator to guide the search. More recently, Ru et al. (2020) proposed CoCaBO [18], a hierarchical method where discrete choices are treated as a multi-armed bandit problem and continuous variables are optimized with BO once a discrete choice is fixed. CoCaBO also introduced a kernel that partially integrates discrete and continuous inputs. However, a bandit approach requires pulling each discrete "arm" (catalyst option) at least once, which becomes intractable as the number of options grows exponentially, and it does not guarantee convergence to a global optimum. Wan et al. (2021) developed CASMOPOLITAN [19], adopting the hierarchical framework of CoCaBO but adding a trust-region method for the discrete space. In this approach, the algorithm focuses on a localized subset of discrete choices at a time, dynamically adjusting the neighborhood size in terms of Hamming distance to balance exploration and exploitation in the categorical domain.

Here, we present a Spectral Mixture Kernel Bayesian Optimization (SMKBO) framework that unifies discrete and continuous variables within a Gaussian Process (GP) surrogate model. Our key innovation is a composite spectral mixture kernel that combines two types of spectral components: Gaussian distributions and Cauchy distributions. The motivation for this design comes from Bochner's theorem [20] in harmonic analysis, which guarantees that any stationary kernel can be represented as a mixture of sinusoids with a certain spectral density. By using a mixture of Gaussian and Cauchy spectral density functions, we create a kernel with dual characteristics: the Gaussian component captures smooth, local variations in the response surface, while the Cauchy component, with its heavy tails, captures global, non-smooth behavior and long-range correlations. Intuitively, this means our GP surrogate can model both gentle slopes and sharp jumps in the catalyst performance landscape. This flexibility in the surrogate greatly alleviates the challenge posed by discrete variables. In effect, the continuous kernel components shoulder more of the burden in explaining variability, allowing the discrete kernel components to handle only the truly combinatorial aspects of the problem. By integrating this spectral mixture kernel with a Hamming-distance-based covariance for categorical inputs (inspired by the kernels used in CoCaBO [18] and CASMOPOLITAN [19]), we obtain a unified GP model that treats the entire input (catalyst types + reaction conditions) holistically.

Furthermore, to ensure efficient search in the discrete space of catalyst candidates, our framework incorporates a trust-region strategy adapted to categorical variables. Inspired by how an amateur experimenter might first screen diverse catalyst families and then hone in on the best one, we impose a dynamic neighborhood constraint on the discrete choices rather than freely jumping among all possible catalyst options. Such constraint not only enables rapid identifications of discrete variables with consistently higher performance in an initial broad search, followed by fine-tuning of the

reaction conditions (continuous variables) to maximize the performance, but also sustains exploratory wills by occasionally examining alternative discrete combinations even after convergence to avoid trapping in local optima. Using experimental data collected from high-throughput experimentations for two important catalytic processes including oxidative coupling of methane (12,708 data points for 59 distinct catalyst formulations under various conditions) [21] and urea-selective catalytic reduction (XXX) as virtual test grounds, our SMKBO achieved higher performance with much fewer experiment iterations than state-of-the-art mixed-variable BO algorithms reported to date, with calculation speed remains comparable to simple BO methods. To the best of our knowledge, this is one of the first BO frameworks to jointly optimize over discrete and continuous experiment variables without splitting the problem, such that the correlations between catalyst composition and reaction conditions, as deeply coupled as they always are, can be fully leveraged.

In summary, we introduce a Bayesian optimization framework for mixed discrete-continuous spaces that enables accelerated catalyst discovery. By capturing both local and global patterns in the experimental landscape via a Cauchy-Gaussian spectral mixture kernel, our approach overcomes the traditional hurdles of mixed-variable optimization. We not only improve the efficiency of finding high-performance catalysts for methane oxidative coupling, but also glean valuable strategic insight into how a smart search algorithm navigates complex design spaces. The generality of this framework can be readily applicable to many other mixed-variable optimization domains, including but not limited to drug discovery, material deisgns, robotic motion planning, supply chain network design etc. providing a powerful tool to accelerate innovation in experimental science.

2 Results

2.1 Optimization of catalysts for the oxidative coupling of methane(OCM).

Oxidative coupling of methane is an important class of reactions where methane gas is directly upgraded to high-value products such as ethane and ethylene, often accompanied by by-products such as CO and CO_2. From a chemoinformatics perspective, research in this field is already quite mature, with abundant experimental data. In 2020, Thanh Nhat Nguyen et al. [21] used high-throughput equipment to obtain a total of 12,708 real experimental data points for 59 different catalysts under varying metal ratios and reaction conditions, providing strong data support for our work.

This dataset includes, for each catalyst, information on the metal species and proportions, support type, reaction temperature, and reaction atmosphere. It also records the selectivity and yield of each product as evaluation metrics. For multiobjective optimization problems, a common approach is to use specialized multi-objective acquisition functions to obtain the Pareto front. Another approach is to aggregate multiple objectives into a single objective, allowing the use of standard single-objective optimization tools. The advantage of the former lies in its ability to produce a set of non-dominated solutions, offering various trade-offs for decision makers without relying on comparability or predefined weights among objectives. However, it tends to

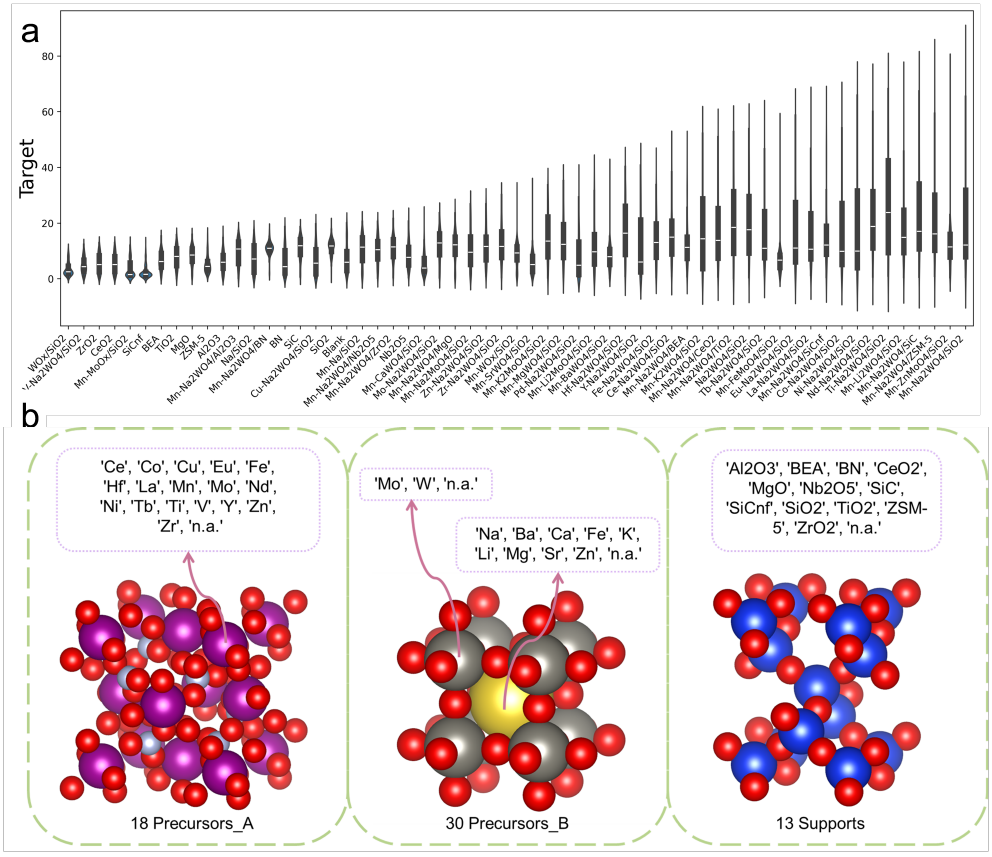


Fig. 1 (a)Distribution of target values corresponding to different catalyst compositions, with the white line indicating the median. The compositions are arranged in ascending order according to the highest target value reported in the database. (b) Overview of the catalyst library employed in the study, including 18 precursors_A, 30 precursors_B, and 13 supports.

be computationally expensive and selecting a final solution still requires additional decision making. The latter approach, although it sacrifices the diversity of trade off solutions and requires careful consideration in how the objectives are combined, is generally more efficient once the aggregation method is determined.

In this work, we adopt the latter strategy by integrating the selectivity and yield of different products into a single objective function. To balance and integrate selectivity and yield, the objective function is defined as follows:

$$Target = Conv_{CH_4} \frac{2Y_{C_2H_4} + Y_{C_2H_6}}{2Y_{CO_2} + Y_{CO}} \tag{1}$$

where $Conv_{CH_4}$ represents the conversion of CH_4, Y represents yield. The coefficient is determined by the value of each product, reflecting the idea of how much value-generating product can be produced at a given level of byproduct. In theory, every possible combination of multiple objectives can define a distinct chemical space. This work also explores different ways of defining the target, and the results show that as long as the target exhibits a reasonable trend, our method can deliver excellent performance.

2.2 Treatment of discrete parameters

In the context of the OCM optimization problem, four discrete parameters were initially considered, corresponding to the selection of three distinct metal components and one catalyst support type.(Fig. 1b) To enable joint optimization with continuous parameters, we employed an exponentially decaying overlap kernel in conjunction with Hamming distance based discrete search space defined in (Eq. 3), thereby embedding both discrete and continuous domains within a unified surrogate modeling framework in (Eq. 4).

Subsequently, the four discrete parameters were aggregated into composite categorical entries, each representing a unique combination of metal-support configurations (e.g., Mn—Na—W—SiC) rather than being treated as independent parameter dimensions. This reformulation effectively reduced the discrete optimization subspace to a single categorical variable, simplifying the search landscape while preserving the combinatorial complexity inherent to catalyst composition, ensuring the validity of the results.

In this implicit chemical space, we conducted five independent optimization runs, each with 150 iterations. During the early exploration phase (iterations 0–40), MVRSM and SMKBO performed similarly and outperformed the other methods. Subsequently, the performance of CAS and CoCaBO improved markedly, reaching parity with MVRSM; however, the incumbent solution achieved by SMKBO remained clearly higher than the others whose final target performance converged to approximately 65.(Fig 2**b**)

Then the four discrete parameters were separated and treated as four independent dimensions for optimization.(Fig 2**a**) During the early stage of optimization (iterations 0–65), all algorithms, except for random search, exhibited similar optimization speeds. However, beyond this point, CAS, TPE, and MVRSM gradually converged to comparable performance levels, while the average optimization incumbent solution of SMKBO continued to increase, ultimately converging to a value of approximately 60.

To quantify the performance of optimization methods, the concepts of Enhancement Factor (EF) and Acceleration Factor (AF) are introduced.(Section 5.4) The results show that the SMKBO methods outperform other approaches both before and after the integration of discrete parameters. Compared with the optimization configuration where discrete parameters are dispersed, the integrated configuration leads to significant improvements in both enhancement and acceleration factors, with only the TPE method showing a slight decline in performance.(Fig.2**c,d**)

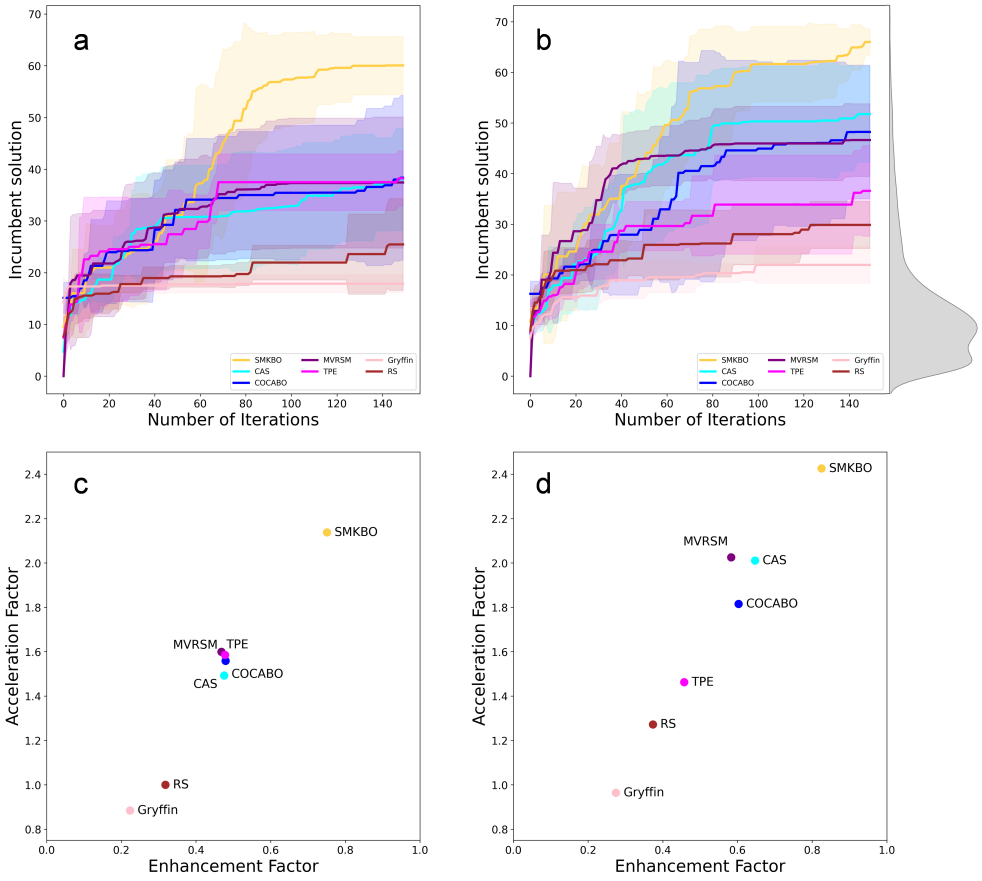


Fig. 2 (a)Incumbent solution values for the separate discrete parameters configuration over 150 iterations (b)Incumbent solution values for the combined discrete parameters configuration over 150 iterations. (c) (d) Enhancement and acceleration factors of all methods for separated discrete parameters configuration and combination discrete parameters configuration. Shaded regions in (a) and (b) represent standard deviation, while scatter points in (c) and (d) indicate individual method performances.

