# DO CANDIDATE-POOL BENCHMARKS FAITHFULLY EVALUATE LLM OPTIMIZERS?

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

A benchmark for a generative optimizer should evaluate the designs that the optimizer proposes. In scientific optimization, however, new measurements are often unavailable, and an optimizer is instead scored by *Selecting* from a pool of measured candidates or by *Mapping* its proposals to the nearest stored design. We ask whether these candidate-pool protocols preserve conclusions about prospective optimization. Across four materials benchmarks, larger search steps by large language model (LLM) optimizers are associated with fewer evaluations to reach the target region. In a controlled prompt experiment, a one-line instruction changes this reaching time without changing the underlying model. These observations motivate the *trajectory-imprinting* hypothesis: a pool can inherit geometric structure from the optimization runs that produced it. To test fidelity beyond the pool, we directly evaluate optimizer-proposed designs using physics-based simulations of planar MOSFETs and GaAs deposition. In GaAs deposition, Gaussian-process (GP) optimizers reach lower nonuniformity under Generation, whereas LLM optimizers reach the pool optimum first under Selection. Direct evaluation of mapped proposals reveals changes in measured-and-feasible status in 38.4% of MOSFET and 19.0% of deposition pairs. Among pairs feasible on both sides, mean projection credit is larger for GP optimizers than for LLM optimizers. We construct core-set pools from designs collected during previous optimization runs, keeping the initial design and the best candidate fixed. These pools have different measured geometric properties, but optimizer orderings under Selection still differ from those under Generation. Proxy fidelity must therefore be assessed for the optimizer set, pool, protocol, and metric rather than assumed from pool geometry alone.

## 1 INTRODUCTION

An optimizer for scientific discovery must propose an experiment worth performing. Large language models (LLMs) offer a direct interface to this task: the objective is described in natural language, previous experiments and their outcomes are supplied as context, and a new design is requested (Yang et al., 2024; Ramos et al., 2023; Liu et al., 2024). Assessing this ability is difficult when each measurement requires laboratory work or an expensive simulation, so retrospective benchmarks make evaluation affordable by reusing a finite dataset of measured designs (Liang et al., 2021; Häse et al., 2021). Reusing a dataset, however, also changes the task that the optimizer is asked to solve.

Two protocols are commonly used to bridge this gap. In *Selection*, the optimizer chooses among the visible inputs of a candidate pool and receives the stored outcome. In *Mapping*, the optimizer proposes an arbitrary design, but receives the recorded outcome of the nearest pool member (Figure 1). Both protocols support informative studies of scientific optimization (Wang et al., 2026; Demir et al., 2026), and both are appropriate when choosing among fixed alternatives is the intended application. The stronger claim that they evaluate the ability to *Generate* useful new designs requires evidence that their conclusions transfer to direct evaluation.

There are two reasons why such transfer may fail. First, a candidate pool need not be a neutral sample of the design space. A pool may be assembled from earlier experiments, optimization trajectories, and selection decisions. When its candidates come from optimization trajectories, they may concentrate along paths toward regions found rewarding by those searches. We call this inherited



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.

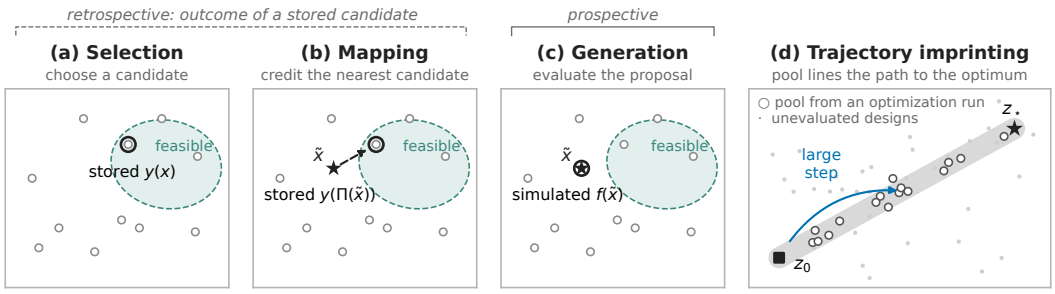


Figure 1: **Evaluation protocols (schematic).** (a–c) The same pool and proposal $\tilde{x}$ under each protocol; here Mapping credits an infeasible proposal with a feasible stored outcome. (d) Trajectory imprinting: a pool from an optimization run lies along the segment from $z_0$ to $z_\star$.

geometric structure *trajectory imprinting* (TI). Such structure may provide information about the objective through candidate locations, without establishing that an optimizer can generate equally useful designs outside the pool. Second, Mapping substitutes one design for another. A nearby design need not have a similar objective value or satisfy the same constraints, and a substitution that favors some optimizers more than others distorts a comparison even when every optimizer is subject to the same rule.

Our question is therefore one of *benchmark fidelity*: when a retrospective protocol ranks one optimizer above another, does prospective evaluation support the same conclusion? This differs from asking whether an LLM is competitive within a finite pool, and it complements tests of whether LLM optimizers use experimental feedback at all (Gupta et al., 2025; Kristiadi et al., 2024). Neither a strong pool score nor sensitivity to feedback establishes that a proposed design would succeed when evaluated directly. Offline model-based optimization confronts a related distinction between data-supported and prospectively evaluated designs (Trabucco et al., 2022); here we study the consequences of replacing the evaluated proposal and restricting the admissible search space.

Our contributions are as follows. (i) Across 12 LLMs and a controlled prompt intervention, we show that candidate-pool performance is associated with search step size, motivating TI as a hypothesis rather than establishing it as a causal mechanism. (ii) Using two constrained simulation tasks as oracles, we show that optimizer comparisons under Selection and Mapping can differ from those under Generation. (iii) We directly measure the substitution performed by Mapping and find frequent measured-and-feasible status changes and optimizer-dependent projection credit. (iv) We show that coverage-based core-set pools change the measured geometry but still yield Selection orderings that differ from Generation. Geometric coverage of an optimization-derived reservoir alone therefore does not certify proxy fidelity. Performance at choosing among recorded designs and performance at proposing new designs are separate claims whose agreement must be measured.

## 2 Related Work

**LLMs as optimizers.** LLMs have been used as black-box optimizers through prompting (Yang et al., 2024), as surrogates or acquisition components within Bayesian optimization (Liu et al., 2024; Ramos et al., 2023), and as general-purpose priors for experiment design (Song et al., 2024). LLM-driven search has also been applied directly to materials and chemical optimization (Reinhart & Statt, 2024; MacKnight et al., 2025; Ranković et al., 2026; Abhyankar et al., 2026). Evaluations in chemistry and materials science have questioned whether LLM components add value over conventional surrogates (Kristiadi et al., 2024) and whether LLM optimizers respond to experimental feedback (Gupta et al., 2025). We propose no new optimizer; we ask whether the benchmarks used to compare optimizers measure the claimed capability.

**Pool-based benchmarks and core-sets.** Because experiments are expensive, optimizers for materials and chemistry are frequently compared on lookup datasets of measured designs (Liang et al., 2021; Häse et al., 2021). Recent LLM studies adopt the same practice, either by restricting the optimizer to the pool (Demir et al., 2026) or by mapping free-form proposals to the nearest measured design (Wang et al., 2026); these protocols are the object of our study. Offline model-based optimization benchmarks (Trabucco et al., 2022) likewise separate data-supported from prospec-

tively evaluated designs, but score a final proposal with an oracle rather than replacing it with a stored neighbor. We use farthest-first $k$-center-greedy selection (Sener & Savarese, 2018) to build three pool variants that differ in eligibility and allocation rules (Section 6.2). These variants test coverage-based construction within the evaluated reservoir.

# 3 Motivating Observation: Candidate-Pool Performance Tracks Search Step Size

**Materials candidate pools.** Four retrospective datasets are used as candidate pools: BN, growth recipes for hexagonal boron nitride films from a laboratory MOCVD campaign (55 rows, 13 inputs; maximize the recorded score); Membrane, polysulfone membranes cast by nonsolvent-induced phase separation (73 rows, 10 inputs; maximize elastic modulus) (Wang et al., 2026); P3HT/CNT, poly(3-hexylthiophene)/carbon nanotube composite films (233 rows, 5 inputs; maximize electrical conductivity) (Bash et al., 2021); and Perovskite, halide perovskite compositions (139 rows, 3 inputs; minimize an instability index that measures compositional degradation) (Sun et al., 2021). The latter three are the public datasets used by Wang et al. (2026). Outcomes for new designs are unavailable, so these pools support Selection only, and rows, including repeated input configurations, are the evaluation units. The *target region* of a pool is its top-10% objective region; the number of evaluations to the target region counts every evaluation up to and including the one that first enters it, and the final benchmark target is a separate, stricter endpoint (Appendix C).

Figure 2 summarizes the materials experiments. Panel (a) is a leaderboard of 12 LLMs and GP-UCB on BN by mean evaluations to the final target, and panel (b) relates these outcomes to search behavior. We define the *step size* of a decision as the percentile of the selected candidate's distance from the current best design among all currently available candidates, so that it measures how far the optimizer moves relative to what the pool offers rather than raw displacement. Across the 12 LLMs, the mean step size before reaching the target region is strongly associated with fewer evaluations to reach it (Spearman $\rho = -0.97$). Panel (c) shows that within a model, adding prior information (PI), retrieval from domain papers or textbooks (RAG), or both to the prompt shifts early step size and reaching time together across 48 model–condition contrasts (conditions in Appendix C.1). Panel (e) shows that the Sol–Luna ordering by step size and reaching time is preserved on the three other pools.

A one-line prompt intervention changes reaching time without changing the model or supplying new measurements. On BN, adding a `far` instruction to the prompt reduces the mean number of evaluations to the target region from 1.89 to 1.29 for Sol and from 4.36 to 1.67 for Luna (GPT-5.6-Luna), whereas a `local` instruction raises these means to 10.44 and 8.78 (panel (d); 45 runs per arm; instructions in Appendix C). The instruction does not identify the optimum. Faster entry into the target region does not necessarily improve the final-target metric: for Sol (GPT-5.6-Sol), the mean number of evaluations to the final target rises from 6.44 under the control prompt to 9.42 under `far`.

