SMS-SIZING: AN AGENTIC SEARCH FRAMEWORK FOR FEASIBLE ANALOG CIRCUIT SIZING AND ARCHIVE EXPANSION

Anonymous authors
Paper under double-blind review

ABSTRACT

Analog circuit sizing selects transistor dimensions, bias currents, capacitor values, and other device parameters for a fixed topology so that simulated performance satisfies all specifications simultaneously. Each candidate design requires a circuit simulation to determine whether it meets all specifications. Our goal is to find multiple feasible designs with distinct parameter configurations within a limited simulation budget. Some existing sizing methods simplify the search by restricting parameter ranges or discretizing design variables. These restrictions can exclude feasible designs and limit the diversity of the solutions found. We propose Sensitivity-Manifold Search (SMS-SIZING). Compared with prior work, our study makes three contributions. First, we formulate the task as expanding one verified feasible design into a set of distinct alternatives within the original parameter space, and define metrics for feasible yield and design diversity. Second, we develop a search framework that generates candidates from verified designs, uses surrogate predictions to guide candidate selection and simulation budget allocation, and checks region updates against simulation evidence. Third, we use baseline comparisons and ablation studies to assess performance and examine how different components affect feasible yield and design diversity. All designs added to the archive are verified by circuit simulation. Across six amplifier topologies under eight simultaneous constraints, starting from a single feasible seed and using a budget of 100 simulations, SMS-SIZING reliably expands the design archive. Anchored candidates achieve a 59.4% conversion rate, while uniform random search finds no feasible designs in our experiments. In head-to-head comparisons, SMS-SIZING maintains up to $2.8\times$ greater spatial reach and $2.1\times$ larger worst-case constraint margins, preserving candidate diversity across spatial separation resolutions.

1 INTRODUCTION

Analog circuit sizing chooses device parameters so that a fixed topology simultaneously meets multiple circuit specifications. Each candidate requires one circuit simulation, which makes the problem a constrained black-box search with an expensive evaluation cost. Existing sizing methods have largely targeted a single high-quality design. In practice, however, a single feasible point is not always the desired endpoint: downstream corner, layout, reliability and implementation stages may need several acceptable parameter vectors with different device allocations and constraint margins. A useful search output can therefore be a set of distinct feasible designs rather than one optimum.

This paper studies the following more specific question. Given a simulator-verified feasible design x_0 and a total budget of N simulation calls, including initial verification, how effectively can search expand the feasible archive around and beyond the neighborhood of x_0? We call this feasible-seed-conditioned archive expansion. Every archive member must satisfy all circuit specifications under simulation, and additional designs should be sufficiently separated from existing members to represent distinct alternatives.

This task is challenging because feasibility is sparse in the original design domain. We keep the legal continuous/integer parameter ranges specified for the circuit rather than first reducing or discretizing

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.

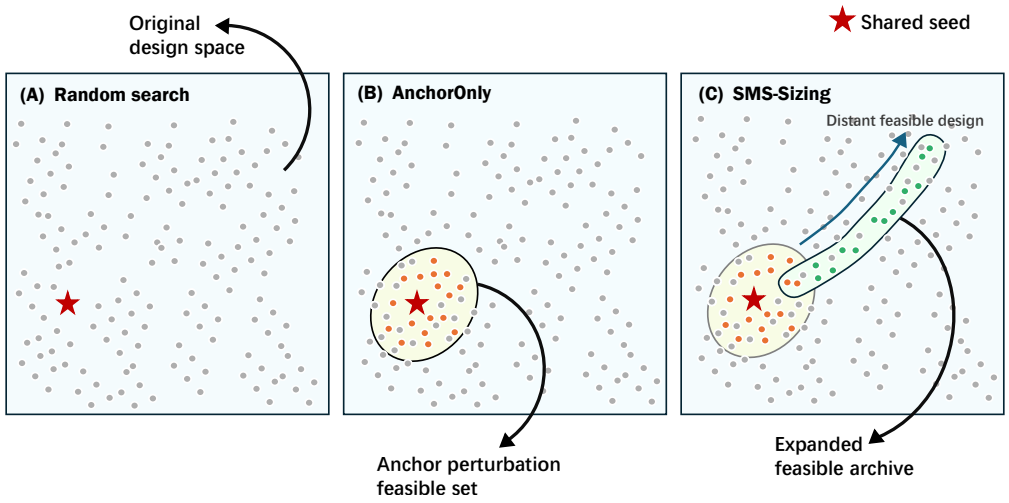


Figure 1: **Conceptual comparison of feasible-archive expansion from a shared verified seed.** **(A)** Random search samples the original design space but rarely finds additional feasible designs. **(B)** AnchorOnly perturbs measured anchors to produce a local cluster of feasible designs. **(C)** SMS-SIZING combines anchor perturbation with surrogate-guided selection to extend the feasible archive toward distant designs.

the space (Section 2). An independent audit of 3,000 uniform draws from these original ranges, across the six amplifier topologies studied here, finds no design that satisfies all eight specifications simultaneously, although individual specifications often pass at a much higher rate (Appendix A). In this setting a simulator-verified feasible point carries information that uninformed global sampling lacks: it locates a position in the feasible set and supplies a directly measured starting point for further exploration.

Our central observation is that a verified feasible point should be treated as a productive search object, establishing a strong operational foothold in an otherwise sparse design space. As illustrated in Figure 1, repeatedly perturbing a single seed confines discovery to an immediate local cluster, whereas an effective search must systematically adapt the search space to navigate across the feasible manifold. To expand beyond the starting neighborhood, we decouple the search into two complementary functions:

1. **Feasible candidate generation** efficiently produces valid designs by sparsely perturbing verified anchors on the feasible manifold.
2. **Surrogate-guided frontier navigation** steers candidate selection and evaluation quotas toward structurally distant and high-margin regions.

SMS-SIZING realizes this decoupled architecture through sparse anchor perturbation, coordinate-influence screening, and an evidence-gated quota controller. A pretrained surrogate ensemble provides fast coordinate ranking and candidate margin scores, which the policy controller translates into targeted evaluation quotas. An empirical evidence gate enforces the retention of observed representatives, while the circuit simulator remains the sole authority on feasibility. This design ensures that high-yield anchored proposals are systematically steered outward by surrogate-scored quotas, transforming local feasible stability into diverse, wide-reach manifold expansion (Section 4).

Compared with prior work, our study makes three contributions:

- **A formulation of archive expansion for sparse feasibility.** We formulate budget-constrained analog sizing as expanding a verified feasible archive in the original unreduced design space, introducing rigorous endpoints (H_N^+, U_N^{new}, and resolution-filtered U_N^τ) that separate genuine manifold coverage from redundant local sampling (Section 3.1).
- **A decoupled manifold-expansion architecture.** We introduce SMS-SIZING, which pairs high-yield sparse anchor perturbation with surrogate-scored quota allocation and evidence

gating, actively guiding search toward wide spatial reach and robust constraint margins (Section 3.2).

- **Mechanistic validation of constrained search.** Through candidate-stream accounting, surrogate-free controls, and baseline re-executions across six amplifier topologies, we demonstrate that anchor conditioning sustains feasibility (59.4% yield vs. 0% uniform) while surrogate ranking drives placement, providing transferable design principles for expensive constrained optimization (Section 4).

2 Related Work

Expensive constrained search. Constrained Bayesian optimization allocates evaluations using objective and constraint models (Gardner et al., 2014; Gelbart et al., 2014). TuRBO and SCBO adapt local trust regions and use partial-coordinate perturbations (Eriksson et al., 2019; Eriksson & Poloczek, 2021); random embeddings and sparse priors address high dimensionality (Wang et al., 2013; Eriksson & Jankowiak, 2021). SMS-SIZING shares their use of local search evidence and targets expansion of a feasible archive. Our internal selector controls isolate batch ordering; LEDRO (Kochar et al., 2025) supplies a released trust-region comparison within the common circuit interface.

Neural surrogates and LLM-assisted sizing. Learning-based sizing uses reinforcement learning, neural surrogate models, and language-model agents to organize costly simulation (Settaluri et al., 2020; Wang et al., 2020; Budak et al., 2021; Liu & Chitnis, 2026; Ahmadzadeh et al., 2025; Yu et al., 2026; Zhang et al., 2026). Recent advances in neural circuit modeling, such as autoregressive neural simulators (Poddar et al., 2025), unified transformer embeddings for zero-shot circuit evaluation (Yang et al., 2025), and layout-constrained machine learning frameworks (Mehradfar et al., 2025), demonstrate the growing capability of deep surrogates in capturing complex circuit dynamics. In optimization workflows, methods differ in how they manage search domains: AutoSizer groups transistors and reduces the space before optimizing within it (Yu et al., 2026); LEDRO reduces parameter ranges via language models and applies trust-region search (Kochar et al., 2025); and TopoSizing updates regions from search feedback (Wei et al., 2026). In contrast, SMS-SIZING operates directly in the original mixed continuous/integer box throughout, employing pretrained surrogates not as black-box arbiters but as efficient ordinal ranking and quota filters within an evidence-gated anchor loop. AnalogGym supplies the circuit benchmark (Li et al., 2024); Appendix D.16 details adaptations needed to execute released baselines in that environment.

Quality-diversity and multi-solution search. Novelty search and quality-diversity methods seek collections of solutions (Lehman & Stanley, 2011; Cully et al., 2015), and MAP-Elites indexes archives by behavioural descriptors (Mouret & Clune, 2015). Our archive uses joint simulator feasibility for membership and parameter separation and constraint margin for evaluation. Appendix E compares these task definitions; quality-diversity methods provide conceptual context for the multi-design endpoint.

3 Method

3.1 Problem formulation

Given one simulator-verified feasible seed x_0, SMS-SIZING seeks additional feasible designs within N simulator calls. A topology has d sizing variables:

$$\mathcal{X} = \prod_{i=1}^{d} \mathcal{X}_i, \qquad \mathcal{X}_i = \begin{cases} [l_i, u_i], & i \in \mathcal{I}_{\text{cont}}, \\ \mathcal{Z}_i, & i \in \mathcal{I}_{\text{int}}, \end{cases} \tag{1}$$

where $\mathcal{Z}_i$ contains the legal integer multiplicities. One AC/DC simulation measures power, DC gain, gain-bandwidth product (GBW), phase margin, common-mode rejection ratio (CMRR), positive and negative power-supply rejection ratio (PSRR), and input-referred offset. For metric g_k and threshold t_k, set $c_k(x) = s_k(g_k(x) - t_k)$, with $s_k = 1$ for upper bounds and $s_k = -1$ for lower bounds. Then

$$\mathcal{F} = \{x \in \mathcal{X} : c_k(x) \le 0,\ k = 1, \dots, 8\}, \qquad x_0 \in \mathcal{F}. \tag{2}$$

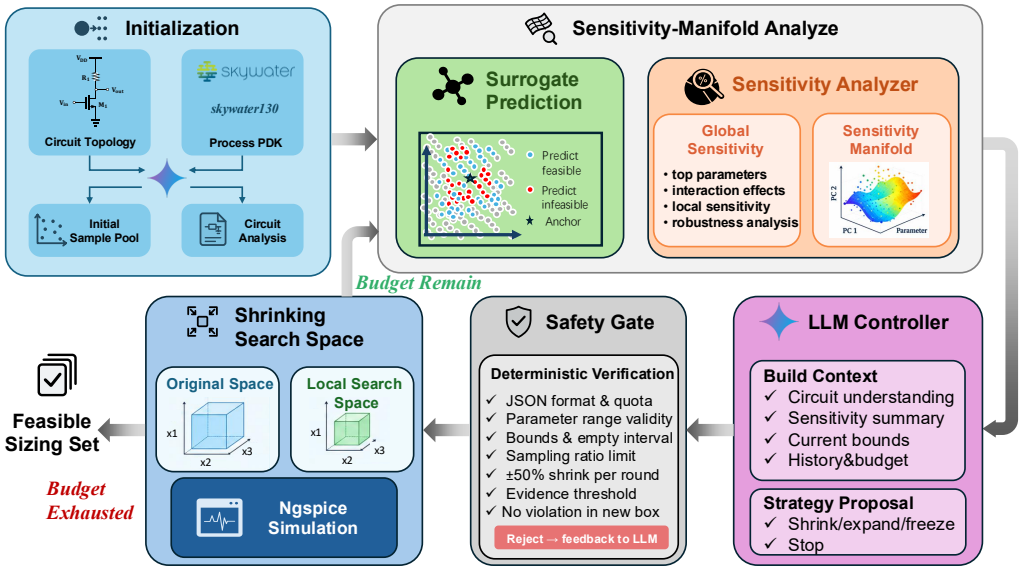


Figure 2: **The online-controller workflow of SMS.** Circuit topology and process information initialize the sample pool and circuit analysis. Surrogate predictions and sensitivity summaries inform the LLM controller. Proposed search-space updates pass the Safety Gate before ngspice evaluates candidates. The loop continues while budget remains and returns the feasible sizing set when the budget is exhausted.