2.3 Robustness analysis

To verify the stability of the proposed method, Gaussian noise with a mean of 0 and standard deviations of 1% and 5% of the global optimum target (69.9) was added to the target values at each optimization iteration, thereby simulating measurement errors encountered in real experiments.

In the robustness analysis, SMKBO consistently outperforms competing algorithms across both low- and high-noise conditions. When the discrete parameters are treated separately (Fig 3**a,c**), SMKBO shows a clear efficiency advantage, particularly under low noise, where the performance gap with other algorithms widens steadily with the number of iterations. When the discrete parameters are integrated (Fig 3**b,d**), the

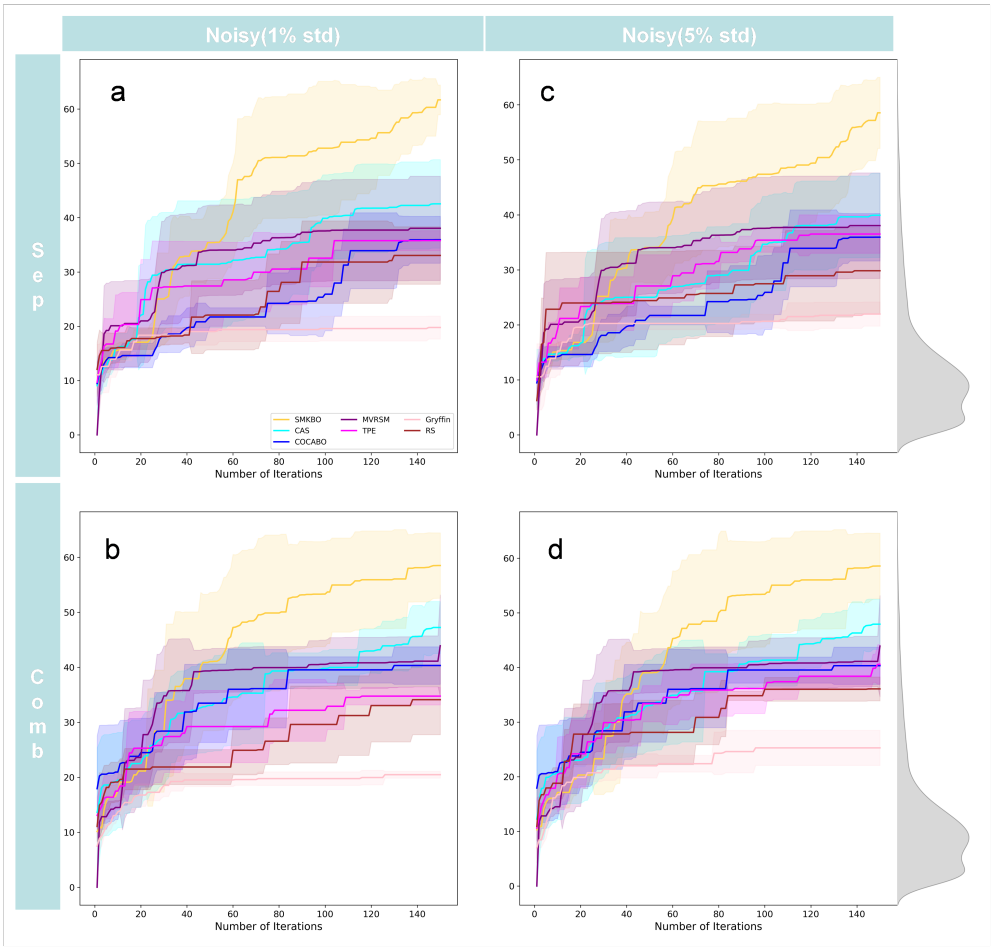


Fig. 3 (a)(b) Incumbent solution values for the separate discrete parameters configuration over 150 iterations under 1% and 5% global standard deviation noise. (c)(d)Incumbent solution values for the separate integrated parameters configuration over 150 iterations under 1% and 5% global standard deviation noise. Shaded regions represent standard deviation

dimensionality reduction allows other algorithms, such as CAS, to focus more effectively on continuous parameters, thereby narrowing the performance gap. However, even under these conditions, SMKBO maintains lower fluctuation levels than other methods, demonstrating superior robustness. As the noise level increases, the benefits of parameter integration become more apparent for all algorithms, yet SMKBO remains the most stable method. These results highlight that SMKBO not only excels in efficiency under challenging mixed variable scenarios but also exhibits stronger robustness against noise, making it particularly suited for practical optimization tasks.

2.4 Different Number of Cauchy and Gaussian Mixtures

To illustrate the performance differences resulting from varying the number of mixture components, we further compare the performance of spectral mixture kernels combining Cauchy and Gaussian components, as illustrated in Fig. 4.

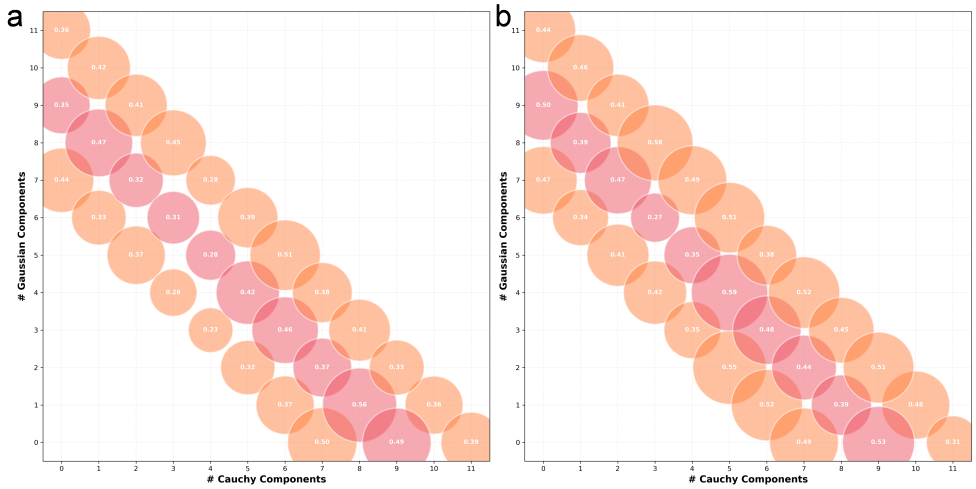


Fig. 4 Results for spectral mixture kernels with different number of Cauchy and Gaussian mixtures. The x-axis represents the number of Cauchy components, and the y-axis represents the number of Gaussian components. Different colors are used to distinguish mixtures based on their total number of components (7,9,11). The bubble size represents the AUC area. (a) is for separation configuration, (b) is for combination configuration

3 Discussion

3.1 Optimization Path Analysis in Discrete Parameters

During the first 50 iterations, the algorithm had not yet gained sufficient knowledge about the metal composition, and the obtained target values remained relatively low. After 50 iterations, the algorithm converged to the Ti–Mg–W-ZSM-5 combination, but it still sparsely replaced the values of discrete parameters to enable exploration. Interestingly, both Component2 and the Support appeared to fall into local optima, with values of Mg and ZSM-5, respectively. However, as the number of iterations increased, both eventually escaped from the local optima. Nevertheless, in some other iterations, the ZSM-5 support was also found to yield optimal values.(Fig.5)

Moreover, after initial convergence, the algorithm tends to alter only one discrete variable at a time while keeping the others fixed. This behavior parallels the common strategy employed by experimental scientists, who usually vary a single factor to assess its impact on the outcome. In addition, when a notable performance improvement is observed, the algorithm not only proceeds with iterating and switching discrete

values, but also takes some "make sure " moves which are reverting to the previous convergence point to verify whether the observed improvement is indeed attributable to the change in the corresponding discrete variable immediately.

From the perspective of convergence results, among all the discrete parameter combinations to which SMKBO ultimately converged, 65% converged to the Mn–Na–W combination, which corresponds to the best C_2 yield reported in the original database and exhibited excellent performance on SiC, SiO_2, CeO_2, ZSM-5, and $SiCnf$. The next most combination is the Ti–Na–W–SiO_2 combination, to which SMKBO converged with a proportion of 15.7%.

3.2 Optimization Path Analysis in Continuous Parameters

To understand the optimization process, all continuous parameters were projected into a two-dimensional space using t-SNE. The background denotes the predicted target values obtained via regression with XGBoost [22], where the discrete parameters were fixed to the optimal combination corresponding to each set of continuous parameters to ensure the consistency of the regression surface. (Fig.6)

The results indicate that the algorithm starts from the lower right region with relatively low target values and progresses almost diagonally toward the upper left region with higher target values. During the first 20 iterations, the continuous parameters exhibited substantial fluctuations, suggesting that the algorithm was simultaneously learning the discrete parameters while exploring the continuous variable space as extensively as possible. Moreover, within the first 50 iterations, two optimization paths for the continuous parameters were identified. After 70 iterations, the algorithm discovered a better performing combination of continuous parameters, entering the red region in the figure. Subsequently, following the 76th iteration (when the Support had just converged to SiC), the algorithm kept the continuous parameters nearly unchanged from iterations 76 to 79, in order to verify whether the observed performance improvement was indeed attributable to the discrete variable.

4 Conclusion

In this study, we replaced the surrogate model in BO with a mixed spectral kernel constructed from GSM and CSM, where the kernel function is defined in the frequency domain based on the corresponding spectral density functions. This approach achieved performance surpassing previous work in OCM tasks. Furthermore, the SMKBO method was also applied to the NO_x reduction over zeolite-based SCR catalysts scenario, and it consistently demonstrated superior performance compared to the other approaches.

These results demonstrate that the SMKBO exhibits a stronger capability in capturing complex relationships in the experimental parameter space. Intuitively, the hybrid kernel combines the characteristics of both Gaussian and Cauchy components: the GSM decays exponentially with the square of distance, effectively capturing smooth regions of the parameter space, while the CSM decays exponentially with

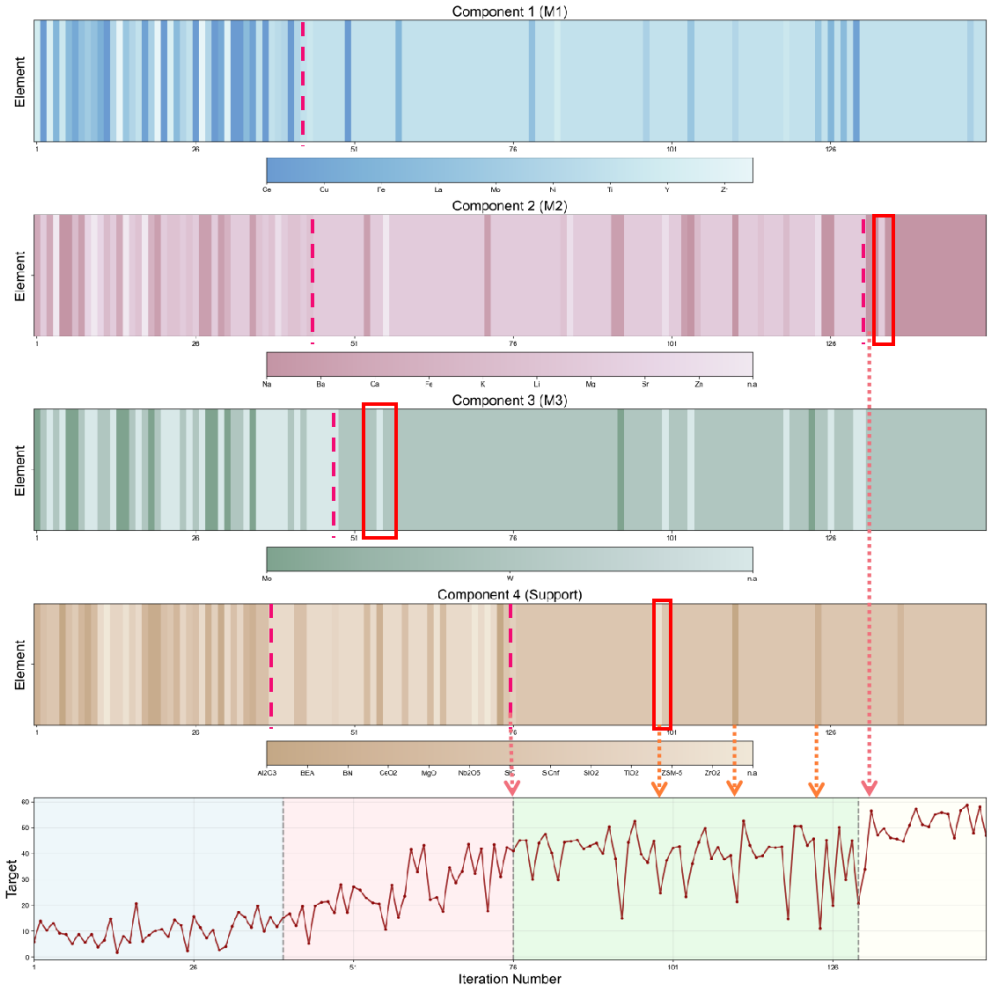


Fig. 5 Evolution of discrete parameters during optimization for (a) Component 1 (M1), (b) Component 2 (M2), (c) Component 3 (M3), and (d) Component 4 (Support). Each vertical stripe corresponds to the selected element/support at a given iteration. (e) Target value as a function of iteration number, showing the overall improvement and convergence trend. Pink arrows highlight partially successful attempts, whereas orange arrows highlight partially failed attempts. Red boxes highlight the "make sure" moves.

distance, enabling better utilization of outliers in the observed data. This dual behavior contributes to the superior performance of the proposed method over existing optimization algorithms.

