WHEN NOT TO TRUST A GLOBAL GAUSSIAN PROCESS POSTERIOR: LOCAL SCORING AND PROGRESSIVE DISSONANCE FOR MULTI-MODAL SCIENTIFIC DISCOVERY

Anonymous authors
Paper under double-blind review

ABSTRACT

Gaussian-process Bayesian optimization (GP-BO) is a standard approach to budget-constrained black-box optimization, but its acquisition decisions rely on the quality of a global GP posterior. We study when this reliance becomes unreliable in scientific discovery problems with phase boundaries, feasibility gaps, threshold-derived responses, and disconnected high-value regions. In these settings, a globally smooth posterior can be miscalibrated precisely in high-response regions: on LGPS, nominal 95% GP coverage drops from 0.887 overall to 0.403 in the top response decile. We introduce two alternatives that reduce dependence on global posterior ranking: **ProSe**, which combines local k-nearest-neighbor scoring with spatial anti-clustering, and **ATLAS**, which searches for progressive disagreement with a hierarchy of simple models while probing spatial frontiers. Our results show a regime split rather than universal superiority. On a smooth synthetic control, several GP methods and random search attain Recall@100 — the fraction of distinct target peaks recovered within 100 evaluations — of 1.00. In contrast, on a strict LGPS compositional-family holdout, ATLAS achieves Recall@50/100 of $0.53/0.63$ and reaches successful discoveries in a median of 23 evaluations versus 41 for MAP-Elites, though its advantage over MAP-Elites is directional and borderline after Holm correction ($p = 0.064$). On discriminative steel search, ATLAS and SAGE — a k-NN-scored baseline that anchors part of its candidates on observed points — each significantly outperform eight of ten alternative GP-based acquisition, quality-diversity, and randomized-search baselines after Holm correction. Finally, an audit of five look-up benchmarks constructed from public materials datasets finds that four show no separation among methods under our protocol, due to saturation, positional-encoding collapse, or random-search matching. These results suggest using global posterior acquisition on smooth, calibrated landscapes, and favoring locality, dissonance, and diversity when its modeling assumptions are violated.

1 INTRODUCTION

Sequential black-box optimization under limited evaluation budgets is central to scientific discovery, engineering design, and hyperparameter tuning (1; 2). In materials and chemical discovery, a single evaluation can require synthesis, characterization, and substantial experimental time; consequently, an optimizer must identify useful candidates within tens rather than thousands of evaluations. Gaussian-process Bayesian optimization (GP-BO) is a standard approach to this setting (3; 4). Acquisition functions such as Expected Improvement (EI), upper confidence bounds, Thompson sampling, and entropy search rank future evaluations through the GP's predictive mean and uncertainty, and this is highly sample-efficient when the posterior is well calibrated. The practical question we study is therefore not whether GPs are useful, but *when an optimizer should stop relying on a global GP posterior as its dominant signal for deciding where to sample next*.

We frame this question as one of *model trust*. Not every weak BO trajectory indicates that a GP is inappropriate: limited data can destabilize hyperparameter estimation, while high-dimensional

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.

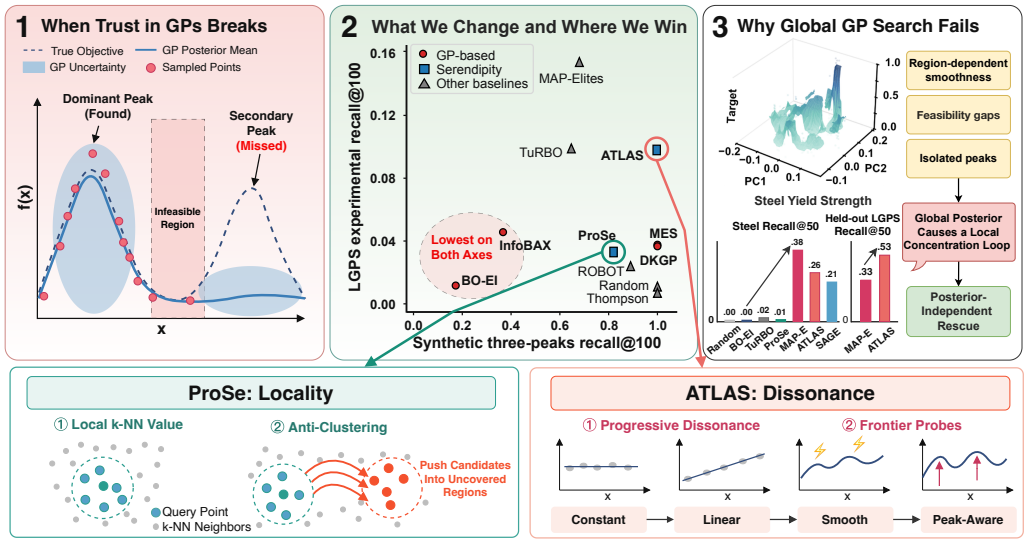


Figure 1: Overview. (1) Trusting a global GP posterior concentrates sampling on the dominant peak and misses a secondary peak behind an infeasible region. (2) Mean multi-peak recall@100 on smooth synthetic three-peaks (x) versus real experimental LGPS (y), 100 repetitions (Table 1): the synthetic axis saturates, the LGPS axis separates methods. (3) Where benchmarks discriminate—steel yield strength (recall@50, Table 6; landscape: element-aligned PC1–PC2, 8-NN surface, Fig. 10) and the held-out LGPS family (recall@50, family block of Table 1)—posterior-independent and quality-diversity search lead. Bottom: ProSe (local k-NN value, anti-clustering) and ATLAS (null-model ladder under dissonance, frontier probes).

geometry can make uncertainty estimates or global exploration ineffective (7; 10). Such difficulties can motivate better priors, initialization, subspace methods, or local trust regions. Our focus is a distinct regime in which the response violates the modeling primitives of a globally smooth surrogate. Scientific objectives may contain phase boundaries, disconnected feasibility regions, inactive gaps, threshold-derived measurements, and heterogeneous local variation. A globally smooth GP can then extrapolate confidently where smooth interpolation is unwarranted. Because acquisition functions inherit the posterior's candidate ranking, this error can create a self-reinforcing concentration pattern: the optimizer further samples an already favored mode while leaving disconnected viable modes under-observed. We treat this mechanism as a diagnostic hypothesis rather than a unique explanation for every GP failure; Figure 2 illustrates the motivating support geometry and local variation.

Rather than introducing another acquisition function built on the same global posterior, we investigate two complementary ways to reduce posterior dependence. **ProSe** (Property-aware Serendipity) is surrogate-free: it combines local k-nearest-neighbor value estimates with K-means anti-clustering for coverage. **ATLAS** (Adaptive Theory-Layered Anomalous Search) identifies locations where observations progressively disagree with a hierarchy of deliberately simple models and uses frontier pressure to distribute queries spatially. ProSe asks what nearby observations suggest; ATLAS asks where the current explanation breaks down.

Our experiments support a regime split rather than universal dominance (Fig. 1). On the controlled synthetic three-peak task, several GP-based and randomized methods saturate Recall@100 (§5.1), so posterior-independent search is not uniformly preferable. On the strict LGPS compositional-family holdout, ATLAS is the top method, though its advantage over MAP-Elites is directional and borderline after correction (§5.2). On the discriminative steel look-up task, MAP-Elites, ATLAS and SAGE each significantly outperform most GP-based and randomized alternatives after Holm correction (§5.3). Thus, our evidence supports reducing global-posterior dependence when responses

have disconnected, threshold-like, or otherwise non-smooth structure, rather than replacing GP-BO everywhere.

Finally, method comparisons are meaningful only when a benchmark can distinguish methods: four of the five public look-up benchmarks we audit fail to support a reliable ranking, through ceiling or floor saturation, positional-encoding collapse, or matching random search (§5.3). We therefore treat benchmark validity as a prerequisite for claims about optimizer quality.

Our contributions are as follows:

1. We characterize when global GP trust breaks: the best GP methods reach Recall@100 of 1.00 on smooth three-peaks, yet global-posterior acquisitions achieve at most 0.07 strict-LGPS family recall by 100 evaluations (BO-EI and DKGP: 0.00), and nominal 95% GP coverage drops from 0.887 overall to 0.403 in the LGPS top decile (§3.2, §5.1).
2. We introduce two complementary alternatives to global-posterior ranking: ProSe combines local k-NN scoring with anti-clustering, and ATLAS combines progressive dissonance with frontier probing (§3–§3.5).
3. We establish a regime split rather than universal superiority: GP-based and randomized methods saturate the smooth control, whereas ATLAS and quality-diversity search lead the discriminative strict-family and steel tasks (§5.1–§5.3).
4. We audit benchmark discriminative power before ranking methods; four of the five public-data look-up benchmarks fail that audit (§5.3, Appendix C).

2 Related Work

2.1 Gaussian-Process Optimization in Scientific Discovery

Gaussian-process (GP) surrogates dominate sample-efficient black-box optimization because they couple a predictive mean with calibrated uncertainty (1; 2; 3). This foundation supports value-based acquisition, posterior sampling, and information-theoretic criteria such as MES and InfoBAX (32; 33). Scientific applications have extended the same paradigm through local trust regions (TuRBO), sparse or expanding subspaces (SAASBO, BAxUS, Bounce), learned kernels, and robust or amortized posteriors (7; 8; 9; 10; 34; 23; 24). These advances primarily improve posterior estimation under high dimension, small data, or heterogeneous noise. They do not remove the optimizer's dependence on a globally meaningful posterior ranking when the response itself is discontinuous, undefined on infeasible regions, or only locally smooth. Our work targets that latter boundary rather than GP optimization in general.

2.2 Diversity-Oriented and Posterior-Independent Search

Several neighboring literatures show that discovery can benefit from objectives other than posterior improvement. Quality-diversity methods preserve solutions across behavioral niches (MAP-Elites), with surrogate-assisted variants improving sample efficiency (SAIL, BOP-Elites); ROBOT and BEACON similarly make diversity or novelty explicit within Bayesian search (13; 14; 25; 26; 11; 16). Space-filling designs, novelty search, and Lipschitz or zooming bandits instead use geometric coverage or local regularity without requiring one global probabilistic model (17; 19; 20; 21). These results motivate our two complementary designs: ProSe couples local non-parametric value estimates with anti-clustering, whereas ATLAS combines progressive null-model failure with frontier coverage. Both retain directionality while reducing the authority of a single global GP posterior.

2.3 Scientific Response Landscapes and Benchmark Validity

Scientific optimization differs from generic smooth-function benchmarks in ways that couple representation, feasibility, and response geometry. Composition spaces are constrained simplices; experimental campaigns sample them non-uniformly; phase boundaries create abrupt changes; failed synthesis leaves gaps; and threshold-derived measurements can flatten or discretize the target (6; 31; 5). Materials studies have accordingly explored memory-based initialization, quality diversity, and non-GP surrogates, while benchmarking work finds that model choice and representation can matter as much as acquisition design (12; 15; 27). Mixed-variable BO addresses discontinuity in the search

domain (38); our focus is response irregularity after the domain has already been made continuous. This distinction also makes benchmark auditing essential: look-up encodings can collapse distinct materials, and relative acceleration can increase even when all methods are poor in absolute terms (41). We therefore evaluate smooth controls and real scientific regimes separately, and report when a benchmark cannot support a method ranking. HEBO, PFNs4BO, SAIL, and BOP-Elites remain plausible but untested alternatives; Appendix I states these comparison boundaries and Appendix A gives the component-level positioning.

3 METHOD

3.1 PROBLEM SETTING

Appendix A gives a pipeline diagram (Fig. 4).

We consider sequential optimization of an unknown function $f : \mathcal{X} \to \mathbb{R}$ over a bounded domain $\mathcal{X} \subseteq \mathbb{R}^d$ with evaluation budget N, starting from N_{init} initial random samples. Our primary domain is the compositional simplex $\mathcal{X} = \{\mathbf{x} \in \mathbb{R}^d : x_i \geq 0, \sum_i x_i = 1\}$, though the method applies to any bounded domain with a distance metric.

3.2 A MODEL-TRUST VIEW OF ACQUISITION DESIGN

We view acquisition design as a question of how much the optimizer should trust a fitted posterior, which separates three regimes that are often conflated. In the *calibrated-posterior regime* the model class is appropriate and posterior uncertainty ranks candidates reliably, so GP acquisition functions (EI, UCB, MES, Thompson sampling) are natural. In the *small-data calibration regime* the model class may be appropriate but hyperparameters, priors or initial coverage are unstable, and better initialization, trust regions or richer priors can help. In the *modeling-primitives violation regime* the objective violates the assumptions that make global posterior uncertainty meaningful—discontinuities, feasibility gaps, threshold-derived measurements or region-dependent smoothness—so a globally smooth GP posterior is confidently wrong in exactly the regions an optimizer should inspect.

Informal proposition. Consider an acquisition score $a_t(\mathbf{x})$ whose ordering is dominated by a globally fitted posterior score $q_t(\mathbf{x})$ (for example posterior mean, variance, entropy, or a deterministic transformation of them). If q_t is miscalibrated outside a data-rich region, then maximizing a_t can inherit the same concentration bias: samples improve the posterior where it is already trusted while leaving secondary modes under-observed. A local or progressive score avoids this dependence when its candidate ranking is not dominated by the global posterior. ProSe does this by replacing q_t with nearest-neighbor value estimates plus coverage pressure; ATLAS does this by ranking points according to dissonance from a hierarchy of deliberately simple null models.

This is not a regret theorem: formal regret would need assumptions on mode separation, noise, smoothness and the observation process that our experimental settings do not satisfy. The point is diagnostic—when the posterior is calibrated, uncertainty-aware GP methods should be strong; when the modeling primitive itself fails, replacing posterior trust can beat refining the acquisition function—and it places ProSe closest to Lipschitz and non-parametric bandits, ATLAS closest to model selection and anomaly-driven experimental design.

3.3 PROSE: PROPERTY-AWARE SERENDIPITY

At each step t with n observations:

Step 1: Region identification. K-means clustering ($k = \min(4, \lfloor n/2 \rfloor)$) on observed points identifies explored regions with centers $\{c_j\}$.

Step 2: Distance scoring. For each of M random candidates $\mathbf{x} \sim \text{Uniform}(\mathcal{X})$:

$$d(\mathbf{x}) = w_c \cdot \min_j \|\mathbf{x} - c_j\| + (1 - w_c) \cdot \min_i \|\mathbf{x} - \mathbf{x}_i\| \tag{1}$$

where $w_c = 0.7 \cdot 0.90^{0.05t}$ decays from cluster-center avoidance to point-level avoidance.

Step 3: Value-aware blending. An inverse-distance-weighted k-NN estimate (k=3) predicts the function value at each candidate:

$$\hat{f}(\mathbf{x}) = \frac{\sum_{j \in \text{kNN}} w_j \cdot f(\mathbf{x}_j)}{\sum_j w_j}, \quad w_j = \frac{1}{\|\mathbf{x} - \mathbf{x}_j\| + \epsilon} \tag{2}$$

The final score blends distance and value:

$$s(\mathbf{x}) = (1 - w_p) \cdot d(\mathbf{x}) + w_p \cdot \hat{f}_{\text{norm}}(\mathbf{x}) \tag{3}$$

where ϵ=10^{-6}, $\hat{f}_{\text{norm}}$ is $\hat{f}$ min–max normalised over the candidate pool, and $w_p = 0.3 \cdot (0.5 + 0.5 \cdot t/N)$ gradually increases value influence. The full pseudocode is given in Appendix B. **SAGE** is ProSe with the same scoring rule but a different candidate pool and metric: 60% of the M candidates are anchored on observations (35% local jitter, 25% sparse swaps) and 40% are uniform, and distances switch to CLR (Aitchison) geometry when the observations are sparse or concentrated (Appendix E).

3.4 Why ProSe Breaks the Concentration Loop

The key insight is the difference between *global* and *local* value information. A GP surrogate fits a single global model; when one mode has more data, the GP's predictions are dominated by that mode, creating the self-reinforcing loop. The k-NN estimator (k=3) consults only three nearby observations, which may come from *different* modes depending on the candidate's location. This locality ensures each mode's influence is proportional to its proximity, not its data volume—breaking the concentration bias at its root.