These findings are consistent with TI: in a pool assembled from earlier searches, the good region lies at a distance from the initialization and many candidates are arranged along the way, so an optimizer that moves far in the direction suggested by the pool reaches the target region quickly. The evidence is behavioral, since step-size averages depend on the trajectory and its stopping time and a prompt can change behavior in ways other than step size, so causal mediation by geometry is not established. More fundamentally, no experiment restricted to measured rows can reveal whether an optimizer could generate useful designs outside the pool; that missing outcome motivates the simulator study.

# 4 Evaluation Protocols and Experimental Setup

## 4.1 Three evaluation protocols

Let $\mathcal{X}$ be a bounded design space and $\mathcal{P} \subset \mathcal{X}$ a finite pool with recorded outcomes. An evaluator returns an objective $f(x)$ and constraint measurements when available, or a failure record; $F(x)$ indicates that the evaluation succeeds and satisfies the task's constraints. At step $t$ (one evaluation), the optimizer uses the observation history $H_t$, and $\mathcal{P}_t$ denotes the pool rows not yet observed.

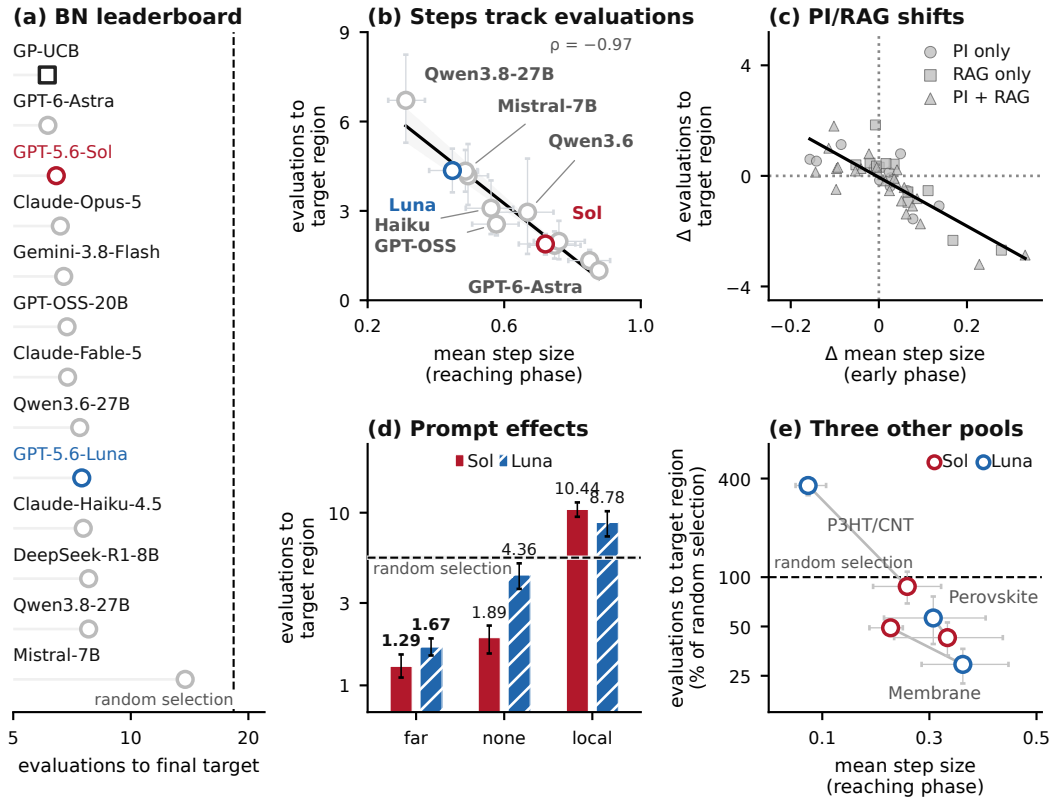


Figure 2: **Candidate-pool performance tracks search step size in materials pools.** (a) BN leaderboard: mean evaluations to the final target for 12 LLMs and GP-UCB; the dashed line is the random-selection expectation. (b) Mean step size before reaching the target region versus evaluations to the target region across the 12 LLMs ($\rho = -0.97$); the line is an ordinary least-squares fit and shading its pointwise 95% initial-block bootstrap interval. (c) Within-model changes under prior information (PI), domain-paper/textbook retrieval (RAG), or both: 48 contrasts across six models relative to the names-revealed baseline (Appendices C and C.1). (d) BN prompt interventions (`far`, none, `local`; none is the control prompt without a search instruction) for Sol and Luna, 45 runs per arm; bars are means with 95% intervals. (e) Sol and Luna on Membrane, P3HT/CNT, and Perovskite, 50 runs per mean; evaluations are expressed relative to the exact expectation for random selection without replacement, and gray segments connect the two models within a pool. Step size is the current-best-relative distance percentile among available candidates, averaged before entry into the target region in (b,e) and over the early phase in (c); error bars are 95% intervals.

**Generation** proposes $x_t \in \mathcal{X}$ and evaluates that proposal with the simulator. **Selection** chooses $x_t \in \mathcal{P}_t$ from the visible inputs and reveals its stored outcome. **Mapping** proposes $\tilde{x}_t \in \mathcal{X}$ but reveals the outcome of

$$x_t = \Pi_{\mathcal{P}_t}(\tilde{x}_t) = \arg\min_{x \in \mathcal{P}_t} d\big(z(x), z(\tilde{x}_t)\big), \tag{1}$$

where $z(\cdot)$ is a task-dependent coordinate transformation and $d$ a metric in the transformed space; we use RMS Euclidean distance, with task-specific transforms given in Appendix D. The mapped input and outcome enter the next history, so Mapping changes both the credited design and subsequent feedback. Unobserved outcomes remain hidden, Mapping does not expose the pool, and both pool protocols sample without replacement.

### 4.2 Tasks and outcome status

Unlike the materials pools, the two tasks below are defined by physics-based simulators, so any proposed design can be evaluated directly under Generation; Selection and Mapping can also be evaluated on the same tasks once a candidate pool is specified (Section 4.3).

**Planar MOSFET process optimization (MOSFET).** Synopsys Sentaurus Process and Device evaluate a nine-input fabrication process that maximizes on-current subject to a leakage constraint (Synopsys, Inc., 2024):

$$\max_x I_{\text{on}}(x) \quad \text{subject to} \quad I_{\text{off}}(x) \leq 1.71 \times 10^{-14}\ \text{A}, \tag{2}$$

A valid simulation above the leakage limit is measured infeasible; a process, device, or numerical failure without a valid measurement is failed. Simulation and current-extraction details are given in Appendix B.

**GaAs deposition process optimization.** A COMSOL Multiphysics model evaluates a seven-input MOCVD process that minimizes growth-rate nonuniformity subject to a minimum mean growth rate (COMSOL AB, 2023). With $g$ the growth-rate profile across the wafer,

$$\min_x \text{NU}(x) = 100\,\frac{g_{\max}(x) - g_{\min}(x)}{g_{\max}(x) + g_{\min}(x)} \quad \text{subject to} \quad \bar{g}(x) \geq 3\ \mu\text{m/h}. \tag{3}$$

Here, $\bar{g}(x)$ denotes the mean growth rate across the wafer. A valid result below the growth-rate threshold is measured infeasible; solver or numerical-validity failure is failed (Appendix B).

**Failures are retained.** Measured infeasible points keep their physical outputs; failed points keep their inputs and diagnostics with missing outputs, and no objective value is imputed. Both consume evaluation budget and cannot improve the feasible incumbent. In the binary status audit, $F = 0$ for both, so a change in $F$ denotes a change in measured-and-feasible status, not necessarily a measured constraint crossing.

### 4.3 Optimizers, trajectory pools, and metrics

**Optimizers and Generation runs.** We compare two LLM optimizers, Luna and Sonnet (Claude Sonnet 5), with two Gaussian-process (GP) baselines, GP-UCB (Srinivas et al., 2010) and GP-EI (Jones et al., 1998). The materials experiments additionally include Sol (GPT-5.6-Sol) and ten further LLMs listed in Figure 2. Each simulator task has five Generation repeats per optimizer with 50 new evaluations each, all starting from the same task-specific initial design, which is supplied as a free observation. LLMs receive real parameter names and the chronological observation history, with tools and retrieval disabled. MOSFET uses a joint GP over on-current, leakage, and measurement success; GaAs deposition uses separate objective and constraint GPs, and the retrospective GP baselines reuse the corresponding formulation (Appendix A).

**Trajectory pools A–E.** For each simulator task we construct five 50-candidate pools, each consisting of the designs evaluated along one optimization trajectory from the shared initial design: Pool A from a human expert, Pool B from Luna, Pool C from Sonnet, Pool D from GP-UCB, and Pool E from GP-EI, all excluding the free initial observation. Each optimizer is run on each pool under Selection and under Mapping with five repeats, giving 25 retrospective runs per optimizer and protocol on fixed case-study pools.

**Best-feasible trajectories.** Comparing protocols requires the same performance quantity on both sides. For a run, let $B_t$ be the best feasible incumbent after evaluation $t$: the best feasible on-current for MOSFET and the best feasible nonuniformity for GaAs deposition, undefined before the first feasible observation. The free initial observation defines $B_0$ but is not counted as an evaluation. Pool runs stop at the exact feasible pool optimum, which is carried forward to evaluation 50. For each optimizer and protocol we plot the median of $B_t$ over five runs for $t = 1, \ldots, 50$, left undefined until all five runs have a feasible incumbent; the curves show both the attainable level and the speed of improvement.