Every evaluation, including seed verification, failure and timeout, consumes budget. With $A_N = \{x_n : x_n \in \mathcal{F},\ n \le N\}$, we measure

$$H_N = \frac{1}{N}\sum_{n=1}^{N} \mathbf{1}\{x_n \in \mathcal{F}\}, \quad U_N = |A_N|, \quad U_N^{\text{new}} = |A_N \setminus \{x_0\}|. \tag{3}$$

The beyond-initialization rate H_N^+ deletes every evaluation of x_0 from both numerator and denominator. Formally,

$$H_N^+ = \frac{\sum_{n=1}^{N} \mathbf{1}\{x_n \in \mathcal{F},\ x_n \neq x_0\}}{\sum_{n=1}^{N} \mathbf{1}\{x_n \neq x_0\}}, \qquad \sum_{n=1}^{N} \mathbf{1}\{x_n \neq x_0\} > 0. \tag{4}$$

This distinguishes new search yield from repeated seed measurements.

Archive utility also depends on separation and margin. Normalize coordinates by their original legal ranges, $z_i(x) = (x_i - l_i)/(u_i - l_i)$, and use $\Delta(x, x') = \|z(x) - z(x')\|_2/\sqrt{d}$. We report within-archive median pairwise distance, median distance to the seed, and worst-case margin $m(x) = \min_k[-c_k(x)/|t_k|]$. The resolution count U_N^τ greedily retains new feasible designs in evaluation order when they are farther than τ from all retained designs; $U_N^0 = U_N^{\text{new}}$. This fixed-order separation rule measures archive resolution without changing the search.

3.2 Overview and Initialization

The search maintains a current region $B_r \subseteq \mathcal{X}$, a record of simulator measurements, and a feasible archive A_r. The loop starts with the original legal box and validates the provided seed. The initial batch combines global samples with local candidates. Subsequent rounds reuse measured feasible designs as anchors, construct a candidate pool, obtain evaluation quotas and region proposals, verify constraints using an empirical evidence gate, and simulate the selected candidates (Figure 2). Appendix B.1 gives the complete execution order. The search couples two distinct operations: **Feasible anchor generation.** Simulator-verified feasible designs are sparsely perturbed to produce high-yield

candidate populations. **Surrogate-scored quota guidance.** The surrogate ensemble ranks coordinate influence and predicts candidate margins, enabling the controller to allocate simulation quotas toward high-margin boundaries and Pareto frontiers.

3.3 Coordinate influence and anchored candidate generation

A five-member surrogate ensemble predicts eleven metrics. Eight define feasibility; slew rate, temperature coefficient and settling time provide auxiliary circuit responses. For pool coordinates X and predictions Y, the analyzer computes the mean absolute Pearson correlation

$$a_i = \frac{1}{11} \sum_{j=1}^{11} |\mathrm{corr}(X_{\cdot i}, Y_{\cdot j})| \,. \tag{5}$$

Recalled simulator records enter this analysis with measured values and zero predictive spread. We employ a_i as an efficient online coordinate-influence screening score: it rapidly prioritizes coordinates with strong linear association across predicted metrics without requiring expensive global variance decompositions. Averaging all eleven responses provides a balanced ranking across circuit behaviours. The ranking is highly robust: recomputing it with the eight constraints yields a mean rank correlation of 0.969, and replacing Pearson with Spearman correlation yields 0.980 (Appendix D.10). Circuit simulation establishes authoritative feasibility throughout.

Each pool mixes original-box global sentinels (20%), sparse anchor moves (15% of the non-global remainder), and Latin-hypercube fill of B_r. Up to eight distinct measured anchors are chosen, prioritizing verified feasible designs. Near-feasible anchors must pass at least six constraints and have total deficit $D(x) = \sum_k \max\{c_k(x)/|t_k|, 0\} \leq 1$; lower deficit and larger $m(x)$ break ties. For an anchor projected into $B_r = \prod_i [L_i, U_i]$, denoted $\bar{x}$, we choose $q = \max(1, \mathrm{round}(0.05d))$ coordinates uniformly without replacement. Each active coordinate is sampled within

$$[\max\{L_i, \bar{x}_i - 0.10(U_i - L_i)\}, \min\{U_i, \bar{x}_i + 0.10(U_i - L_i)\}] \,. \tag{6}$$

Integer coordinates use legal integers inside this interval; inactive coordinates retain their projected values. Uniform coordinate selection across the active subset maintains broad exploration over valid local directions, while the joint simulator test evaluates the resulting non-linear parameter interactions. Appendix B.2 specifies clipping and eligibility details.

3.4 Evidence-gated policy updates

The policy controller reads sensitivity rankings, predicted margins, measured records, bounds and remaining budget. Its JSON response proposes bound adjustments and five evaluation quotas: global, boundary, Pareto, uncertainty and final. A served language model or an offline rule can implement this interface. Candidate generation and simulator admission are explicit procedures outside the controller.

For predicted mean $\hat{g}_k$ and spread σ_k, define

$$\hat{m}_k = -s_k(\hat{g}_k - t_k)/|t_k|, \qquad m_k^{\pm} = \hat{m}_k \pm w_k \sigma_k / |t_k|, \tag{7}$$

with $w_k = 1$ for GBW, phase margin and offset and 0.25 elsewhere. The tracker links candidates passing either margin test within a standardized distance of $0.35\sqrt{d}$ and retains a representative of each empirical component.

Both controller and automatic analyzer proposals pass the evidence gate. Checks cover schema validity, legal parameter names and bounds, evidence counts, contraction caps, and retention of tracked representatives. Four conservative records or two distinct measured feasible designs permit up to 50% width reduction per round. With less evidence, at most two coordinates may shrink by 25% each. Rejected policies receive diagnostic feedback for up to three attempts in total; without an accepted proposal the box stays fixed. The gate strictly enforces retention of observed feasible representatives. Crucially, our matched controls reveal that search stability and archive expansion are highly robust to bounding box volume: feasible yield remains steady between 19.70% and 19.97% across three orders of magnitude of final box volume (Appendix D.11), demonstrating that surrogate-scored quota selection, rather than rigid geometric confinement, reliably drives candidate placement (Section 4.3).

Table 1: **Design alternatives from adapted released implementations.** 18 runs per method and tier, $N = 100$. New: mean distinct feasible count beyond the seed. Reach: mean within-run median normalized distance to the seed. Margin: mean of each run's best worst-case normalized margin. Here, n.a. denotes an undefined metric (e.g., reach when no new feasible design is found).

	Simple			Medium			Hard		
Implementation	New	Reach	Margin	New	Reach	Margin	New	Reach	Margin
SMS-M	35.3	**.127**	**.060**	27.5	**.070**	**.025**	**10.2**	**.085**	**.041**
LEDRO	**56.1**	.038	.042	**28.4**	.025	.012	8.6	.040	.020
EEsizer	36.9	.042	.048	8.6	.032	.009	2.1	.035	.015
AutoSizer	0.0	n.a.	.037	0.0	n.a.	.003	0.0	n.a.	n.a.

3.5 Simulation and archive update

The quota selector reserves 30% of non-global slots for available anchored candidates and uses margin and diversity ordering to fill the batch. Every selected design is simulated, including candidates outside predicted-feasibility masks. New feasible measurements enter A_{r+1}; the last 100 distinct successful simulations are recalled for the next analysis. Accepted bounds carry into the next pool, coupling measured archive growth to future move radii. Global sentinels continue to cover the original legal box. Appendices B and C specify the quotas, recall rule, surrogate and conditional volume analysis.

4 Experiments

4.1 Experimental Setup

We evaluate archive expansion along three primary axes: the yield of newly discovered feasible designs from a single seed, their spatial reach and constraint margins, and the underlying mechanisms sustaining expansion. A fourth diagnostic tests controller substitution. All benchmarks use six operational amplifier topologies from AnalogGym under eight simultaneous constraints simulated in `ngspice` with the SkyWater sky130 PDK at nominal `tt`. Each run is allocated a budget of $N = 100$ simulator calls (timeouts and failures fully charged). A simulator-verified design serves as the initial seed across all difficulty tiers (Appendix D.14).

Evaluations are partitioned into two strictly separated cohorts: (1) **Published-agent comparison cohort (**18 **runs per method and tier):** Benchmarking against adapted implementations of LEDRO, EEsizer, and AutoSizer across three random seeds. (2) **Component ablation cohort (**30 **runs per arm):** Five seeds across six circuits, testing systematic ablations of coordinate direction, boundary exploration, memory recall, and selectors. SMS-M is the primary method; SMS-B provides the matched reference for component ablations. Inference treats topology as the unit, summarizing paired differences via cluster bootstrap (Appendix D.15). Pretrained across 16 topologies ($60{,}000$ evaluations each), surrogate pre-training costs are treated as an amortized asset. All reported feasible designs are confirmed by full ngspice simulation.

4.2 Baseline Comparisons

Expanding the feasible archive from a single verified seed. Analog circuit sizing typically contends with an acutely sparse feasible space in the original, unreduced parameter domain. Given only a single simulator-verified seed, an effective multi-point search must generate a substantial archive of distinct feasible alternatives without becoming trapped at the starting point. As shown in Table 1, SMS-M consistently expands the feasible archive across all difficulty tiers within the limited 100-evaluation budget. It discovers tens of distinct feasible designs beyond the seed on both the `simple` and `medium` tiers, and achieves the highest count of successful runs and feasible designs on the challenging `hard` tier. In contrast to single-point optimizers that stall once the seed is verified, SMS-M maintains active exploration, steadily converting simulation budget into a broader portfolio of valid circuit dimensions.

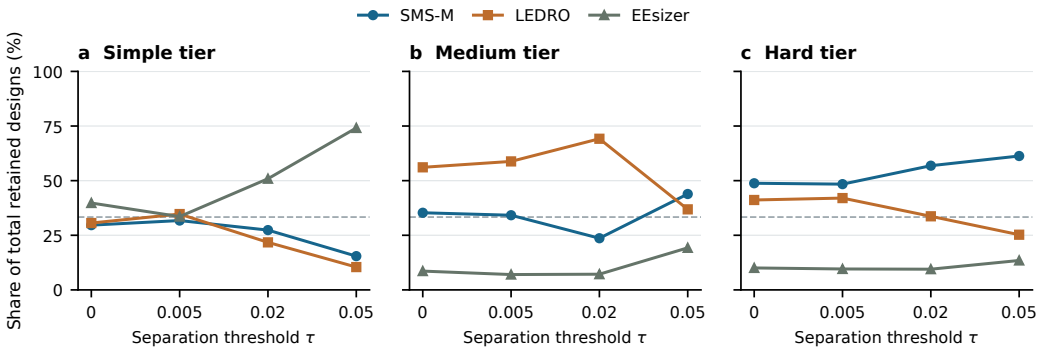


Figure 3: **Relative shares of retained designs across separation thresholds.** Panels show the `simple`, `medium` and `hard` tiers. Curves report each method's share (%) of retained designs at $\tau \in \{0, 0.005, 0.02, 0.05\}$ for SMS-M, LEDRO and EEsizer. The dashed horizontal line marks an equal three-way share (33.3%). Each tier uses 18 runs per method at $N = 100$. AutoSizer is omitted because it returns no new feasible design.

Reaching distant designs across spatial scales. Archive counts at $\tau = 0$ can obscure spatial dispersion. Downstream physical design requires mutually distinct alternatives rather than tightly clustered points. Sweeping the separation threshold τ (Figure 3) highlights contrasting allocation behaviors:

- **Trade-offs between local density and manifold coverage:** On `simple`, LEDRO returns more raw designs (56.1 vs. 35.3) by concentrating samples in a dense cluster near the starting seed (reach of 0.038 vs. 0.127 for SMS-M). Figure 3 shows how the methods' relative contributions change as the separation threshold increases.
- **Sustained reach away from the seed:** In contrast, SMS-M targets structural diversity along the feasible manifold, achieving greater reach (0.070 vs. 0.025) and higher worst-case margin (0.025 vs. 0.012) on `medium`. At the largest displayed threshold, $\tau = 0.05$, its share of retained designs exceeds LEDRO's; the ordering differs at finer resolutions.

Continuous domain mobility vs. categorical discretization. The adapted AutoSizer agent returns no feasible designs beyond the initial seed across all 54 runs (Table 1). AutoSizer's architecture pre-groups parameters and discretizes continuous search ranges into a coarse categorical grid. In continuous mixed-domain spaces, this discretization fractures narrow feasible corridors. Without continuous coordinate mobility or sensitivity-guided directions, discrete grid steps repeatedly step across constraint boundaries into infeasibility, demonstrating that continuous parameter mobility is essential in tight analog feasibility manifolds.

4.3 Ablation and Mechanism Attribution

Each experiment in our evaluation fulfills a distinct evidential role: (1) **Global sparsity audit:** 3,000 uniform draws yield zero feasible designs (0/3,000, Appendix A), proving unguided search cannot recover valid sizing vectors. (2) **Stream accounting:** Anchored proposals convert at 59.4% (774/1,302), supplying 95.0% of feasible evaluations (774/815) from 43.4% of budget, while box-fill converts 1.0% (11/1,052) and uniform search converts 0.0% (0/616). (3) **Stepwise component isolation:** Isolating components reveals what factors drive volume versus spatial spread. (4) **Multi-method baselines:** Comparing against published agents maps density-versus-coverage trade-offs across resolutions.

Decoupling count from placement: anchor perturbation vs. surrogate scoring. The combined matrix in Table 2 provides a clean, auditable attribution of SMS-SIZING's mechanisms:

- **Anchored perturbation drives feasible count:** `AnchorOnly` retains sparse anchor perturbations without surrogate models or bound updates. It matches SMS-M in raw volume

Table 2: **Ablation matrix and step-by-step mechanism attribution on `medium` ($N = 100$, 30 runs per arm). Top:** Component cohort search trajectories ($45/40/15$ schedule; new designs exclude the seed; H_N^+ removes seed evaluations). **Bottom:** Stepwise isolation of archive geometry (final box volume ratio, reach to seed, and best worst-case margin across completed $30/30$ paired runs).

