ANCHORED SCENARIO COVERAGE FOR FAILURE-AWARE FIRST-HIT BATCH INVERSE DESIGN

Anonymous authors
Paper under double-blind review

ABSTRACT

Early discovery of at least one valid design satisfying a target requirement is a central objective in failure-prone closed-loop inverse design. A natural batch strategy ranks candidate designs by a product-form marginal valid-hit score, but selecting the highest-ranked candidates independently within a batch can produce redundant recommendations under predictive uncertainty. We introduce ARC-SC (Anchored Risk-Constrained Scenario Coverage), which preserves strong marginal candidates as anchors and allocates the remaining batch positions by maximizing complementary coverage over predictive target scenarios under a risk-support constraint. For fixed anchors and a fixed admissible pool, the scenario-coverage objective is monotone submodular and admits the standard greedy-completion guarantee. In frozen-oracle closed-loop simulations, ARC-SC yields a statistically supported improvement in first-hit discovery, while remaining competitive with directionally favorable performance on higher-dimensional tasks. These results support scenario-aware batch diversification as a principled complement to strong marginal acquisition for early valid-target discovery under experimental failure.

1 INTRODUCTION

Closed-loop inverse design iteratively proposes candidates, evaluates them through experiments or simulations, and uses the outcomes to guide subsequent decisions (Xue et al., 2016; Kusne et al., 2020; Wu et al., 2026). In many scientific settings the immediate objective is not the terminal optimum, but to discover, using as few experimental rounds as possible, at least one valid candidate satisfying a predefined target requirement. This first-hit objective is especially important when evaluations are costly, destructive, or performed in parallel batches.

A natural failure-aware batch strategy ranks candidate designs by a product-form marginal valid-hit score: the estimated probability of satisfying the target times the estimated probability of yielding a valid measurement. Selecting the top-B candidates by this score, which we call POF,[1] is appropriate when these scores are calibrated and candidate outcomes are effectively independent at the batch level.

In practice, the calibration and conditional-independence assumptions underlying this marginal argument can be unreliable in small-data closed-loop design. In Bayesian optimization, posterior uncertainty can become misleading under model misspecification, while reliable calibration is itself difficult in the small-sample regimes typical of sequential optimization (Neiswanger & Ramdas, 2021; Foldager et al., 2023). Moreover, batch Bayesian-optimization methods have long recognized that simultaneously selected evaluations interact and that independently maximizing marginal acquisition values can produce redundant batches, motivating local-penalization and diversity-aware selection strategies (González et al., 2016; Kathuria et al., 2016). Consequently, several candidates may receive high marginal scores while representing strongly overlapping posterior opportunities. Selecting all of them can produce a batch with strong individual scores but limited protection against surrogate error: if the common predictive explanation is wrong, the entire batch may miss the target. This is particularly consequential for first-hit discovery, where the value of a batch depends primarily on whether at least one member succeeds.

[1] Throughout this paper, POF denotes the implemented top-B marginal valid-hit baseline, not the conventional probability of feasibility in constrained Bayesian optimization.

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.

Related Work and Positioning. Batch Bayesian optimization has long recognized that parallel recommendations should not be obtained solely by repeating a sequential acquisition rule, but should account for interactions among batch members. Existing approaches model batch interactions through simulation matching, information-theoretic objectives, local penalization, and diversity-promoting constructions such as determinantal point processes (Azimi et al., 2010; Shah & Ghahramani, 2015; González et al., 2016; Kathuria et al., 2016; Wu & Frazier, 2016). These methods establish the value of batch complementarity, but are typically designed for conventional black-box optimization rather than valid-target first-hit discovery under experimental failure. Constrained and failure-aware Bayesian optimization instead learns or reasons about feasibility, using probabilities of constraint satisfaction, boundary exploration, pessimistic handling of failed observations, or stochastic feasibility models (Gardner et al., 2014; Hernández-Lobato et al., 2015; Bergmann & Graichen, 2020; Wakabayashi et al., 2022; Tian et al., 2024; Hickman et al., 2025; Iwazaki et al., 2025). These approaches improve candidate-level feasibility awareness, but do not directly ask whether simultaneously selected valid-hit opportunities are complementary.