**Proposal–projection audit.** For every recorded Mapping decision, we evaluate the original proposal with the simulator and pair it with the stored outcome that was credited; duplicate proposals reuse the same verified evaluation, and nothing is fed back to the optimizer. Objective differences use only pairs feasible on both sides, averaged within a trajectory and then equally across the 25 trajectories per optimizer; feasibility-status changes use all pairs, including failures.

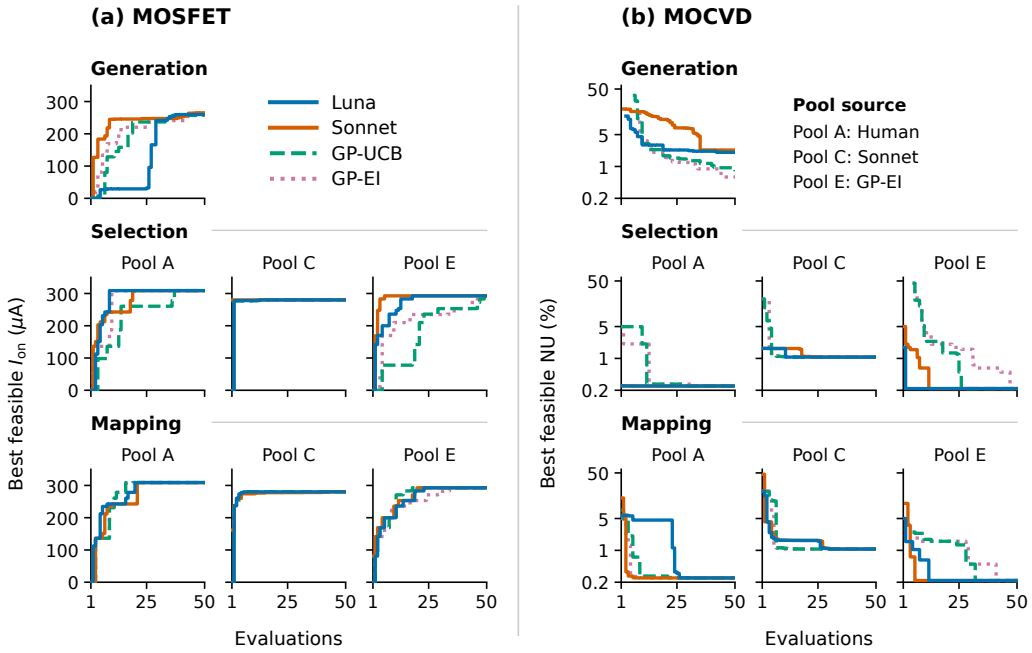


Figure 3: Optimizer progress under Generation, Selection, and Mapping for (a) MOSFET and (b) GaAs deposition. Pool protocols show A, C, and E; all five pools are in Appendix F. Curves are median best-feasible incumbents over five runs for evaluations 1–50. MOSFET $I_{\text{on}}$ is maximized on a linear axis and GaAs NU minimized on a logarithmic axis. Medians remain undefined until all five runs have a feasible incumbent; early-stopped pool runs carry the optimum through evaluation 50.On MOSFET Pool C (from an LLM run), 22 of 26 feasible candidates lie within 5% of the optimum, so best-feasible trajectories are flat from the first evaluation and do not discriminate among optimizers.

## 5 Simulator Oracles Expose Limits of Proxy Fidelity

### 5.1 Optimizer Comparisons Differ Between Generation and Pool Protocols

An oracle supports two comparisons that a fixed dataset cannot: running Generation reveals what each optimizer achieves when its proposals receive their own outcomes, and evaluating Mapping proposals reveals what the projection substituted for those outcomes. Figure 3 reports the first comparison as median best-feasible trajectories, with Selection and Mapping shown for Pools A, C, and E; all five pools are shown in Appendix F.

Optimizer comparisons under the pool protocols do not consistently follow Generation. In GaAs deposition, GP-EI and GP-UCB attain lower nonuniformity under Generation than Luna and Sonnet, with median endpoints of 0.59% and 0.78% versus 2.04% and 2.29%, respectively. Under Selection, however, the LLM optimizers identify the pool optimum substantially earlier. On Pool A, Luna and Sonnet reach the optimum at evaluations 2 and 7, compared with 31 and 33 for GP-UCB and GP-EI; on Pool E, the corresponding evaluations are 2 and 12 for the LLMs and 26 and 47 for the GPs. Thus, the optimizers that generate the best designs under direct evaluation are not necessarily those that identify the best available candidate most quickly under Selection. For MOSFET, Section 6.3 separately compares core-set Selection and Generation orderings using a common run-level metric.

A different discrepancy appears in MOSFET. Under Generation, Luna improves latest and remains near 30 $\mu$A until evaluation 26, whereas under Mapping on Pool A the differences among the four optimizers are strongly compressed, with all four median trajectories reaching the pool optimum between evaluations 18 and 21. Because Mapping replaces each proposed design with a stored candidate and feeds the mapped outcome back into the subsequent trajectory, these trajectory differences alone do not identify the effect of the substitution. We therefore examine the proposal–projection pairs directly in the next section.

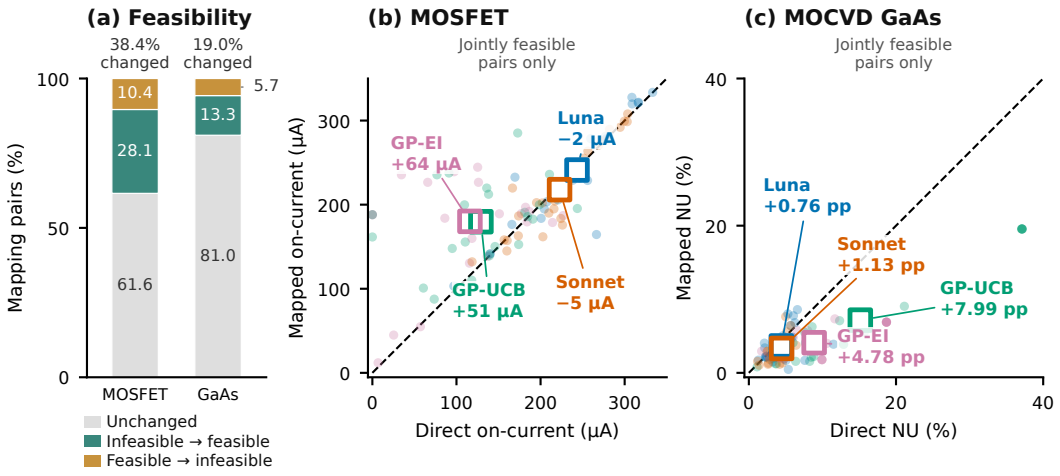


Figure 4: Mapping changes feasibility status and gives optimizer-dependent objective credit. (a) Proposal-to-projection status changes over all pairs, with failures counted as not feasible: 38.4% of MOSFET and 19.0% of GaAs pairs change status. (b,c) For jointly feasible pairs, faint points are trajectory means and large squares are means over 25 trajectories per optimizer (Pools A–E $\times$ five repeats); dashed lines denote equality. Labels report mean projection credit, with positive values indicating better credited performance. Objective means use only the 785 MOSFET and 1,777 GaAs jointly feasible pairs; missing measurements are not imputed. These are substitutions along Mapping histories, not trajectories under direct feedback.

Selection evaluates how efficiently an optimizer identifies good members of an existing candidate set, whereas Generation evaluates the optimizer's own proposed designs. Mapping allows arbitrary proposals but evaluates a stored candidate in place of each proposal. Figure 3 therefore shows that comparative performance under candidate-pool protocols need not be preserved when the optimizers' own designs are evaluated directly.

### 5.2 Projection changes feasibility and allocates unequal objective credit

The audit contains 2,063 MOSFET and 2,344 GaAs deposition proposal–projection pairs. Measured-and-feasible status differs between proposal and projection in 793 MOSFET pairs (38.4%) and 445 GaAs deposition pairs (19.0%; Figure 4a). In MOSFET, 214 transitions lose feasible status and 579 gain it; in GaAs deposition the counts are 134 and 311. These totals include failed evaluations as $F = 0$ and quantify the status that the benchmark credits; transitions among feasible, measured infeasible, and failed outcomes are given by optimizer in Appendix G.

Objective comparisons use the 785 MOSFET and 1,777 GaAs deposition pairs that are feasible on both sides. With $\Delta_f = f(\tilde{x}) - f(\Pi(\tilde{x}))$, Figure 4 reports the *projection credit* $C = -\Delta_f$ for the maximized on-current and $C = \Delta_f$ for the minimized nonuniformity, so that positive credit means that Mapping assigns a better value than direct evaluation of the proposal would. The credit is asymmetric across optimizers. In MOSFET, the mean run-level credits are $-2.15$ and $-4.76\,\mu$A for Luna and Sonnet but $+50.63$ and $+64.13\,\mu$A for GP-UCB and GP-EI. In GaAs deposition all four means are positive, but the GP credits (7.99 and 4.78 percentage points) exceed the LLM credits (0.76 and 1.13). Projection yields larger mean credit for the tested GP optimizers than for the tested LLM optimizers on jointly feasible pairs. This credit is measured relative to each optimizer's own proposals and is distinct from its Mapping ordering or its performance under Generation feedback.

This is direct evidence that projection error is not neutral across optimizers on their Mapping histories. The audit does not reconstruct Generation, because the audited proposals were produced after mapped feedback and replacing that feedback would change later proposals; nor do conditional objective means describe pairs excluded by feasibility, which is why the status transitions are reported alongside the conditional credit.

# 6 DO COVERAGE-BASED CORE-SET POOLS RECOVER GENERATION AGREEMENT?

**Geometry diagnostics.** To characterize pool geometry relevant to the TI hypothesis, we use two descriptive statistics of a pool relative to its initial design $z_0$ and best observed design $z_\star$. *Alignment* $A$ is the median cosine between $z_i - z_0$ and $z_\star - z_0$ over candidates $i$, excluding the target and zero-length vectors, and the *target-distance percentile* is the percentile of $\|z_\star - z_0\|$ among the initial-to-candidate distances. High alignment with a distant target is consistent with candidates being arranged along a path from the initialization toward the optimum, as can occur in pools assembled from search trajectories (Appendix D). Appendix H provides a reference-path formulation of TI that accommodates curved paths and clarifies its relation to these endpoint-based diagnostics.