3.5 ATLAS: Progressive Dissonance-Driven Exploration

While ProSe breaks the concentration loop via locality, ATLAS takes a complementary model-theoretic approach inspired by scientific discovery: build the simplest possible model, find where it fails, investigate those failures, and refine.

Four-level null model hierarchy. ATLAS maintains a hierarchy of models with increasing complexity:

- **Level 0 (Constant):** $M_0(\mathbf{x}) = \bar{y}$ — detects high/low-value regions.
- **Level 1 (Linear):** $M_1(\mathbf{x}) = \mathbf{w}^\top \mathbf{x} + b$ — detects non-linear interactions.
- **Level 2 (Smooth):** $M_2(\mathbf{x})$ = KernelRidge(RBF, ℓ=0.2) — detects phase boundaries and discontinuities.
- **Level 3 (Peak-Aware):** $M_3(\mathbf{x}) = M_2(\mathbf{x}) + \sum_k G_k(\mathbf{x})$ with Gaussian peaks — detects remaining unknown peaks.

Dissonance and level transitions. At each step, ATLAS computes the *dissonance* $D(\mathbf{x}) = |f(\mathbf{x}) - M_\ell(\mathbf{x})|/\sigma_{\text{residual}}$ for each observation. A level transition fires when at least two of the last ten dissonances exceed the median plus one standard deviation computed over all current dissonances, subject to a five-step cooldown.

Candidate generation. The current ATLAS default uses compute-fair configurable channels: dissonance probes (30%, near high-dissonance points), frontier probes (45%, farthest-point sampling), and random Dirichlet proposals (25%). The value-guided local channel is off in the single configuration used for every experiment in this paper; it was fixed once, after early development runs, for the discovery objective (a withheld family need not be high-value under the observed target), and Appendix E reports the sensitivity to switching it back on. Sparse-swap and low-value local probes are retained as optional ablations. All candidates are scored by a weighted combination:

$$s(\mathbf{x}) = w_d \cdot \hat{D}(\mathbf{x}) + w_v \cdot \hat{f}(\mathbf{x}) + w_f \cdot d_{\text{frontier}}(\mathbf{x}) - w_\rho \cdot \text{density}(\mathbf{x}) \tag{4}$$

where $\hat{D}$ and $\hat{f}$ are k-NN estimates (k=3) of dissonance and value respectively. The value coefficient increases modestly as the budget is spent, but dissonance and frontier coverage remain the dominant positive terms throughout; Appendix E gives the exact schedules and normalization.

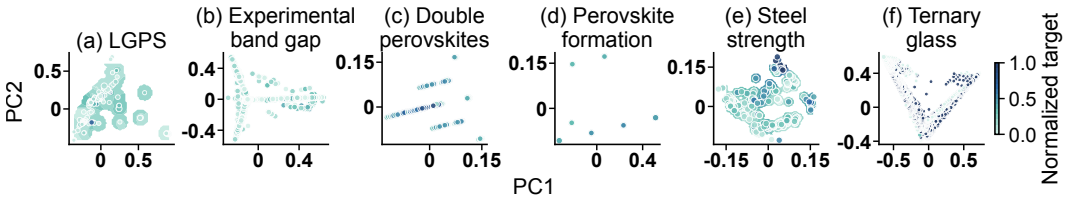


Figure 2: Projected response landscapes for the six scientific datasets after element alignment. Points mark observed compositions, and color shows the normalized target; construction details and 3-D views are in Appendix C (Fig. 10).

Why ATLAS avoids concentration. The progressive null model ensures that ATLAS targets *different* landscape features at each level: Level 0 finds trends, Level 1 finds non-linearities, Level 2 finds boundaries, and Level 3 finds unknown peaks. This multi-scale anomaly detection naturally distributes samples across the landscape.

3.6 Connections to Non-Parametric Optimization

ProSe is closest to Lipschitz bandit optimization (21): the k-NN estimate uses only local information, while K-means distance scoring acts like a data-adaptive discretization that keeps sampling pressure away from already-covered regions. We use this connection only as intuition, not as a formal guarantee.

4 Experiments

4.1 Baselines

BO-EI: GP (RBF + WhiteKernel) with Expected Improvement, $M{=}1000$ candidates. **TuRBO** (7): Local trust region BO with adaptive sizing. **ROBOT** (11): BO with diversity-constrained portfolio. **Thompson Sampling**: one independent draw from the GP's marginal posterior at each candidate (the joint covariance is not sampled), selecting the arg-max. **MES** (32): Max-value Entropy Search via BoTorch's `qMaxValueEntropy`. **InfoBAX** (level-set) (33): Bayesian Algorithm Execution targeting the super-level set $A(f) = \{\mathbf{x} : f(\mathbf{x}) \geq \tau\}$. **DKGP** (34): Deep-Kernel GP with an MLP feature extractor and EI acquisition on the learned embedding.

MAP-Elites (13): quality-diversity archive over a 2-D PCA behaviour descriptor fitted on the initial design; each step mutates an archive elite (Appendix E). **Random**: Uniform sampling (lower bound). All baselines use standard configurations from their original publications; full settings in Appendix E. Every GP baseline and ProSe maximize their acquisition score over a pool of $M{=}1000$ uniform Dirichlet candidates drawn afresh at each step, not by continuous optimization over the simplex; TuRBO draws its pool inside the trust region, and MAP-Elites, SAGE and ATLAS generate candidates near observations (Appendix C, candidate-pool control).

4.2 Experimental Data: Kanno et al. LGPS

To validate on real experimental data, we use Kanno et al.'s real experimental LGPS dataset: 302 wet-lab synthesis trials (299 within the modelled seven-element Li–S–P–Si–Sn–Ge–Ga space) with XRD-derived target (31; 6), spanning the Li_2S–P_2S_5–MS_2–Ga_2S_3 precursor space; the property is LGPS phase purity (a proxy for ionic conductivity). Each method is evaluated via virtual experiment (Table 1: the controlled-screen block uses the peak hold-out, $N{=}100$, 100 reps, and the strict-family block uses the family hold-out, 30 paired reps, 200 for the follow-up ATLAS–MAP-Elites comparison). The retrospective family target is small—three trials at two distinct compositions—and lies in the lower part of the observed target distribution. The protocol shares paired initial designs across methods and counts a hit only when the queried composition's nearest reference member belongs to that held-out family, so nearby train-region queries are not counted as new-family discoveries.

Table 1: **Evidence map across three regimes.** GP and randomized baselines saturate smooth controls; ATLAS leads strict-family recall; steel provides clear separation among public look-ups. Evaluated over 100/30/100 runs (look-up: recall@50). Bold indicates column maxima; superscripts count the other methods beaten after Holm correction; gray highlights benchmark saturation. Pairwise statistics are in Secs. 5.1–5.3. Method taxonomy follows Fig. 1.

	Controlled screen				Strict LGPS family holdout			Public look-up: recall@50				
	Synthetic 3-peaks		Experimental LGPS		Family recall			Steel	Band gap	Dbl. perov.	Glass	Perov.
Method	**R@100**	**Hit**	**R@100**	**Hit**	**R@20**	**R@50**	**R@100**	**Top-5% recall ↑**				
GP-based: global posterior acquisition												
BO-EI	.173	.000	.011	.000	.00	.00	.00	.00	.25	.20	1.00	0.00
InfoBAX	.367	.000	.045	.001	–	–	–	.00	.43	.20	1.00	0.00
DKGP	**1.000**	.030	.036	.002	.00	.00	.00	.00	.97[5]	.18	1.00	0.00
MES	**1.000**	**.380**	.037	.002	.03	.07	.07	.00	**.98**[6]	.26[2]	1.00	0.00
GP-based: refinements (trust region / posterior sampling / diversity portfolio)												
TuRBO	.647	.063	.099	.008	**.23**	.23	.27	.02	.58[1]	.15	1.00	0.00
ROBOT	.890	.143	.024	.000	–	–	–	.01	.54[1]	.13	1.00	0.00
Thompson	**1.000**	.020	.007	.000	–	–	–	–	–	–	–	–
Other baselines (quality-diversity and randomized)												
MAP-Elites	.680	.020	**.154**	**.009**	**.23**	.33	.57	**.38**[8]	.84[5]	**.50**[9]	1.00	0.00
Random	**1.000**	.013	.011	.000	.00	.03	.03	.00	**.98**[6]	.29[3]	1.00	0.00
Ours												
ProSe	.820	.000	.033	.000	.00	.00	.00	.01	.96[5]	.14	1.00	0.00
ATLAS	.997	.090	.098	**.009**	**.23**	**.53**	**.63**	.26[8]	.95[5]	.32[4]	1.00	0.00
SAGE	–	–	.113	.004	.07	.20	.33	.21[8]	.59[1]	.14	1.00	0.00

Reading guide. GP-based methods are those whose decisions are driven by a fitted GP posterior; the refinements subgroup (TuRBO's trust region, ROBOT's diversity portfolio, Thompson's marginal posterior sampling) modifies how that posterior is used. R = recall and Hit = threshold-hit rate (τ=0.7). The screen and look-up sweeps used different 11-method rosters; dashes are unrun cells, not zeros (SAGE: LGPS only, same paired initial designs as ProSe). Strict-family medians among successful runs were 23 steps for ATLAS and 41 for MAP-Elites. Glass and perovskite formation are respectively ceiling- and floor-saturated; double perovskites collapses to 16 coordinates.

4.3 Synthetic Benchmarks

Five distributions on a quaternary simplex (CMBD: Compositional Multi-modal Benchmark Distributions): **Original** (single peak), **Three Peaks** (multi-modal), **Linear Gradient**, **Step-Like** and **Boundary**, Gaussian noise σ=0.03; 30 paired-seed repetitions per method–distribution pair, N=50, N_{init}=5 (auxiliary Table 10, means only), while the extended benchmark of Table 1 uses N=100 and 100 repetitions. Every significance statement in the paper uses paired one-sided sign tests with Holm correction, since all runs share paired seeds.

4.4 Real-World Benchmarks

Five publicly available datasets treated as black-box look-up benchmarks via k-NN interpolation (k=3, noise σ=0.02): experimental band gaps (29) (4,604 compounds), double perovskite band gaps (30) (1,306), perovskite formation energies (30) (18,824), steel yield strengths (312 alloys) and ternary glass-forming ability (7,191), each at **100 repetitions** under a top-5% hold-out protocol with the full roster; four of the five cannot support a per-method claim, largely because our positional sorted-fraction encoding merges chemically distinct formulas (Appendix C).

Scope controls. Appendix G reports peak-count scaling and standard Branin/Hartmann controls beyond compositional domains.

5 Results

5.1 Two-Axis Regime Split

We first test the model-trust thesis against uncertainty-aware Bayesian baselines on two regimes—a smooth synthetic three-peaks landscape and Kanno et al.'s real experimental LGPS dataset in its 7-D representation—scoring **multi_peak_recall@100** and **threshold_hit_rate** (τ=0.7).

Table 1 (Fig. 1, panel 2) reveals a *regime split*, but not a universal ranking. On smooth synthetic three-peaks, the recall metric saturates: MES, DKGP, Random, and Thompson all reach 1.000, with ATLAS nearly saturated and ProSe below the saturated group. The threshold metric also favors MES, so this regime should be read as a case where posterior-based or broad randomized exploration is already reliable. On value-oriented real experimental LGPS recall, the ranking changes: MAP-Elites leads, while SAGE, TuRBO, and ATLAS are the next strongest group and all are above the evaluated global-GP refinements, whereas ProSe (0.033) sits with the global-GP group. Paired sign tests support ATLAS over MES ($p_{\text{Holm}} = 4.0\times10^{-5}$) and TuRBO ($p_{\text{Holm}} = 0.017$), and SAGE over DKGP ($p_{\text{Holm}} = 0.029$) and InfoBAX ($p_{\text{Holm}} = 1.0\times10^{-9}$); SAGE versus MES is directionally positive but borderline after correction ($p_{\text{Holm}} = 0.068$). We therefore read this table as an assumption-violation screen rather than a ranking. A group-held-out calibration audit supports this reading mechanistically: 95% coverage falls from .887 overall to .403 in the LGPS top decile (.865 to .691 on steel; Fig. 3), so the failures concentrate exactly where exploration decisions matter. The effect is not variance shrinkage alone: top-decile residuals show $\mathbb{E}[z] = 2.85$ (LGPS) and 1.82 (steel), i.e. systematic underprediction of high responses on top of underdispersed intervals (Appendix D.1).

5.2 Strict LGPS Family Holdout

To isolate true experimental behavior from look-up interpolation, we also evaluate a strict family holdout on Kanno et al.'s LGPS dataset (31), whose genuine experimental noise, non-uniform coverage and highly multi-modal phase-purity landscape make it the paper's hardest case. Unlike the value-oriented table above, this protocol asks a discovery question: after hiding the LGPS family, how often does a method query a composition whose nearest reference member is from that held-out family?

The family block of Table 1 is the primary new-phase evidence. In the full screen, ATLAS is top at budgets 50 and 100 and reaches the hidden family earlier than MAP-Elites (median step 23 vs 41 among successful budget-100 runs). Because the 30-repetition screen is underpowered after correction, we ran a follow-up 200-repetition paired ATLAS–MAP-Elites comparison: ATLAS wins more matched repetitions at recall@50 (46 vs 29) and recall@100 (43 vs 27), one-sided sign-test $p = 0.032$ and $p = 0.036$ (unchanged by that run's within-budget Holm correction; 0.064 for both when the two nested budgets form one family, Appendix G). This is directional evidence favoring ATLAS, rather than a corrected claim of superiority; it does not extend to ProSe, and under the same seeds on Linux ATLAS and MAP-Elites tie at recall@100 $= 0.47$ (Appendix C), so the ordering is not established platform-independently.

5.3 Real Look-up Benchmarks and Benchmark Discriminative Power

We also run the full eleven-method roster on five look-up benchmarks built from public real-materials datasets at 100 repetitions under a top-5% hold-out protocol (Appendix C, Table 6). The first finding concerns the benchmarks, not the methods: *four of the five cannot support a per-method claim.* Ternary glass-forming ability saturates at recall 1.00 for all eleven methods and perovskite formation energies at 0.00 for all eleven; on band gaps random search ties MES for the best score (0.98). A representation audit traces much of this to our positional sorted-fraction encoding—a common way to place variable-element formulas on one fixed simplex—which merges chemically distinct formulas: 1,306 double perovskites occupy 16 distinct points, 18,824 perovskites 11, and 4,604 band-gap compounds 1,342. Double perovskites is therefore degenerate as well; element-aligned re-encoding reverses the other three's ceiling/floor regimes and scrambles the method ranking, and a row-matched six-element control attributes much of that 55–81-D degradation to ambient dimension rather than to coordinate semantics, consistent with the advantage of near-data proposers. We report discriminative power before any per-method claim.

On the one benchmark that discriminates, steel yield strength, the family-level ordering of Table 1 reappears with more power and separates our two variants sharply: SAGE (0.21), ATLAS (0.26) and MAP-Elites (0.38) each beat *eight* of the other ten methods after Holm correction—every global-GP and randomized baseline, and ProSe—with five at exactly zero; the ordering survives one-column-per-element re-encoding and the anchored-pool, CLR-coordinate, random-forest-surrogate

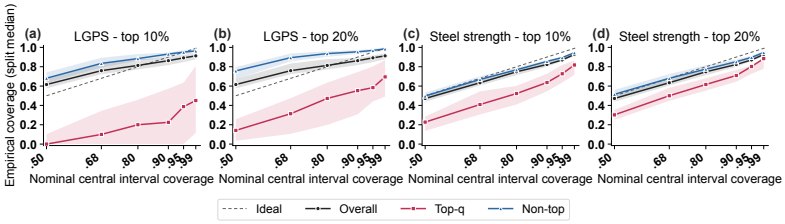


Figure 3: **Optimization-conditioned reliability for LGPS and steel.** Top-q regions ($q \in \{10\%, 20\%\}$, red) drop severely below both the ideal diagonal and non-top baselines, revealing that global evaluation masks severe miscalibration concentrated in the high-value optimization tail. Split medians across 20 held-out runs (shaded: IQR). Extended six-dataset results are in Appendix D.1.

and budget-500 controls of Appendix C. Quality-diversity and posterior-independent search therefore lead where the benchmark can tell methods apart, but MAP-Elites tops it.