Part A: Search Trajectory and Component Ablations

Configuration	New designs	H_N (%)	H_N^+ (%)	Warm-up	Local	Final
SMS-M	**26.2**	$\mathbf{27.17 \pm 10.26}$	**26.43**	2.7	45.3	**52.2**
Matched reference for component ablations: SMS-B-L						
SMS-B-L	18.7	19.73 ± 7.57	18.92	3.6	44.9	1.1
`NoMemory`	19.0	19.97 ± 8.04	19.16	2.7	46.1	2.0
`NoBoundary`	13.6	14.63 ± 6.85	13.77	2.4	33.7	0.4
`NoDirection`	0.6	1.60 ± 1.40	0.61	2.7	0.9	0.0
`NoDimReduction`	0.7	1.67 ± 1.63	0.68	2.7	1.1	0.2
Sobol/LHS	0.0	4.00 ± 0.00	0.00	4.4	5.0	0.0

Part B: Stepwise Mechanism Attribution of Archive Geometry

Ablation Arm	Box updates	Surrogate	Controller	Final Vol.	Reach	Margin
SMS-M	Contracted	Yes	Served LLM	1.65e-2	**0.0706**	**0.0255**
`NoContraction`	Frozen (1.0)	Yes	Served LLM	1.0000	**0.0712**	0.0216
`NoGate`	Ungated	Yes	Served LLM	7.00e-4	0.0692	0.0207
SMS-M, offline	Contracted	Yes	Offline rule	1.65e-2	0.0431	0.0215
`AnchorOnly`	Frozen (1.0)	None	Offline rule	1.0000	0.0146	0.0106

(26.6 vs. 26.2 new designs across $30/30$ completed runs, Appendix D.8). Sparse perturbation of verified feasible anchors operates as the primary volume generation engine.

- **Surrogate-scored selection drives reach and margin:** Without surrogate scoring, `AnchorOnly`'s reach drops nearly $5\times$ ($0.0706 \rightarrow 0.0146$) and its margin drops from 0.0255 to 0.0106. When an offline rule replaces the served LLM controller while keeping the surrogate, reach shifts from 0.0706 to 0.0431. Thus, surrogate margin ranking and uncertainty quotas act as the navigation system that steers candidates outward along the manifold.
- **Stability across bounding box volume:** Freezing the bounding box at the full domain (`NoContraction`, volume ratio 1.0) yields a reach of 0.0712, matching SMS-M's 0.0706, while unconstrained contraction (`NoGate`, volume ratio 7.0×10^{-4}) yields 0.0692. Across three orders of magnitude of box volume, spatial reach remains flat at ≈ 0.071. Candidate placement is therefore governed by surrogate-scored quota filtering rather than geometric box boundaries.
- **Sensitivity guidance in local moves:** While `NoDirection` removes anchor proposals entirely, isolating coordinate selection alone (`SensitivityDirectedLocal`, Appendix D.10) shows that biasing coordinate choices by sensitivity scores yields 19.63% hit rate versus 19.73% for uniform coordinate selection ($p = 0.34$). Sensitivity guidance functions primarily by screening coordinate impact for pool filtering, while uniform coordinate draws maintain local exploration diversity.

Final-phase allocation sustains late-stage expansion. While SMS-M and SMS-B show similar hit rates during local exploration ($\approx 45\%$), SMS-B encounters yield exhaustion in the final 15 calls (1.1%), whereas SMS-M's quota-and-margin selector sustains a 52.2% rate (Table 2 and Figure 4). Allocating evaluations to boundary candidates (`NoBoundary`) improves pairwise separation, and measurement recall (`NoMemory`) stabilizes post-search confirmations.

4.4 Analysis

Synthesizing the baseline comparisons and ablation trajectories isolates why SMS-SIZING succeeds where prior approaches falter: it decouples high-conversion candidate generation from directional manifold navigation. While localized trust-region methods (e.g., LEDRO) cluster redundantly near

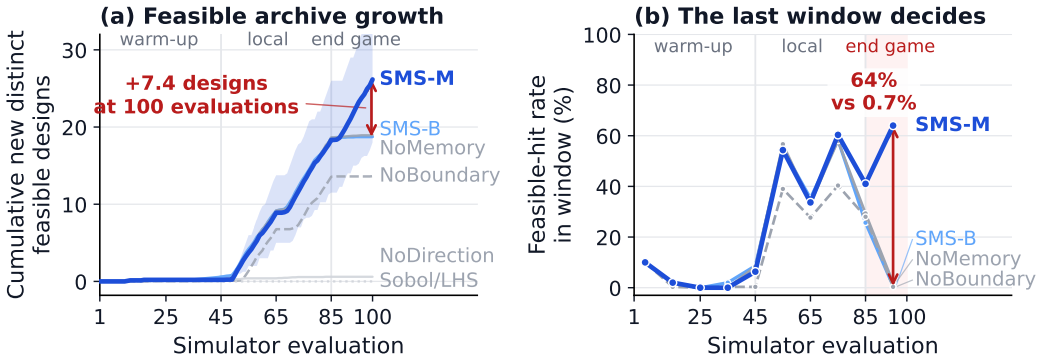


Figure 4: **The final selector sustains archive expansion.** Means over 30 component-study runs; boundaries mark evaluations 45 and 85. (a) Cumulative new distinct feasible designs, excluding the seed; shading is the SMS-M interquartile range. (b) Hit rate in ten-evaluation windows. The 64% and 0.7% annotations refer to the final ten-evaluation window. SMS-M and SMS-B have similar local-phase yield and diverge in the final phase.

the starting seed and unguided perturbations (`AnchorOnly`) collapse in spatial dispersion (reach 0.015), SMS-SIZING leverages sparse anchor perturbation to maintain a 59.4% feasible conversion engine while using surrogate-scored quota selection to steer candidates outward, achieving nearly 5× greater reach (0.071) and doubling constraint margins. Crucially, aligning with modern neural circuit modeling (Poddar et al., 2025; Yang et al., 2025; Mehradfar et al., 2025), the surrogate functions as an efficient ordinal ranking and coordinate screening filter rather than a hard feasibility gate; search yield is decoupled from surrogate regression fit ($r_{\text{fit}} = 0.03$), preventing model approximation errors from discarding physically viable designs while SPICE simulation provides authoritative ground truth. Furthermore, preserving continuous parameter mobility avoids the catastrophic boundary failures of categorical discretization seen in AutoSizer, and controller substitution confirms that search efficiency stems from the gated anchored architecture itself (25.36% offline vs. 26.20% served LLM) rather than any proprietary language model.

5 CONCLUSION

SMS-SIZING provides a sample-efficient framework for analog circuit sizing: expanding a single simulator-verified seed into an extensive feasible archive under tight simulation budgets. By decoupling sparse anchor perturbation from surrogate-scored frontier quota allocation in the original continuous mixed-domain space, SMS-SIZING discovers tens of mutually distinct, simulator-confirmed feasible designs across benchmark operational amplifiers. In nominal `tt` simulations across six topologies, anchored candidates convert at 59.4% (774/1,302), supplying 95% of all feasible evaluations, while the full method achieves 2.8× the spatial reach and 2.1× the worst-case constraint margin of LEDRO on the medium tier. Guided by learned rankings rather than geometric confinement, verified feasible designs serve as dynamic search anchors for expensive black-box search on constrained engineering manifolds.

6 FUTURE WORK

Future work will test whether the candidate archives improve multi-point yield after layout extraction and under process-voltage-temperature (PVT) variation and electromagnetic constraints. We will also assess bandgaps, phase-locked loops, and data converters, as well as FinFET and FD-SOI nodes, beyond the SkyWater 130nm amplifier benchmarks studied here.

AI USE STATEMENT

Generative AI assisted with manuscript polishing, editing, figure redrawing, code inspection and consistency checks against project records. Quantitative tables and plots use archived simulation records. A language model also serves as the policy controller studied in Section 3; the offline comparison in Section 4.4 measures controller substitution. The authors reviewed all AI-assisted material and remain responsible for the final text, figures, code and scientific claims.

REPRODUCIBILITY STATEMENT

The main text and appendices specify the problem definition, execution order, algorithm constants, prompt schemas, baseline adaptations and analysis procedures. The anonymous source code, packaged surrogate ensemble, recorded evaluation trajectories, and analysis scripts required to reproduce all reported tables and figures are publicly available at `https://anonymous.4open.science/r/SMS-sizing-9141/`.

REFERENCES

Mohsen Ahmadzadeh, Kaichang Chen, and Georges Gielen. AnaFlow: Agentic LLM-based workflow for reasoning-driven explainable and sample-efficient analog circuit sizing. In *IEEE/ACM International Conference on Computer Aided Design (ICCAD)*, pp. 1–7, 2025. doi: 10.1109/ICCAD66269.2025.11240818. URL https://arxiv.org/abs/2511.03697.

Ahmet Faruk Budak, Prateek Bhansali, Bo Liu, Nan Sun, David Z. Pan, and Chandramouli V. Kashyap. DNN-Opt: An RL inspired optimization for analog circuit sizing using deep neural networks. In *58th ACM/IEEE Design Automation Conference (DAC)*, pp. 1219–1224, 2021. doi: 10.1109/DAC18074.2021.9586139.

Antoine Cully, Jeff Clune, Danesh Tarapore, and Jean-Baptiste Mouret. Robots that can adapt like animals. *Nature*, 521(7553):503–507, 2015. doi: 10.1038/nature14422.

David Eriksson and Martin Jankowiak. High-dimensional Bayesian optimization with sparse axis-aligned subspaces. In *Proceedings of the 37th Conference on Uncertainty in Artificial Intelligence*, volume 161 of *Proceedings of Machine Learning Research*, pp. 1508–1517, 2021. URL https://proceedings.mlr.press/v161/eriksson21a.html.

David Eriksson and Matthias Poloczek. Scalable constrained Bayesian optimization. In *Proceedings of the 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pp. 730–738, 2021. URL https://proceedings.mlr.press/v130/eriksson21a.html.

David Eriksson, Michael Pearce, Jacob R. Gardner, Ryan Turner, and Matthias Poloczek. Scalable global optimization via local Bayesian optimization. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Jacob R. Gardner, Matt J. Kusner, Zhixiang Xu, Kilian Q. Weinberger, and John P. Cunningham. Bayesian optimization with inequality constraints. In *Proceedings of the 31st International Conference on Machine Learning*, volume 32 of *Proceedings of Machine Learning Research*, pp. 937–945, 2014. URL https://proceedings.mlr.press/v32/gardner14.html.

Michael A. Gelbart, Jasper Snoek, and Ryan P. Adams. Bayesian optimization with unknown constraints. In *Proceedings of the 30th Conference on Uncertainty in Artificial Intelligence*, pp. 250–259, 2014. URL https://arxiv.org/abs/1403.5607.

Dimple Vijay Kochar, Hanrui Wang, Anantha Chandrakasan, and Xin Zhang. LEDRO: LLM-enhanced design space reduction and optimization for analog circuits. In *2025 IEEE International Conference on LLM-Aided Design (ICLAD)*, 2025. doi: 10.1109/ICLAD65226.2025.00011. URL https://arxiv.org/abs/2411.12930.

Joel Lehman and Kenneth O. Stanley. Abandoning objectives: Evolution through the search for novelty alone. *Evolutionary Computation*, 19(2):189–223, 2011. doi: 10.1162/EVCO_a_00025.

Jintao Li, Haochang Zhi, Ruiyu Lyu, Wangzhen Li, Zhaori Bi, Keren Zhu, Yanhan Zhen, Weiwei Shan, Changhao Yan, Fan Yang, Yun Li, and Xuan Zeng. AnalogGym: An open and practical testing suite for analog circuit synthesis. In *IEEE/ACM International Conference on Computer-Aided Design (ICCAD)*, pp. 1–9, 2024. doi: 10.1145/3676536.3697117. URL https://arxiv.org/abs/2409.08534.

Chang Liu and Danial Chitnis. EEsizer: LLM-based AI agent for sizing of analog and mixed signal circuit. *IEEE Transactions on Circuits and Systems I: Regular Papers*, pp. 3825–3834, 2026. doi: 10.1109/TCSI.2025.3646359. URL https://arxiv.org/abs/2509.25510.

Asal Mehradfar, Xuzhe Zhao, Yilun Huang, Emir Ceyani, Yankai Yang, Shihao Han, Hamidreza Aghasi, and Salman Avestimehr. FALCON: An ML framework for fully automated layout-constrained analog circuit design. In *Advances in Neural Information Processing Systems 38*, 2025. doi: 10.52202/085713-1819. URL https://arxiv.org/abs/2505.21923.

Jean-Baptiste Mouret and Jeff Clune. Illuminating search spaces by mapping elites. *arXiv preprint arXiv:1504.04909*, 2015. URL https://arxiv.org/abs/1504.04909.

Sounak Poddar, Youngmin Oh, Yao Lai, Hanqing Zhu, Bosun Hwang, and David Z. Pan. INSIGHT: A universal neural simulator framework for analog circuits with autoregressive transformers. In *2025 62nd ACM/IEEE Design Automation Conference (DAC)*, pp. 1–7, 2025. doi: 10.1109/DAC63849.2025.11133292. URL https://arxiv.org/abs/2407.07346.