## 6.1 TRAJECTORY POOLS SHOW GEOMETRY CONSISTENT WITH THE TI HYPOTHESIS

Figure 5 places every pool in the plane of alignment $A$ and target-distance percentile. The four materials pools have alignment of approximately $0.65$, $0.53$, $0.98$, and $0.90$ for BN, Membrane, P3HT/CNT, and Perovskite, with target-distance percentiles of $87.7$, $81.4$, $84.4$, and $77.5$: their best designs lie far from the benchmark initializations, and most candidates point toward them. The MOSFET trajectory pools show the same signature, with alignment between $0.78$ (Pool D) and $1.00$ (Pools B and C). This is the geometry under which large steps are rewarded, consistent with Section 3.

## 6.2 COVERAGE-BASED CORE-SET POOLS

To test whether changing the geometry associated with TI improves agreement with Generation, we construct three 50-candidate coverage-based pools from a reservoir of 5,798 evaluated MOSFET designs, while fixing the initial design and the Pool A optimum ($308.903\,\mu$A). After excluding the initial design and feasible designs better than this optimum, candidates are selected by farthest-first $k$-center greedy selection (Sener & Savarese, 2018):

$$x_{j+1} = \arg\max_{x \in \mathcal{R}_j \setminus C_j} \min_{c \in C_j} \|z(x) - z(c)\|_2, \qquad C_{j+1} = C_j \cup \{x_{j+1}\}. \tag{4}$$

We consider three constructions. **Pure** selects from all eligible designs, maximizing geometric coverage while retaining infeasible and failed evaluations. **Feasible-only** restricts selection to feasible designs, removing feasible-region discovery from Selection. **Stratified** enforces a mixture of feasible quality levels, measured-infeasible designs, and failures to avoid either extreme. Exact eligibility counts and stratum allocations are given in Appendix E.

Because the optimum filter and stratification use recorded outcomes, these constructions are not outcome-blind, and they remain restricted to support that was previously evaluated.

## 6.3 CORE-SETS CHANGE THE MEASURED GEOMETRY BUT NOT THE RANK AGREEMENT

The three core-set pools of Section 6.2 occupy a different region of the plane: alignment falls to $0.02$, $0.20$, and $0.33$ for Pure, Feasible-only, and Stratified, and the fixed optimum lies at target-distance percentiles of $1$, $3$, and $13$. All simulator pools in Figure 5 are measured in the declared-bound coordinates of the Mapping distance (Appendix D).

For this comparison, we summarize each run by $\mathrm{AUC}_{50}$, the mean best-feasible incumbent over evaluations 1–50, and compare the resulting Selection and Generation optimizer orderings using Spearman's $\rho$. The Selection–Generation correlations are $+0.40$, $-0.20$, and $-0.40$ for Pure, Feasible-only, and Stratified, respectively (Appendix E).

The reservoir itself was produced by optimization runs from the same initial design, so core-set construction remains restricted to previously evaluated support. Early discovery of the fixed optimum can also compress $\mathrm{AUC}_{50}$ differences; optimizer-specific discovery times are reported in Appendix E.

This comparison does not separate inherited reservoir geometry from the effects of fixing the optimum and changing candidate composition. Thus, changing these geometric properties alone does

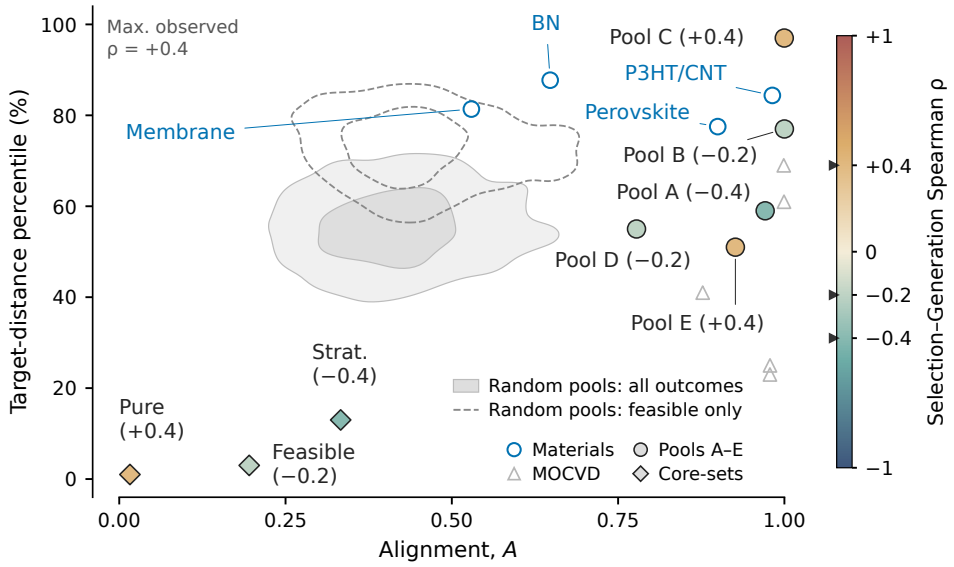


Figure 5: Geometry and Selection–Generation agreement across trajectory and core-set pools. Alignment $A$ versus target-distance percentile for four materials pools, five GaAs trajectory pools, five MOSFET trajectory pools A–E, and three MOSFET core-sets. Trajectory pools use their own best feasible design as the reference target; core-sets use the fixed Pool A optimum. Marker color and labels give Selection–Generation Spearman $\rho$ across the four optimizers under $\mathrm{AUC}_{50}$. Filled and dashed contours show the 50% and 90% density regions of 1,000 random 50-candidate MOSFET pools from the full reservoir and feasible subset. These random pools use the Pool A optimum only as a fixed reference target and may contain better designs; no optimizer performance was evaluated on them. Coordinates are defined in Appendix D.

not establish agreement with Generation; fidelity should be tested directly under the same metric and evaluation budget.

## 7 Discussion

Retrospective pools remain useful for comparing search over fixed candidates; our results concern the further inference to generative design.

We recommend reporting the evaluated and fed-back object and, when an oracle is available, proposal–projection feasibility and objective differences while preserving failures and conditional denominators. Rank transfer should be tested under matched metrics and budgets, with pool provenance, initialization, coordinate scaling, and attainable optimum disclosed.

**Limitations.** Prospective evidence is limited to two simulation models, four optimizers, one shared initialization per task, and five Generation repeats per optimizer; it does not establish performance in physical experiments. Rank estimates from four optimizers are coarse, and the five trajectory pools are case studies rather than a sample of possible pools. The geometry diagnostics and core-set comparisons do not identify a unique causal mechanism, so TI remains a hypothesis supported by behavioral and geometric evidence.

## 8 Conclusion

Candidate-pool protocols evaluate outcomes attached to previously evaluated designs, whereas Generation evaluates optimizers' own proposals. Across two simulation tasks, optimizer comparisons under Selection and Mapping are not consistently preserved under Generation, and Mapping can change feasibility status and assign optimizer-dependent projection credit. Proxy fidelity should therefore be assessed for the optimizer set, pool, protocol, and metric rather than inferred from retrospective performance alone.

### AI USE STATEMENT

Large language models are the object of study in this paper: the LLMs listed in Figure 2 and Section 4.3 were queried as optimizers under the protocols and prompts described in Sections 3–4 and Appendices A and C. In addition, LLM-based assistants were used to help write analysis and plotting scripts and to edit the text of the manuscript. All experimental designs, analyses, numerical results, and claims were checked by the authors, who take full responsibility for the content of the paper.

### ETHICS STATEMENT

This work involves no human subjects, personal data, or animal experiments. The BN data come from a laboratory hBN growth campaign described in Appendix C, and the Membrane, P3HT/CNT, and Perovskite datasets are publicly released with the cited work; all are used as recorded. The simulator tasks are computational process models run with licensed software, and no devices or films were fabricated for this study. The paper evaluates benchmarking methodology rather than introducing new optimization capabilities, and its intended effect is to reduce overstated claims about LLM optimizers in scientific discovery.

### REPRODUCIBILITY STATEMENT

The supplementary material contains the frozen trajectories, the pool definitions A–E and core-set pools, the proposal–projection ledger, the recorded prompt packages, the figure-source tables, and the scripts that regenerate every figure and table in this paper from recorded outcomes without new LLM or simulator calls. Appendices A–H specify every evaluation, aggregation, simulator, coordinate, geometry, core-set, and ranking convention. New physical simulations require licensed Synopsys Sentaurus and COMSOL Multiphysics installations.

## REFERENCES

Nikhil Abhyankar, Sanchit Kabra, Saaketh Desai, and Chandan K. Reddy. LLEMA: Evolutionary search with LLMs for multi-objective materials discovery. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=TIqzhBvCNB.

Daniil Bash, Yongqiang Cai, Vijila Chellappan, Swee Liang Wong, Xu Yang, Pawan Kumar, Jin Da Tan, Anas Abutaha, Jayce J. W. Cheng, Yee-Fun Lim, et al. Multi-fidelity high-throughput optimization of electrical conductivity in P3HT-CNT composites. *Advanced Functional Materials*, 31(36):2102606, 2021. doi: 10.1002/adfm.202102606.

COMSOL AB. COMSOL Multiphysics v. 6.2. COMSOL AB, Stockholm, Sweden. https://www.comsol.com, 2023.

Dino-Rober Demir, Florian Le Bronnec, and Rio Yokota. LLMs as acquisition policies for finite-pool materials optimization: A controlled study, 2026. URL https://arxiv.org/abs/2608.19790.

John C. Gower. A general coefficient of similarity and some of its properties. *Biometrics*, 27(4): 857–871, 1971.