Our two variants sit at opposite ends of a coverage–locality trade-off, and the benchmark decides which wins: SAGE leads by 0.20 on steel, where random scores 0.00; ProSe leads by 0.37 on band gaps, where random already scores 0.98 and the encoding merges most compounds—a benchmark property. The steel gap exceeds most between-family gaps in Table 6 (Fig. 11, Appendix C).

6 Discussion

Two complementary paradigms. ProSe and ATLAS break global model dominance via locality and via progressive modeling with frontier pressure; the results locate where this matters: the real-data assumption-violation screen, not saturated synthetic recall. ProSe remains competitive on three-peaks (0.82) and band gap (0.96), but neither supports a ProSe-specific advantage (the former saturates; Random reaches 0.98 on the latter); on experimental LGPS, true ProSe reaches $0.033/0.000$ on the screen and $0/30$ on the strict family, whereas SAGE supplies the previously reported LGPS successes and partially overlaps ATLAS (Appendix A). The strict-family evidence therefore supports ATLAS rather than ProSe, subject to the oracle limitation below.

GP assumption violations on real experimental data. Phase boundaries can produce jump discontinuities, synthesis-feasibility gaps leave the target undefined, XRD-derived scores show threshold behavior, and smoothness varies across the simplex (Fig. 2). The conditioned-calibration audit (Fig. 3; details in Appendix D.1) localizes a corresponding failure in the high-response tail. Together with the optimizer rankings, this is consistent with posterior miscalibration limiting the evaluated GP methods. It does not establish calibration error as the causal mediator or uniquely attribute it to discontinuity rather than small n, heteroscedasticity, or omitted descriptors.

Limitations. The strongest real-experimental evidence is one small retrospective LGPS family hold-out (three trials at two compositions), and ATLAS's fixed configuration was developed with that objective in view; broader prospective validation is needed. The public look-up results depend on a k-NN oracle and our encodings; smooth and high-dimensional controls can saturate or collapse toward random performance, so these diagnose benchmark limits rather than universal rankings. Neither method has a regret guarantee, key schedules are hand-tuned (Appendices E, C), and GP acquisition maximization is pool-based ($M{=}1000$ candidates per step), under-approximating the true optimum as dimension grows. Finally, the LGPS oracle's tie-break at repeated compositions is platform-dependent (Appendix C).

7 Conclusion

We demonstrate that reducing global-GP dependence, through local scoring (ProSe, SAGE) or ATLAS's progressive dissonance, can rescue optimization when posterior trust breaks down. Conditioned calibration audits and the observed regime split agree: GP methods remain strong on smooth controls, while posterior-independent methods excel on irregular, real-world data. Acquisition design must reflect landscape realities: trust the GP posterior when its assumptions hold, and abandon it when they do not.

REPRODUCIBILITY STATEMENT

Pseudocode, baseline configurations, k-NN ablations, auxiliary benchmark tables, dataset distinctions and a speed-to-target variant (Appendix J) are provided in the appendix. Every method, hold-out protocol, budget grid, repetition count and statistical test is specified in enough detail to be reimplemented from the paper and appendix alone. The LGPS look-up oracle is platform-dependent at repeated compositions: the LGPS values were produced on macOS/arm64 (package versions at the time not recorded) and are regenerated bit-for-bit by the current macOS stack (numpy 2.4.6, scikit-learn 1.8.0); a same-seed Linux re-run, which gives ATLAS 0.47 and MAP-Elites 0.47 instead of 0.63 and 0.57, is reported in Appendix C. The implementation itself—source code, unit and integration tests, the retrospective benchmark runner, the paired-statistics code, and instructions for both quick smoke benchmarks and the longer runs behind the controlled-screen and family blocks of Table 1—is packaged and ready, but is not attached to this submission because it has not yet cleared our institution's code-release process. We will make it available to reviewers on request through an anonymous channel during the discussion period and, subject to institutional approval, publicly on publication.

AI USE STATEMENT

We disclose our use of large language models (LLMs) specifically rather than generically, as required. LLM assistants were used through an agentic coding interface across the full lifetime of this project, from initial method prototyping to final manuscript preparation; the majority of commits in our repository carry co-authorship trailers recording this, so the extent of assistance is auditable from the commit history, which accompanies the code when it is released (see the Reproducibility Statement).

Uses in categories that require disclosure.

- *Refining hypotheses.* The model-trust framing of acquisition design (§3) and the separation of the three GP failure modes—in particular the distinction between sample-complexity limits, small-data calibration failure, and modeling-primitives violation—were sharpened through dialogue with LLM assistants.
- *Designing experiments.* The retrospective hold-out protocols (peak hold-out, top-5% hold-out, LGPS family hold-out), repetition counts, budget grids, and the choice of paired sign tests with Holm correction were developed with LLM assistance.
- *Implementing methods.* ProSe, ATLAS, the baseline implementations (MES, InfoBAX, Deep-Kernel GP, TuRBO, ROBOT, MAP-Elites) and the control variants of Appendix C (anchored-pool, CLR and random-forest baselines), the retrospective benchmark runner, and the paired-statistics code were written with LLM assistance under a test-first workflow. Unit and integration tests accompany the implementation.
- *Data cleaning.* Preparation of the LGPS dataset and the public look-up benchmarks, including removal of identifying metadata columns from the tables prepared for release.
- *Interpreting results.* The two-axis regime-split reading of the benchmark results (§5.1) and the benchmark-saturation interpretation of the high-dimensional results were developed with LLM assistance.

Uses in categories where disclosure is recommended. Analysis and plotting scripts; figure generation; drafting, condensing, and copy-editing of manuscript text; literature search and related-work triage.

Categories that require disclosure and do not apply. No generative AI tool was used to generate any dataset: the synthetic benchmark values, look-up tables, and LGPS measurements all come from executed code or laboratory records, not from model output. No mathematical claims or proofs are made in this paper, so the corresponding categories are not applicable; no translation or qualitative/thematic data analysis was performed.

Author responsibility and verification. The authors take full responsibility for the contents of this submission. Every numerical claim in the paper was re-verified by the authors directly against the stored raw result files rather than against LLM-generated summaries, and every citation was checked against the cited work. LLM assistants were not used to generate experimental data: all reported numbers come from executed code in our benchmark implementation. Where LLM-proposed analyses did not survive verification, they were discarded rather than reported.

REFERENCES

[1] B. Shahriari *et al.*, "Taking the human out of the loop," *Proc. IEEE*, 2016, **104**, 148.

[2] P. I. Frazier, "A tutorial on Bayesian optimization," *arXiv:1807.02811*, 2018.

[3] C. E. Rasmussen and C. K. I. Williams, *Gaussian Processes for Machine Learning*, MIT Press, 2006.

[4] J. Snoek, H. Larochelle, and R. P. Adams, "Practical Bayesian optimization of machine learning algorithms," in *NeurIPS*, 2012.

[5] Y. Wu, A. Walsh and A. M. Ganose, *Digital Discovery*, 2024, **3**, 1086.

[6] N. Kamaya *et al.*, *Nat. Mater.*, 2011, **10**, 682.

[7] D. Eriksson *et al.*, "Scalable global optimization via local Bayesian optimization," in *NeurIPS*, 2019.

[8] D. Eriksson and M. Jankowiak, "High-dimensional Bayesian optimization with sparse axis-aligned subspaces," in *UAI*, 2021.

[9] L. Papenmeier, L. Nardi and M. Poloczek, "Increasing the scope as you learn," in *NeurIPS*, 2022.

[10] L. Papenmeier, L. Nardi and M. Poloczek, "Bounce: Reliable high-dimensional Bayesian optimization," in *NeurIPS*, 2023.

[11] N. Maus *et al.*, "Discovering diverse solutions in deep reinforcement learning by maximizing expected novelty," in *AISTATS*, 2023.

[12] A. E. Siemenn *et al.*, *npj Comput. Mater.*, 2023, **9**, 80.

[13] J.-B. Mouret and J. Clune, "Illuminating search spaces by mapping elites," *arXiv:1504.04909*, 2015.

[14] J. K. Pugh, L. B. Soros and K. O. Stanley, *Front. Robot. AI*, 2016, **3**, 40.

[15] J. Verhellen and J. Van den Abeele, *Chem. Sci.*, 2020, **11**, 11485.

[16] W.-T. Tang, A. Chakrabarty and J. A. Paulson, *arXiv:2406.03616*, 2024.

[17] K. Crombecq *et al.*, *SIAM J. Sci. Comput.*, 2011, **33**, 1948.

[18] J. Aitchison, "The statistical analysis of compositional data," *J. R. Stat. Soc. B*, 1982, **44**, 139.

[19] V. R. Joseph, E. Gul and S. Ba, *Biometrika*, 2015, **102**, 371.

[20] J. Lehman and K. O. Stanley, "Abandoning objectives: Evolution through the search for novelty alone," *Evol. Comput.*, 2011, **19**, 189.

[21] R. Kleinberg, A. Slivkins and E. Upfal, "Multi-armed bandits in metric spaces," in *STOC*, 2008.

[22] W. Gao, R. Mercado and C. W. Coley, "Sample efficiency matters: A benchmark for practical molecular optimization," in *NeurIPS*, 2022.

[23] A. I. Cowen-Rivers *et al.*, "HEBO: Pushing the limits of sample-efficient hyper-parameter optimisation," *J. Artif. Intell. Res.*, 2022, **74**, 1269.

[24] S. Müller, M. Feurer, N. Hollmann, and F. Hutter, "PFNs4BO: In-context learning for Bayesian optimization," in *ICML*, 2023.

[25] A. Gaier, A. Asteroth, and J.-B. Mouret, "Data-efficient design exploration through surrogate-assisted illumination," *Evol. Comput.*, 2018, **26**, 381.

[26] P. Kent, A. Gaier, J.-B. Mouret, and J. Branke, "BOP-Elites, a Bayesian optimisation approach to quality diversity optimisation with black-box descriptor functions," *IEEE Trans. Evol. Comput.*, 2024.

[27] Q. Liang *et al.*, *npj Comput. Mater.*, 2021, **7**, 188.

[28] F. Hutter, H. H. Hoos, and K. Leyton-Brown, "Sequential model-based optimization for general algorithm configuration," in *LION*, 2011.

[29] A. Dunn *et al.*, *npj Comput. Mater.*, 2020, **6**, 138.

[30] I. E. Castelli *et al.*, *Energy Environ. Sci.*, 2012, **5**, 5814.

[31] S. Hori *et al.*, "Synthesis, structure, and ionic conductivity of solid solution, $Li_{10+\delta}M_{1+\delta}P_{2-\delta}S_{12}$ (M = Si, Sn)," *Faraday Discuss.*, 2014, **176**, 83.

[32] Z. Wang and S. Jegelka, "Max-value entropy search for efficient Bayesian optimization," in *ICML*, 2017, pp. 3627–3635.

[33] W. Neiswanger, K. A. Wang and S. Ermon, "Bayesian algorithm execution: Estimating computable properties of black-box functions using mutual information," in *ICML*, 2021, pp. 8005–8015.

[34] A. G. Wilson, Z. Hu, R. Salakhutdinov and E. P. Xing, "Deep kernel learning," in *AISTATS*, 2016, pp. 370–378.

[35] N. Srinivas, A. Krause, S. M. Kakade and M. Seeger, "Gaussian process optimization in the bandit setting: No regret and experimental design," in *ICML*, 2010, pp. 1015–1022.

[36] A. D. Bull, "Convergence rates of efficient global optimization algorithms," *J. Mach. Learn. Res.*, 2011, **12**, 2879–2904.

[37] D. Eriksson and M. Jankowiak, "High-dimensional Bayesian optimization with sparse axis-aligned subspaces," in *UAI*, 2021, pp. 493–503.

[38] Y. Zhang, T. John, M. Stosiek and P. Rinke, "Bayesian optimization for mixed-variable problems in the natural sciences," *arXiv:2604.07416*, 2026.

[39] J. Meindl, Y. Tian, T. Cui, V. Thost, Z.-W. Hong, J. Chen, W. Matusik and M. Konaković Luković, "SemanticOpt: Towards LLM-based semantic black-box optimization," *arXiv:2510.25404*, 2026.

[40] N. Abhyankar, S. Kabra, S. Desai and C. K. Reddy, "LLEMA: Evolutionary search with LLMs for multi-objective materials discovery," *arXiv:2510.22503*, 2026.

[41] A. D. Adesiji, J. Wang, C.-S. Kuo and K. A. Brown, "Benchmarking self-driving labs," *arXiv:2508.06642*, 2025.

A METHOD COMPARISON

Table 2 summarizes how ProSe and ATLAS differ from the evaluated baselines along five capability axes.

In the LGPS controlled screen at budget 100, SAGE and ATLAS both succeed on 100 matched held-out cases, SAGE alone on 125 and ATLAS alone on 96 (Jaccard 0.312). True ProSe and ATLAS both succeed on 12 of 2,000 matched cases, with 53 ProSe-only and 184 ATLAS-only cases (Jaccard 0.048); because true ProSe's aggregate LGPS recall is low, the latter overlap should not be read as evidence of complementary strength.

Table 2: Comparison of multi-modal optimization approaches.

Method	Surr.-free	Local val.	Anti-clust.	Prog. model	Multi-modal
BO-EI	×	×	×	×	×
TuRBO	×	×	×	×	×
ROBOT	×	×	✓	×	partial
MES	×	×	×	×	partial
InfoBAX	×	×	×	×	partial
DKGP	×	×	×	×	×
MAP-Elites	✓	×	✓	×	✓
LOLA-Voronoi	✓	✓	✓	×	×
Novelty Search	✓	×	✓	×	×
ProSe	✓	✓	✓	×	✓
ATLAS	✓	×	✓	✓	✓

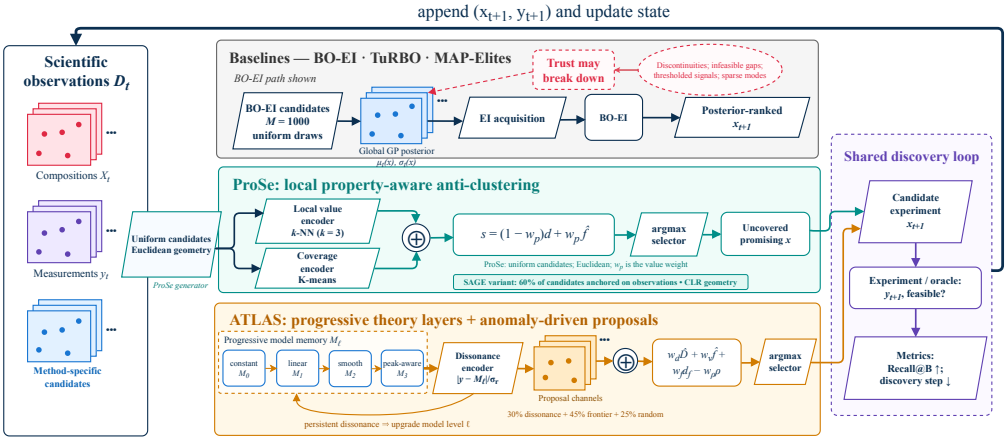


Figure 4: Method pipeline. All methods share the observations, budget, hold-out protocol and evaluation; they differ in how candidates are generated and scored. Top: a GP baseline (BO-EI shown) scores $M{=}1000$ uniform draws with a global posterior, whose trust breaks down on discontinuities, infeasible gaps, thresholded signals and sparse modes. Middle: ProSe scores uniform candidates in Euclidean geometry with a local k-NN value encoder ($k{=}3$) and a K-means coverage encoder, $s = (1 - w_p)\, d + w_p \hat{f}$; SAGE keeps the scoring rule but anchors 60% of the candidates on observations and switches to CLR geometry. Bottom: ATLAS keeps a progressive model memory M_0–M_3 (constant, linear, smooth, peak-aware), scores dissonance $|y - M_\ell|/\sigma_r$, proposes through dissonance (30%), frontier (45%) and random (25%) channels, and upgrades the model level on persistent dissonance.

