EDGE: HIGH-DIMENSIONAL MIXED-VARIABLE BAYESIAN OPTIMIZATION WITH EIGEN-DETECTED GRADIENT EMBEDDINGS

Anonymous authors
Paper under double-blind review

ABSTRACT

Hyperparameter optimization and physical-system design often require searching high-dimensional, mixed-variable spaces, yet expensive evaluations leave little room for wasted exploration. Bayesian optimization (BO) handles the expense, but existing subspace methods often rely on random embeddings or fix the number of gradient-informed directions without testing whether they carry signal. We propose EDGE, a high-dimensional and mixed-variable BO algorithm that uses a differentiable tabular foundation model (TFM) as the surrogate and thus allows us to ask the model itself which directions of the search space matter. EDGE combines spectral direction selection without a prescribed subspace rank with optimal-transport probabilistic reparameterization for geometry-aware optimization of discrete variables. We evaluate EDGE on 32 benchmark problems, including mixed-variable and combinatorial domains, spanning 3$\sim$450 dimensions under a budget of 300 evaluations after initialization. Across these diverse benchmarks, EDGE achieves the best average rank among seven methods. EDGE ranks first on both applied-science and engineering tasks and HPO tasks, while Bounce leads on the synthetic benchmarks. A separate combinatorial study examines the limits of the approach.

1 INTRODUCTION

Optimization of expensive black-box objective functions is common in hyperparameter optimization (HPO), engineering design, and scientific discovery (Snoek et al., 2012; Zhang et al., 2020; Yu et al., 2025). These applications often require searching over a mixed-variable design space with both continuous parameters (e.g., numerical hyperparameters) and discrete choices (e.g., model architecture choices), and are often high-dimensional (Ru et al., 2020a; Wan et al., 2021; Daulton et al., 2022; Papenmeier et al., 2023). Bayesian optimization (BO) is the method of choice in these settings due to its sample efficiency (Shahriari et al., 2015; Binois & Wycoff, 2022; Papenmeier et al., 2023). Yet, extending its sample efficiency to mixed-variable problems with hundreds of dimensions requires search strategies that can exploit information shared across variables while preserving the discrete choices that define a valid configuration.

Existing methods offer complementary ways to address this challenge. Mixed-variable BO methods such as Casmopolitan (Wan et al., 2021) and probabilistic reparameterization (Daulton et al., 2022) (PR) provide mechanisms for modeling and optimizing over discrete choices. Embedding methods reduce the dimension of the search, with Bounce (Papenmeier et al., 2023) extending adaptive random embeddings to mixed and combinatorial domains. However, random-embedding methods do not adapt their directions to the observed objective (Wang et al., 2016; Nayebi et al., 2019; Papenmeier et al., 2022; 2023), while existing surrogate-informed subspace methods select learned directions without testing whether they separate from finite-sample spectral noise and are primarily formulated for continuous domains (Constantine, 2015; Papenmeier et al., 2022; Yu et al., 2026b).

Tabular foundation models (TFMs) provide a natural basis to address the gap above (Hollmann et al., 2025). Pretrained on collections of tabular tasks, they produce predictive distributions conditioned on observed data without updating their weights during optimization (Müller et al., 2023). Differentiable TFMs also expose how those predictions vary across the input space. GIT-BO demon-

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

strates how predictive-mean gradients can guide search in continuous domains (Yu et al., 2026b). For mixed-variable optimization, translating this information into candidate configurations requires both a rule for selecting search directions from finite gradient samples and an acquisition optimizer that can refine discrete decisions.

We introduce EDGE, a BO framework that combines adaptive search directions with optimal-transport probabilistic reparameterization (OTPR) for high-dimensional mixed-variable problems. Our key idea is to use a differentiable surrogate's acquisition gradients to guide candidate generation and probability distributions to refine discrete choices. A spectral rule selects and weights search directions without prescribing a subspace rank. Optimal transport incorporates distances between admissible levels into probability updates, connecting directional search to mixed-variable acquisition optimization within a trust region. Together, these components adapt the search to the observations using a frozen tabular foundation model, and we evaluate the method across real-world design and HPO problems.

Our contributions are:

- We introduce EDGE, a mixed-variable Bayesian optimization framework that constructs adaptive search directions from the acquisition gradients of a frozen, original-space foundation-model surrogate. EDGE combines Marchenko–Pastur(MP)-referenced spectral selection, overlap-inspired proposal weighting, and a random-sketch fallback, avoiding a prescribed subspace rank.
- We develop a transport-aware extension of probabilistic reparameterization that combines ground-cost smoothing with graph-transport natural-gradient updates, integrating discrete probability refinement with acquisition-guided candidate generation and trust-region search.
- We evaluate EDGE on a 32-problem suite spanning 3–450 dimensions, with separate analyses of engineering, HPO, and synthetic tasks. Component ablations examine surrogate, rank, and refinement choices, while a combinatorial study investigates limitations in surrogate suitability and encoding sensitivity.

2 BACKGROUND

Mixed-variable Bayesian optimization and probabilistic reparameterization We maximize an expensive black-box function f over $\mathcal{X} \times \mathcal{Z}$, where $\mathcal{X} \subseteq \mathbb{R}^{d_c}$ is continuous and $\mathcal{Z}$ holds d_z discrete (binary, ordinal, or categorical) variables. At each iteration, BO conditions a probabilistic surrogate on the observed data and selects the next point by maximizing an acquisition function (AF) $\alpha(x, z)$ that balances predicted performance and uncertainty (Jones et al., 1998; Srinivas et al., 2010). The difficulty is the inner problem $\max_{(x,z)} \alpha(x, z)$, since standard continuous gradient-based optimization cannot be applied directly to discrete choices. Several existing BO methods target this setting. Casmopolitan (Wan et al., 2021) combines trust regions with kernels tailored to categorical and continuous inputs. CoCaBO (Ru et al., 2020a) treats the categorical part as a multi-armed bandit. BODi (Deshwal et al., 2023) embeds combinatorial variables through a dictionary of anchor points. A complementary approach is to optimize continuous probability distributions over admissible values.

This approach predates its use in BO: variational optimization optimizes the expectation of a potentially discrete or non-differentiable objective under a parameterized distribution (Staines & Barber, 2012). Probabilistic reparameterization (PR) (Daulton et al., 2022) applies this construction to mixed-variable acquisition optimization: it replaces the discrete point z by a distribution $p(Z \mid \theta)$ with continuous parameters θ, and maximizes the probabilistic objective $\mathbb{E}_{Z \sim p(\cdot|\theta)}[\alpha(x, Z)]$, which is differentiable in (x, θ) when the acquisition and parameterization have the required differentiability. For PR's distribution families, with α continuous in x for each z, sampling from an exact maximizer of the probabilistic objective yields an acquisition maximizer. A BO policy driven by an exact maximizer of the probabilistic objective inherits the regret guarantees of the underlying acquisition when their assumptions hold; this does not guarantee exact maximization with finitely many gradient steps. PR ascends this objective with a REINFORCE estimator (Williams, 1992; Mohamed et al., 2020) through a softmax-temperature map. The score-function estimator differentiates distribution parameters without differentiating through discrete samples.

Because the optimization variables represent probability distributions, their geometry can inform the updates. Natural-gradient methods account for the geometry of a statistical model (Amari, 1998), while Wasserstein natural gradients additionally incorporate a ground metric between outcomes (Li & Montúfar, 2018). For nominal categories, this ground metric requires an explicit modeling choice; category labels alone do not define an ordering.

High-dimensional Bayesian optimization with random embeddings When an objective depends primarily on a low-dimensional subspace, BO can search in fewer dimensions than the ambient input space. REMBO (Wang et al., 2016) uses a random linear projection. HeSBO (Nayebi et al., 2019) uses a count-sketch, which hashes each input dimension into one of a few buckets with a random sign. BAxUS (Papenmeier et al., 2022) grows a nested count-sketch embedding over time and Bounce (Papenmeier et al., 2023) extends this method to mixed and combinatorial spaces. Random embeddings reduce the search dimension without estimating objective-specific directions, motivating an approach that derives its search basis from the surrogate. Embeddings also affect surrogate modeling: Letham et al. (2020) show that linear projections can change the kernel structure needed to model the embedded function. This motivates distinguishing the coordinates used for surrogate modeling from the directions used to generate candidates.

Tabular foundation models and gradient-informed subspaces Prior-data fitted networks (PFNs) are trained on tasks sampled from a prior to approximate posterior predictive distributions from an observed context (Müller et al., 2022). TabPFN develops this approach for tabular prediction (Hollmann et al., 2023; 2025), and PFNs4BO applies it to BO (Müller et al., 2023). Their weights remain fixed during optimization, avoiding per-iteration weight training while still requiring inference on the growing context.

Active-subspace methods identify directions of average variation using the leading eigenvectors of an expected gradient outer-product matrix (Constantine et al., 2014). GIT-BO (Yu et al., 2026b) uses predictive-mean gradients from a differentiable TabPFN surrogate to build an active subspace: it averages gradient outer products and searches the top-r eigenspace. GIT-BO selects a subspace using a fixed rank or an explained-variance threshold, leaving open how to calibrate direction selection to finite-sample uncertainty. The Marchenko–Pastur law describes the limiting sample-covariance spectrum under a noise model (Marčenko & Pastur, 1967). Under a spiked-covariance model, random matrix theory relates spectral separation to eigenvector alignment (Baik et al., 2005; Paul, 2007), motivating EDGE's Marchenko–Pastur selection rule.

These developments suggest combining surrogate-derived search directions with probabilistic optimization of discrete choices, while accounting for uncertainty in the estimated gradient spectrum. The next section describes how EDGE combines acquisition-gradient directions with optimal-transport probabilistic refinement within a trust region.

3 The EDGE Algorithm

EDGE conditions a differentiable original-space surrogate, selects directions from its acquisition-gradient spectrum, initializes candidates using those directions, and refines feasible mixed-variable proposals. When the spectral rule selects no directions, it uses a random-sketch proposal branch. Figure 1 illustrates the pipeline, and Algorithm 1 summarizes the BO loop.

3.1 Which directions matter: the gradient outer-product matrix

EDGE conditions a differentiable surrogate on observations in the original d-dimensional space and uses acquisition gradients to guide candidate generation. We normalize the inputs to $[0, 1]^d$, rescaling continuous variables and encoding each discrete variable with $k \geq 2$ levels at k equally spaced points. The surrogate defines a continuous relaxation of the acquisition α_{UCB}, whose input gradients are obtained by automatic differentiation. For discrete coordinates, these are gradients of the relaxed acquisition.

EDGE uses a trust region, a bounded neighborhood of the best observed point, to focus each search on a promising part of the domain (Eriksson et al., 2019). Acquisition gradients can vary across the domain, so we estimate directions within the trust region used for candidate generation. We

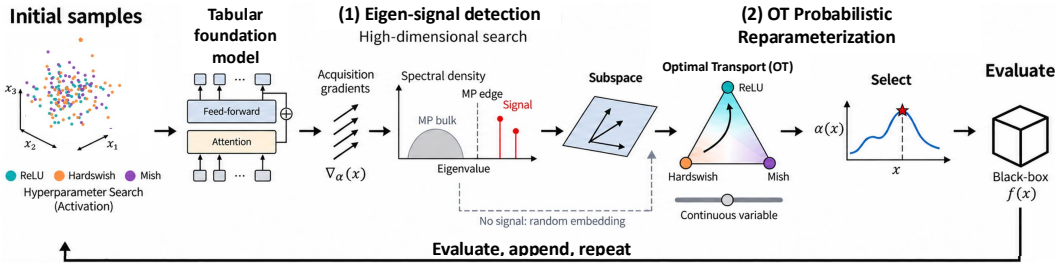

Figure 1: Overview of EDGE. At each BO iteration, acquisition gradients at probe locations define an uncentered second-moment matrix. An MP-referenced spectral rule selects directions for candidate initialization. Clipping, discrete snapping, and OTPR produce feasible proposals within the specified trust region. A random signed count-sketch supplies proposals when no direction is selected.

draw M probe locations and perturb them with Gaussian jitter, using the implementation's specified boundary handling. Jitter changes the probe distribution and is a numerical design choice. It does not, by itself, identify our estimator with the smoothed-gradient estimator of (Yuan et al., 2025) or establish that estimator's convergence rate. We form the uncentered gradient second-moment matrix as follows.

$$\widehat{C} = \frac{1}{M}\sum_{j=1}^{M} g_j g_j^\top, \qquad g_j = \nabla_x \alpha(x_j) \in \mathbb{R}^d, \qquad \gamma = d/M. \tag{1}$$

Directions in which the acquisition surface tilts strongly produce large eigenvalues of $\widehat{C}$; flat directions produce small ones. The aspect ratio γ measures how sample-starved the estimate is, and it is the quantity Section 3.2 runs on. The entire measurement costs one batched backward pass, requiring no objective evaluation.

3.2 Which directions do we select: an MP-referenced spectral rule

Let $\widehat{C} = M^{-1}\sum_{j=1}^{M} g_j g_j^\top$, and write $\widehat{\lambda}_1 \geq \cdots \geq \widehat{\lambda}_d \geq 0$ for its eigenvalues. We use the Marchenko–Pastur (MP) upper edge as a reference for selecting directions of acquisition variation. For noise scale σ_g^2 and aspect ratio $\gamma = d/M$, the reference edge is $\sigma_g^2(1+\sqrt{\gamma})^2$. With the scale estimate $\widehat{\sigma}_g^2$ returned by our bulk-fitting procedure, the default rule at $M = 256$ is

$$\widehat{r} = \#\{i : \widehat{\lambda}_i > \widehat{\sigma}_g^2(1+\sqrt{\gamma})^2\}. \tag{2}$$

This rule avoids prescribing the search rank. The scale estimator, numerical tolerances, and any edge multiplier remain algorithm settings; the limiting edge alone is not a finite-sample significance threshold.

Theorem 1 (Gaussian spiked reference model)**.** *Consider a sequence of models with fixed r, fixed pairwise distinct spike strengths $\ell_1 > \cdots > \ell_r > 1$, deterministic orthonormal directions $u_{1,d}, \ldots, u_{r,d}$, and fixed $\sigma_g^2 > 0$. Let*

$$C_d = \sigma_g^2\left(I_d + \sum_{i=1}^{r}(\ell_i - 1)u_{i,d}u_{i,d}^\top\right), \qquad g_j = C_d^{1/2} z_j, \qquad z_j \overset{\text{iid}}{\sim} N(0, I_d).$$

As $d, M \to \infty$ with $d/M \to \gamma \in (0,\infty)$, let $(\widehat{\lambda}_i, \widehat{u}_i)$ be the ordered sample eigenpairs. For each spike,

$$\frac{\widehat{\lambda}_i}{\sigma_g^2} \longrightarrow \begin{cases} (1+\sqrt{\gamma})^2, & \ell_i \leq 1+\sqrt{\gamma}, \\ \ell_i\left(1 + \dfrac{\gamma}{\ell_i - 1}\right), & \ell_i > 1+\sqrt{\gamma}, \end{cases} \tag{3}$$

$$|\widehat{u}_i^\top u_{i,d}|^2 \longrightarrow \begin{cases} 0, & \ell_i \leq 1+\sqrt{\gamma}, \\ \dfrac{1-\gamma/(\ell_i-1)^2}{1+\gamma/(\ell_i-1)}, & \ell_i > 1+\sqrt{\gamma}, \end{cases} \tag{4}$$

almost surely. The strict supercritical condition can equivalently be written $M > d/(\ell_i - 1)^2$ at the level of the limiting aspect-ratio comparison. It is not a finite-sample sample-size guarantee.

These limits follow from Gaussian spiked-covariance theory (Paul, 2007; Benaych-Georges & Nadakuditi, 2011). The general framework for multiplicative perturbations also accommodates $\gamma > 1$, where the MP law has an atom at zero. Repeated spikes require an eigenspace formulation rather than alignment with an individually identified population eigenvector.

For a retained normalized eigenvalue $y_i = \widehat{\lambda}_i/\widehat{\sigma}_g^2 > (1+\sqrt{\gamma})^2$, inverting the outlier-location relation gives the plug-in quantities

$$a_i = y_i - 1 - \gamma, \qquad \widehat{\ell}_i = 1 + \frac{a_i + \sqrt{a_i^2 - 4\gamma}}{2},$$
$$\widehat{c}_i = \left(\frac{1 - \gamma/(\widehat{\ell}_i - 1)^2}{1 + \gamma/(\widehat{\ell}_i - 1)} \right)^{1/2}. \tag{5}$$

We use $\widehat{c}_i$ as an overlap-inspired proposal weight. This choice is motivated by eigenvector alignment; it is not an optimal BO step-length result. The reference theorem does not establish that the implemented acquisition gradients satisfy its Gaussian assumptions, that fitted-scale admissions have a controlled false-positive rate, or that discarded directions cannot improve optimization. Its population directions describe acquisition-gradient variation, not automatically the active subspace of the objective.

3.3 How do we refine subspace-initialized candidates: optimal-transport probabilistic reparameterization

Given the retained eigenvectors $U_{\widehat{r}}$ and weights $\widehat{c}$, we initialize proposals around the incumbent x_{ctr}:

$$x_{\mathrm{prop}} = \mathrm{clip}_{[\mathrm{lo},\mathrm{hi}]}\left(x_{\mathrm{ctr}} + L U_{\widehat{r}}(\widehat{c} \odot \zeta)\right), \qquad \zeta \in \mathrm{Sobol}([-1,1]^{\widehat{r}}). \tag{6}$$

For nonbinary coordinates, the bounds are $\mathrm{lo} = \max(x_{\mathrm{ctr}} - L, 0)$ and $\mathrm{hi} = \min(x_{\mathrm{ctr}} + L, 1)$; binary coordinates retain bounds $[0, 1]$. We reserve 25% of the candidate pool for full-dimensional proposals, snap discrete coordinates to admissible levels, and refine the highest-scoring feasible candidates. The affine-span restriction applies before clipping and snapping. These operations and subsequent probability updates can leave that span. Full binary bounds also do not guarantee a nonzero binary move: for $u = \mathbf{1}/\sqrt{d}$, every rank-one informed proposal snaps back to a binary incumbent whenever $L\widehat{c}/\sqrt{d} < 1/2$.