Rushil Gupta, Jason Hartford, and Bang Liu. LLMs for Bayesian optimization in scientific domains: Are we there yet? In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 15482–15510. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.findings-emnlp.838. URL https://aclanthology.org/2025.findings-emnlp.838/.

Florian Häse, Matteo Aldeghi, Riley J. Hickman, Loïc M. Roch, Melodie Christensen, Elena Liles, Jason E. Hein, and Alán Aspuru-Guzik. Olympus: A benchmarking framework for noiseless optimization of physical and chemical experiments. *Machine Learning: Science and Technology*, 2(3):035021, 2021. doi: 10.1088/2632-2153/abedc8.

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

Agustinus Kristiadi, Felix Strieth-Kalthoff, Marta Skreta, Pascal Poupart, Alán Aspuru-Guzik, and Geoff Pleiss. A sober look at LLMs for material discovery: Are they actually good for Bayesian optimization over molecules? In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, 2024. URL https://arxiv.org/abs/2402.05015.

Qiaohao Liang, Aldair E. Gongora, Zekun Ren, Armi Tiihonen, Zhe Liu, Shijing Sun, James R. Deneault, Daniil Bash, Flore Mekki-Berrada, Saif A. Khan, Kedar Hippalgaonkar, Benji Maruyama, Keith A. Brown, John III Fisher, and Tonio Buonassisi. Benchmarking the performance of Bayesian optimization across multiple experimental materials science domains. *npj Computational Materials*, 7:188, 2021. doi: 10.1038/s41524-021-00656-9.

Tennison Liu, Nicolás Astorga, Nabeel Seedat, and Mihaela van der Schaar. Large language models to enhance Bayesian optimization. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://arxiv.org/abs/2402.03921.

Robert MacKnight, Jose Emilio Regio, Jeffrey G. Ethier, Luke A. Baldwin, and Gabe Gomes. Pretrained knowledge elevates large language models beyond traditional chemical reaction optimizers. *arXiv preprint arXiv:2509.00103*, 2025. doi: 10.48550/arXiv.2509.00103.

Mayk Caldas Ramos, Shane S. Michtavy, Marc D. Porter, and Andrew D. White. Bayesian optimization of catalysts with in-context learning. *arXiv preprint arXiv:2304.05341*, 2023. URL https://arxiv.org/abs/2304.05341.

Bojana Ranković, Ryan-Rhys Griffiths, and Philippe Schwaller. Large language models as uncertainty-calibrated optimizers for experimental discovery. *Nature Machine Intelligence*, 8: 1466–1477, 2026. doi: 10.1038/s42256-026-01283-z.

Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 3982–3992, 2019.

Wesley F. Reinhart and Antonia Statt. Large language models design sequence-defined macromolecules via evolutionary optimization. *npj Computational Materials*, 10:262, 2024. doi: 10.1038/s41524-024-01449-6.

Ozan Sener and Silvio Savarese. Active learning for convolutional neural networks: A core-set approach. In *International Conference on Learning Representations*, 2018. URL https://arxiv.org/abs/1708.00489.

Xingyou Song, Yingtao Tian, Robert Tjarko Lange, Chansoo Lee, Yujin Tang, and Yutian Chen. Position: Leverage foundational models for black-box optimization. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, 2024. URL https://arxiv.org/abs/2405.03547.

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

Shijing Sun, Armi Tiihonen, Felipe Oviedo, Zhe Liu, Janak Thapa, Yicheng Zhao, Noor Titan P. Hartono, Anuj Goyal, Thomas Heumueller, Clio Batali, et al. A data fusion approach to optimize compositional stability of halide perovskites. *Matter*, 4(4):1305–1322, 2021. doi: 10.1016/j.matt.2021.01.008.

Synopsys, Inc. Sentaurus Process user guide, version 2024.03. Synopsys, Inc., Mountain View, CA, 2024.

Brandon Trabucco, Xinyang Geng, Aviral Kumar, and Sergey Levine. Design-bench: Benchmarks for data-driven offline model-based optimization. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 21658–21676. PMLR, 2022. URL https://proceedings.mlr.press/v162/trabucco22a.html.

Hongchen Wang, Rafael Espinosa Castañeda, Jay R. Werber, Yao Fehlis, Edward Kim, and Jason Hattrick-Simpers. Training-free active learning framework in materials science with large language models. *npj Computational Materials*, 12:265, 2026. doi: 10.1038/s41524-026-02136-4.

Chengrun Yang, Xuezhi Wang, Yifeng Lu, Hanxiao Liu, Quoc V. Le, Denny Zhou, and Xinyun Chen. Large language models as optimizers. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=Bb4VGOWELI.

# A PROTOCOL AND METRIC DETAILS

**Models and prompts.** The simulator LLMs are `gpt-5.6-luna` and `claude-sonnet-5`, both at medium reasoning effort, accessed through their recorded command-line runtimes. Each request is a system message and a JSON user message containing the objective and constraint, the evaluation budget of 50, the chronological observations with real parameter names, the declared input bounds, and the required output schema; unobserved outcomes and pool-source identities are not supplied. Inputs are shown as raw physical values without rescaling or rounding; in MOSFET both times are expressed in seconds, including the gate-oxidation time that the simulator deck records in minutes. The simulator prompts state targets of $300\,\mu$A for MOSFET and $0.2\%$ NU for GaAs deposition. These are prompt constants rather than feasibility thresholds or stopping rules, and Generation always continues to 50 new evaluations.

Mapping sends the same request as Generation and credits the nearest unselected candidate. Selection reuses the Generation system message, replaces the task instruction and output schema with the choice of one candidate identifier, and adds the unobserved candidates under neutral identifiers with the recorded presentation schedule. Proposals that violate the input schema are handled by the harness before simulator evaluation and are not recorded as physical failures; a valid simulator attempt that fails remains an evaluated observation. Repeated rows and reuse of the same audited proposal evaluation do not create independent physical measurements. The full text of every recorded request is included in the supplementary material.

**Initial designs.** The MOSFET initial design has well concentration $10^{18}\,\mathrm{cm}^{-3}$, gate oxidation for 600 s at $950\,^{\circ}$C, LDD dose $10^{14}\,\mathrm{cm}^{-2}$ at 30 keV, source/drain dose $10^{15}\,\mathrm{cm}^{-2}$ at 15 keV, and a 1 s anneal at $1000\,^{\circ}$C; it is feasible with $I_{\mathrm{on}} = 1.65 \times 10^{-12}$ A and $I_{\mathrm{off}} = 6.95 \times 10^{-16}$ A. The GaAs deposition initial design has susceptor temperature 1173 K, pressure 20 torr, inlet velocity 0.5 m/s, tilt $2.5^{\circ}$, inlet height 0.03 m, nitrogen fraction 0, and wafer center 0.14 m; it is measured infeasible with $\mathrm{NU} = 93.4\%$ and $\bar{g} = 0.41\,\mu$m/h.

**GP models and acquisition.** MOSFET uses one correlated three-output GP over affine-scaled $I_{\mathrm{on}}$, transformed leakage, and a Bernoulli-probit measurement-success latent. The covariance combines a Matérn-5/2 ARD input kernel with a rank-2-plus-diagonal output covariance, fitted with a joint variational objective (Adam, 1,200 steps, followed by L-BFGS). Missing currents contribute no regression likelihood terms; failures inform measurement success. With feasible utility $U = \max(I_{\mathrm{on}}, 0)\,\mathbb{1}(\text{measured and } I_{\mathrm{off}} \le 1.71 \times 10^{-14}\,\mathrm{A})$, GP-UCB maximizes $\mathbb{E}[U] + 2\sqrt{\mathrm{Var}[U]}$ and GP-EI maximizes $\mathbb{E}[(I_{\mathrm{on}} - B)_+\,\mathbb{1}(\cdot)]$ with $B$ the best feasible on-current, both over 1,024 scrambled Sobol candidates drawn in a trust region (initial side 0.1 of the unit box) and over the whole box at every fifth evaluation; exact duplicates are excluded. GaAs deposition fits separate BoTorch GPs with Matérn-5/2 ARD kernels to the objective $(-\mathrm{NU}/100)$ and to the constraint $(1 - \bar{g}/3$, with failures labeled as violations). GP-UCB maximizes $\mu + 2\sigma$ subject to a feasibility probability of at least 0.5, and GP-EI maximizes $\log \mathrm{EI} + \log P(\text{feasible})$; before the first feasible observation both maximize the feasibility probability. Acquisitions are optimized from 1,024 Sobol starting points with continuous refinement (eight restarts). Neither task uses random or space-filling initial evaluations: the GPs start from the free initial observation. These are task-specific baselines rather than a shared implementation across both simulators. In the pool protocols, the GP acquisition policies are deterministic given the history, so their Selection repeats on a pool coincide or differ only through tie handling.

**Trajectory pools and aggregation.** Each of Pools A–E contains 50 evaluated designs and excludes the free initial observation. The MOSFET sources are the human expert path, Luna repeat 1 and Sonnet repeat 2 with the $300\,\mu$A prompt target, GP-UCB repeat 2, and GP-EI repeat 5. The GaAs deposition sources are the human expert path, Luna repeat 4, Sonnet repeat 1, GP-UCB repeat 5, and GP-EI repeat 3. Pool runs stop upon finding the exact feasible pool optimum or exhausting the pool; the stopping target need not equal the prompt target. Each pool is a fixed case study; the five pools are not independent draws of a pool-construction process.

For $\mathrm{AUC}_{50}$, the discovered pool optimum is carried forward after early stopping and the incumbents at evaluations 1–50 are averaged. Generation uses the same horizon and includes its free initial

observation in the incumbent. Failed and measured infeasible observations leave the incumbent unchanged.