Algorithm 1 ProSe: Property-Aware Serendipity

Sample N_{init} random points; evaluate f
for $t = N_{\text{init}} + 1, \ldots, N$ **do**
  K-means on observations $\rightarrow$ centers C
  Generate M random candidates
  for each candidate $\mathbf{x}$ **do**
    $d(\mathbf{x}) \leftarrow$ distance score (Eq. 1)
    $\hat{f}(\mathbf{x}) \leftarrow k$-NN value estimate (Eq. 2)
    $s(\mathbf{x}) \leftarrow$ blended score (Eq. 3)
  end for
  Select $\mathbf{x}_t = \arg\max s(\mathbf{x})$; evaluate $f(\mathbf{x}_t)$
end for

B ProSe Pseudocode

C Real-World Look-up Benchmarks

We additionally evaluate the full eleven-method roster on five public materials datasets treated as black-box functions via k-NN interpolation (k=3, noise σ=0.02): experimental band gaps (29) (4,604 compounds), double perovskite band gaps (30) (1,306), perovskite formation energies (30) (18,824), steel yield strengths (312 alloys), and ternary glass-forming ability (7,191). These are look-up benchmarks — the k-NN interpolation preserves the landscape structure of real materials data but does not capture synthesis-process noise, distinguishing them from the Kanno LGPS dataset (Table 1) which uses true wet-lab measurements. The oracle returns the inverse-distance-weighted mean of the k=3 nearest reference rows ($w_j = 1/(d_j + 10^{-6})$) plus $\mathcal{N}(0, 0.02^2)$ noise, clipped to $[0, 1]$; it is continuous near the data but flat far from it, where the three nearest rows no longer change—the plateau on which the GP baselines' best-so-far values are reached (candidate-pool control below)—and at collided coordinates the k=3 set among equidistant rows is whatever the nearest-neighbour search returns first, which is platform-dependent (see *Oracle tie-break and platform reproducibility* below).

Composition encoding and coordinate collisions. Our preparation script (part of the implementation described in the Reproducibility Statement) stores each formula's element fractions sorted alphabetically and written positionally (`frac_0` is the alphabetically first element *of that formula*, not a fixed element), truncated to 3–10 columns. Chemically distinct formulas therefore share coordinates: the 4,604 band-gap compounds occupy 1,342 distinct coordinates, the 1,306 double perovskites 16, the 18,824 perovskites 11, and the 7,191 glasses 2,264. Steel (312 alloys) has no collisions but keeps ten of fourteen elements—V is dropped in 310 alloys (median 0.26% of atoms, at most 4.8%)—and its column meanings shift in 69 alloys whose element sets differ from the majority. Because the look-up value at a coordinate is a k-NN average over whichever colliding formulas the neighbour search returns, double perovskites and perovskites reduce to 16 and 11 effective points, which is the mechanical reason for their degeneracy in Table 6; band-gap and glass collisions likewise inflate what a uniform sample can reach. The LGPS dataset uses element-indexed columns and is unaffected. The encoding is ours, not a property of the source databases; we keep these files as the benchmark actually run and report the audit here rather than silently re-encoding. Fig. 10 shows the element-aligned data of all six datasets as descriptive surfaces.

Element-aligned re-run of steel. Because steel carries the paper's main look-up result, we re-ran the full eleven-method roster on a re-encoded steel table with one column per element (14 columns, no truncation) under the identical protocol and budgets, with seeds drawn independently for each encoding (the runner derives them from the dataset name), alongside the original encoding on the same code. The original encoding reproduces Table 6 to two decimals. Under the aligned encoding the ordering is unchanged and the separation widens: at budget 50, ATLAS reaches recall 0.48, MAP-Elites 0.40 and SAGE 0.14, beating 9, 9 and 8 of the other ten methods after Holm correction (smallest $p = 2 \times 10^{-13}$), while every global-GP and randomized baseline and ProSe score exactly 0.00. ATLAS and MAP-Elites are statistically tied under both encodings; both separate from SAGE

only under the aligned one ($p = 1.2 \times 10^{-6}$ and 8.7×10^{-4}). The Spearman rank correlation of method means between the two encodings is 0.81 at budget 50.

Paired high-element representation sensitivity. We also re-ran band gap, perovskite formation and glass with fixed element columns—all eleven methods, 100 repetitions, budgets 20/50/100—pairing each positional run with its aligned twin (same initial rows, method seeds and observation-noise stream; all 3,300 matched trajectories verified from the traces). This is deliberately a representation-*and*-dimension sensitivity rather than a corrected physical benchmark: alignment expands the inputs from 4/5/3 to 81/56/55 dimensions (Table 3), and a common fixed-system control is impossible because the largest element systems hold only 23 band-gap and four perovskite rows (glass alone has 323). The encoding changes the benchmark regime (Table 4, Fig. 5). Band gap turns from a random-search ceiling into a sparse task in which only TuRBO records nonzero aligned recall ($0.03/0.07/0.17$), separating from all ten alternatives only at budget 100. Glass turns from a ceiling—89% of positional repetitions already "hit" the hold-out in the five-point initialization because distinct formulas share a coordinate—into a discriminative task: MAP-Elites reaches $0.60/0.74/0.83$ and beats all ten alternatives at every budget, while Random stays at 0.09, all of it initial-coordinate aliasing. Perovskite formation turns from an all-zero floor into a nominal 0.16 for ten methods and $0.18/0.20/0.27$ for TuRBO, but the common 0.16 is again the initialization alias, TuRBO's extra $0.02/0.04/0.11$ does not separate after Holm correction, and the aligned coordinates still collide at 49.3% because composition alone cannot distinguish structures or polymorphs. Method-rank Spearman correlations between encodings are $-0.20/-0.20/-0.31$ (band gap) and $-0.12/0.23/0.23$ (glass) across the three budgets, and undefined for the tied positional perovskite results: the historical high-element per-method readings are not robust to coordinate meaning. The aligned runs diagnose the benchmark construction; they do not establish physical landscape rankings, because representation repair and a large dimension increase are confounded. They do agree with the candidate-pool controls above: in 55–81 dimensions only the methods that propose near observations (TuRBO's trust region, MAP-Elites' archive) find anything, and every uniform-candidate method scores zero. LGPS is untouched (its columns were already element-indexed), and the aligned steel re-run above preserves the paper's headline look-up ordering. The row-matched six-element control below separates the two changes.

Table 3: Representation and support audit for the paired high-element sensitivity. "Unique" counts exact coordinate rows; collision is the fraction of rows sharing a coordinate; the largest fixed element system shows why a common fixed-system protocol cannot cover all three datasets.

Dataset	Dim. (pos./aligned)	Unique (pos./aligned)	Collision (pos./aligned)	Largest system	Rows
Band gap	4/81	1,342/4,601	70.9%/0.1%	Ge–Si	23
Perovskite formation	5/56	11/9,542	99.9%/49.3%	As–N–O–Sc	4
Glass	3/55	2,264/5,680	68.5%/21.0%	Al–La–Ni	323

Table 4: Top-5% recall in the paired representation sensitivity (100 repetitions; mean $\pm$ s.e.m.). "Holm wins" counts the other methods beaten by the aligned leader after correction over the 55 within-representation pairs; Random is shown under both encodings; ρ is the Spearman correlation of the eleven method means across encodings. All method means with error bars are in Fig. 5.

Dataset	N	Positional leader	Aligned leader	Random (pos./aligned)	ρ	Holm wins
Band gap	20	MES (0.72 ± 0.05)	TuRBO (0.03 ± 0.02)	0.71/0.00	-0.20	0
Band gap	50	DKGP/Random (1.00 ± 0.00)	TuRBO (0.07 ± 0.03)	1.00/0.00	-0.20	0
Band gap	100	DKGP/MES/ProSe/Random (1.00 ± 0.00)	TuRBO (0.17 ± 0.04)	1.00/0.00	-0.31	10
Perovskite formation	20	all eleven (0.00)	TuRBO (0.18 ± 0.04)	0.00/0.16	–	0
Perovskite formation	50	all eleven (0.00)	TuRBO (0.20 ± 0.04)	0.00/0.16	–	0
Perovskite formation	100	all eleven (0.00)	TuRBO (0.27 ± 0.04)	0.00/0.16	–	0
Glass	20	nine-way tie (1.00)	MAP-Elites (0.60 ± 0.05)	1.00/0.09	-0.12	10
Glass	50	ten-way tie (1.00)	MAP-Elites (0.74 ± 0.04)	1.00/0.09	0.23	10
Glass	100	ten-way tie (1.00)	MAP-Elites (0.83 ± 0.04)	1.00/0.09	0.23	10

Six-element row-matched decomposition. To separate coordinate semantics from ambient dimension, we searched element universes E^* and retained a row only when its complete elemental support satisfied $E_i \subseteq E^*$; no unlisted element was discarded. Exhaustive three- through five-element searches and an exact six-element certificate selected Al–As–Ga–In–P–Sb for band gap (83 rows), Cs–N–Na–O–Sr–Ti for perovskite formation (56 rows), and Al–B–Fe–La–Ni–Zr for glass

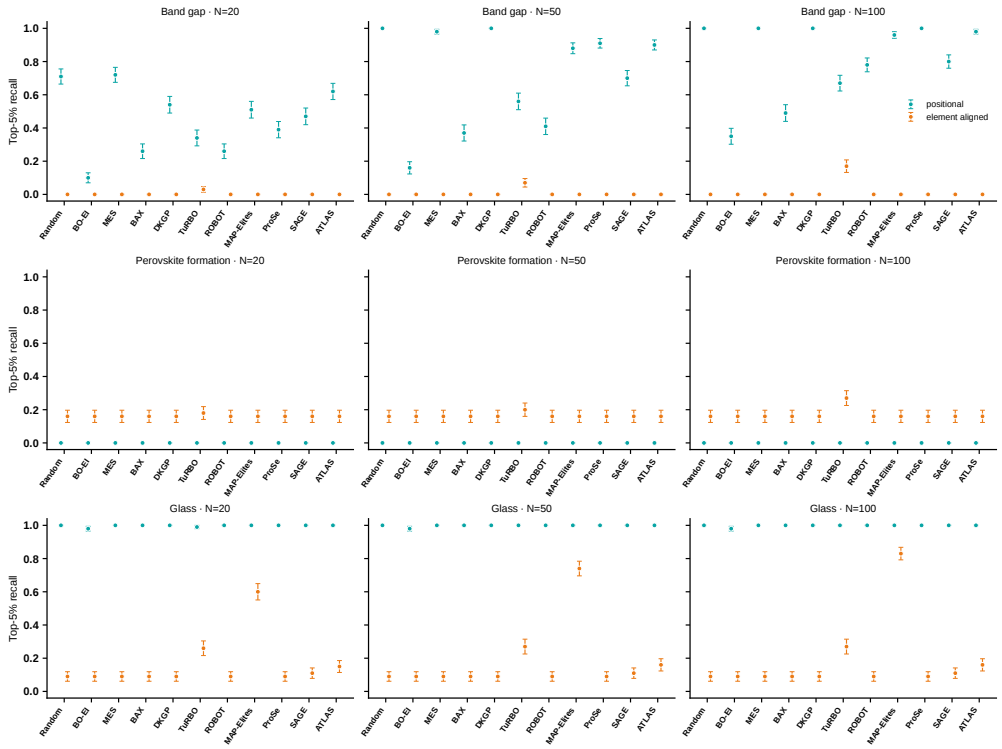


Figure 5: Positional versus element-aligned top-5% recall for all eleven methods (mean ± s.e.m., 100 paired repetitions; rows: datasets, columns: budgets 20/50/100). Pairing fixes initial rows, method seeds and observation-noise streams. Alignment removes the positional coordinate-meaning error but raises the dimension to 55–81, so the differences are benchmark-design sensitivity rather than physical-landscape effects.

Table 5: Maximum retained-row support under the strict complete-support rule $E_i \subseteq E^*$ for three- through six-element universes. The reported six-element perovskite universe is one member of a many-way 56-row optimum tie, selected for target coverage; no element is removed from any retained row.

Dataset	$\lvert E^*\rvert = 3$	$\lvert E^*\rvert = 4$	$\lvert E^*\rvert = 5$	$\lvert E^*\rvert = 6$	Locked six-element universe
Band gap	28	40	62	83	Al–As–Ga–In–P–Sb
Perovskite formation	2	12	30	56	Cs–N–Na–O–Sr–Ti
Glass	434	669	1,016	1,396	Al–B–Fe–La–Ni–Zr

(1,396 rows; Table 5). The subsets contain 83, 32 and 885 unique aligned coordinates, respectively. The perovskite universe is one of many support ties and retains 24 coordinate groups with discordant targets, so it is not a structure-complete representation.

On the identical retained rows we compared four arms: historical positional coordinates at native dimension, the same positional vector zero-padded to six dimensions, six fixed element columns, and the same aligned rows embedded in the full 81/56/55-dimensional element spaces. All arms shared hold-outs, five-point initial designs, method seeds and observation-noise streams across eleven methods, 100 paired repetitions and budgets 20/50/100 (13,200 verified trajectories). Budget 100 is executable through the continuous k-NN look-up but exceeds finite row support for band gap and perovskite, so those endpoints are support-limited.

The four-arm absolute results (Figs. 6–8) expose the ceiling/floor structure and the method-specific performance under each representation. The paired mechanism analysis (Fig. 9) shows that both mechanisms contributed to the former two-arm result. Positional padding was adverse for band gap and glass but approximately neutral for perovskite formation. The raw fixed-six-dimensional

semantics contrast is contaminated by positional initial-design aliases; after descriptively subtracting those initial apparent hits, median semantics effects across methods were $+0.20/+0.44/+0.64$ for band gap, $0.00/-0.01/-0.03$ for perovskite formation and $+0.44/+0.41/+0.49$ for glass at budgets 20/50/100. Finally, increasing aligned ambient dimension from six to the full element space was strongly adverse for band gap (median $-0.20/-0.45/-0.70$) and glass ($-0.64/-0.76/-0.80$), and non-positive for perovskite formation. With correct semantics at a fixed six dimensions the global-GP baselines become the strongest methods on band gap (at budget 100, BO-EI and ROBOT reach 0.98, TuRBO 0.94, ProSe 0.70, ATLAS 0.40, MAP-Elites 0.31), whereas in the full element space only the near-data proposers survive—MAP-Elites on glass (0.85) and TuRBO on band gap (0.37)—the candidate-pool mechanism of the controls above. Thus the historical 55–81-D figure remains valid as a combined benchmark-sensitivity screen, while the controlled experiment identifies ambient dimension as a major contributor and finds dataset-specific effects of semantic repair. None of these figures is a physical optimizer leaderboard.

Table 6: Real-materials look-up benchmarks under the top-5% hold-out protocol (the LGPS screens use peak and family hold-outs, §5.1): recall at budget 50, **100 repetitions**, full baseline roster plus both of our variants. Superscripts give the number of the other ten methods that this method beats with Holm-corrected $p < 0.05$ (out of 10); the correction is applied within each cell over all 55 method pairs. The last three columns are reported for completeness: on glass and perovskites every method scores identically, and double perovskites has only 16 distinct coordinates (see the encoding paragraph above), so none supports a per-method claim.

	discriminative or partly so		degenerate or collapsed		
Method	Steel	Band Gap	Dbl Perov.	Glass	Perovsk.
ProSe	0.01	0.96^{5}	0.14	1.00	0.00
SAGE	0.21^{8}	0.59^{1}	0.14	1.00	0.00
ATLAS	0.26^{8}	0.95^{5}	0.32^{4}	1.00	0.00
MAP-Elites	$\mathbf{0.38}^{8}$	0.84^{5}	$\mathbf{0.50}^{9}$	1.00	0.00
TuRBO	0.02	0.58^{1}	0.15	1.00	0.00
ROBOT	0.01	0.54^{1}	0.13	1.00	0.00
MES	0.00	$\mathbf{0.98}^{6}$	0.26^{2}	1.00	0.00
BAX	0.00	0.43	0.20	1.00	0.00
DKGP	0.00	0.97^{5}	0.18	1.00	0.00
BO-EI	0.00	0.25	0.20	1.00	0.00
Random	0.00	$\mathbf{0.98}^{6}$	0.29^{3}	1.00	0.00
spread	0.38	0.73	0.37	0.00	0.00

Why our two variants diverge. ProSe draws all M candidates uniformly on the simplex; SAGE anchors 60% of them on observed points (35% local jitter, 25% sparse swap) and switches to CLR/Aitchison geometry when the observations are sparse or concentrated. Random search identifies which behaviour a benchmark rewards, and the two variants reverse accordingly. On steel yield strength random scores 0.00: alloys are $\sim 73\%$ Fe plus trace additions, so uniform candidates essentially never land near a real composition, and only the anchored variant finds held-out targets (0.21 against ProSe's 0.01, itself indistinguishable from random). On experimental band gaps random scores 0.98, so uniform coverage is already near-optimal, and the ordering flips: ProSe reaches 0.96 while SAGE gives up coverage for nothing at 0.59. The geometry selector is re-evaluated at every step and selects CLR on the large majority of them on all five datasets (76.5% of steps on band gaps rising to 97.0% on perovskite formation energies), falling back to Euclidean mainly early on, while the observation set is still small; so the divergence between our two variants is driven by the candidate pool rather than by the geometry. Neither variant dominates; the choice is a property of the landscape.

