TagBO: LLM-Driven Task-Aware Graph Bayesian Optimization for Scientific Discovery

Xinzhe Yuan[2,1,*] **Zimu Zeng**[3,7,*] **Zhuo Chen**[4,1,*] **Jianshu Zhang**[4] **Jinzong Dong**[5]
Jingyi Chen[6] Nanyang Ye[4,†] Huan Xiong[2,†] Qinying Gu[1,†]

[1]Shanghai Artificial Intelligence Laboratory
[2]Harbin Institute of Technology
[3]Fudan University
[4]Shanghai Jiao Tong University
[5]Central South University
[6]Huazhong Agricultural University
[7]Shenzhen Loop Area Institute
[*]Equal contribution. [†]Corresponding authors.

Abstract

Mixed-variable optimization in scientific discovery often requires exploring large combinatorial spaces under costly experimental evaluations, making sample-efficient Bayesian optimization (BO) essential. However, existing BO methods typically rely on structural descriptors to define similarity among categorical variables, implicitly assuming that structural similarity reflects similarity in task performance. In scientific optimization, this assumption is often violated, leading to surrogate representations that are poorly aligned with the optimization objective. To address this limitation, we propose **TagBO** (**T**ask-**a**ware **G**raph **B**ayesian **O**ptimization), a method that integrates LLM-derived relational priors with descriptor-based structural information for mixed-variable Bayesian optimization. Rather than directly using LLM outputs as predictive scores, TagBO treats LLM outputs as a potentially noisy source of task-aware relational priors and extracts coarse ordinal relational signals from LLM reasoning to construct task-aware similarity structures for scientific optimization. TagBO then performs constrained Gaussian process Bayesian optimization over the resulting joint discrete-continuous representation space. Experiments on diverse chemistry and materials formulation tasks demonstrate that TagBO consistently improves sample efficiency under fixed evaluation budgets across LLMs of varying capabilities.

1 Introduction

Accelerating scientific discovery through intelligent experimental design has become a central goal in AI for Science [16, 40]. In many scientific optimization problems, practitioners must jointly select categorical variables (e.g., reagent types, reaction pathways) and tune continuous parameters (e.g., concentrations, temperatures), with each evaluation typically requiring costly real-world experiments, resulting in mixed-variable problem in which each experimental evaluation is typically costly and time-consuming [7, 14, 37, 42]. Bayesian optimization (BO) has emerged as a principled strategy for such problems, and a central challenge it must address is how to measure similarity among categorical choices so that the surrogate model can generalize across the combinatorial space [5, 10, 27].

A rich body of work has tackled this challenge using domain-specific descriptors—molecular fingerprints, chemical group encodings, learned embeddings, and other structural representations [17, 29, 39]. These methods have been widely adopted and are effective in many settings. **Yet** they

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

often rely on a task-agnostic notion of proximity, assuming that nearby candidates in descriptor space exhibit similar task behavior—an assumption that can fail in the presence of task-specific mechanisms or activity cliffs [25, 33]. This assumption implicitly defines the geometry over which BO generalizes. When it fails, the surrogate model may extrapolate along directions that are irrelevant, or even misleading, for the target task. For example, two solvents with nearly identical molecular structures may exhibit drastically different electrochemical stability because the mechanisms governing task performance are not captured by global structural descriptors.

These observations suggest that the challenge is not merely to design better descriptors, but to enable task-aware generalization in Bayesian optimization. In mixed-variable BO, the surrogate model generalizes according to assumptions induced by descriptor representations. When these assumptions fail to reflect the functional relationships relevant to the target task, BO may generalize along misleading directions.

Interestingly, scientific domains already contain extensive task-specific relational knowledge in the literature, including experimental findings, empirical heuristics, and mechanistic insights. Large language models (LLMs), pre-trained on broad scientific corpora, can implicitly encode how different materials or components relate under specific task contexts [6, 8, 36]. **However**, directly using raw LLM outputs to define optimization spaces remains challenging due to noise and inconsistency across candidates. At the same time, structural descriptors alone often fail to capture task-specific functional relationships. These observations suggest that neither source alone is sufficient.

To address these limitations, we propose **TagBO** (**T**ask-**a**ware **G**raph **B**ayesian **O**ptimization), a method that integrates LLM-derived relational priors with structural information for mixed-variable Bayesian optimization. Rather than directly using LLM outputs as predictive scores, TagBO extracts ordinal relational signals from LLM reasoning to represent task-aware relationships between categorical choices. These signals are combined with structural information derived from physicochemical descriptors, enabling the optimization space to reflect both task-specific functional behavior and local structural similarity. We evaluate TagBO on diverse scientific optimization benchmarks involving chemistry and materials formulation tasks. Experimental results show that TagBO consistently improves sample efficiency under fixed experimental budgets and remains effective across LLMs of varying capabilities. Our contributions are as follows:

- *Task-aware similarity modeling for scientific optimization.* We propose TagBO, a method that incorporates LLM-derived relational knowledge to construct task-aware similarity structures for mixed-variable Bayesian optimization.
- *Ordinal relational modeling from LLM reasoning.* TagBO treats LLM outputs as a potentially noisy source of task-aware relational priors, and extracts coarse ordinal signals from LLM reasoning for task-aware similarity modeling in scientific optimization problems.
- *Improved sample efficiency under fixed budgets.* Across diverse mixed-variable benchmarks, TagBO consistently improves optimization performance under fixed experimental budgets and remains robust across LLMs of varying capabilities.

2 Related Work

Structured representations in Bayesian optimization. Many approaches construct similarity structures from domain-specific descriptors, learned embeddings, or latent representations [18, 15] to support Bayesian optimization in scientific design problems, such as LADDER, LOLBO [11, 23]. These methods enable the surrogate model to generalize across large combinatorial or molecular search spaces through representation-induced similarity. However, these approaches are primarily designed for optimizing or generating individual molecular structures in continuous latent spaces, and are less suitable for mixed-variable formulation problems involving constrained combinations of heterogeneous categorical and continuous variables.

In contrast, prior work in discrete and combinatorial Bayesian optimization has focused on defining explicit similarity structures to enable generalization across candidate configurations. For example, COMBO [26] represents the search space as a graph Cartesian product and constructs kernels over this structure, allowing the surrogate model to generalize via graph-induced similarity. CoCaBO [30] further addresses mixed continuous and categorical optimization through kernels that share information across categorical variables. Across these approaches, such representations are typically

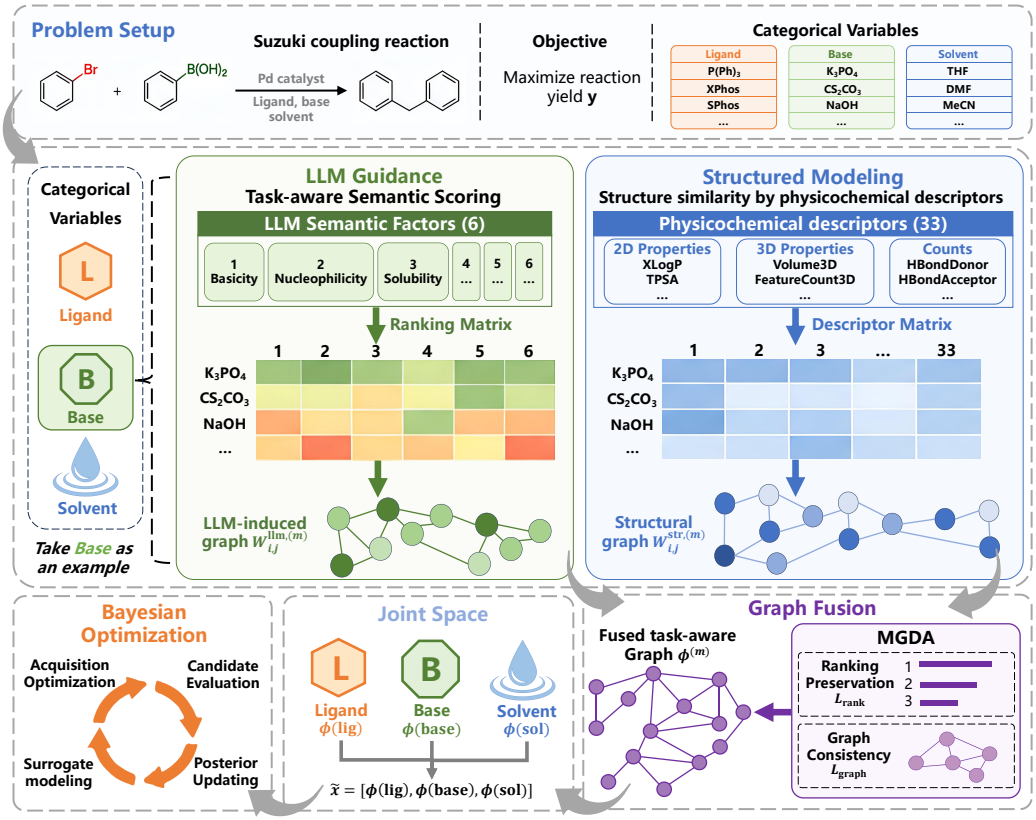


Figure 1: Overview of TagBO on Suzuki coupling reaction optimization. For each categorical variable, TagBO constructs an LLM-derived task-aware relational graph from semantic rankings under the target task, alongside a structural graph derived from physicochemical descriptors. The two signals are fused to form a task-aware optimization space, which is combined with continuous variables and used for Bayesian optimization.

task-agnostic and are not explicitly aligned with the task-specific structure of the objective, which can lead to model misspecification and misleading exploration [31], limiting their ability to perform targeted and diverse exploration.

LLM-enhanced Bayesian optimization. Recent work has explored integrating LLMs into Bayesian optimization to guide candidate generation and improve search efficiency. LLAMBO [22] leverages LLMs to propose promising initial points and generate candidate solutions during optimization, while relying on standard acquisition functions for selection. BOPRO [1] performs optimization in a latent space induced by pretrained LLM embeddings. CAKE [35] employs LLMs to adapt Gaussian process kernels during optimization, enabling more flexible inductive biases for Bayesian optimization. More recently, LGBO [41] further incorporates LLMs to guide the search process through generation or scoring mechanisms. And LABO [9] incorporates LLM-generated predictions into Bayesian optimization through a confidence-gated dual-surrogate framework. However, existing approaches do not explicitly construct interpretable task-aware structures over the search space, which is particularly important in scientific discovery problems where understanding relationships among candidate configurations is crucial.

3 Method

Problem Setup. We consider black-box optimization problems arising in scientific discovery, where the goal is to identify high-performing experimental configurations from a mixed search space.