For fixed continuous block x, OTPR minimizes $F(\phi) = -\mathbb{E}_{Z \sim p_\varepsilon(\cdot|\phi)}\alpha(x, Z)$. For each discrete factor, let $q(\phi) = \mathrm{softmax}((\phi - 0.5)/\tau)$, and let $D_{ab} = d(a, b)^2$ be a specified ground cost. Define the Gibbs-kernel map

$$p_\varepsilon(\phi) = \frac{Kq(\phi)}{\mathbf{1}^\top Kq(\phi)}, \qquad K_{ab} = \exp(-D_{ab}/\varepsilon). \tag{7}$$

The joint distribution is the product of the factors. This is a normalized kernel smoothing map, not the solution of an entropic optimal-transport coupling problem. The ground metric is a modeling choice; nominal labels alone do not supply one.

We combine this map with a graph-transport metric. On each level set choose a connected graph and specify its positive interior edge conductances, forming a weighted Laplacian $L(p)$. In independent probability parameters, with $J_\phi p_\varepsilon$ of full column rank, set

$$G_W^{\mathrm{comp}}(\phi) = (J_\phi p_\varepsilon)^\top L(p_\varepsilon)^\dagger J_\phi p_\varepsilon, \qquad \phi_{k+1} = \phi_k + \eta_k (G_W^{\mathrm{comp}}(\phi_k))^{-1} \widehat{g}_k, \tag{8}$$

where $\widehat{g}_k$ estimates $-\nabla_\phi F(\phi_k)$. Redundant softmax logits must be removed or handled by an equivalent constrained formulation. Numerical damping, projection, and stochastic gradient estimation are implementation choices, not consequences of the exact-flow analysis.

Building on the graph-transport natural gradient of Li & Montúfar (2018), Appendix D.5 establishes a uniform smoothing-error bound (Proposition 1) and conditional contraction of the exact flow at fixed x (Theorem 3). The contraction requires gradient dominance and existence of the flow in its smooth domain. It does not give a finite-step rate for the implemented optimizer. Sampled decoding is treated in Lemma 3, while search coverage and finite optimization errors remain explicit in Theorem 4.

Algorithm 1 EDGE (TabPFN surrogate and UCB by default)

Require: objective f, budget I, initial design size n_{init}, probe count M, acquisition α

1: Draw n_{init} initial points (shared Sobol design); evaluate; $\mathcal{D} \leftarrow \{(x_i, y_i)\}_{i=1}^{n_{\text{init}}}$; initialize the trust region.
2: **for** $n = 1$ **to** I **do**
3: Fit the surrogate on all of $\mathcal{D}$ in the original space; build $\alpha(x) = \mu(x) + \beta^{1/2}\sigma(x)$.
4: Draw M jittered probes in the trust region; $g_j \leftarrow \nabla_x \alpha(x_j)$; form $\widehat{C}$ by equation 1.
5: Estimate the gradient noise scale $\widehat{\sigma}_g^2$; select $\widehat{r}$ using equation 2; compute proposal weights $\widehat{c}_i$ using equation 5.
6: **if** $\widehat{r} \geq 1$ **then** generate the candidate pool from the overlap-weighted subspace $U_{\widehat{r}}(c \odot \zeta)$
7: **else** generate the pool from a random signed count-sketch of dimension $d/5$
8: **end if**
9: Add 25% full-dimensional Sobol candidates; snap all candidates to the discrete grids.
10: Refine the top pool candidates with OTPR (Section 3.3) inside the trust region.
11: Evaluate the winner x_{next}; append $(x_{\text{next}}, f(x_{\text{next}}))$ to $\mathcal{D}$; update the trust region; checkpoint.
12: **end for**
13: **return** $x^\star = \arg\max_{(x,y)\in\mathcal{D}} y$

3.4 When no direction is selected: random-sketch fallback

When $\widehat{r} = 0$, EDGE generates candidates with a random signed count-sketch, following HeSBO (Nayebi et al., 2019). Each original coordinate is assigned to a search coordinate with a random sign; the target sketch dimension and rounding rule are implementation settings. Both proposal branches retain the full-dimensional candidate component and use the same refinement procedure. This is an alternative proposal mechanism when spectral selection is inconclusive, rather than a finite-budget guarantee or evidence that the objective lacks structure.

4 Experimental Setup

Baseline algorithms We compare EDGE against six baselines that span the mixed-variable and high-dimensional families: random search (Bergstra & Bengio, 2012), PR (Daulton et al., 2022), Casmopolitan (Wan et al., 2021), BODi (Deshwal et al., 2023), SMAC3 (Lindauer et al., 2022), and Bounce (Papenmeier et al., 2023) from the state-of-the-art (SOTA) PyTorch-based BO library BoTorch (Balandat et al., 2020) . Details of implementations are in Appendix F.

Test problems We evaluate EDGE on 32 benchmarks spanning 3∼450 dimensions: 25 real-world problems, comprising 16 applied-science and engineering tasks and nine HPO tasks, and seven synthetic benchmarks. The suite covers both mixed-variable and fully discrete search spaces, combining established BO benchmarks with application-specific adaptations. Appendix G lists each problem's source, dimensions, variable types, and implementation changes. Additional combinatorial and relabeling studies appear in Appendix E.

Experiment Protocol Every method solves every problem with 10 seeds and a budget of 300 evaluations after a shared initial sample $n_{\text{init}} = 20$. For each problem and seed, all methods start from the same initial points. To run this extensive benchmarking process, we use a distributed server infrastructure featuring Intel Xeon Platinum 8480+ CPUs and NVIDIA H200 GPUs. For all algorithms, each individual experiment (run) was conducted with the same amount of compute allocated: a single H200 GPU node with 16 CPU cores and 200GB RAM.

Evaluation metrics We compare methods using average rank and normalized regret. For each problem, we average the best observed objective over seeds and rank the methods, with rank 1 indicating the best mean outcome. We then average these ranks with equal weight per problem. To also measure the size of the remaining objective gap, let $b_{j,m,s}(t)$ be the best maximization score

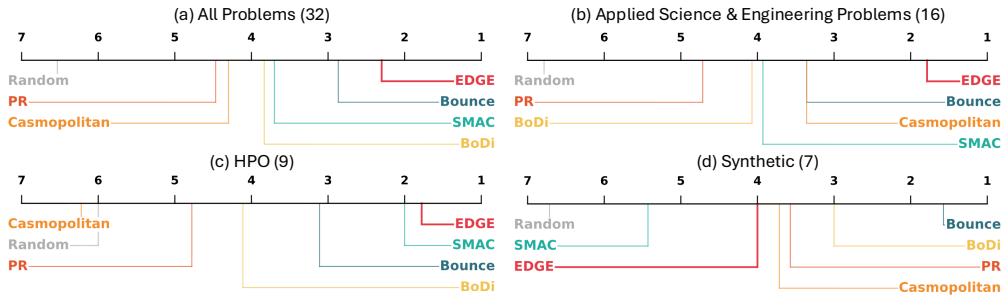

Figure 2: Average ranks at iteration 300 for (a) the full benchmark suite, (b) applied-science and engineering tasks, (c) HPO tasks, and (d) synthetic benchmarks. Lower ranks are better. EDGE is highlighted in red. Panel titles give the number of included problems. The complete statistical comparisons are reported in Appendix C.

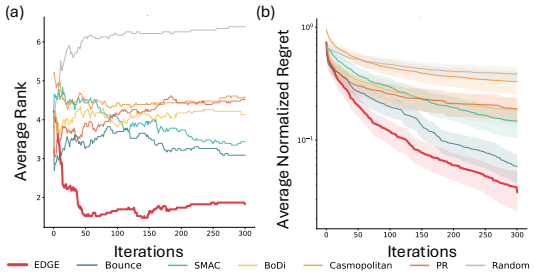

Figure 3: Progress with the evaluation budget: (a) average rank and (b) average normalized incumbent regret, with the latter plotted on a logarithmic scale. Lower values are better. EDGE is shown in red. Shaded bands indicate unit standard deviation.

observed by method m on problem j in seed s by budget t. We define

$$r^{\text{norm}}_{j,m,s}(t) = \frac{f^{\text{ref}}_j - b_{j,m,s}(t)}{s_j}, \qquad s_j > 0, \tag{9}$$

where f^{ref}_j and s_j are fixed task-specific reference values shared by all methods, seeds, and budgets. We average first over seeds within each problem and then equally over problems. Lower values are better. When the reference is empirical rather than a known optimum, this metric measures a gap to that reference, not certified regret to the global optimum. The statistical testing procedure and pairwise results are reported in Appendix C.

5 Results

Overall ranking EDGE achieves the lowest average rank in the aggregate comparison, with Bounce as the closest competitor (Figure 2a). The subgroup results reveal where this advantage arises: EDGE leads on both applied-science and engineering tasks and HPO tasks, whereas Bounce leads on the synthetic benchmarks. EDGE's aggregate advantage therefore reflects complementary strengths across the suite rather than uniform dominance on every problem. Appendix C reports the per-problem curves and statistical comparisons.

Where the advantage lives EDGE ranks first in both the applied-science and engineering group and the HPO group, while Bounce leads on the synthetic benchmarks (Figure 2b–d). The benchmark inventory also differs in variable composition: five of seven synthetic tasks are fully discrete, compared with six of 25 real-world tasks (Table 6). For EDGE, useful acquisition gradients must translate into useful changes on this feasible domain. The controlled surrogate comparison in Appendix E.3 identifies surrogate suitability as one contributor to the gap on MaxSAT and cluster expansion. This provides a task-specific explanation; it does not isolate the causes of the broader difference between benchmark groups.

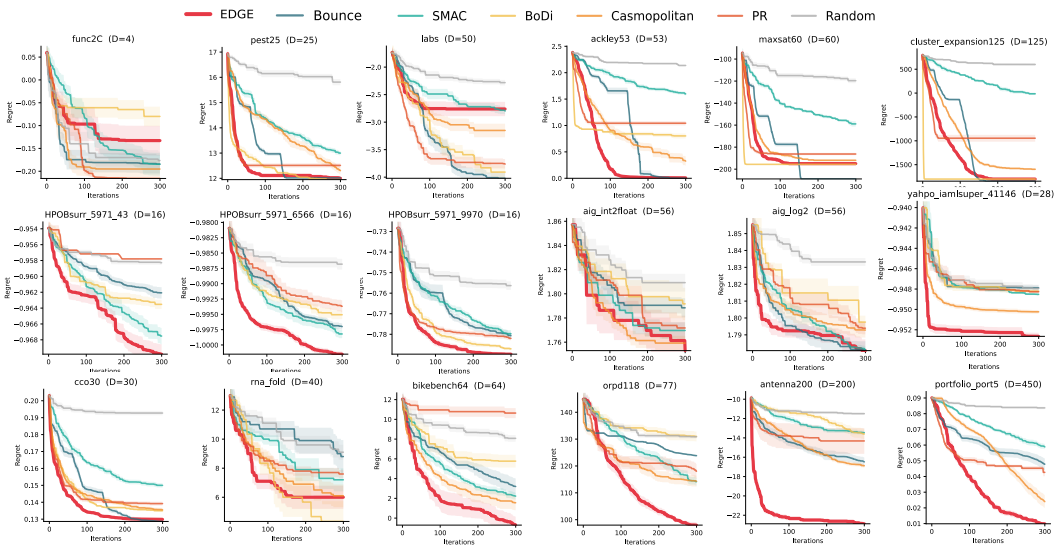

Figure 4: Per-problem convergence on 18 selected benchmarks from the 32-problem suite. Panel titles give the dimension D. Curves show mean regret values in the plotting convention, oriented so that lower is better. The top row contains synthetic tasks; the remaining rows contain real-world tasks. EDGE is shown in red.

Performance across dimensions EDGE's strongest visual separations occur on antenna design ($d = 200$) and portfolio optimization ($d = 450$), where it continues improving after several baselines slow down (Figure 4). At lower dimensions, the gains vary by task, with close outcomes on some circuit and coverage problems. The cluster-expansion task ($d = 125$) shows that high dimension alone does not favor EDGE: other methods reach better or comparable final values there. These comparisons identify promising applications for the complete EDGE pipeline without attributing its performance to dimension alone.

Progress under limited budgets EDGE's advantage develops during the optimization run, rather than appearing only at the final evaluation. The aggregate optimization trajectories in Figure 3 show sustained reductions in normalized regret and a favorable average rank after the initial search phase. This behavior is relevant when evaluations are expensive and the run may stop before exhausting its budget. These curves compare progress per evaluation; the separate runtime comparison in Appendix C.2 measures the computation required to choose each evaluation.

Sensitivity to discrete encodings Relabeling preserves the underlying optimization problem but can change EDGE's search behavior. In the separate combinatorial study, EDGE's mean final value decreases from 1765 to 884 on the original and relabeled cluster-expansion problem, whereas Bounce remains at 1765 (Table 4). This exposes a limitation of the current pipeline and motivates encoding-aware refinements, though it does not identify which component causes the sensitivity.

Component ablations The ablations identify which choices influence EDGE's performance. With TabPFN retained as the surrogate, the detected-rank configuration achieves better final aggregate performance than the fixed-$r = 10$ variant (Appendix B.2). OTPR also improves the final objective relative to Euclidean PR on all five displayed tasks (Appendix B.5). Surrogate rankings vary across problems, showing that EDGE can benefit from different backends (Appendix B.1). Finally, using fallback dimension $d/4$ improves the mean outcome on seven of ten ablation tasks, indicating scope for adaptive fallback settings (Appendix B.4).

6 CONCLUSION

EDGE combines acquisition-gradient candidate generation, adaptive spectral direction selection, and transport-aware probabilistic refinement for high-dimensional mixed-variable Bayesian optimization. Its original-space surrogate supports changing search directions without changing the representation of past observations. The empirical comparison shows its strongest performance on real-world science and engineering application and HPO benchmarks, while the combinatorial study identifies limitations in surrogate suitability and encoding sensitivity. The theoretical analysis clarifies the assumptions linking spectral estimation, probability refinement, and outer-loop regret, with finite search errors retained explicitly.

Limitations: EDGE trades additional computation for progress under an evaluation budget: in the runtime study, it requires 26.20 s per iteration versus 3.92 s for Bounce (Appendix C.2). Its performance also depends on surrogate suitability, and relabeling exposes sensitivity on some combinatorial tasks (Appendix E). Adaptive fallback dimensions and encoding-aware candidate generation are promising directions. The reference-model analysis does not establish a calibrated detector for arbitrary surrogate gradients or a finite-budget guarantee for the fallback. Finally, the main suite contains only two real-world tasks above 100 dimensions (antenna design and portfolio optimization) so broader evaluation at that scale remains important.

REFERENCES

Moustafa Alzantot, Yash Sharma, Supriyo Chakraborty, Huan Zhang, Cho-Jui Hsieh, and Mani B Srivastava. Genattack: Practical black-box attacks with gradient-free optimization. In *Proceedings of the genetic and evolutionary computation conference*, pp. 1111–1119, 2019.

Shun-Ichi Amari. Natural gradient works efficiently in learning. *Neural computation*, 10(2):251–276, 1998.

Luca Amarú, Pierre-Emmanuel Gaillardon, and Giovanni De Micheli. The EPFL combinational benchmark suite. In *Proceedings of the 24th International Workshop on Logic & Synthesis (IWLS)*, 2015.

Sebastian Ament, Samuel Daulton, David Eriksson, Maximilian Balandat, and Eytan Bakshy. Unexpected improvements to expected improvement for bayesian optimization. In *Advances in Neural Information Processing Systems*, volume 36, pp. 20577–20612, 2023.

Hossain M Amir and Takashi Hasegawa. Nonlinear mixed-discrete structural optimization. *Journal of Structural Engineering*, 115(3):626–646, 1989.

Jeff Anderson-Lee, Eli Fisker, Vineet Kosaraju, Michelle Wu, Justin Kong, Jeehyung Lee, Minjae Lee, Mathew Zada, Adrien Treuille, Rhiju Das, et al. Principles for predicting RNA secondary structure design difficulty. *Journal of molecular biology*, 428(5):748–757, 2016.

Fahiem Bacchus, Matti Järvisalo, and Ruben Martins. Maxsat evaluation 2018: New developments and detailed results. *Journal on Satisfiability, Boolean Modelling and Computation*, 11(1):99–131, 2019.

Zhidong Bai, Jack W Silverstein, et al. *Spectral analysis of large dimensional random matrices*, volume 20. Springer, 2010.

Jinho Baik, Gérard Ben Arous, and Sandrine Péché. Phase transition of the largest eigenvalue for nonnull complex sample covariance matrices. *The Annals of Probability*, 33(5):1643 – 1697, 2005.

Maximilian Balandat, Brian Karrer, Daniel R. Jiang, Samuel Daulton, Benjamin Letham, Andrew Gordon Wilson, and Eytan Bakshy. BoTorch: A Framework for Efficient Monte-Carlo Bayesian Optimization. In *Advances in Neural Information Processing Systems*, volume 33, pp. 21524–21538, 2020.

John E Beasley. Or-library: distributing test problems by electronic mail. *Journal of the operational research society*, pp. 1069–1072, 1990.

Florent Benaych-Georges and Raj Rao Nadakuditi. The eigenvalues and eigenvectors of finite, low rank perturbations of large random matrices. *Advances in Mathematics*, 227(1):494–521, 2011.

James Bergstra and Yoshua Bengio. Random search for hyper-parameter optimization. *Journal of Machine Learning Research*, 13(10):281–305, 2012.

Mickael Binois and Nathan Wycoff. A survey on high-dimensional gaussian process modeling with application to bayesian optimization. *ACM Transactions on Evolutionary Learning and Optimization*, 2(2):1–26, 2022.

Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. In *2017 IEEE symposium on security and privacy (sp)*, pp. 39–57, 2017.

T-J Chang, Nigel Meade, John E Beasley, and Yazid M Sharaiha. Heuristics for cardinality constrained portfolio optimisation. *Computers & Operations Research*, 27(13):1271–1302, 2000.

Sayak Ray Chowdhury and Aditya Gopalan. On kernelized multi-armed bandits. In *International Conference on Machine Learning*, pp. 844–853. PMLR, 2017.

Paul G Constantine. *Active subspaces: Emerging ideas for dimension reduction in parameter studies*. SIAM, 2015.

Paul G. Constantine, Eric Dow, and Qiqi Wang. Active subspace methods in theory and practice: Applications to kriging surfaces. *SIAM Journal on Scientific Computing*, 36(4):A1500–A1524, 2014.