Four of the five benchmarks cannot support a per-method claim. On glass-forming ability every method reaches recall 1.00, and on perovskite formation energies every method reaches 0.00: the spread across all eleven methods is exactly zero, so neither can support any per-method claim. Double perovskites has a nominal spread (0.37) but only 16 distinct coordinates, so its column ranks methods on an artefact of the encoding rather than on chemistry. On experimental band gaps the spread is large, but *random search ties for the best score* (0.98, with MES), so the benchmark

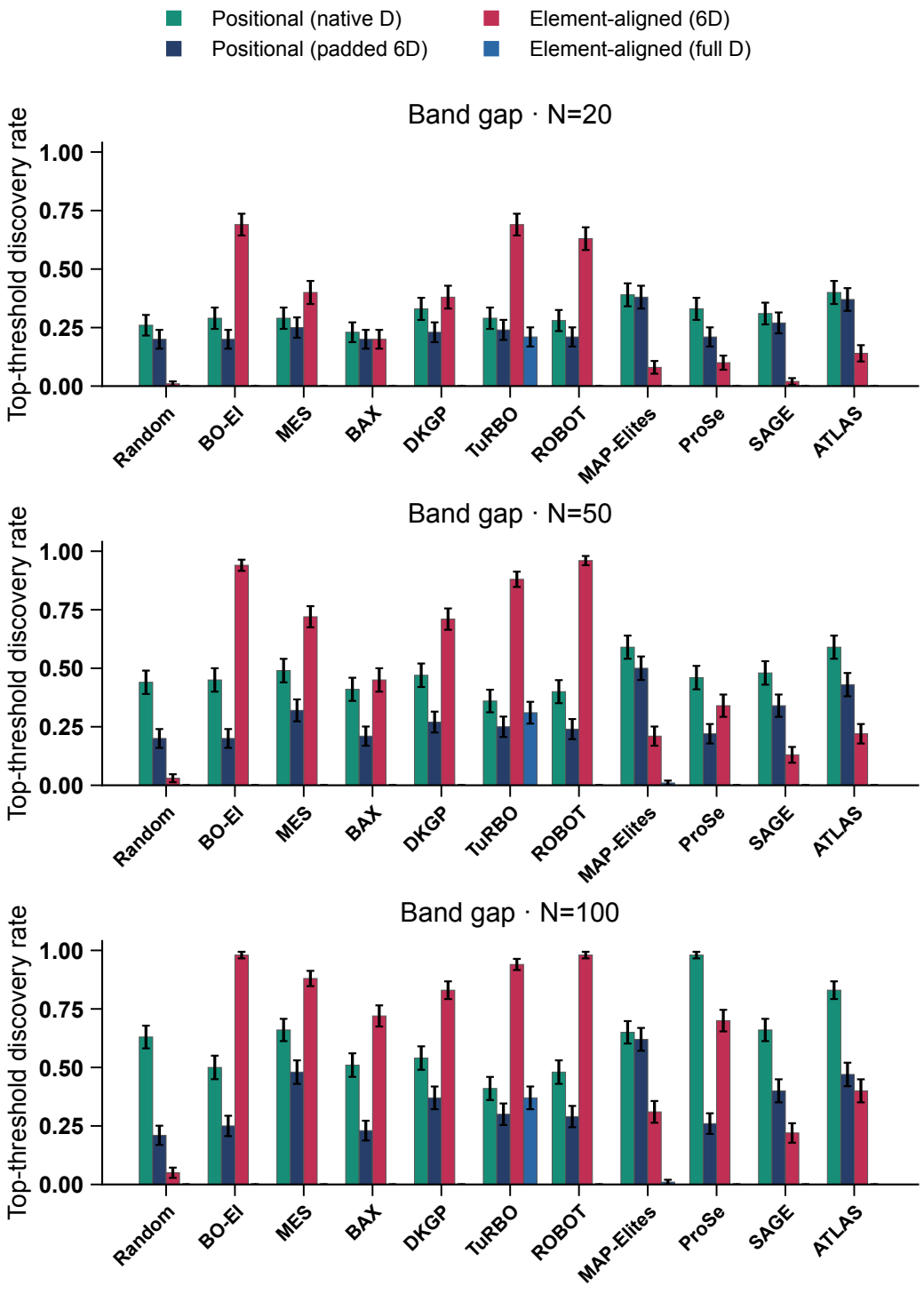


Figure 6: Absolute top-threshold discovery rate for band gap on the strict six-element subset under native positional coordinates, the same positional coordinates padded to six dimensions, six fixed element coordinates, and the same aligned rows embedded in the full element space (mean ± s.e.m. over 100 paired repetitions). All arms use the same rows, hold-outs, five-point initial designs, seeds and observation-noise streams. Positional values include apparent discoveries caused by initial coordinate aliases; the figure therefore compares benchmark representations rather than physical optimizer quality.

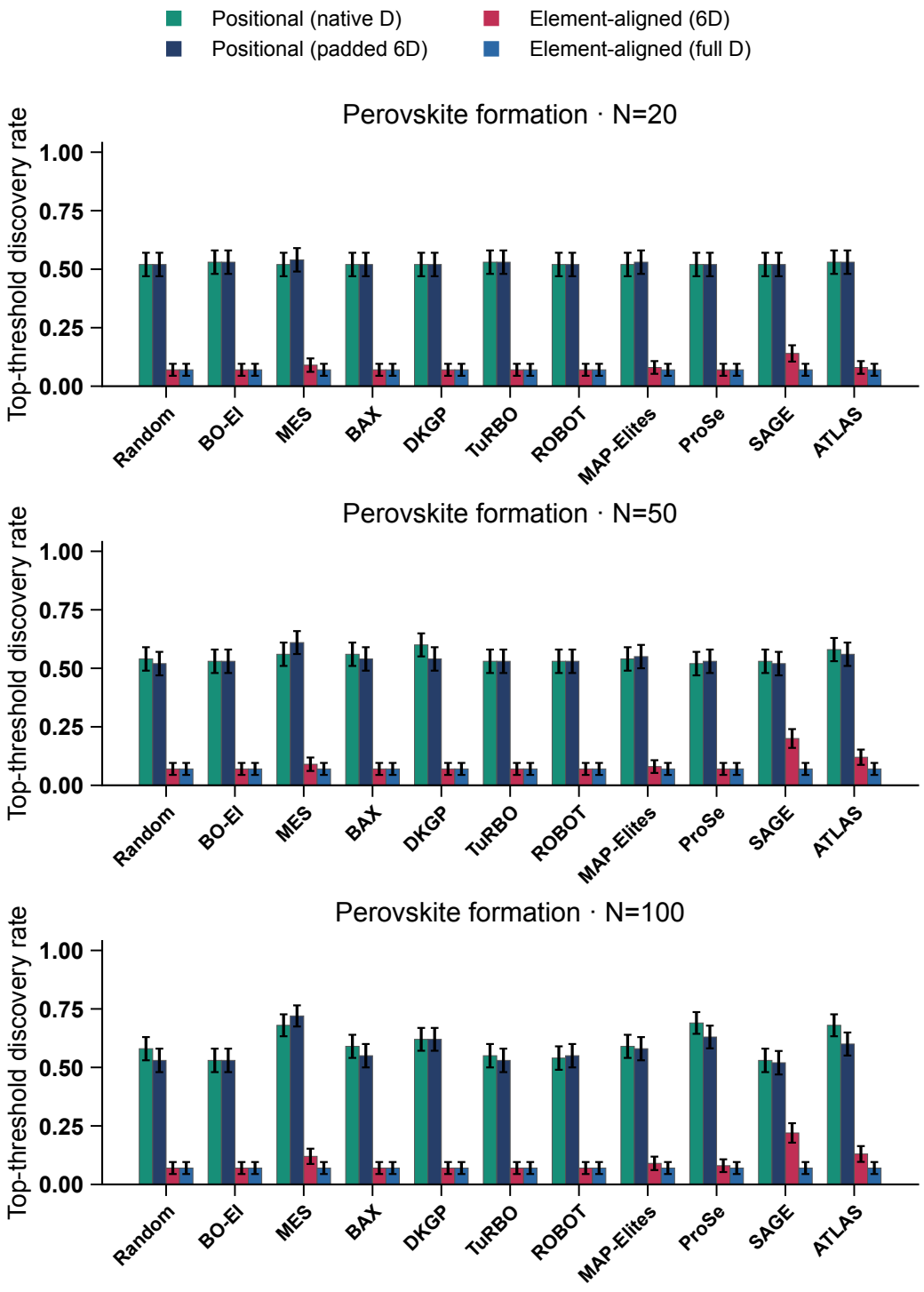


Figure 7: Absolute top-threshold discovery rate for perovskite formation under the same four-arm control as Fig. 6 (mean ± s.e.m. over 100 paired repetitions). Composition retains 24 discordant-target coordinate groups and does not encode structure identity; the aligned values therefore remain representation diagnostics.

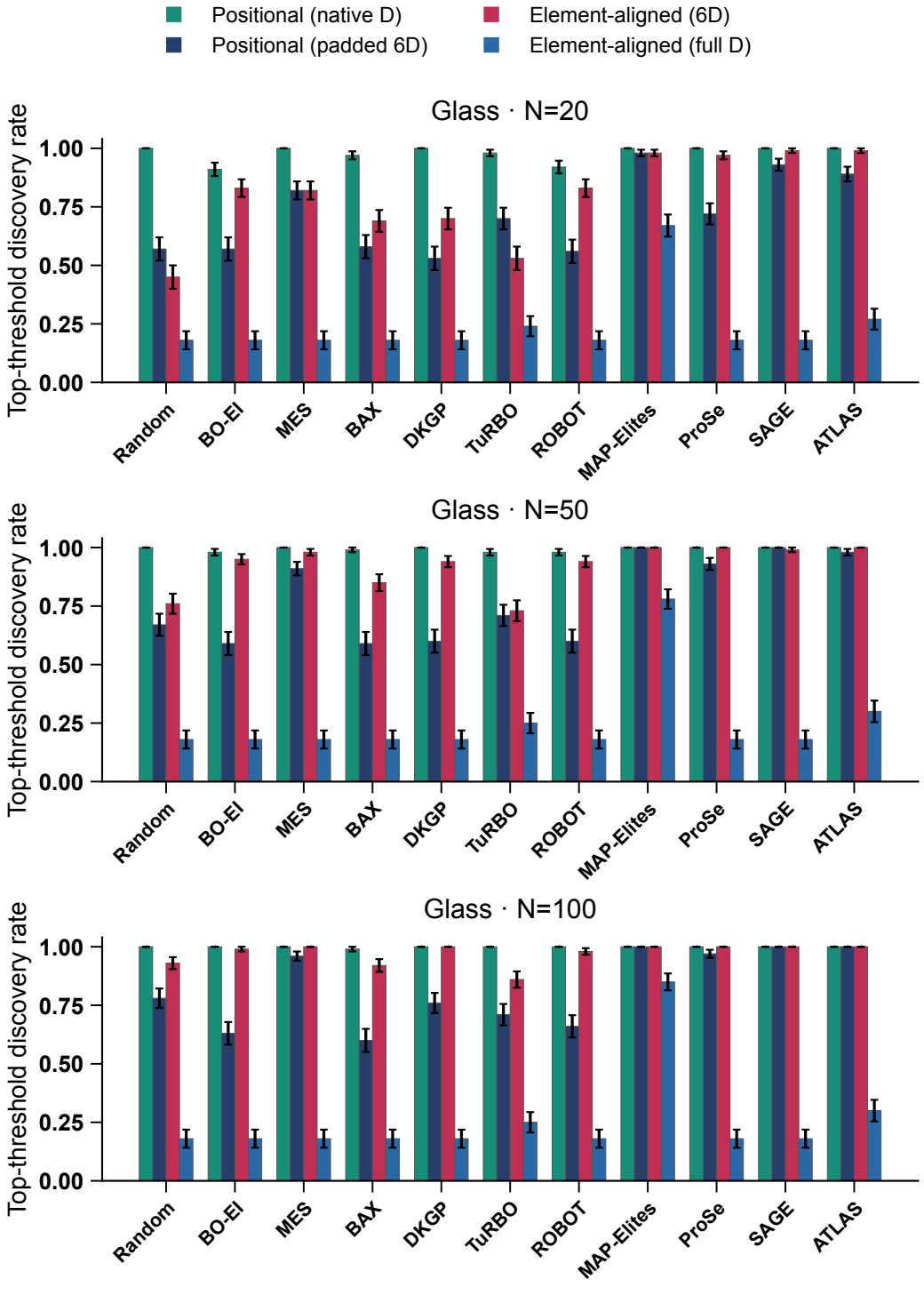


Figure 8: Absolute top-threshold discovery rate for glass under the same four-arm control as Fig. 6 (mean $\pm$ s.e.m. over 100 paired repetitions). The binary target makes the top-5% rule a positive-class hold-out rather than a literal five-percent set.

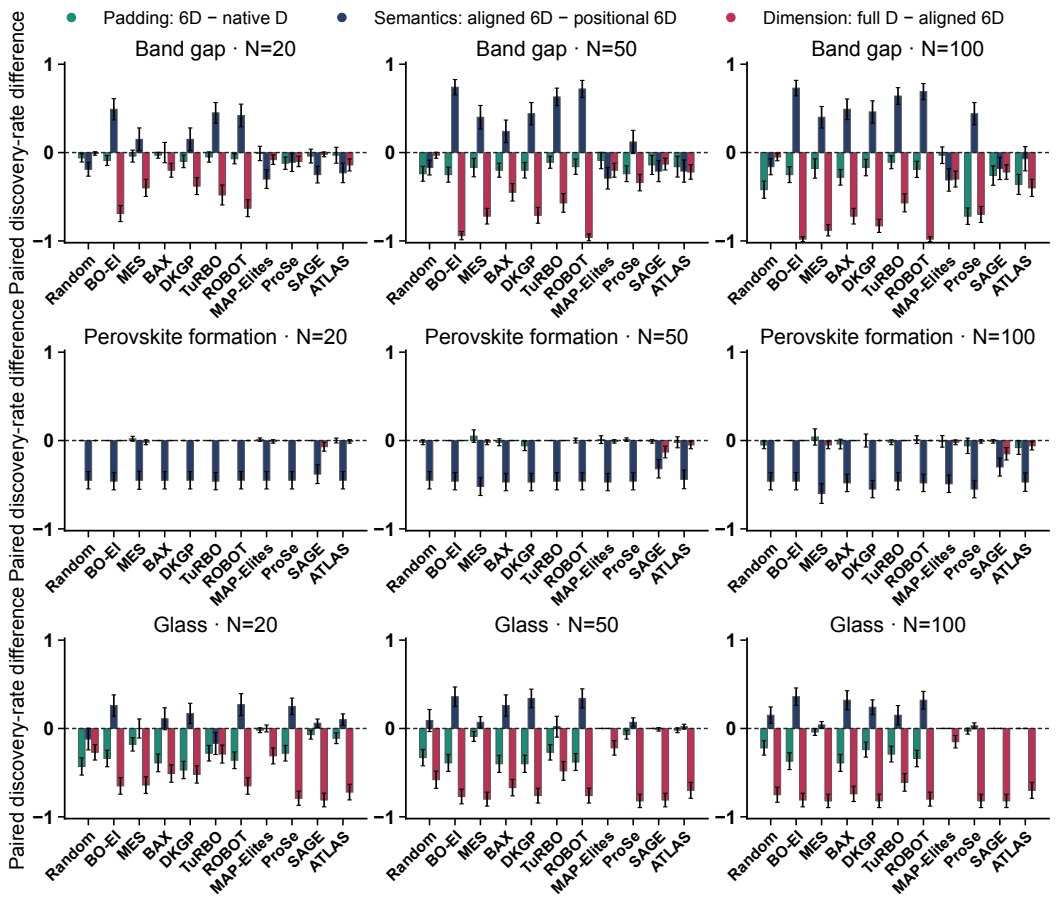


Figure 9: Mechanism decomposition across datasets and budgets. Clustered bars are method-specific mean paired discovery-rate differences for positional padding (six-dimensional minus native positional coordinates), element semantics at fixed six dimensions (element-aligned minus positional), and ambient dimension under fixed semantics (full element space minus six-dimensional aligned coordinates); whiskers are 95% normal-approximation intervals from the paired s.e.m., and the horizontal line marks zero. The raw semantics contrast includes positional initial-design aliases; post-initial values are reported in the text.

separates weak methods from strong ones without identifying anything a uniform sample would not find. These are three distinct failure modes—flat endpoint, coordinate collapse, and a random-search ceiling—and should not be read as one claim that a dataset intrinsically lacks discriminative power. We report this first because it changes how the remaining columns — and, we suspect, a good deal of published look-up-benchmark evidence — should be read: on this evidence, benchmark choice dominates method choice.

