A Bayesian Framework Coupling Discrete 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

Content

1 • Background

2 • Method Details

3 • Results

4 • Conclusion

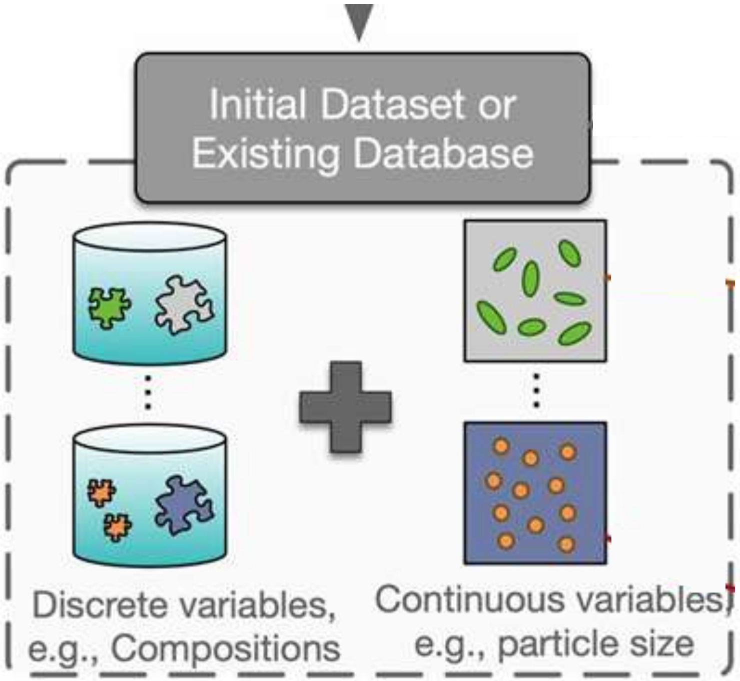
Background

Challenges in General Experimental Design

High-dimensional design space: The number of parameters grows rapidly, making exhaustive search impractical.

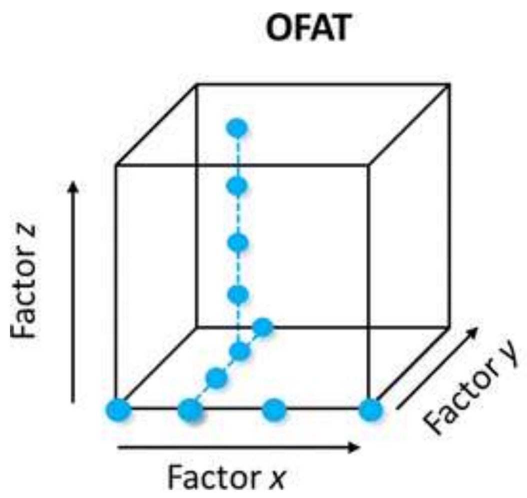
Mixed discrete and continuous variables: The coexistence of categorical (e.g., catalyst type) and continuous parameters (e.g., temperature, pressure) complicates modeling.

High experimental cost & limited samples: Each experiment is time- and resource-intensive, leading to data scarcity and unreliable model training.



High-dimensional design space

Background

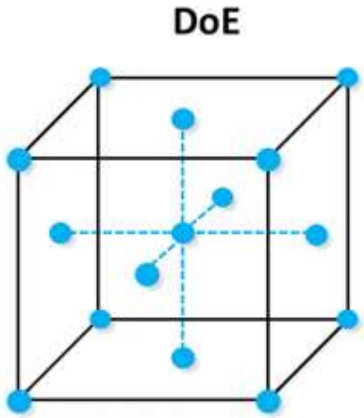


One-Factor-at-a-Time (OFAT) Method：

1. Change only **one** parameter per experiment while keeping all others constant.
2. Observe the effect of that **single** factor on the outcome.

Limitations:

1. Fails to capture **interaction effects** between variables.
2. Becomes inefficient or infeasible in **high-dimensional** design spaces.



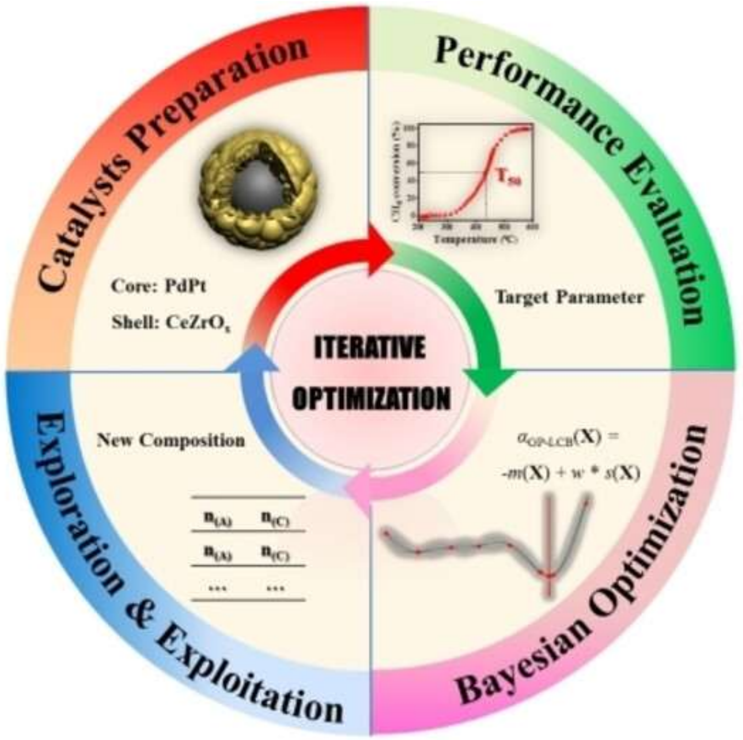
Design of Experiments (DOE) Method：

1. Systematically varies multiple factors simultaneously according to a statistical design
2. Analyzes both main effects and interactions among factors.

Limitations:

1. The number of experiments still grows exponentially with dimensionality.
2. Struggles with mixed discrete and continuous variables common in materials or chemical systems.