Keertana Settaluri, Ameer Haj-Ali, Qijing Huang, Kourosh Hakhamaneshi, and Borivoje Nikolic. AutoCkt: Deep reinforcement learning of analog circuit designs. In *Design, Automation and Test in Europe Conference (DATE)*, pp. 490–495, 2020. URL https://arxiv.org/abs/2001.01808.

Hanrui Wang, Kuan Wang, Jiacheng Yang, Linxiao Shen, Nan Sun, Hae-Seung Lee, and Song Han. GCN-RL circuit designer: Transferable transistor sizing with graph neural networks and reinforcement learning. In *57th ACM/IEEE Design Automation Conference (DAC)*, pp. 1–6, 2020. doi: 10.1109/DAC18072.2020.9218757.

Ziyu Wang, Masrour Zoghi, Frank Hutter, David Matheson, and Nando de Freitas. Bayesian optimization in a billion dimensions via random embeddings. In *Proceedings of the 30th International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research*, pp. 1022–1030, 2013.

Ziming Wei, Zichen Kong, Yuan Wang, David Z. Pan, and Xiyuan Tang. TopoSizing: An LLM-aided framework of topology-based understanding and sizing for AMS circuits. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, 2026. doi: 10.1109/TCAD.2026.3685525. URL https://arxiv.org/abs/2509.14169.

Xiaomeng Yang, Jian Gao, Yanzhi Wang, and Xuan Zhang. ZeroSim: Zero-shot analog circuit evaluation with unified transformer embeddings. In *2025 IEEE/ACM International Conference on Computer-Aided Design (ICCAD)*, pp. 1–9, 2025. doi: 10.1109/ICCAD66269.2025.11240890. URL https://arxiv.org/abs/2511.07658.

Xi Yu, Dmitrii Torbunov, Soumyajit Mandal, and Yihui Ren. AutoSizer: Automatic sizing of analog and mixed-signal circuits via large language model (LLM) agents. *arXiv preprint arXiv:2602.02849*, 2026. URL https://arxiv.org/abs/2602.02849.

H. Zhang, W. Fan, X. Gao, B. Liu, R. Wang, and Y. Lin. PANDA: An LLM-enhanced performance-driven analog design framework bridging design intent and layout generation. *arXiv preprint arXiv:2606.15052*, 2026. URL https://arxiv.org/abs/2606.15052.

A Target Construction and Feasibility Diagnostics

Target construction. The benchmark constructs thresholds by quantile tightening in each metric's favorable direction, using levels 0.2, 0.35 and 0.5 for its simple, medium and hard tiers. The component study uses medium; the released-code comparison reports across all three tiers (simple, medium, and hard). Across 48 benchmark tasks and 384 task–constraint pairs, the supplied audit finds 24 reference metrics lying exactly on a constraint surface, and the maximum relative difference between a reference metric and its threshold is 0.9845. The targets are therefore distribution quantiles rather than the reference design's own metrics, and clearing them is a property of the design rather than of the construction. Appendix D.1 audits the archived default initialization separately.

Metric conventions. The benchmark maximizes DC gain (dB), GBW (Hz), and phase margin (degrees). It minimizes power (μW), offset (V), and the stored signed CMRR/PSRR quantities (dB). Rejection values therefore must be compared in their recorded sign convention; replacing them by positive rejection magnitudes would reverse the inequalities. Slew rate, temperature coefficient and settling time are additional surrogate outputs, not hard constraints in this AC/DC experiment.

Why the conjunction is the binding difficulty. Individual constraints are not what makes this space hard. Under uniform sampling, phase margin passes up to 94.0% of the time and GBW up to 50.8%, yet no draw in $3{,}000$ satisfies all eight at once (Table 3). Figure 5 places that result next to the slack of the designs the search does find: uniform feasibility is sparse and discovered designs often have small non-offset margins. This motivates reuse of verified anchors.

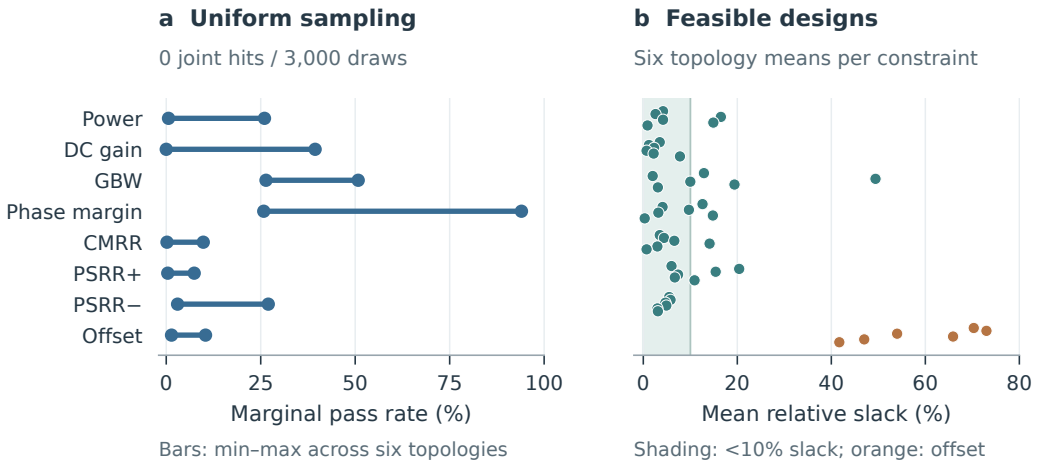


Figure 5: **Sparse feasibility and small measured slack.** (a) Range of marginal pass rates across six topologies in the uniform-sampling audit; no draw satisfies all eight constraints jointly. (b) Mean relative slack in each topology–constraint cell, from 815 feasible records. Each point is a cell. The shaded range marks slack below 10%. Tables 3 and 4 give the underlying numbers.

Table 3: Marginal feasibility under uniform sampling. Entries are the minimum and maximum pass rates over six topologies; the joint observation pools 3,000 draws. The product of marginals is the value the conjunction would take if the constraints were independent; they share circuit parameters, so it is a reference point rather than a bound.

Constraint	Minimum (%)	Maximum (%)
Phase margin	25.8	94.0
GBW	26.4	50.8
DC gain	0.0	39.4
Negative PSRR (PSRN)	3.0	27.0
Power	0.6	26.0
Offset (`vos`)	1.4	10.4
CMRR (`cmrrdc`)	0.2	9.8
Positive PSRR (PSRP)	0.4	7.4
Jointly feasible	$0/3000$	
Product of marginals (probability)	0 to 4.27×10^{-7}	

Slack of discovered designs. Table 4 summarizes the 815 feasible records from the primary $N = 100$ matrix, averaging relative slack within each topology and constraint. These are the designs the method reaches, and they sit against their thresholds: the median non-offset cell has 4.8% slack and 31 of 42 cells fall below 10%. Offset is the exception at 41.7%–73.0%, so its average slack is larger in this sample. The measurement characterizes the region the anchored moves occupy; Section 4.2 reports how far into the interior those moves subsequently travel.

Table 4: Mean relative slack (%) for each topology–constraint cell. PM denotes phase margin. The median over the 42 non-offset cells is 4.8%; 31 cells have slack below 10%. Offset ranges from 41.7% to 73.0%.

Topology	d	Power	Gain	GBW	PM	CMRR	PSRP	PSRN	Offset
Alfio_RAFFC	27	4.2	3.5	12.9	12.6	3.5	6.0	5.5	70.3
Peng_IAC	31	2.6	1.2	2.0	4.1	4.4	20.4	5.8	73.0
Fan_SMC	23	16.5	2.3	49.4	9.7	6.6	15.4	4.7	54.0
Sau_CFCC	29	4.2	0.7	10.0	3.2	14.1	7.4	4.9	65.9
Leung_DFCFC1	30	14.9	2.2	19.4	14.8	3.0	6.7	3.0	47.0
Qu2017_AZC	38	0.9	7.8	3.1	0.3	0.7	10.9	3.1	41.7

B Algorithm Details and Conditional Contraction Analysis

B.1 Round procedure

Figure 6 states the per-round procedure in the order the implementation executes it. Given the random seed, simulator environment and controller decisions, the procedure specifies candidate generation, evidence checks and archive updates.

Inputs: current box B_r, original box $\mathcal{X}$, feasible archive A_r, full measurement history, random-generator state and remaining budget.

1. Select up to eight distinct measured anchors, preferring feasible designs.
2. Form the pool from original-box sentinels, sparse anchor perturbations, and Latin hypercube samples of B_r.
3. Predict the eleven metrics and their uncertainty with the surrogate.
4. Append the last 100 distinct successful simulator records, using measured metrics and zero predictive uncertainty for those records.
5. Compute sensitivity rankings, automatic shrink suggestions, feasibility masks, and empirical candidate components.
6. Obtain one controller policy containing actions, bound proposals, and quotas.
7. Apply the evidence gate. Accept admissible bounds as B_{r+1}; otherwise retain B_r.
8. Fill the simulator batch using the evaluation quotas. Identify global candidates by sampling origin.
9. Simulate each selected candidate. Record failures against the budget.
10. Store every measurement in history and add feasible designs to A_{r+1} before round $r + 1$.

Output: updated bounds, measured archive, and remaining budget.

Figure 6: Disclosed per-round procedure. The surrogate-scoring step supplies subsequent analysis and ordering; final feasibility is measured by the simulator.

B.2 Module contracts

Sensitivity and automatic proposals. The analyzer receives a candidate-by-parameter matrix, eleven surrogate metrics per candidate, and a conservative-feasibility mask. It computes per-metric absolute correlations and their mean over outputs (Eq. 5). An automatic shrink suggestion requires at least eight finite coordinate values, at least four conservatively feasible values, and aggregate effect at most 0.05. Their 5th and 95th percentiles define the proposed interval. The interval must be more than 5% narrower than the original metadata range. The loop then intersects suggestions with the current box before gate validation. At most ten automatic suggestions are emitted. Controller-generated proposals are a separate channel.

Anchors and local moves. Anchors come from distinct archive records whose simulations completed successfully. Feasible points take priority. Eligible near-feasible points pass at least six of the eight constraints and have total normalized deficit at most one; smaller deficit and larger worst-case margin determine their ordering. Local sampling uses Eq. 6. The center is first clamped into the current box. Coordinates outside the active subset remain at that clamped center. Consequently, a retained coordinate need not equal its historical measured value if its anchor lies outside the updated box.

Region tracker. The analyzer constructs the conservative and optimistic feasibility masks using Eq. 7, with measured records assigned zero spread. The tracker groups their union by distances in coordinates divided by the standard deviation of the candidate feasible points. A linkage threshold of $0.35\sqrt{d}$ gives each empirical component a representative and a bounding box. The same scored pool produces the same grouping. The guarantee of retaining a representative refers to this empirical grouping.

Evidence gate. The gate receives proposed actions, original and current bounds, evidence counts, and tracked representatives. With at least four conservative records or two distinct simulator-

validated feasible points, the maximum single-round width reduction is 50%. Under weaker evidence, at most two coordinates can shrink, by at most 25% each. Bounds must stay inside the original box and retain every tracked representative. Counting distinct measured vectors prevents repeated recall of one design from increasing the validated-evidence count.

B.3 Sampling, recall, and freezing

Pool generation and evaluation allocation are separate operations. The default pool of $50{,}000$ candidates contains $10{,}000$ global sentinels, one default, and, when anchors exist, $6{,}000$ local proposals. The latter count is $\mathrm{round}[0.15(50{,}000-10{,}000-1)]$. Latin hypercube sampling fills the remainder; before an anchor is available, the local stream is empty. A five-way policy allocates the simulator batch among global, boundary, Pareto, uncertainty, and final categories. Global eligibility follows sampling origin; fallback batch fill ignores predicted feasibility. Before filling the boundary, Pareto, uncertainty and final categories, the selector reserves $\lceil 0.30(n_r - n_{r,\mathrm{global}}) \rceil$ slots for available local candidates, where n_r is the round's evaluation budget. Local and boundary candidates are ordered by optimistic constraint count, total normalized deficit and worst margin, followed by diversity selection from a short list. This ranking uses a full spread on every metric, separately from the analyzer's asymmetric weights in Eq. 7. Parameter keys rounded to 14 decimal places remove already attempted designs and within-batch duplicates on the agent path.

Recall supplies measured evidence to sensitivity analysis, region classification, and the gate. It also supports a Pareto set that can contain predicted and measured records, with ties favoring the simulator-validated point. An optional topology-understanding pass supplies interacting parameter groups. Freezing requires a width no greater than 5% of the original. Members of an identified interacting group must satisfy this condition jointly. No freezing is reported in the evaluated matrix.

B.4 Reproducibility constants

Table 5 lists every algorithm constant the loop reads, including the quota fractions, the anchor cap and the evidence thresholds the gate applies. None is tuned per topology; the evaluated matrix uses these values throughout.

Table 5: Disclosed defaults and scheduling constants. Evaluated configurations can override defaults as specified in Section 4.1.