Formally, we optimize an unknown objective function over

$$\max_{(\mathbf{x},\mathbf{c})\in\mathcal{X}\times\mathbb{R}^N} f(\mathbf{x},\mathbf{c}), \tag{1}$$

where $\mathbf{x} = (x^{(1)}, \ldots, x^{(M)}) \in \mathcal{X} = \prod_{m=1}^{M} \mathcal{X}^{(m)}$ denotes a configuration of M discrete decision variables, and $\mathbf{c} \in \mathbb{R}^N$ denotes continuous experimental variables. Each $\mathcal{X}^{(m)}$ is a finite set of candidates with explicit semantic meaning (e.g., reagents, solvents, molecular structures, or material components), where each candidate corresponds to a physically interpretable component in the experimental system. The continuous variables correspond to controllable conditions such as concentration, temperature, or reaction time.

The main difficulty lies in defining a task-aware similarity structure over the discrete space $\mathcal{X}$, which limits the applicability of standard Bayesian optimization methods. In addition, in scientific optimization, each categorical candidate corresponds to a physically meaningful entity, so the learned representation should preserve a one-to-one correspondence with the original candidate set. To this end, TagBO adopts a graph-based representation, where nodes correspond to valid categorical candidates and edges encode task-aware relational similarities. The key idea is to construct this graph by combining coarse-grained ordinal relational guidance from LLMs with fine-grained local structural information from external descriptors, thereby inducing a task-aware similarity structure over $\mathcal{X}$. Bayesian optimization is then performed over the resulting representation while respecting the feasibility constraints of the original mixed domain.

We organize the method as follows. Sec. 3.1–3.3 construct task-aware representations for different discrete variables. Sec. 3.4 builds the joint representation and integrates it with Bayesian optimization. Sec. 3.5 provides representation-Level interpretation.

3.1 Coarse-grained LLM Guidance

Direct numerical scores produced by LLMs can be noisy or inconsistent across candidates and tasks. To obtain a more stable form of task-aware prior information, we use LLMs to extract coarse-grained ordinal relational guidance instead of precise predictive scores.

For each discrete variable type $\mathcal{X}^{(m)}$, we leverage an LLM to extract a set of task-relevant factors and to characterize candidates along these dimensions. Specifically, given the candidate set $\{x_i^{(m)}\}_{i=1}^{N} \subset \mathcal{X}^{(m)}$, the LLM first identifies F key factors that are likely to influence the target objective based on its prior knowledge and the task context. Subsequently, the LLM ranks the N candidates for each factor f according to their relative strengths, resulting in a ranking matrix $R^{(m)} \in \{1, \ldots, N\}^{N\times F}$, where $R_{i,f}^{(m)}$ denotes the rank of candidate $x_i^{(m)}$ under factor f. We represent each candidate using its ranking vector

$$\mathbf{r}_i^{(m)} = (R_{i,1}^{(m)}, \ldots, R_{i,F}^{(m)}) \in \{1, \ldots, N\}^F,$$

which provides a coarse relational representation of candidate $x_i^{(m)}$. We interpret these ranking vectors as elements in a task-aware space $\mathcal{Z}^{(m)} := \mathbb{R}^F$ and normalize them to obtain continuous embeddings $\mathbf{z}_i^{(m)} \in \mathcal{Z}^{(m)}$. We then define pairwise distances $d_{ij}^{(m)} = \|\mathbf{z}_i^{(m)} - \mathbf{z}_j^{(m)}\|_2$, and construct a similarity matrix using an RBF kernel:

$$W_{ij}^{\mathrm{llm},(m)} = \exp\left(-\gamma \|\mathbf{z}_i^{(m)} - \mathbf{z}_j^{(m)}\|_2^2\right), \tag{2}$$

where $\gamma > 0$ is a bandwidth parameter controlling the decay rate. The resulting matrix $W^{\mathrm{llm},(m)} \in \mathbb{R}^{N\times N}$ defines a weighted graph over categorical candidates, where edge weights encode task-aware relational similarity induced from LLM reasoning. This construction converts coarse-grained ordinal relationships into a task-aware similarity structure, which later serves as relational guidance for graph regularization and Bayesian optimization.

3.2 Fine-grained Structured Modeling

While the LLM-induced graph represents coarse-grained ordinal relational information, ranking-based signals provide only relative task-aware structure and may contain noise from LLM reasoning. They therefore do not fully characterize fine-grained local consistency among candidates. To stabilize

the resulting representation space, we introduce descriptor-based structural smoothness as a complementary inductive bias. Specifically, we incorporate domain-specific descriptors obtained from external databases (e.g., PubChem [20]) for each discrete variable type $\mathcal{X}^{(m)}$. Given the candidate set $\{x_i^{(m)}\}_{i=1}^N \subset \mathcal{X}^{(m)}$, we represent each candidate by a structural attribute vector $\mathbf{a}_i^{(m)} \in \mathbb{R}^K$, where K denotes the number of descriptor features. We measure pairwise similarity in this structural space using an RBF kernel:

$$W_{ij}^{\text{str},(m)} = \exp\left(-\gamma_{\text{str}} \|\mathbf{a}_i^{(m)} - \mathbf{a}_j^{(m)}\|_2^2\right), \tag{3}$$

where $\gamma_{\text{str}} > 0$ is a bandwidth parameter. The resulting matrix $W^{\text{str},(m)} \in \mathbb{R}^{N \times N}$ defines a structural similarity graph over the same candidate set as $W^{\text{llm},(m)}$, and thus shares the same dimensionality. This provides a complementary view of candidate relationships, capturing fine-grained consistency that may not be reflected in the LLM-derived task-aware space.

3.3 Interpretation as Implicit Graph Fusion.

To obtain a stable task-aware representation for Bayesian optimization, we combine coarse-grained ordinal relational guidance from the LLM with fine-grained structural smoothness from descriptor-based similarities. For each discrete variable type $\mathcal{X}^{(m)}$, let $\mathbf{z}_i^{(m)} \in \mathcal{Z}^{(m)}$ denote the task-aware representation of candidate $x_i^{(m)}$ defined in Sec. 3.1. Our goal is to learn a refined representation that preserves the coarse relational ordering induced by the LLM while remaining locally smooth over structurally similar candidates. Specifically, we consider the following two objectives:

(1) Ranking preservation. We define a soft ranking loss that encourages the learned representation to preserve the relative ordering induced by the LLM. For each factor f, the ranking matrix $R^{(m)}$ provides pairwise ordering information. For any pair of candidates (i, j) such that $R_{i,f}^{(m)} < R_{j,f}^{(m)}$ (i.e., $x_i^{(m)}$ is ranked higher than $x_j^{(m)}$ under factor f), we enforce that their learned representations maintain this ordering. The ranking loss is defined as:

$$\mathcal{L}_{\text{rank}}^{(m)}(\mathbf{z}) = \sum_{f=1}^{F} \sum_{i,j} \mathbb{I}(R_{i,f}^{(m)} < R_{j,f}^{(m)}) \cdot \log\left(1 + \exp\left(-(z_{i,f}^{(m)} - z_{j,f}^{(m)})\right)\right), \tag{4}$$

where $\mathbb{I}(\cdot)$ is the indicator function. This loss penalizes violations of pairwise ordering, encouraging the learned representation to remain consistent with the LLM-induced task-aware ranking while allowing flexibility under structural constraints.

(2) Graph consistency. We enforce smoothness over both the task-aware graph $W^{\text{llm},(m)}$ and the structural similarity graph $W^{\text{str},(m)}$:

$$\mathcal{L}_{\text{graph}}^{(m)}(\mathbf{z}) = \sum_{i,j} \left(\sqrt{W_{ij}^{\text{llm},(m)} \, W_{ij}^{\text{str},(m)}} \, \|\mathbf{z}_i^{(m)} - \mathbf{z}_j^{(m)}\|_2\right)^2. \tag{5}$$

The multiplicative form can be interpreted as a geometric-mean style fusion between the task-aware relational graph and the structural similarity graph. As a result, smoothness is primarily enforced between candidate pairs that are simultaneously supported by both the coarse ordinal prior and the local structural similarity. Then, the overall problem is thus formulated as a multi-objective optimization:

$$\min_{\mathbf{z}^{(m)}} \quad \left(\mathcal{L}_{\text{rank}}^{(m)}(\mathbf{z}),\ \mathcal{L}_{\text{graph}}^{(m)}(\mathbf{z})\right). \tag{6}$$

By jointly optimizing these two objectives using the Multi-Gradient Descent Algorithm [12], we obtain a refined task-aware representation that balances coarse ordinal guidance from the LLM with fine-grained structural smoothness from descriptor similarities. Rather than manually specifying a fixed trade-off coefficient, MGDA adaptively balances these two inductive biases during optimization.

3.4 Joint Optimization Space and Constrained BO

We now construct a joint optimization space by combining the refined representations of all variable types. Suppose there are M types of discrete decision variables, where the refined representation of a candidate from the m-th type lies in $\mathcal{Z}^{(m)} = \mathbb{R}^{F_m}$. A joint configuration in the representation space is obtained by concatenating the embeddings of selected candidates:

$$\tilde{\mathbf{x}} = [\mathbf{z}^{(1)}, \mathbf{z}^{(2)}, \ldots, \mathbf{z}^{(M)}] \in \mathbb{R}^{\sum_{m=1}^{M} F_m}, \tag{7}$$

where each $\mathbf{z}^{(m)} \in \mathcal{Z}^{(m)}$ corresponds to the refined representation of a selected candidate $x^{(m)} \in \mathcal{X}^{(m)}$. In addition, continuous decision variables $\mathbf{c} \in \mathbb{R}^N$ are appended to $\tilde{\mathbf{x}}$, yielding a unified representation space:

$$\tilde{\mathcal{X}} = \left\{ [\mathbf{z}^{(1)}, \ldots, \mathbf{z}^{(M)}, \mathbf{c}] \;\middle|\; x^{(m)} \in \mathcal{X}^{(m)},\ \mathbf{z}^{(m)} = \phi^{(m)}(x^{(m)}),\ \mathbf{c} \in \mathcal{C} \right\}, \tag{8}$$

where $\phi^{(m)}$ denotes the representation mapping defined in Sec. 3.1–3.3, and $\mathcal{C}$ denotes the domain of continuous variables. We perform Gaussian process Bayesian optimization in this representation space, with acquisition optimization restricted to the feasible set:

$$\tilde{\mathbf{x}}_{t+1} = \arg\max_{\tilde{\mathbf{x}} \in \tilde{\mathcal{X}}} \alpha_t(\tilde{\mathbf{x}}), \tag{9}$$

ensuring that only valid configurations are considered during the search.

Remark 1 (Search Complexity)**.** *Restricting optimization to $\tilde{\mathbf{x}}$ ensures that only valid configurations are evaluated. As $\tilde{\mathbf{x}}$ coincides with the original discrete candidate set, both the search space cardinality and the number of acquisition evaluations remain unchanged. Thus, the proposed representation introduces no additional computational overhead.*

3.5 Representation-Level Interpretation

We provide several representation-level interpretations to better understand why task-aware representations may improve Bayesian optimization performance in mixed-variable scientific search. Our discussion focuses on three aspects: (i) Gaussian process modeling under representation warping, (ii) the relationship between representation quality and surrogate approximation behavior, and (iii) the adaptive trade-off between task-aware relational guidance and structural smoothness. Due to space limitations, technical assumptions and derivations are deferred to Appendix B.