## B SIMULATION IMPLEMENTATION

The MOSFET campaign records Sentaurus version 2024.03 (Synopsys, Inc., 2024). The process and device decks, bias points, current extraction, and nine-parameter schema are fixed across all optimizer comparisons. Currents are extracted at $V_d = 1$ V with $V_g = 1$ V ($I_{\mathrm{on}}$) or $V_g = 0$ V ($I_{\mathrm{off}}$) in the two-dimensional $1\,\mu$m-width convention. The nine inputs are `NWell`, `GOxTime`, `GOxTemp`, `LDD_Dose`, `LDD_Energy`, `SD_Dose`, `SD_AnnealTime`, `SD_AnnealTemp`, and `SD_Energy`. The two temperatures use linear protocol coordinates; the other seven use $\log_{10}$ coordinates. Feasibility is checked against the unrounded current. The declared bounds are $10^{10}$–$10^{20}\,\mathrm{cm}^{-3}$ for well concentration, $10^{10}$–$10^{20}\,\mathrm{cm}^{-2}$ for each implantation dose, 0.1–1000 keV for each implantation energy, $1.6667 \times 10^{-5}$–1666.6667 min ($0.001$–$10^5$ s) for oxidation time, $0.001$–$10^5$ s for anneal time, and 0–2000 °C for both temperatures. These broad bounds do not imply that every recipe has a valid process solution, which is why numerical and process failures are part of the benchmark definition.

The GaAs deposition configuration records COMSOL Multiphysics 6.2 (COMSOL AB, 2023). The reduced reactor model uses a $0.20$ m channel and a $0.10$ m wafer interval centered at the chosen wafer position. Inputs span susceptor temperature 773–1173 K, pressure 20–760 torr, inlet velocity $0.05$–$0.5$ m/s, tilt 0–5 degrees, inlet height $0.02$–$0.03$ m, nitrogen fraction 0–1, and wafer center $0.06$–$0.14$ m. The outlet height follows the imposed tilt and channel length. The nonuniformity is recorded as $100\%$ when any wafer location has nonpositive growth. Transport uses a reduced precursor/intermediate model with gas-phase and surface reactions and a gallium desorption contribution; the oracle is a specified process model rather than a claim to simulate all GaAs chemistry. The numerical validity policy rejects outputs whose absolute mass-balance residual is at least $0.1\%$. Pressure alone is logarithmic in the protocol coordinates.

## C MATERIALS EXPERIMENTS

**Datasets, initial designs, and targets.** BN is a 55-row table from a laboratory campaign of hexagonal boron nitride growth by MOCVD; its 0–5 score is used as supplied. Membrane, P3HT/CNT, and Perovskite are the public datasets distributed with Wang et al. (2026); P3HT/CNT and Perovskite also appear in Liang et al. (2021) and originate from Bash et al. (2021) and Sun et al. (2021). Perovskite instability is scaled by $10^{-3}$, and Boolean Membrane inputs are encoded as 0/1. The nine BN initial designs are the nine rows with score 0; the final target is a score of 5.0 (two rows), and the target region is a score of at least 4.5 (nine rows, including ties at the top-10% boundary). Each of the other three pools uses ten initial rows chosen by seeded random selection, and its final target is the best row of the dataset: elastic modulus $296.35$ for Membrane, $1243.67$ S/cm for P3HT/CNT, and instability $23.707$ for Perovskite, with target regions of at least $233.72$, at least $660$ S/cm, and at most $88.158$.

**Prompts.** The BN system message is, verbatim: "You are selecting the next BN MOCVD experiment from a finite candidate pool. Use only the data in this prompt. Do not inspect files, call tools, infer scores from candidate identifiers, or invent candidates. Candidate identifiers are freshly reassigned every cycle and carry no information. Choose exactly one available candidate expected to reach score 5.0 as quickly as possible. Return only JSON." The user message contains the task (target score and the instruction to select one candidate identifier), the observed recipes with their 13 inputs and scores, the candidate pool with inputs only, and the required output. The `far` and `local` arms add a `search_guidance` field; the control arm has none. The `far` instruction is "Search distance: look farther. Prefer an available candidate farther from the current best observed recipe, while still aiming to reach score 5.0 as quickly as possible." and the `local` instruction is "Local refinement: look nearby. Prefer an available candidate making a small local change around the current best observed recipe, while still aiming to reach score 5.0 as quickly as possible."

**Step size and aggregation.** The BN leaderboard includes 12 LLMs with 45 trajectories each (nine shared initial designs $\times$ five repeats) and a GP-UCB reference. Panel (a) of Figure 2 uses the final

benchmark target; panels (b,d,e) use first entry into the target region, and the entering decision is counted. For a decision with available set $\mathcal{P}_t$, incumbent $b_t$, and selected $x_t$, the step-size percentile is

$$s_t = \frac{\#\{x \in \mathcal{P}_t : d(x, b_t) < d(x_t, b_t)\} + \frac{1}{2}\#\{x \in \mathcal{P}_t : d(x, b_t) = d(x_t, b_t)\}}{|\mathcal{P}_t|}. \tag{5}$$

Continuous-feature distances use pool-range normalization; the mixed-feature Membrane analysis uses Gower distance (Gower, 1971). Step size is averaged within the relevant phase of each run, then within each initial design, and then equally across initial designs. This statistic must be distinguished from the target-distance percentile of Figure 5, which is a property of the pool rather than of a trajectory.

Panel (c) averages step size over the early phase, before the incumbent reaches the top-30% region, whereas panels (b,e) average over the whole phase before entry into the target region. Its 48 contrasts come from six models and the eight retained conditions of Appendix C.1. English-literature and optimization-paper retrieval and their combinations are excluded, as are DeepSeek context runs whose experimental conditions were incorrect; DeepSeek's leaderboard and step-size runs remain included. The pooled fit is recomputed after exclusions. The three additional pools use ten initial designs and five repeats for Sol and Luna. Model-mean intervals, the fitted band, and cross-pool intervals use shared-initial bootstrap resampling with 10,000 draws. Random-selection expectations for reaching the target region are $8.11$, $9.32$, and $9.27$ evaluations for Membrane, P3HT/CNT, and Perovskite. For Luna, the mean number of evaluations to the final target is $7.49$ under the control prompt, $6.47$ under `far`, and $11.09$ under `local`; for Sol the corresponding values are $6.44$, $9.42$, and $13.31$. Differences in the mean number of evaluations to the target region between each prompt arm and the control prompt are tested by permuting arm labels within each initial design (200,000 permutations, two-sided), with Holm correction over the four contrasts. The `far` instruction lowers the mean by $0.60$ evaluations for Sol (initial-block bootstrap 95% interval $[-0.87, -0.31]$; adjusted $p = 0.0013$) and by $2.69$ for Luna ($[-3.49, -1.82]$; $p < 10^{-4}$), and the `local` instruction raises it by $8.56$ and $4.42$ ($[7.51, 9.64]$ and $[3.02, 5.82]$; both $p < 10^{-4}$). An exact sign-flip test on the nine initial-design mean differences gives $p = 0.016$ for Sol under `far` and $p = 0.0039$, its smallest attainable value, for the other three contrasts. Because the Sol control mean is already $1.89$ and one evaluation is the minimum, the Sol `far` effect is bounded by a floor.

### C.1 Prior information and retrieval conditions

The context conditions of Figure 2(c) add text to the names-revealed BN prompt; the system message, observation history, and candidate pool are unchanged.

**Prior information (PI).** PI is placed in a `process_information` field. *Brief PI* contains six one-line descriptions of the input groups (annealing stage, growth conditions, cycle exposure, purge, precursor and carrier flows, and susceptor rotation) and states no direction of effect. *Detailed PI* is a 15,620-character document describing the reactor configuration, precursors, and operating ranges, the growth objective and how growth outcomes are characterized, and empirical knowledge from earlier process development, including qualitative effects of temperature, interruption time, and precursor ratio. Detailed PI therefore supplies domain knowledge beyond variable definitions, and its independence from the historical experiments behind the benchmark is not established.

**Retrieval (RAG).** Retrieval adds four passages in an `external_scientific_context` field, with the instruction "Use these retrieved passages only as external scientific background. Do not treat statements in them as observed experiment scores, and do not infer candidate identity from source metadata." Passages come from one of two corpora: *domain papers*, 38 articles on hBN growth by MOCVD, MOVPE, or CVD published between 1986 and 2026 (3,558 chunks), and *textbooks*, seven books on chemical vapor deposition, crystal growth, thin-film deposition, and two-dimensional materials (6,018 chunks). Text is extracted page by page, split into 512-character chunks without overlap, embedded with `all-MiniLM-L6-v2` (Reimers & Gurevych, 2019), and ranked by cosine similarity to one fixed query on the effects of growth temperature, pressure, precursor ratio, flows, pulse and purge times, carrier gas, annealing, and growth cycles on hBN crystalline quality. The query is not shown to the model, so the four passages are fixed for each corpus and reused at every decision. The top-4 similarities are $0.71$–$0.79$ for domain papers and $0.55$–$0.61$ for textbooks, none of whose retrieved passages concerns hBN growth.

**Retained conditions.** The eight conditions of panel (c) are brief PI, detailed PI, domain-paper retrieval, textbook retrieval, and each PI level combined with each corpus, each compared with the names-revealed baseline of the same model. Two further corpora, 21 optimization papers and five public-domain novels used as a random-text control, were run but are not included in panel (c).

## D COORDINATES, GEOMETRY DEFINITIONS, AND REFERENCE POPULATIONS

**Coordinate scaling by use.** Table 1 lists the input coordinates used in each computation. Within each task, the Mapping distance, core-set construction, and GP models share one transform and one set of declared bounds; the geometry diagnostics of Figure 5 use the same coordinates, and the LLM prompts use unnormalized physical values.