On the one benchmark that discriminates. Steel yield strength gives the cleanest separation in this paper. SAGE (0.21), ATLAS (0.26), and MAP-Elites (0.38) each beat *eight* of the other ten methods after Holm correction (smallest $p = 2.0 \times 10^{-10}$) — every global-GP and randomized baseline, and ProSe — and five of them score exactly zero. This is consistent with the family-level ordering of the real experimental LGPS result in Table 1 — {ATLAS, MAP-Elites} $\gg$ {TuRBO} $\gg$ {BO-EI, MES, DKGP, Random} — on an independent real-materials dataset at 100 repetitions rather than 30. It is corroboration rather than a replication: the two analyses use different rosters (eleven methods here against the LGPS screen's roster in Table 1) and different endpoints (top-5% recall against held-out-family recall), and SAGE is third on LGPS while ProSe is at the random level on both. We do not read double perovskites per method (16 distinct coordinates, above). Quality-

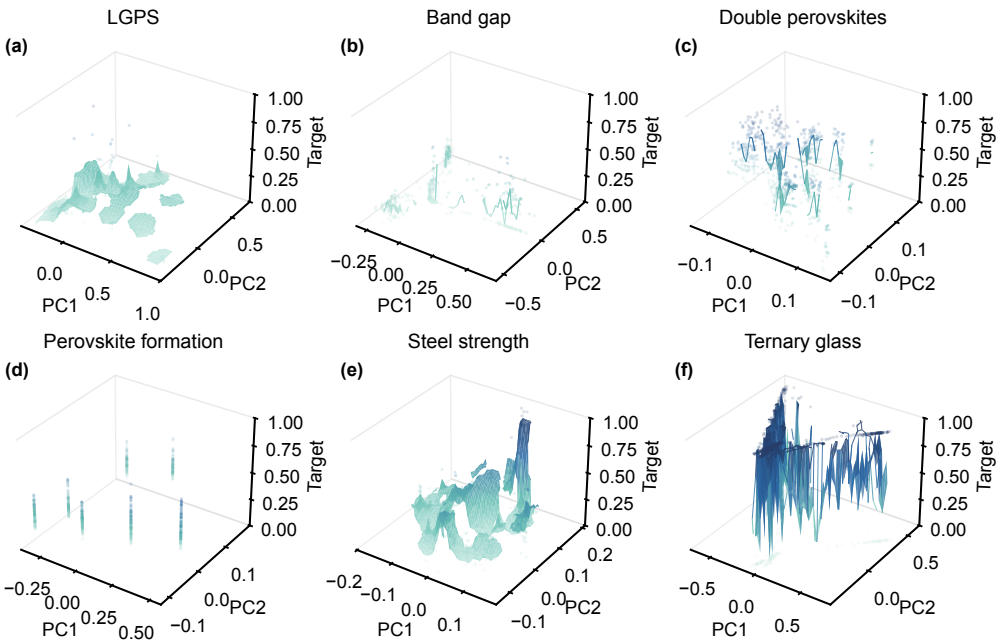


Figure 10: Three-dimensional views of the element-aligned composition–response data for (a) LGPS, (b) experimental band gap, (c) double-perovskite band gap, (d) perovskite formation energy, (e) steel yield strength and (f) ternary-glass formation. Horizontal axes are the first two principal-component scores of the element-aligned composition, the vertical axis the normalised target. Surfaces are descriptive eight-nearest-neighbour interpolations within the observed support—not GP posterior means and not the k=3 benchmark oracle—and are shown only to make the feasibility gaps, isolated peaks and region-dependent smoothness of the real data visible. Grid cells farther from the data than the 90th-percentile nearest-neighbour radius are left blank, so the plotted extent is the observed support.

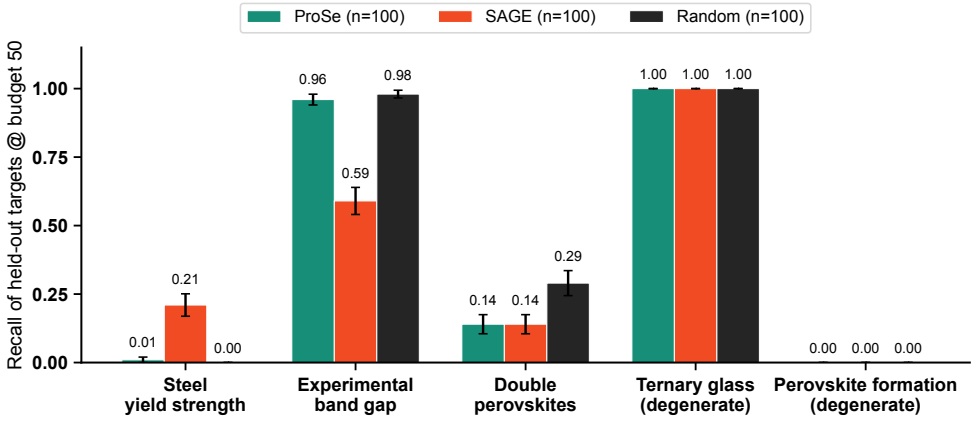


Figure 11: Our two variants on the look-up benchmarks, recall at budget 50 over 100 repetitions with random search as the scale reference. ProSe draws candidates uniformly; SAGE anchors 60% of them on observed points and uses log-ratio geometry on sparse or concentrated data. Where random search already scores highly (band gap, 0.98) uniform coverage wins; where it scores zero (steel) only the anchored variant finds held-out targets. Error bars are s.e.m. over repetitions.

diversity and posterior-independent search therefore lead where the benchmark can tell methods apart, but neither of our variants is uniformly ahead, and MAP-Elites — not ProSe or SAGE — tops the one discriminative column.

Candidate-pool control. A natural objection to the steel and LGPS results is that the methods differ in *where they propose candidates* rather than in how they score them: MAP-Elites mutates archive members, SAGE anchors 60% of its pool on observations and ATLAS's channels probe near the data, whereas every GP baseline and ProSe score $M{=}1000$ uniform Dirichlet draws, which land at a median Euclidean distance of 0.60 (steel) and 0.30 (LGPS) from the nearest reference composition and within 0.1 of one essentially never. We therefore re-ran BO-EI and MES with SAGE's candidate generator (25% sparse swap, 35% local jitter, 40% uniform; GP, kernel and acquisition unchanged) in one paired run with the reference methods (100 repetitions, budgets 20/50/100). The pool does not close the gap. On steel, recall@100 is 0.00 for BO-EI-Anchored and 0.09 for MES-Anchored against 0.47 (SAGE), 0.51 (MAP-Elites) and 0.58 (ATLAS) in the same run; on the LGPS family hold-out it is 0.02 and 0.19 against 0.11, 0.49 and 0.64. MES-Anchored does improve on MES (paired one-sided sign test at budget 100: 9–0 wins on steel, 18–1 on LGPS; Holm $p{=}0.041$ and 0.001), but the traces show why the gain is small: with 60% of the pool anchored, MES-Anchored still places only 6% (steel) and 21% (LGPS) of its queries within 0.1 of a reference composition, and BO-EI-Anchored's best query sits a median 0.86 from the nearest alloy—the acquisition prefers the far-off uniform candidates, where posterior variance is largest and the look-up oracle extrapolates the nearest alloys' values. Posterior-driven scoring, not the candidate generator, is what sends the GP baselines off-data; the same pool under ProSe-style scoring (SAGE) reaches 0.47.

Surrogate and geometry controls. Two further objections are that the GP baselines are handicapped by their raw-fraction inputs (a simplex-aware geometry might behave differently) or by the GP itself (a tree ensemble, whose uncertainty does not grow off-data, might not). We ran both in one paired run with the reference methods (100 repetitions, budgets 20/50/100, same hold-outs): BO-EI and MES on CLR (Aitchison) coordinates (18) with multiplicative replacement 10^{-3}, and RF-EI—the BO-EI loop with a 100-tree random forest as surrogate and the spread of the per-tree predictions as σ, in the SMAC construction (28)—on the uniform pool and on SAGE's anchored pool. CLR coordinates change nothing: BO-EI-CLR and MES-CLR score 0.00 on both datasets and tie their raw-fraction counterparts in 98–100 of 100 paired repetitions at every budget, never winning one. The forest surrogate on the uniform pool changes nothing either (RF-EI: 0.02 on LGPS, 0.00 on steel). RF-EI on the anchored pool is the one control that moves: recall@100 of 0.32 on both LGPS and steel—above SAGE on LGPS (0.13; 28–9 wins, Holm $p{=}0.038$), level with SAGE on steel (0.39; 20–27, n.s.), and below MAP-Elites and ATLAS on both (LGPS 5–26 and 5–29 at budget 100; steel 8–37 and 16–27). Its queries sit where the anchored candidates are (61% within 0.1 of a trial on LGPS and 39% on steel; best query a median 0.06 and 0.18 away), whereas the GP on the same pool wandered off (above). The three controls together localise the failure as an interaction: a GP posterior scores zero on either pool, a uniform pool scores zero under either surrogate, and only a bounded-uncertainty surrogate given near-data candidates recovers part—not all—of the quality-diversity methods' separation. That is consistent with our reading that trust in posterior variance, rather than surrogate family or coordinate geometry, is what drives the baselines off-data, and it is the caveat that an anchored RF-EI is a stronger baseline than any GP we ran.

Budget scan. The budgets in this paper are wet-lab scale ($N \leq 100$), and on the smooth synthetic three-peaks landscape the GP baselines saturate the recall metric already at that budget (Table 1), so a fair question is whether they would simply catch up on the real data given more evaluations. We ran BO-EI, BO-EI-Anchored, MES, TuRBO, MAP-Elites, ATLAS and Random to budget 500 (a separate run of 30 paired repetitions with its own seeds, so its budget-100 values differ from Table 1; recall read at 100/250/500, same hold-outs; LGPS has 140 distinct compositions and steel 312 alloys, so the 500-query arm largely re-samples a small support). They do not. On the LGPS family hold-out, ATLAS and MAP-Elites rise from 0.57 and 0.40 at budget 100 to 0.77 and 0.80 at 500, while BO-EI reaches 0.10, MES 0.07, BO-EI-Anchored 0.03 and TuRBO—the one GP method that proposes locally—0.30; on steel ATLAS and MAP-Elites reach 0.83 and 0.67 and every GP baseline finishes at 0.00–0.03, level with random. Re-running BO-EI and MES on CLR coordinates to budget 500 (same seed namespace, 30 paired repetitions) changes nothing either: BO-EI-CLR and MES-CLR score 0.00 on the LGPS family hold-out and 0.03 on steel at budget 500, tied with their raw-fraction twins in 27–30 of 30 repetitions at every budget and never significantly different. At budget 500 ATLAS beats BO-EI in 20–0 (LGPS) and 24–0 (steel) discordant repetitions and MES in 21–0 and 24–0. The GP baselines' best-so-far *values* do keep rising (0.985–0.991 on steel at 500, above MAP-Elites' 0.985 and ATLAS's 0.951), but they are reached a median 0.58–0.84

from any real alloy, on the extrapolated plateau of the look-up oracle rather than at a real material. The separation on these two datasets is therefore not a small-budget artefact.

Oracle tie-break and platform reproducibility. The LGPS table has 299 trials at 141 bitwise-distinct composition vectors (140 physical compositions after merging one pair that differs by 1.1×10^{-16}), and repeated compositions carry different measured targets because the seven-element coordinates omit synthesis and heat-treatment conditions. When a query coincides with a repeated composition, the k=3 neighbour set among the equidistant rows is whatever the nearest-neighbour search returns first. That order is deterministic on a given platform but differs between macOS/arm64—the environment that produced the reported runs (June 2026; package versions not recorded) and whose current stack (numpy 2.4.6, scikit-learn 1.8.0) regenerates all stored non-ProSe traces bit-for-bit—and Linux/x86_64 (numpy 2.2.6, scikit-learn 1.7.2): with identical seeds the two draw identical initial designs, yet the observed values at the initial rows differ in 29 of 30 repetitions, and every value-driven method then follows a different trajectory (only Random's queries are unchanged). Re-running the family block on Linux with the same seeds gives recall@20/50/100 of $0.33/0.43/0.47$ for ATLAS, $0.20/0.27/0.47$ for MAP-Elites, $0.07/0.10/0.10$ for TuRBO, $0.07/0.07/0.07$ for MES, $0.00/0.03/0.03$ for Random and 0.00 throughout for BO-EI and DKGP, against $0.63/0.57/0.27/0.07/0.03/0.00/0.00$ at budget 100 in Table 1. The two platforms therefore evaluate different effective objectives at repeated compositions. The family values in Table 1, the medians of Sec. 5.2, the 200-repetition p-values and the LGPS controlled-screen column are results under the macOS oracle, and the present evidence does not establish a platform-independent ordering between ATLAS and MAP-Elites. A deterministic oracle (one value per physical composition under a pre-declared aggregate, with a stable tie-break) and a re-run of both LGPS protocols are planned; the platform-specific traces are retained for provenance.

D Auxiliary GP Failure Modes

For completeness we distinguish two sources of GP failure that affect GP-based methods broadly but are categorically distinct from the principled assumption violations discussed in the Discussion. *(i) Theoretical sample-complexity bounds:* GP-UCB regret scales as $\mathcal{O}(\sqrt{T\gamma_T})$ with $\gamma_T = \mathcal{O}((\log T)^{d+1})$ for RBF kernels (35), and posterior contraction for functions in H^s decays only as $n^{-s/(s+d)}$ (36). These affect all GP methods and apply even when the modeling assumptions hold. *(ii) Empirical calibration failures:* at small n, MLE of kernel hyperparameters is ill-conditioned and length-scales often saturate at user-imposed bounds (we observed this in our LGPS runs). In principle these can be mitigated with richer priors (SAASBO (37)), but the required sample sizes typically exceed the $N \leq 100$ budgets in materials discovery. The principled assumption violations described in the Discussion (jump discontinuities, undefined-function gaps, threshold-derived targets, non-stationarity) are different: they are not remedied by the GP refinements evaluated here.

D.1 Where does GP miscalibration occur?