The optimization logic of SMKBO resembles that of human scientists. It follows an explore-first, exploit-later strategy, allowing extensive exploration of the search space before focusing on high-potential regions. Moreover, verifications are conducted after

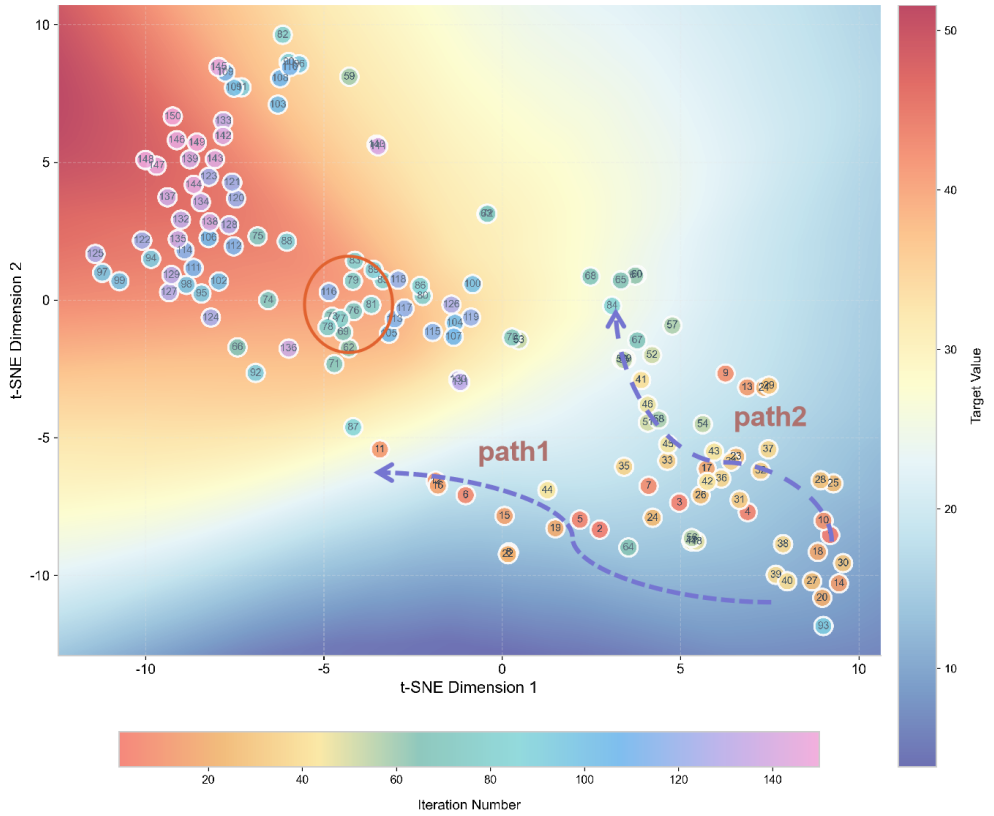


Fig. 6 All continuous parameters were projected into a two-dimensional space using t-SNE. The background color represents the regression surface of the target value, with discrete parameters fixed to the optimal combination corresponding to each set of continuous parameters. Each point denotes an iteration, colored according to the iteration number.

switching the convergence criterion, ensuring that the obtained solutions are both stable and reproducible.

5 Method Overview

5.1 Dataset Construction

The data were originally collected from the work of Thanh Nhat Nguyen et al. in 2020 [21]. In this study, the authors developed a high-throughput screening instrument capable of automatically evaluating the performance of 20 catalysts under 216 different reaction conditions. A total of 12,708 data points were collected from 59 combinations of catalysts and varying experimental parameters.

Each condition includes temperature (°C), total flow (mL/min), flows of Ar, CH_4, and O_2, the molar amount of metal in the precursor (mol), contact time (seconds),

and the CH_4/O_2 molar ratio. The performance of catalysis is represented by Coversion rate, yield and selectivity of C_2H_4, C_2H_6, CO and CO_2. In this work, the performance indicators of the catalysts were recombined to enable a comprehensive representation of catalyst selectivity and productivity. (Eq. 1) Besides, the optimization parameters include temperature (°C), flows of Ar, CH_4, and O_2, the molar amount of metal in the precursor (mol) and contact time (seconds).

5.2 Bayesian Optimization

Consider the problem of maximizing an unknown and expensive function f, which can be formulated as:

$$x^* = \arg\max_{x \in \mathcal{X}} f(x)$$

where $\mathcal{X} \subset \mathbb{R}^d$ denotes the search/decision space of interest and x^* represents the global minimum. Starting with a limited set of observations, BO builds a probabilistic surrogate model, typically a GP, to estimate the unknown objective function. It then uses an acquisition function to evaluate and prioritize candidate solutions based on the model's posterior distribution. The acquisition function guides the search by selecting the most promising points to evaluate next. After each new evaluation, the surrogate model is updated, and the process continues to refine the search. The method seeks to identify the optimal solution within the given constraints.

5.2.1 Continuous Search Space: Spectral Mixture Kernel

The choice of kernel significantly influences BO performance by shaping the structure of the underlying GP surrogate. Most existing BO solvers use conventional kernels (e.g., squared exponential, rational quadratic, and Matérn), which, despite their simplicity, often fail to capture the complexity inherent in practical applications. [23].

In this paper, we propose a spectral mixture kernel, motivated by Bochner's theorem [24], which establishes that every stationary kernel corresponds to a symmetric spectral measure. This leads to the following general form:

$$k_x(\tau) = \sum_{q=1}^{Q_g} w_q^g \exp\left(-2\pi^2 \tau^\top \mathbf{\Sigma}_q \tau\right) \cos\left(2\pi \tau^\top \mu q\right) + \sum_{q=1}^{Q_c} w_q^c \exp\left(-2\pi |\tau^\top \gamma_q|\right) \cos\left(2\pi \tau^\top \mathbf{x}_{0q}\right). \tag{2}$$

A detailed treatment of the spectral mixture kernel is provided in Section 6, where we present its mathematical derivation, discuss its theoretical properties, and demonstrate how its flexibility allows it to approximate a wide class of kernels. We further show that it achieves superior empirical performance in continuous search spaces compared to conventional alternatives.

5.2.2 Mixed Search Space: Composite Kernel

In addition to the purely continuous problems, our spectral mixture kernel also generalizes to mixed categorical-continuous spaces, a setting frequently encountered in the context chemistry and experimentation but hitherto under-explored in literature.

For categorical inputs, we modify the Hamming kernel $k(\mathbf{h}, \mathbf{h}') = \frac{\sigma}{d_h} \sum_{i=1}^{d_h} \delta(h_i, h_i')$, in Ru et al. [25] and Kondor and Lafferty [26]:

$$k_h(\mathbf{h}, \mathbf{h}') = \exp\left(\frac{1}{d_h} \sum_{i=1}^{d_h} \ell_i \delta(h_i, h_i')\right), \tag{3}$$

where $\{\ell_i\}_i^{d_h}$ are the lengthscale(s), and $\delta(\cdot, \cdot)$ is the Kronecker delta function.

To handle mixed input $z = [h, x]$, we combine the spectral mixture kernel and Hamming kernel together and propose the composite kernel:

$$k(\mathbf{z}, \mathbf{z}') = \lambda\Big(k_x(\mathbf{x}, \mathbf{x}')k_h(\mathbf{h}, \mathbf{h}')\Big) + (1 - \lambda)\Big(k_h(\mathbf{h}, \mathbf{h}') + k_x(\mathbf{x}, \mathbf{x}')\Big), \tag{4}$$

where $\lambda \in [0, 1]$ is a trade-off parameter, k_x is defined in Eq. (2) and k_h is defined in Eq. (3). This formulation therefore allows us to use composite kernels that are most appropriate for the mixed input types while still flexibly capturing the possible additive and multiplicative interactions between them.

5.2.3 Acquisition Function

Given our unified Gaussian Process handles both categorical and continuous inputs, we propose a novel acquisition strategy that alternates between optimizing these input types. At each optimization step, we perform one local search step for categorical inputs followed by one gradient-based optimization step for continuous inputs. This alternation repeats until convergence or until reaching a predefined maximum number of iterations.

Local Search for Categorical Inputs

Optimizing categorical spaces presents unique challenges as standard surrogate models tend to over-explore. We employ a Trust Region (TR) approach [27] with two phases: First, the unified GP (with kernel in Eq. (4)) identifies a promising center solution. Then, a local GP trained on solutions within this trust region performs refined optimization to select the final candidate. This hierarchical strategy balances global exploration with local exploitation.

Center Solution Selection

The unified GP selects the center solution $\mathbf{h}_i^{(0)}$ using an Upper Confidence Bound (UCB) acquisition strategy:

$$\mathbf{h}_i^{(0)} = \arg\max_{\mathbf{h} \in \mathcal{H}} \left[\mu_{\text{uni}}(\mathbf{h}; D_i) + \sqrt{\beta_i}\sigma_{\text{uni}}(\mathbf{h}; D_i)\right]$$

where $\mu_{\text{uni}}(\mathbf{h}; D_i)$ and $\sigma^2_{\text{uni}}(\mathbf{h}; D_i)$ are the posterior mean and variance of the unified GP trained on historical data D_i, and β_i controls the exploration-exploitation trade-off. This center solution serves as the starting point for trust region exploration.

Trust Region Construction

The trust region defines a constrained neighborhood around center solution $\mathbf{h}^*$:

$$\mathrm{TR}_h(\mathbf{h}^*)_{L^h} = \left\{ \mathbf{h} \mid \sum_{i=1}^{d_h} \delta(h_i, h_i^*) \leq L^h \right\}$$

The radius L^h adapts dynamically, expanding when the best function value f_T^* improves and contracting otherwise. The bounds $L^h_{\min} = 0$ and $L^h_{\max} = d_h$ correspond to the Hamming distance limits.

5.3 AutoML

Based on the dataset and prediction target, models such as XGBoost, LightGBM, and Random Forest were automatically integrated by AutoGluon-Tabular [28] to capture the implicit chemical space. This approach achieved a mean absolute error (MAE) of 2.15% on the test set, with a training-to-test split ratio of 7:3.

AutoGluon-Tabular is an open-source automated machine learning (AutoML) framework specifically designed for tabular data. It integrates widely used models, including XGBoost, LightGBM, CatBoost, Neural Networks, and Random Forests. By employing advanced techniques such as multi-layer stack ensembling and repeated k-fold bagging, it enables both accurate and efficient model fitting. In this work, we used AutoGluon version 1.3.1.

5.4 Enhancement factor and acceleration factor

The EF is employed to quantify the optimization performance of an algorithm after a given number of iterations.

The AF is employed to evaluate the speed improvement of an optimization algorithm relative to random search. Specifically, it is defined as the ratio between the area under the iteration–incumbent solution curve of the algorithm and that of random search.

$$EF = \frac{Y_{Incumbent}}{Y_{best}} \quad AF = \frac{AUC_{Method}}{AUC_{RS}}$$

This definition was originally inspired by the work of Qiaohao Liang et al.[29] In their study, the authors introduced the concepts of the enhancement factor and the acceleration factor, formulated with reference to the Top% metric.

6 Spectral Mixture Kernel

6.1 Kernel Derivation

Our proposed spectral mixture kernel stems from Bochner's Theorem, which provides a foundational result in characterizing positive definite kernels in terms of their spectral representations.

Theorem 1 (Bochner's Theorem [30]) *A complex-valued function k on $\mathbb{R}^P$ is the kernel of a weakly stationary, mean square continuous complex-valued random process on $\mathbb{R}^P$ if and only if it can be represented as*

$$k(\tau) = \int_{\mathbb{R}^P} \exp(2\pi i s^\top \tau) d\psi(s),$$

where ψ is a positive finite Borel measure on $\mathbb{R}^P$.

Specifically, Bochner's theorem states that the Fourier transform of any stationary covariance function on $\mathbb{R}^P$ is proportional to a probability measure, and conversely, the inverse Fourier transform of a probability measure yields a stationary covariance function [24, 30]. The measure ψ is called the *spectral measure* of k. If ψ has a density S, then S is referred to as the *spectral density* or *power spectrum* of k. The covariance function k and the spectral density S forms a Fourier pair [31]:

$$k(\tau) = \int S(s) \exp(2\pi i s^\top \tau)\, ds, \qquad S(s) = \int k(\tau) \exp(-2\pi i s^\top \tau)\, d\tau. \tag{5}$$

6.1.1 Gaussian Spectral Density

A natural choice for constructing a space of stationary kernels is to use a mixture of Gaussian distributions [23] to represent the spectral density $S(s)$:

$$\phi_{\mathrm{g}}(s) = \sum_{q=1}^{Q_g} w_q \mathcal{N}(s; \mu_q, \mathbf{\Sigma}_q), \quad S(s) = \frac{\phi_g(s) + \phi_g(-s)}{2}, \tag{6}$$

where the construction of S ensures symmetry, and the wights w_q determine the contribution of each of the Q_g components. By taking the inverse Fourier transform in Eq. (5), the resulting spectral mixture kernel induced by Gaussian distributions (Gaussian spectral mixture kernel, GSM) is given by:

$$k_{\mathrm{g}}(\tau) = \sum_{q=1}^{Q_g} w_q \exp\left(-2\pi^2 \tau^\top \mathbf{\Sigma}_q \tau\right) \cos\left(2\pi \tau^\top \mu_q\right). \tag{7}$$

Inspecting Eq. (7), we observe that the covariance function induced by a Gaussian mixture spectral density is infinitely differentiable. However, this choice may generate overly smooth sample paths [32].

6.1.2 Cauchy Spectral Density

To address the issue of overly smooth sample paths in GSM, we introduce a different family of distributions, *i.e.*, Cauchy distributions, to construct a class of continuous but finitely differentiable covariance functions.

Theorem 2 *If $\phi_c(s)$ is a mixture of Q_c Cauchy distributions on $\mathbb{R}^P$, where the q^{th} component has a position parameter vector $\mathbf{x_{0}}_q = (x_{0q}^{(1)}, \ldots, x_{0q}^{(P)})$ and scale parameter $\gamma_\mathbf{q} = diag(\gamma_q^{(1)}, \ldots, \gamma_q^{(P)})$, and τ_p is the p^{th} component of the P-dimensional vector $\tau = \mathbf{x} - \mathbf{x}'$. The Fourier dual of spectral density $\phi_c(s)$ is*

$$k_{\mathrm{c}}(\tau) = \sum_{q=1}^{Q_c} w_q \exp\left(-2\pi|\tau^\top \gamma_\mathbf{q}|\right) \cos\left(2\pi\tau^\top \mathbf{x_{0q}}\right).$$

The spectral mixture kernel induced by Cauchy distributions is referred to as the Cauchy spectral mixture kernel (CSM). We offer a brief outline of the key ideas here, and a detailed proof is provided in Appendix E.1.