LLM relational priors as surrogate priors. Our method can be interpreted as transforming ordinal relational knowledge derived from LLM reasoning into a Gaussian process prior over the search space. Rather than directly using LLM outputs as predictive scores, TagBO incorporates relative ranking relations into the learned representation space prior to surrogate modeling. Let $w : \mathcal{X} \to \mathcal{Z}$ denote the learned representation map induced by the fused relational-structural graph, and let $g \sim \mathcal{GP}(m, K)$ be a Gaussian process defined on the representation space $\mathcal{Z}$. Defining $f(x) = g(w(x))$, induces a valid Gaussian process on the original input space with kernel

$$K_w(x, x') = K(w(x), w(x')).$$

Under this interpretation, LLM-derived ordinal relations are not used as direct predictions, but instead shape the similarity structure and inductive bias of the surrogate model itself. This connects task-aware relational guidance with probabilistic surrogate modeling in Bayesian optimization.

LLM ordinal priors and surrogate approximation. Bayesian optimization critically depends on whether the surrogate model can generalize smoothly across the search space. However, structural similarity alone may fail to reflect task-relevant behavior in scientific optimization problems. This motivates using LLM-derived ordinal relations as an additional task-aware source of information. Here, we analyze the representation-level effect of the LLM-induced representation relative to purely structural and one-hot representations; the subsequent graph-smoothing formulation further regularizes this signal with structural similarity. To formalize this intuition, let $\mathcal{D}_W = \mathbb{E}\left[\|Z^\star - \mathbb{E}[Z^\star \mid W]\|^2\right]$ denote the latent reconstruction error associated with a representation W, where $Z^\star$ denotes a latent task-relevant variable. Under the representation-quality assumptions described in Appendix B,

if the LLM-induced ordinal representation preserves task-relevant variation more effectively than task-agnostic structural descriptors, then

$$\mathcal{D}_L \leq \mathcal{D}_S \leq \mathcal{D}_O.$$

Furthermore, if the task link $f = h(Z^\star)$ is L_h-Lipschitz, the corresponding surrogate approximation upper bounds satisfy

$$\mathcal{U}_L \leq \mathcal{U}_S \leq \mathcal{U}_O.$$

This interpretation does not claim that LLM information is always superior; rather, it formalizes the condition under which LLM-derived ordinal priors can provide a more task-aware surrogate prior than purely structural similarity. The role of the structural graph in TagBO is complementary: it regularizes and smooths this LLM-induced relational signal during representation learning.

Adaptive trade-off interpretation. Our multi-objective formulation jointly optimizes ordinal relational consistency and structural graph smoothness in (6). Under standard MGDA [12] optimality conditions (Appendix B), a Pareto-stationary solution $\mathbf{z}^\star$ admits an equivalent adaptive scalarization:

$$\nabla \left(\mathcal{L}_{\text{rank}}(\mathbf{z}^\star) + \lambda^\star \mathcal{L}_{\text{graph}}(\mathbf{z}^\star) \right) = 0,$$

for some adaptive coefficient $\lambda^\star > 0$. The ranking objective preserves coarse-grained ordinal relations induced by LLM reasoning, while the structural objective regularizes these relations through local similarity constraints. This interpretation suggests that the learned representation emerges from balancing coarse-grained ordinal relational guidance with fine-grained local structural smoothness.

4 Experiments

We evaluate the performance of TagBO on multiple optimization datasets from the domains of pharmaceuticals, chemistry, and materials science. These tasks are designed to provide a unified comparison of optimization efficiency and stability under limited evaluation budgets, benchmarking against both traditional Bayesian optimization methods and recent LLM-guided approaches. More experiments for stability and ablation are also provided in Appendix C.

4.1 Experimental Setup

Baselines. The following methods are used as baselines:

1) **One-hot Feature BO:** Discrete variables are encoded using one-hot representations, and the concatenated sparse vectors are fed into a Bayesian optimization model for iterative search.
2) **Property Feature BO:** Physicochemical properties of candidate substances are extracted and concatenated into continuous feature vectors, which are then fed into a Bayesian optimization model for iterative search.
3) **COMBO [26]:** A graph structure is constructed over the discrete combinatorial space, and Gaussian process modeling and optimization are performed based on graph kernels.
4) **BOPRO [1]:** Iteratively updates the LLM prompt using selected past generations to sample new candidates that explore or exploit different regions of the search space.
5) **LGBO [41]:** This method incorporates LLM predictions as low-cost evaluation signals into the BO method alongside real experimental results. A gating mechanism is used to dynamically balance the use of LLM predictions and real evaluations, thereby improving optimization efficiency.

Tasks. We evaluate all methods on six optimization tasks spanning chemistry, materials science, and energy-related applications. Detailed data descriptions are provided in C.1:

1) **Organic Photovoltaics [34] (2200 combinations):** Maximize spectral overlap predicted by DFT by optimizing donor–acceptor building blocks in organic photovoltaic materials.
2) **Suzuki–Miyaura Reaction [28] (384 combinations):** Maximize reaction yield on a fixed-substrate Suzuki–Miyaura reaction subset by optimizing three categorical reaction conditions: ligand, base, and solvent.

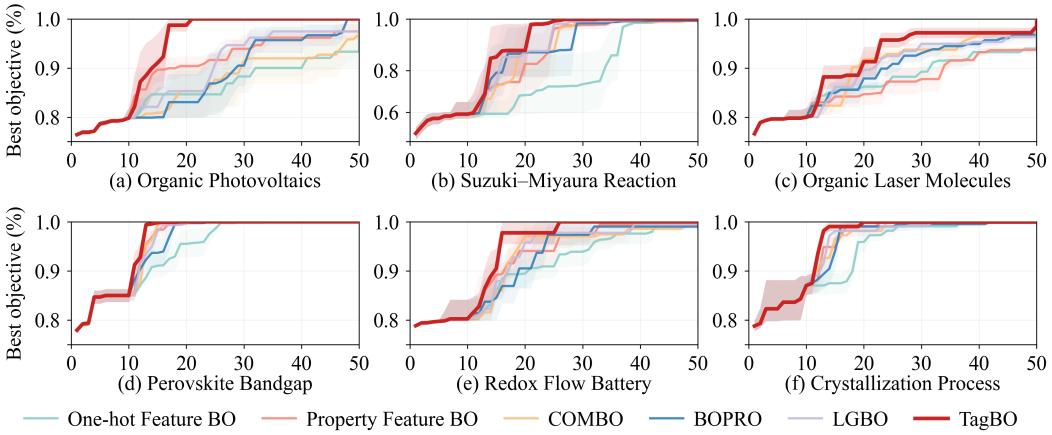


Figure 2: Optimization trajectories on six tasks. The y-axis reports the best objective value found so far, normalized by the oracle optimum, and the x-axis reports the iteration number. TagBO is compared against One-hot Feature BO, Property Feature BO, COMBO, BOPRO, and LGBO.

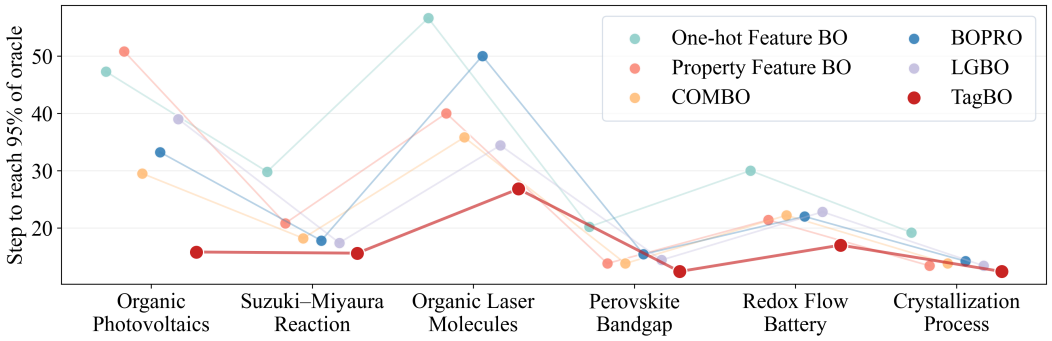


Figure 3: Comparison of the average number of iterations required to reach 95% of the oracle optimum on six benchmark tasks. Lower is better.

3) **Organic Laser Molecules [32] (3458 combinations):** Optimize the photophysical performance of symmetric pentamer molecules by selecting three molecular fragments, with the goal of maximizing peak score and fluorescence rate while minimizing spectral overlap.
4) **Perovskite Bandgap [19] (192 combinations):** Predict and tune the band gap of hybrid organic–inorganic perovskites by selecting the organic, cation, and anion components.
5) **Redox Flow Battery [2] (1408 combinations):** Optimize redoxmer candidates for non-aqueous redox flow batteries by tuning four molecular substituents to minimize absorption wavelength difference, reduction potential, and solvation free energy.
6) **Crystallization Process [21] (mixed-variable):** Optimize the crystallization outcome in a vapor diffusion experiment by adjusting solvent identity, concentrations, vial volumes, reaction time, and temperature to maximize the ordinal crystal score.

Implementation details. All methods employ Gaussian process (GP)-based surrogate models with an RBF kernel, and optimization for each task is initialized with 10 starting points. In the main experiments, the LLM used is Intern S1 241B Thinking, a large language model pretrained on scientific knowledge. All experiments were conducted using 4 NVIDIA H200 GPUs.

4.2 Main Results

As shown in Fig. 2 and Fig. 3, TagBO consistently improves sample efficiency across all six optimization tasks, particularly during the early and middle stages of optimization. Under the same

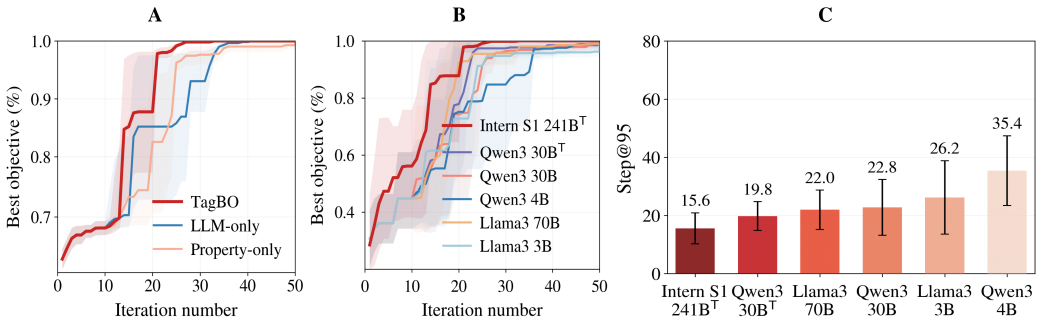


Figure 4: Ablation analysis on the Suzuki dataset. (A) Comparison with two reduced variants using only LLM-derived scoring features or only physicochemical property features. (B) Optimization trajectories by different LLMs, where curves report the best objective value found so far, normalized by the oracle optimum. (C) Mean Step@95% under different LLMs. $*^T$ denote Thinking variants.