Apple-to-apple conditioned diagnostic. We re-evaluate the saved pointwise predictions from the calibration audit without refitting or changing any prediction. Thus the representation, 20 deterministic group-held-out splits, $N_{\text{train}} = 100$, Matérn-$5/2$ GP, posterior mean and standard deviation, central Gaussian intervals, and standardized residual $z = (y-\mu)/\max(\sigma, 10^{-6})$ are identical to the overall analysis. Within each held-out split, the top-q subset contains exactly $\lceil qn_{\text{test}} \rceil$ observations ranked by true y (prediction-independent deterministic tie-breaking), and its non-top complement is the matched comparator. We use $q \in \{.10, .20\}$ as primary thresholds and $.05$ as a sensitivity analysis. For every subset we compute $C_{68}, C_{90}, C_{95}, C_{99}$, the mean absolute calibration error across those levels, $\mathbb{E}[z]$, $\mathbb{E}[z^2]$, $P(|z| > 2)$, $P(|z| > 3)$, posterior-standard-deviation over target-standard-deviation, and pointwise Gaussian negative log predictive density. Split-paired effects are reported both against overall and against the non-top complement. Inference uses 20,000 split-level bootstrap resamples for 95% confidence intervals and two-sided Wilcoxon signed-rank tests, with Holm correction within each metric/comparator family across the six datasets and three q values. The two-dataset reliability figure for this audit appears in the main text (Fig. 3); the six-dataset version and the paired effects follow below.

Table 7: Optimization-conditioned calibration at $q = .10$ (means over 20 paired splits). ΔC_{95} is top minus non-top; p_{Holm} is the Holm-adjusted two-sided Wilcoxon result. Positive $\mathbb{E}[z]$ indicates underprediction.

Dataset	C_{95}^{all}	C_{95}^{top}	C_{95}^{non}	ΔC_{95}	$\mathbb{E}[z]_{\text{top}}$	$\mathbb{E}[z^2]_{\text{top}}$	$P(\lvert z\rvert > 3)_{\text{top}}$	p_{Holm}
LGPS	.887	.403	.943	-.540	2.850	11.661	.492	.001
Band gap	.928	.463	.979	-.517	2.415	8.106	.222	$< .001$
Double perovskites	.946	.622	.984	-.362	1.822	3.452	.000	$< .001$
Perovskite formation	.942	.483	.993	-.510	2.165	5.269	.158	$< .001$
Steel strength	.865	.691	.885	-.194	1.823	8.076	.175	$< .001$
Ternary glass	.961	.997	.957	.040	.618	.462	.000	.010

Interpretation and evidence boundary. Miscalibration is concentrated in the high-response region, not distributed uniformly. LGPS is the strongest case: mean C_{95} drops from .887 overall to .403 in the top decile, versus .943 in the complement ($\Delta = -.540$, Holm-adjusted $p = .001$), with $\mathbb{E}[z] = 2.850$ and 49.2% of top-decile points beyond 3σ. Steel shows the same, weaker pattern (.865 overall, .691 top, .885 non-top; $\Delta = -.194$, adjusted $p < .001$). Band gap, double perovskites, and perovskite formation also show significant top-region undercoverage, so the enrichment is not specific to the two datasets where discovery is emphasized. Ternary glass is the exception because its maximum target is saturated and heavily tied; within that boundary plateau, a "top" ranking is not identifiable and its reversed effect should not be interpreted as superior tail calibration. The .05 and .20 sensitivity analyses preserve the substantive pattern.

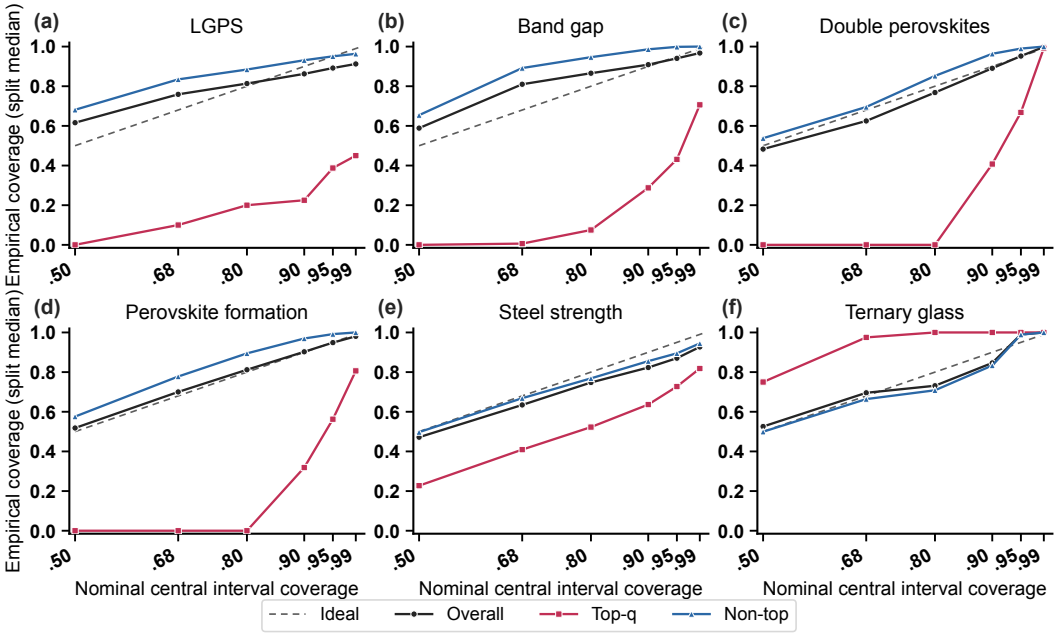


Figure 12: Top-decile reliability across all six look-up datasets. Each curve is the median over 20 paired splits. Five non-saturated landscapes show the same ordering—top-region coverage below overall and non-top coverage—whereas ternary glass is qualitatively different because a large tied plateau occupies its maximum target value. The repeated pattern establishes a general optimization-conditioned diagnostic, not an LGPS/steel-specific biomarker.

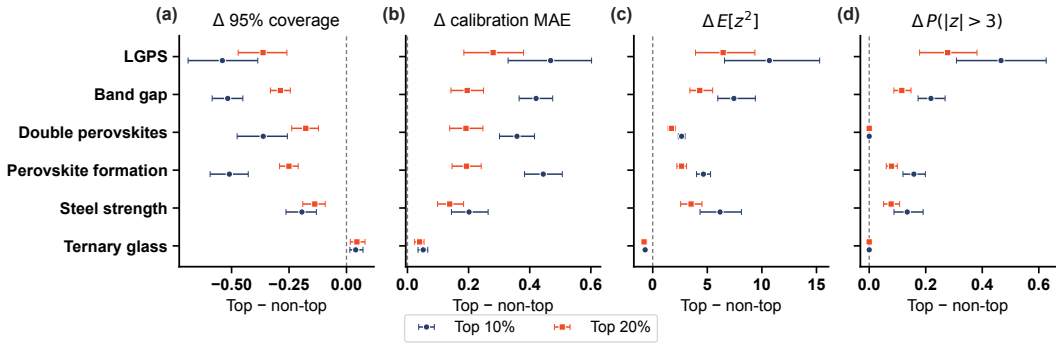


Figure 13: Paired top-minus-non-top effects for the 10% and 20% thresholds. Points are mean paired-split effects and whiskers are 20,000-resample bootstrap 95% confidence intervals. Negative ΔC_{95} and positive changes in calibration MAE, $\mathbb{E}[z^2]$, and $P(|z| > 3)$ indicate worse top-region calibration.

Threshold sensitivity and failure mechanism. For LGPS, the top-minus-non-top C_{95} gap changes monotonically from $-.650$ at $q = .05$ to $-.540$ at $q = .10$ and $-.363$ at $q = .20$; the corresponding steel effects are $-.270$, $-.194$, and $-.138$. The increasingly severe failure toward the extreme tail is accompanied by strongly positive $\mathbb{E}[z]$ (2.850 for LGPS and 1.823 for steel at $q = .10$), so the result is not accurately described as variance underestimation alone. It combines systematic underprediction of high responses with posterior uncertainty that does not cover that conditional mean error (Figs. 3–13). This diagnostic therefore separates three claims. First, global miscalibration is directly observed. Second, top-region enrichment is directly supported by paired, prediction-preserving comparisons. Third, mode-conditioned calibration is *unavailable*: the offline split predictions do not contain mode labels aligned with sequential discovered/undiscovered status, so no such result is imputed. Because these are post-hoc held-out posteriors rather than the posterior at each BO decision, the analysis localizes a high-response failure but does not establish that calibration error causally produced the sequential discovery ranking.

E Baseline Implementation Details

Table 8 lists every method's configuration; the ATLAS paragraph below gives the definitions that §3 states only qualitatively.

F k-NN Neighbor Count Ablation

k=3 and k=5 perform comparably on Three Peaks (0.680 vs 0.681), confirming that small k is optimal for multi-modal landscapes. k=10 degrades to 0.528 (-22%), demonstrating that large k recreates GP-like global smoothing. k=1 is unstable (higher variance). We select k=3 as the best locality–robustness trade-off (Fig. 14).

ATLAS details. The score of Eq. (4) uses $w_d = 0.42 - 0.04\,p$, $w_v = 0.08 + 0.04\,p$, $w_f = 0.35 + 0.02\,p$ and $w_\rho = 0.15$ with $p = t/N$ the fraction of the budget spent; $\hat{D}$ and $\hat{f}$ are inverse-distance-weighted k=3 estimates over the observations, and each term is min–max normalised over the candidate pool before blending. Dissonance is $D(\mathbf{x}) = |f(\mathbf{x}) - M_\ell(\mathbf{x})|/\sigma_{\text{resid}}$ with σ_{resid} the standard deviation of the current level's residuals over *all* observations. A level transition fires when at least two of the last ten dissonances exceed the median plus one standard deviation of all dissonances, with a five-step cooldown. The frontier term $d_{\text{frontier}}(\mathbf{x})$ is the distance to the nearest observation (in the active metric); the density term is the number of observations within 0.1 of the candidate, divided by its maximum over the pool. Level 2 is kernel ridge regression (RBF, ℓ=0.2, α=0.01); it is $O(n^3)$ like a GP, which at $n \leq 100$ is immaterial—the ladder exists for trust, not cost. Level 3 adds Gaussian peaks to the Level 2 fit: among observations whose dissonance exceeds the transition threshold, at most three greedy clusters of radius 0.15 are formed per transition, each

Table 8: Baseline configurations. All methods: N_{init}=5 and a fresh pool of M=1000 Dirichlet candidates per step unless stated; budgets and repetition counts are those of the experiment in which a method appears (§4).

Method	Configuration
BO-EI	GP: RBF (length_scale=0.1) + WhiteKernel (noise=0.01). EI acquisition. M=1000.
TuRBO	GP: Matérn(ν=2.5) + WhiteKernel. Trust region: L_{init}=0.8, L_{min}=0.05, L_{max}=1.6. M=1000.
ROBOT	GP: Matérn(ν=2.5) + WhiteKernel. Diversity: threshold=0.15, portfolio=5, $w_{\text{div}} = \max(0.1, 0.5(1 - \text{progress}))$. M=1000.
Thompson	GP: RBF + WhiteKernel. Independent marginal posterior draw per candidate (no joint sample); arg-max over the pool. M=1000.
MES	BoTorch `qMaxValueEntropy` on a SingleTaskGP with Normalize/Standardize input and outcome transforms; 64 max-value samples.
InfoBAX	Level-set target $A(f) = \{\mathbf{x} : f(\mathbf{x}) \geq \tau\}$ with τ the 0.8-quantile of observed y; acquisition is the binary entropy $H(P[f(\mathbf{x}) \geq \tau])$.
DKGP	$d \to 16 \to 16 \to 2$ MLP feature extractor, RBF kernel on the 2-D embedding, EI acquisition; 200 Adam iterations per step.
Random	Uniform (Dirichlet) sampling; no model.
ProSe	K-means $k \leq 4$. Decay: $w_c = 0.7 \cdot 0.90^{0.05t}$. k-NN: k'=3. $w_p = 0.3(0.5 + 0.5t/N)$. M=1000.
SAGE	ProSe scoring with an anchored candidate pool and compositional geometry: M=1000 as 25% sparse swap (swap mass 0.15) / 35% local jitter (σ=0.06(1−0.5t/N)) / 40% uniform; distances use CLR/Aitchison geometry (18) on any step where the observations reach zero-fraction ≥ 0.20 or mean largest component ≥ 0.70, and Euclidean otherwise, re-evaluated from all observations at each step; constant 0.7/0.3 distance blend; $w_p = 0.2(0.5 + 0.5t/N)$. A gradient-triggered GP-exploitation branch exists in the implementation but is disabled in every run reported here.
ATLAS	4-level null models (Constant→Linear→Smooth→Peak-Aware). Transition: high-dissonance events in a 10-step recency window. Cooldown: 5 steps. Candidate channels, one fixed configuration in every experiment of this paper: dissonance 30%, value-guided 0%, frontier 45%, random 25% (sparse-swap and low-value channels off; the value-channel sensitivity control is described below). M=1000.

Table 9: Mean best value (± s.d.) by k and distribution (30 reps, N=50).

k	Original	Lin. Grad.	Step-Like	Boundary	3-Peaks
1	.416 ± .034	.964 ± .023	.926 ± .022	.959 ± .021	.626 ± .146
3	.419 ± .051	.969 ± .017	.933 ± .025	.955 ± .021	**.680 ± .142**
5	.428 ± .064	.969 ± .019	.932 ± .022	.953 ± .016	.681 ± .106
7	.405 ± .028	.969 ± .018	.934 ± .018	.960 ± .016	.648 ± .157
10	.430 ± .065	.966 ± .020	.934 ± .017	.962 ± .019	.528 ± .098

peak centred on the cluster's highest-value member with height $y_{\max} - M_2(\text{centre})$ and width the mean cluster radius clipped to $[0.03, 0.15]$; a peak is registered if its height exceeds 0.01 and it lies more than 0.1 from an existing peak. The candidate channels are dissonance 30%, value-guided 0%, frontier 45% and random 25% in every experiment of this paper; As a sensitivity control, ATLAS-ValueProbes moves 15% of the frontier budget to value-guided probes near high-value observations (30/15/30/25) and changes nothing else. In a paired run against ATLAS (100 repetitions, budgets 20/50/100, seed namespace `atlas-value-probe-v1`), recall@100 is 0.40 against 0.43 on the LGPS family hold-out and 0.41 against 0.43 on steel; neither difference is significant (one-sided sign tests 13–16 and 18–20 discordant repetitions, Holm p=1), the variant still beats SAGE on LGPS (0.40 vs 0.16; 33–9, Holm p=7 × 10^{-4}) and ties it on steel, and MAP-Elites leads both ATLAS variants in this run (0.45 / 0.57). The channel choice therefore does not produce the discovery result; the configuration was fixed once and is not tuned per dataset. Numbers differ from Tables 1 and 6 because this is a separate seeded run.

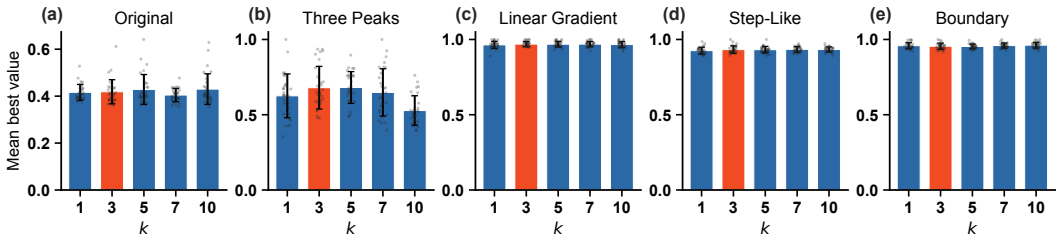


Figure 14: k-NN ablation across five distributions (30 repetitions each). Coral: k=3 (default); grey points are individual repetitions and error bars are ± 1 standard deviation.

Thompson sampling variant. The Thompson baseline of Table 1 draws one independent sample from the marginal posterior at each of the M=1000 candidates and takes the arg-max, so the posterior covariance across candidates is not used. As a control we ran a joint-sample variant (one draw from the full predictive covariance over the pool, through a jittered Cholesky factor) paired with the marginal version under the controlled-screen protocol (100 repetitions, budgets 20/50/100, seed namespace `thompson-joint-v1`): recall@100 is 1.00 against 1.00 on three-peaks and 0.008 against 0.011 on LGPS, with no significant paired difference at any budget (two-sided sign tests, Holm $p \geq 0.22$), and the threshold-hit rates are likewise indistinguishable. The marginal approximation therefore does not drive the Thompson row.