We begin by considering a simplified case where the probability density function of a univariate Cauchy distribution is given by

$$\mathcal{C}(s; x_0, \gamma) = \frac{1}{\pi} \frac{\gamma}{(s - x_0)^2 + \gamma^2}, \tag{8}$$

and the spectral density S follows Eq. (6), while replacing the Gaussian distribution with Cauchy distribution. Noting that S is symmetric [33], substituting S into Eq. (5) yields

$$k_c(\tau) = \frac{1}{\pi} \int \frac{\gamma}{(s - x_0)^2 + \gamma^2} \exp\left(2\pi i s \tau\right) ds.$$

Notice that $\int_{-\infty}^{\infty} \frac{\gamma}{(s-x_0)^2+\gamma^2}\, ds = \pi$, we have

$$\mathcal{F}\left[\frac{\gamma}{(s - x_0)^2 + \gamma^2}\right] = \exp(-2\pi|\gamma\tau|) \exp\left(2\pi i x_0 \tau\right),$$

where $\mathcal{F}$ denotes the Fourier Transform. This gives us the following kernel

$$k_{\mathrm{c}}(\tau) = \exp\left(-2\pi|\gamma\tau|\right) \cos(2\pi\tau x_0).$$

Now, if $\phi(s)$ is a mixture of Q_c Cauchy distributions as described in Theorem 2, with its spectral density given by $\phi_{\mathrm{c}} = \sum_{q=1}^{Q_c} w_q \mathcal{C}(s; \mathbf{x}_{0q}, \gamma_q)$, we obtain

$$k_{\mathrm{c}}(\tau) = \sum_{q=1}^{Q_c} w_q \prod_{p=1}^{P} \exp\left(-2\pi|\tau_p \gamma_q^{(p)}|\right) \cos(2\pi\tau_p x_{0q}^{(p)}) = \sum_{q=1}^{Q_c} w_q \exp\left(-2\pi|\tau^\top \gamma_\mathbf{q}|\right) \cos\left(2\pi\tau^\top \mathbf{x_{0q}}\right). \tag{9}$$

6.1.3 Cauchy and Gaussian Spectral Mixture Kernel

To leverage the complementary properties of both distributions, we define a spectral density $S(s)$ as a mixture of Gaussian and Cauchy components:

$$\phi_{\text{cg}}(s) = \sum_{q=1}^{Q_g} w_q^g \mathcal{N}(s; \mu_q, \mathbf{\Sigma}_q) + \sum_{q=1}^{Q_c} w_q^c \mathcal{C}(s; \mathbf{x}_{0q}, \gamma_q). \tag{10}$$

The resulting Cauchy-Gaussian Spectral Mixture (CSM+GSM) then follows:

$$k_{\text{cg}}(\tau) = \sum_{q=1}^{Q_g} w_q^g \exp\left(-2\pi^2 \tau^\top \mathbf{\Sigma}_q \tau\right) \cos\left(2\pi \tau^\top \mu_q\right) + \sum_{q=1}^{Q_c} w_q^c \exp\left(-2\pi |\tau^\top \gamma_q|\right) \cos\left(2\pi \tau^\top \mathbf{x}_{0q}\right), \tag{11}$$

which is exactly the same as the general form in (2).The CSM+GSM kernel offers several key advantages over using either component alone. First, it maintains spectral interpretability, where each component's location parameters (μ_q and $\mathbf{x}_{0q}$) correspond to distinct frequency modes and their weights represent relative energy contributions. Second, it achieves adaptive smoothness through multi-scale modeling capability. The Gaussian components capture smooth global trends via their bandwidth parameters $\mathbf{\Sigma}_q$ while the Cauchy components model local variations through their heavy-tailed distributions controlled by γ_q.

6.2 Information Gain and Regret Bound

We first provide upper bounds on the maximum information gains of the spectral mixture kernels, which measure how fast the objective function f can be learned in an information-theoretic sense. We refer readers to Appendix E.2 for the detailed proofs of the results in this section.

The maximum information gain $\gamma(T)$ achieved by sampling T points in a GP defined over a set $\mathcal{X} \subset \mathbb{R}^d$ with a kernel k is defined as:

$$\gamma(T) := \max_{A \subset \mathcal{X}: |A| = T} I(y_A; f_A),$$

where $I(y_A; f_A) = \frac{1}{2} \log |I + \sigma^{-2} K_A|$. $K_A = [k(x, x')]_{x,x' \in A}$ is the covariance matrix of $f_A = [f(x)]_{x \in A}$ associated with the samples A, and σ^2 is the noise variance.

Theorem 3 *The upper bounds on the maximum information gain of CSM and GSM kernels are:*

a) Cauchy spectral mixture (CSM): $\gamma_c(T) = \mathcal{O}\left(T^{\frac{d^2+d}{d^2+d+1}} (\log T)\right)$;

b) Gaussian spectral mixture (GSM): $\gamma_g(T) = \mathcal{O}\left((\log T)^{d+1}\right)$.

where d denotes the input dimension.

Using maximum information gain, we apply Theorem 1 from Srinivas et al. [34] to derive the cumulative regret bound when pairing with a UCB acquisition function.

Proposition 4 *Let* $\delta \in (0,1), \beta_t = 2\log\left(|\mathcal{X}|t^2\pi^2/6\delta\right)$*, where* β_t *is the hyperparameter for the UCB acquisition function. Suppose the objective function* $f : \mathcal{X} \to \mathbb{R}$ *is sampled from* $\mathcal{GP}(\mathbf{0}, k(x, x'))$*. With high probability, BO using the UCB acquisition function obtains a cumulative regret bound of*

a) Cauchy spectral mixture (CSM): $\mathcal{O}\left(T^{\frac{2d^2+2d+1}{2(d^2+d+1)}}\sqrt{\log T \cdot \log|\mathcal{X}|}\right)$,

b) Gaussian spectral mixture (GSM): $\mathcal{O}\left(\sqrt{T}\cdot(\log T)^{\frac{d+1}{2}}\cdot\sqrt{\log|\mathcal{X}|}\right)$.

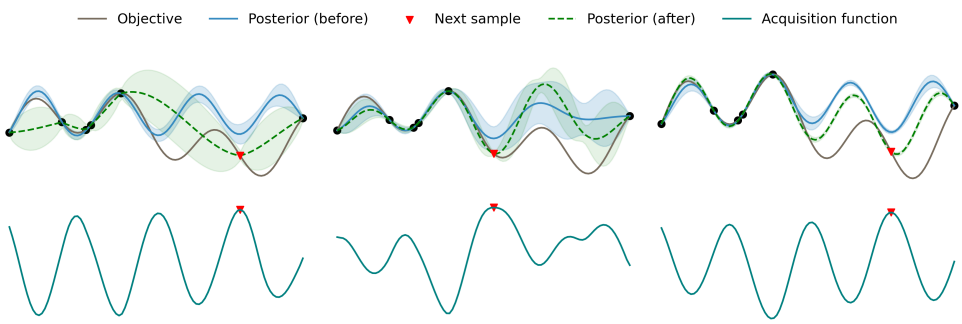
The information gain for the CSM kernel grows at a sub-polynomial rate in T. The exponent $\frac{d^2+d}{d^2+d+1}$ is slightly less than 1, implying that the information gain increases rapidly, but at a diminishing rate as T grows. For the GSM kernel, the information gain grows logarithmically in T, with the growth rate amplified by the dimensionality d, indicating that higher dimensionality increases the rate of information gain. Due to its logarithmic growth in T, the GSM kernel more effectively controls cumulative regret, providing more stable long-term performance.

While our theoretical analysis establishes regret bounds for pure CSM and GSM, the CSM+GSM kernel suggests intriguing potential behavior. Intuitively, its mixed components imply dual-phase characteristics: in the initial optimization stages, the Cauchy components may promote high-frequency exploration, leading to polynomial information gain, while the asymptotic behavior is likely to transition to logarithmic scaling as the Gaussian components, which capture low-frequency patterns, become dominant. This phased behavior could provide a practical balance, achieving faster initial convergence compared to pure GSM while offering better long-term stability than pure CSM.

Example 1 Figure 7 visualizes one iteration of BO using different kernels based on 6 random samples from a 1D test function $f(x) = \sin(x) + \sin(\frac{10}{3}x)$.

The results reveal significant divergence in the behavior of GP surrogates employing different kernels. CSM excels in capturing high-frequency components (rapid oscillations and narrow confidence bands around sharp peaks), aligning with its heavy-tailed spectral density that preserves high-frequency content; while GSM better models low-frequency trends (smooth fitting and wider confidence intervals) due to its exponentially decaying spectral density that attenuates high frequencies. The CSM+GSM[1] achieves superior performance by maintaining CSM's precise high-frequency tracking and GSM's stable low-frequency extrapolation, as visually confirmed by its balanced error distribution and theoretically explained by its dual-peaked spectral energy distribution.

[1] CSM, GSM, and CSM+GSM in the main text denote spectral mixture kernels with 7 Cauchy, 7 Gaussian, and 6 Cauchy plus 1 Gaussian components, respectively. We verify the robustness of our results to alternative component specifications in Appendix C.6.



(a) CSM (b) GSM (c) CSM+GSM

Fig. 7 Comparison of predictive distributions for the objective $f(x)$ using different kernels, before and after conditioning on the sampled point. The upper subplots show the GP surrogate before and after adding a new sample point, while the lower subplots display the corresponding acquisition function values using UCB and indicate the next point to sample.

6.3 Approximate Conventional Kernels

Theorem 1 allows us to approximate an arbitrary stationary covariance function by approximating (*e.g.*, modeling or sampling from) its spectral density [32]. Since mixtures of Cauchy and Gaussian distributions can be used to construct a wide range of spectral densities [35], these mixtures enable us to approximate any stationary covariance function. We considered the approximation of conventional kernels of various complexities. The parameters for source kernels that generate the data are shown in Table D2 (Appendix C).

We evaluate the approximation ability of six distinct kernels, including the standard RQ, MA12, and MA52, alongside our proposed CSM, GSM, and CSM+GSM, through both quantitative and functional analyses. Table 1 reveals that our proposed kernels achieve superior marginal log likelihood (MLL) scores compared to conventional kernels. Figure 8 further demonstrates that our proposed kernels provide the closest approximation to the true kernel's correlation structure.

6.4 Optimization Tasks

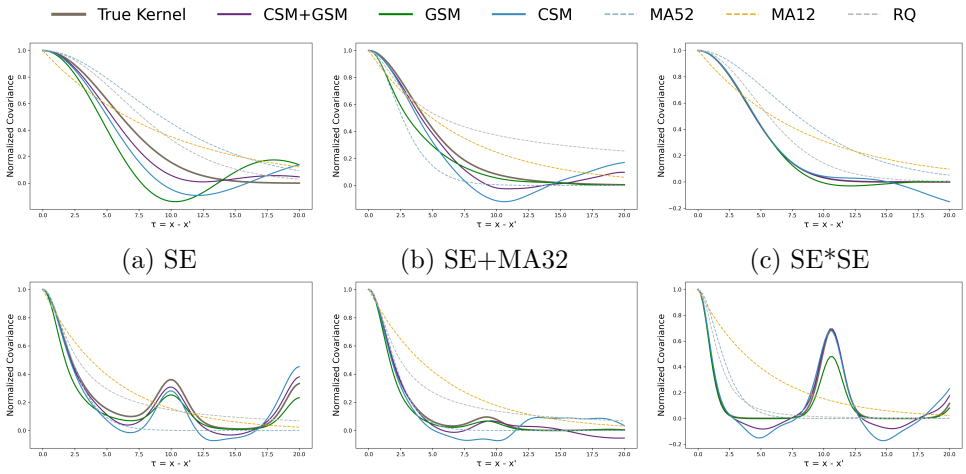
To further illustrate the efficay of spectral mixture kernel, we validate our approach against several baselines across a diverse set of optimization tasks, with results summarized in Table 2 and details deferred to Appendix C. The findings indicate that both CSM and GSM consistently outperform other methods in terms of convergence speed and optimality gap. Notably, the combined CSM+GSM kernel achieves the best overall performance across most benchmark functions, and this advantage becomes increasingly pronounced as the problem dimensionality grows.

For conventional kernels, performance varies substantially depending on the specific task, underscoring their limited adaptability. In contrast, as we shift to more practical and challenging objective functions (Robot 4d, Portfolio 5d), the benefits of employing more flexible and computationally efficient kernels become evident. All variants of the spectral mixture kernels outperform conventional baselines, with the

Table 1 MLL of training on the sampled data. A higher MLL indicates a better approximation of the true kernel. Each experiment was repeated 10 times using different random seeds. The table shows the mean MLL values with standard deviations in parentheses.

	CSM+GSM	CSM	GSM	RQ	MA52	MA12	**True**
SE	**2.751** (0.067)	2.719 (0.075)	2.702 (0.059)	2.498 (0.058)	2.437 (0.058)	2.256 (0.061)	2.752 (0.044)
SE+MA32	1.021 (0.055)	1.012 (0.061)	**1.042** (0.059)	0.872 (0.066)	1.040 (0.058)	0.891 (0.053)	1.134 (0.051)
SE*SE	**2.499** (0.059)	2.271 (0.054)	2.398 (0.056)	2.292 (0.065)	1.933 (0.058)	1.728 (0.063)	2.517 (0.045)
PE+SE+MA32	**0.885** (0.057)	0.821 (0.050)	0.776 (0.053)	0.527 (0.061)	0.530 (0.051)	0.459 (0.053)	0.920 (0.044)
SE*(PE+MA32)	**0.569** (0.049)	0.521 (0.053)	0.557 (0.044)	0.418 (0.052)	0.534 (0.042)	0.377 (0.043)	0.575 (0.041)
PE*(SE+MA32)	**0.983** (0.049)	0.965 (0.052)	0.894 (0.048)	0.634 (0.057)	0.681 (0.053)	0.685 (0.056)	1.053 (0.044)

hybrid kernel that integrates both Gaussian and Cauchy components delivering the most robust gains. In particular, this kernel reduces the log optimality gap by nearly two orders of magnitude compared to standard Bayesian optimization methods that rely on conventional kernels.