The objective of discovering target members is also related to active search, which allocates queries to identify as many members of a valuable class as possible (Garnett et al., 2012; Jiang et al., 2017; Lookman et al., 2019), and to batch active-learning analyses based on adaptive submodularity (Chen & Krause, 2013). A close recent work studies diverse failure discovery in paired proxy–target autonomous systems using proxy-corrected risk estimates, support-aware mutual information, and geometric clustering (Parashar et al., 2026). ARC-SC addresses a different setting: first valid-target discovery under unknown experimental validity. Rather than covering the observed input geometry, it covers sampled predictive target scenarios, preserves strong marginal valid-hit candidates through explicit anchors, and imposes a batch-level validity/support risk constraint.

In summary, the resulting gap lies at the intersection of first-hit search, batch complementarity, and failure awareness. A suitable acquisition rule should preserve strong immediate opportunities, avoid spending an entire batch on overlapping predictive explanations, and control exposure to candidates that are likely invalid or weakly supported.

We address this problem with ARC-SC, an Anchored Risk-Constrained Scenario Coverage method for failure-aware first-hit batch inverse design. ARC-SC constructs each batch in two stages. It first selects risk-admissible anchors using the same marginal valid-hit criterion as POF, then fills the remaining batch positions by maximizing a noisy-OR coverage objective over predictive target scenarios. We prove that the fixed scenario-coverage objective is normalized, monotone, and submodular, giving the standard greedy-completion guarantee conditional on fixed anchors and a fixed admissible pool. In frozen-oracle simulations constructed from the UCI Superconductivity and JARVIS CFID materials-property datasets, ARC-SC yields more earlier first-hit outcomes than later ones relative to POF, with strongest gains under feature-interaction and low-validity failures.

Our contributions are:

1. We formulate failure-aware batch inverse design around the first valid target hit and propose ARC-SC, which combines marginal valid-hit anchoring, predictive scenario coverage, and batch-level risk-support control.
2. We prove that the scenario-coverage objective is normalized, monotone, and submodular, and derive the corresponding fixed-domain greedy-completion guarantee.
3. We evaluate ARC-SC on two frozen-oracle materials benchmarks under controlled failure mechanisms, showing more earlier first-hit wins than losses relative to ordinary batch POF, with the strongest gains under feature-interaction and low-validity failures.

2 METHODOLOGY

2.1 PROBLEM FORMULATION

We consider a closed-loop batch inverse-design problem in which experimental evaluations are sequentially performed to discover a design satisfying predefined target requirements. At iteration t, the optimizer has observed $\mathcal{D}_t = \{(x_i, y_i, v_i)\}_{i=1}^{n_t}$, where $x_i \in \mathcal{X}$ is a candidate design, $v_i \in \{0, 1\}$ indicates whether the evaluation returns a valid measurement, and y_i is observed only when $v_i = 1$. An invalid evaluation therefore provides a failure label but no usable target measurement.

Given batch size B, the optimizer selects $\mathcal{B}_t = \{x_{t,1}, \ldots, x_{t,B}\}$ and seeks to identify, as early as possible, at least one candidate satisfying both validity and target requirements, $v(x) = 1$ and $y(x) \geq T$:

$$\tau_{\text{hit}} = \min\{t : \exists x \in \mathcal{B}_t,\ v(x) = 1,\ y(x) \geq T\},$$

with $\tau_{\text{hit}} = \infty$ if no such candidate is found. Because the batch succeeds once any member is a valid target hit, ARC-SC optimizes batch-level complementarity rather than treating all selected candidates as independent marginal decisions.

2.2 Anchored Risk-Constrained Scenario Coverage

ARC-SC decomposes batch construction into two complementary stages, $\mathcal{B}_t = \mathcal{B}_{\text{anchor}} \cup \mathcal{B}_{\text{cover}}$. The anchor stage preserves candidates with strong marginal valid-hit potential. The completion stage then uses the remaining batch capacity to cover alternative plausible target opportunities under a risk-support budget. Intuitively, ARC-SC retains exploitation when the marginal model is reliable while hedging against the possibility that several top-ranked candidates rely on the same uncertain predictive explanation.

Algorithm 1 summarizes ARC-SC; a schematic workflow is provided in Appendix Figure 2.

2.3 Candidate Modeling and Risk Estimation

At each iteration, ARC-SC evaluates acquisition quantities on a finite candidate pool $\mathcal{C}_t = \{x_j\}_{j=1}^{N} \subseteq \mathcal{X}$. All candidate-level quantities and batch-selection decisions below are computed on $\mathcal{C}_t$; we write $\mathcal{C}$ when the iteration index is clear from context for notational simplicity. Candidate generation and model hyperparameters are given in Appendix A.1.