Table 1: **Input coordinates by use.** "Log" denotes $\log_{10}$; "declared" denotes min–max scaling to $[0, 1]$ using the fixed bounds in Appendix B, rather than pool-specific extrema. Figure 5 uses the corresponding GP/Mapping input coordinates for all MOSFET and GaAs pools, irrespective of the optimizer that generated each pool. LLM prompts use unnormalized physical values.

| Use | MOSFET | GaAs deposition |
|---|---|---|
| Figure 5 geometry | Seven inputs log, two temperatures linear; declared; Euclidean | Pressure log, six inputs linear; declared; Euclidean |
| Mapping distance (Eq. 1) | Seven inputs log, two temperatures linear; declared; root-mean-square Euclidean | Pressure log, six inputs linear; declared; root-mean-square Euclidean |
| Core-set construction (Eq. 4) | Same input coordinates as Mapping; Euclidean | No core-sets constructed |
| GP models | Same input coordinates as Mapping | Same input coordinates as Mapping |
| LLM prompts | Raw physical values with declared bounds; both times in seconds | Raw physical values with declared bounds |

The root-mean-square and Euclidean distances differ by a constant factor and select the same nearest neighbor or farthest point. Materials pools use observed linear min–max coordinates because no declared bounds exist for them.

**Definitions.** Let $D_i = z_i - z_0$, $L = \|z_\star - z_0\|_2$, and $v = (z_\star - z_0)/L$. For each initial–target pair,

$$A = \underset{i \neq \star,\, \|D_i\| > 0}{\text{median}} \frac{D_i^\top v}{\|D_i\|_2}, \qquad \text{TDP} = \frac{100}{K}\left[\#\{i : \|D_i\|_2 < L\} + \tfrac{1}{2}\#\{i : \|D_i\|_2 = L\}\right], \quad (6)$$

where TDP is the target-distance percentile and the comparison population is all $K$ candidate rows, including the target, failed and measured infeasible rows, and any initial row already in the pool. Rows are not deduplicated. Ties use relative tolerance $10^{-12}$ and absolute tolerance $10^{-14}$; an isolated nearest target has percentile $50/K$. With multiple equally best targets, statistics are averaged within an initialization and then equally over initializations. Each trajectory pool uses its own feasible optimum; the core-sets share the forced Pool A optimum. Materials values are medians over their nine or ten benchmark initializations, after averaging over equally best targets within each initialization.

**Simulator pool values.** Table 2 lists both diagnostics for every simulator pool in the declared-bound coordinates of Table 1. Every MOSFET trajectory pool has alignment of at least $0.78$ and target-distance percentile of at least $51$, whereas the core-sets have alignment of at most $0.33$ and percentile of at most $13$.

**Random reference contours.** The figure uses 1,000 random subsets of size 50 from each of two reference populations: the full 5,798-design reservoir and its feasible subset. Each subset includes the fixed Pool A optimum and 49 sampled designs, excludes the initial design from sampling, and is expressed, together with that initial design, in the same declared-bound coordinates. Unlike the core-sets, these populations permit designs better than the anchor, so the anchor is not necessarily the best member. A Scott-bandwidth kernel density estimate with reflection at percentile boundaries 0 and 100 supplies the 50% and 90% probability contours. No optimizer was evaluated on these random draws.

Table 2: **Pool geometry in Figure 5.** Alignment ($A$) and target-distance percentile (TDP) are computed using the fixed-bound log/linear input coordinates specified in Table 1. Each pool uses its own initialization and best feasible design as the reference target. TDP is reported as a percentage; values are rounded for display.

| Pool | Alignment $A$ | TDP (%) |
|---|---|---|
| MOSFET A (Human) | 0.97 | 59 |
| MOSFET B (Luna) | 1.00 | 77 |
| MOSFET C (Sonnet) | 1.00 | 97 |
| MOSFET D (GP-UCB) | 0.78 | 55 |
| MOSFET E (GP-EI) | 0.93 | 51 |
| Core-set Pure | 0.02 | 1 |
| Core-set Feasible-only | 0.20 | 3 |
| Core-set Stratified | 0.33 | 13 |
| GaAs A (Human) | 1.00 | 69 |
| GaAs B (Luna) | 1.00 | 61 |
| GaAs C (Sonnet) | 0.98 | 23 |
| GaAs D (GP-UCB) | 0.88 | 41 |
| GaAs E (GP-EI) | 0.98 | 25 |

## E CORE-SET EVALUATION DETAILS

**Pool construction.** All core-set pools are constructed from a reservoir of 5,798 distinct evaluated MOSFET inputs. To keep the attainable optimum matched to Pool A, we exclude the initial design and the 158 feasible designs exceeding the fixed Pool A optimum ($308.903\,\mu$A), leaving 5,639 eligible designs, of which 2,726 are feasible. Each pool contains 50 candidates including the forced optimum, and farthest-first selection uses the declared-bound logarithmic/linear coordinates defined in Appendix D.

**Pure** applies farthest-first selection over all eligible designs. The resulting pool contains 2 feasible, 10 measured-infeasible, and 38 failed evaluations. **Feasible-only** applies the same selection rule to the 2,726 feasible designs, so feasible-region discovery is removed from the Selection task. **Stratified** contains 30 feasible, 17 measured-infeasible, and 3 failed designs. Its feasible designs are allocated across on-current bins below 50, 50–100, 100–150, 150–200, 200–250, 250–300, and at least 300 $\mu$A with counts $5, 5, 5, 5, 5, 4, 1$, respectively; the final bin contains the forced optimum. Greedy selection fills each feasible bin and then the remaining outcome strata while computing distance against all centers already selected.

The optimum filter and stratification use recorded outcomes, so these constructions are not outcome-blind. Moreover, farthest-first selection can redistribute coverage only within the evaluated reservoir and cannot create support in regions that were never evaluated.

**Evaluation.** The three pools share the same initial observation and forced optimum. The two LLMs have five Selection repeats per pool. GP Selection on the core-sets is deterministic under the implemented candidate policy and is evaluated once per pool; its single run is not treated as five independent observations. All runs stop at the exact pool optimum or at pool exhaustion, and the discovered optimum is carried forward to the common 50-evaluation horizon.

Table 3 reports mean Selection $\mathrm{AUC}_{50}$. The Generation row reports mean $\mathrm{AUC}_{50}$ over the five Generation runs of each optimizer. Its observed optimizer ordering is Sonnet, GP-EI, GP-UCB, Luna, and $\rho$ compares each Selection ordering with this Generation ordering.

**Optimum discovery.** In Pure, where only two of the 50 candidates are feasible, Luna and Sonnet select the fixed optimum at the first or second evaluation in every repeat, GP-EI at the second, and GP-UCB at the thirteenth. In Feasible-only and Stratified, the LLMs require between 1 and 18 evaluations. Early optimum discovery can compress $\mathrm{AUC}_{50}$ differences because the discovered optimum is carried forward to evaluation 50.

Pure and Feasible-only therefore answer different questions. Pure retains failures among the candidates while controlling geometric coverage, whereas Feasible-only removes feasible-region discovery from the candidate-selection task. Stratified prescribes a mixture of feasible quality lev-

Table 3: **Selection on core-set pools under $\mathrm{AUC}_{50}$ (MOSFET, $\mu$A).** Means over five repeats for the LLMs and one deterministic run for each GP. The Generation row gives means over five Generation runs. Brackets rank the observed optimizer means; $\rho$ is Spearman correlation with the observed Generation ordering.

| Pool | Luna | Sonnet | GP-UCB | GP-EI | $\rho$ |
|---|---|---|---|---|---|
| Generation | 132.3 [4] | 215.0 [1] | 187.8 [3] | 201.9 [2] | – |
| Pure | 306.4 [2] | 307.7 [1] | 234.8 [4] | 302.7 [3] | +0.40 |
| Feasible-only | 308.7 [1] | 308.4 [2] | 284.3 [4] | 296.5 [3] | −0.20 |
| Stratified | 304.5 [1] | 298.5 [2] | 297.1 [3] | 263.2 [4] | −0.40 |

els, measured-infeasible designs, and failures so that comparisons are dominated neither by an all-feasible pool nor by the concentration of failures selected by Pure. None of these pools is a uniform sample of the continuous design space, and these comparisons therefore do not isolate geometry as the sole cause of disagreement between Selection and Generation.

## F Protocol comparisons on all pools

Figures 6 and 7 extend Figure 3 to all five trajectory pools with the same plotting code, data conventions, and axes. In every pool, the Selection and Mapping medians converge to that pool's optimum, whereas the order in which the optimizers reach it varies across pools and between the two protocols.

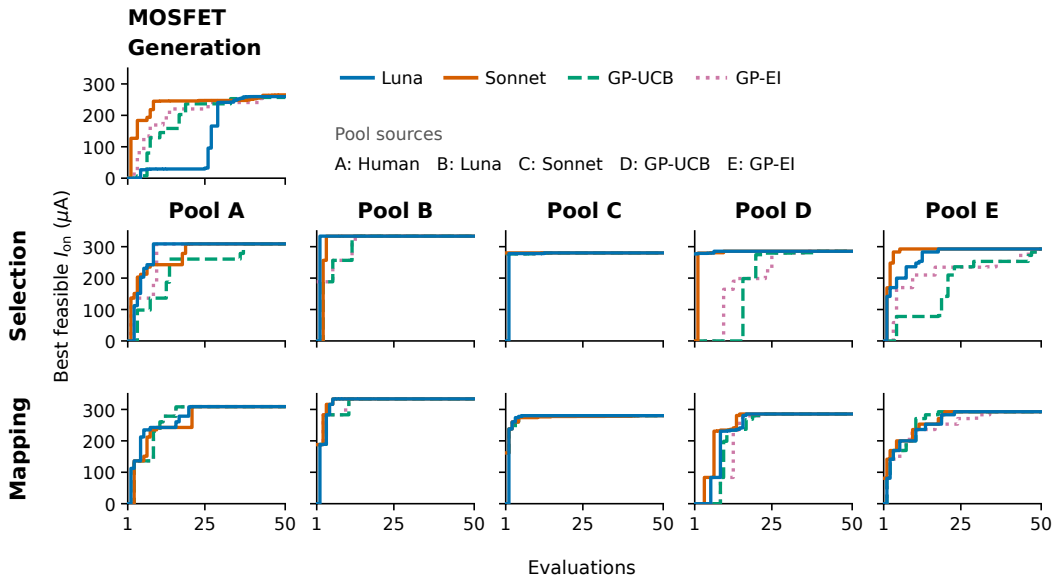


Figure 6: **MOSFET optimization progress under Generation and all five candidate pools.** Generation is shown at the top, Selection in the middle row, and Mapping in the bottom row. Pools A–E are derived from the human expert, Luna, Sonnet, GP-UCB, and GP-EI, respectively; line colors and styles identify the optimizer evaluated. Curves show the median best-feasible incumbent across five runs; on-current is shown in $\mu$A on a linear axis (higher is better). A median is shown only when all five runs have a feasible incumbent, and early-stopped runs carry forward their last incumbent through evaluation 50.