(a) SE (b) SE+MA32 (c) SE*SE

(f) PE+SE+MA32 (d) SE*(PE+MA32) (e) PE*(SE+MA32)

Fig. 8 Learned correlation function in kernel approximation. The horizontal axis denotes the Euclidean distance between two points, while the vertical axis represents the corresponding covariance distance. The darker solid line denotes the kernel that generates sampling points. Spectral mixture kernels more closely approximate the true kernel.

Table 2 Results for different test functions and algorithms across 10 repetitions using UCB acquisition function. In each cell, the first number represents the mean results, and the second number (inside parentheses) indicates the standard error. ABO failed in Rosenbrock-20d and Levy-30d.

Objective	Dim	RBF	RQ	MA52	ABO	ADA	SDK	SINC	CSM	GSM	CSM+GSM
Branin	2	-2.29 (0.02)	-2.25 (0.03)	-2.33 (0.02)	-0.80 (0.03)	-2.29 (0.01)	2.56 (0.15)	3.02 (0.08)	-1.98 (0.03)	-2.29 (0.03)	**-2.34** (0.01)
Hartmann	3	-1.43 (0.06)	-1.63 (0.05)	-0.91 (0.10)	0.46 (0.15)	-2.14 (0.11)	-1.77 (0.07)	-0.49 (0.06)	-3.21 (0.10)	-1.84 (0.12)	**-7.22** (0.20)
Exponential	5	3.23 (0.05)	2.19 (0.09)	3.64 (0.06)	-0.71 (0.13)	0.76 (0.07)	0.23 (0.04)	2.97 (0.10)	-0.61 (0.14)	**-0.89** (0.09)	-0.87 (0.06)
Hartmann	6	-1.41 (0.08)	0.74 (0.09)	-2.36 (0.12)	-2.43 (0.15)	-2.34 (0.10)	-2.36 (0.08)	-0.55 (0.06)	-3.14 (0.12)	-2.59 (0.13)	**-3.14** (0.15)
Exponential	10	2.17 (0.14)	2.81 (0.16)	2.21 (0.14)	2.28 (0.13)	2.48 (0.19)	1.46 (0.13)	2.06 (0.16)	1.37 (0.18)	1.33 (0.17)	**0.72** (0.18)
Rosenbrock	20	7.97 (0.42)	7.97 (0.46)	7.94 (0.45)	–	7.86 (0.45)	7.93 (0.47)	7.96 (0.46)	3.97 (0.40)	**3.68** (0.38)	4.11 (0.39)
Levy	30	3.47 (0.28)	3.57 (0.24)	3.62 (0.25)	–	3.51 (0.27)	3.59 (0.27)	3.66 (0.23)	3.59 (0.22)	3.49 (0.25)	**3.34** (0.21)
Robot	4	2.08 (0.05)	1.51 (0.06)	1.84 (0.07)	1.94 (0.05)	0.87 (0.04)	0.91 (0.10)	1.62 (0.11)	0.87 (0.05)	0.71 (0.03)	**-0.41** (0.04)
Portfolio	5	20.02 (1.01)	20.61 (0.97)	15.49 (0.92)	18.84 (0.87)	17.61 (1.07)	16.27 (0.98)	18.79 (1.02)	23.32 (0.90)	21.86 (0.91)	**25.62** (0.87)

References

[1] Reymond, J.-L., Van Deursen, R., Blum, L.C., Ruddigkeit, L.: Chemical space as a source for new drugs. MedChemComm **1**(1), 30 (2010)

[2] Osolodkin, D.I., Radchenko, E.V., Orlov, A.A., Voronkov, A.E., Palyulin, V.A., Zefirov, N.S.: Progress in visual representations of chemical space. Expert Opinion on Drug Discovery **10**(9), 959–973 (2015)

[3] Medina-Franco, J.L., Sánchez-Cruz, N., López-López, E., Díaz-Eufracio, B.I.: Progress on open chemoinformatic tools for expanding and exploring the chemical space. Journal of Computer-Aided Molecular Design **36**(5), 341–354 (2022)

[4] Oprea, T.I., Gottfries, J.: Chemography: The Art of Navigating in Chemical Space. Journal of Combinatorial Chemistry **3**(2), 157–166 (2001)

[5] Cai, J., Liang, Q., Luo, M.: Synergistic Acceleration of Adsorbent Material Development by DFT and ML for CO_2 Capture. Chemical Engineering & Technology **48**(7) (2025)

[6] Chang, Y., Benlolo, I., Bai, Y., Reimer, C., Zhou, D., Zhang, H., Matsumura, H., Choubisa, H., Li, X.-Y., Chen, W., Ou, P., Tamblyn, I., Sargent, E.H.: High-entropy alloy electrocatalysts screened using machine learning informed by quantum-inspired similarity analysis. Matter, 2590238524005289 (2024)

[7] Özönder, Ş., Küçükkartal, H.K.: Rapid Discovery of Graphene Nanoflakes with Desired Absorption Spectra Using DFT and Bayesian Optimization with Neural Network Kernel. The Journal of Physical Chemistry A **129**(20), 4591–4600 (2025)

[8] Agarwal, G., Doan, H.A., Robertson, L.A., Zhang, L., Assary, R.S.: Discovery of Energy Storage Molecular Materials Using Quantum Chemistry-Guided Multiobjective Bayesian Optimization. Chemistry of Materials **33**(20), 8133–8144 (2021)

[9] Aldulaijan, N., Marsden, J.A., Manson, J.A., Clayton, A.D.: Adaptive mixed variable Bayesian self-optimisation of catalytic reactions. Reaction Chemistry & Engineering **9**(2), 308–316 (2024)

[10] Herbol, H.C., Poloczek, M., Clancy, P.: Cost-effective materials discovery: Bayesian optimization across multiple information sources. Materials Horizons **7**(8), 2113–2123 (2020)

[11] Ramirez, A., Lam, E., Gutierrez, D.P., Hou, Y., Tribukait, H., Roch, L.M., Copéret, C., Laveille, P.: Accelerated exploration of heterogeneous CO2 hydrogenation catalysts by Bayesian-optimized high-throughput and automated experimentation. Chem Catalysis, 100888 (2024)

[12] Wahab, H., Jain, V., Tyrrell, A.S., Seas, M.A., Kotthoff, L., Johnson, P.A.: Machine-learning-assisted fabrication: Bayesian optimization of laser-induced graphene patterning using in-situ Raman analysis. Carbon **167**, 609–619 (2020)

[13] Wang, K., Dowling, A.W.: Bayesian optimization for chemical products and functional materials. Current Opinion in Chemical Engineering **36**, 100728 (2022)

[14] Xie, Y., Zhang, C., Deng, H., Zheng, B., Su, J.-W., Shutt, K., Lin, J.: Accelerate Synthesis of Metal–Organic Frameworks by a Robotic Platform and Bayesian Optimization. ACS Applied Materials & Interfaces **13**(45), 53485–53491 (2021)

[15] Zhang, Y., Apley, D.W., Chen, W.: Bayesian Optimization for Materials Design with Mixed Quantitative and Qualitative Variables. Scientific Reports **10**(1), 4924 (2020)

[16] Cuesta Ramirez, J., Le Riche, R., Roustant, O., Perrin, G., Durantin, C., Gliere, A.: A comparison of mixed-variables bayesian optimization approaches. Advanced Modeling and Simulation in Engineering Sciences **9**(1), 6 (2022)

[17] Bergstra, J.S., Bardenet, R., Bengio, Y., Kégl, B.: Algorithms for Hyper-Parameter Optimization (2011)

[18] Ru, B., Alvi, A., Nguyen, V., Osborne, M.A., Roberts, S.: Bayesian optimisation over multiple continuous and categorical inputs. In: International Conference on Machine Learning, pp. 8276–8285 (2020). PMLR

[19] Wan, X., Nguyen, V., Ha, H., Ru, B., Lu, C., Osborne, M.A.: Think Global and Act Local: Bayesian Optimisation over High-Dimensional Categorical and Mixed Search Spaces. arXiv (2021)

[20] Stein, M.L.: Interpolation of Spatial Data: Some Theory for Kriging. Springer, (1999)

[21] Nguyen, T.N., Nhat, T.T.P., Takimoto, K., Thakur, A., Nishimura, S., Ohyama, J., Miyazato, I., Takahashi, L., Fujima, J., Takahashi, K., *et al.*: High-throughput experimentation and catalyst informatics for oxidative coupling of methane. Acs Catalysis **10**(2), 921–932 (2019)

[22] Chen, T., Guestrin, C.: Xgboost: A scalable tree boosting system. In: Proceedings of the 22nd Acm Sigkdd International Conference on Knowledge Discovery and Data Mining, pp. 785–794 (2016)

[23] Wilson, A.G., Adams, R.P.: Gaussian process kernels for pattern discovery and extrapolation. In: Proceedings of the 30th International Conference on International Conference on Machine Learning - Volume 28. ICML'13, pp. 1067–1075. JMLR.org, (2013)

[24] Stein, M.L.: Interpolation of Spatial Data: Some Theory for Kriging, pp. 1–249. Springer, (1999)

[25] Ru, B., Alvi, A.S., Nguyen, V., Osborne, M.A., Roberts, S.J.: Bayesian optimisation over multiple continuous and categorical inputs. In: Proceedings of the 37th International Conference on Machine Learning. ICML'20. JMLR.org, (2020)

[26] Kondor, R.I., Lafferty, J.D.: Diffusion kernels on graphs and other discrete input spaces. In: Proceedings of the Nineteenth International Conference on Machine Learning. ICML '02, pp. 315–322. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA (2002)

[27] Eriksson, D., Pearce, M., Gardner, J.R., Turner, R., Poloczek, M.: Scalable global optimization via local Bayesian optimization. Curran Associates Inc., Red Hook, NY, USA (2019)

[28] Erickson, N., Mueller, J., Shirkov, A., Zhang, H., Larroy, P., Li, M., Smola, A.: Autogluon-tabular: Robust and accurate automl for structured data. arXiv preprint arXiv:2003.06505 (2020)

[29] Liang, Q., Gongora, A.E., Ren, Z., Tiihonen, A., Liu, Z., Sun, S., Deneault, J.R., Bash, D., Mekki-Berrada, F., Khan, S.A., *et al.*: Benchmarking the performance of bayesian optimization across multiple experimental materials science domains. npj Computational Materials **7**(1), 188 (2021)

[30] Trust, S.: Lectures on Fourier Integrals. Princeton University Press, Princeton (1960)

[31] Chatfield, C.: The Analysis of Time Series: An Introduction, pp. 1–329. CRC Press, (2016)

[32] Garnett, R.: Bayesian Optimization. Cambridge University Press, (2023)

[33] Rasmussen, C.E., Williams, C.K.I.: Gaussian Processes for Machine Learning. The MIT Press, (2006)

[34] Srinivas, N., Krause, A., Kakade, S.M., Seeger, M.W.: Information-theoretic regret bounds for gaussian process optimization in the bandit setting. IEEE Transactions on Information Theory **58**(5), 3250–3265 (2012)

[35] Plataniotis, K., Hatzinakos, D.: Gaussian mixtures and their applications to signal processing, p. 36. CRC Press, (2000)

[36] Kandasamy, K., Schneider, J., Póczos, B.: High dimensional bayesian optimisation and bandits via additive models. In: Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37. ICML'15, pp. 295–304. JMLR.org, (2015)

[37] Gardner, J., Guo, C., Weinberger, K., Garnett, R., Grosse, R.: Discovering and Exploiting Additive Structure for Bayesian Optimization. In: Singh, A., Zhu, J. (eds.) Proceedings of the 20th International Conference on Artificial Intelligence and Statistics. Proceedings of Machine Learning Research, vol. 54, pp. 1311–1319. PMLR, (2017)

[38] Malkomes, G., Garnett, R.: Automating bayesian optimization with bayesian optimization. In: Proceedings of the 32th International Conference on Neural Information Processing Systems. NIPS'18, pp. 5988–5997. Curran Associates Inc., Red Hook, NY, USA (2018)

[39] Roman, I., Santana, R., Mendiburu, A., Lozano, J.A.: An experimental study in adaptive kernel selection for bayesian optimization. IEEE Access **7**, 184294–184302 (2019)

[40] Lázaro-Gredilla, M., Quiñonero-Candela, J., Rasmussen, C.E., Figueiras-Vidal, A.R.: Sparse spectrum gaussian process regression. J. Mach. Learn. Res. **11**, 1865–1881 (2010)

[41] Samo, Y.-L.K., Roberts, S.: Generalized Spectral Kernels (2015)

[42] Wilson, A.G., Gilboa, E., Nehorai, A., Cunningham, J.P.: Fast kernel learning for multidimensional pattern extrapolation. In: Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 2. NIPS'14, pp. 3626–3634. MIT Press, Cambridge, MA, USA (2014)

[43] Tobar, F.: Band-limited gaussian processes: the sinc kernel. Curran Associates Inc., Red Hook, NY, USA (2019)

[44] Vargas-Hernández, R., Gardner, J.: Gaussian Processes with Spectral Delta kernel for higher accurate Potential Energy surfaces for large molecules (2021)

[45] Parra, G., Tobar, F.: Spectral mixture kernels for multi-output gaussian processes. In: Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R. (eds.) Advances in Neural Information Processing Systems, vol. 30. Curran Associates, Inc., (2017)

[46] Altamirano, M., Tobar, F.: Nonstationary multi-output gaussian processes via harmonizable spectral mixtures. In: Camps-Valls, G., Ruiz, F.J.R., Valera, I. (eds.) Proceedings of The 25th International Conference on Artificial Intelligence and Statistics. Proceedings of Machine Learning Research, vol. 151, pp. 3204–3218. PMLR, (2022)

[47] Simpson, F., Boukouvalas, A., Cadek, V., Sarkans, E., Durrande, N.: The minecraft kernel: Modelling correlated gaussian processes in the fourier domain. In: Banerjee, A., Fukumizu, K. (eds.) Proceedings of The 24th International Conference on Artificial Intelligence and Statistics. Proceedings of Machine Learning