Bayesian Optimization



1. Initial points
2. Fit **Surrogate Model**
3. Optimize **Acquisition Function**
4. Query next point
5. Add new point to dataset
6. Return to Step 2 until finish

Adaptive

Sample-efficient

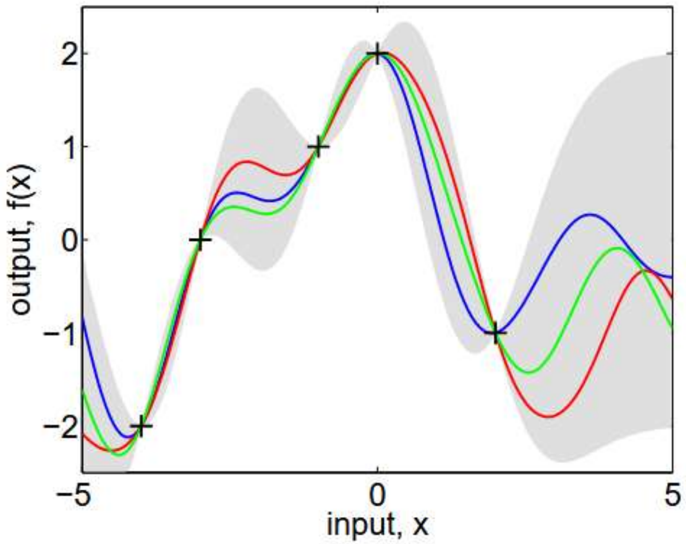
Uncertainty-aware

Details in BO

Surrogate Model (Gaussian Process)

Consider a finite collection of data pairs D_n = (X,**y**). Under the assumption of linear regression model with Gaussian noise, we have

$$\begin{bmatrix} \mathbf{y} \\ \mathbf{f}_* \end{bmatrix} \sim \mathcal{N}\left(\mathbf{0}, \begin{bmatrix} K(X,X)+\sigma_n^2 I & K(X,X_*) \\ K(X_*,X) & K(X_*,X_*) \end{bmatrix}\right).$$



Where σ_n denotes the standard variance of noise; K denotes the kernel matrix, which represents the property of the space. f_* is the value we want to predict, like performance of catalysts recipe.

$$\bar{\mathbf{f}}_* \triangleq \mathbb{E}[\mathbf{f}_*|X,\mathbf{y},X_*] = K(X_*,X)[K(X,X)+\sigma_n^2 I]^{-1}\mathbf{y},$$

$$\mathrm{cov}(\mathbf{f}_*) = K(X_*,X_*) - K(X_*,X)[K(X,X)+\sigma_n^2 I]^{-1}K(X,X_*).$$

Details in BO

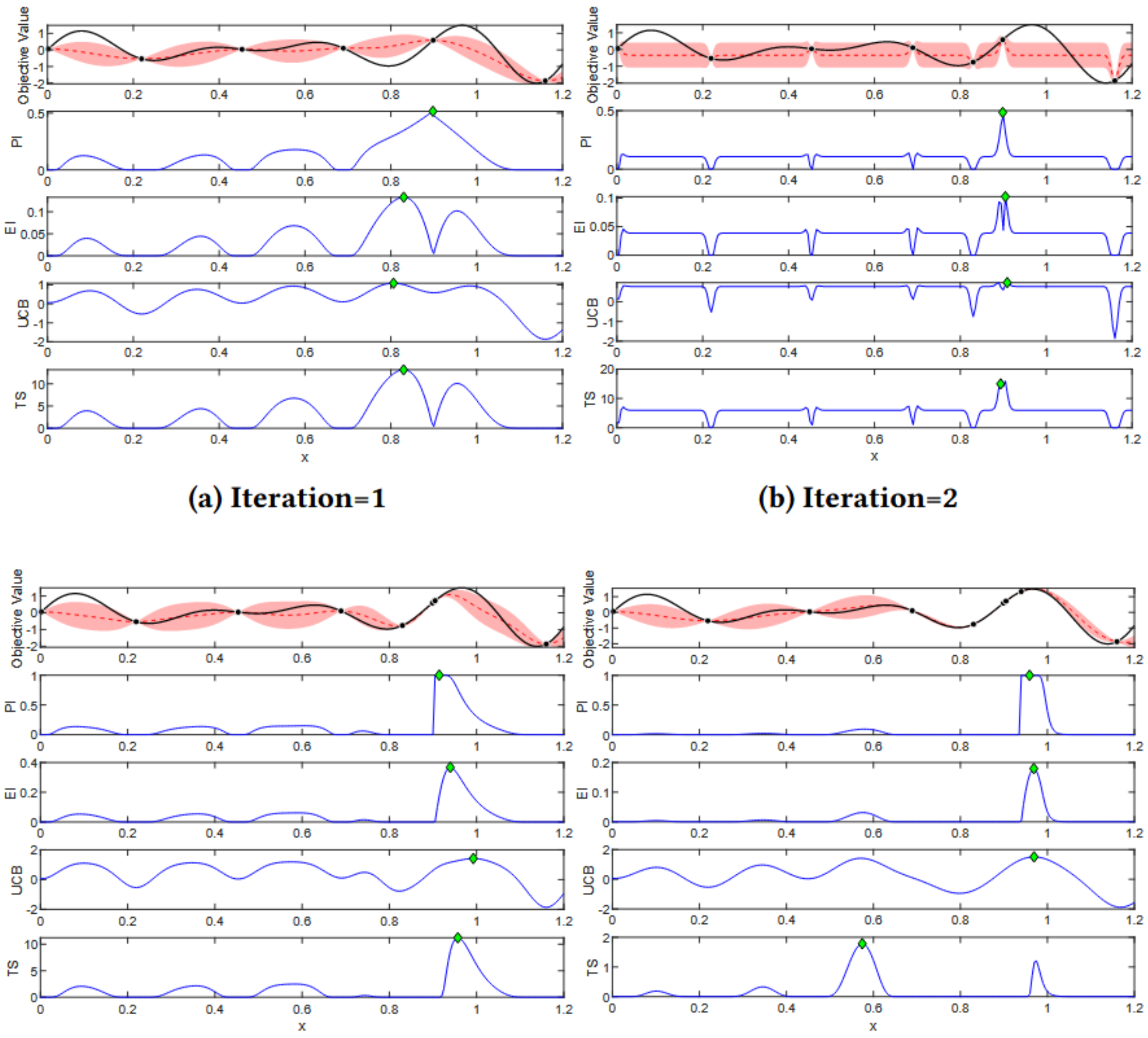
Acquisition Function (AFs)

Acquisition functions are the utility functions that guide the search to reach the optimum of the objective function by identifying where to sample next. The guiding principle behind AFs is to strike a balance between exploration and exploitation.