Jhouben Cuesta Ramirez, Rodolphe Le Riche, Olivier Roustant, Guillaume Perrin, Cedric Durantin, and Alain Gliere. A comparison of mixed-variables bayesian optimization approaches. *Advanced Modeling and Simulation in Engineering Sciences*, 9(1):6, 2022.

Sam Daulton, Xingchen Wan, David Eriksson, Maximilian Balandat, Michael A Osborne, and Eytan Bakshy. Bayesian optimization over discrete and mixed spaces via probabilistic reparameterization. In *Advances in Neural Information Processing Systems*, 2022.

Chandler Davis and W. M. Kahan. The rotation of eigenvectors by a perturbation. iii. *SIAM Journal on Numerical Analysis*, 7(1):1–46, 1970.

Aryan Deshwal, Sebastian Ament, Maximilian Balandat, Eytan Bakshy, Janardhan Rao Doppa, and David Eriksson. Bayesian optimization over high-dimensional combinatorial spaces via dictionary-based embeddings. In *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*, volume 206 of *Proceedings of Machine Learning Research*, pp. 7021–7039. PMLR, 2023.

Kamil Dreczkowski, Antoine Grosnit, and Haitham Bou Ammar. Framework and benchmarks for combinatorial and mixed-variable bayesian optimization. In *Advances in Neural Information Processing Systems*, volume 36, pp. 69464–69489, 2023.

Ryan M Dreifuerst, Samuel Daulton, Yuchen Qian, Paul Varkey, Maximilian Balandat, Sanjay Kasturia, Anoop Tomar, Ali Yazdan, Vish Ponnampalam, and Robert W Heath. Optimizing coverage and capacity in cellular networks using machine learning. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 8138–8142. IEEE, 2021.

David Eriksson and Martin Jankowiak. High-dimensional Bayesian optimization with sparse axis-aligned subspaces. In *Proceedings of the Thirty-Seventh Conference on Uncertainty in Artificial Intelligence*, volume 161 of *Proceedings of Machine Learning Research*, pp. 493–503. PMLR, 2021.

David Eriksson, Michael Pearce, Jacob R Gardner, Ryan Turner, and Matthias Poloczek. Scalable global optimization via local Bayesian optimization. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, 2019.

Amir Hossein Gandomi, Xin-She Yang, and Amir Hossein Alavi. Mixed variable structural optimization using firefly algorithm. *Computers & Structures*, 89(23-24):2325–2336, 2011.

Léo Grinsztajn, Klemens Flöge, Oscar Key, Felix Birkel, Philipp Jund, Brendan Roof, Mihir Manium, Shi Bin Hoo, Magnus Bühler, Anurag Garg, Dominik Safaric, Jake Robertson, Benjamin Jäger, Simone Alessi, Adrian Hayler, Vladyslav Moroshan, Lennart Purucker, Philipp Singer, Alan Arazi, Julien Siems, Jan Hendrik Metzen, Georg Grab, Nick Erickson, Siyuan Guo, Eliott Kalfon, Simon Bing, David Salinas, Clara Cornu, Lilly Charlotte Wehrhahn, Diana Kriuchkova, Kursat Kaya, Lydia Sidhoum, Marie Salmon, Jerry Chen, Madelon Hulsebos, Yann LeCun, Samuel Müller, Bernhard Schölkopf, Sauraj Gambhir, Noah Hollmann, and Frank Hutter. Tabpfn-3: Technical report, 2026.

Antoine Grosnit, Cedric Malherbe, Rasul Tutunov, Xingchen Wan, Jun Wang, and Haitham Bou Ammar. Boils: Bayesian optimisation for logic synthesis. In *2022 Design, Automation & Test in Europe Conference & Exhibition (DATE)*, pp. 1193–1196. IEEE, 2022.

L Gu, RJ Yang, Cheng-Ho Tho, M Makowskit, O Faruquet, and Y Li Y. Li. Optimisation and robustness for crashworthiness of side impact. *International journal of vehicle design*, 26(4): 348–360, 2001.

Randy L Haupt. Thinned arrays using genetic algorithms. *IEEE transactions on antennas and propagation*, 42(7):993–999, 1994.

Noah Hollmann, Samuel Müller, Katharina Eggensperger, and Frank Hutter. TabPFN: A transformer that solves small tabular classification problems in a second. In *The Eleventh International Conference on Learning Representations*, 2023.

Noah Hollmann, Samuel Müller, Lennart Purucker, Arjun Krishnakumar, Max Körfer, Shi Bin Hoo, Robin Tibor Schirrmeister, and Frank Hutter. Accurate predictions on small data with a tabular foundation model. *Nature*, 637(8045):319–326, 2025.

Yingjie Hu, JianQiang Hu, Yifan Xu, Fengchun Wang, and Rong Zeng Cao. Contamination control in food supply chain. In *Proceedings of the 2010 Winter Simulation Conference*, pp. 2678–2681. IEEE, 2010.

Donald R Jones, Matthias Schonlau, and William J Welch. Efficient global optimization of expensive black-box functions. *Journal of Global optimization*, 13(4):455–492, 1998.

Ben Letham, Roberto Calandra, Akshara Rai, and Eytan Bakshy. Re-examining linear embeddings for high-dimensional bayesian optimization. In *Advances in Neural Information Processing Systems*, volume 33, pp. 1546–1558, 2020.

Wuchen Li and Guido Montúfar. Natural gradient via optimal transport. *Information Geometry*, 1 (2):181–214, 2018.

Marius Lindauer, Katharina Eggensperger, Matthias Feurer, AndrÃ© Biedenkapp, Difan Deng, Carolin Benjamins, Tim Ruhkopf, RenÃ© Sass, and Frank Hutter. SMAC3: A versatile Bayesian optimization package for hyperparameter optimization. *Journal of Machine Learning Research*, 23(54):1–9, 2022.

Ronny Lorenz, Stephan H Bernhart, Christian Höner zu Siederdissen, Hakim Tafer, Christoph Flamm, Peter F Stadler, and Ivo L Hofacker. Viennarna package 2.0. *Algorithms for molecular biology*, 6(1):26, 2011.

Vladimir A Marčenko and Leonid Andreevich Pastur. Distribution of eigenvalues for some sets of random matrices. *Mathematics of the USSR-Sbornik*, 1(4):457–483, 1967.

Shakir Mohamed, Mihaela Rosca, Michael Figurnov, and Andriy Mnih. Monte Carlo gradient estimation in machine learning. *Journal of Machine Learning Research*, 21(132):1–62, 2020. URL `http://jmlr.org/papers/v21/19-346.html`.

Samuel Müller, Noah Hollmann, Sebastian Pineda Arango, Josif Grabocka, and Frank Hutter. Transformers can do bayesian inference. In *International Conference on Learning Representations*, 2022.

Samuel Müller, Matthias Feurer, Noah Hollmann, and Frank Hutter. PFNs4BO: In-context learning for Bayesian optimization. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 25444–25470. PMLR, 23–29 Jul 2023.

Amin Nayebi, Alexander Munteanu, and Matthias Poloczek. A framework for Bayesian optimization in embedded subspaces. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 4752–4761. PMLR, 2019.

Ryan O'Donnell. *Analysis of boolean functions*, volume 2. Cambridge University Press Cambridge, 2014.

Changyong Oh, Jakub Tomczak, Efstratios Gavves, and Max Welling. Combinatorial Bayesian optimization using the graph Cartesian product. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Leonard Papenmeier, Luigi Nardi, and Matthias Poloczek. Increasing the scope as you learn: Adaptive bayesian optimization in nested subspaces. In *Advances in Neural Information Processing Systems*, volume 35, pp. 11586–11601, 2022.

Leonard Papenmeier, Luigi Nardi, and Matthias Poloczek. Bounce: Reliable high-dimensional bayesian optimization for combinatorial and mixed spaces. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

Debashis Paul. Asymptotics of sample eigenstructure for a large dimensional spiked covariance model. *Statistica Sinica*, pp. 1617–1642, 2007.

Florian Pfisterer, Lennart Schneider, Julia Moosbauer, Martin Binder, and Bernd Bischl. Yahpo gym - an efficient multi-objective multi-fidelity benchmark for hyperparameter optimization. In *Proceedings of the First International Conference on Automated Machine Learning*, volume 188 of *Proceedings of Machine Learning Research*, pp. 3/1–39. PMLR, 2022.

Sebastian Pineda-Arango, Hadi S. Jomaa, Martin Wistuba, and Josif Grabocka. HPO-B: A large-scale reproducible benchmark for black-box HPO based on openml. *Neural Information Processing Systems (NeurIPS) Track on Datasets and Benchmarks*, 2021.

Jingang QU, David Holzmüller, Gaël Varoquaux, and Marine Le Morvan. TabICLv2: A better, faster, scalable, and open tabular foundation model. In *Forty-third International Conference on Machine Learning*, 2026.

Lyle Regenwetter, Yazan Abu Obaideh, Fabien M. C. Chiotti, Ioanna Lykourentzou, and Faez Ahmed. Bikebench: A bicycle design benchmark for generative models with objectives and constraints. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2026.

Olivier Roustant, Esperan Padonou, Yves Deville, Aloïs Clément, Guillaume Perrin, Jean Giorla, and Henry Wynn. Group kernels for gaussian process metamodels with categorical inputs. *SIAM/ASA Journal on Uncertainty Quantification*, 8(2):775–806, 2020.

Binxin Ru, Ahsan Alvi, Vu Nguyen, Michael A. Osborne, and Stephen Roberts. Bayesian optimisation over multiple continuous and categorical inputs. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 8276–8285. PMLR, 2020a.

Binxin Ru, Adam Cobb, Arno Blaas, and Yarin Gal. BayesOpt adversarial attack. In *International Conference on Learning Representations*, 2020b.

Bobak Shahriari, Kevin Swersky, Ziyu Wang, Ryan P Adams, and Nando De Freitas. Taking the human out of the loop: A review of bayesian optimization. *Proceedings of the IEEE*, 104(1): 148–175, 2015.

Ruslan Shaydulin, Changhao Li, Shouvanik Chakrabarti, Matthew DeCross, Dylan Herman, Niraj Kumar, Jeffrey Larson, Danylo Lykov, Pierre Minssen, Yue Sun, Yuri Alexeev, Joan M. Dreiling, John P. Gaebler, Thomas M. Gatterman, Justin A. Gerber, Kevin Gilmore, Dan Gresh, Nathan Hewitt, Chandler V. Horst, Shaohan Hu, Jacob Johansen, Mitchell Matheny, Tanner Mengle, Michael Mills, Steven A. Moses, Brian Neyenhuis, Peter Siegfried, Romina Yalovetzky, and Marco Pistoia. Evidence of scaling advantage for the quantum approximate optimization algorithm on a classically intractable problem. *Science Advances*, 10(22):eadm6761, 2024.

Jasper Snoek, Hugo Larochelle, and Ryan Adams. Practical bayesian optimization of machine learning algorithms. *Advances in neural information processing systems*, 25, 2012.

Niranjan Srinivas, Andreas Krause, Sham Kakade, and Matthias Seeger. Gaussian process optimization in the bandit setting: No regret and experimental design. In *Proceedings of the 27th International Conference on International Conference on Machine Learning*, pp. 1015–1022, 2010.

Joe Staines and David Barber. Variational optimization, 2012.

Leon Thurner, Alexander Scheidler, Florian Schäfer, Jan-Hendrik Menke, Julian Dollichon, Friederike Meier, Steffen Meinecke, and Martin Braun. pandapower—an open-source python tool for convenient modeling, analysis, and optimization of electric power systems. *IEEE Transactions on Power Systems*, 33(6):6510–6521, 2018.

Joel A. Tropp. *An Introduction to Matrix Concentration Inequalities*, volume 8. Now Publishers Inc., 2015.

Xingchen Wan, Vu Nguyen, Huong Ha, Binxin Ru, Cong Lu, and Michael A. Osborne. Think global and act local: Bayesian optimisation over high-dimensional categorical and mixed search spaces. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 10663–10674. PMLR, 2021.

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.

Ronald J Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*, 8(3):229–256, 1992.

Rosen Yu, Cyril Picard, and Faez Ahmed. Fast and accurate bayesian optimization with pre-trained transformers for constrained engineering problems. *Structural and Multidisciplinary Optimization*, 68(3):66, 2025.

Rosen Ting-Ying Yu, Christophe Hatterer, Advaith Narayanan, Cyril Picard, and Faez Ahmed. Bocode: Engineering-centered benchmarking for bayesian optimization, 2026a.

Rosen Ting-Ying Yu, Cyril Picard, and Faez Ahmed. GIT-BO: High-dimensional bayesian optimization with tabular foundation models. In *The Fourteenth International Conference on Learning Representations*, 2026b.

Yi Yu, Tengyao Wang, and Richard J Samworth. A useful variant of the davis–kahan theorem for statisticians. *Biometrika*, 102(2):315–323, 2015.

Gan Yuan, Mingyue Xu, Samory Kpotufe, and Daniel Hsu. Efficient estimation of the central mean subspace via smoothed gradient outer products. *SIAM Journal on Mathematics of Data Science*, 7(3):1241–1264, 2025.

Olivier Zahm, Tiangang Cui, Kody Law, Alessio Spantini, and Youssef Marzouk. Certified dimension reduction in nonlinear Bayesian inverse problems. *Mathematics of Computation*, 91(336): 1789–1835, 2022.

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.

Ray Daniel Zimmerman, Carlos Edmundo Murillo-Sánchez, and Robert John Thomas. Matpower: Steady-state operations, planning, and analysis tools for power systems research and education. *IEEE Transactions on power systems*, 26(1):12–19, 2010.

TABLE OF CONTENTS FOR APPENDICES

A AI Use, Ethics, and Reproducibility Statements **15**

B Ablation Studies **15**

B.1 Why TabPFN v3? — Ablation on EDGE's surrogate model 15

B.2 Why select the rank adaptively? — Ablate against fixed rank value 15

B.3 Why UCB? — Ablation on the acquisition family 15

B.4 How sensitive is EDGE to the fallback dimension? — Ablation on fallback dimension 16

B.5 Does transport-aware refinement improve EDGE? — OTPR vs. PR 17

B.6 Why 25% Sobol samples — Ablation on balancing directional search and full-dimensional exploration 17

C Performance and Runtime Results Analysis **18**

C.1 Full results on the 32-problem suite 18

C.2 Runtime on fixed hardware 18

D Theoretical Analysis **20**

D.1 Preliminaries 20

D.2 Assumptions 20

D.3 Confidence Bounds for TabPFN-based Surrogates 21

D.4 Subspace Detection and Alignment Analysis 22

D.5 Optimal-transport probabilistic reparameterization and acquisition optimization . . 24

D.6 Main Regret Bounds 27

D.7 Information Gain Bounds for High-Dimensional Subspaces 28

D.8 Convergence Rate 28

E Structure, representation, and learning in combinatorial search **29**

E.1 Benchmark patterns and encoding 29

E.2 Objective structure and surrogate information 29

E.3 Learning and using variable importance 30

E.4 Implications for mixed-variable search 30

E.5 Containment guarantees for typed embeddings 31

F Baseline Algorithms Implementation Details **34**

G Benchmark Problems Implementation Details **35**

H EDGE Algorithm Implementation Details **38**

A AI USE, ETHICS, AND REPRODUCIBILITY STATEMENTS

AI use statement AI and large language models (LLMs) were used in this work as an editing assistant for the context, math proofs, and plot editing as well as to help orchestrate the execution of experiments (job scheduling and monitoring) under the authors' instruction. All algorithms, proofs, and experimental designs are the authors', and all reported numbers come from the authors' code. No LLM-generated result is reported.

Ethics statement We do not foresee any immediate negative ethical or societal consequences beyond the environmental impact of running inference on foundation models and standard considerations in automating engineering workflows.

Reproducibility statement Benchmark definitions, baseline implementations, and EDGE implementation details are provided in Appendices G, F, and H, respectively. Code and data will be released upon acceptance.

B ABLATION STUDIES

B.1 WHY TABPFN V3? — ABLATION ON EDGE'S SURROGATE MODEL

EDGE can use different surrogate families while retaining the same detection and refinement machinery. Figure 5 compares TabPFN(v3) (Grinsztajn et al., 2026), TabICL(v2) (QU et al., 2026), and SAAS-GP (Eriksson & Jankowiak, 2021) on five problems. Their strengths are complementary: TabPFN leads on `antenna200`, TabICL converges fastest on Ackley and cluster expansion, and SAAS-GP leads on `orpd118` and portfolio optimization. The comparison illustrates EDGE's flexibility: the surrogate supplies the predictive structure, while EDGE provides a common way to detect and search promising directions.

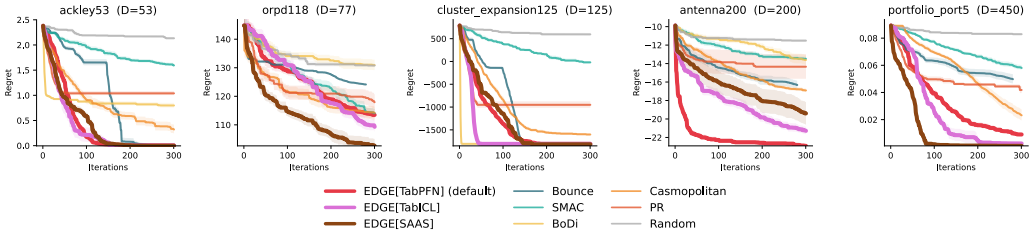

Figure 5: Surrogate backend comparison on five problems (53–450 dimensions). EDGE uses TabPFN, TabICL, or SAAS-GP; thin curves show baselines. Lower is better; shaded bands show variation across seeds.

B.2 WHY SELECT THE RANK ADAPTIVELY? — ABLATE AGAINST FIXED RANK VALUE

Figure 6 compares EDGE's detected-rank configuration with fixed-$r = 10$ variants. The comparison using TabPFN(v3) (Grinsztajn et al., 2026) in both arms shows better final average rank and normalized regret with detected rank, supporting adaptive direction selection on this ablation set. The SAAS-GP (Eriksson & Jankowiak, 2021) and TabICL(v2) (QU et al., 2026) curves additionally vary the surrogate and therefore measure the performance of alternative configurations rather than the isolated effect of rank. A fixed rank of ten is also used in GIT-BO's main experiments (Yu et al., 2026b).