Setting	Value
Surrogate pool size	50,000 candidates
Automatic effect cutoff	0.05
Automatic shrink interval	$[5\%, 95\%]$ quantiles
Global pool fraction	0.20
Local sampling setting	0.15 of non-global, non-default remainder
Local evaluation reservation	30% of non-global slots, rounded up
Local radius	0.10 of current coordinate width
Local active fraction	0.05 of dimension, at least one coordinate
Maximum anchors	8 distinct measured points
Component linkage radius	$0.35\sqrt{d}$ in standardized coordinates
Recall depth	Last 100 distinct successful simulations
Full-shrink evidence	4 conservative or 2 distinct validated feasible records
Weak-evidence limit	2 coordinates, 25% width reduction each
Maximum width reduction	50% per round
Freeze condition	Width $\leq 5\%$ of original
Controller bound actions	At most 6 per round
Evaluation categories	Global, boundary, Pareto, uncertainty, final
Schedule for $N = 100$	45 warm-up, 40 adaptive, 15 final evaluations
Schedule for $N = 200$	40 warm-up, 120 adaptive, 40 final evaluations

B.5 When contraction increases uniform hit probability

Box contraction trades search volume against retained feasible mass, and the trade is exact. For the continuous relaxation, normalize B_0 to unit volume and let the feasible set have positive volume fraction μ. For an axis set S, write the retained width ratios as $\rho_i \in (0, 1]$ and $v_S = \prod_{i \in S} \rho_i$, and define the retained feasible fraction

$$\delta(S) = \frac{\mathrm{vol}(\mathcal{F} \cap B_S)}{\mathrm{vol}(\mathcal{F} \cap B_0)}.$$

This is a statement about volume, distinct from the fraction of an adaptive archive that the gate retains.

Proposition 1 (Uniform-sampling identity) *For $\mu > 0$ and $v_S > 0$, the ratio of uniform feasible-hit probabilities in B_S and B_0 is*

$$G(S) = \frac{\delta(S)}{\prod_{i \in S} \rho_i}. \tag{8}$$

The feasible probabilities are $\delta(S)\mu/v_S$ and μ, so contraction pays exactly when $\delta(S) > v_S$: a box that discards more volume than feasible mass. The identity separates volume concentration from feasible-mass retention. The implemented gate checks finite-pool representatives; those checks do not estimate $\delta(S)$ or establish that it exceeds v_S.

Proposition 2 (Conditional sensitivity to axis selection) *Consider two k-axis choices S and S^* with $e = |S^* \setminus S| = |S \setminus S^*|$ replacements and the same retained width on every shared coordinate. If $\delta(S^*) > 0$ and $\rho_{\min} = \min_{i \in S^*} \rho_i$, then*

$$\frac{G(S)}{G(S^*)} \geq \frac{\delta(S)}{\delta(S^*)} \, \rho_{\min}^{e}.$$

Shared widths cancel; each removed factor is at least $\rho_{\min}$ and each added factor is at most one. Under the additional condition $\delta(S) = \delta(S^*)$, the lower bound becomes $\rho_{\min}^{e}$. This is a conditional comparison of uniform sampling in continuous boxes. It does not establish ranking correctness or retention for mixed integer parameters, sparse anchor moves, or the empirical gate. Those behaviours are measured by the component and ranking studies in Sections 4.3 and 4.4.

C Surrogate and Execution Details

Data and ensemble. The surrogate-development dataset contains 60,000 rows for each of 16 topologies. Its metadata records a deterministic topology–parameter-group hash split into 770,800 training, 95,227 validation and 93,973 test rows. This split evaluates new parameter combinations within represented topologies; it is not a held-out-topology evaluation. The package uses five neural members with topology-by-metric weights selected on validation data and averages their predictions in physical units. The eleven outputs are power, DC gain, GBW, phase margin, CMRR, positive and negative PSRR, offset, slew rate, temperature coefficient, and settling time. Only the first eight enter the hard feasibility test. The package's offline expert-fusion test cache is distinct from its deployable ensemble and is not used as an oracle for new search candidates.

Accuracy of the deployed ensemble. Over all 16 training topologies and 11 outputs the deployed ensemble reaches a macro R^2 of 0.8586, a median cell R^2 of 0.9394 and a macro SMAPE of 10.86% on the blind test split; the validation split gives 0.8664, 0.9469 and 10.97%. For the 48 cells covering the six evaluated topologies and eight hard constraints, the same ensemble gives a macro R^2 of 0.853, a median cell R^2 of 0.941 and a macro SMAPE of 5.00%. Per-topology macro R^2 over the eight constraints is 0.922 (`Alfio_RAFFC`), 0.922 (`Fan_SMC`), 0.890 (`Peng_IAC`), 0.848 (`Sau_CFCC`), 0.821 (`Leung_DFCFC1`) and 0.717 (`Qu2017_AZC`). Table 6 records the lowest-R^2 cells, including five evaluated hard-constraint cells below 0.5. The model-selection metadata uses the status `target_not_reached`. These diagnostics describe the predictor used by the search; rank repeatability is measured separately in Table 15. Simulator measurements determine archive membership.

Table 6: The twelve lowest-R^2 topology–metric cells of the deployed ensemble. $\star$ marks a cell that is both on one of the six evaluated topologies and on one of the eight hard constraints. `Yan_AZ`, `Peng_ACBC`, `Tan_CLIA` and `Song_DACFC` belong to the surrogate's 16-topology training corpus but are not evaluated in this paper; TC, SR and settling time are auxiliary outputs that never enter the feasibility test.

Topology	Metric	R^2
Yan_AZ	GBW	0.004
Qu2017_AZC	DC gain $\star$	0.025
Qu2017_AZC	Phase margin $\star$	0.119
Yan_AZ	Offset	0.230
Leung_DFCFC1	DC gain $\star$	0.305
Sau_CFCC	Offset $\star$	0.316
Sau_CFCC	Temperature coefficient	0.342
Yan_AZ	Settling time	0.365
Peng_ACBC	DC gain	0.377
Tan_CLIA	DC gain	0.437
Qu2017_AZC	Slew rate	0.446
Song_DACFC	Temperature coefficient	0.543

Uncertainty updates. Predicted means remain unchanged during online calibration. For positive spread, the implementation accumulates the paired ratios $r_k = |g_k - \hat{g}_k|/\sigma_k$. After at least four pairs, it computes

$$\gamma_k = \mathrm{clip}\left(\frac{\mathrm{median}(r_k)}{0.67448975},\, 0.5,\, 8\right) \tag{9}$$

and blends this target into the stored metric scale using pair counts. The constant converts a median absolute Gaussian residual to a scale estimate; it does not imply that residuals are Gaussian. Zero-spread cases additionally use empirical relative-error floors. Logged affine offsets are diagnostic and do not alter the mean predictions. These updates inform the asymmetric margins in Eq. 7; the retrospective filter in Appendix D.17 instead applies the same κ to every metric.

Finite-pool region estimates. The component routine uses at most $2{,}000$ finite candidates from the union of conservative and optimistic masks, chosen by deterministic subsampling. It standardizes each coordinate and connects neighbors within $0.35\sqrt{d}$. The gate protects representatives of these empirical components. The retention check is defined over those representatives; unobserved components are outside that finite-pool check.

Historical final selector. The implemented selector ranks a fixed pool using

$$A(x) = p(x)^2\big(1 + 0.25v(x) + 0.1|b(x)|\big). \tag{10}$$

Here p averages clipped, spread-normalized per-constraint safety scores; $v = p(1 + 0.25u)$ uses a pool-normalized objective utility u; and on the executed code path b averages the raw metric uncertainty mapping, which mixes physical units. The executed routine is a fixed-pool ranking heuristic; it fits no Gaussian process. Manuscript references to SMS-B's historical selector denote this rule, including where legacy logs use an acquisition-function label. SMS-M substitutes quota and margin selection in the final phase. The separate surrogate-enhanced BO control and the released TuRBO kernel used by LEDRO are different comparators.

Controller provenance and runtime. The recorded launch configuration names Gemini-3.8-Flash as the served controller, reached through an OpenAI-compatible interface; the secondary matrix records the provider identifier `gemini-3.8-flash` directly, while the primary artifacts log the controller's decisions, attempts and request latency without re-recording the model identifier per run. Settings are temperature 0.1, an initial output limit of 4096 tokens and a $180\,\mathrm{s}$ request timeout, with retries able to raise the limit to $32{,}768$. The offline-controller arm of Section 4.4 runs the same loop with no served model and uses a deterministic policy with the recorded random seeds. Archived job scripts request four CPUs and 32 GB per task; these are resource requests rather than measured utilization. The reported budget counts simulator evaluations and excludes surrogate pretraining, pool inference and controller calls.

Policy schema and gate outcomes. The controller returns one JSON object per round with nine required fields: `action` (one of `continue`, `shrink`, `expand`, `rollback`, `stop`), `exploration_ratio`, `exploitation_ratio`, `uncertainty_ratio`, `parameter_actions` (a list of per-coordinate bound proposals), `real_eval_quota` (the five-way evaluation quota), `confidence`, `evidence` (free-text justification, not parsed), and `stop_reason`. A response missing any field, using an unsupported action, or proposing a malformed quota is rejected and returned to the controller with the specific failing field named, for up to three attempts in total. Pooled over the 30 SMS-M `medium`-tier runs, the gate accepts 121 of 139 recorded verdicts; the 18 rejections are dominated by two failure modes, a proposed contraction that would delete a region holding a competitive candidate (14) and a proposed contraction that would drop a tracked feasible island's representative (10, some verdicts triggering more than one reason). No response was rejected for malformed JSON or an unsupported action in this sample, so the gate's binding criteria here are the evidence and retention checks rather than the schema check.

Controller prompt templates. The online LLM agent operates under two formal prompt templates:

1. **Circuit Understanding Prompt (`CIRCUIT_UNDERSTANDING_PROMPT`):** Formatted during initialization to produce a static qualitative baseline for parameter roles and signal flow:

```
You are an expert analog integrated-circuit designer.
Analyze one fixed amplifier topology for parameter sizing.
Base every statement on the provided netlist, testbench,
parameter schema, and metric definitions.  The 11 metric
names are exact and must not be renamed.
Topology:  {topology}
Netlist:  {netlist}
Testbench:  {testbench}
Active parameter schema:  {parameter_schema}
Metric schema:  {metric_schema}
Return one JSON object with exactly these fields:
"circuit_topology_overview",
"optimization_variables_mapping",
"optimization_variables_impact", "variable_interactions",
"key_insights_for_optimization", and "confidence".  Return
JSON only.  Do not propose numeric bounds and do not
generate code.
```

2. **Search Policy Prompt (`SEARCH_POLICY_PROMPT`):** Executed at each adaptive round to propose regional bounds and evaluation allocations:

```
You control search strategy for one fixed analog topology.
Use both the circuit-understanding evidence and the
numerical sensitivity evidence.  Circuit understanding is
qualitative; numerical sensitivity, feasibility, OOD, and
real-evaluation results are quantitative.  Do not invent
parameter names or metrics.  Do not output code and do not
directly execute a space update.
Context JSON: {context_json}
If `safety_gate_feedback` is present in the context, it is
a hard correction request.  Address every listed reason
using the current search_space bounds before returning a
new policy.
Return JSON only with this schema:  "action" (continue|
shrink|expand|rollback|stop), "exploration_ratio",
"exploitation_ratio", "uncertainty_ratio",
"parameter_actions" (["parameter", "action", "lower",
"upper", "reason"]), "real_eval_quota" ("global",
"boundary", "pareto", "uncertainty", "final"), "confidence",
"evidence", "stop_reason".
The three sampling ratios must each lie in [0, 1] and sum
to 1.  Include at most six parameter actions.  The sum of
real_eval_quota fields must not exceed both
budget.remaining and budget.stage_limits[budget.stage].
```

```
Every proposed shrink must stay inside search_space.bounds
and search_space.original_bounds, and may reduce each
parameter's current width by no more than 50% in this round.
If numerical_sensitivity.shrink_evidence reports fewer than
4 conservative feasible samples and fewer than 2 distinct
validated feasible points, treat the evidence as weak:
leave the space open or shrink at most two parameters by at
most 25% each.  A Python validator and safety gate will
reject violations, so return a corrected policy rather than
relying on repair.
```

Reproduction scope. Evaluation trajectories, states, decisions, gate events and analysis scripts support reconstruction of the archived summaries. The grouped training split measures within-topology prediction; all six evaluation topologies are represented in training. An independent row-level overlap audit is not supplied, so the results are reported as search on represented topologies. Online replay additionally depends on the provider runtime: the configured identifier is retained, while an immutable served-model snapshot and raw completion cache are not part of these artifacts. The offline rule and frozen surrogate support reruns with the specified simulation environment and seeds.

D Additional experimental details

D.1 Evaluation protocol and configuration mapping

Tasks and simulation. The component study uses the medium tier of the six Amplifier topologies; the released-code comparison spans all three tiers. Targets are favourable-direction distribution quantiles with levels 0.2, 0.35 and 0.5 for the simple, medium and hard tiers, respectively (Appendix A). An AC/DC `ngspice` simulation uses the sky130 PDK at the `tt` corner with a $300\,\mathrm{s}$ timeout. A hit must satisfy Power, `dcgain`, GBW, phase margin, CMRR, positive and negative PSRR, and offset constraints simultaneously. Timeouts and non-convergence consume one evaluation and count as failures.