$$\mathrm{PI}(\mathrm{x}) = P\left(f(\mathrm{x}) \geq f^*\right) = \Phi\left(\frac{\mu(\mathrm{x}) - f^*}{\sigma(\mathrm{x})}\right)$$

$$\begin{aligned}\mathrm{EI}(\mathrm{x}) &= \mathbb{E}\left[\max\left(0, f^* - f(\mathrm{x})\right)\right] \\ &= \left(f^* - \mu(\mathrm{x})\right)\Phi\left(\frac{f^* - \mu(\mathrm{x})}{\sigma(\mathrm{x})}\right) + \sigma(\mathrm{x})\phi\left(\frac{f^* - \mu(\mathrm{x})}{\sigma(\mathrm{x})}\right)\end{aligned}$$

$$\mathrm{UCB}(\mathrm{x}) = \mu(\mathrm{x}) + \beta\sigma(\mathrm{x})$$



(a) Iteration=1 (b) Iteration=2

(c) Iteration=3 (d) Iteration=4

Limitation of BO

Handling Mixed Variables

- Real experiments involve both **discrete and continuous** parameters (e.g., catalyst type, temperature).
- Standard GP-based BO **assumes smooth continuous spaces**, causing instability with categorical inputs.

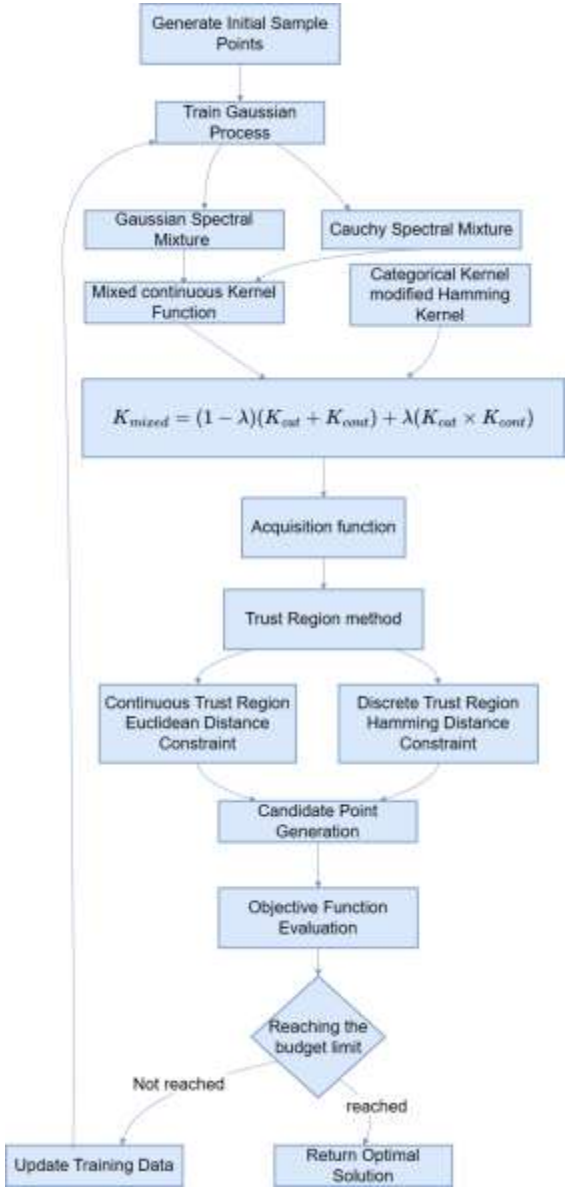
Balancing Multiple Objectives

- Experimental optimization often requires improving **several conflicting metrics** (e.g., activity vs. stability).
- However, acquisition design and convergence in Pareto fronts become **more complex** as objectives increase.

Integrating Domain Knowledge

- Incorporating **priors** into surrogate models improves efficiency but risks introducing bias.
- Hybrid BO frameworks aim to balance **data-driven** learning and **expert-guided** constraints.

SMKBO Methods Details



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

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

Where ψ *is a positive finite Borel measure on* $\mathbb{R}^d$.

The measure ψ is called ***spectral measure*** of k, if ψ has a density **S**, then **S** is referred to as ***spectral density*** or ***power spectrum*** of k

$$k(\tau)=\int S(s)\exp(2\pi i s\ \tau)\,ds, \qquad S(s)=\int k(\tau)\exp(-2\pi i s\ \tau)\,d\tau.$$

SMKBO Methods Details

Gaussian Spectral Density:

$$\phi_g(s) = \sum_{q=1}^{Q_g} w_q \mathrm{N}\,(s; \boldsymbol{\mu}_q, \boldsymbol{\Sigma}_q), \quad S(s) = \frac{\phi_g(s) + \phi_g(-s)}{2}, \longrightarrow k_g(\tau) = \sum_{q=1}^{Q_g} w_q \exp\left(-2\pi^2 \tau \;\; \boldsymbol{\Sigma}_q \tau\right) \cos\left(2\pi\tau \;\; \boldsymbol{\mu}_q\right).$$

Cauchy Spectral Density:

$$\phi_c(s) = \sum_{q=1}^{Q_c} w_q \mathrm{C}(s; \mathbf{x}_{0q}, \boldsymbol{\gamma}_q), \quad \mathrm{C}(s; x_0, \gamma) = \frac{1}{\pi} \frac{\gamma}{(s - x_0)^2 + \gamma^2}, \longrightarrow k_c(\boldsymbol{\tau}) = \sum_{q=1}^{Q_c} w_q \exp\left(-2\pi \,|\, \boldsymbol{\tau} \;\; \boldsymbol{\gamma}_\mathbf{q} \,|\right) \cos\left(2\pi \boldsymbol{\tau} \;\; \mathbf{x}_{\mathbf{0q}}\right).$$

Cauchy-Gaussian Spectral Mixture:

$$k_{cg}(\boldsymbol{\tau}) = \sum_{q=1}^{Q_g} w_q^g \exp\left(-2\pi^2 \boldsymbol{\tau} \;\; \boldsymbol{\Sigma}_q \boldsymbol{\tau}\right) \cos\left(2\pi\boldsymbol{\tau} \;\; \boldsymbol{\mu}_q\right) + \sum_{q=1}^{Q_c} w_q^c \exp\left(-2\pi \,|\, \boldsymbol{\tau} \;\; \boldsymbol{\gamma}_q \,|\right) \cos\left(2\pi\boldsymbol{\tau} \;\; \mathbf{x}_{0q}\right),$$

SMKBO Methods Details

For categorical inputs, we modify the Hamming kernel: $k_h(\mathbf{h},\mathbf{h}') = \exp\left(\frac{1}{d_h}\sum_{i=1}^{d_h}\ell_i\delta(h_i,h_i')\right),$