B.3 WHY UCB? — ABLATION ON THE ACQUISITION FAMILY

UCB balances predicted value with an explicit uncertainty bonus, while LogEI and NoisyLogEI (Ament et al., 2023) target expected improvement over the incumbent. In Figure 7, the default UCB

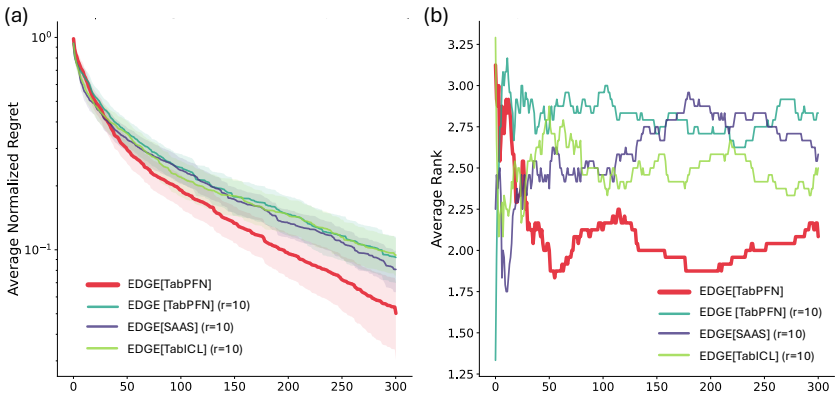

Figure 6: EDGE's default detected-rank (red) compared with the plotted fixed-rank variants ($r = 10$): (a) average normalized regret on a logarithmic scale and (b) average rank, against BO iterations. Lower is better. The default configuration reaches lower final regret and a better average rank. Shaded bands unit standard deviation.

configuration continues improving after the EI configurations slow down, reaching lower final regret on all eight panels. This motivates retaining explicit exploration in EDGE; isolating its contribution requires a comparison with the surrogate held fixed.

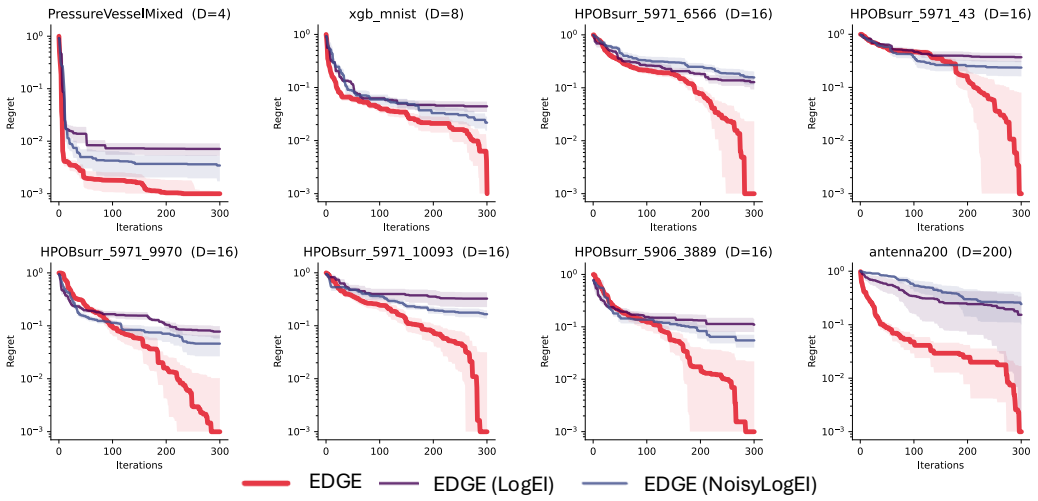

Figure 7: Acquisition configurations on eight problems (4–200 dimensions): default UCB, LogEI, and NoisyLogEI (Ament et al., 2023). Normalized regret is shown on a log scale; lower is better. Shaded bands show variation across seeds.

B.4 How sensitive is EDGE to the fallback dimension? — Ablation on fallback dimension

Table 1 shows that fallback dimension can affect performance: increasing it from $d/5$ to $d/4$ improves the mean final objective on seven of ten problems. No single setting dominates: $d/2$ achieves the highest mean on four problems, $d/4$ on three, and the default on three. These results suggest room for improvement through task-adaptive fallback dimensions, while our fixed-default results do not exhaust the performance attainable within the EDGE framework. The comparisons are descriptive and do not establish statistical significance.

Table 1: Fallback-dimension sensitivity on ten ablation tasks: default EDGE ($d/5$), $d/2$, and $d/4$. Larger mean final objective values are better. Bold marks the largest mean in each row.

	EDGE	fallback dim d/2	fallback dim d/4
HPOBsurr_5971_43	$\mathbf{0.9675 \pm 0.00417}$	0.9661 ± 0.00468	0.9648 ± 0.0038
HPOBsurr_6767_146064	1.057 ± 0.174	0.9369 ± 0.299	$\mathbf{1.091 \pm 0.0218}$
aig_cavlc	-1.793 ± 0.0779	$\mathbf{-1.706 \pm 0.029}$	-1.742 ± 0.0914
aig_priority	-0.9638 ± 0.0249	-0.9874 ± 0.00549	$\mathbf{-0.9541 \pm 0.0366}$
contamination25	-21.56 ± 0.181	$\mathbf{-21.48 \pm 0.0808}$	-21.5 ± 0.0786
labs	2.761 ± 0.434	$\mathbf{3.575 \pm 0.194}$	3.29 ± 0.417
labsf	3.083 ± 0.28	$\mathbf{3.672 \pm 1.1}$	3.488 ± 0.819
maxsat60	177.4 ± 10.4	173.3 ± 2.24	$\mathbf{179.5 \pm 9.03}$
pest25	$\mathbf{-12.2 \pm 0.212}$	-12.57 ± 0	-12.47 ± 0.224
svm	$\mathbf{-0.1673 \pm 0.00333}$	-0.1738 ± 0.00848	-0.1774 ± 0.00892

B.5 Does transport-aware refinement improve EDGE? — OTPR vs. PR

We compare EDGE's OTPR refinement with Euclidean PR (Daulton et al., 2022), holding the remaining pipeline fixed. OTPR combines ground-cost smoothing with a graph-transport natural gradient, so its updates account for the geometry of probability redistribution (Section 3.3). Figure 8 shows better final objective values with OTPR on all five displayed tasks. This supports the combined refinement design on these tasks; the comparison does not separate the contributions of smoothing and metric preconditioning. Appendix D.5 analyzes their distinct mathematical roles.

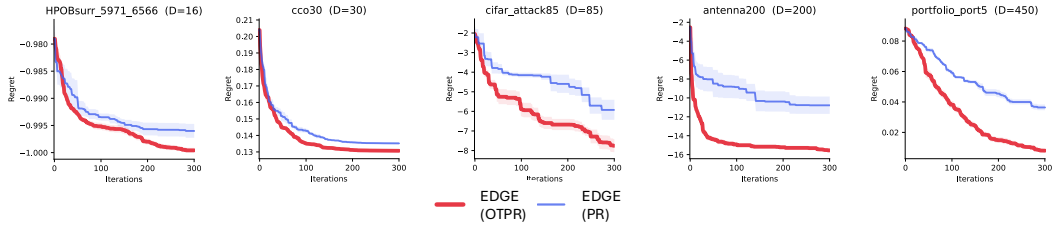

Figure 8: Discrete refinement within EDGE: OTPR versus Euclidean PR (Daulton et al., 2022), with the remaining pipeline fixed, on five problems (16–450 dimensions). Lower is better; shaded bands show variation across seeds.

B.6 Why 25% Sobol samples — Ablation on balancing directional search and full-dimensional exploration

We vary the fraction of full-dimensional Sobol candidates from the default 25% to 0%, 12.5%, 50%, and 75% (Figure 9). The default achieves the lowest final average normalized regret, approximately 0.015 compared with 0.04 for the next-best setting, 75%. Removing full-dimensional candidates increases aggregate regret, while allocating a larger fraction to them also reduces performance. The two displayed HPO tasks illustrate the benefit of the default mixture: EDGE continues improving after the alternative settings largely plateau. On cluster expansion, all settings approach a similar plotted regret floor, with differences primarily in convergence speed. These results support retaining both directional and full-dimensional proposals, with 25% providing the strongest aggregate performance among the tested settings.

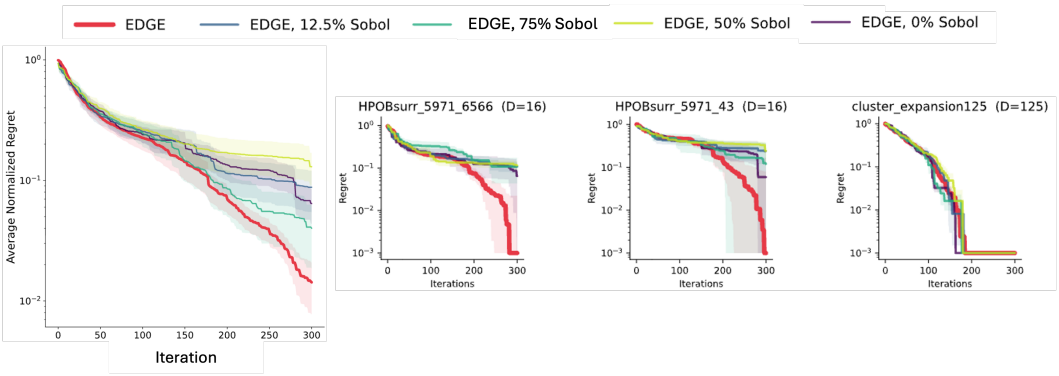

Figure 9: Ablation of the full-dimensional Sobol candidate fraction. Default EDGE reserves 25% of its candidate pool for full-dimensional proposals; alternatives use 0%, 12.5%, 50%, or 75%. Left: average normalized regret across the ablation suite. Right: individual trajectories on two HPO tasks and cluster expansion. Lower is better; vertical axes use logarithmic scales. The default mixture achieves the lowest final aggregate regret, while the individual trajectories illustrate variation across tasks.

C Performance and Runtime Results Analysis

C.1 Full results on the 32-problem suite

Figure 10 complements the aggregate results in Section 5 with per-problem convergence curves. Table 2 reports EDGE's strongest performance on the 25 real-world tasks, with an average rank of 1.77 versus 3.32 for Bounce. On the nine HPO tasks, EDGE achieves rank 1.44 versus 3.22 for Bounce, with a Holm-adjusted Wilcoxon $p = 0.027$. The curves further illustrate sustained improvements on antenna design ($D = 200$) and portfolio optimization ($D = 450$). Together, these results support EDGE's value for evaluation-budget-limited optimization across real-world tasks.

Table 2: Average rank (lower is better) and Holm-corrected Wilcoxon p against EDGE with EDGE on the 32-problem suite. Real-world and synthetic tasks partition the suite; applied science & engineering and HPO partition the real-world subset. Lower ranks are better. Small values beside baseline ranks are two-sided Wilcoxon signed-rank p-values computed from per-problem mean final incumbents, with Holm correction across baseline comparisons within each group. Bold marks the best displayed average rank in each row.

Group	n	EDGE	Bounce	SMAC	BoDi	Casmopolitan	PR	Random
All 32 problems *(32)*	**2.30**	2.87 0.839	3.67 0.007	4.00 0.069	4.23 0.017	4.40 0.001	6.53 <0.001	
Real-world *(25)*	**1.78**	3.26 0.052	3.13 0.010	4.30 0.004	4.39 0.012	4.65 <0.001	6.48 <0.001	
Synthetic *(7)*	4.00	**1.57** 0.094	5.43 0.438	3.00 0.891	3.71 0.891	3.57 0.891	6.71 0.156	
Applied science & engineering *(16)*	**2.00**	3.29 0.426	3.79 0.043	4.43 0.074	3.14 0.238	4.57 0.002	6.79 <0.001	
HPO *(9)*	**1.44**	3.22 0.023	2.11 0.359	4.11 0.023	6.33 0.023	4.78 0.023	6.00 0.023	

C.2 Runtime on fixed hardware

Table 3 reports EDGE's computational cost in the runtime study. EDGE averages 26.20 s per iteration, approximately 6.7× Bounce's 3.92 s, while using 30.8% less time than BODi's 37.84 s. EDGE's gains under an evaluation budget therefore come with additional search overhead relative to Bounce. Whether this additional computation is worthwhile depends on objective-evaluation cost and the resulting solution quality.

Table 3: Per-iteration runtime cost of every method. The table is order from the fastest method (top) to the slowest method (bottom).

Method	n	mean s/iter	median	std
Random	3	0.04	0.04	0.01
SMAC	3	1.50	1.44	0.15
Casmopolitan	3	1.63	1.66	0.10
PR	3	3.01	1.82	2.44
Bounce	3	3.92	4.02	0.28
EDGE	3	26.20	27.05	4.61
BoDi	3	37.84	37.92	1.17

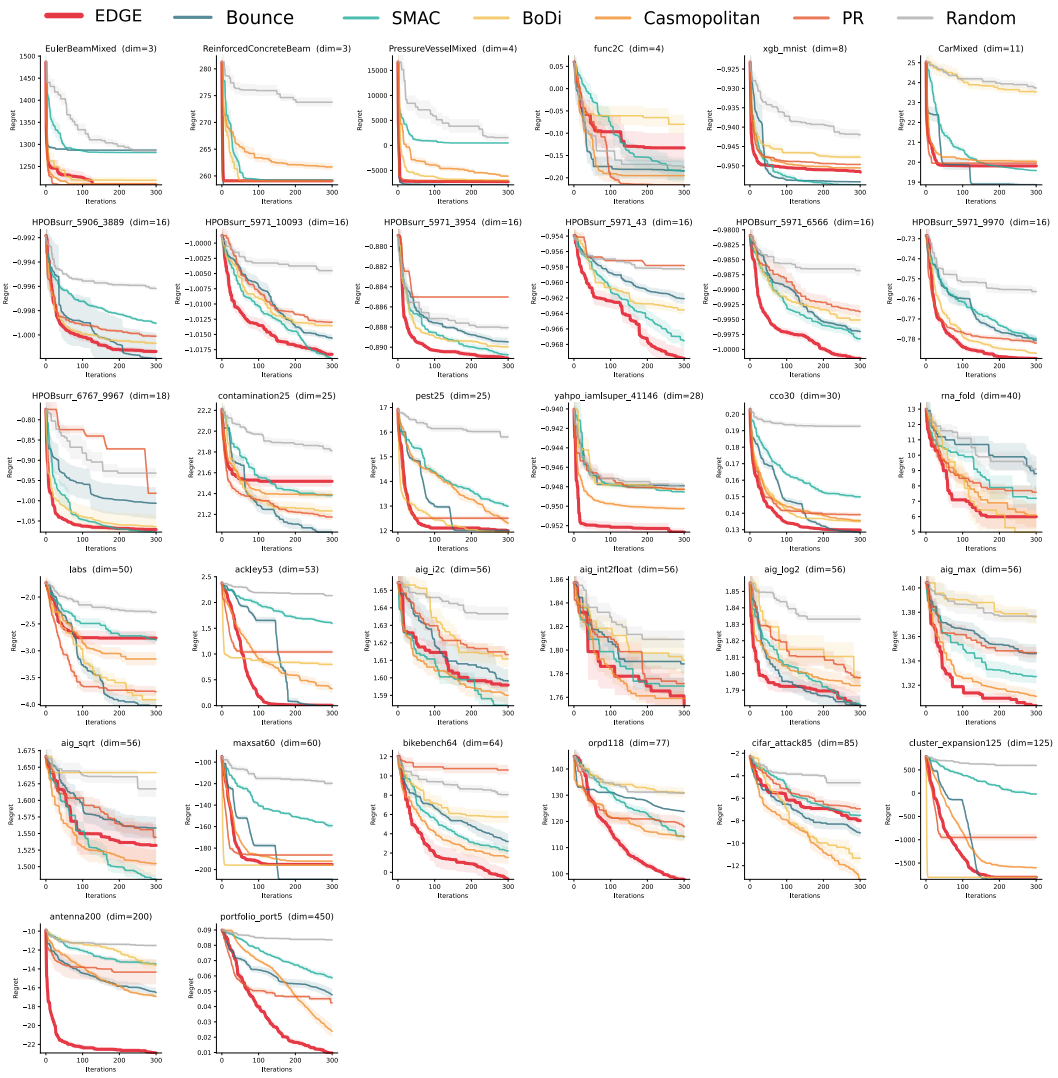

Figure 10: Average regret vs. iterations (# function evaluations) with a budget of 300 iterations for all benchmarks. Average regrets are illustrated by solid lines, with shaded bands denoting unit standard deviation. EDGE finds the optimal value for 18 out of the total 32 problems.

D THEORETICAL ANALYSIS

The analysis provides conditional foundations for three distinct parts of EDGE. A Gaussian spiked reference model motivates spectral selection, and a separate finite-sample perturbation theorem relates matrix error to subspace error under independent bounded probes. The acquisition analysis studies graph-transport flow and Gibbs smoothing at fixed continuous coordinates. The outer-loop analysis transfers a reference confidence event to the surrogate and retains explicit coverage, decoding, and optimization errors. We do not prove that the implementation satisfies all of these hypotheses or that its finite search errors vanish.

D.1 PRELIMINARIES

Problem setup. Let $\mathcal{X} = \mathcal{X}_c \times \prod_{j=1}^{d_z} \mathcal{Z}_j \subseteq [0,1]^d$ be the feasible mixed domain, where $\mathcal{X}_c \subseteq [0,1]^{d_c}$ and each $\mathcal{Z}_j$ is a finite normalized level set. The full cube is a continuous relaxation used for differentiation; objective queries belong to $\mathcal{X}$. We seek $x^\star \in \arg\max_{x\in\mathcal{X}} f(x)$ and observe $y_t = f(x_t)+\epsilon_t$, where x_t is selected using past observations and ϵ_t is conditionally R-sub-Gaussian. Predictions indexed by t are available before observing y_t. For clarity the theoretical index counts observations from the first query, including any initialization.

TabPFN surrogate properties. Following Yu et al. (2026b), let $q_\theta(y \mid x, \mathcal{D}_t)$ denote the surrogate's posterior predictive distribution given the observed data $\mathcal{D}_t$, with predictive mean $\mu_t(x)$ and variance $\sigma_t^2(x)$. The network weights are frozen. Conditioning happens in context, and μ_t, σ_t, and the acquisition built from them are differentiable in the query x by automatic differentiation.