Target model. The target surrogate is trained on valid observations $\mathcal{D}_t^y = \{(x_i, y_i) : v_i = 1\}$ and provides the surrogate-estimated posterior probability that x satisfies the target threshold T,

$$p_y(x) = P(y(x) \geq T | x, \mathcal{D}_t). \tag{1}$$

Validity model. Failed experiments provide binary information about the unknown feasible region. ARC-SC trains a validity model on $\mathcal{D}_t^v = \{(x_i, v_i)\}_{i=1}^{n_t}$ and estimates

$$p_v(x) = P(v(x) = 1 | x, \mathcal{D}_t). \tag{2}$$

Here $p_v(x)$ is the predicted validity probability, corresponding to probability of feasibility in constrained Bayesian optimization.

Marginal valid-hit score and POF baseline. We define the product-form marginal valid-hit score

$$h(x) = p_y(x) p_v(x). \tag{3}$$

Under conditional independence between target satisfaction and validity, $h(x)$ is the corresponding joint posterior probability; otherwise it is used as a marginal acquisition score.

The POF baseline selects the B largest values of this score:

$$\mathcal{B}_{\text{POF}} = \text{TopB}_{x \in \mathcal{C}}\, h(x). \tag{4}$$

POF is therefore an independent marginal-ranking baseline with no explicit redundancy control among batch members.

Empirical support estimation. Validity prediction can be unreliable far outside the observed design distribution, especially in small-data regimes. ARC-SC therefore uses an empirical support score $p_s(x) = \exp(-d_{\text{kNN}}(x)^2/(2\tau_s^2))$, where $d_{\text{kNN}}(x)$ is the average distance to nearby observed designs in normalized feature space.

Risk-support score. A candidate has low risk when it is both predicted to be valid and sufficiently supported by the observed design distribution. We define the candidate-level risk-support score as

$$c(x) = 1 - p_v(x)p_s(x). \tag{5}$$

For a nonempty partial batch $\mathcal{S}$, ARC-SC evaluates the average batch risk:

$$C(\mathcal{S}) = \frac{1}{|\mathcal{S}|} \sum_{x \in \mathcal{S}} c(x).$$

ARC-SC enforces $C(\mathcal{S}) \leq R$ during selection whenever a risk-feasible extension exists, and R is a user-defined risk budget; otherwise it uses the fallback rule described in Appendix A.1.5.

2.4 POF-Anchored Candidate Selection

ARC-SC uses the same marginal valid-hit score $h(x)$ as the POF baseline to select its first K batch members, subject to the risk-admissible extension set

$$\mathcal{F}_R(\mathcal{S}) = \{x \in \mathcal{C} \setminus \mathcal{S} : C(\mathcal{S} \cup \{x\}) \leq R\}.$$

Starting from $\mathcal{S}_0^{\mathrm{a}} = \varnothing$, ARC-SC selects

$$a_k \in \arg\max_{x \in \mathcal{F}_R(\mathcal{S}_{k-1}^{\mathrm{a}})} h(x), \quad \mathcal{S}_k^{\mathrm{a}} = \mathcal{S}_{k-1}^{\mathrm{a}} \cup \{a_k\},$$

for $k = 1, \ldots, K$, whenever a feasible extension exists, and sets $\mathcal{B}_{\mathrm{anchor}} = \mathcal{S}_K^{\mathrm{a}}$.

Because $\mathcal{F}_R$ depends on the partial batch, the anchors need not be identical to the first K members of $\mathcal{B}_{\mathrm{POF}}$; they preserve the same exploitation criterion within the current risk-admissible set. This matters because, when the predictive model is reliable, marginal selection can already be strong; replacing it entirely with uncertainty-driven diversification could unnecessarily sacrifice immediate success probability. The anchors therefore preserve strong marginal opportunities while reserving only part of the batch for scenario-based diversification.

2.5 Predictive Scenario Construction

After selecting anchors, ARC-SC models uncertainty in the unknown target landscape through multiple predictive scenarios. Ideally, one would sample complete posterior functions $f^{(1)}, ..., f^{(M)} \sim P(f|\mathcal{D}_t)$. For computational scalability, the current implementation constructs marginal posterior-predictive scenarios on the finite candidate pool:

$$f^{(m)}(x_j) = \mu_j + \sigma_{y,j} z_{m,j}, \quad z_{m,j} \overset{\mathrm{iid}}{\sim} \mathcal{N}(0,1), \tag{6}$$

where $\mu_j = \mu(x_j)$ and $\sigma_{y,j} = \sigma_y(x_j)$ denote the predictive mean and standard deviation returned by the target surrogate for each $x_j \in \mathcal{C}$. These scenarios preserve the marginal predictive uncertainty at each candidate but do not explicitly retain cross-candidate posterior covariance.