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

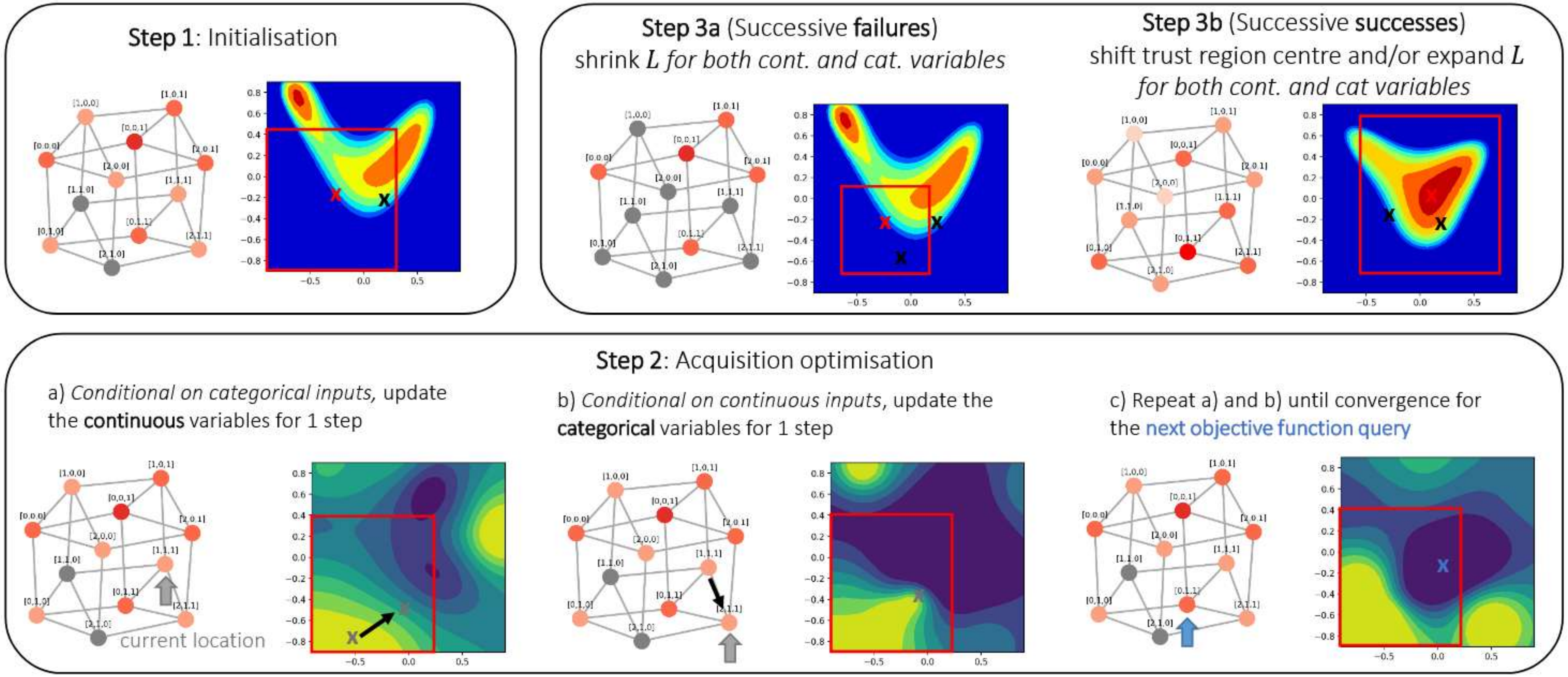
$$k(\mathbf{z},\mathbf{z}') = \lambda\Big(k_x(\mathbf{x},\mathbf{x}')k_h(\mathbf{h},\mathbf{h}')\Big) + (1-\lambda)\Big(k_h(\mathbf{h},\mathbf{h}') + k_x(\mathbf{x},\mathbf{x}')\Big),$$