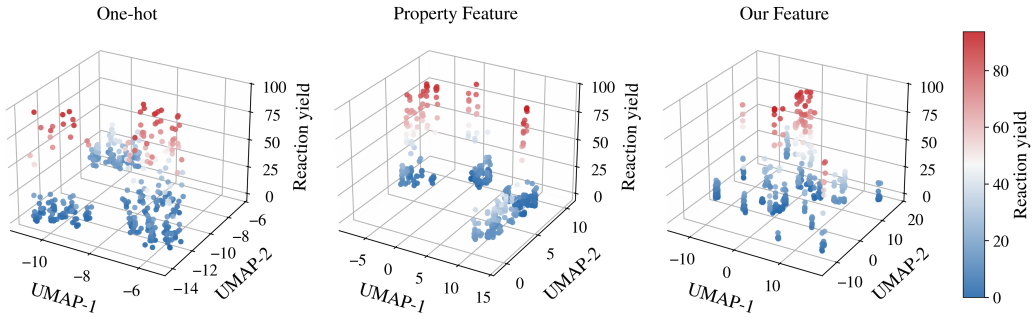


Figure 5: 3D UMAP visualization of feature spaces. Each point represents a reaction candidate; the horizontal axes show the UMAP embedding, and the vertical axis encodes the reaction yield.

evaluation budget, TagBO identifies high-quality candidates with fewer evaluations while maintaining competitive final performance. This advantage is reflected quantitatively by lower Step@95% values across all datasets, indicating faster convergence toward near-optimal solutions. The improvement is especially notable on challenging discrete optimization tasks such as Organic Photovoltaics and Organic Laser Molecules These results suggest that the proposed representation effectively incorporates task-relevant relational priors into the BO search space, improving early-stage search efficiency in expensive scientific optimization settings.

In addition, to understand why the learned representation improves optimization performance, we visualize the reaction space induced by different feature sets using a 3D UMAP embedding [24], as shown in Fig. 5. Compared with one-hot and property-feature baselines, the representation learned by TagBO induces a more connected and smoother high-performing region in the projected search space. High-yield candidates become more locally clustered and exhibit improved neighborhood continuity, suggesting that the learned representation forms a more optimization-friendly landscape for Gaussian-process-based surrogate modeling. Such structure is beneficial for acquisition-guided search, as local interpolation and uncertainty estimation become more aligned with task performance.

4.3 Ablation Experiences

Representation source ablation. To investigate the roles of relational guidance and structural smoothness, we compare the full TagBO model with two reduced variants that use only LLM-derived relational features or only physicochemical property features. As shown in Fig. 4(A), the LLM-only variant achieves strong early-stage optimization performance, indicating that ordinal relational priors provide effective task-aware guidance during early search. However, its performance gradually saturates in later stages, suggesting that relational priors alone are insufficient for fine-grained surrogate refinement. In contrast, the structural-only variant exhibits consistently weaker performance throughout the optimization process. By combining coarse-grained relational guidance

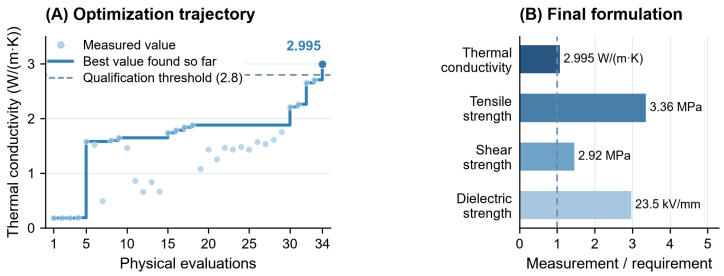


Figure 6: Real-world formulation optimization. **(A)** Thermal conductivity and the best value found so far over 34 physical evaluations. **(B)** Final measurements relative to their qualification thresholds; values above one satisfy the requirements.

with fine-grained structural smoothness, TagBO achieves both strong early-stage exploration and improved later-stage refinement.

Effect of LLM prior quality. We compare LLM backbones with varying scales and reasoning capabilities, including **Intern-S1-241B** [4], **Llama-3.3-70B** [38], **Qwen3-30B** [3](Instruct and Thinking), as well as smaller variants such as **Qwen3-4B** and **Llama-3.2-3B**. As shown in Fig. 4(B) and Fig. 4(C), stronger models generally achieve lower Step@95% values and improved final sample efficiency. In particular, **Intern-S1-241B Thinking** consistently reaches 95% of the oracle optimum using the fewest evaluations, suggesting that domain-relevant scientific pretraining provides more informative relational priors for guiding optimization. Interestingly, optimization trajectories from different LLM backbones remain relatively close during early search, suggesting that the proposed ordinal relational formulation is less sensitive to precise numerical calibration from a specific LLM. Stronger models mainly improve later-stage refinement by providing more informative task-aware relational structure.

4.4 Real-world Formulation Experiment

Task setting. This task aims to maximize the measured **thermal conductivity** of a thermally conductive adhesive while satisfying mechanical and electrical performance requirements. The mixed-variable search space involves selecting 10 fillers from a library of 29 candidate materials and optimizing their 10 corresponding continuous composition variables, as summarized in Table 7. The thermal conductivity qualification threshold is $2.8\,\mathrm{W}/(\mathrm{m}\cdot\mathrm{K})$. The formulation must also satisfy tensile strength $\geq 1\,\mathrm{MPa}$, shear strength $\geq 2\,\mathrm{MPa}$, and dielectric strength $\geq 7.9\,\mathrm{kV}/\mathrm{mm}$. The experiment includes 34 physical evaluations, each requiring approximately three days.

Results. As shown in Fig. 6, the best measured thermal conductivity increases from $0.1823\,\mathrm{W}/(\mathrm{m}\cdot\mathrm{K})$ at the first evaluation to $2.995\,\mathrm{W}/(\mathrm{m}\cdot\mathrm{K})$ after only 34 evaluations, exceeding the qualification threshold. The final formulation achieves a tensile strength of $3.36\,\mathrm{MPa}$, a shear strength of $2.92\,\mathrm{MPa}$, and a dielectric strength of $23.5\,\mathrm{kV}/\mathrm{mm}$, satisfying all three performance requirements. Under the $n_{\mathrm{init}} = 2d$ setting used in standard TuRBO [13] implementations, a 20-variable problem would require 40 evaluations for initialization alone. In this large combinatorial and continuous search space, TagBO thus meets all four engineering requirements within a total experimental budget smaller than this initial design, demonstrating its sample efficiency for expensive formulation optimization.

5 Conclusion

In this work, we proposed TagBO, a task-aware graph Bayesian optimization framework for mixed-variable scientific optimization. Unlike conventional approaches that rely primarily on task-agnostic structural similarity, TagBO incorporates coarse-grained ordinal relational priors extracted from LLM reasoning and combines them with fine-grained structural information through adaptive multi-objective optimization.

References

[1] D. Agarwal, M. G. Arivazhagan, R. Das, S. Swamy, S. Khosla, and R. Gangadharaiah. Searching for optimal solutions with llms via bayesian optimization. In *The Thirteenth International Conference on Learning Representations*, 2025.

[2] Garvit Agarwal, Hieu A Doan, Lily A Robertson, Lu Zhang, and Rajeev S Assary. Discovery of energy storage molecular materials using quantum chemistry-guided multiobjective bayesian optimization. *Chemistry of Materials*, 33(20):8133–8144, 2021.

[3] Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023.

[4] Lei Bai, Zhongrui Cai, Yuhang Cao, Maosong Cao, Weihan Cao, Chiyu Chen, Haojiong Chen, Kai Chen, Pengcheng Chen, Ying Chen, et al. Intern-s1: A scientific multimodal foundation model. *arXiv preprint arXiv:2508.15763*, 2025.

[5] Ricardo Baptista and Matthias Poloczek. Bayesian optimization of combinatorial structures. In *International conference on machine learning*, pages 462–471. PMLR, 2018.

[6] Andres M Bran, Sam Cox, Oliver Schilter, Carlo Baldassari, Andrew D White, and Philippe Schwaller. Chemcrow: Augmenting large-language models with chemistry tools. *arXiv preprint arXiv:2304.05376*, 2023.

[7] Keith T Butler, Daniel W Davies, Hugh Cartwright, Olexandr Isayev, and Aron Walsh. Machine learning for molecular and materials science. *Nature*, 559(7715):547–555, 2018.

[8] Cayque Monteiro Castro Nascimento and André Silva Pimentel. Do large language models understand chemistry? a conversation with chatgpt. *Journal of Chemical Information and Modeling*, 63(6):1649–1655, 2023.

[9] Zhuo Chen, Xinzhe Yuan, Jianshu Zhang, Jinzong Dong, Ruichen Zhou, Yingchun Niu, Tianhang Zhou, Yu Yang Fredrik Liu, Yuqiang Li, Nanyang Ye, et al. Labo: Llm-accelerated bayesian optimization through broad exploration and selective experimentation. *arXiv preprint arXiv:2605.22054*, 2026.

[10] Aryan Deshwal, Syrine Belakaria, and Janardhan Rao Doppa. Bayesian optimization over hybrid spaces. In *International conference on machine learning*, pages 2632–2643. PMLR, 2021.

[11] Aryan Deshwal and Jana Doppa. Combining latent space and structured kernels for bayesian optimization over combinatorial spaces. *Advances in neural information processing systems*, 34:8185–8200, 2021.

[12] Jean-Antoine Désidéri. Multiple-gradient descent algorithm (mgda). 2009.

[13] David Eriksson, Michael Pearce, Jacob Gardner, Ryan D Turner, and Matthias Poloczek. Scalable global optimization via local bayesian optimization. *Advances in neural information processing systems*, 32, 2019.

[14] Peter I Frazier. A tutorial on bayesian optimization. *arXiv preprint arXiv:1807.02811*, 2018.

[15] Peter I. Frazier and Jialei Wang. Bayesian optimization for materials design. In *Information Science for Materials Discovery and Design*, pages 45–75. Springer International Publishing, Cham, 2015.

[16] Alireza Ghafarollahi and Markus J Buehler. Sciagents: Automating scientific discovery through multi-agent intelligent graph reasoning, 2024. *URL https://arxiv. org/abs/2409.05556*, 2409.

[17] Justin Gilmer, Samuel S Schoenholz, Patrick F Riley, Oriol Vinyals, and George E Dahl. Neural message passing for quantum chemistry. In *International conference on machine learning*, pages 1263–1272. Pmlr, 2017.