Research, vol. 130, pp. 1945–1953. PMLR, (2021)

[48] Kaelbling, L.P., Lozano-Pérez, T.: Learning composable models of parameterized skills. In: 2017 IEEE International Conference on Robotics and Automation (ICRA), pp. 886–893 (2017)

[49] Nguyen, Q.P., Dai, Z., Low, B.K.H., Jaillet, P.: Value-at-risk optimization with gaussian processes. In: Meila, M., Zhang, T. (eds.) Proceedings of the 38th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 139, pp. 8063–8072. PMLR, (2021)

[50] Boyd, S., Busseti, E., Diamond, S., Kahn, R.N., Koh, K., Nystrup, P., Speth, J., (2017)

[51] Cakmak, S., Astudillo, R., Frazier, P., Zhou, E.: Bayesian optimization of risk measures. In: Proceedings of the 34th International Conference on Neural Information Processing Systems. NIPS '20. Curran Associates Inc., Red Hook, NY, USA (2020)

[52] Owen, A.B.: Scrambling sobol' and niederreiter–xing points. Journal of Complexity **14**(4), 466–489 (1998)

[53] Surjanovic, S., Bingham, D.: Virtual Library of Simulation Experiments: Test Functions and Datasets (2025)

[54] Zhu, C., Byrd, R.H., Lu, P., Nocedal, J.: Algorithm 778: L-bfgs-b: Fortran subroutines for large-scale bound-constrained optimization. ACM Trans. Math. Softw. **23**(4), 550–560 (1997)

[55] Widom, H.: Asymptotic behavior of the eigenvalues of certain integral equations. Transactions of the American Mathematical Society **109**, 278–295 (1963)

[56] Zhu, H., Williams, C.K.I., Rohwer, R., Morciniec, M.: Gaussian regression and optimal finite dimensional linear models. Technical report, Birmingham (July 1997)

[57] Seeger, M.W., Kakade, S.M., Foster, D.P.: Information consistency of nonparametric gaussian process methods. IEEE Transactions on Information Theory **54**(5), 2376–2382 (2008)

Appendix A Primer on Gaussian Process

A.1 Gaussian Process

Gaussian process defines a distribution over functions $f : \mathcal{X} \to \mathbb{R}$, parameterized by a mean function $m(\cdot)$ and a covariance function $k(\cdot, \cdot)$:

$$f \sim \mathcal{GP}(m(x), k(x, x')),$$

where $x \in \mathcal{X}$ is an arbitrary input variable, and the mean function $m(x)$ and covariance function $k(x, x')$ are defined as

$$m(x) = \mathbb{E}[f(x)], \quad k(x, x') = \mathrm{cov}(f(x), f(x')).$$

For any finite set $\{x_1, \ldots, x_n\} \subset \mathcal{X}$, the function values $f = (f(x_1), f(x_2), \ldots, f(x_n))$ follow a multivariate Gaussian distribution:

$$(f(x_1), f(x_2), \ldots, f(x_n))^\top \sim \mathcal{N}(\boldsymbol{\mu}, \mathbf{K}),$$

where the $n \times n$ covariance matrix $\mathbf{K}$ has entries $K_{ij} = k(x_i, x_j)$, and the mean vector $\boldsymbol{\mu}$ has entries $\mu_i = m(x_i)$. GP depends on specifying a kernel function, which measures the similarity between input points. A typical example is the SE kernel in Eq. (A1). Functions drawn from a GP with this kernel are infinitely differentiable. Another example is the Matérn kernel with degrees of freedom $\nu = p + 1/2$:

$$k_\nu(r) = e^{-\frac{\sqrt{2\nu}r}{\ell}} \frac{\Gamma(p+1)}{\Gamma(2p+1)} \sum_{i=0}^{p} \frac{(p+i)!}{i!(p-i)!} \left(\frac{\sqrt{8\nu}r}{\ell} \right)^{p-i},$$

which is k-times differentiable only if $k < \nu$.

A.2 Conventional Kernels

Squared Exponential, also known as Radial Basis Function (RBF):

$$k_{\mathrm{SE}}(x, x') = \exp\left(-\frac{\|x - x'\|^2}{2\ell^2}\right), \tag{A1}$$

Rational Quadratic:

$$k_{\mathrm{RQ}}(\tau) = \left(1 + \frac{\tau^2}{2\alpha\ell^2}\right)^{-\alpha}, \tag{A2}$$

Periodic:

$$k_{\mathrm{PE}}(\tau) = \exp\left(-2\sin^2(\pi\tau\omega)/\ell^2\right). \tag{A3}$$

Matérn kernel with $\nu = \frac{3}{2}$:

$$k_{\nu=3/2}(\tau) = \left(1 + \frac{\sqrt{3}\tau}{\ell}\right) \exp\left(-\frac{\sqrt{3}\tau}{\ell}\right), \tag{A4}$$

Matérn kernel with $\nu = \frac{5}{2}$:

$$k_{\nu=5/2}(\tau) = \left(1 + \frac{\sqrt{5}\tau}{\ell} + \frac{5\tau^2}{3\ell^2}\right) \exp\left(-\frac{\sqrt{5}\tau}{\ell}\right). \tag{A5}$$

We refer readers to Rasmussen & Williams [33] for a comprehensive catalog of different kernels.

Appendix B Related Work of Spectral Mixture Kernel

B.1 BO with Kernel Designs

Conventional kernels, such as the SE or Matérn kernels, assume a fixed structure that may not capture the complexities of high-dimensional objective functions. To address this, several approaches have been proposed to design flexible kernels through automatic construction and adaptation. Kandasamy et al. [36] introduced an additive structure for the objective function, decomposing it into a sum of lower-dimensional functions. This approach allows for efficient optimization in high-dimensional spaces by reducing the effective dimensionality of the problem. Building upon this, Gardner et al. [37] proposed a method to automatically discover and exploit additive structures using a Metropolis-Hastings sampling algorithm. Their approach demonstrated improved performance by identifying hidden additive components in the objective function.

Malkomes et al. [38] introduced a dynamic approach to kernel construction during the BO process. They used a predefined grammar to iteratively combine basic kernels, enabling the exploration of a wide range of kernel compositions. While this approach offers greater flexibility, it also presents challenges in computational efficiency and risks overfitting due to the complexity of the kernel structures [23]. To address the limitations of static kernel choices, adaptive kernel selection strategies have been explored [39]. These strategies maintain a set of candidate kernels and dynamically select the most suitable one at each iteration based on six adaptive criteria. This adaptability allows the surrogate model to better respond to newly acquired data, which is particularly beneficial in the early stages of optimization when data is limited.

Despite the advancements in kernel design for BO, several challenges remain. The vast number of possible kernel compositions can make the search computationally expensive, and overly complex kernel structures may lead to overfitting and difficulties with hyperparameter inference.

B.2 Spectral Kernels

Recent developments in spectral kernels have notably enhanced the capabilities of GP modeling. A key advancement is the concept of sparse spectrum kernels [40], which are derived by sparsifying the spectral density of a full GP, resulting in a more efficient, sparse alternative. However, this class of kernels is prone to overfitting and

implicitly assumes that the covariance between two points does not decay as their distance increases, an assumption that may not hold in many real-world, non-periodic applications [41].

Wilson and Adams [23] defined a space of stationary kernels using a family of Gaussian mixture distributions in the Fourier domain to represent the spectral density in GP regression. This formulation was later extended to handle multidimensional inputs by incorporating a Kronecker structure for scalability [42]. However, the covariance functions induced by Gaussian mixture spectral densities are infinitely differentiable[2], which may be unrealistic for modeling certain physical processes [24]. While being introduced for regression tasks, the role of spectral mixture kernels in BO remains underexplored. In this paper, we propose and investigate the combination of Cauchy and Gaussian kernels into a spectral mixture, demonstrating state-of-the-art performance in optimization tasks.

The Sinc kernel [43] is another notable advancement, parameterizing the GP's power spectral density as a rectangular function. This kernel has shown exceptional performance in signal processing, particularly in tasks like band-limited frequency recovery and anti-aliasing. The Spectral Delta kernel [44] approximates stationary kernels through a finite sum of cosine basis functions, offering computational efficiency by avoiding Fourier integrals. However, its expressiveness is constrained by the discrete frequency sampling and fixed amplitude scaling.

To address multi-channel dependencies, researchers have generalized spectral kernels to multi-output GP regressions. The Multi-Output Spectral Mixture kernel [45] explicitly models cross-covariances by representing them as complex-valued spectral mixtures, capturing inter-channel correlations within a parametric framework. Building on this, Altamirano and Tobar [46] proposed a nonstationary harmonizable kernel family, enabling time-varying cross-spectral density estimation for nonstationary processes. The Minecraft kernel [47] further innovates by structuring cross-covariances via block-diagonal spectral representations with rectangular step functions, enhancing interpretability for high-dimensional outputs.

Appendix C Optimization Experiments

In further validate the efficacy of spectral mixture kernel in continuous domain, we validate our approach against several baselines across a wide range of optimization tasks.

C.1 Test Functions

We implement BO with CSM, GSM, and CSM+GSM kernels, and consider three sets of baseline models:

- **Off-the-shelf BO implementations**: SE, RQ, and MA52 kernels;
- **Advanced BO methods**: automatic Bayesian optimization (ABO) [38] and Adaptive Kernel Selection (ADA) [39];

[2] A kernel is "differentiable" means functions drawn from a GP with this kernel are differentiable.

- **Other spectral kernels**: spectral Delta kernel (SDK) [44] and SINC kernel (SINC) [43].

We consider a wider range of optimization problems with increasing dimensionality and complexity:

- **Synthetic problems**: Branin-2d, Hartmann-3d, Exponential-5d, Hartmann-6d, Exponential-10d, Rosenbrock-20d, Levy-30d;
- **Real-world problems**: Robot pushing-4d, Portfolio optimization-5d.

C.1.1 Synthetic Functions

Our first set of experiments involves test functions commonly used for optimization. The hyperparameters to be optimized are summarized in Table C1.

C.1.2 Simulated Problems

Robot Pushing

We consider a Robot Pushing problem widely used in recent literature [38, 48, 49]. This problem addresses an active learning task for the pre-image learning problem in robotic pushing. The goal is to determine an optimal pre-image for pushing the robot to a desired location, with the pushing action as the input and the distance from the goal location as the output. We test a 4-dimensional input function: robot location (r_x, r_y, r_θ), and pushing duration t_r.

Portfolio Optimization

Another real-world problem is portfolio optimization. Our goal is to tune the hyperparameters of a trading strategy so as to maximize investment return. We simulate and optimize the evolution of a portfolio over a period of four years using open-source market data.

Since the simulator CVXPortfolio [50] is expensive to evaluate, with each evaluation taking around 3 minutes, evaluating the performance of the various algorithms becomes prohibitively expensive. Therefore, following Cakmak et al. [51], we do not use the simulator directly in the experiments. Instead, we build a surrogate function obtained as the mean function of a GP trained using evaluations of the actual simulator across 3000 points chosen according to a Sobol sampling design [52].

C.2 Performance Metric

For portfolio optimization tasks, we directly evaluate performance using investment returns. For other optimization problems, we employ the log-optimality gap metric:

$$gap = \log(|f_n^* - f_{\text{opt}}|),$$

where f_n^* is the optimal solution found by the model, and f_{opt} is the true global optimum.

To quantify algorithmic improvements, we introduce two comparative metrics.

Relative improvement increase:

$$\frac{(RI_{\text{spectral}} - RI_{\text{baseline}})}{RI_{\text{baseline}}}, \quad RI = \frac{f_n^* - f_0}{f_{\text{opt}} - f_0};$$

Optimality gap reduction:

$$\frac{(OG_{\text{baseline}} - OG_{\text{spectral}})}{OG_{\text{baseline}}}, \quad OG = f_n^* - f_{\text{opt}}.$$

Table C1 Test functions used in our experiments. The analytic form of synthetic test functions, as well as the global minima, are available in [53].

Objective Function	Type	Dimension	Iterations	Input Domain
Branin	min	2	15	$x \in [-3, 3]^2$
Hartmann	min	3	30	$x \in [0, 1]^3$
Exponential	min	5	60	$x \in [-5.12, 5.12]^5$
Hartmann	min	6	80	$x \in [0, 1]^6$
Exponential	min	10	150	$x \in [-5.12, 5.12]^{10}$
Rosenbrock	min	20	200	$x \in [-2.048, 2.048]^{20}$
Levy	min	30	200	$x \in [-5, 5]^{30}$
Robot Pushing	min	4	150	$x_0 \in [-5, 5]$ (x-position), $x_1 \in [-5, 5]$ (y-position), $x_2 \in [1, 30]$ (pushing duration), $x_3 \in [0, 2\pi]$ (pushing angle),
Portfolio	max	5	200	$x_0 \in [0.1, 1000]$ (risk parameter), $x_1 \in [5.5, 8.0]$ (trade aversion parameter), $x_2 \in [0.1, 100]$ (holding cost multiplier), $x_3 \in [10^{-4}, 10^{-2}]$ (bid-ask spread), $x_4 \in [10^{-4}, 10^{-3}]$ (borrow cost)

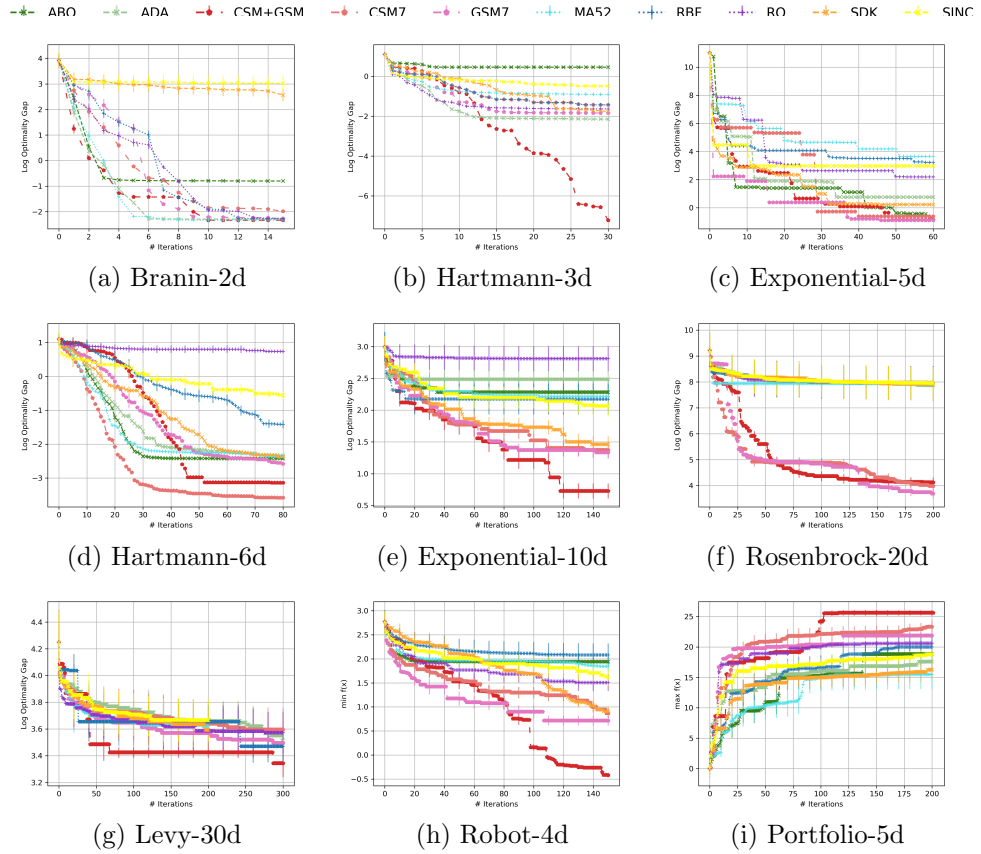
C.3 Result Analysis

Figure C1 illustrates the performance of different methods across various optimization tasks using UCB acquisition function. In nearly all experiments, our method consistently outperforms existing baselines in both low- and high-dimensional settings, achieving an increase of over 11% in relative improvement, and a reduction of 76% in optimality gap. We also provide a summarized result in Table 2.