$\lambda \in [0,1]$ is a trade-off parameter

SMKBO Methods Details

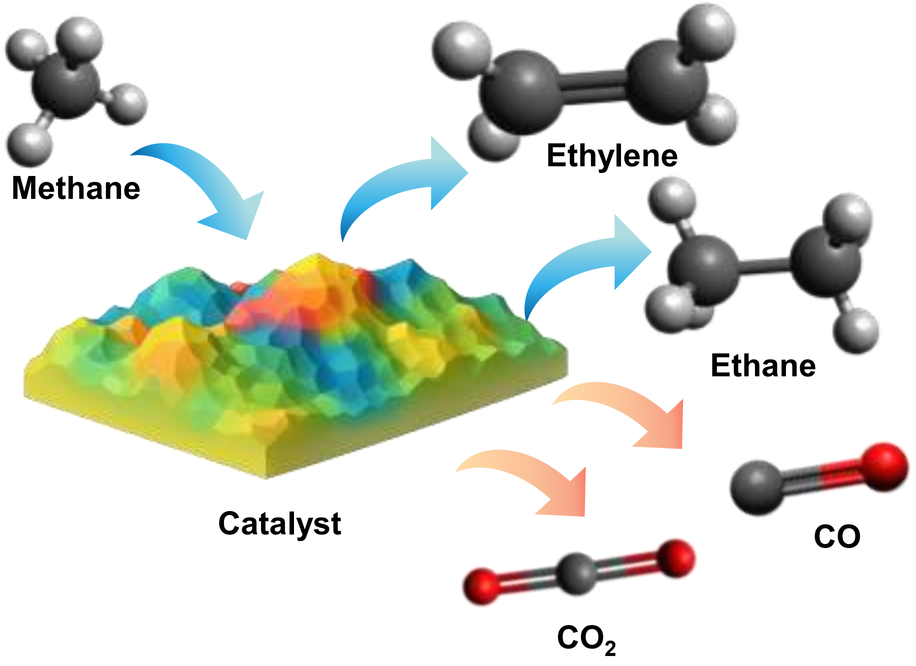
Trust Region & Alternating Optimization

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

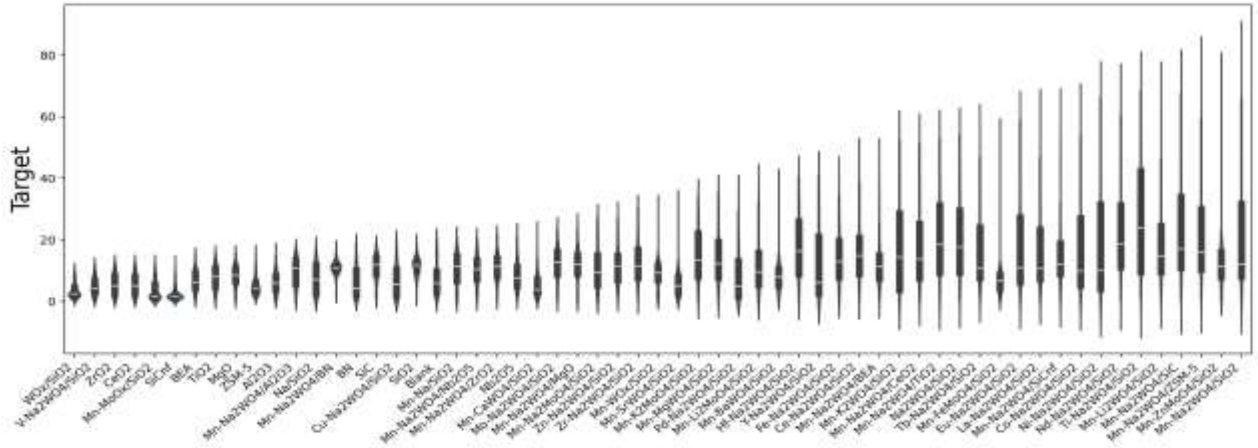


SMKBO Methods Details

Oxidative Coupling of Methane



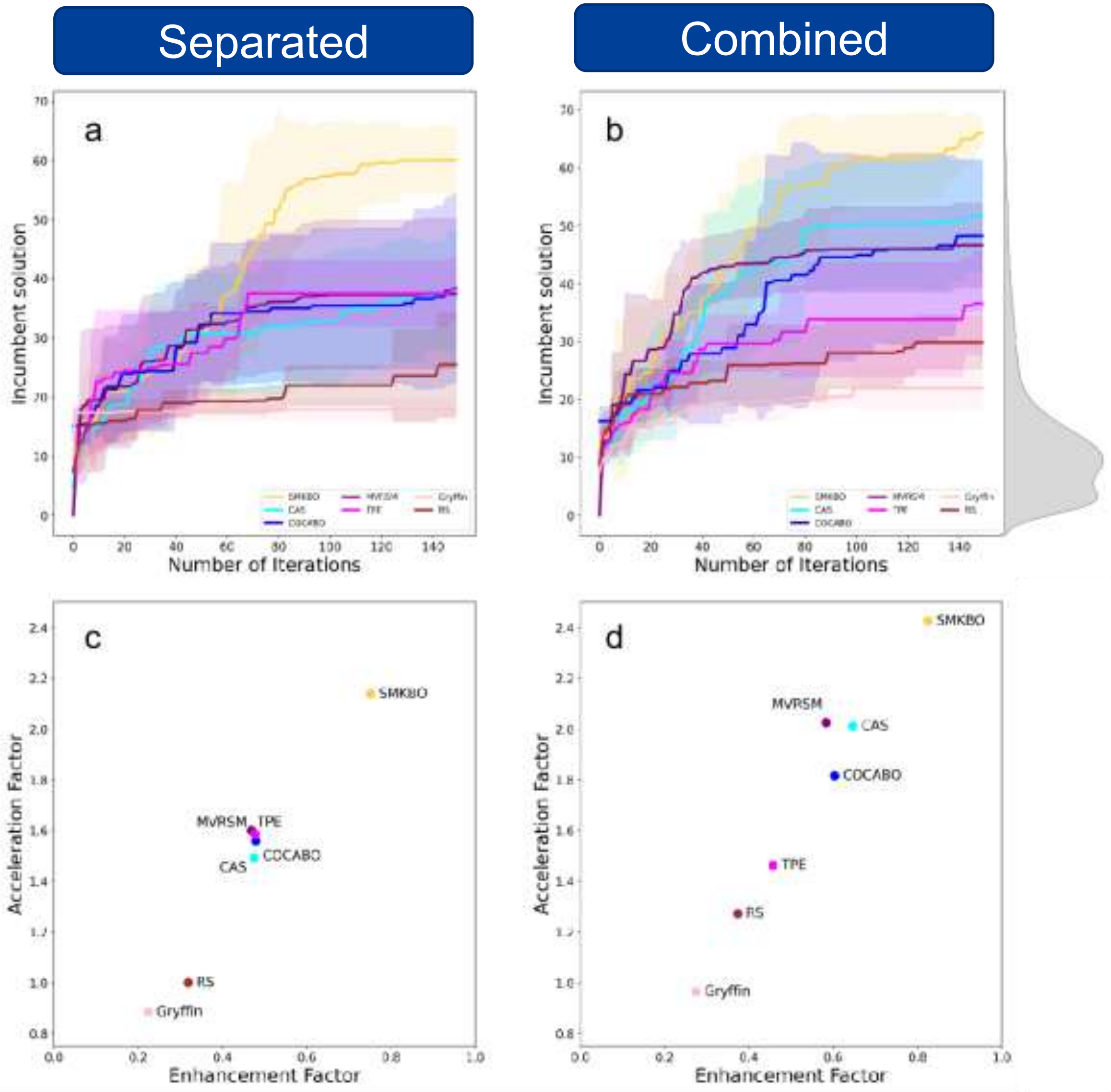
$$Target = Conv_{CH_4}\frac{2Y_{C_2H_4}+Y_{C_2H_6}}{2Y_{CO_2}+Y_{CO}}$$