Budgets and endpoints. The primary matrix contains ten harness arms representing nine configurations. Each is run on six topologies with seeds $\{7, 17, 29, 43, 73\}$, for 300 runs of 100 search evaluations. The published-agent comparison of Table 1 is a separate matrix with four arms (SMS-M and the three released implementations) on the same six topologies with seeds $\{7, 17, 29\}$ and all three difficulty tiers, for 216 focal runs of 100 evaluations. The source directories also contain 54 SMS-D runs, bringing the full audit to 270; all reached budget. The superseded proxy matrix contains five arms, including the excluded proxy described in Appendix D.16, for 150 runs of 200 evaluations. Feasible-hit rate divides feasible evaluations, including repeats, by the corresponding search budget. Unique feasible count deduplicates parameter vectors. First-feasible index is $N + 1$ when no hit is found. The primary matrix adds 15 confirmation evaluations after search; the released-code and historical proxy matrices have no confirmation stage. Confirmation uses the same nominal simulation setting and is distinct from process-corner or fabrication robustness testing.

Initialization in the archived runs. The benchmark supplies an initial verified feasible design that is evaluated as call 1 in every matched run across all difficulty tiers and charged against N. All 30 runs of every SMS-SIZING arm and every component ablation begin with it. Sobol/LHS re-evaluates the same design at indices 1, 26, 51 and 76, which accounts for its four feasible evaluations and single distinct feasible vector per run. The archived executions of random search and surrogate-enhanced BO predate this convention and did not receive it; both were re-executed with the initialization candidate injected as evaluation 1, and Table 8 lists both executions. Because the initialization candidate is shared, the main table also reports every endpoint with all evaluations of that design deleted, identified by parameter vector rather than by the runner's origin tag (`analysis/make_mechanism_numbers.py`). The two conventions bracket the comparison: SMS-M moves from 27.17% to 26.43%, while every selection-only arm moves to exactly zero.

Supplementary comparisons. The supplementary controls include `AnchorOnly`, `RankOnly`, `RandomPick`, matched-start random search and matched-start surrogate-enhanced BO.

AnchorOnly covers the same 30 topology–seed cells as SMS-M. RankOnly selects simulator batches by surrogate ranking with contraction and anchored proposals disabled; RandomPick draws uniformly from the same scored candidate pool. Initialization is identified by parameter-vector identity.

Configurations. Table 7 maps manuscript names to archived identifiers. SMS-M disables the BO-style selector and sets local ratio 0.15, active fraction 0.05, and at most eight anchors. SMS-B retains these local settings and enables the selector. SMS-B-L and SMS-B-S are two archived harness arms that share that selector and differ only in their recorded identifier; both are reported wherever either is used. The four original component ablations enable the historical selector in their archived states, so their matched selector reference is SMS-B, not SMS-M. Every ablation row of Table 8 must be read against SMS-B's 18.7 new designs rather than SMS-M's 26.2, because the selector change and the ablated component would otherwise be confounded. NoDirection removes anchored proposals while allowing box contraction. NoDimReduction removes shrink suggestions and the anchored stream and frees all coordinates. NoBoundary reallocates the boundary quota to the global quota while preserving the representative-retention rule. NoMemory removes recalled measurements from analysis while preserving the anchor path.

Table 7: Mapping between manuscript configurations and recorded implementation names.

Configuration	Recorded identifier	Final selector
SMS-M	Ours_NoBO Ours_SensitivityManifold (summary)	Quota and margin
SMS-B-L	LLMOrchestrated	Historical pool score
SMS-B-S	SensitivityFeasibleBO	Historical pool score
AnchorOnly	AnchorOnly	Quota and margin
Four component ablations	Named by the disabled component	Historical pool score

Pairing and statistical unit. Comparisons pair the same topology and seed. Five seeds share a circuit, so the unit for inference across circuits is the topology. Reported intervals first take the median paired difference over five seeds within each topology. They resample these six topology medians 10,000 times and compute their mean in each resample. Percentiles 2.5 and 97.5 form the interval. The topology-level Wilcoxon test uses the per-topology median paired difference and Holm correction across the executed comparisons. Detailed statistics are given in Appendix D.15.

D.2 Complete primary-matrix outcomes

What the selection-only arms discovered. Every feasible evaluation recorded by the seven selection-only arms is the benchmark default design. This was checked by comparing parameter vectors against dataset/Amplifier/topologies/*/parameters/default.json rather than by trusting the origin tags, and it holds for all 254 feasible evaluations across the 191 completed runs, including all 31 recorded by the original surrogate-enhanced BO executions, whose mean first-hit index of 74.0 reflects how long its pool ranking took to re-select the design the metadata already contained. Their "unique feasible" counts in Table 8 are therefore counts of that one design.

Table 8: Primary matrix and supplementary arms, $N = 100$. Values are means and standard deviations over the completed runs. First hit is the first-feasible evaluation index. Confirmation is the feasible count among 15 post-search evaluations; n.a. means these confirmations were unavailable for the supplementary arms. Rows marked "matched start" receive the shared default as evaluation 1; the two rows above them are the original executions that did not.

Method	Runs	Feasible hits (%)	Unique feasible	First hit	Confirmation
SMS-M	30	27.17 ± 10.26	27.2 ± 10.3	1.0	14.5
NoMemory	30	19.97 ± 8.04	20.0 ± 8.0	1.0	6.2
SMS-B-L	30	19.73 ± 7.57	19.7 ± 7.6	1.0	13.5
SMS-B-S	30	19.27 ± 8.04	19.3 ± 8.0	1.0	13.3
NoBoundary	30	14.63 ± 6.85	14.6 ± 6.8	1.0	11.7
NoDimReduction	30	1.67 ± 1.63	1.7 ± 1.6	1.0	1.7
NoDirection	30	1.60 ± 1.40	1.6 ± 1.4	1.0	1.6
SobolLHS	30	4.00 ± 0.00	1.0 ± 0.0	1.0	1.0
SurrogateEnhancedBO, original	30	1.03 ± 1.59	0.3 ± 0.5	74.0	0.3
RandomSearch, original	30	0.00 ± 0.00	0.0 ± 0.0	101.0	0.0
SurrogateEnhancedBO, matched start	19	2.16 ± 1.80	1.0 ± 0.0	1.0	n.a.
RandomSearch, matched start	30	1.00 ± 0.00	1.0 ± 0.0	1.0	n.a.
RankOnly	22	1.23 ± 1.63	0.4 ± 0.5	67.0	n.a.
RandomPick	30	0.17 ± 0.38	0.2 ± 0.4	96.5	n.a.

D.3 Historical data provenance

The $N = 200$ proxy matrix uses SMS-D and paper-inspired local controllers. Its trajectories remain in the project archive. The current published-agent comparison uses adapted released implementations at $N = 100$, with provenance in Appendix D.16; proxy results are not used for those comparative claims.

D.4 Per-topology outcomes

SMS-M exceeds SMS-B and the reported search controls on each topology (Table 9). The rate varies from 15.4% to 38.0%. Across these six circuits, its Spearman rank correlation with parameter dimension is $\rho = 0.086$ ($p = 0.872$); the reported comparison is stratified by topology. Sobol/LHS obtains one distinct feasible vector and four feasible evaluations per run on every topology.

Table 9: Mean feasible-hit rate (%) over five seeds. BO denotes SurrogateEnhancedBO, NoDim denotes `NoDimReduction`, and SMS-B-L is the primary execution. BO and Random are the original executions without injected initialization; matched-start controls are listed in Table 8.

Topology	d	SMS-M	SMS-B	NoDim	Sobol/LHS	BO	Random
Fan_SMC	23	**20.6**	14.8	4.8	4.0	0.0	0.0
Alfio_RAFFC	27	**38.0**	27.4	1.0	4.0	2.2	0.0
Sau_CFCC	29	**15.4**	10.0	1.0	4.0	0.0	0.0
Leung_DFCFC1	30	**36.0**	26.8	1.0	4.0	0.0	0.0
Peng_IAC	31	**31.6**	21.4	1.2	4.0	0.0	0.0
Qu2017_AZC	38	**21.4**	18.0	1.0	4.0	4.0	0.0

D.5 Budget partition by candidate stream

Every evaluation record stores the stream that produced its candidate, so the simulator budget partitions exactly. Table 10 gives the partition over the whole run and Table 11 over the end game alone; both are recomputed from the per-evaluation records by `analysis/make_mechanism_numbers.py`. The pooled row of Table 10 covers the seven SMS-SIZING-family arms of the primary matrix, which are the arms that tag candidate origin; Sobol/LHS, random search and surrogate-enhanced BO do not, and their evaluations are excluded from it.

Table 10: Budget partition by candidate stream, pooled over 30 runs per arm (first 100 evaluations). "Anchored" is a sparse perturbation of a measured anchor, "box fill" is Latin-hypercube sampling of the current box, "uniform" is a draw from the original box. Entries are feasible/evaluated; the last column is the anchored stream's share of feasible hits. Here, n.a. indicates that no anchored proposals were made, so the anchored share is undefined.

Arm	Anchored	Box fill	Uniform	Anchored share (%)
SMS-M	774/1302 (59.4%)	11/1052	0/616	95.0
NoMemory	559/1011 (55.3%)	10/1335	0/624	93.3
SMS-B-L	555/1008 (55.1%)	7/1368	0/594	93.8
SMS-B-S	539/1014 (53.2%)	9/1344	0/612	93.3
NoBoundary	405/844 (48.0%)	4/1045	0/1081	92.3
NoDimReduction	n.a.	20/2299	0/671	n.a.
NoDirection	n.a.	18/2328	0/642	n.a.
Pooled	**2832/5179 (54.7%)**	79/10771 (0.73%)	**0/4840**	

Table 11: End game (evaluations 86–100), pooled over 30 runs per arm: 450 evaluations per arm. The final rule decides how much of the budget reaches the anchored stream and how well it is ordered.

Arm	Anchored share (%)	Anchored hit (%)	Box-fill hit (%)	Phase yield (%)
SMS-M	77.8	66.0	18.2	**52.2**
SMS-B-S	17.8	13.8	0.3	2.7
NoMemory	16.0	11.1	0.3	2.0
SMS-B-L	11.6	9.6	0.0	1.1
NoBoundary	18.4	1.2	0.3	0.4

D.6 PHASE-RESOLVED ABLATIONS

Table 12 reports feasible evaluations divided by the evaluations within each phase, pooled over 30 runs. The phase lengths are 45, 40 and 15, so a phase-length-weighted average recovers the aggregate hit rate up to rounding. The warm-up column is similar across the anchored arms, because the first batch is generated before the process initialization has been measured and consists of the process initialization, box fill and global sentinels; the arms separate exactly when the anchored stream switches on, and separate again when the final rule decides how much budget reaches it (Table 11). Table 12 lists the values and Figure 7 plots them.

Table 12: Within-phase feasible-hit rates (%) for the $N = 100$ matrix. Local exploration is called `manifold` in the run summaries. Δ is the aggregate drop from SMS-M in percentage points.

Method	Warm-up	Local exploration	End game	Δ (pp)
SMS-M	2.7	**45.3**	**52.2**	0.00
SMS-B-L	3.6	44.9	1.1	7.43
SMS-B-S	2.9	43.9	2.7	7.90
NoDirection	2.7	0.9	0.0	25.57
NoDimReduction	2.7	1.1	0.2	25.50
NoBoundary	2.4	33.7	0.4	12.53
NoMemory	2.7	46.1	2.0	7.20
SobolLHS	4.4	5.0	0.0	23.17
SurrogateEnhancedBO	0.9	1.2	1.1	26.13
RandomSearch	0.0	0.0	0.0	27.17

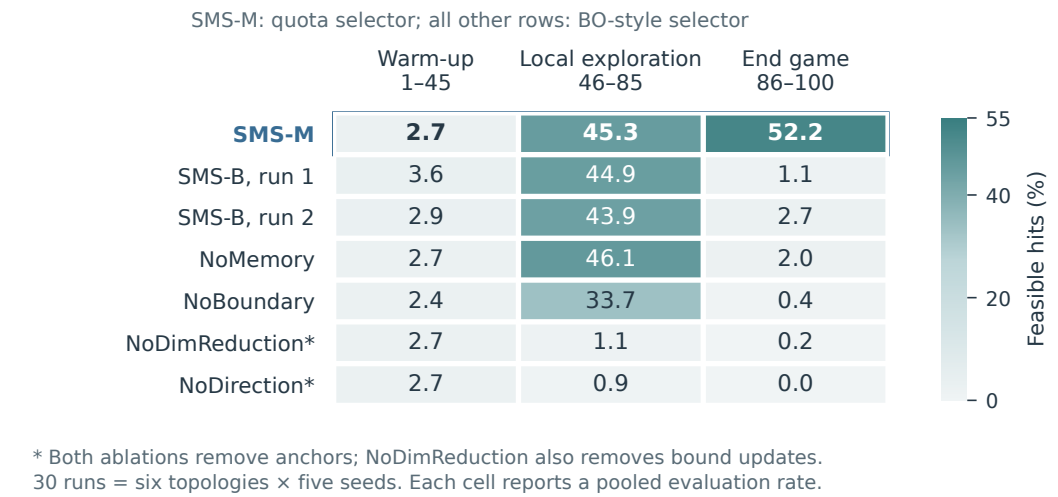


Figure 7: **Ablations separate during local exploration and the end game.** Feasible-hit rate pooled over 30 runs within warm-up (evaluations 1–45), local exploration (46–85) and end game (86–100). The two ablations that disable anchored proposals hold near 1% throughout local exploration. The end-game rule separates SMS-M from the two SMS-B executions; the other ablations also retain the BO-style selector, and NoDimReduction additionally removes anchors. Full values appear in Table 12.