Probabilistic objective. For the discrete variables, PR (Daulton et al., 2022) replaces z by a distribution $p(Z \mid \theta)$ and maximizes $\mathbb{E}_{Z\sim p(\cdot|\theta)}[\alpha(x,Z)]$, which is differentiable in (x,θ) and whose maximizers correspond to maximizers of α (their Theorem 1) when α is continuous in x for every z.

Reference GP class and information gain. For a fixed kernel k and regularization $\sigma^2_{\mathrm{ref}} > 0$, define

$$\gamma_{T,\mathrm{full}} = \sup_{u_1,\dots,u_T\in\mathcal{X}} \frac{1}{2}\log\det\!\left(I + \sigma_{\mathrm{ref}}^{-2}K_T\right), \qquad (K_T)_{ij} = k(u_i,u_j).$$

The supremum permits repeated queries. This information-gain quantity controls the reference posterior variance sum in GP-UCB analyses (Srinivas et al., 2010; Chowdhury & Gopalan, 2017).

D.2 ASSUMPTIONS

Following the reference-GP approximation setup of Yu et al. (2026b, Appendix A.2), we state the required surrogate approximation as an explicit event along the sequential observation history. Lemma 1 retains its contribution as an additive confidence error.

Assumption 1 (Surrogate approximation event)**.** For a fixed reference kernel and regularization parameter, let $\mu_t^{\mathrm{GP}}, \sigma_t^{\mathrm{GP}}$ denote reference predictions before query t. On an explicitly specified event, uniformly over $u \in \mathcal{X}$ and $t \geq 1$,

$$|\mu_t(u) - \mu_t^{\mathrm{GP}}(u)| \leq e_t, \qquad |\sigma_t^2(u) - (\sigma_t^{\mathrm{GP}}(u))^2| \leq e_t, \qquad e_t \geq 0.$$

We take $\sigma_t(u) \geq 0$. The bounds above are hypotheses on the sequential prediction problem, not consequences of finite empirical comparisons. No decay of e_t is assumed here; sufficient decay conditions are stated separately in Appendix D.8.

Assumption 2 (Reference function and observation model)**.** The objective belongs to the RKHS of a fixed kernel k, with $\|f\|_k \leq B$ and $k(u,u) \leq 1$. Query selection is predictable from the observation history and the noise is conditionally R-sub-Gaussian. The reference posterior uses fixed regularization $\sigma^2_{\mathrm{ref}} > 0$. A simultaneous reference confidence event must be obtained from a confidence construction valid for these choices.

Assumption 3 (Gaussian spectral reference)**.** For the asymptotic spectral result only, the gradients satisfy the Gaussian spiked model in Theorem 1. This is a reference model, not an asserted distributional property of EDGE's measured gradients. The bounded-probe perturbation result has its own separate hypotheses. Gaussian probes do not satisfy an almost-sure finite norm bound.

Assumption 4 (Graph metric and independent coordinates)**.** Each probability factor lies in the relative interior of its simplex. Its connected level graph has smooth positive interior conductances, with Laplacian $L(p)$. In independent coordinates θ, the Jacobian $J_\theta p$ has full column rank and

$$G_W(\theta) = (J_\theta p)^\top L(p)^\dagger J_\theta p$$

is positive definite. For a smoothed objective, p and the Jacobian refer to the composed smoothed distribution. The graph and conductances must be specified in addition to the ground-cost matrix.

Assumption 5 (Conditional flow hypothesis)**.** Fix the continuous block and the probability map being optimized. For $F(\theta) = -\mathbb{E}_{p(\cdot|\theta)}\alpha(x, Z)$ and $F^\star = \inf_{\theta\in\Theta} F(\theta)$, suppose the exact flow exists for all nonnegative times in its smooth domain and

$$\nabla F(\theta)^\top G_W(\theta)^{-1}\nabla F(\theta) \geq 2\lambda(F(\theta) - F^\star), \qquad \lambda > 0.$$

This condition is imposed only for the conditional contraction result. It is not a generic property of probabilistic acquisition objectives.

Objective structure is a separate hypothesis. An optional ridge representation $f(u) = h(U_*^\top u) + r(u)$ with $\sup_{u\in\mathcal{X}} |r(u)| \leq \varepsilon_{\text{ridge}}$ describes the objective. It does not identify U_* with eigenvectors of a surrogate acquisition second moment. The regret decomposition below does not require this representation, and no search coverage bound is deduced from it here.

D.3 Confidence Bounds for TabPFN-based Surrogates

Let β_t be nondecreasing and chosen so that the reference confidence event

$$\mathcal{E}_{\text{GP}} = \left\{|f(u) - \mu_t^{\text{GP}}(u)| \leq \sqrt{\beta_t}\,\sigma_t^{\text{GP}}(u) \text{ for all } u \in \mathcal{X},\ t \geq 1\right\} \tag{10}$$

has probability at least $1 - \delta_{\text{GP}}$. Under the fixed-kernel RKHS model, one conservative kernelized self-normalized confidence choice is

$$\sqrt{\beta_t} = B + \frac{R}{\sigma_{\text{ref}}}\sqrt{2\,(\gamma_{t-1,\text{full}} + 1 + \log(1/\delta_{\text{GP}}))}, \tag{11}$$

where the information gain uses the same regularization σ_{ref}^2 and permits repeated queries. This is the standard RKHS self-normalized construction after regularization scaling (Chowdhury & Gopalan, 2017). The fixed-kernel, noise, and measurability hypotheses are essential. An implementation using a different UCB coefficient does not inherit this confidence claim without separate justification.

Lemma 1 (Additive transfer of reference confidence)**.** *On $\mathcal{E}_{\text{GP}}$ and the approximation event of Assumption 1, set $b_t = e_t + \sqrt{\beta_t e_t}$. Then, simultaneously over t and u,*

$$|f(u) - \mu_t(u)| \leq \sqrt{\beta_t}\,\sigma_t(u) + b_t, \qquad |\alpha_t(u) - \alpha_t^{\text{GP}}(u)| \leq b_t, \tag{12}$$

where $\alpha_t = \mu_t + \sqrt{\beta_t}\sigma_t$ and $\alpha_t^{\text{GP}} = \mu_t^{\text{GP}} + \sqrt{\beta_t}\sigma_t^{\text{GP}}$.

Proof. For nonnegative variances, $|\sqrt{a} - \sqrt{b}| \leq \sqrt{|a - b|}$. Consequently $|\sigma_t - \sigma_t^{\text{GP}}| \leq \sqrt{e_t}$. The triangle inequality and the reference confidence event give

$$|f - \mu_t| \leq \sqrt{\beta_t}\sigma_t^{\text{GP}} + e_t \leq \sqrt{\beta_t}\sigma_t + e_t + \sqrt{\beta_t e_t}.$$

Applying the same discrepancy bounds to the two acquisitions proves the second assertion. □

If the approximation event has probability at least $1 - \delta_{\text{app}}$, the joint statement has probability at least $1 - \delta_{\text{GP}} - \delta_{\text{app}}$; no independence is needed. If its bounds hold deterministically, only δ_{GP} is used. An additive term is necessary: at a query with $\sigma_t = 0$, a permitted nonzero mean discrepancy cannot be absorbed into any finite multiplicative confidence parameter.

D.4 Subspace Detection and Alignment Analysis

The Gaussian reference model describes a phase transition in sample eigenstructure. It motivates a selection rule and overlap-inspired weights, but does not certify the implemented rank, noise-scale estimator, or BO proposal lengths. In particular, a Gaussian reference model and a separate bounded-probe concentration theorem are alternative analytical settings; we do not assume that Gaussian vectors have bounded norms almost surely.

The model restrictions matter beyond the covariance and independence across probes. For example, let $h_j \sim N(0, C)$ and $B_j \sim \mathrm{Bernoulli}(1/2)$ be independent, and set $g_j = \sqrt{2} B_j h_j$. These independent vectors have second moment C and, for bounded population eigenvalues, uniformly bounded coordinate fourth moments. Nevertheless,

$$\widehat{C} = \frac{2}{M} \sum_{j:B_j=1} h_j h_j^\top$$

uses asymptotically $M/2$ nonzero samples. Its isotropic noise edge is $\sigma_g^2(1 + \sqrt{2\gamma})^2$, not $\sigma_g^2(1 + \sqrt{\gamma})^2$. Thus the second-moment specification alone does not imply the Gaussian spectral limits.

The selected directions concern variation of the surrogate acquisition under the probe distribution. We next separate a population approximation result from a finite-sample perturbation bound. Neither result, by itself, bounds the loss incurred by EDGE's feasible candidate search.

A conditional population approximation bound. This construction is an analytical comparison, not EDGE's candidate-sampling law. Let μ be a reference probability measure on the continuous relaxed domain, set $Z_\alpha = \int e^{\alpha(u)}\, d\mu(u) \in (0, \infty)$, and define $dq/d\mu = e^\alpha / Z_\alpha$. For an orthogonal projector P, define the full-support ridge-density approximation

$$\frac{dq_P}{d\mu}(u) = \frac{\mathbb{E}_\mu[e^{\alpha(X)} \mid PX = Pu]}{Z_\alpha}.$$

Lemma 2 (Ridge-density approximation under conditional LSI)**.** *Suppose the conditional reference measures on fibers $PX = y$ satisfy a logarithmic Sobolev inequality with common constant κ, in the convention $\mathrm{Ent}(h^2) \le 2\kappa \mathbb{E}\|\nabla_\perp h\|^2$. Assume the required derivatives and integrals exist, and set $H = \mathbb{E}_q[\nabla\alpha \nabla\alpha^\top]$. Then*

$$\mathrm{KL}(q \| q_P) \le \frac{\kappa}{2} \operatorname{tr}((I - P)H). \tag{13}$$

If P projects onto the leading r eigenvectors of H, the right-hand side is $\frac{\kappa}{2} \sum_{i>r} \lambda_i(H)$.

Proof. The measures q and q_P have the same marginal law of PX, while the conditional law of q_P on each fiber equals that of μ. The chain rule for relative entropy therefore expresses $\mathrm{KL}(q\|q_P)$ as the q-marginal average of the conditional relative entropies. On a fiber the conditional density ratio is proportional to e^α, so the conditional logarithmic Sobolev inequality bounds this entropy by $\frac{\kappa}{2}\mathbb{E}_q[\|(I - P)\nabla\alpha\|^2 \mid PX]$. Integrating and using $P = P^\top = P^2$ yields the trace bound. The eigenvalue tail follows for the stated choice of P. □

This is the conditional/subspace log-Sobolev construction underlying Zahm et al. (2022) (Theorem 2.10 and Corollary 2.11). An ordinary global LSI alone does not supply the stated uniform fiber hypothesis. Also, q_P is not a distribution supported only on the affine search subspace; such a lower-dimensional support would generally give $\mathrm{KL}(q\|q_P) = \infty$. Finally, H is defined under q, whereas EDGE estimates C under its probe law. For an arbitrary detector projector the trace bound applies if its hypotheses hold, but the minimizing eigenvalue-tail expression cannot be assigned to that projector without relating the two matrices.

From matrix estimation to subspace alignment. The next step is to relate the error in $\widehat{C}$ to the error in its eigenvectors. The spectral gap matters because nearby population eigenvalues allow a larger rotation of the associated subspace under the same matrix perturbation.

Theorem 2 (Conditional subspace perturbation bound)**.** *Suppose the probe gradients are independent with common second moment C and satisfy $\|g_j\|^2 \le B_g$ almost surely for a deterministic B_g.*

For $1 \le r < d$ and $0 < \delta < 1$, let $\mathrm{gap}_r = \lambda_r(C) - \lambda_{r+1}(C) > 0$. With probability at least $1-\delta$, matrix Bernstein and a population-gap Davis–Kahan bound give, in operator norm,

$$\begin{aligned} \|\sin\Theta(\widehat{U}_r, U_r)\| &\le \frac{2\|\widehat{C} - C\|}{\mathrm{gap}_r}, \\ \|\widehat{C} - C\| &\lesssim \sqrt{\frac{B_g\|C\|\log(2d/\delta)}{M}} + \frac{B_g\log(2d/\delta)}{M}. \end{aligned} \tag{14}$$

Both terms must be controlled to obtain a prescribed subspace error.

Proof. All norms below are operator norms. We work at one BO iteration, conditional on the observed data and fitted surrogate; the independence and boundedness assumptions are required under this conditional distribution. Write $A_j = g_j g_j^\top$ and $X_j = A_j - C$, so that $\widehat{C} - C = M^{-1}\sum_{j=1}^M X_j$ and $\mathbb{E}X_j = 0$.

Step 1: bound the summands and their variance. Since $0 \preceq A_j \preceq B_g I$, taking expectations gives $0 \preceq C \preceq B_g I$. Therefore $-B_g I \preceq X_j \preceq B_g I$ and $\|X_j\| \le B_g$. Moreover, $A_j^2 = \|g_j\|^2 A_j \preceq B_g A_j$, and hence

$$0 \preceq \mathbb{E}X_j^2 = \mathbb{E}A_j^2 - C^2 \preceq \mathbb{E}A_j^2 \preceq B_g C.$$

Consequently, the matrix variance parameter satisfies $v := \|\sum_j \mathbb{E}X_j^2\| \le M B_g\|C\|$.

Step 2: control the empirical matrix. The two-sided self-adjoint matrix Bernstein inequality (Tropp, 2015), applied to the independent centered matrices X_j, gives, for $s > 0$ in the nondegenerate case,

$$\mathbb{P}\left(\left\|\sum_j X_j\right\| \ge s\right) \le 2d\exp\left(-\frac{s^2}{2(v + B_g s/3)}\right).$$

Set $t = \log(2d/\delta)$ and $s = \sqrt{2vt} + 2B_g t/3$. Direct expansion shows $s^2 \ge 2t(v + B_g s/3)$, so division by M yields, with probability at least $1-\delta$,

$$\|\widehat{C} - C\| \le \sqrt{\frac{2B_g\|C\|t}{M}} + \frac{2B_g t}{3M}. \tag{15}$$

If $B_g\|C\| = 0$, the positive-semidefinite matrices A_j are zero almost surely, and the matrix deviation statement is immediate. Equation 15 establishes the second line of equation 14 with explicit constants.

Step 3: transfer matrix error to subspace error. Let $E = \widehat{C} - C$, $e = \|E\|$, and let U_r and $\widehat{U}_r$ contain orthonormal leading r eigenvectors of C and $\widehat{C}$, respectively. Let $U_\perp$ contain the remaining population eigenvectors, with eigenvalue matrix $\Lambda_\perp$, and let $\widehat{\Lambda}_r$ contain the leading sample eigenvalues. The largest sine of the principal angles is $\|\sin\Theta(\widehat{U}_r, U_r)\| = \|U_\perp^\top \widehat{U}_r\|$. If $e \ge \mathrm{gap}_r/2$, the claimed bound follows from this norm being at most one. Otherwise, Weyl's eigenvalue inequality gives

$$\lambda_r(\widehat{C}) - \lambda_{r+1}(C) \ge \mathrm{gap}_r - e > \mathrm{gap}_r/2.$$

Multiplying $(C+E)\widehat{U}_r = \widehat{U}_r\widehat{\Lambda}_r$ on the left by $U_\perp^\top$ gives the Sylvester equation

$$Y\widehat{\Lambda}_r - \Lambda_\perp Y = R, \qquad Y = U_\perp^\top \widehat{U}_r, \quad R = U_\perp^\top E\widehat{U}_r.$$

Writing $a = \lambda_{r+1}(C)$, its solution is

$$Y = \int_0^\infty e^{u(\Lambda_\perp - aI)} R\, e^{-u(\widehat{\Lambda}_r - aI)}\, du.$$

Indeed, differentiating the integrand and integrating between zero and infinity verifies the equation; the positive spectral separation ensures convergence and uniqueness. Since the two exponential norms are bounded by 1 and $e^{-u(\lambda_r(\widehat{C}) - a)}$, respectively, we obtain

$$\|Y\| \le \frac{\|R\|}{\lambda_r(\widehat{C}) - a} \le \frac{e}{\mathrm{gap}_r - e} \le \frac{2e}{\mathrm{gap}_r}.$$

Substituting $Y = U_\perp^\top \widehat{U}_r$, $e = \|\widehat{C} - C\|$, and $t = \log(2d/\delta)$ back into the two bounds gives, on the same event of probability at least $1 - \delta$,

$$\|\sin\Theta(\widehat{U}_r, U_r)\| \le \frac{2\|\widehat{C} - C\|}{\mathrm{gap}_r},$$

$$\|\widehat{C} - C\| \le \sqrt{\frac{2B_g\|C\|\log(2d/\delta)}{M}} + \frac{2B_g\log(2d/\delta)}{3M} \lesssim \sqrt{\frac{B_g\|C\|\log(2d/\delta)}{M}} + \frac{B_g\log(2d/\delta)}{M}.$$

The first and last lines are exactly the two assertions of equation 14. Here $\lesssim$ hides only a universal multiplicative constant (one may use $\sqrt{2}$). This completes the Davis–Kahan-type argument (Davis & Kahan, 1970; Yu et al., 2015) and the proof of the stated inequalities.

For completeness, for $0 < \varepsilon < 1$, the sufficient probe budget

$$M \ge \max\left\{\frac{32B_g\|C\|t}{\varepsilon^2\mathrm{gap}_r^2}, \frac{8B_g t}{3\varepsilon\mathrm{gap}_r}\right\}$$

makes each term in equation 15 at most $\varepsilon\mathrm{gap}_r/4$, and therefore gives subspace error at most ε with probability at least $1 - \delta$. □

Relation to the implementation. The proof requires a deterministic almost-sure bound B_g, not the maximum of the observed gradient norms. It does not establish these assumptions for Sobol probes or transfer a smoothed-estimator rate merely from adding jitter (Yuan et al., 2025). Its role is to state the conditions under which probe accuracy implies subspace accuracy; calibration of the logged ratio is a separate question.

EDGE logs an estimated matrix-error-to-gap ratio. This is a stability diagnostic, not a certified upper bound on subspace error. A sample maximum gradient norm is not the deterministic almost-sure bound required by the perturbation theorem, and a sample eigengap is not automatically a lower bound on the population gap. Interpreting the diagnostic as a certificate would require valid uncertainty bounds for both quantities. Its relation to alignment and optimization is evaluated empirically.

Candidate generation when no direction is selected. The spectral rule also determines when EDGE uses its random-sketch branch. Its relation to random-embedding methods concerns the candidate mechanism.

Random-sketch branch. When $\widehat{r} = 0$, EDGE generates candidates using a random signed count-sketch of the trust region (Nayebi et al., 2019).