For synthetic problems, the results indicate that both CSM and GSM outperform other methods in terms of convergence and optimality gap. Notably, CSM+GSM demonstrates superior performance across most baseline functions. This performance advantage grows with the dimensionality of the problem. For the conventional kernels, performance varies significantly depending on the task. Specifically, single kernels

perform as good on Branin-2d, a simple, low-dimensional problem where all methods show similar performance.

As we transition to more practical objective functions, the advantages of a more flexible and computationally efficient model become clear. Each type of spectral mixture kernel outperforms all other methods, with the one using both Gaussian and Cauchy components delivering the best performance. As a result, the log optimality gap is nearly two orders of magnitude smaller compared to standard BO methods using conventional kernels.



(a) Branin-2d (b) Hartmann-3d (c) Exponential-5d

(d) Hartmann-6d (e) Exponential-10d (f) Rosenbrock-20d

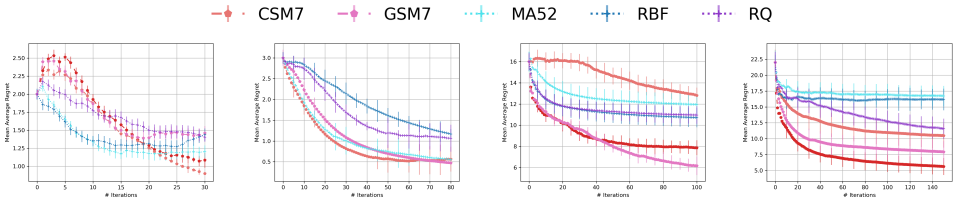
(g) Levy-30d (h) Robot-4d (i) Portfolio-5d

Fig. C1 Performance of different test functions and algorithms across 10 repetitions using UCB acquisition function.

C.4 Alternative Performance Metric

To align with our theoretical analysis, we also conducted experiments using the mean average regret metric, in addition to the original optimal value and optimality gap metrics. The results, depicted in Figure C2, demonstrate consistent advantages of the

proposed CSM and GSM kernels over conventional approaches (MA52, RBF, RQ) across multiple test functions.



(a) Hartmann-3d (b) Hartmann-6d (c) Robot-3d (D) Robot-4d

Fig. C2 Results for the average mean regret over iterations using UCB as the acquisition function.

C.5 Alternative Acquisition Functions

To address potential bias from acquisition function selection, we repeated experiments using EI and PI alongside UCB. Results show that our approach outperformed baselines regardless of the choice acquisition function in nearly all cases, as shown in Figures C3 and C4.

C.6 Different Number of Spectral Mixtures

To illustrate the performance differences resulting from varying the number of mixture components, we further compare the performance of spectral mixture kernels combining Cauchy and Gaussian components, as illustrated in Figure C5. Across all tasks, there is a noticeable trend where specific configurations of Cauchy and Gaussian components yield superior optimization performance, as indicated by the larger bubbles. Additionally, the performance varies across tasks, highlighting the interplay between kernel structure and task-specific characteristics. The results emphasize the utility of balancing Cauchy and Gaussian components to achieve optimal performance in diverse optimization scenarios.

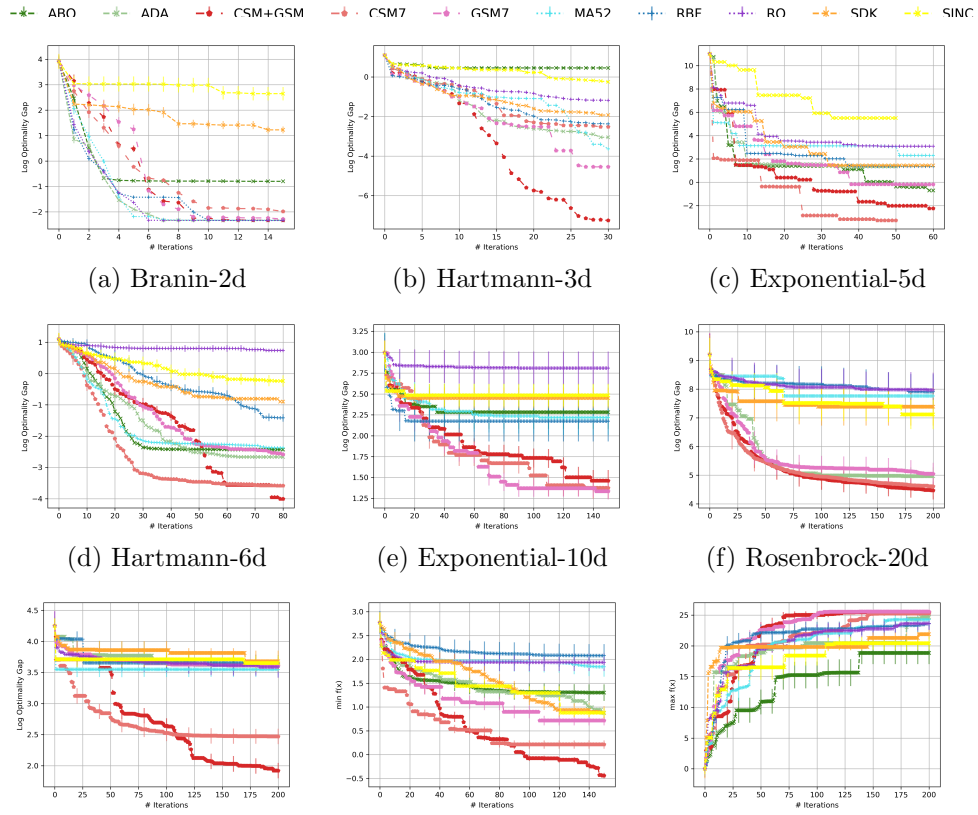
Appendix D Implementation Details

Approximation

To validate the flexibility of our proposed spectral mixture kernel, we considered the approximation of more sophisticated conventional kernels in Section 6.3. The parameters for source kernels that generate the data are shown in Table D2.

Optimization

The objective is to identify the global optimum of each test function within a limited number of evaluations. We optimize each acquisition function using the LBFGS method [54]. We conduct 10 trials for all problems. Mean and standard errors are



(a) Branin-2d (b) Hartmann-3d (c) Exponential-5d

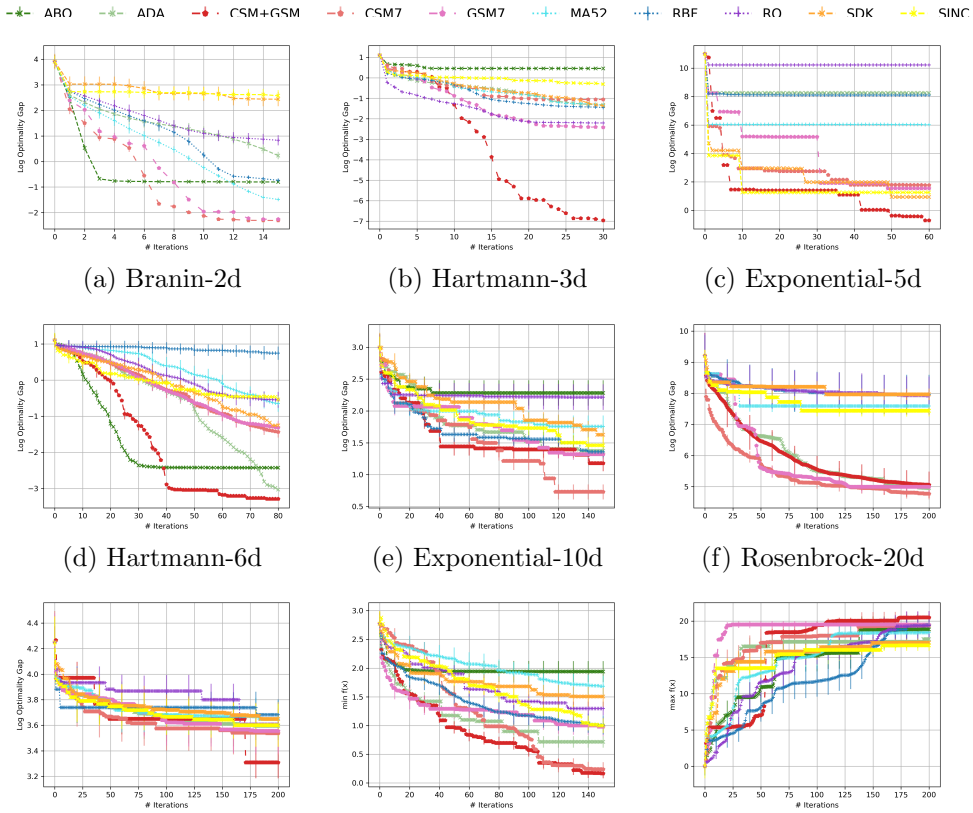
(d) Hartmann-6d (e) Exponential-10d (f) Rosenbrock-20d

(g) Levy-30d (h) Robot-4d (i) Portfolio-5d

Fig. C3 Optimization performance of different test functions and algorithms across 10 repetitions using EI acquisition function.

Table D2 Parameters for source kernels generating the training data. The lengthscale and outputscale parameters are the same for all MA52, MA32, and PE.

Parameters	Value
number of data	80
number of repetitions	10
number of mixture	10
lengthscale	5
outputscale	4
period	10



(a) Branin-2d (b) Hartmann-3d (c) Exponential-5d

(d) Hartmann-6d (e) Exponential-10d (f) Rosenbrock-20d

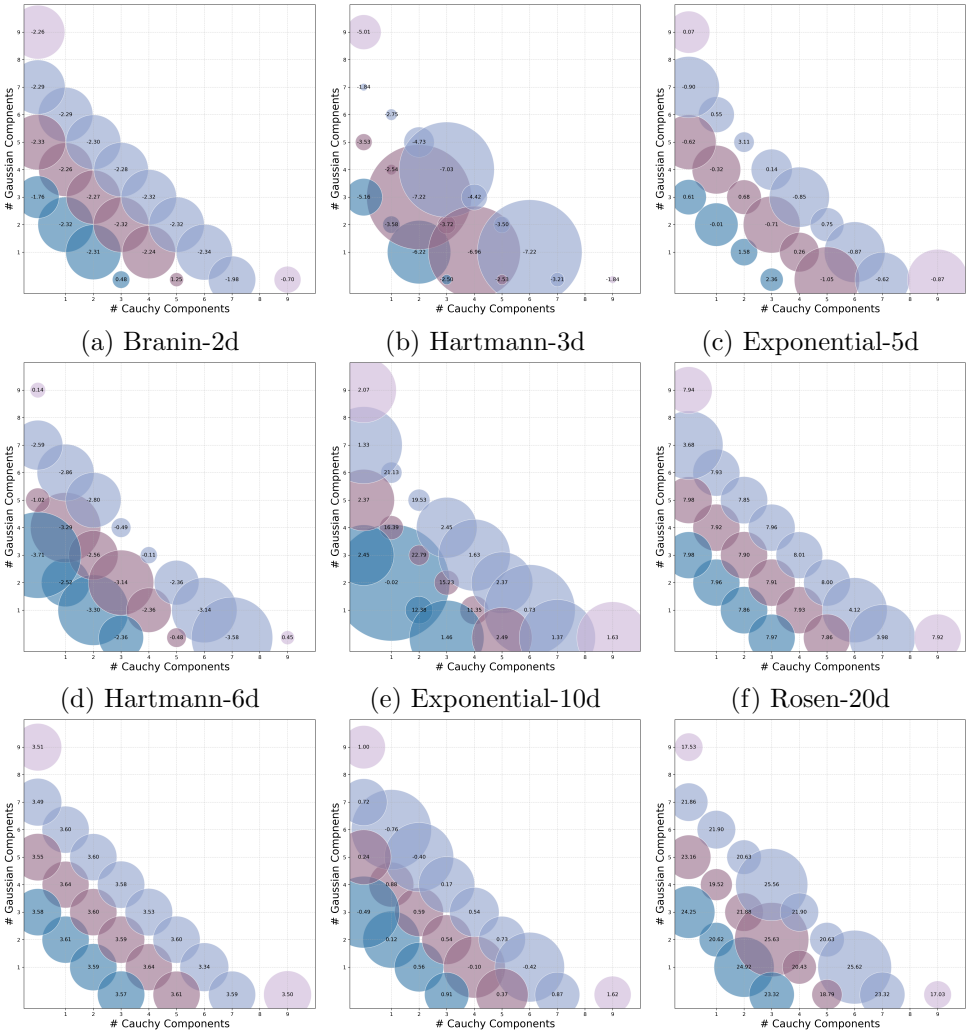
(g) Levy-30d (h) Robot-4d (i) Portfolio-5d

Fig. C4 Optimization performance of different test functions and algorithms across 10 repetitions using PI acquisition function. Implementation constraint: original implementation of ABO exclusively uses EI; reported PI results for ABO are EI-based to enable cross-algorithm comparison.

reported in all cases. All models employ UCB as the acquisition function. The kernel hyperparameters, as well as the observation noise, are inferred via marginal likelihood maximization after each function evaluation.