[18] Rafael Gómez-Bombarelli, Jennifer N Wei, David Duvenaud, José Miguel Hernández-Lobato, Benjamín Sánchez-Lengeling, Dennis Sheberla, Jorge Aguilera-Iparraguirre, Timothy D Hirzel, Ryan P Adams, and Alán Aspuru-Guzik. Automatic chemical design using a data-driven continuous representation of molecules. *ACS central science*, 4(2):268–276, 2018.

[19] C. Kim, T. D. Huan, S. Krishnan, and R. Ramprasad. A hybrid organic-inorganic perovskite dataset. *Scientific data*, 4(1):1–11, 2017.

[20] Sunghwan Kim, Paul A Thiessen, Evan E Bolton, Jie Chen, Gang Fu, Asta Gindulyte, Lianyi Han, Jane He, Siqian He, Benjamin A Shoemaker, et al. Pubchem substance and compound databases. *Nucleic acids research*, 44(D1):D1202–D1213, 2016.

[21] Zhi et al. Li. Robot-accelerated perovskite investigation and discovery. *Chemistry of Materials*, 32(13):5650–5663, 2020.

[22] T. Liu, N. Astorga, N. Seedat, and M. van der Schaar. Large language models to enhance bayesian optimization. In *The Twelfth International Conference on Learning Representations*, 2024.

[23] Natalie Maus, Haydn Jones, Juston Moore, Matt J Kusner, John Bradshaw, and Jacob Gardner. Local latent space bayesian optimization over structured inputs. *Advances in neural information processing systems*, 35:34505–34518, 2022.

[24] Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018.

[25] Jose L Medina-Franco. Scanning structure–activity relationships with structure–activity similarity and related maps: from consensus activity cliffs to selectivity switches. *Journal of chemical information and modeling*, 52(10):2485–2493, 2012.

[26] Changyong Oh, Efstratios Gavves, and Max Welling. Combinatorial bayesian optimization using the graph cartesian product. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

[27] Changyong Oh, Jakub Tomczak, Efstratios Gavves, and Max Welling. Combinatorial bayesian optimization using the graph cartesian product. *Advances in Neural Information Processing Systems*, 32, 2019.

[28] B. J. Reizman, Y. M. Wang, S. L. Buchwald, and K. F. Jensen. Suzuki–miyaura cross-coupling optimization enabled by automated feedback. *Reaction chemistry & engineering*, 1(6):658–666, 2016.

[29] David Rogers and Mathew Hahn. Extended-connectivity fingerprints. *Journal of chemical information and modeling*, 50(5):742–754, 2010.

[30] Binxin Ru, Ahsan Alvi, Vu Nguyen, Michael A Osborne, and Stephen Roberts. Bayesian optimisation over multiple continuous and categorical inputs. In *International Conference on Machine Learning*, pages 8276–8285. PMLR, 2020.

[31] Peter Sollich. Can gaussian process regression be made robust against model mismatch? In *International Workshop on Deterministic and Statistical Methods in Machine Learning*, pages 199–210, Berlin, Heidelberg, 2004. Springer.

[32] Felix et al. Strieth-Kalthoff. Delocalized, asynchronous, closed-loop discovery of organic laser emitters. *Science*, 384(6697):eadk9227, 2024.

[33] Dagmar Stumpfe, Huabin Hu, and Jurgen Bajorath. Evolving concept of activity cliffs. *ACS omega*, 4(11):14360–14368, 2019.

[34] Wenbo et al. Sun. Machine learning–assisted molecular design and efficiency prediction for high-performance organic photovoltaic materials. *Science advances*, 5(11):eaay4275, 2019.

[35] Richard Cornelius Suwandi, Feng Yin, Juntao Wang, Renjie Li, Tsung-Hui Chang, and Sergios Theodoridis. Adaptive kernel design for bayesian optimization is a piece of cake with llms. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

[36] Ross Taylor, Marcin Kardas, Guillem Cucurull, Thomas Scialom, Anthony Hartshorn, Elvis Saravia, Andrew Poulton, Viktor Kerkez, and Robert Stojnic. Galactica: A large language model for science. *arXiv preprint arXiv:2211.09085*, 2022.

[37] Gary Tom, Stefan P Schmid, Sterling G Baird, Yang Cao, Kourosh Darvish, Han Hao, Stanley Lo, Sergio Pablo-García, Ella M Rajaonson, Marta Skreta, et al. Self-driving laboratories for chemistry and materials science. *Chemical Reviews*, 124(16):9633–9732, 2024.

[38] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

[39] Ziyu Wang, Frank Hutter, Masrour Zoghi, David Matheson, and Nando De Feitas. Bayesian optimization in a billion dimensions via random embeddings. *Journal of Artificial Intelligence Research*, 55:361–387, 2016.

[40] Hongliang Xin, John R Kitchin, and Heather J Kulik. Towards agentic science for advancing scientific discovery. *Nature Machine Intelligence*, 7(9):1373–1375, 2025.

[41] Xinzhe Yuan, Zhuo Chen, Jianshu Zhang, Huan Xiong, Nanyang Ye, Yuqiang Li, and Qinying Gu. Unleashing llms in bayesian optimization: Preference-guided framework for scientific discovery. In *International Conference on Learning Representations (ICLR)*, 2026. Poster.

[42] Yichi Zhang, Daniel W Apley, and Wei Chen. Bayesian optimization for materials design with mixed quantitative and qualitative variables. *Scientific reports*, 10(1):4924, 2020.

Acknowledgements

This work was supported by the New Generation Artificial Intelligence National Science and Technology Major Project (No. 2025ZD0121802). It was also supported by the New Generation Artificial Intelligence National Science and Technology Major Project (No. 2025ZD0122901) and the National Natural Science Foundation of China (No. 62572313).

Appendix

A Notation and Variables

Symbol	Description
Spaces and mappings	
$\mathcal{X}$	Input space
$\mathcal{Z}$	Latent / transformed space
$w : \mathcal{X} \to \mathcal{Z}$	Warping (representation mapping)
$x \in \mathcal{X}$	Input variable
$z \in \mathcal{Z}$	Transformed input
Gaussian process construction	
$g \sim \mathcal{GP}(m, K)$	Gaussian process defined on $\mathcal{Z}$
$m(\cdot)$	Mean function
$K(\cdot, \cdot)$	Covariance kernel
$f(x) = g(w(x))$	Induced process on $\mathcal{X}$
$K_w(x, x')$	Induced kernel $K(w(x), w(x'))$
Latent-variable formulation	
$Z^\star \in \mathbb{R}^d$	Task-relevant latent variable
$h : \mathbb{R}^d \to \mathbb{R}$	Task link function ($f = h(Z^\star)$)
L_h	Lipschitz constant of h
X	Random input variable
f	Target output variable
Representations	
$L = w_{\mathrm{llm}}(X)$	LLM-based representation
$S = w_{\mathrm{str}}(X)$	Structural representation
$O = w_{\mathrm{oh}}(X)$	One-hot representation
Error measures	
$\mathcal{D}_W$	Latent reconstruction error
$\mathcal{R}_W$	Bayes risk
$\mathcal{U}_W$	Surrogate upper bound ($L_h^2 \mathcal{D}_W$)
Assumption parameters and mappings	
σ_L^2	Upper bound on LLM decoding error
ρ_S^2	Irreducible uncertainty under structural representation
ψ_L	Decoder from L to $Z^\star$
T	Mapping from S to O

Table A.1: Summary of notation used throughout the appendix.

B Proof

B.1 Warping induces a Gaussian process prior

Proposition 1 (Warping induces a Gaussian process prior)**.** *Let $\mathcal{X}$ be the input space and $\mathcal{Z}$ be a measurable space. Let $w : \mathcal{X} \to \mathcal{Z}$ be a measurable mapping. Suppose*

$$g \sim \mathcal{GP}(m, K)$$

is a Gaussian process defined on $\mathcal{Z}$, with mean function $m : \mathcal{Z} \to \mathbb{R}$ and covariance kernel $K : \mathcal{Z} \times \mathcal{Z} \to \mathbb{R}$.

Define a stochastic process $f : \mathcal{X} \to \mathbb{R}$ by

$$f(x) := g(w(x)).$$

Then f is a Gaussian process on $\mathcal{X}$, and

$$f \sim \mathcal{GP}(m \circ w,\ K_w),$$

where

$$(m \circ w)(x) = m(w(x)), \qquad K_w(x, x') = K\big(w(x), w(x')\big).$$

In particular, K_w is a valid positive definite kernel on $\mathcal{X}$, and hence a fixed warping w induces a Gaussian process prior over functions on $\mathcal{X}$.

Proof. To show that f is a Gaussian process, it suffices to verify that for any finite set of inputs $x_1, \ldots, x_n \in \mathcal{X}$, the random vector

$$(f(x_1), \ldots, f(x_n))$$

is jointly Gaussian.

Let $z_i := w(x_i) \in \mathcal{Z}$ for $i = 1, \ldots, n$. Then by definition,

$$(f(x_1), \ldots, f(x_n)) = (g(z_1), \ldots, g(z_n)).$$

Since $g \sim \mathcal{GP}(m, K)$, the vector $(g(z_1), \ldots, g(z_n))$ is multivariate Gaussian with mean

$$\big(m(z_1), \ldots, m(z_n)\big) = \big(m(w(x_1)), \ldots, m(w(x_n))\big),$$

and covariance matrix

$$\big(K(z_i, z_j)\big)_{i,j=1}^{n} = \big(K(w(x_i), w(x_j))\big)_{i,j=1}^{n}.$$

Therefore $(f(x_1), \ldots, f(x_n))$ is Gaussian for any finite collection of inputs, and hence f is a Gaussian process. Its mean and covariance functions follow directly:

$$\mathbb{E}[f(x)] = m(w(x)), \qquad \mathrm{Cov}(f(x), f(x')) = K(w(x), w(x')).$$

This proves the result. □

Remark 2 (Positive definiteness and modeling space)**.** *The construction above preserves positive definiteness of the kernel. Specifically, if K is a positive definite kernel on $\mathcal{Z}$, then the induced kernel*

$$K_w(x, x') := K\big(w(x), w(x')\big)$$

is also positive definite on $\mathcal{X}$. This follows directly from the definition of positive definiteness: for any $x_1, \ldots, x_n \in \mathcal{X}$ and $c_1, \ldots, c_n \in \mathbb{R}$,

$$\sum_{i,j=1}^{n} c_i c_j K_w(x_i, x_j) = \sum_{i,j=1}^{n} c_i c_j K\big(w(x_i), w(x_j)\big) \geq 0.$$

Consequently, the Gaussian process prior induced by w is well-defined.

From a modeling perspective, this construction can be interpreted as performing Bayesian inference in the transformed space $\mathcal{Z}$, where a standard kernel K is applied, while the warping w maps the original inputs into this space.