D.7 Separation of the returned designs

Table 13 quantifies what the distinct-design count of Eq. 3 contains. Each coordinate is mapped to $[0, 1]$ by its legal range and Euclidean distances are divided by $\sqrt{d}$, so that the root mean square distance between independent uniform draws is $1/\sqrt{6} = 0.408$ and the numbers below are directly comparable across topologies of different dimension. The worst-case normalized margin is $\min_k m_k(x)$ with $m_k(x) = -c_k(x)/|t_k|$; the process initialization holds 0.34%. `NoDimReduction` and the selection-only arms are absent because they return too few distinct designs per run for a within-run spread to be defined.

Table 13: Separation and constraint margin of the distinct feasible designs returned per run. Distances are normalized as described above; the uniform root mean square reference is 0.408. "Moved" is the fraction of coordinates whose value spans more than 5% of its legal range across the returned set. Margins are worst-case normalized margins, in percent.

		Separation			Worst-case margin (%)		
Arm	Runs	pairwise	to initialization	moved	median	best	$>$ initialization (%)
SMS-M	30	0.0477	0.0706	0.60	1.26	2.55	76.1
NoMemory	30	0.0575	0.0758	0.54	1.21	2.10	69.7
SMS-B-L	30	0.0454	0.0708	0.50	1.22	2.13	68.8
NoBoundary	30	0.0310	0.0489	0.43	1.08	1.97	63.3

SMS-M combines the largest best worst-case margin and the largest fraction of returned designs improving the initialization margin in this component table. Pairwise spread is of the same scale across the anchored arms (SMS-M 0.0477; NoMemory 0.0575; SMS-B-L 0.0454), while NoBoundary's 0.0310 reflects its different quota allocation. These descriptive statistics define the nominal archive endpoints; corner and layout evaluation form the next validation stage.

D.8 The surrogate-free anchored control

`AnchorOnly` compares the anchored sparse-perturbation sampler in a surrogate-free system setting. It retains SMS-M's local settings exactly: 0.15 local pool fraction, 0.05 active-coordinate

fraction, eight-anchor cap, 0.10 radius, the quota-and-margin final selector, the same $45/40/15$ schedule, the same simulator, budget, seeds and default initialization. It removes all surrogate predictions, and with them the sensitivity ranking of Equation 5, the predicted feasibility masks, and the predicted-margin ordering of Equation 7; the served controller; the gate's bound decision; and all box contraction, holding the box at the full legal range for the whole run.

The configurations cover the same 30 topology–seed cells as the component cohort.

Table 14: **What the anchored sampler supplies, and what the rest of the loop adds.** `Medium` tier, $N = 100$, on the 30 topology–seed cells completed in both arms. `AnchorOnly` keeps only the anchored sparse-perturbation sampler at SMS-M's local settings, with no surrogate, no served controller, no gate decision and no box contraction. New designs exclude the shared default by parameter identity. Reach and pairwise separation are normalised distances divided by $\sqrt{d}$; margin is each run's best worst-case normalised margin.

Configuration	Runs	Hits (%)	New	Reach	Pairwise	Margin
`AnchorOnly`	30	27.60 ± 5.13	26.6	0.0146	0.0203	0.0106
SMS-M	30	27.17 ± 10.26	26.2	0.0706	0.0477	0.0255

The two arms return similar mean counts: 26.6 new feasible designs for `AnchorOnly` and 26.2 for SMS-M, with feasible-hit rates of 27.60% and 27.17%, respectively. Their archive geometry differs: median reach is 0.0146 versus 0.0706, pairwise separation is 0.0203 versus 0.0477, and best worst-case margin is 0.0106 versus 0.0255. The count, reach, separation and margin endpoints summarize the returned archives.

D.9 Ranking diagnostics

Split-half agreement tests rank stability within a sampled surrogate pool; cross-seed agreement tests stability across five independent search executions, which are exposed to different pools and different accepted contractions (Table 15). These measurements test repeatability of the pool statistic. The local response probe below measures a different property, agreement with simulation. The comparison of cost is a comparison of two diagnostic procedures: central finite differences for one scalar response at one anchor consume $2d = 46$–76 simulator evaluations, while the pool statistic reuses predictions the round has already computed.

Table 15: Reported rank repeatability. Cross-seed values average pairwise rank correlations among the five seeds; split-half values compare rankings from two halves of a pool.

Topology	Split-half Spearman ρ	Cross-seed Spearman ρ
Fan_SMC	0.976	0.725
Alfio_RAFFC	0.977	0.983
Sau_CFCC	0.931	0.957
Leung_DFCFC1	0.984	0.971
Peng_IAC	0.964	0.969
Qu2017_AZC	0.954	0.981

A separate probe perturbs each of the 27 Alfio_RAFFC parameters by $+20\%$ around one discovered feasible design and evaluates the same perturbed points with the single-model `core` surrogate and with `ngspice`. The two agree on which coordinates respond: zero change for the same 15 coordinates and nonzero change for the same 12. They rank the responding coordinates consistently at Spearman $\rho = 0.937$, with effect magnitudes within a factor of 0.79–2.57 (Table 16). Five of the 15 silent coordinates sit at a bound, where a forward-only step is absorbed by clipping or by discrete parameter handling, so the probe measures agreement on this design rather than physical relevance in general. It tests local response prediction and is independent of the pool-correlation ranking of Eq. 5.

A descriptive aggregation by device role places the bias-current source at mean rank 1.7 and compensation capacitor at 7.7, followed by gain-stage parameters at 12.3–21.1 and output-load parameters

Table 16: Local response-change diagnostic on Alfio_RAFFC, using a single model, one anchor and forward 20% perturbations. Zero/nonzero refers to this probe.

Diagnostic	Value
Coordinates with zero change in both responses	15 of 27
Coordinates with nonzero change in both responses	12 of 27
Response-change Spearman rank agreement on nonzero subset	$\rho = 0.937$
Surrogate/measured effect magnitude ratio	0.79–2.57

at 18.4. The prominence of bias and compensation is consistent with their influence on operating point and frequency response. The precise ordering also depends on parameterization, bounds and the sampled operating regime.

D.10 Ranking robustness

Two retrospective checks evaluate the ranking's dependence on correlation statistic and output set. Both use the archived round-0 candidate pools of 30 SMS-M medium runs, with 5,000 recorded designs per run before any contraction. They recompute orderings without executing new search trajectories.

Pearson versus Spearman. Recomputing the ranking with Spearman rank correlation in place of Pearson gives a mean rank correlation of 0.980 against the reported ordering (minimum 0.962 across the 30 runs), a mean top-10 overlap of 93.7%, and the same top-ranked coordinate in every run. This quantifies agreement of the two orderings on the recorded pools.

Hard constraints alone versus the full 11-metric average. Recomputing the ranking from the 8 hard-constraint metrics only, excluding the three unconstrained objectives (`Power`, `dcgain`, `GBW` are themselves constrained; the excluded metrics are `sr`, `TC`, `setting_time`) gives a mean rank correlation of 0.969 against the as-reported ranking (minimum 0.925), a mean top-10 overlap of 87.7%, and again an unchanged top-ranked coordinate in every run. The unchanged leading coordinate and top-ten overlap quantify robustness to output aggregation; they do not measure the yield of a re-executed policy.

Directing the perturbed coordinate. Section 3.3 notes that the sensitivity ranking governs the box, not which coordinate a local candidate perturbs, which is chosen uniformly at random among the $q = \max(1, \mathrm{round}(0.05d))$ active slots. `SensitivityDirectedLocal` biases that choice toward higher-ranked coordinates instead, weighting the draw by the previous round's `parameter_effects`, with every other setting matched to SMS-B. Over 30 runs on the `medium` tier it reaches $19.63\% \pm 7.22$ against SMS-B's 19.73% (15 wins, 15 losses of 30 paired cells, median topology difference -1.0 points, $p = 0.34$), so the measured yields are similar under the two coordinate-selection rules. This experiment isolates that sampling choice while keeping the SMS-B update and selector configuration fixed.

D.11 Gate and contraction controls

`NoGate` computes the diagnostic verdict but commits controller bounds without enforcing it. `NoContraction` keeps the original box throughout. Each covers six topologies and five seeds at $N = 100$.

The final box ratios span more than three orders of magnitude, while the measured reach is 0.0692 for `NoGate` and 0.0712 for `NoContraction`. Table 2 reports the archive geometry alongside the other configurations.

Table 17: Update-control outcomes. Hit rate is mean ± standard deviation across 30 medium-tier runs; reach and pairwise separation are within-run median normalized distances over new feasible designs; margin is the mean run-wise best worst-case normalized margin. Volume is the median final/original mixed-domain box ratio.

Configuration	Hits (%)	Reach	Pairwise	Margin	Final box ratio
NoGate	19.70 ± 7.65	0.0692	0.0527	0.0207	0.000701
NoContraction	19.97 ± 7.91	0.0712	0.0378	0.0216	1.00000

D.12 Contraction summaries and action attribution

Table 18 summarizes the contraction records over five seeds per topology. Retained widths come from the parameter-level records pooled across seeds; the space factor is a geometric mean of per-seed search-space factors, which multiply continuous width ratios by ratios of counts of legal integer values and are therefore distinct from both a product of pooled widths and the sum of continuous log-width reductions used for rank attribution. Figure 8(a) and the Top 5/Top 10 width groups order coordinates by the final ranking of seed 73; the correlation column averages per-seed Pearson correlations between each seed's own ranks and widths, and the last column pools log-width reductions using each seed's own top-ten set. Rank labels use one-based ordinals; the stored records index from zero.

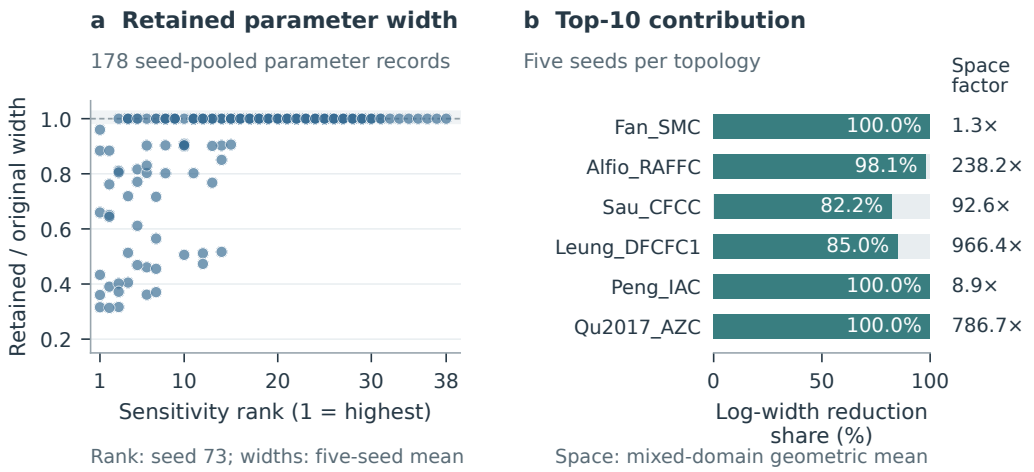


Figure 8: **Contraction follows the sensitivity ranking.** (a) Retained widths averaged over five seeds, ordered by the final sensitivity ranking of seed 73; the highest-ranked coordinates are the ones that narrow. (b) Top-ten share of summed log-width reduction, using each seed's own ranks. Space factors are geometric means of mixed continuous/discrete search-space reductions. The panels describe committed contractions, which arrive through the controller channel (Table 19).

The two bound-proposal channels do not contribute equally. Joining each round's final accepted gate event to that round's automatic shrink suggestions across the 30 `LLMOrchestrated` runs, and crediting a committed shrink to the analyzer when its parameter appears in the suggestion list, assigns every committed shrink to the controller channel (Table 19): the analyzer's automatic rule emits a nonempty suggestion in only 6 of 124 rounds, because its eligibility test requires a *low*-effect coordinate ($a_i \leq 0.05$) and a sufficient count of conservatively feasible samples. Contraction in practice therefore comes from the controller acting on the ranking, and it concentrates where the ranking is highest: ranks 1–5 carry 57.7% of the committed log-volume reduction and ranks 6–10 a further 28.8%. The offline rule supplies an alternative implementation of this channel.

Table 18: Five-seed contraction summaries. Contracted counts the union of coordinates narrowed across seeds. Retained width is final/original width; r averages per-seed rank–width Pearson correlations, omitting seeds with constant widths. Space factors are geometric means. The final column attributes summed log-width reduction to the top ten ranks.

			Retained width					
Topology	d	r	Top 5	Top 10	Rest	Contracted	Space	Share (%)
Fan_SMC	23	.489	.954	.977	1.000	2	1.3×	100.0
Alfio_RAFFC	27	.686	.530	.647	1.000	8	238×	98.1
Sau_CFCC	29	.520	.695	.770	.957	11	93×	82.2
Leung_DFCFC1	30	.575	.688	.707	.935	10	966×	85.0
Peng_IAC	31	.587	.730	.836	1.000	8	8.9×	100.0
Qu2017_AZC	38	.650	.457	.619	.991	10	787×	100.0

Table 19: Action attribution for the 30 `LLMOrchestrated` runs. Rank bands use one-based ordinals; n.a. denotes rounds without a committed shrink.