G Additional Results

Table 10: Auxiliary CMBD mean best value, all methods 30 repetitions on the current code with paired seeds (regenerated 2026-09; the previous version mixed runs from different code states). Bold marks the column maximum among the eight methods shown. This table is auxiliary evidence; the main claim relies on the two-axis retrospective comparison in Table 1.

Method	Orig.	Lin.G.	Step	Bnd.	3-Peaks
Random	**.430**	.793	.877	.931	.510
BO-EI	.417	.985	.948	.960	.366
Thompson	.415	.978	.935	.928	.527
TuRBO	.394	**1.00**	**.961**	**.966**	.409
ROBOT	.408	.988	.941	.953	.379
ATLAS	.404	.926	.911	.951	.523
ProSe	.417	.959	.930	.947	**.696**
SAGE	.420	.994	.944	.962	.521

Table 10 reports the auxiliary CMBD screen on the best-value metric; Fig. 15 shows the dimension sweep. On this best-value score the GP methods lead the smooth landscapes and ProSe leads Three Peaks; ATLAS, which scored 0.72 on Three Peaks in an earlier version of this table, now scores 0.52 because its default channels were re-tuned for held-out-family discovery (frontier and dissonance probes; Appendix E) after that run. Best value on a smooth mixture and early discovery of a hidden family are different targets, which is the two-axis point of §5.1–§5.3; the paper's other tables use the re-tuned ATLAS throughout.

Multiplicity in the 200-repetition LGPS comparison. This paragraph documents the correction behind the p-values quoted in §5.1. The planned comparison ran ATLAS, MAP-Elites and Random for 200 paired repetitions under the family hold-out, so the runner's Holm correction acts within each budget over three method pairs. Both comparisons against Random are far smaller ($p < 10^{-18}$ at budgets 50 and 100), so the ATLAS–MAP-Elites value is the largest in each within-budget family and the correction leaves it unchanged: one-sided paired sign-test $p = 0.032$ (46 wins vs 29, 125 ties) at budget 50 and $p = 0.036$ (43 vs 27, 130 ties) at budget 100; the two-sided values are 0.064 and 0.072. Recall@50 and recall@100 are nested read-outs of the same 200 runs; treating the two

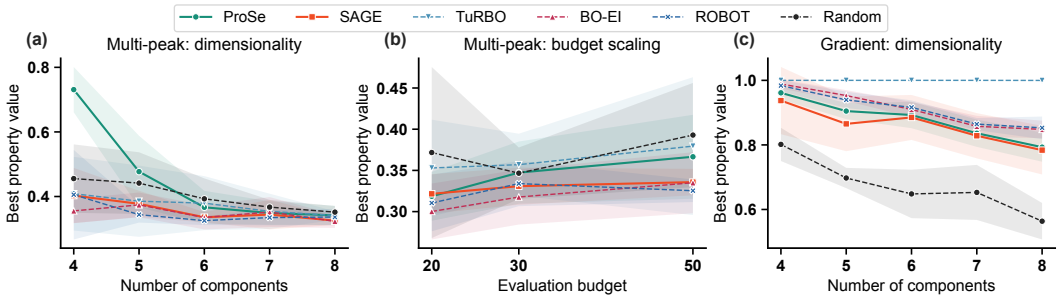


Figure 15: Scalability across 4–8 components, 10 repetitions; lines are means and same-colour bands are ± 1 standard deviation. (a) Multi-peak: ProSe leads the *strongest* alternative at each dimension by $+61\%$ (4D) and $+8\%$ (5D); at $\geq$6D every method converges to random performance and ProSe falls marginally below it, indicating a *benchmark saturation regime* (no per-method verdict drawable; reported as a methodological negative result). (b) Budget scaling at 6D. (c) Gradient landscapes: BO-EI degrades with dimension while TuRBO does not. Data regenerated 2026-08-04; the previous version of this figure was produced before the 2026-03-27 seed-control and exception-handling fixes to the baselines and its TuRBO curve differs materially.

budgets as one inferential family gives Holm-adjusted one-sided $p = 0.064$ for both. At budget 20 the comparison is uninformative (27 vs 23, $p = 0.34$). We therefore read the 200-repetition result as consistent directional evidence at both informative budgets: significant per budget, borderline family-wise.

H Standard BO Benchmark Results

Table 11: Standard BO benchmarks (30 reps, N=50). Branin: d=2, 3 global minima. Hartmann: d=6, 6 local optima.

Method	Branin-2D	Hartmann-6D
BO-EI	$1.035 \pm .017$	$0.271 \pm .059$
ProSe	$1.053 \pm .014$	$0.375 \pm .060$
TuRBO	$1.040 \pm .022$	$0.392 \pm .137$
ROBOT	$1.037 \pm .018$	$\mathbf{0.426 \pm .056}$
Random	$1.036 \pm .017$	$0.283 \pm .062$

On Branin, all methods saturate due to low dimensionality (d=2) and sufficient budget (N=50). On Hartmann-6D, ROBOT outperforms ProSe because the GP accurately models this smooth function—the diversity bonus is effective when the surrogate does not degrade. This validates our claim: ProSe's advantage is specific to settings where GP concentration bias is the bottleneck (multi-modal compositional landscapes), not where the GP itself is the bottleneck (smooth standard functions).

I Extended Related Work and Positioning

This appendix expands three distinctions that Section 2 states only briefly, because each concerns a recent line of work that could otherwise be read as contradicting our thesis.

Domain discontinuity versus response discontinuity. Zhang et al. (38) extend probabilistic reparameterization to non-equidistant discrete variables and report efficient optimization of highly discontinuous, discretized objectives, explicitly targeting autonomous laboratories. This is complementary to, not in conflict with, our thesis, and the distinction is worth stating precisely. Their discontinuity is a property of the *search domain*: variables that are mixed, discrete, or unevenly

spaced, so that a naive continuous relaxation misrepresents the geometry. A suitable reparameterization absorbs exactly that, and the GP posterior over the reparameterized domain can remain well calibrated. The failure mode we target is a property of the *response* on a domain that is already continuous: on a compositional simplex, phase boundaries introduce jump discontinuities in the target, synthesis-feasibility gaps leave the function undefined on subsets of the domain, and XRD-derived scores are thresholded. No reparameterization of the inputs restores posterior calibration in that case, because the violated assumptions — smoothness of the response, and support over the whole domain — are assumptions about the function, not about the coordinates. A useful test is whether the problem remains hard after the domain is made continuous and evenly spaced; for the landscapes studied here it does.

Why LLM-guided optimizers are not evaluated here. Recent work injects language-model priors into the optimization loop, either by fine-tuning on BO trajectories augmented with natural-language context (39) or by coupling LLM chemistry knowledge with evolutionary search for multi-objective materials discovery (40). These change *where prior knowledge enters* the loop; our question is whether a fitted posterior should remain the dominant decision signal once it is fitted. The axes are orthogonal, and in principle an LLM prior could be combined with either a GP acquisition or with ProSe/ATLAS. We do not evaluate these methods, for two reasons that would make a quick comparison misleading rather than informative. First, their advantage presumes a pretrained model that already encodes relevant domain knowledge, so a single benchmark measures the pairing of method and pretraining corpus rather than the method. Second, and more seriously for materials work, on public benchmarks that advantage is confounded with memorization of the very datasets used for evaluation: the retrospective hold-outs used here remove a region from a surrogate's training data, but they cannot remove it from a pretrained model's weights. Constructing an evaluation that controls this — for instance on unpublished experimental data — is future work we regard as necessary before such a comparison would mean anything.

Relative acceleration versus absolute saturation. Adesiji et al. (41) propose acceleration and enhancement factors as a common currency for self-driving-lab performance, reporting a median acceleration factor of $\sim$6, enhancement peaking at 10–20 experiments per dimension, and — most relevant here — acceleration *increasing* with parameter-space dimensionality, which they describe as a blessing of dimensionality. Read quickly, this appears to contradict the high-dimensional saturation we report in §5.1. The two are compatible, and the reason is instructive. Acceleration is defined *relative* to a reference strategy: as dimensionality grows, a naive reference degrades quickly, so the ratio can rise even while both arms approach uselessness in absolute terms. Our saturation finding is an absolute statement — beyond roughly six dimensions at $N{=}50$, every evaluated method, including random search, approaches the same floor — and its consequence is about the benchmark rather than about any method: once all arms sit at the floor, the benchmark has no power to separate them, so no per-method conclusion should be drawn from it. We therefore report it as a methodological negative result. The enhancement-factor peak at 10–20 experiments per dimension is consistent with this reading and with our budget choices: at $d{\geq}6$ and $N{=}50$ we are well below that regime. Separately, Liang et al. (27) find that random forests match anisotropic GPs across five experimental materials domains while both beat isotropic GPs, which supports the broader claim of this paper that fit to a landscape's assumptions, rather than surrogate family per se, is what drives performance.

J A Value-Weighted RF Blend and a Warmup-Delay Fix

This appendix reports an extension along a different axis from the rest of the paper: an alternative score blend that trades ProSe/SAGE's spatial-coverage emphasis for faster progress toward a good *value*, and a schedule fix for the late-budget cost that trade incurs. It is evaluated at the same 100-repetition protocol as the main results on all eleven benchmark datasets used in this paper, but it targets speed-to-target on a single best value rather than the multi-peak coverage the main text measures, so we keep it separate.

A whole-budget value-weighted variant. ProSe and SAGE both cap their k-NN/RF value term (RF: a random-forest value estimate fitted to the observations, used by the blend in place of the k-NN estimate) at a bounded fraction of the score ($w_p \leq 0.3$, §3, Appendix E). We evaluate an alternative that instead draws its distance-vs-value split from the *entire* remaining score budget:

$\mathrm{dist_weight}(t) = \max(0.1,\ 0.5(1 - t/N))$, so the k-NN/RF value terms rise from 0.5 to 0.9 of the score as the run progresses, ported onto both bases ("ProSe-RF" and "SAGE-RF" below) with matching min-max normalization of the distance score.

An early-speed, late-coverage trade-off. We measure speed-to-target with an area-under-the-best-so-far-curve ratio (variant AUC / base AUC) at budgets 20/50/100, a metric distinct from the recall and threshold-hit metrics used elsewhere in this paper, and pair it with the raw best-so-far value at each budget. For the unmodified blend on the ProSe base, the mean ratio exceeds 1 on every synthetic dataset at budgets 50 and 100 (three-peaks, 4-D, 7-D, 12-D, 20-D: 1.03–1.29 at budget 50, 1.09–1.14 at budget 100) but sits at parity or slightly below at budget 20 (0.98–1.13), and the gain shifts to later budgets as dimension grows. Trace inspection explains the mechanism: because the value terms already hold half the score at step 0, the search concentrates onto the current best region early and stops adding new regions thereafter — an early-speed, late-coverage trade-off rather than a uniform improvement. Its cost shows in the final value: on three-peaks the unmodified variant finishes budget 100 a mean 0.007 *below* its own base (100 repetitions), and on two real datasets (experimental LGPS and band gaps) the base is significantly ahead at budget 100.

A warmup-delay fix. We hold $\mathrm{dist_weight}$ at its floor (0.5) for a warmup window of W evaluations, then resume the same decay measured from the end of the window: $\mathrm{dist_weight}(t) = 0.5$ if $t < W$, else $\max(0.1,\ 0.5(1 - (t - W)/(N - W)))$. We evaluate this fix on both bases ("ProSe-RF-Warmup", "SAGE-RF-Warmup") at $W{=}20$, at 100 repetitions and budgets 20/50/100, on the five synthetic dimension-sweep datasets (three-peaks, 4-D, 7-D, 12-D, 20-D; peak hold-outs, 3/6/38/308/408 held-out peaks respectively), the five public look-up datasets of Appendix C (top-5% hold-out), and experimental LGPS (family hold-out) — 769 held-out cases per comparison. The unmodified blend on the ProSe base is run alongside as the reference.

Full-scale results. The warmup fix retains at least 93% of the unmodified variant's own AUC-ratio advantage in every one of the 66 (dataset, base, budget) cells (93.6–126%), and its mean AUC ratio on the synthetic datasets at budget 100 is 1.06–1.13 on the ProSe base and 1.15–1.41 on the SAGE base, with the ProSe-base gain smallest at 20-D (1.06) and the SAGE-base gain still 1.17 there. Second, the late-budget cost is removed on the SAGE base and reduced on the ProSe base: in a per-hold-out paired sign test on the best-so-far value at budget 100 (100 repetitions per held-out case; $2{,}304$ of $2{,}307$ cases have enough discordant pairs to reach significance), the variants are significantly ahead of their base in $2{,}223$ cases and the base is significantly ahead in 4. SAGE-RF-Warmup has 764 significant wins and *no* significant losses out of 768 powered cases; ProSe-RF-Warmup has 725 wins and 2 losses (LGPS and band gaps — the same two cases on which the unmodified blend also loses). The speed gain is not free early in high dimension: at budget 20 on 12-D the base is significantly ahead in 56–60 of 308 held-out peaks for both warmup variants, reversing by budget 50 (191 and 308 of 308 variant wins); at 20-D the early cost has largely gone (base ahead in 20 of 408 peaks on the ProSe base, none on the SAGE base; 358 and 408 of 408 variant wins by budget 50). On the real look-up datasets the picture is close to parity (0.98–1.12 mean AUC ratio); the clearest positive case is band gaps (1.05–1.11 at every budget, both bases), the ternary-glass benchmark is at its ceiling for all methods (0.999–1.003), and the SAGE-base LGPS ratio (1.08–1.12) is driven by a minority of repetitions (median exactly 1). Figure 16 shows the three-peaks curves: at step 100 the unmodified blend sits 0.007 below ProSe while the warmup variants finish $+0.022$ (ProSe base) and $+0.073$ (SAGE base) above their bases.

Scope. These results concern speed toward a single best value, not the multi-peak coverage the main text measures; on the same runs the blend carries a multi-peak-recall cost where that metric is powered (at budget 100 the base is significantly ahead in 22 held-out cases versus 6 for ProSe-RF-Warmup, and 6 versus 1 for SAGE-RF-Warmup; no 20-D cell is powered on that metric), which is the trade the mechanism makes by design and which we leave to future work. Only $W{=}20$ was evaluated. The SAGE-base variant has no unmodified reference of its own, so its retention figures are measured against the ProSe-base reference.

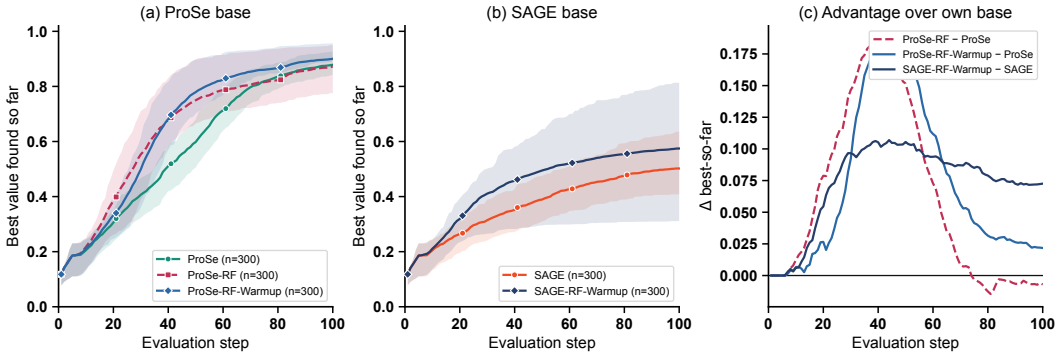


Figure 16: Value-weighted RF-blend variant and the warmup-delay fix, synthetic three-peaks, 100 repetitions $\times$ 3 held-out peaks, budget 100 (mean best-so-far, IQR band). (a) ProSe base: the original mechanism leads early (peak advantage near step 40) then falls behind ProSe by step 100. (b) SAGE base: the fixed variant leads throughout. (c) Each variant's advantage over its own base across the evaluation budget — the original mechanism's advantage crosses zero near step 75 and finishes negative; the warmup-delayed fix on either base stays positive through budget 100.