B.2 Ordered Upper Bounds via Task-Relevant Signal Dominance

B.2.1 Setup

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. Let $X \in \mathcal{X}$ denote the input, and suppose there exists a latent task-relevant variable

$$Z^\star \in \mathbb{R}^d$$

such that

$$f = h(Z^\star),$$

where $h : \mathbb{R}^d \to \mathbb{R}$ is measurable.

We consider three representations:

$$L := w_{\mathrm{llm}}(X), \qquad S := w_{\mathrm{str}}(X), \qquad O := w_{\mathrm{oh}}(X).$$

B.2.2 Definitions

Definition 1 (Latent reconstruction error)**.** *For any representation W, define*

$$\mathcal{D}_W := \mathbb{E}\left[\|Z^\star - \mathbb{E}[Z^\star \mid W]\|^2\right].$$

Definition 2 (Bayes risk)**.** *For any representation W, define*

$$\mathcal{R}_W := \mathbb{E}\left[(f - \mathbb{E}[f \mid W])^2\right].$$

B.2.3 Assumptions

Assumption 1 (Lipschitz task link)**.** *The function h is L_h-Lipschitz, i.e.,*

$$|h(z_1) - h(z_2)| \le L_h \|z_1 - z_2\|, \qquad \forall z_1, z_2 \in \mathbb{R}^d.$$

Our ordering result is based on an explicit modeling hypothesis about the relative task-relevance of the three representations. Recall that $Z^\star$ denotes a latent task-relevant variable such that $f = h(Z^\star)$. We interpret the LLM representation L as encoding task-aware information aligned with the downstream task, the structural representation S as encoding task-agnostic structural regularities, and the one-hot representation O as encoding only input identity. The following assumptions formalize this interpretation.

Assumption 2 (LLM representation preserves task-relevant information)**.** *Because the LLM-derived representation is assumed to encode task-relevant information, we assume that $Z^\star$ can be decoded from L with bounded mean-squared error. That is, there exists a measurable function ψ_L such that*

$$\mathbb{E}\left[\|Z^\star - \psi_L(L)\|^2\right] \le \sigma_L^2.$$

Here, σ_L^2 measures the residual task-irrelevant noise in the LLM representation.

Assumption 3 (Structural representation is task-agnostic)**.** *The structural representation S is assumed to capture generic relational or topological patterns of the input domain, rather than task-specific predictive content. Accordingly, even after observing S, a nontrivial amount of uncertainty about the task-relevant latent variable $Z^\star$ remains. We formalize this by assuming that there exists $\rho_S^2 > 0$ such that*

$$\mathcal{D}_S = \mathbb{E}\left[\|Z^\star - \mathbb{E}[Z^\star \mid S]\|^2\right] \ge \rho_S^2.$$

Thus, ρ_S^2 represents an irreducible latent reconstruction error associated with the task-agnostic structural representation.

Assumption 4 (Structural representation contains one-hot information)**.** *The one-hot representation O only records input identity, whereas the structural representation S is assumed to retain at least this identity information together with additional structural information. Therefore, we assume that there exists a measurable function T such that*

$$O = T(S).$$

Equivalently, the information carried by O is contained in that carried by S.

Assumption 5 (Comparative representation assumption)**.** *We assume that the task-relevant information contained in the LLM representation is stronger than that available from the task-agnostic structural representation. Quantitatively, this means that the LLM decoding error level is no larger than the irreducible latent uncertainty under S:*

$$\sigma_L^2 \leq \rho_S^2.$$

This assumption expresses the intended inductive bias that task-aware features are more informative about $Z^\star$ than purely structural features.

B.2.4 Auxiliary Results

Lemma 1 (Bayes risk as conditional variance)**.** *For any representation W,*

$$\mathcal{R}_W = \mathbb{E}\left[\mathrm{Var}(f \mid W)\right].$$

Proof. Since $\mathbb{E}[f \mid W]$ is the L^2 projection of f onto the σ-field generated by W, we have

$$f = \mathbb{E}[f \mid W] + \big(f - \mathbb{E}[f \mid W]\big),$$

where the two terms are orthogonal in L^2. Taking expectations of squares yields

$$\mathcal{R}_W = \mathbb{E}\left[(f - \mathbb{E}[f \mid W])^2\right] = \mathbb{E}\left[\mathrm{Var}(f \mid W)\right].$$

□

Lemma 2 (Transfer from latent reconstruction to target prediction)**.** *Under Assumption 1, for any representation W,*

$$\mathcal{R}_W \leq L_h^2 \mathcal{D}_W.$$

Proof. Let $\delta_W := \mathbb{E}[f \mid W]$. By optimality of conditional expectation,

$$\mathcal{R}_W = \mathbb{E}\left[(f - \delta_W)^2\right] \leq \mathbb{E}\left[(f - a(W))^2\right]$$

for any W-measurable function $a(W)$.

Choose

$$a(W) := h\big(\mathbb{E}[Z^\star \mid W]\big).$$

Then

$$\mathcal{R}_W \leq \mathbb{E}\left[\Big(h(Z^\star) - h(\mathbb{E}[Z^\star \mid W])\Big)^2\right].$$

By the Lipschitz property of h,

$$\Big(h(Z^\star) - h(\mathbb{E}[Z^\star \mid W])\Big)^2 \leq L_h^2 \left\|Z^\star - \mathbb{E}[Z^\star \mid W]\right\|^2.$$

Taking expectations yields the result. □

Lemma 3 (LLM upper bound)**.** *Under Assumption 2,*

$$\mathcal{D}_L \leq \sigma_L^2.$$

Proof. By definition,

$$\mathcal{D}_L = \inf_{\phi} \mathbb{E}\left[\|Z^\star - \phi(L)\|^2\right].$$

Taking $\phi = \psi_L$ yields

$$\mathcal{D}_L \leq \mathbb{E}\left[\|Z^\star - \psi_L(L)\|^2\right] \leq \sigma_L^2.$$

□

Lemma 4 (Structural representation dominates one-hot)**.** *Under Assumption 4,*

$$\mathcal{D}_S \leq \mathcal{D}_O.$$

Proof. Since $O = T(S)$, any O-measurable function is also S-measurable. In particular, $\mathbb{E}[Z^\star \mid O]$ is S-measurable.

By optimality of conditional expectation,

$$\mathbb{E}\left[\|Z^\star - \mathbb{E}[Z^\star \mid S]\|^2\right] \leq \mathbb{E}\left[\|Z^\star - a(S)\|^2\right]$$

for any S-measurable function $a(S)$. Choosing $a(S) = \mathbb{E}[Z^\star \mid O]$ yields

$$\mathcal{D}_S \leq \mathcal{D}_O.$$

□

B.2.5 Main Result

Proposition 2 (Ordered surrogate upper bounds)**.** *Let*

$$\mathcal{D}_W = \mathbb{E}\left[\|Z^\star - \mathbb{E}[Z^\star \mid W]\|^2\right]$$

denote the latent reconstruction error of a representation W, i.e., the minimal mean-squared error for recovering the task-relevant latent variable $Z^\star$ from W.

Let

$$\mathcal{U}_W := L_h^2 \mathcal{D}_W$$

denote the latent-reconstruction-induced surrogate upper bound on the Bayes risk associated with representation W.

Let L, S, and O denote the LLM representation, the structural representation, and the one-hot representation, respectively.

Under Assumptions 2–5, the latent reconstruction errors satisfy

$$\mathcal{D}_L \le \mathcal{D}_S \le \mathcal{D}_O.$$

Equivalently, the induced surrogate upper bounds satisfy

$$\mathcal{U}_L \le \mathcal{U}_S \le \mathcal{U}_O.$$

In particular, the LLM representation induces the smallest surrogate upper bound, the structural representation induces a larger one, and the one-hot representation induces the largest one.

Moreover, under Assumption 1, each Bayes risk is controlled by its corresponding surrogate upper bound:

$$\mathcal{R}_W \le \mathcal{U}_W, \qquad W \in \{L, S, O\}.$$

Proof. We first prove the ordering of latent reconstruction errors.

Since the LLM representation L is assumed to admit a decoder for the task-relevant latent variable $Z^\star$, Lemma 3 gives

$$\mathcal{D}_L \le \sigma_L^2.$$

Here, σ_L^2 is the assumed upper bound on the latent reconstruction error achievable from the LLM representation.

Since the structural representation S is assumed to be task-agnostic, Assumption 3 gives

$$\mathcal{D}_S \ge \rho_S^2.$$

Here, ρ_S^2 is the irreducible latent uncertainty remaining after observing the structural representation.

By the task-relevant signal dominance assumption, Assumption 5, we have

$$\sigma_L^2 \le \rho_S^2.$$

Combining these inequalities yields

$$\mathcal{D}_L \le \sigma_L^2 \le \rho_S^2 \le \mathcal{D}_S.$$

Therefore,

$$\mathcal{D}_L \le \mathcal{D}_S.$$

It remains to compare the structural representation S with the one-hot representation O. By Assumption 4, there exists a measurable function T such that

$$O = T(S).$$

Hence every O-measurable estimator is also S-measurable. In particular,

$$\mathbb{E}[Z^\star \mid O]$$

is an admissible estimator of $Z^\star$ based on S.

Since

$$\mathbb{E}[Z^\star \mid S]$$

is the optimal S-measurable estimator under squared loss, we have

$$\mathbb{E}\left[\|Z^\star - \mathbb{E}[Z^\star \mid S]\|^2\right] \le \mathbb{E}\left[\|Z^\star - \mathbb{E}[Z^\star \mid O]\|^2\right].$$

By the definition of latent reconstruction error, this is exactly

$$\mathcal{D}_S \le \mathcal{D}_O.$$

Combining the two comparisons gives

$$\mathcal{D}_L \le \mathcal{D}_S \le \mathcal{D}_O.$$

Now recall that

$$\mathcal{U}_W := L_h^2 \mathcal{D}_W.$$

Since $L_h^2 \ge 0$, multiplying the latent reconstruction ordering by L_h^2 gives

$$\mathcal{U}_L \le \mathcal{U}_S \le \mathcal{U}_O.$$

Finally, by Lemma 2, under the Lipschitz task-link assumption,

$$\mathcal{R}_W \le L_h^2 \mathcal{D}_W = \mathcal{U}_W, \qquad W \in \{L, S, O\}.$$

This completes the proof. □

Remark 3 (Implications for Bayesian optimization)**.** *The ordering of surrogate upper bounds provides a representation-level explanation for the effectiveness of Bayesian optimization.*

In Bayesian optimization, the performance of the optimization procedure critically depends on the ability of the surrogate model (e.g., a Gaussian process) to accurately model the underlying objective function and provide well-calibrated uncertainty estimates. As emphasized by Snoek et al. (2014), failures in modeling, such as violations of stationarity, can significantly degrade optimization performance, while improving the fidelity of the surrogate model can lead to faster convergence and more reliable results.