More dimensions than probes. For $\gamma = d/M > 1$, write $G = [g_1, \ldots, g_M]$ so that $\widehat{C} = GG^\top/M$. Its nonzero eigenvalues equal those of the companion matrix $G^\top G/M$; the remaining eigenvalues include at least $d - M$ zeros. The MP upper edge remains $\sigma^2(1 + \sqrt{\gamma})^2$ (Bai et al., 2010). This covers $d = 450$ with $M = 256$; structural zeros must be accounted for when estimating the noise scale. For $d > M$, estimating the scale must account for the structural zero mass; using only positive eigenvalues requires the corresponding normalized positive spectral law. At very small d, a fitted bulk is weakly determined. The fitting initialization, eigenvalue subset, stopping rule, and behavior when no usable bulk remains are part of the algorithm specification. Numerical failure to estimate a noise scale must be distinguished from a successful fit that selects no directions.

D.5 Optimal-transport probabilistic reparameterization and acquisition optimization

The analysis separates acquisition refinement from outer-loop progress. We first identify the geometry and approximation error of the probabilistic objective, then derive a regret bound that retains the errors of finite mixed-variable search.

Probabilistic reparameterization makes discrete acquisition optimization continuous. We analyze the metric used to update its probabilities and the approximation introduced by ground-cost smoothing.

Fix a BO iteration and the continuous block x. For finite discrete level sets, define

$$A(\theta) = \sum_z p(z \mid \theta)\alpha(x, z), \qquad F(\theta) = -A(\theta).$$

Differentiability in θ follows from differentiability of the probability map; differentiability in x additionally requires differentiability of $\alpha(\cdot, z)$. PR relates the global optimum of this expectation to the original acquisition when the probability family includes point masses, or approaches them in its closure (Daulton et al., 2022).

Geometry of the update. Choose a connected graph on each factor's levels and a smooth weighted Laplacian $L(p)$ with positive edge conductances for interior p. The graph, its ground-cost weights, and the interpolation defining its conductances are part of the metric specification. In independent coordinates, assume $J_\theta p$ has full column rank and set

$$G_W(\theta) = (J_\theta p)^\top L(p)^\dagger J_\theta p, \qquad \dot\theta = -G_W(\theta)^{-1}\nabla F(\theta). \tag{16}$$

This is the pullback construction for a graph transport metric (Li & Montúfar, 2018). Its local squared distance satisfies $d_{G_W}(\theta, \theta + v)^2 = v^\top G_W(\theta)v + o(\|v\|^2)$. Here d_{G_W} denotes the induced Riemannian distance, rather than the Kantorovich W_2 distance on a finite level set. Redundant softmax logits must first be reduced to independent coordinates.

For the two-level, unit-distance case, choosing logarithmic-mean conductance gives

$$\theta_{01}(p) = \frac{1-2p}{\log((1-p)/p)}, \quad \theta_{01}(1/2) = 1/2, \qquad L(p) = \theta_{01}(p)\begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix}, \quad G_p = \theta_{01}(p)^{-1}. \tag{17}$$

Indeed, the probability vector is $(1-p, p)$, and substitution of its Jacobian into equation 16 gives G_p. Consequently, $\theta_{01}(p)$ is the inverse metric in the probability coordinate p; changing coordinates also changes the preconditioner.

Theorem 3 (Conditional contraction of the exact flow)**.** *Let G_W be positive definite and let the exact flow in equation 16 exist for all $s \geq 0$ in its smooth domain. Suppose, throughout that domain,*

$$\nabla F(\theta)^\top G_W(\theta)^{-1}\nabla F(\theta) \geq 2\lambda\big(F(\theta) - F^\star\big), \qquad \lambda > 0, \tag{18}$$

where $F^\star$ is the infimum over the domain. Then

$$F(\theta_s) - F^\star \leq e^{-2\lambda s}\big(F(\theta_0) - F^\star\big).$$

Proof. The chain rule gives $\frac{d}{ds}(F(\theta_s) - F^\star) = -\nabla F(\theta_s)^\top G_W(\theta_s)^{-1}\nabla F(\theta_s) \leq -2\lambda(F(\theta_s) - F^\star)$. Multiplying by $e^{2\lambda s}$ and integrating proves the claim. A sufficient condition for equation 18 is λ-geodesic strong convexity on a geodesically convex domain with an attained minimizer: the first-order strong-convexity inequality gives $F - F^\star \leq \|\operatorname{grad} F\|_{G_W} d - \lambda d^2/2 \leq \|\operatorname{grad} F\|^2_{G_W}/(2\lambda)$. □

The curvature condition is an additional assumption on the acquisition and chosen metric. The result concerns exact continuous-time optimization at fixed x; finite stochastic updates and optimization of the continuous block require separate error control.

The gradient-dominance hypothesis is restrictive even for a binary linear objective. In the probability coordinate $p \in (0, 1)$, take $F(p) = -p$, $F^\star = -1$, and the logarithmic-mean conductance

$$\theta_{01}(p) = \frac{1-2p}{\log((1-p)/p)}.$$

Then $\|\operatorname{grad} F(p)\|^2_G = \theta_{01}(p) \to 0$ as $p \downarrow 0$, while $F(p) - F^\star \to 1$. Hence no positive constant λ makes the gradient-dominance inequality hold on the whole interior. A restricted invariant domain may admit an appropriate condition, but clipping probabilities or keeping finite logits does not establish it automatically.

Ground-cost smoothing and composition. For a factor with C levels, let $q(\phi) = \mathrm{softmax}((\phi - 0.5)/\tau)$ and use the ground cost $D_{ab} = d(a,b)^2$ from the geometry assumption. For $\varepsilon > 0$, consider

$$p_\varepsilon(\phi) = \frac{Kq(\phi)}{\mathbf{1}^\top Kq(\phi)}, \qquad K_{ab} = \exp(-D_{ab}/\varepsilon). \tag{19}$$

This normalized Gibbs-kernel map favors nearby levels while assigning positive mass to every level when costs are finite. It is evaluated by a matrix-vector product and normalization; no Sinkhorn scaling is involved. Assume $D_{aa} = 0$ and $D_{ab} > 0$ for $a \neq b$.

Proposition 1 (Uniform smoothing error)**.** *For factor j with $C_j \geq 2$ levels, put $m_j = \min_{a \neq b} D^{(j)}_{ab} > 0$ and $b_j = (C_j - 1)e^{-m_j/\varepsilon}$. Let p_ε and p_0 denote the products of the smoothed and base factors, respectively. For fixed x,*

$$|\mathbb{E}_{p_\varepsilon}\alpha(x,Z) - \mathbb{E}_{p_0}\alpha(x,Z)| \leq B_\varepsilon(x) := \mathrm{osc}_z\,\alpha(x,z)\,\min\left\{1, \sum_j \frac{b_j}{1+b_j}\right\}, \tag{20}$$

uniformly over the base probabilities. Singleton factors contribute zero.

Proof. For one factor, write $K = I + E$ and $a = \mathbf{1}^\top Eq$. Since E has nonnegative entries and column sums at most b, $0 \leq a \leq b$. For $a > 0$, $r = Eq/a$ is a probability vector and $p_\varepsilon = (q + ar)/(1+a)$, so $\mathrm{TV}(p_\varepsilon, q) \leq a/(1+a) \leq b/(1+b)$. The $a = 0$ case is immediate. Telescoping the product distributions gives $\mathrm{TV}(\bigotimes_j p_{\varepsilon,j}, \bigotimes_j q_j) \leq \min\{1, \sum_j \mathrm{TV}(p_{\varepsilon,j}, q_j)\}$. The expectation of a bounded function differs by at most its range times total variation, proving equation 20. □

Thus smoothing vanishes uniformly as $\varepsilon \to 0$. If $\widehat{\phi}$ is δ_{in}-optimal for the smoothed expectation, its gap for the base expectation is at most $\delta_{\mathrm{in}} + 2B_\varepsilon(x)$, by applying equation 20 at $\widehat{\phi}$ and at a maximizing sequence. Recovery of the original discrete optimum additionally uses PR's point-mass condition. Selecting a feasible point from the distribution requires its own optimization or sampling-error accounting.

The metric must use the distribution appearing in the objective:

$$G^{\mathrm{comp}}_W(\phi) = (J_\phi p_\varepsilon)^\top L(p_\varepsilon)^\dagger J_\phi p_\varepsilon. \tag{21}$$

For a two-level factor, write $k = e^{-1/\varepsilon}$, $s = (1-k)/(1+k)$, and $p_\varepsilon = (k + (1-k)q)/(1+k)$ for the upper-level probability. In a scalar coordinate ϕ,

$$G^{\mathrm{comp}}_W(\phi) = \frac{[sq'(\phi)]^2}{\theta_{01}(p_\varepsilon)}, \qquad (G^{\mathrm{comp}}_W(\phi))^{-1} = \frac{\theta_{01}(p_\varepsilon)}{[sq'(\phi)]^2}.$$

In probability coordinates $q'(\phi) = 1$; in a single-logit coordinate, $q'(\phi) = q(1-q)/\tau$. The contraction theorem applies to this composed metric when its rank, flow-existence, and curvature assumptions hold.

For fixed logits with a unique maximizer and gap $\Delta > 0$, $\partial q_a/\partial \phi_b = \tau^{-1} q_a(\mathbf{1}_{a=b} - q_b) = O(\tau^{-1}e^{-\Delta/\tau})$. At fixed ε, the derivative of the normalized kernel map is bounded on the simplex, so composition retains this pointwise vanishing Jacobian. Smoothing alone therefore supplies no general guarantee against temperature-induced saturation.

A conditional bridge to sampled decoding. The following observation distinguishes an expectation gap from the acquisition of a feasible sampled point.

Lemma 3 (Sampling from a nearly optimal smoothed distribution)**.** *Fix x and a finite feasible discrete domain. Suppose the closure of the base probability family contains every point mass, and let $\alpha^\star_x = \max_z \alpha(x,z)$. Let $A_\varepsilon(\phi) = \mathbb{E}_{p_\varepsilon(\cdot|\phi)}\alpha(x,Z)$. Assume the uniform smoothing bound $|A_\varepsilon - A_0| \leq B_\varepsilon(x)$ and $\sup_\phi A_\varepsilon(\phi) - A_\varepsilon(\widehat{\phi}) \leq \delta_{\mathrm{in}}$. Conditional on this optimized distribution, draw $N \geq 1$ independent samples $Z_1, \ldots, Z_N$ from $p_\varepsilon(\cdot \mid \widehat{\phi})$. For any $h > 0$,*

$$\mathbb{P}\left[\alpha^\star_x - \max_{n \leq N} \alpha(x, Z_n) > h\right] \leq \min\left\{1, \frac{\delta_{\mathrm{in}} + B_\varepsilon(x)}{h}\right\}^N.$$

Proof. The point-mass closure gives $\sup_\phi A_0(\phi) = \alpha_x^\star$. Uniform smoothing implies $\sup_\phi A_\varepsilon(\phi) \geq \alpha_x^\star - B_\varepsilon(x)$, so $\mathbb{E}[\alpha_x^\star - \alpha(x, Z)] \leq \delta_{\text{in}} + B_\varepsilon(x)$. The gap is nonnegative. Markov's inequality bounds its one-sample tail; conditional independence gives the Nth power for the best sample. □

This lemma does not apply to deterministic rounding of probabilities and does not prove that a finite optimizer achieves δ_{in}. It controls discrete decoding only at the fixed continuous block; it does not control outer-loop search coverage.

D.6 Main Regret Bounds

The inner analysis concerns a distribution at fixed continuous coordinates. To connect it to BO, we account separately for surrogate approximation, search coverage, and the acquisition value of the point actually evaluated.

Use $\mathcal{X}$ for the feasible mixed domain and index predictions before evaluation t. Fix the reference kernel $k(u, u) \leq 1$ and regularization $\sigma_{\text{ref}}^2 > 0$. On the simultaneous reference event,

$$|f(u) - \mu_t^{\text{GP}}(u)| \leq \sqrt{\beta_t}\, \sigma_t^{\text{GP}}(u) \quad \text{for all } u \in \mathcal{X},\ t \geq 1. \tag{22}$$

The confidence parameter must obey a valid construction such as that in Section D.3. On the uniform approximation event, the additive transfer lemma gives

$$b_t = e_t + \sqrt{\beta_t e_t}, \qquad |f(u) - \mu_t(u)| \leq \sqrt{\beta_t}\sigma_t(u) + b_t. \tag{23}$$

The nonnegative search and optimization errors defined below use the actual feasible point evaluated, including decoding. Their averages are not proved to vanish for the implementation.

Let $S_t \subseteq \mathcal{X}$ be the feasible search set at round t, including any full-dimensional candidates, and let $x_t \in S_t$ be the evaluated point. With $\alpha_t = \mu_t + \sqrt{\beta_t}\sigma_t$, define

$$\rho_t = \sup_{u \in \mathcal{X}} \alpha_t(u) - \sup_{u \in S_t} \alpha_t(u), \qquad \eta_t = \sup_{u \in S_t} \alpha_t(u) - \alpha_t(x_t). \tag{24}$$

These nonnegative quantities distinguish coverage of the acquisition landscape from optimization within the search set. The evaluated point, including discrete decoding, determines η_t.

Theorem 4 (Regret with search and surrogate errors)**.** *Under the preceding assumptions, on the intersection of the reference confidence event and the surrogate approximation event,*

$$R_T \leq \sqrt{C_1 T \beta_T \gamma_{T,\text{full}}} + \sum_{t=1}^{T}(2b_t + \rho_t + \eta_t), \qquad C_1 = \frac{8}{\log(1 + \sigma_{\text{ref}}^{-2})}. \tag{25}$$

Here $\gamma_{T,\text{full}}$ is the maximal information gain of the reference kernel over feasible sequences of T points, including repeated evaluations. If the approximation event has probability at least $1 - \delta_{\text{app}}$, the joint event has probability at least $1 - \delta_{\text{GP}} - \delta_{\text{app}}$. For deterministic approximation bounds, only δ_{GP} is required.

Proof. Write $\alpha_t^{GP} = \mu_t^{GP} + \sqrt{\beta_t}\sigma_t^{GP}$. The discrepancy assumptions imply $|\alpha_t - \alpha_t^{GP}| \leq b_t$. For an objective maximizer $x^\star$, confidence and equation 24 give

$$\begin{aligned} f(x^\star) &\leq \alpha_t^{GP}(x^\star) \leq \alpha_t(x^\star) + b_t \leq \alpha_t(x_t) + \rho_t + \eta_t + b_t \\ &\leq \alpha_t^{GP}(x_t) + \rho_t + \eta_t + 2b_t. \end{aligned}$$

Combining with the lower confidence bound at x_t yields $r_t \leq 2\sqrt{\beta_t}\sigma_t^{GP}(x_t) + 2b_t + \rho_t + \eta_t$. For $v_t = (\sigma_t^{GP}(x_t))^2 \leq 1$, concavity of the logarithm gives $v_t \leq \log(1 + \sigma_{\text{ref}}^{-2} v_t)/\log(1 + \sigma_{\text{ref}}^{-2})$. The GP determinant identity yields $\sum_t \log(1 + \sigma_{\text{ref}}^{-2} v_t) \leq 2\gamma_{T,\text{full}}$. Cauchy–Schwarz and monotonicity of β_t now give

$$\sum_t 2\sqrt{\beta_t}\sigma_t^{GP}(x_t) \leq 2\sqrt{T\beta_T \sum_t v_t} \leq \sqrt{C_1 T \beta_T \gamma_{T,\text{full}}},$$

which proves equation 25. □

The result accommodates adaptive bases and random-sketch candidate generation through S_t. Subspace-angle control alone does not bound ρ_t: that step additionally requires coverage of high-acquisition feasible configurations. Likewise, the fixed-x flow theorem controls an idealized inner objective, rather than the full η_t of finite mixed optimization.

D.7 Information Gain Bounds for High-Dimensional Subspaces

A lower-dimensional search can reduce information gain when the evaluation sequence is contained in a common subspace. The reduction depends on both this containment and the kernel.

Lemma 4 (Information gain on a fixed subspace)**.** *Let $S_r = \{x_0 + Uz : z \in \mathcal{Z}\}$ for a fixed orthonormal $U \in \mathbb{R}^{d\times r}$ and compact convex $\mathcal{Z} \subset \mathbb{R}^r$. Assume the restricted kernel is a squared-exponential or Matérn kernel with fixed parameters and diagonal bounded by one. Then the classical bounds of Srinivas et al. (2010, Theorem 5) give*

$$\gamma_{T,S_r} = \begin{cases} O((\log T)^{r+1}), & \textit{squared exponential}, \\ O(T^{r(r+1)/(2\nu+r(r+1))}\log T), & \textit{Matérn}, \ \nu > 1. \end{cases} \tag{26}$$

The same upper bounds hold for a feasible subset of S_r.

Proof. The kernel matrix at $x_0+Uz_1, \ldots, x_0+Uz_T$ equals that of the restricted kernel at $z_1, \ldots, z_T$. Hence their log-determinant information gains agree. Apply the stated kernel bounds in dimension r; restricting the admissible sequences cannot increase the maximum. □

If all evaluation points lie in this common S_r, the variance-sum step of Theorem 4 can use γ_{T,S_r}. The search-coverage term still accounts for optimization relative to the full domain. For a sequence of changing subspaces or additional full-dimensional candidates, the applicable bound is on the full sequence of evaluations; individual ranks $\widehat{r}_t$ do not establish a common r-dimensional domain.

D.8 Convergence Rate

The regret decomposition gives explicit sufficient conditions for convergence and identifies which errors must decrease with the evaluation budget.

Corollary 1 (Sufficient conditions for vanishing average regret)**.** *Under Theorem 4, if*

$$\frac{\beta_T \gamma_{T,\mathrm{full}}}{T} \longrightarrow 0, \qquad \frac{1}{T}\sum_{t=1}^{T}(2b_t + \rho_t + \eta_t) \longrightarrow 0, \tag{27}$$

then $R_T/T \to 0$ on the confidence event. Under the fixed-subspace conditions, the corresponding restricted information gain may be used instead.

Proof. Divide equation 25 by T and apply equation 27 to its two nonnegative terms. □

For example, $e_t = O(t^{-\xi})$ for any $\xi > 0$ and polylogarithmic β_t imply $b_t \to 0$, and hence vanishing average surrogate error. Coverage and optimization errors must also vanish on average. A bound leaving a fixed positive approximation term yields a residual upper bound, rather than a proof of zero average regret.