Computational Resources

All experiments were performed on an Ubuntu 20.04 server equipped with an AMD Ryzen 9 5950X CPU (16 cores, 32 threads), 125 GiB RAM, and an NVIDIA RTX 3090 GPU (24 GiB VRAM). The primary storage was a 1.8 TB NVMe SSD.



(a) Branin-2d (b) Hartmann-3d (c) Exponential-5d

(d) Hartmann-6d (e) Exponential-10d (f) Rosen-20d

(g) Levy-30d (h) Robot Pushing-4d (i) portfolio-5d

Fig. C5 Results for spectral mixture kernels with different number of Cauchy and Gaussian mixtures. The x-axis represents the number of Cauchy components, and the y-axis represents the number of Gaussian components. Different colors are used to distinguish mixtures based on their total number of components (3, 5, 7, 9). The bubble size represents the inverse of the optimality results, with larger bubbles indicating better performance.

Appendix E Technical Proofs

E.1 Proof of Theorem 2

Spectral density $S(s)$ of Cauchy distribution is given in Eq. (8), substituting it into Eq. (5):

$$k(\tau) = \int_{-\infty}^{\infty} \frac{1}{\pi\gamma\left[1 + \left(\frac{s-x_0}{\gamma}\right)^2\right]} e^{2\pi i s\tau}\, ds.$$

To simplify the integral, let:

$$u = \frac{s - x_0}{\gamma}, \quad ds = \gamma du, \quad s = \gamma u + x_0.$$

We have

$$k(\tau) = \int_{-\infty}^{\infty} \frac{1}{\pi\gamma(1+u^2)} e^{2\pi i(\gamma u + x_0)\tau} \gamma du = \frac{1}{\pi}\int_{-\infty}^{\infty} \frac{e^{2\pi i\gamma u\tau} e^{2\pi i x_0\tau}}{1+u^2} du = e^{2\pi i x_0\tau}\frac{1}{\pi}\int_{-\infty}^{\infty} \frac{e^{2\pi i\gamma u\tau}}{1+u^2} du.$$

Next, we calculate the standard integral using contour integration:

$$I(\tau) = \int_{-\infty}^{\infty} \frac{e^{2\pi i\gamma u\tau}}{1+u^2} du.$$

where the integrand is:

$$f(u) = \frac{e^{2\pi i\gamma u\tau}}{1+u^2}.$$

We are interested in integrating this function along the real line. To apply contour integration, we extend the function to the complex plane. The function $\frac{1}{1+u^2}$ has simple poles at $u = i$ and $u = -i$. The residues at these poles are:

$$\text{Res}(f, i) = \lim_{u\to i}(u-i)\frac{e^{2\pi i\gamma u\tau}}{1+u^2} = \frac{e^{-2\pi\gamma\tau}}{2i},$$

$$\text{Res}(f, -i) = \lim_{u\to -i}(u+i)\frac{e^{2\pi i\gamma u\tau}}{1+u^2} = \frac{e^{2\pi\gamma\tau}}{-2i}.$$

We use the residue theorem, which states that the integral of a meromorphic function around a closed contour is $2\pi i$ times the sum of the residues inside the contour. By closing the contour in the upper half-plane (since $e^{2\pi i\gamma u\tau}$ decays for large u in the upper half-plane), we get:

$$\int_{-\infty}^{\infty} \frac{e^{2\pi i\gamma u\tau}}{1+u^2} du = 2\pi i \cdot \text{Res}(f, i).$$

Thus,

$$I(\tau) = \pi e^{-2\pi|\gamma\tau|}.$$

Substituting this result into the expression for $k(\tau)$, we get:

$$k(\tau) = exp(2\pi i x_0 \tau) exp(-2\pi|\gamma\tau|) = [\cos(2\pi x_0\tau) + i\sin(2\pi x_0\tau)]\exp(-2\pi|\gamma\tau|).$$

Exploiting the symmetry of $S(s)$ gives

$$k(\tau) = \exp(-2\pi|\gamma\tau|)\cos(2\pi\tau x_0).$$

E.2 Proof of Theorem 3

We first state a theorem that gives an upper bound for information gain $\gamma(T)$ with a particular kernel k, given that $B_k(T_*)$ is known.

Theorem 5 (Srinivas et al. [34]) *Suppose that $D \subset \mathbb{R}^d$ is compact. Let $B_k(T_*) = \sum_{s>T_*} \lambda_h$, where $\{\lambda_h\}$ is the operator spectrum of k with respect to the uniform distribution over D. Pick $\tau > 0$, and let $n_T = C_4 T^\tau (\log T)$ with $C_0 = 2\mathcal{V}(D)(2\tau+1)$. Then, the following bound holds true:*

$$\gamma(T) \leq \frac{1/2}{1-e^{-1}} \max_{r=1,\ldots,T} \left(T_* \log\left(\frac{r n_T}{\sigma^2}\right) + C_0\sigma^{-2}(1-r/T)(\log T)\left(T^{\tau+1}B_k(T_*)+1\right)\right) + \mathcal{O}(T^{1-\tau/d})$$

for any $T_* \in \{1,\ldots,n_T\}$.

To obtain the tail bound on $B_k(T_*) = \sum_{h>T_*} \lambda_h$, we further draw on a theorem of Widom [55], which gives the asymptotic behavior of the operator spectrum $\{\lambda_h\}$.

Assumption 6 *We assume that the covariate distribution μ has a bounded density, such that*

$$\int I_{\{\|x\|\leq T\}} \mu(x)^{d/(2\nu+d)}\, dx \leq \tilde{C},$$

where $\tilde{C}$ is a constant independent of $T > 0$.

Assumption 6 provides a controlled growth of $\mu(x)$, ensuring that the covariates are not overly concentrated in any specific region, which could lead to numerical instabilities or biased estimation. Distributions that satisfy this assumption include uniform, Gaussian, and truncated power-law distributions, among others.

Theorem 7 (Widom [55]) *Define*

$$\psi(\varepsilon) = (2\pi)^{-d} \int I_{\{\mu(x)S(s)>(2\pi)^{-d}\varepsilon\}}\, dx\, d\omega, \tag{E6}$$

where $S(s)$ is the spectral density and $h = h(\varepsilon) = \min\{h' \mid \lambda_{h'} > \varepsilon\}$. we have

$$\psi(\varepsilon) \sim h(\varepsilon), \quad \varepsilon \to 0.$$

where both $\psi(\varepsilon)$ and $h(\varepsilon)$ are non-increasing and $h(\varepsilon)$ is unbounded as $\varepsilon \to 0$.

When $\psi(\varepsilon)$ is strictly decreasing and $\psi^{-1}(h+o(h)) \sim \psi^{-1}(h)$ for $h \to \infty$. The asymptotic distribution of eigenvalues λ_h satisfies

$$\lambda_h \sim \psi^{-1}(h), \quad h \to \infty.$$

Theorem 5 gives a upper bound on $\gamma(T)$, provided that $B_k(T_*) = \sum_{h>T_*} \lambda_h$ is known. Theorem 7 gives the asymptotic behavior of λ_h, so we can compute $B_k(T_*)$. Following this path, we first show how λ_h is derived.

Lemma 8 *Let $K(\tau)$ be the CSM kernel with d-dimensioanl inputs. Define the bounded support measure μ_T with density $\mu_T(x) = I_{\{\|x\| \le T\}}\mu(x)$, and let $\{\lambda_h\}$ be the spectrum of $K(\tau)$ w.r.t. μ_T. Then, for all $T > 0$ large enough, there exists a h_0 such that*

$$\lambda_h \le Ch^{-(d+1)/d} \quad \forall h \ge h_0.$$

Here, C is a constant independent of T.

Proof The joint distribution of d independent standard Cauchy is given by

$$\lambda(\eta) = \frac{\Gamma\left(\frac{d+1}{2}\right)}{\pi^{\frac{d+1}{2}}} \cdot \frac{1}{(1+||\eta||^2)^{\frac{d+1}{2}}}.$$

To upper bound Eq. (E6) for the measure μ_T, we first transform $\psi_T(\varepsilon)$ into polar coordinates. Recall that $d\omega = A^{d-1}\eta^{d-1}d\eta d\sigma$ with $d\sigma$ the uniform distribution on the unit sphere, and $A^{d-1} = 2\pi^{d/2}/\Gamma(d/2)$. If $y = 1 + ||\eta||^2, q = (d+1)/2$, then

$$\psi_T(\varepsilon) = C_1 \int_0^\infty \int_{\|x\| \le T} I_{\{y^{-q}\mu(x) > c_1\varepsilon\}} \eta^{d-1}\, dx\, d\eta, \tag{E7}$$

where

$$C_1 = \frac{2^{1-d}\pi^{-d/2}}{\Gamma(d/2)}, \quad c_1 = \pi.$$

Let $\rho = (c_1\varepsilon)^{-1}$. Note that $\rho \to \infty$ as $\varepsilon \to 0$. Now,

$$\eta^{d-1}(d\eta) = \frac{1}{2}(y-1)^{\frac{d}{2}-1}\, dy,$$

so that

$$\psi_T(\varepsilon) = C_2 \int_{\|x\| \le T} \int_1^\infty I_{\{y^q < \rho\mu(x)\}} G(y)\, dy\, dx \tag{E8}$$

with $C_2 = \frac{1}{2}C_1, G(y) = (y-1)^{\frac{d}{2}-1}$. Integrating out y, we have that

$$\psi_T(\varepsilon) \sim C_2 \frac{2}{d}\rho^{\frac{d}{2q}} \int I_{\{\|x\| \le T\}}\mu(x)^{\frac{d}{2q}}\, dx.$$

The integration leaves us with $(\rho\mu(x))^{1/q} - 1)^{2/d}$. We can use the binomial theorem in order to write that as a polynomial in $(\rho\mu(x))^{1/q}$, which is dominated by the highest degree term as $\varepsilon \to 0$. Moreover, since $(y-1)^{2/d} \le y^{2/d}$ for $y \ge 1$, the right-hand side is also an exact upper bound once $\rho \ge \rho_0 := \sup\{\mu(x)^{-1} \mid \|x\| \le T\}$. Note that $q = (d+1)/2$ If $C_3 = C_2\frac{2}{d}\tilde{C}c_1^{-2q/d}$, then $\psi_T(\varepsilon) \le C_3(1+o(1))\varepsilon^{-2q/d}$ as $\varepsilon \to 0$. Widom's theorem gives that $s - 1 \le C_3(1+o(1))(\lambda_h)^{-d/2q}$. The lemma follows by solving for λ_h. □

Therefore, $B_k(T_*) = \sum_{h>T_*} = \mathcal{O}\left(T_*^{1-(d+1)/d}\right)$. Following Srinivas et al. [34], we choose $T_* = (Tn_T)^{d/(1+d)}(\log(Tn_T))^{-(1+d)/d}$, so that the upper bound $\gamma(T)$ becomes:

$$\max_{r=1,\ldots,T}\left(T_*\log\left(\frac{rn_T}{\sigma^2}\right) + \sigma^{-2}(1-r/T) \times (C_3T_*(\log(Tn_T)) + C_0(\log T))\right) + \mathcal{O}(T^{1-\tau/d})$$

with $n_T = C_0 T^\tau (\log T)$. The maximum of upper bound $\gamma(T)$ over r is $\mathcal{O}(T_* \log(T n_T)) = \mathcal{O}(T^{(\tau+1)d/(d+1)}(\log T))$. Next, we choose $\tau = d/(d^2+d+1)$ to match this term with $\mathcal{O}(T^{1-\tau/d})$. Plugging this in, we can obtain $\gamma(T)$.

Lemma 9 *Let $K(\tau)$ be the GSM kernel with d-dimensional inputs. Then, for all $T > 0$ large enough, there exists a h_0 such that*

$$\lambda_h \leq C \cdot B^{h^{1/d}} \quad \forall h \geq h_0.$$

where $B < 1$, C is a constant independent of T.

Proof The joint distribution of d independent standard Gaussian distributions is given by

$$\lambda(\eta) = \frac{1}{(\sqrt{2\pi})^d} \exp(-\frac{||\eta||^2}{2})$$

For covariate distributions $\mu(x)$ that satisfy Assumption 6, the same decay rate holds for λ_h, while the constants might change [34]. Therefore, without loss of generality, we assume $\mu(x) = N(x|0, \boldsymbol{I})$. In this case, the eigenexpansion of k is known explicitly for $d = 1$ [56]:

$$\lambda_h = (\frac{1}{2A})^{1/2+h}, \quad A = \frac{1}{4} + \frac{1}{2} + \sqrt{\frac{5}{16}}.$$

Following the analysis of Seeger et al [57], in cases where $d > 1$, a tight bound can be obtained on λ_h:

$$\lambda_h \leq (\frac{1}{2A})^{d/2} (B)^{h^{1/d}}$$

where $B < 1$, A is a constant independent of T.

Because the same decay rate holds for any $\mu(x)$ satisfying Assumption 6. The lemma follows by keeping the decay rate $B^{h^{1/d}}$ while changing the constant term. □

Upper bound on information gain *w.r.t* λ_h satisfying Lemma 9 is given in Srinivas et al. [34]. This completes our proof for Theorem 3.