The present theorem shows that the choice of representation W directly controls an upper bound on the prediction error via $\mathcal{U}_W = L_h^2 \mathcal{D}_W$. Therefore, representations that better preserve the task-relevant latent variable $Z^\star$ induce tighter surrogate upper bounds, which in turn correspond to more accurate function models.

B.3 Adaptive coarse-to-fine trade-off

Proposition 3 (Adaptive scalarization induced by MGDA). Let

$$\mathcal{L}_1(\mathbf{z}) = \mathcal{L}_{\text{rank}}(\mathbf{z}), \qquad \mathcal{L}_2(\mathbf{z}) = \mathcal{L}_{\text{graph}}(\mathbf{z}),$$

where both objectives are differentiable with respect to $\mathbf{z}$. Suppose that $\mathbf{z}^\star$ is a Pareto-stationary point obtained by MGDA for the bi-objective problem

$$\min_{\mathbf{z}} \left(\mathcal{L}_1(\mathbf{z}), \mathcal{L}_2(\mathbf{z})\right).$$

Then there exists $\alpha^\star \in [0, 1]$ such that

$$\alpha^\star \nabla \mathcal{L}_1(\mathbf{z}^\star) + (1 - \alpha^\star) \nabla \mathcal{L}_2(\mathbf{z}^\star) = 0. \tag{10}$$

Moreover, if $\alpha^\star > 0$, then there exists an adaptive coefficient

$$\lambda^\star = \frac{1 - \alpha^\star}{\alpha^\star}$$

such that

$$\nabla \left(\mathcal{L}_{\text{rank}}(\mathbf{z}^\star) + \lambda^\star \mathcal{L}_{\text{graph}}(\mathbf{z}^\star)\right) = 0. \tag{11}$$

Therefore, $\mathbf{z}^\star$ is a first-order stationary point of a scalarized objective whose trade-off coefficient is determined adaptively by MGDA rather than manually specified.

Proof. MGDA finds a common descent direction by solving

$$\min_{\alpha\in[0,1]} \|\alpha\nabla\mathcal{L}_1(\mathbf{z}) + (1-\alpha)\nabla\mathcal{L}_2(\mathbf{z})\|_2^2 . \tag{12}$$

At a Pareto-stationary point $\mathbf{z}^\star$, there exists no direction that can simultaneously decrease both objectives. This condition is equivalent to the origin belonging to the convex hull of the gradients of the objectives. Hence, for the two-objective case, there exists $\alpha^\star \in [0,1]$ such that

$$\alpha^\star\nabla\mathcal{L}_1(\mathbf{z}^\star) + (1-\alpha^\star)\nabla\mathcal{L}_2(\mathbf{z}^\star) = 0. \tag{13}$$

If $\alpha^\star > 0$, dividing both sides by $\alpha^\star$ yields

$$\nabla\mathcal{L}_1(\mathbf{z}^\star) + \frac{1-\alpha^\star}{\alpha^\star}\nabla\mathcal{L}_2(\mathbf{z}^\star) = 0. \tag{14}$$

By setting

$$\lambda^\star = \frac{1-\alpha^\star}{\alpha^\star},$$

we obtain

$$\nabla\left(\mathcal{L}_{\mathrm{rank}}(\mathbf{z}^\star) + \lambda^\star\mathcal{L}_{\mathrm{graph}}(\mathbf{z}^\star)\right) = 0. \tag{15}$$

This proves the claim.

Remark 4 (Adaptive coarse-to-fine trade-off.)**.** *The above result provides an interpretation of our multi-objective formulation as an adaptive trade-off between coarse-grained task-aware information and fine-grained structural consistency. The ranking objective $\mathcal{L}_{\mathrm{rank}}$ preserves the coarse, task-aware ordinal structure induced by the LLM, while the graph objective $\mathcal{L}_{\mathrm{graph}}$ enforces fine-grained smoothness over the structural similarity graph. MGDA adaptively determines the effective coefficient $\lambda^\star$ at the gradient level, rather than requiring a manually fixed weighting parameter. Thus, the learned representation can be viewed as a Pareto-consistent compromise between LLM-derived task-aware guidance and data-driven fine structural regularization.*

C More Experiments

This section provides additional experimental details and results for TagBO. We first describe the six benchmark tasks used in our experiments. For each task, we summarize the optimization objective, search space, input variables, and output variables. We then present supplementary experimental results, including an ablation study on different feature components and a stability analysis of repeated feature extraction runs.

C.1 Dataset

This section provides detailed descriptions of the six experiments used in this study. Each task focuses on the optimization of a particular system or material, with the goal of improving performance based on various parameters.

Organic Photovoltaics [34]. This task aims to optimize the choice of molecular building blocks for an organic photovoltaic candidate, where the goal is to maximize the **PredictedSO**, a DFT-predicted spectral overlap score. The search space is fully discrete and consists of selecting one **DB_block** from 22 options and one **A_block** from 100 options, yielding 2200 possible combinations. The optimization therefore focuses on identifying the pair of building blocks whose combined electronic and molecular characteristics lead to the highest predicted spectral overlap.

Input/Output	Name	Range / Count	Description	Type
Inputs	DB_block	22	Donor-bridge molecular building block	Categorical
	A_block	100	Acceptor molecular building block	Categorical
Output	PredictedSO	$[0, +\infty)$	DFT-predicted spectral overlap score	Continuous

Table 1: Input and Output variables for Organic Photovoltaics task.

Suzuki–Miyaura Reaction [28]. This task aims to maximize reaction yield on a fixed-substrate Suzuki–Miyaura reaction subset by optimizing three categorical reaction conditions: **ligand**, **base**, and **solvent**. In this subset, the electrophile, nucleophile, and catalyst are fixed, and the search space is fully discrete, consisting of 12 ligand options, 8 base options, and 4 solvent options, yielding 384 possible reaction conditions. The optimization therefore focuses on identifying the combination of ligand, base, and solvent that gives the highest **yield** for the fixed Suzuki–Miyaura coupling reaction.

Input/Output	Name	Range / Count	Description	Type
Inputs	ligand	12	Ligand selected from 12 categorical options	Categorical
	base	8	Base selected from 8 categorical options	Categorical
	solvent	4	Solvent selected from 4 categorical options	Categorical
Output	yield	$[0, 100]$	Reaction yield to be maximized	Continuous

Table 2: Input and Output variables for Suzuki–Miyaura Reaction task.

Organic Laser Molecules [32]. This task aims to optimize the choice of molecular fragments for symmetric pentamer candidates designed for organic dye laser applications. The search space is fully discrete and consists of selecting one **frag_a** from 14 options, one **frag_b** from 13 options, and one **frag_c** from 19 options, yielding 3458 possible molecular combinations. The optimization is multi-objective: it seeks to maximize the **peak_score**. The task therefore focuses on identifying fragment combinations that jointly improve lasing-related photophysical performance.

Input/Output	Name	Range / Count	Description	Type
Inputs	frag_a	14	Molecular fragment selected for the first fragment position	Categorical
	frag_b	13	Molecular fragment selected for the second fragment position	Categorical
	frag_c	19	Molecular fragment selected for the third fragment position	Categorical
Outputs	peak_score	$[0, 1]$	Photophysical peak score to be maximized	Continuous

Table 3: Input and Output variables for Organic Laser Molecules task.

Perovskite Bandgap [19]. This task aims to optimize the composition of hybrid organic/inorganic perovskite materials, where the goal is to minimize the **hse_gap**, the band gap computed at the HSE06 level of theory. The search space is fully discrete and consists of selecting one **organic** component from 16 options, one **cation** from 3 options, and one **anion** from 4 options, yielding 192 possible compositions. The optimization therefore focuses on identifying the combination of organic cation, metal cation, and halide anion that leads to the smallest predicted band gap, while reflecting how the physicochemical properties of the individual perovskite components influence the electronic structure of the material.

Input/Output	Name	Range / Count	Description	Type
Inputs	organic	16	Organic component	Categorical
	cation	3	Metal cation	Categorical
	anion	4	Halide anion	Categorical
Output	hse_gap	$[0, +\infty)$	HSE06-computed band gap	Continuous

Table 4: Input and Output variables for Perovskite Bandgap task.

Redox Flow Battery [2]. This task aims to optimize the choice of substituent groups for redoxmer candidates used in non-aqueous redox flow batteries. The search space is fully discrete and consists of selecting one **r1_label** from 2 options, one **r3_label** from 8 options, one **r4_label** from 8 options, and one **r5_label** from 11 options, yielding 1408 possible molecular combinations. The optimization objective is to minimize the normalized average of **abs_lam_diff**, **ered**, and **gsol**, thereby identifying substituent combinations with improved overall redoxmer suitability for battery applications.

Input/Output	Name	Range / Count	Description	Type
Inputs	r1_label	2	Substituent group selected for the R1 position	Categorical
	r3_label	8	Substituent group selected for the R3 position	Categorical
	r4_label	8	Substituent group selected for the R4 position	Categorical
	r5_label	11	Substituent group selected for the R5 position	Categorical
Outputs	abs_lam_diff	$[0, +\infty)$	Absolute wavelength difference to be minimized	Continuous
	ered	$(-\infty, +\infty)$	Reduction potential to be minimized	Continuous
	gsol	$(-\infty, +\infty)$	Solvation free energy to be minimized	Continuous

Table 5: Input and Output variables for Redox Flow Battery task.

Crystallization Process [21]. This task aims to optimize the conditions of a vapor diffusion crystallization experiment, where the goal is to maximize the ordinal **crystal_score**. The search space is mixed-variable and includes the choice of **organic** cation and **solvent**, together with **concentration**,

volume, **time**, and **temperature** settings. Specifically, the dataset contains one categorical organic component, four continuous composition and volume variables, one categorical solvent choice, and two discrete operating conditions for reaction time and reaction temperature. The task focuses on identifying feasible crystallization conditions that promote better crystal formation outcomes.

Input/Output	Name	Range / Count	Description	Type
Inputs	organic	17	Organic ammonium component	Categorical
	organic_molarity	[0, 10]	Molarity of the organic component	Continuous
	solvent	3	Solvent	Categorical
	solvent_molarity	[0, 15]	Molarity of the solvent component	Continuous
	inorganic_molarity	[0, 5]	Molarity of the inorganic precursor	Continuous
	acid_molarity	[0, 25]	Molarity of the acid additive	Continuous
	alpha_vial_volume	[0, 0.001]	Volume of the alpha vial mixture	Continuous
	beta_vial_volume	[0, 0.001]	Volume of the beta vial mixture	Continuous
	reaction_time	3	Reaction time	Discrete
	reaction_temperature	3	Reaction temperature	Discrete
Output	crystal_score	4	Ordinal crystallization outcome with four ordered levels from clear solution to large crystallites	Ordinal

Table 6: Input and Output variables for Crystallization Process task.