For the exact flow in Theorem 3, an initial objective gap at most H is reduced to $\delta_{\mathrm{in}} > 0$ after flow time $s \geq (2\lambda)^{-1}\log(H/\delta_{\mathrm{in}})$ when $H > \delta_{\mathrm{in}}$. Together with equation 20, this separates optimization accuracy from finite smoothing. Converting flow time to a number of stochastic updates requires step-size, smoothness, and gradient-noise assumptions; it does not establish a wall-clock complexity bound for EDGE.

E STRUCTURE, REPRESENTATION, AND LEARNING IN COMBINATORIAL SEARCH

Purely combinatorial problems separate three questions that are closely coupled in mixed-variable optimization: what structure the objective contains, which configurations the search representation can express, and what the optimizer can learn from its observations. We examine these questions through a separate combinatorial benchmark comparison, an analysis of optimizer containment under typed embeddings, and controlled surrogate and binning comparisons. Together, they clarify when learning useful directions can translate into effective discrete search.

E.1 BENCHMARK PATTERNS AND ENCODING

Table 4 compares the complete optimizers on MaxSAT, LABS, contamination control, cluster expansion, their four relabeled variants, and a logic-rewriting task. The original–relabeled pairs provide a useful comparison because relabeling preserves the underlying optimization problem while changing its input representation. On cluster expansion, EDGE's reported mean final value changes from 1765 to 884, while Bounce retains 1765 on both encodings. The corresponding MaxSAT values are 194 and 178 for EDGE, and 196 on both encodings for Bounce. These outcomes show that the complete search pipeline can respond differently to equivalent encodings. The component studies below examine the information and representation choices involved.

Table 4: Mean final incumbent $\pm$ standard deviation over seeds on the nine purely combinatorial problems (5 seeds, 300 evaluations after initialisation; larger is better; best mean per row in bold; number of complete seeds per cell in parentheses).

	EDGE	Bounce	PR	Casmopolitan	BODi	SMAC
maxsat60	194 ± 1.79 (5)	$\mathbf{196 \pm 0}$ (5)	185 ± 2.95 (5)	192 ± 2.42 (5)	$\mathbf{196 \pm 0}$ (5)	156 ± 6.55 (5)
labs	2.8 ± 0.489 (5)	3.77 ± 0.214 (5)	3.73 ± 0.321 (5)	2.89 ± 0.274 (5)	$\mathbf{3.84 \pm 0.309}$ (5)	2.52 ± 0.143 (5)
contamination25	-21.6 ± 0.118 (5)	$\mathbf{-21.2 \pm 0.0492}$ (5)	-21.2 ± 0.0513 (5)	-21.4 ± 0.169 (5)	-21.2 ± 0.115 (5)	-21.4 ± 0.0683 (5)
maxsat60f	178 ± 13.9 (5)	$\mathbf{196 \pm 0}$ (5)	187 ± 5.2 (5)	187 ± 8 (5)	$\mathbf{196 \pm 0}$ (5)	165 ± 10.1 (5)
labsf	3.1 ± 0.203 (5)	$\mathbf{3.83 \pm 0.386}$ (5)	3.64 ± 0.408 (5)	3.45 ± 0.299 (5)	3.62 ± 0.486 (5)	2.65 ± 0.453 (5)
contamination25f	-21.6 ± 0.122 (5)	-21.3 ± 0.115 (5)	-21.2 ± 0.0871 (5)	-21.5 ± 0.0838 (5)	$\mathbf{-21.2 \pm 0.00447}$ (5)	-21.4 ± 0.11 (5)
cluster_expansion125	1765 ± 0 (5)	$\mathbf{1765 \pm 0}$ (5)	960 ± 145 (5)	1566 ± 66.5 (5)	$\mathbf{1765 \pm 0}$ (5)	59.7 ± 78.6 (5)
cluster_expansion125f	884 ± 103 (5)	$\mathbf{1765 \pm 0}$ (5)	564 ± 231 (5)	1520 ± 147 (5)	$\mathbf{1765 \pm 0}$ (5)	132 ± 145 (5)
mig_seq_opt	-0.695 ± 0.0083 (5)	$\mathbf{-0.693 \pm 0.0257}$ (5)	-0.728 ± 0.0221 (5)	-0.705 ± 0.011 (5)	-0.749 ± 0.0126 (5)	-0.702 ± 0.013 (5)

E.2 OBJECTIVE STRUCTURE AND SURROGATE INFORMATION

The Walsh expansion provides a common language for describing structure in a binary objective. For $s \in \{-1, +1\}^d$,

$$f(s) = \sum_{S \subseteq [d]} \widehat{f}(S) \prod_{i \in S} s_i.$$

Under the uniform measure on the hypercube, these basis functions are orthonormal: coefficients with $|S| = 1$ describe additive effects, and higher-order coefficients describe interactions (O'Donnell, 2014). For a nonconstant objective, the fraction

$$\frac{\sum_{|S|=1} \widehat{f}(S)^2}{\sum_{S \neq \varnothing} \widehat{f}(S)^2}$$

measures the additive share of its variance. Excluding the constant term makes this quantity insensitive to an arbitrary objective offset.

For MaxSAT and cluster expansion, coefficients can be calculated directly from the specified objective terms without fitting a surrogate. This calculation distinguishes structure present in the benchmark from structure learned from limited observations. Additive structure need not be low-dimensional, since many variables can contribute additive effects, and even a small interaction component can change which configuration is optimal.

EDGE observes neither these coefficients nor the objective's variable influences directly. It selects directions from gradients of a learned acquisition on a continuous relaxation. The empirical question is therefore whether the surrogate and search procedure turn the available observations into useful ordering of feasible configurations. Section E.3 tests this question through a surrogate substitution and through changes to how variables are grouped and probed.

E.3 Learning and using variable importance

Controlled comparisons. We use the eight original and relabeled binary problems, with 20 initial evaluations followed by a budget of 200 BO evaluations. The logic-rewriting task is excluded from this study. Bounce+TabPFN replaces Bounce's GP surrogate; Bounce+IB changes the bin assignment using importance estimated from the observed history; Bounce+AP uses paid single-bit probes to inform the assignment. EDGE and EDGE-comb provide contextual comparisons: EDGE-comb uses an informed embedding with a different search loop and therefore does not isolate bin assignment within Bounce.

Table 5 reports available-seed means and standard deviations. Paired counts use the seeds shared with Bounce: 80 pairs for EDGE, EDGE-comb and Bounce+IB, and 240 for Bounce+AP. The separate surrogate-swap comparison has 79 pairs. Task-level signed-rank tests use eight task means and are reported without multiplicity adjustment as exploratory summaries; original and relabeled instances share an underlying problem. The endpoint means and paired counts can use different seed subsets, since Bounce has 30 available seeds and several arms have ten.

Surrogate suitability. Replacing the surrogate inside Bounce provides a direct way to examine the value of its predictive model within the same search framework. Bounce+TabPFN records 10 wins, 49 losses and 20 ties against Bounce's GP configuration over 79 paired cells. The reported task-level signed-rank p-value is 0.018. The offline prediction diagnostic points in the same direction on the studied MaxSAT and cluster-expansion instances. With 100 training evaluations and 10,000 held-out configurations, the reported Spearman correlations are 0.81 versus 0.52 on MaxSAT and 0.58 versus 0.30 on cluster expansion, for the GP and TabPFN respectively. On MaxSAT, the GP's predicted best 100 configurations contain 41 of the true best 100, compared with 17 for TabPFN. On LABS and its relabeling, both models have correlations within 0.01 of zero at this training size. These comparisons identify surrogate suitability as one contributor on the studied instances, rather than measuring acquisition-gradient accuracy.

Identifying variables that move together. The containment analysis motivates isolating effective variables, but the embedding also determines what the observation history can identify. If variables in a bin remain tied throughout the observations used for estimation, then $x_i^{(t)} = s_i v^{(t)}$ for their fixed signs s_i. An additive model consequently observes

$$\sum_{i \in B} \theta_i x_i^{(t)} = \left(\sum_{i \in B} s_i \theta_i \right) v^{(t)}.$$

It can identify the signed sum of effects but cannot separate the members' coefficients from these observations alone. Bounce+IB estimates importance with ridge regression on degree-one Walsh features and changes the assignment within Bounce's loop. Its 19 wins, 23 losses and 38 ties over 80 paired cells illustrate the practical question left open by the containment guarantee: a rule for isolating variables needs observations that distinguish them. The separate EDGE-comb comparison changes the search loop as well, so its outcomes cannot resolve that question by themselves.

Paying for local information. Bounce+AP obtains additional information through single-bit flips from the incumbent for members of a selected bin. These probes reveal local effects without forcing the members to move together, and their evaluations count against the same budget, capped at 33 of the 220 total evaluations. The final comparison records 48 wins, 54 losses and 138 ties over 240 paired cells. Bounce+AP matches Bounce's reported mean and zero standard deviation on both MaxSAT encodings and both cluster-expansion encodings; the other tasks account for the remaining variation in final outcomes. This comparison evaluates the value of additional local information after charging for its acquisition, rather than treating variable identification as free. The reported results do not establish an overall improvement at this budget.

E.4 Implications for mixed-variable search

The theory and experiments distinguish representational capacity from the information needed to use it. Randomized discrete embeddings can retain a fixed optimal pattern even when effective variables

Table 5: Controlled comparisons on eight binary tasks after 20 initial and 200 BO evaluations. Entries give mean final incumbent ± standard deviation, with seed counts in parentheses; larger is better. Bold marks the best mean before display rounding. Paired counts are wins–losses–ties against Bounce over shared seeds. The final row reports unadjusted exploratory tests over eight task means.

	Bounce	EDGE	EDGE-comb	Bounce+IB	Bounce+AP
maxsat60	$\mathbf{196 \pm 0}$ (30)	$\mathbf{196 \pm 0}$ (10)	166 ± 9.49 (10)	195 ± 1.03 (10)	$\mathbf{196 \pm 0}$ (30)
maxsat60f	$\mathbf{196 \pm 0}$ (30)	174 ± 7.07 (10)	167 ± 8 (10)	$\mathbf{196 \pm 0}$ (10)	$\mathbf{196 \pm 0}$ (30)
labs	$\mathbf{3.86 \pm 0.41}$ (30)	2.71 ± 0.318 (10)	3.05 ± 0.356 (10)	3.75 ± 0.454 (10)	3.76 ± 0.422 (30)
labsf	3.66 ± 0.23 (30)	2.99 ± 0.303 (10)	3.02 ± 0.314 (10)	$\mathbf{3.8 \pm 0.382}$ (10)	3.73 ± 0.455 (30)
contamination25	$\mathbf{-21.3 \pm 0.111}$ (30)	-21.6 ± 0.195 (10)	-21.5 ± 0.137 (10)	-21.3 ± 0.156 (10)	-21.3 ± 0.121 (30)
contamination25f	-21.3 ± 0.108 (30)	-21.6 ± 0.126 (10)	-21.5 ± 0.102 (10)	-21.3 ± 0.0821 (10)	$\mathbf{-21.3 \pm 0.112}$ (30)
cluster_expansion125	$\mathbf{1765 \pm 0}$ (30)	1646 ± 68.4 (10)	534 ± 132 (10)	1759 ± 20.0 (10)	$\mathbf{1765 \pm 0}$ (30)
cluster_expansion125f	$\mathbf{1765 \pm 0}$ (30)	759 ± 185 (10)	485 ± 171 (10)	1740 ± 43.7 (10)	$\mathbf{1765 \pm 0}$ (30)
paired result vs. Bounce	—	2–67–11	6–74–0	19–23–38	48–54–138
Wilcoxon p (task means)	—	0.00781	0.00781	0.499	0.715

collide, while correctly isolating those variables can increase its containment probability. The controlled studies examine the corresponding learning problem: surrogate choice affects the ordering learned from limited observations, tied variables can obscure individual effects, and active probes consume evaluations to reveal local information. For EDGE, the resulting design requirement is that learned acquisition directions must lead to useful feasible decisions. This links the combinatorial study to mixed-variable search without assigning the broader application-group performance differences to a single mechanism.

E.5 Containment guarantees for typed embeddings

We formalize the representability question used in Section **??**. The analysis concerns a fixed embedding dimension and a specified partition law, following the typed-binning constructions motivated by Bounce, BAxUS and HeSBO (Papenmeier et al., 2023; 2022; Nayebi et al., 2019).

Setup. Let

$$\mathcal{X} = \{-1, +1\}^{n_{\text{bin}}} \times \prod_{i=1}^{n_{\text{cat}}} [c_i], \qquad c_i \geq 2.$$

Variables are partitioned into bins separately by type. A binary member i of bin j is represented as $x_i = s_i v_j$, where s_i is an independent uniform sign. A categorical bin has $\kappa_j = \max_{i \in B_j} c_i$ levels. Its level $k \in [\kappa_j]$ maps member i to $\lceil k c_i / \kappa_j \rceil$, followed by an independent uniform label randomization from a transitive permutation group. Full label permutations and cyclic shifts both satisfy this assumption. Ordinal variables may be represented by this categorical construction; the results concern its finite level sets.

Definition 1 (Effective variables)**.** An objective is a d_e-junta if there is a minimal coordinate set S, $|S| = d_e$, such that $f(x) = g(x_S)$ (O'Donnell, 2014). Fix an optimal pattern $z^\star$ of these effective variables. Write $E_j = B_j \cap S$ and $m_j = |E_j|$ for the effective members and load of bin j. For binary members, use $c_i = 2$ when counting label patterns.

Containing this fixed pattern is sufficient to contain a global optimizer. It is also necessary when the optimal effective-variable pattern is unique. Conditional on the partition, the pattern is representable precisely when every bin can realize its own part, and the independent randomizations make these bin events independent.

Lemma 5 (Per-bin target coverage)**.** *For a bin with $m_j \geq 1$, let A_j be the set of distinct joint labels of its effective members before randomization. For a categorical bin,*

$$A_j = \left\{ \left(\left\lceil \frac{k c_i}{\kappa_j} \right\rceil \right)_{i \in E_j} : k \in [\kappa_j] \right\};$$

for a binary bin, $|A_j| = 2$. Under independent transitive label randomizations, the probability of realizing any fixed target is

$$q_j = \frac{|A_j|}{\prod_{i \in E_j} c_i}.$$

In particular, a singleton effective variable has $q_j = 1$, while equal cardinalities c give $q_j = c^{1-m_j}$.

Proof. The randomizations make the coverage probability equal for every target in the Cartesian product of effective labels. Every randomized image contains exactly $|A_j|$ distinct targets. Summing the common coverage probability over all $\prod_{i \in E_j} c_i$ targets gives the result. The maximum-cardinality map covers every label of a singleton effective variable. For equal cardinalities its effective joint labels form a diagonal set of size c. □

Theorem 5 (All-pattern containment under balanced binning)**.** *For one variable type, assign n variables uniformly to slots in b bins, where $1 \le b \le n$ and the fixed sizes $s_1, \dots, s_b$ sum to n and differ by at most one. If k variables are effective, the probability of representing every effective-variable pattern is*

$$Q_B(n, b, k) = \frac{\displaystyle\sum_{\substack{J \subseteq \{1,\dots,b\} \\ |J|=k}} \prod_{j \in J} s_j}{\binom{n}{k}}.$$

Here $Q_B(n, b, 0) = 1$ and $Q_B(n, b, k) = 0$ for $k > b$. With independent partitions for different variable types, multiply their probabilities; an absent type contributes one.

Proof. The effective variables occupy a uniform k-subset of the n slots. Choosing k distinct bins and one slot in each gives the numerator. An effective singleton reaches every label. If two or more effective variables share a categorical bin, their pre-randomization joint labels form a monotone chain with at most $1 + \sum_{i \in E_j}(c_i - 1) < \prod_{i \in E_j} c_i$ distinct patterns. A binary bin with multiple effective members likewise misses some joint patterns. Label randomization does not change the number represented. Thus all-pattern containment is equivalent to having no effective-variable collisions, and the product rule follows from independent partitions. □

Theorem 6 (Fixed-pattern and optimizer containment)**.** *Under the setup above, a fixed optimal effective-variable pattern satisfies*

$$\Pr(z^\star \text{ is representable}) = \mathbb{E}_\pi \left[\prod_{j : m_j \ge 1} \frac{|A_j|}{\prod_{i \in E_j} c_i} \right], \tag{28}$$

where π is the random partition. This is a lower bound on the probability that the embedded space contains any global optimizer, with equality for a unique optimal effective-variable pattern.

For one type with equal cardinality c, define $w_c(0) = 1$ and $w_c(m) = c^{1-m}$ for $m \ge 1$. Independent uniform hashing of k effective variables into $b \ge 1$ bins gives

$$P_H(b, k; c) = \frac{k!}{b^k} [x^k] \big(c e^{x/c} - (c - 1) \big)^b,$$

where $[x^k]$ extracts the coefficient of x^k. Uniform balanced assignment of n variables gives

$$P_B(n, b, k; c) = \frac{1}{\binom{n}{k}} \sum_{\substack{m_1 + \cdots + m_b = k \\ 0 \le m_j \le s_j}} \prod_{j=1}^{b} \binom{s_j}{m_j} w_c(m_j).$$

Both probabilities are one when $k = 0$.

Proof. Condition on the partition and apply Lemma 5 independently to its occupied effective bins, then average. Under hashing, the effective loads have a multinomial distribution. The generating function of the rescue weights is $\sum_{m \ge 0} w_c(m) x^m / m! = c e^{x/c} - (c - 1)$, giving P_H. Under balanced assignment, the loads have a multivariate hypergeometric distribution, giving P_B. □

For two equal-cardinality effective variables, fixed-pattern containment is $1 - (1 - 1/c) p_{\text{coll}}$: separate bins preserve the target, and a shared bin preserves it with probability $1/c$. With balanced bins at $(n, b, k, c) = (30, 6, 4, 2)$, the formulas give $P_B = 0.64327$ and $Q_B = 0.34209$. The extra coverage comes from target patterns retained despite collisions.