**Conventional summaries.** Table 4 reports the summaries that benchmarks typically use: the best feasible objective after 50 Generation evaluations, and the mean number of evaluations to the exact feasible pool optimum under Selection or Mapping, averaged over Pools A–E and five repeats. These summaries have different endpoints and are not commensurable: the Generation entry measures the attained objective, whereas the pool entries measure how quickly an optimum fixed by the pool is found.

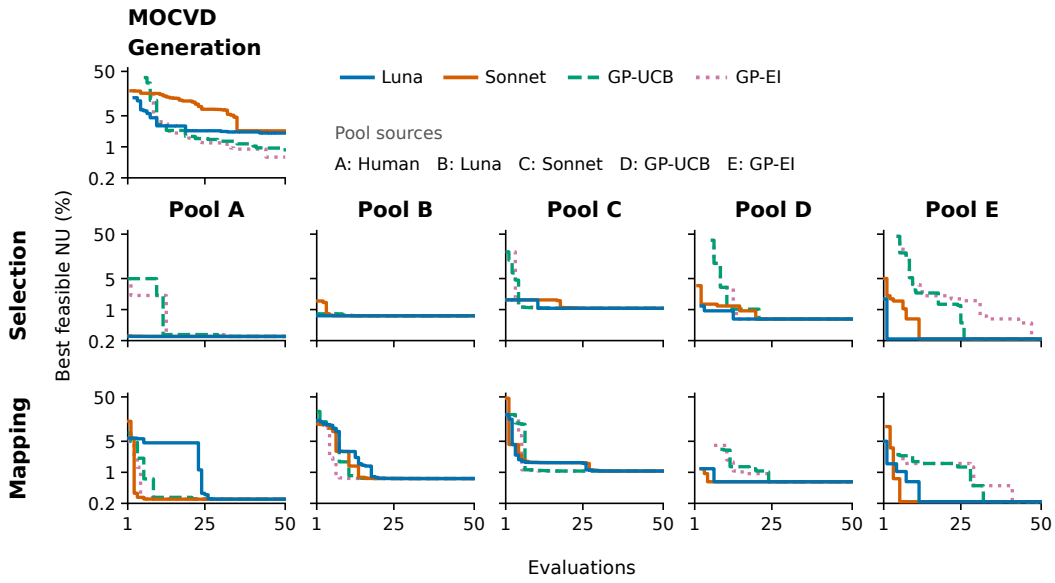


Figure 7: **GaAs deposition optimization progress under Generation and all five candidate pools.** Layout and conventions as in Figure 6; nonuniformity is shown in percent on a logarithmic axis (lower is better).

Table 4: **Conventional summaries with unmatched endpoints.** Generation: best feasible objective after 50 evaluations (mean ± sample SD, five runs). Selection and Mapping: mean evaluations to the exact feasible pool optimum over Pools A–E × five repeats (lower is better).

| Task | Optimizer | Generation | Selection | Mapping |
|---|---|---|---|---|
| MOSFET ($\mu$A) | Luna | $182.9 \pm 143.8$ | 10.44 | 19.60 |
| | Sonnet | $255.1 \pm 57.2$ | 11.96 | 26.92 |
| | GP-UCB | $251.6 \pm 38.8$ | 41.04 | 17.48 |
| | GP-EI | $259.9 \pm 21.5$ | 33.08 | 18.52 |
| GaAs deposition (NU, %) | Luna | $1.92 \pm 0.85$ | 7.68 | 23.68 |
| | Sonnet | $3.76 \pm 3.74$ | 15.00 | 18.92 |
| | GP-UCB | $0.96 \pm 0.39$ | 24.20 | 26.40 |
| | GP-EI | $0.60 \pm 0.31$ | 26.80 | 24.76 |

## G MAPPING AUDIT TRANSITIONS BY OPTIMIZER

Table 5 separates the status transitions of Section 5.2 into feasible (F), measured infeasible (I), and failed (X) outcomes, from proposal to credited projection. In MOSFET, 96 proposals and 37 projections are failed evaluations; the 214 losses of feasible status comprise 202 transitions to measured infeasible and 12 to failed, and the 579 gains comprise 530 from measured infeasible and 49 from failed. The GaAs deposition audit contains no failed evaluations, so every status change there is a measured constraint crossing.

Table 5: **Proposal → projection status transitions by optimizer.** F: feasible; I: measured infeasible; X: failed. Pairs pooled over Pools A–E and five repeats.

| Task | Optimizer | F→F | F→I | F→X | I→F | I→I | I→X | X→F | X→I | X→X |
|---|---|---|---|---|---|---|---|---|---|---|
| MOSFET | Luna | 246 | 81 | 2 | 80 | 70 | 2 | 2 | 2 | 5 |
| | Sonnet | 223 | 61 | 0 | 227 | 161 | 0 | 0 | 1 | 0 |
| | GP-UCB | 164 | 30 | 4 | 93 | 94 | 3 | 27 | 12 | 10 |
| | GP-EI | 152 | 30 | 6 | 130 | 108 | 0 | 20 | 12 | 5 |
| | All | 785 | 202 | 12 | 530 | 433 | 5 | 49 | 27 | 20 |
| GaAs deposition | Luna | 505 | 47 | – | 24 | 16 | – | – | – | – |
| | Sonnet | 439 | 20 | – | 12 | 2 | – | – | – | – |
| | GP-UCB | 362 | 33 | – | 211 | 54 | – | – | – | – |
| | GP-EI | 471 | 34 | – | 64 | 50 | – | – | – | – |
| | All | 1777 | 134 | – | 311 | 122 | – | – | – | – |

## H DEFINITION OF TI

**Setting.** Let $\mathcal{P} = (z_i)_{i=1}^K$ be an indexed pool with transformed candidates $z_i = z(x_i) \in \mathcal{Z} \subseteq \mathbb{R}^m$. We retain repeated rows and fix an initial design $z_0$ and a designated target row $z_\star$, with

$$L := \|z_\star - z_0\|_2 > 0. \tag{7}$$

All distances below use the Euclidean norm in these coordinates. The target is a best observed design in the pool and need not be a global optimum over the continuous design space.

Define the empirical measure on the candidate pool by

$$\mu_{\mathcal{P}} := \frac{1}{|\mathcal{I}|} \sum_{i \in \mathcal{I}} \delta_{z_i}, \qquad \mathcal{I} := \big\{ i \in \{1, \dots, K\} : z_i \notin \{z_0, z_\star\} \big\} \tag{8}$$

where $\delta_z$ denotes the Dirac probability measure at $z$, and we assume $|\mathcal{I}| > 0$. Rows at either fixed endpoint are excluded because these locations lie on every admissible reference path by construction. All other repeated rows retain their multiplicity.

**Definition H.1** (Trajectory imprinting)**.** Let $\gamma : [0,1] \to \mathcal{Z}$ be a prescribed continuous reference path satisfying $\gamma(0) = z_0$ and $\gamma(1) = z_\star$. Define the distance from $z \in \mathcal{Z}$ to the reference path by

$$d_\gamma(z) := \min_{t \in [0,1]} \|z - \gamma(t)\|_2, \tag{9}$$

and the normalized $\varepsilon$-tube around $\gamma$ by

$$\mathcal{T}_\varepsilon(\gamma) := \{ z \in \mathcal{Z} : d_\gamma(z) \le \varepsilon L \}, \qquad \varepsilon \ge 0. \tag{10}$$

Then, the *$\varepsilon$-trajectory-imprinting (TI) score* of the candidate pool $\mathcal{P}$ relative to $\gamma$ is

$$\mathrm{TI}_\gamma(\mathcal{P}; \varepsilon) := \mu_{\mathcal{P}}\big(\mathcal{T}_\varepsilon(\gamma)\big) = \frac{1}{|\mathcal{I}|} \sum_{i \in \mathcal{I}} \mathbf{1}\left\{ d_\gamma(z_i) \le \varepsilon L \right\}. \tag{11}$$

At a fixed radius $\varepsilon$, a larger $\mathrm{TI}_\gamma(\mathcal{P}; \varepsilon)$ indicates stronger TI relative to $\gamma$.

*Remark* H.2 (Reference dependence). Given the endpoints and coordinate system, the reference path must be specified without fitting the remaining candidate locations. If $\gamma$ is obtained by interpolating the same optimization trajectory from which $\mathcal{P}$ is constructed, then $d_\gamma(z_i) = 0$ for every trajectory point $z_i$, and hence $\mathrm{TI}_\gamma(\mathcal{P}; 0) = 1$ trivially.

**Straight reference path.** To relate Definition H.1 to the used diagnostics, choose

$$\gamma_{\mathrm{lin}}(t) := (1-t)z_0 + tz_\star, \qquad t \in [0,1]. \tag{12}$$

We call $\mathrm{TI}_{\gamma_{\mathrm{lin}}}$ the *chordal TI score*. This is a special case of Definition H.1, rather than an assumption that actual optimization trajectories are straight. Alignment and target-distance percentile are endpoint-based geometric summaries, not evaluations of the tube-based TI score. Our empirical analysis uses these diagnostics and directly tests agreement with Generation.