C.2 Real-world Formulation Experiment

Task setting. This task aims to maximize the measured **thermal conductivity** of a thermally conductive adhesive while satisfying mechanical and electrical performance requirements. The mixed-variable search space involves selecting 10 fillers from a library of 29 candidate materials and optimizing their 10 corresponding continuous composition variables, as summarized in Table 7. The thermal conductivity qualification threshold is $2.8\,\mathrm{W}/(\mathrm{m}\cdot\mathrm{K})$. The formulation must also satisfy tensile strength $\geq 1\,\mathrm{MPa}$, shear strength $\geq 2\,\mathrm{MPa}$, and dielectric strength $\geq 7.9\,\mathrm{kV}/\mathrm{mm}$. The experiment includes 34 physical evaluations, each requiring approximately three days.

Input/Output	Name	Range / Count	Description	Type
Inputs	filler_1	29	First filler	Categorical
	filler_2	29	Second filler	Categorical
	filler_3	29	Third filler	Categorical
	filler_4	29	Fourth filler	Categorical
	filler_5	29	Fifth filler	Categorical
	filler_6	29	Sixth filler	Categorical
	filler_7	29	Seventh filler	Categorical
	filler_8	29	Eighth filler	Categorical
	filler_9	29	Ninth filler	Categorical
	filler_10	29	Tenth filler	Categorical
	composition_1	–	Amount of filler_1	Continuous
	composition_2	–	Amount of filler_2	Continuous
	composition_3	–	Amount of filler_3	Continuous
	composition_4	–	Amount of filler_4	Continuous
	composition_5	–	Amount of filler_5	Continuous
	composition_6	–	Amount of filler_6	Continuous
	composition_7	–	Amount of filler_7	Continuous
	composition_8	–	Amount of filler_8	Continuous
	composition_9	–	Amount of filler_9	Continuous
	composition_10	–	Amount of filler_10	Continuous
Outputs	conductivity	–	Thermal conductivity ($\mathrm{W}/(\mathrm{m}\cdot\mathrm{K})$)	Continuous
	tensile	–	Tensile strength (MPa)	Continuous
	shear	–	Shear strength (MPa)	Continuous
	dielectric	–	Dielectric strength (kV/mm)	Continuous

Table 7: Input and Output variables for the Real-world Formulation Experiment.

Results. As shown in Fig. 6, the best measured thermal conductivity increases from $0.1823\,\mathrm{W}/(\mathrm{m}\cdot\mathrm{K})$ at the first evaluation to $2.995\,\mathrm{W}/(\mathrm{m}\cdot\mathrm{K})$ after 34 evaluations, exceeding the qualification threshold. The final formulation achieves a tensile strength of $3.36\,\mathrm{MPa}$, a shear strength of $2.92\,\mathrm{MPa}$, and a dielectric strength of $23.5\,\mathrm{kV}/\mathrm{mm}$, satisfying all three performance requirements.

C.3 More Results

Supplementary ablation experiment. For the ablation study in Fig. 7, We compare the full TagBO framework with two reduced variants using only LLM-derived scoring features or only physicochemical property features across six datasets, and the trajectories show that the full method generally reaches strong objective values in fewer optimization steps while achieving comparable or better final performance in most cases.

To investigate the effect of model scale and reasoning capability, we compare Intern-S1-241B Thinking, Qwen3-235B, Qwen3-30B (Instruct and Thinking), Qwen3-8B, Qwen3-4B, Llama-3-70B, and Llama-3-3B as LLM backbones, as shown in Fig. 8. Larger models and Thinking variants generally require fewer iterations to reach 95% of the oracle optimum, with Intern-S1-241B Thinking achieving the best overall sample efficiency.

Model Stability Analysis. To evaluate repeatability, we perform six independent runs of Intern-S1-241B Thinking on the Suzuki–Miyaura reaction dataset. As shown in Fig. 9, the runs show similar convergence trends, enter the high-performance regime within a narrow iteration range, and reach comparable best-so-far values. The mean Step@95 varies only slightly across runs, indicating limited variation in convergence speed. These results suggest that the proposed feature extraction pipeline is stable under repeated runs.

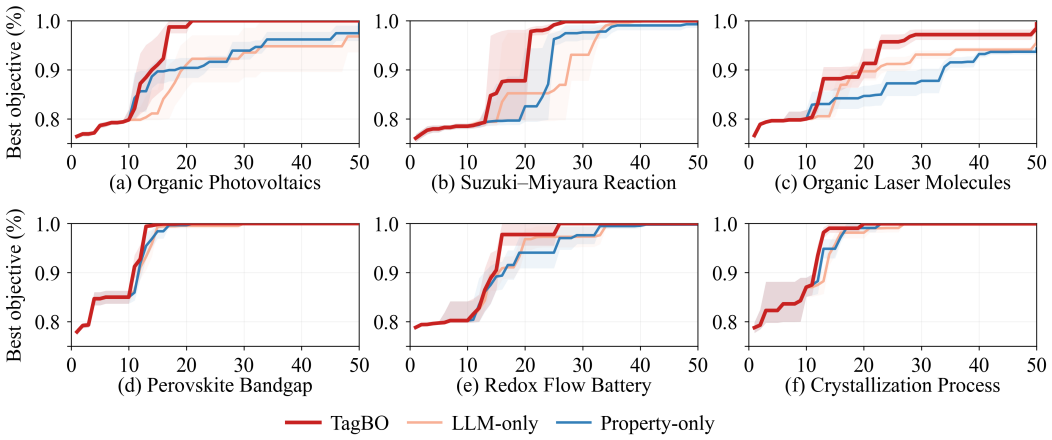


Figure 7: Ablation study of TagBO across six datasets. We compare the full TagBO framework with two reduced variants using only LLM-derived scoring features or only physicochemical property features.

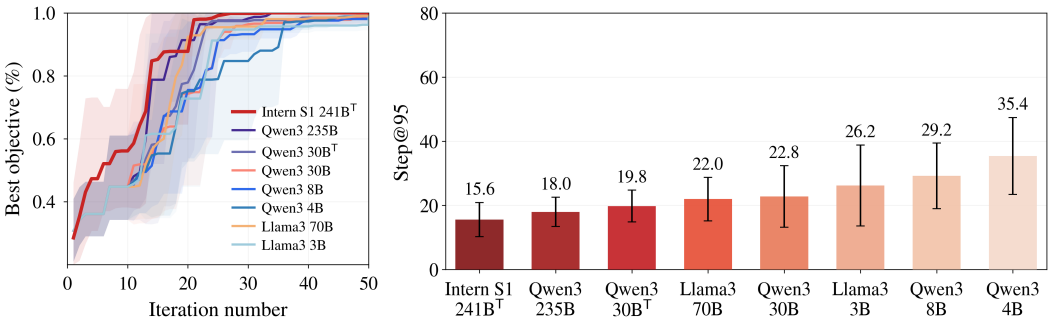


Figure 8: Optimization performance of LLM variants for feature extraction on the Suzuki–Miyaura reaction dataset. **Left**: the optimization trajectories, where curves denote the best objective value found so far, normalized by the oracle optimum. **Right**: the mean Step@95%, where bar height indicates the average number of iterations required to reach 95% of the optimum, and error bars reflect variation across seeds. $*^T$ denotes Thinking variants.

D Limitations

The effectiveness of TagBO depends on the quality of the task-aware relational priors extracted from the LLM. In domains where relevant scientific knowledge is sparse, highly specialized, or poorly represented in the pretraining corpus, the induced ordinal relations may become noisy or misaligned with the downstream optimization objective. We believe that retrieval-augmented generation (RAG) and domain-specific knowledge integration may help alleviate this limitation by grounding relational reasoning in external scientific resources, and we plan to explore this direction in future work.

In addition, TagBO primarily modifies the representation and surrogate prior of Bayesian optimization while leaving the underlying BO optimization procedure unchanged. Consequently, our theoretical analysis focuses on representation-level surrogate behavior rather than cumulative regret guarantees for the overall sequential optimization process. Deriving regret bounds that explicitly characterize the interaction between task-aware representations, kernel geometry, and acquisition dynamics remains an important direction for future research.

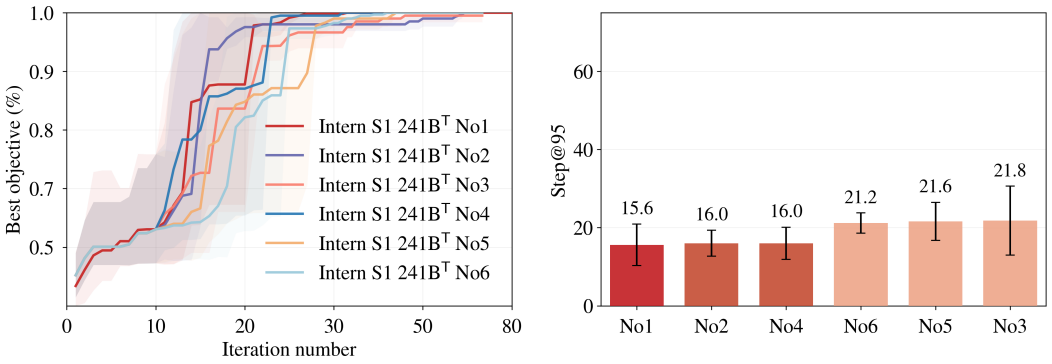


Figure 9: Optimization performance of repeated Intern-S1-241B Thinking variants for feature extraction on the Suzuki–Miyaura reaction dataset. **Left**: the optimization trajectories, where curves denote the best objective value found so far, normalized by the oracle optimum. **Right**: the mean Step@95%, where bar height indicates the average number of iterations required to reach 95% of the optimum, and error bars reflect variation across seeds. $*^{T}$ denote Thinking variants

LLM Usage Statement

Large language models (LLMs) are core components of the proposed method. Specifically. But, LLMs were not used to generate experimental results automatically without human verification. All reported analyses, evaluations, and conclusions were verified by the authors.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction accurately summarize the main contributions and scope of the paper.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The limitations is discussed in Appendix.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

Justification: Proof is provided in Appendix.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We provide code in supplemental materials.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: We provide code in supplemental materials.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: We provide code in supplemental materials.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: Variance across runs is reported using shaded regions in Fig. 2 and Fig. 4(a,b), while box plots are used to illustrate the variability in the ablation studies.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.

- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Variance across runs is reported using shaded regions in Fig. 2 and Fig. 4(a,b), while box plots are used to illustrate the variability in the ablation studies.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research complies with the NeurIPS Code of Ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [N/A]

    Justification: There is no societal impact of the work performed.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The proposed method does not introduce high-risk generative models or sensitive datasets requiring special safeguards.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: All datasets and external codebases used in this work are properly cited, and their licenses and usage terms are respected.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.

- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer:[N/A]

    Justification: This paper does not introduce new public assets such as datasets or models.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: This work does not involve crowdsourcing or human subject experiments.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: This work does not involve human subject research requiring IRB approval.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [Yes]

Justification: LLMs are core components of the proposed framework

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.