Proposition 2 (A tight binary collision example)**.** *Let $f(x) = -x_a x_b$ on $\{-1, +1\}^n$. Conditional on a, b sharing a bin, write $x_a = s_a v$ and $x_b = s_b v$ with independent uniform signs. If $s_a s_b = -1$, every bin value attains the optimum $+1$; otherwise every bin value gives -1. The probability of containing an optimizer is therefore $1/2$ under this collision, and is one if the two variables occupy separate bins.*

Proof. Since $v^2 = 1$, the embedded objective is the constant $-s_a s_b$. The two possible sign products are equally likely. □

Corollary 2 (Containment after singleton isolation)**.** *For one type with equal cardinality c, let n be the number of variables, b the bin budget, and k the number of effective variables. Isolate a specified set D of $h < b$ variables, of which r are effective, into singleton bins. If, conditional on D, the remaining variables are freshly reassigned using the stated partition law and independent label randomizations, fixed-pattern containment is*

$$\begin{cases} P_H(b-h, k-r; c), & \textit{independent hashing}, \\ P_B(n-h, b-h, k-r; c), & \textit{balanced reassignment}. \end{cases}$$

For independent typed partitions, multiply the conditional probabilities within types. The separate construction giving every variable its own singleton bin has containment probability one.

Proof. The isolated singletons realize their target labels with probability one. Apply Theorem 6 to the remaining variables and bins. □

At $b = 6$, $k = 4$ and $c = 2$, correctly isolating $r = 0, 1, 2, 3$ effective variables under hashing gives 0.59664, 0.730, 0.875 and 1. The accuracy of identification matters: isolating four variables of which one is effective leaves two bins for three effective variables, giving 0.4375. These calculations specify both the allocation of bins and the reassignment law; a history-dependent splitting policy need not satisfy that conditional law.

Relation to optimization and continuous variables. These statements concern exact representation at a fixed dimension. They do not give the probability that a finite-budget optimizer finds a represented solution. Containment is nondecreasing for genuinely nested reachable sets, and becomes certain at full dimension. For continuous variables, tying several coordinates confines their joint values to a lower-dimensional set of Lebesgue measure zero. For a fixed embedding, a target drawn from a density is missed with probability one. Special targets or nonunique optimizer sets can behave differently. Finally, the singleton guarantee uses the maximum-cardinality map: if a lone effective variable with c_i levels is represented by a bin with $\kappa_j < c_i$ levels, its target coverage under uniform transitive randomization is κ_j / c_i. Applying these results to an implementation requires matching its bin map and partition law to the stated construction.

F BASELINE ALGORITHMS IMPLEMENTATION DETAILS

We benchmark EDGE against seven baselines. Their implementation sources and experiment-specific modifications are summarized below.

- **Bounce** (Papenmeier et al., 2023): We use the authors' released implementation and configuration (link: `https://github.com/lpapenme/bounce`, license: MIT license, last accessed: Sep 24th, 2026).
- **PR** (Daulton et al., 2022): We use the authors' implementation of probabilistic reparameterization with its published gradient-estimator settings (link: `https://github.com/facebookresearch/bo_pr`, license: MIT license, last accessed: Sep 24th, 2026).
- **Casmopolitan** (Wan et al., 2021): We use the reference builder provided by the MCBO framework (Dreczkowski et al., 2023) (link: `https://github.com/huawei-noah/HEBO/tree/master/MCBO`, license: MIT license, last accessed: Sep 24th, 2026).
- **BODi** (Deshwal et al., 2023): We use the authors' released implementation with $m = 128$ dictionary elements, following the paper's setting (link: `https://github.com/aryandeshwal/BODi`, license: MIT license, last accessed: Sep 24th, 2026). In our implementation, we add an Adam fallback when standard GP fitting fails.
- **SMAC3** (Lindauer et al., 2022): We use the hyperparameter-optimization facade with its own seeded initial design.
- **Sobol random search**: We draw scrambled Sobol points and snap discrete coordinates to their admissible grids, using the same evaluation budget and seed set as the other methods.

G BENCHMARK PROBLEMS IMPLEMENTATION DETAILS

Table 6 lists the benchmark dimensions and variable types. We represent inputs in $[0, 1]^d$ and snap finite-cardinality coordinates to their admissible levels before objective evaluation. All objective scores are oriented for maximization. inimization objectives are negated, and nonnegative constraint penalties reduce the resulting score. The entries below describe benchmark sources and implementation changes. Initialization and evaluation-budget conventions are specified in Section 4.

- **HPO-B surrogates** (Pineda-Arango et al., 2021). Seven problems named by HPO-B's `surrogate-[search_space_id]-[task_id]` convention; spaces `5906`, `5971` and `6767` are XGBoost spaces. The objective is the predicted accuracy of an XGBoost surrogate over the box (HPO-B's continuous mode), with one 15-level ordinal and six or seven binary coordinates per space; which hyperparameter each coordinate controls is not recorded in HPO-B's tables. The official surrogates are no longer downloadable, so we retrain equivalent ones from the released pool. Surrogate results are comparable across methods but not to published HPO-B numbers, and smooth surrogates may flatter gradient-based methods.
- **EPFL logic-synthesis circuits** (Amarú et al., 2015; Dreczkowski et al., 2023). Five circuits posed as MCBO's `aig_optimization_hyp` task (after BOiLS, Grosnit et al., 2022): choose an and-inverter-graph rewriting sequence and the operators' hyperparameters. 20 nominal operator slots, declared nominal to our method, plus 22 integer and 14 boolean hyperparameters.
- **YAHPO-Gym surrogate HPO** (Pfisterer et al., 2022). The `iaml_super` scenario on OpenML task `41146` through YAHPO's ONNX surrogate; objective $1 - \mathrm{mmce}$ at full training-set size; 13 discrete coordinates of cardinality 2 to 2000, names not recorded in our pipeline.
- **Firefly structural design** (Gandomi et al., 2011). `CarMixed` (the car side-impact design of Gu et al., 2001; two binary material choices for the B-pillar inner and floor-side inner panels), `PressureVesselMixed` (shell and head thicknesses in 0.0625-inch steps, $k = 99$) and `ReinforcedConcreteBeam` (Amir & Hasegawa, 1989; bar area from a catalogue, $k = 76$, and an integer section dimension in $\{28, \ldots, 40\}$ inches, bound as the depth in our implementation), ported from BoCoDe (Yu et al., 2026a); constraint violations are subtracted from the negated cost. The Firefly cantilever beam was excluded: the implementation used in our runs carried a mis-transcribed Young's modulus and an incomplete deflection constraint, leaving every method on a constant penalty plateau.
- **Beam with section catalogue** (Cuesta Ramirez et al., 2022; Roustant et al., 2020). `EulerBeamMixed`, ported from BoCoDe: a cantilever beam minimising $PL^3/(3ES^2\tilde{I}) + \alpha LS$ ($P = E = 600$, $\alpha = 60$) over its length, section area and a 12-level categorical profile (four shapes at three hollowness levels), which sets the normalised moment of inertia $\tilde{I}$.
- **BODi / COMBO combinatorial** (Deshwal et al., 2023; Oh et al., 2019). Ported from the BODi code. `maxsat60`: the weighted-CNF instance `frb-frb10-6-4.wcnf` from the 2018 MaxSAT Evaluation (Bacchus et al., 2019). `labs`: the merit factor of a length-50 binary sequence, classically hard at this length (Shaydulin et al., 2024), read as a negative control. `contamination25` (Hu et al., 2010): binary stage controls under a chance constraint, in the implementation shipped with the PR code (Daulton et al., 2022). `pest25`: five-way stage interventions (none or one of four pesticide types).
- **Cluster-expansion MaxSAT** (Papenmeier et al., 2023). `cluster_expansion125`: the materials-science instance `IS1_5.0.5.0.0.5_softer_periodic.wcnf` of the 2018 MaxSAT Evaluation (125 variables, 60,375 soft clauses), verified bit-exact against Bounce's implementation. For the binary problems we also run Bounce's relabelled variants, with one random relabelling mask shared across all methods and seeds so that the comparison stays paired.
- **Casmopolitan / CoCaBO synthetics** (Wan et al., 2021; Ru et al., 2020a). `ackley53` (50 binary and three continuous coordinates) and `func2C` (two categorical coordinates each selecting a sub-function of two continuous coordinates), ported verbatim from the Casmopolitan repository. Deviation: we drop the original small uniform observation noise so that the objective is deterministic.

- **CIFAR-10 sparse attack** (Wan et al., 2021). `cifar_attack85`: the Casmopolitan sparse targeted attack in the BayesOpt-Attack/GenAttack line (Ru et al., 2020b; Alzantot et al., 2019) on a Carlini–Wagner network (Carlini & Wagner, 2017), maximising the negated attack margin on a fixed image. It has 42 categorical pixel-column choices ($k = 14$), one upsampling choice ($k = 3$) and 42 continuous magnitudes. Deviation: the original overrides the score with a constant once the attack succeeds; we always return the continuous margin and expose success separately.
- **MCBO ML and biology tasks** (Dreczkowski et al., 2023). `xgb_mnist`: MCBO's XGBoost space (three binary choices, integer `max_depth`, four continuous rates). Deviation: evaluated on MNIST with fixed stratified splits and 20 trees instead of scikit-learn's `digits`. `rna_fold`: RNA inverse folding to Eterna100 puzzle 65 (Anderson-Lee et al., 2016) with ViennaRNA (Lorenz et al., 2011); 40 four-way base choices, negated Hamming distance to the target structure.
- **Cellular coverage (CCO)** (Dreifuerst et al., 2021; Daulton et al., 2022). `cco30`, in the implementation shipped with the PR code: 15 six-level antenna downtilts and 15 continuous transmit powers; weak-coverage and over-coverage fractions scalarised with equal weights.
- **Portfolio selection** (Chang et al., 2000). `portfolio_port5`: OR-Library instance `port5` (Beasley, 1990), the Nikkei 225; 225 binary selection indicators and 225 continuous weights renormalised over the selected assets, with the risk–return objective at $\lambda = 0.5$ and the cardinality ($K = 10$) and bound constraints as penalties whose weights are our own fixed constants.
- **Thinned and phased antenna array** (Haupt, 1994). `antenna200`: Haupt's 100-element thinning problem (100 binary element switches) minimising the peak sidelobe level; the 100 continuous excitation phases are our extension, for which we claim no source.
- **Reactive power dispatch, IEEE 118-bus** (Thurner et al., 2018; Zimmerman et al., 2010). `orpd118`: loss minimisation on `pandapower`'s `case118` with 54 continuous voltage setpoints, 9 ordinal transformer taps ($k = 17$) and 14 binary shunt banks; voltage violations enter as a quadratic penalty. Deviation: `case118` ships fixed tap ratios, so we impose a uniform 1.25% step over ± 8 positions, the usual convention of the literature but no specific study.
- **BikeBench bicycle design** (Regenwetter et al., 2026). `bikebench64`: 48 continuous frame parameters, four integer component counts, three booleans and nine categorical component types (class labels not recorded on our side), scored by the package's surrogates. The single-objective form is ours: structural mass plus a penalty on the mean normalised violation of the 34 constraints.

Table 6: The 32-problem benchmark table, with total dimension and counts of continuous, binary, categorical, and integer coordinates. The suite contains 21 mixed continuous/discrete tasks and 11 fully discrete tasks. Source groups and implementation adaptations are described in Appendix G.

Problem	d	cont.	bin.	cat.	int.	Mixed-variable or Combinantorial?
BODi / COMBO combinatorial (*Synthetic*) (Deshwal et al., 2023; Oh et al., 2019)						
`maxsat60`	60	0	60	0	0	Combinatorial (fully binary).
`labs`	50	0	50	0	0	Combinatorial (fully binary).
`contamination25`	25	0	25	0	0	Combinatorial (fully binary).
`pest25`	25	0	0	25	0	Combinatorial (fully categorical).
Cluster-expansion MaxSAT (*Synthetic*) (Papenmeier et al., 2023)						
`cluster_expansion125`	125	0	125	0	0	Combinatorial (fully binary).
Casmopolitan / CoCaBO (*Synthetic*) (Wan et al., 2021; Ru et al., 2020a)						
`ackley53`	53	3	50	0	0	Mixed-variable
`func2C`	4	2	0	2	0	Mixed-variable
HPO-B surrogates (*HPO*) (Pineda-Arango et al., 2021)						
`HPOBsurr_6767_9967`	18	8	7	0	3	Mixed-variable
`HPOBsurr_5906_3889`	16	8	7	0	1	Mixed-variable
`HPOBsurr_5971_10093`	16	8	7	0	1	Mixed-variable
`HPOBsurr_5971_3954`	16	8	6	0	2	Mixed-variable
`HPOBsurr_5971_43`	16	8	6	0	2	Mixed-variable
`HPOBsurr_5971_6566`	16	8	7	0	1	Mixed-variable
`HPOBsurr_5971_9970`	16	8	7	0	1	Mixed-variable
YAHPO-Gym surrogate HPO (*HPO*) (Pfisterer et al., 2022)						
`yahpo_iamlsuper_41146`	28	15	2	3	8	Mixed-variable
MCBO ML and biology tasks (Dreczkowski et al., 2023)						
`xgb_mnist` (*HPO*)	8	4	3	1	0	Mixed-variable
`rna_fold` (*Applied science and Engineering design*)	40	0	0	40	0	Combinatorial (fully categorical).
EPFL logic-synthesis circuits *(Applied science and Engineering design)* (Amarú et al., 2015; Dreczkowski et al., 2023)						
`aig_i2c`	56	0	20	24	12	Combinatorial.
`aig_int2float`	56	0	20	24	12	Combinatorial.
`aig_log2`	56	0	20	24	12	Combinatorial.
`aig_max`	56	0	20	24	12	Combinatorial.
`aig_sqrt`	56	0	20	24	12	Combinatorial.
Firefly structural design *(Applied science and Engineering design)* (Gandomi et al., 2011)						
`CarMixed`	11	9	2	0	0	Mixed
`PressureVesselMixed`	4	2	0	0	2	Mixed-variable
`ReinforcedConcreteBeam`	3	1	0	0	2	Mixed-variable
Beam with section catalogue *(Applied science and Engineering design)* (Cuesta Ramirez et al., 2022)						
`EulerBeamMixed`	3	2	0	0	1	Mixed-variable
CIFAR-10 sparse attack *(Applied science and Engineering design)* (Wan et al., 2021)						
`cifar_attack85`	85	42	0	1	42	Mixed-variable
Cellular coverage (CCO) *(Applied science and Engineering design)* (Dreifuerst et al., 2021; Daulton et al., 2022)						
`cco30`	30	15	0	15	0	Mixed-variable
Portfolio selection *(Applied science and Engineering design)* (Chang et al., 2000)						
`portfolio_port5`	450	225	225	0	0	Mixed-variable
Thinned and phased antenna array *(Applied science and Engineering design)* (Haupt, 1994)						
`antenna200`	200	100	100	0	0	Mixed-variable
Reactive power dispatch, IEEE 118-bus *(Applied science and Engineering design)* (Thurner et al., 2018)						
`orpd118`	77	54	14	0	9	Mixed-variable
BikeBench bicycle design *(Applied science and Engineering design)* (Regenwetter et al., 2026)						
`bikebench64`	64	48	3	12	1	Mixed-variable

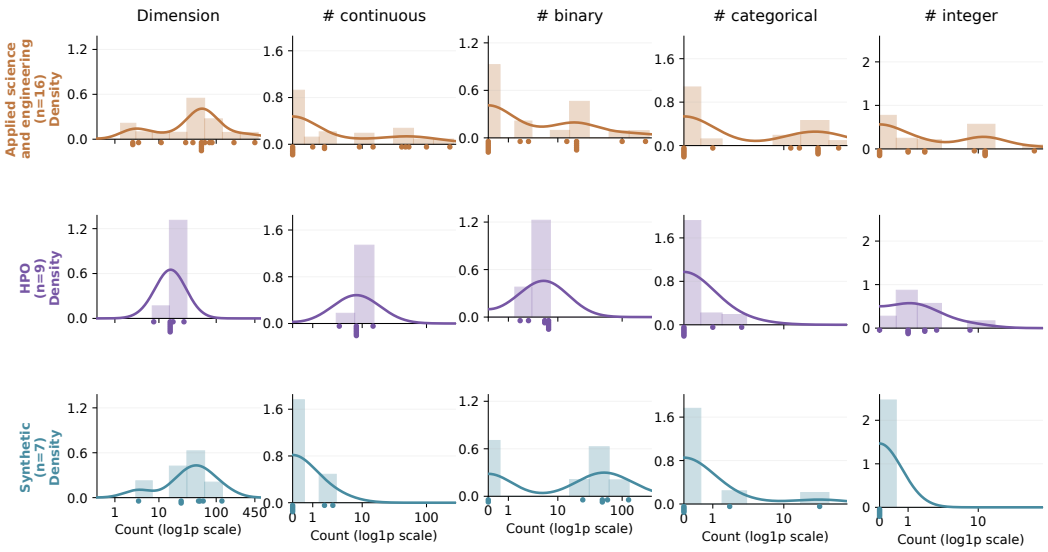

Figure 11: Composition of the 32-benchmark suite: applied science and engineering (16 problems), HPO (9), and synthetic benchmarks (7). Columns show total dimension and counts of continuous, binary, categorical, and integer variables. Histograms and smooth density curves summarize each group; dots mark individual problems. Horizontal axes use a $\log(1 + \text{count})$ scale. Engineering problems span the broadest range of dimensions and variable types.

H EDGE Algorithm Implementation Details

The EDGE algorithm was implemented using these packages: Python 3.11.15, torch 2.6.0+cu124, botorch 0.18.0, gpytorch 1.15.2, numpy 1.26.4, and scipy 1.12.0. Here are the sources of all surrogates we used in the EDGE surrogate ablation experiment (Section B.1):

1. The TabPFN-3 model checkpoint is `tabpfn-v3-regressor-v3_default.ckpt` with library version v8.0.6. (Grinsztajn et al., 2026) (link: `https://huggingface.co/Prior-Labs/tabpfn_3`, license: TABPFN-3 License v1.0, last accessed: Sep 24th, 2026)
2. The TabICL is using the checkpoint and code form v2.1.1. (QU et al., 2026) (link: `https://github.com/soda-inria/tabicl` , license: BSD 3-Clause License, last accessed: Sep 24th, 2026)
3. The SAAS model is the `SaasFullyBayesianSingleTaskGP` model from BoTorch. (Eriksson & Jankowiak, 2021) (link: `https://botorch.readthedocs.io/en/latest/models.html#botorch.models.fully_bayesian.SaasFullyBayesianSingleTaskGP`, license: MIT license, last accessed: Sep 24th, 2026)