Accepted shrink source	Rounds	Share of shrink rounds (%)
Analyzer only	0	0.0
Controller only	37	100.0
Both	0	0.0
No shrink committed	87	n.a.
Rank band	**Parameter shrinks**	**Log-volume share (%)**
Ranks 1–5	98	57.7
Ranks 6–10	48	28.8
Remaining ranks	22	13.5

D.13 FAN_SMC: RETENTION CHECKS ON DISPERSED REPRESENTATIVES

Fan_SMC illustrates how retention checks respond to dispersed empirical components. Cross-seed rank correlation is 0.725, its search-space reduction is $1.3\times$, and three of five seeds commit no bound change. All five seeds record representative-retention conflicts, with four to seven empirical components named (Table 20). SMS-M still reaches 20.6% hits on this circuit, compared with 4.8% for NoDimReduction. The gate constrains updates based on the sampled representatives while anchored proposals continue to explore. These components are defined by finite-pool masks and should not be interpreted as an enumeration of the simulator's feasible set.

Table 20: Reported gate conflicts with retaining estimated feasible-region representatives. Counts refer to seeds with at least one such rejection.

Topology	Seeds with island-related rejection	Distinct islands named
Fan_SMC	5 of 5	4–7 per affected seed
Peng_IAC	1 of 5	1
Each of the remaining four	0 of 5	0

D.14 TARGET TIERS AND BUDGET SCALING

Targets use favourable-direction distribution quantiles (Appendix A). The `simple`, `medium` and `hard` tiers use quantile levels 0.2, 0.35 and 0.5, holding topology, simulator and process corner fixed. Table 1 reports the three-tier comparison at $N = 100$.

Budget compounds where anchors exist. At `medium` with the budget schedule scaled to $N = 200$, SMS-B reaches $32.53 \pm 8.89\%$ and returns 64.1 distinct discovered designs per run, against $19.27 \pm 8.04\%$ and 18.3 at $N = 100$: more than three times the designs for twice the budget. Because each discovered feasible design becomes an anchor for subsequent rounds, the higher-budget result

is consistent with continued archive expansion. This compares two budget settings of the same configuration and is not a continuation of one run: the phase schedule scales with N, so the first 100 evaluations of an $N = 200$ run (30.43%) are not the $N = 100$ arm.

D.15 TOPOLOGY-LEVEL INFERENCE

Five seeds share a circuit and are therefore not independent; the topology is the unit of inference. Table 21 reports, for each comparison, the paired win count over the 30 topology–seed cells, the median over the six topologies of the within-topology median paired difference, and a cluster bootstrap interval. The interval resamples the six topology medians $10{,}000$ times and takes the mean of each resample, so it estimates the *mean* of the topology medians while the effect column reports their *median*; the two columns are different summaries of the same six numbers, and neither equals the all-run mean difference of Table 8.

The primary comparisons have positive within-topology median differences on all six circuits. The exact two-sided signed-rank value for unanimous signs at $n = 6$ is 0.03125; Holm adjustment over nine comparisons gives 0.28125. The corrected value is descriptive directional evidence rather than a claim of statistical significance. Effect sizes, topology-wise directions and cluster intervals are therefore reported together with the test specification.

Table 21: Paired comparisons at the topology level. The effect is the median of per-topology median paired differences. CIs are cluster-bootstrap intervals for the mean of those topology medians, in percentage points. Wins count the 30 topology–seed cells.

Comparator	Wins	Median effect (pp)	Cluster 95% CI (pp)
Primary matrix: SMS-M minus comparator			
SMS-B-S	28/30	7.0	[6.7, 9.3]
SMS-B-L	28/30	8.0	[5.3, 9.3]
NoMemory	28/30	9.0	[5.2, 10.0]
NoBoundary	30/30	13.0	[10.3, 14.8]
SobolLHS	30/30	24.5	[16.7, 30.5]
NoDimReduction	30/30	25.5	[19.0, 33.2]
NoDirection	30/30	25.5	[19.0, 33.2]
RandomSearch	30/30	28.5	[20.7, 34.5]
SurrogateEnhancedBO	30/30	28.5	[19.5, 33.7]

D.16 WHAT THE BASELINE IMPLEMENTATIONS EXECUTE

Each baseline delegates search to released source adapted to the common benchmark interface. The provenance record identifies upstream repositories, commits and checksums for copied source files. EEsizer's notebook is extracted by a recorded script; LEDRO uses a kernel port. The adaptations below define the implementation-level comparison in Table 1.

Table 22: Adaptation matrix for the released-code comparison. Full file-level provenance is recorded in `baselines/VENDOR.md`.

Method	Retained search components	Benchmark adaptation
AutoSizer	Grouping, reduction, optimizer choice and inner optimizers	Bijective parameter naming; pre-layout metrics; fixed budget
EEsizer	Notebook prompts, three-call cycle, parser and history	Parameterized netlist bridge; 100-call cap; stop requests logged
LEDRO	TuRBO kernel, prompt frame and reward	ngspice callback; two levels; integer snapping; failed-call reward

Logged execution. Table 23 separates simulator completion from API requests and netlist parsing. In these records, a successful API response can still require a parser fallback; EEsizer logs 53

Table 23: Execution diagnostics from 54 runs per released baseline across three tiers. All 5,400 calls are charged. OK counts completed simulations, not feasibility. Timeouts are explicitly labelled simulator errors; API errors count non-success request attempts. Calls is the median recorded LLM call count per run; n/r means a parser-failure counter is not recorded.

Method	Calls to sim.	OK	Timeouts	API errors	LLM calls	Parse errors
LEDRO	5400	5362	0	0	2	n/r
EEsizer	5400	5397	0	0	298	53
AutoSizer	5400	5396	0	0	8	n/r

such netlist parse failures. The parser count is not a simulator-failure or transport-error count. All methods use a 300 s simulation timeout; controller settings include temperature 0.1, a 180 s request timeout and token-limit escalation from 4,096 to 32,768. Appendix C specifies policy retry limits. Repeated defaults are identified by parameter-vector identity.

What is common to all arms. Every arm is scored by the same ngspice evaluation of the same eight constraints, receives the same full continuous box, spends the same budget, and reaches the same served model through one retry policy. No arm is handed a pre-reduced or pre-discretised space: reducing that box is the upstream method's own job, and how it chooses to do so is what this comparison measures. Candidates are clamped into the box and integer-typed device multipliers are snapped before simulation, without which the netlist renderer rejects the candidate and charges the arm a failed simulation for a schema error. Where an upstream loop would stop early because of a convergence test or a controller answering `stop`, the arm keeps spending its budget and records that the stop was requested, to enforce the common evaluation budget.

AutoSizer. Yu et al. (2026) is a two-level agent, and both levels run as released: circuit understanding, the transistor grouping its pipeline invokes above twenty variables, the search-space reduction and its parser, the per-iteration decision that selects an inner algorithm and a sample count, the second regeneration cycle, and execution of its inner optimizers. These include Latin hypercube sampling, genetic search, simulated annealing, Gaussian-process and Optuna Bayesian optimization, and adaptive and multi-start search. Its figure of merit is the one its own code generates from the constraint thresholds. Two adaptations are required. Its variable naming convention is $\langle W|L|m\rangle$_$\langle$transistor$\rangle$ and its grouping composes group members by string-composing that convention, so this benchmark's parameter names are translated into it and back through a bijective map recorded per run. Its layout and parasitic-extraction stages have no counterpart here, so it optimizes the pre-layout metrics this benchmark scores.

EEsizer. Liu & Chitnis (2026) ships only notebooks; its agent is extracted verbatim by a script in the package, which drops the notebook's own simulation layer and moves the notebook's top-level driver into a function that is never called, so that importing it performs no work. Its three-call reasoning cycle of observation, optimisation and sizing runs on its own prompt text, its reply parser, its five-entry result history and its five percent convergence tolerance; the reported feasibility never uses that tolerance, which is the agent's internal stopping test. Upstream rewrites netlist text, so the topology's real subcircuit is presented with its sizing as SPICE `.param` lines and the rewritten netlist is parsed back from them: the device lines keep their parameter placeholders, so matched devices stay coupled as upstream's own text editing would leave them, and each rewritten value maps back to exactly one parameter. Upstream evaluates one design per cycle, so its twenty-five iteration cap is replaced by the benchmark budget, and its rate-limit sleeps, which are not part of the algorithm, are dropped.

LEDRO. Kochar et al. (2025) targets Cadence Spectre and FinFET netlists, neither of which is available here, so this arm is a kernel port and is labelled as one. What runs is upstream's TuRBO-1 implementation with its own published settings, upstream's range-reduction prompt frame, and upstream's reward function unchanged. Its two prompt levels become two equal search levels, each preceded by a reduction call, and the first evaluations are the uniform random stage its driver runs before its first prompt so that the prompt has measured points to reason about. Upstream's optimizer calls its objective itself, so it executes on a worker thread and blocks inside that objective until a

simulation returns: its trust-region updates, restarts and Thompson sampling are its own rather than a re-expression of them. Its reward has no branch for a failed simulation, so the adapter supplies that case: a failure is scored at twice the worst reward seen so far. Two parts of its input cannot be supplied: its prompt asks the model to reason about transistor operating regions, which this benchmark does not extract, and its integer device multipliers are snapped after TuRBO proposes them in a continuous box.

Excluded proxies. Earlier paper-inspired proxy controllers remain in the repository with their original trajectories; the published-agent comparison uses only the released implementations above and excludes every proxy, including `PANDA_MOSTAR_proxy`, whose named upstream implementation could not be verified.

D.17 Retrospective surrogate-filter recall

For an upper-bounded metric, a pessimistic filter accepts when

$$\hat{g}_k(x) + \kappa\sigma_k(x) \le t_k \quad \text{for every upper-bounded constraint } k, \tag{11}$$

with $\hat{g}_k(x) - \kappa\sigma_k(x) \ge t_k$ for lower-bounded constraints, the ensemble spread σ_k supplying the margin. The evaluation set is the $5{,}052$ feasible records discovered across the primary and historical proxy matrices. Every rejected design in this set is a simulator-confirmed solution that the filter would have discarded. This rejection rate determines whether prediction can replace simulation for admission. It cannot, which is why admission is simulator-only by construction.

Table 24 sweeps κ from 0 to 3. Even at $\kappa = 0$, where the filter trusts the mean prediction with no safety margin at all, it discards a quarter of the confirmed feasible designs; at $\kappa = 1$ it discards 54.0%.

Table 24: Retrospective retention among $5{,}052$ simulator-confirmed feasible records.

Margin κ	0.00	0.25	0.50	1.00	2.00	3.00
Retained	3763	3484	3093	2323	1239	713
Retention (%)	74.5	69.0	61.2	46.0	24.5	14.1

At $\kappa = 1$ the joint filter retains 46.0%, while the single strictest constraint, phase margin, retains 66.1% on its own (Table 25). Joint filtering compounds exclusions: no single metric accounts for more than a third of the loss, and the remaining constraints compound it. Because the constraints share circuit parameters, their rejection masks overlap, so the product of the marginal retentions is not the joint rate. This common-κ filter is also stricter than the asymmetric margins the analyzer applies in Eq. 7, which weight GBW, phase margin and offset at full spread and the rest at a quarter.

Table 25: Retention when each constraint is applied separately at $\kappa = 1$. PM denotes phase margin.

Constraint	Power	Gain	GBW	PM	CMRR	PSRP	PSRN	Offset
Retention (%)	88.8	95.3	77.5	66.1	88.6	96.4	94.6	95.1

E Extended related work

Local constrained search. Constrained Bayesian optimization models expensive feasibility and objectives (Gardner et al., 2014; Gelbart et al., 2014). TuRBO and SCBO use adaptive regions and partial-coordinate proposals (Eriksson et al., 2019; Eriksson & Poloczek, 2021). SMS combines related local sampling ideas with multiple measured anchors and an archive endpoint. Its historical-selector ablation isolates one pool-ordering rule; it is not an implementation of a complete trust-region Bayesian optimizer.

Region adaptation for sizing. LEDRO uses language-model range reduction with TuRBO, and TopoSizing updates regions from search feedback (Kochar et al., 2025; Wei et al., 2026). AutoSizer

Table 26: Task-level positioning. Entries describe the role of the returned solution set; these are conceptual comparisons, not additional benchmark results.

Method family	Search organization	Endpoint relevant to this comparison
TuRBO / SCBO	Adaptive local regions	Best objective under the specified constraints
LEDRO / TopoSizing	LLM-assisted region updates	Simulator-evaluated sizing performance
AutoSizer / EEsizer	Agent-directed sizing	Designs meeting requested specifications
MAP-Elites	Descriptor-indexed archive	Quality and behavioural coverage
SMS	Measured anchors and gated bounds	Feasible archive expansion from process-generated feasible points

and EEsizer organize simulation feedback through different agent interfaces (Yu et al., 2026; Liu & Chitnis, 2026). Our shared benchmark retains released search components with the adaptations in Appendix D.16. SMS emphasizes empirical retention of tracked representatives and the spatial reach of its accumulated feasible set.

Multi-solution search. Novelty search and MAP-Elites make solution collections explicit (Lehman & Stanley, 2011; Mouret & Clune, 2015; Cully et al., 2015). SMS admits designs using joint circuit feasibility and evaluates diversity in parameter space. It does not use a behavioural descriptor grid, and the reported experiments do not compare against a quality-diversity baseline. The common motivation is a collection that can support later selection under additional criteria.