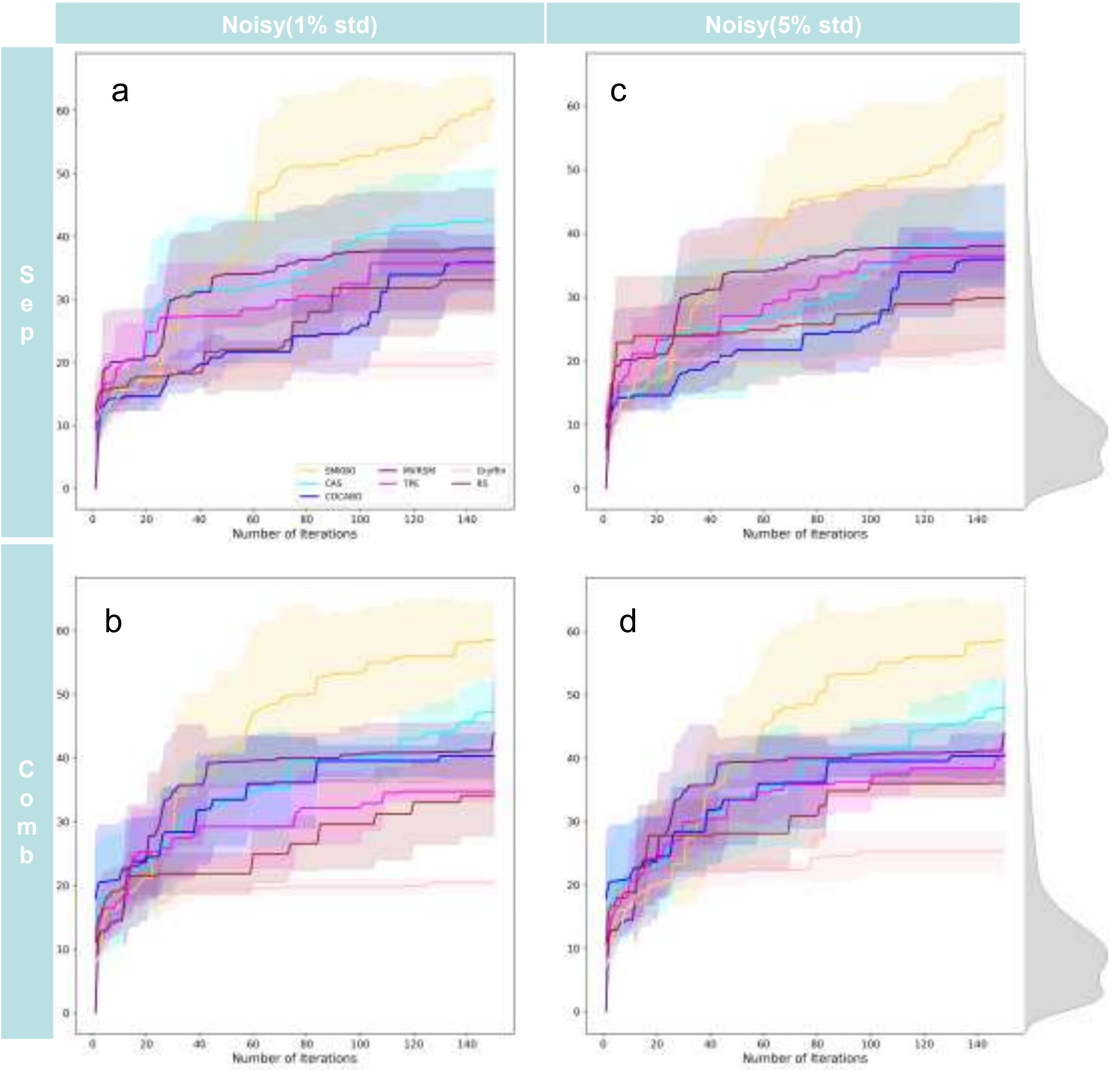
Results

To quantify the performance of optimization methods, the concepts of **Enhancement Factor (EF)** and **Acceleration Factor (AF)** are introduced.

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



Results



*Noise test setup: Gaussian noise (μ = 0, σ = 1% or 5%) of the global optimum target (**69.9**) added to target value (69.9) at each iteration.*

Key observations:

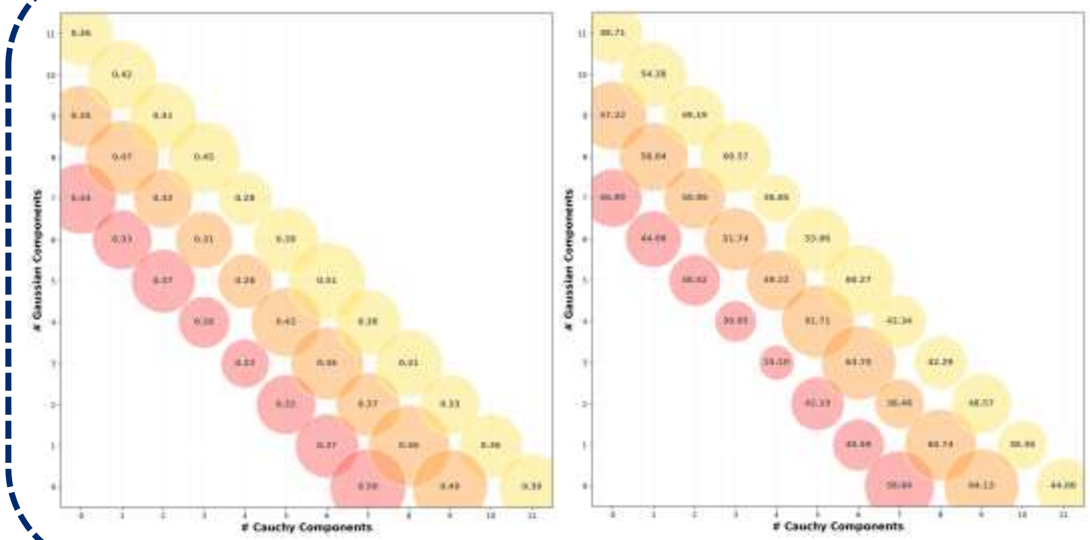
• SMKBO maintains higher robustness at 1% and 5% noise compared with others.

Results

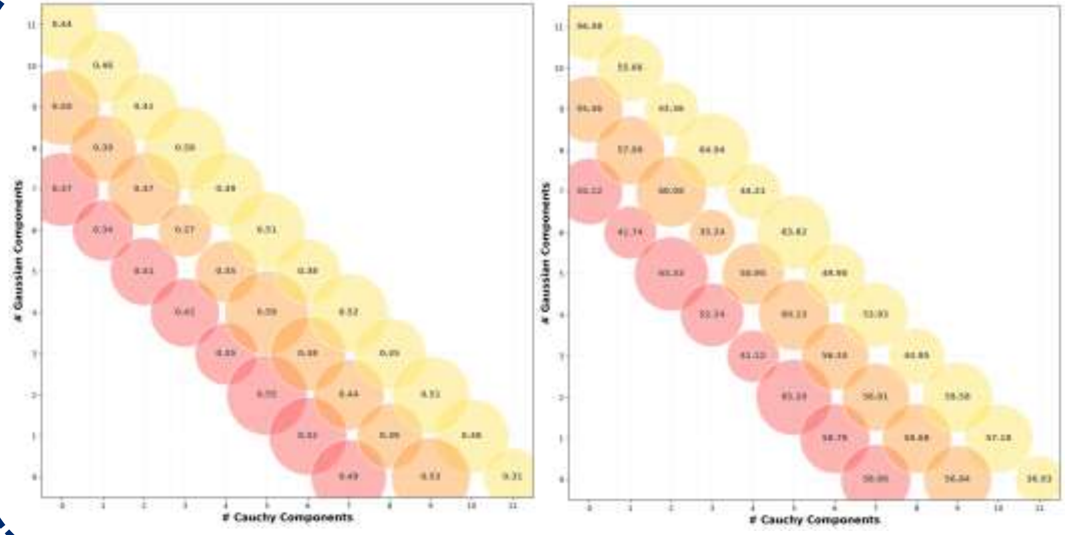
The results after 80 iterations using different combinations of CSK and GSK.

Remaining stable across different compositions

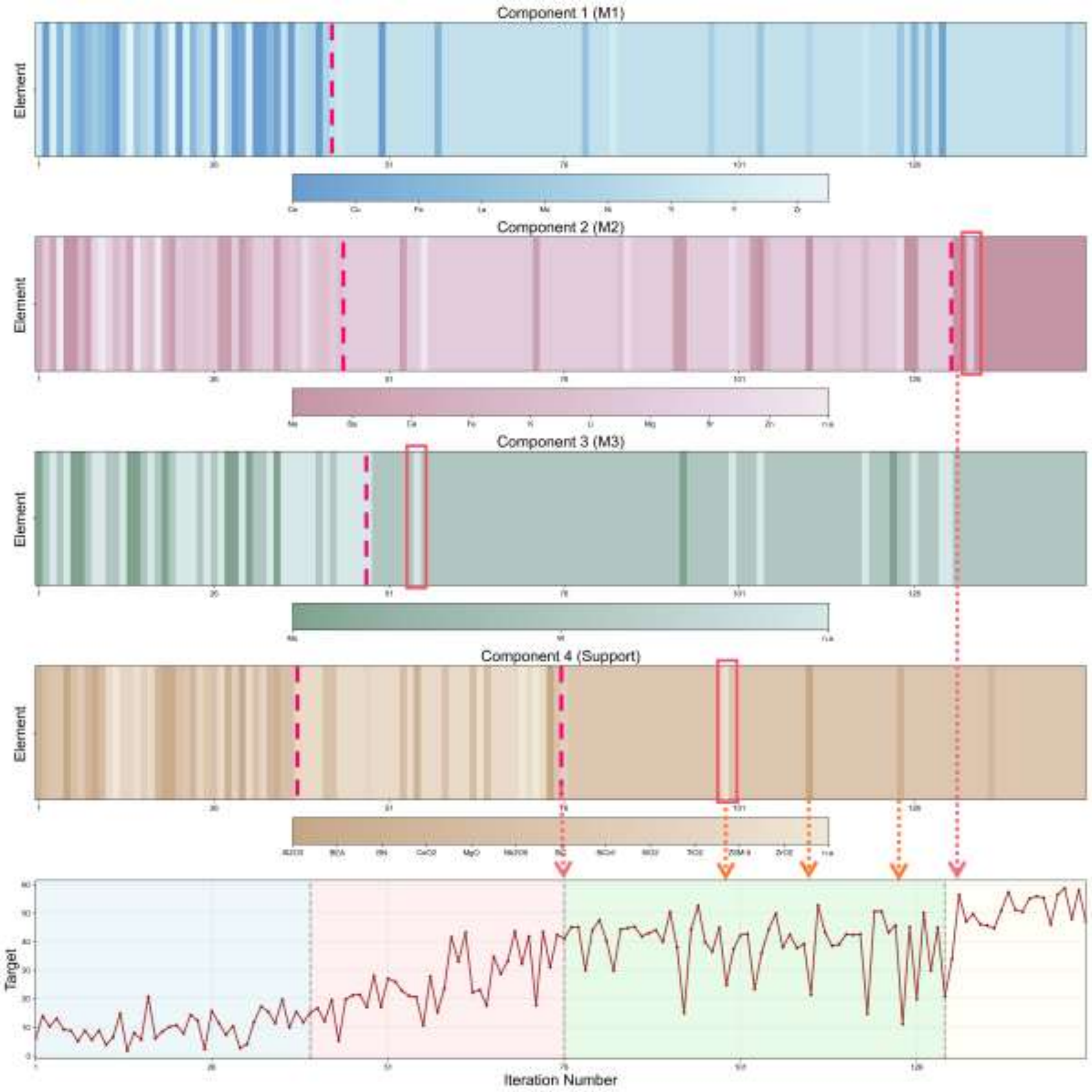
Separated



Combined



Results



1. Explore first, Exploit later.

2. A certain level of exploration is maintained even after convergence.

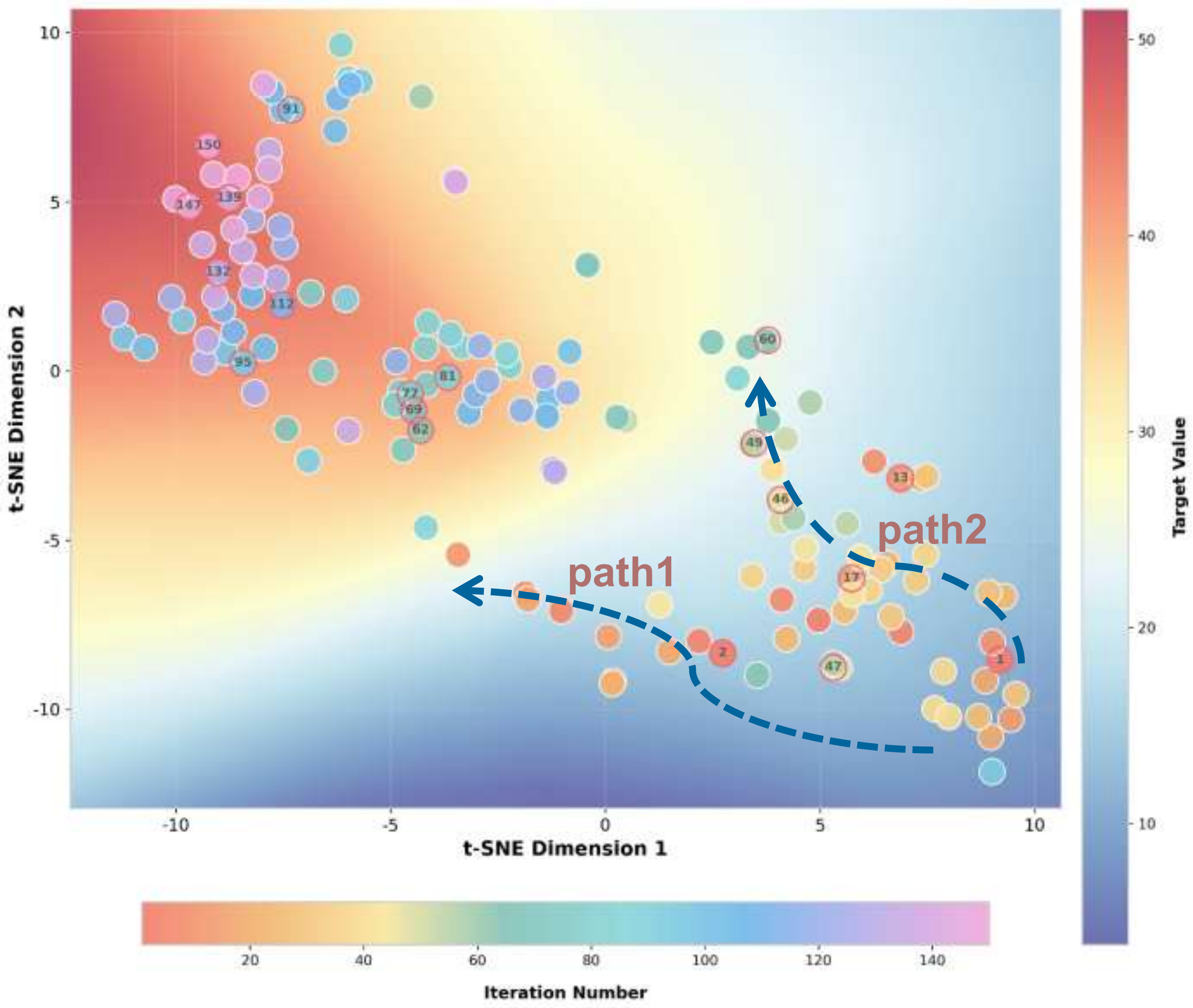
3. Component 2 and the Support temporarily converged to a local optimum before escaping to explore a better region of the search space.

4. verifications are performed after switching the convergence criterion.

Results

1. The algorithm progresses diagonally from low to high target values.

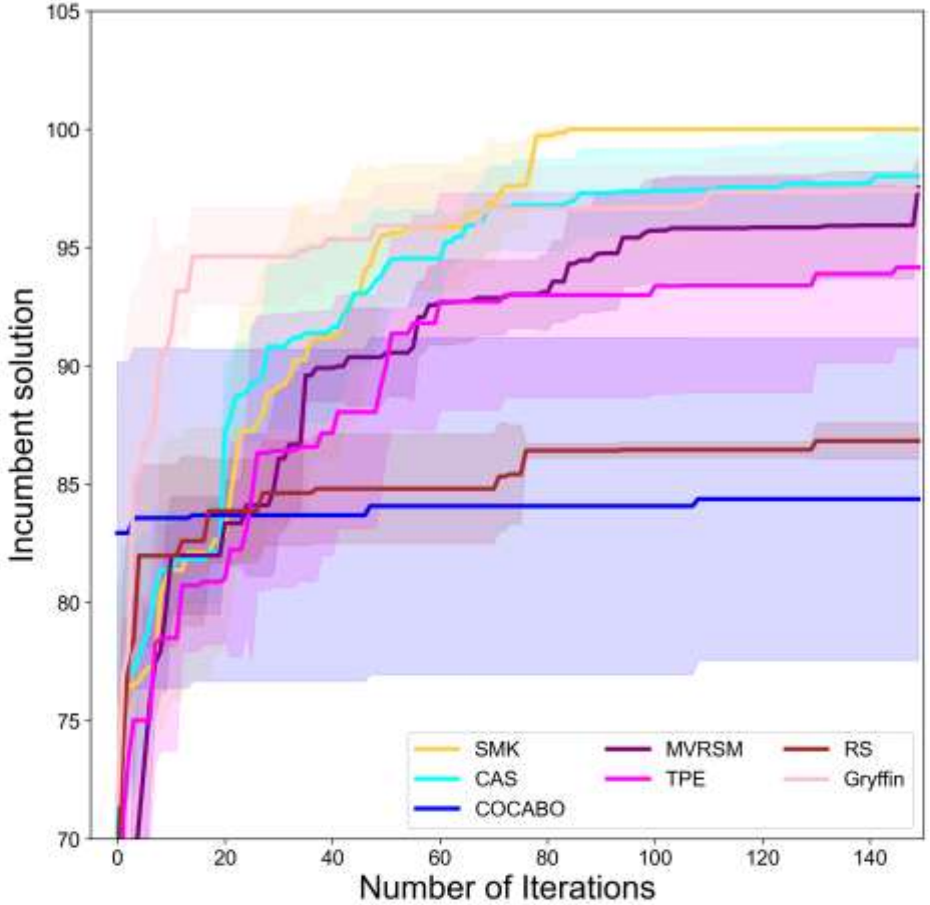
2. Shows two paths leading to high performance.



Numbers indicate points that outperform the previous best performance.

Results

Urea-Selective Catalytic Reduction (SCR)



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)

Results over 10 runs using UCB. Each cell shows mean (± SE). ABO failed on Rosenbrock-20d and Levy-30d.

Conclusion

1. SMKBO exhibits a stronger capability in capturing **complex relationships** in the experimental parameter space.

2. GSM decays with squared distance **(captures smooth regions)**, while CSM decays linearly with distance **(handles outliers better)**.

3. By packing discrete parameters into one variable, BO performance improves as the surrogate handles a single discrete input, **simplifying kernel interactions** and enabling better learning of parameter correlations.

4. SMKBO behaves more like a **human scientist**, featuring an explore-first, exploit-later strategy and performing verifications after switching the convergence criterion.

Publications & Patents

1. Li Z, **Zhao C**, Wang H, et al. Interpreting chemisorption strength with AutoML-based feature deletion experiments[J]. Proceedings of the National Academy of Sciences, 2024, 121(12): e2320232121.

2. Gao Y, Bao M†, **Zhao C**†, et al. Linear scaling relationships between relative diffraction peak intensity and catalytic oxidation of light alkane (under review).

3. Hua C, Zhang Y, **Zhao C**, He Y. Chemical Reaction Optimization Method, System, Medium, and Device Based on Spectral Mixture Kernel. Patent Application No. 2025114285118, 2025; Patent pending.

Thanks！