For each scenario, ARC-SC defines a relaxed threshold and soft target response:

$$T_m^{\mathrm{relax}} = \min\left(T, Q_{1-\pi}\{f^{(m)}(x) : x \in \mathcal{C}\}\right), \tag{7}$$

$$\psi_m(x) = \mathrm{sigm}\left(\frac{f^{(m)}(x) - T_m^{\mathrm{relax}}}{\tau}\right), \tag{8}$$

where $Q_{1-\pi}$ is an empirical quantile over $\mathcal{C}$ and $\mathrm{sigm}(u) = 1/(1+\exp(-u))$. The relaxed threshold prevents degenerate all-low-response scenarios early in the search.

2.6 SCENARIO COVERAGE ACQUISITION

For each scenario m, ARC-SC defines a candidate response:

$$r_m(x) = g_h(x)g_v(x)g_s(x)\psi_m(x). \tag{9}$$

To reduce sensitivity to probability miscalibration, we transform the marginal valid-hit, validity, and support scores into percentile-rank gates. Let $R_h(x)$, $R_v(x)$, and $R_s(x)$ denote the corresponding percentile ranks within $\mathcal{C}$. The gates are

$$g_q(x) = [\epsilon_q + (1-\epsilon_q)R_q(x)]^{\alpha_q}, \quad q \in \{h, v, s\}. \tag{10}$$

The floor ϵ_q and exponent α_q reduce sensitivity to calibration error; parameter settings are in Appendix A.1.6.

For a candidate batch $\mathcal{B}$, ARC-SC defines the scenario coverage objective:

$$A_{\mathrm{SC}}(\mathcal{B}) = \frac{1}{M}\sum_{m=1}^{M}\left[1 - \prod_{x\in\mathcal{B}}(1 - r_m(x))\right]. \tag{11}$$

The noisy-OR term rewards a batch for containing at least one strong candidate in each scenario and discounts repeated coverage. The marginal gain of adding x to partial batch $\mathcal{S}$ is

$$\Delta_{\mathrm{SC}}(x|\mathcal{S}) = \frac{1}{M}\sum_{m=1}^{M} r_m(x)\prod_{s\in\mathcal{S}}(1 - r_m(s)). \tag{12}$$

Thus, a candidate is valuable when it responds strongly in scenarios not already covered by $\mathcal{S}$.

2.7 RISK-CONSTRAINED GREEDY BATCH OPTIMIZATION

Direct optimization of $A_{\mathrm{SC}}(\mathcal{B})$ is combinatorial. Therefore starting from $\mathcal{S} = \mathcal{B}_{\mathrm{anchor}}$, ARC-SC greedily fills the remaining positions by selecting

$$x^\star \in \arg\max_{x\in\mathcal{F}_R(\mathcal{S})} \Delta_{\mathrm{SC}}(x \mid \mathcal{S}).$$

The selected candidate is added to $\mathcal{S}$, and the process continues until $|\mathcal{S}| = B$; if no risk-feasible extension exists, ARC-SC uses the risk-penalized fallback in Appendix A.1.5. The complete procedure is summarized in Algorithm 1.

3 THEORETICAL ANALYSIS

We analyze the one-round batch-selection problem, conditioning on the observed data, candidate pool, predictive scenarios, rank gates, and responses. Thus, during the current batch construction, each $r_m(x)$ is fixed and satisfies $0 \le r_m(x) \le 1$.

3.1 FROM MARGINAL VALID-HIT SELECTION TO BATCH SCENARIO COVERAGE

Let $H_x = \{v(x) = 1\} \cap \{y(x) \ge T\}$ be the event that x is a valid target hit. The ideal current-round utility is

$$F(\mathcal{B}) = \Pr\left(\bigcup_{x\in\mathcal{B}} H_x \,\middle|\, \mathcal{D}_t\right). \tag{13}$$

If $h(x) = p_y(x)p_v(x)$ is a calibrated marginal probability of H_x and hit events are conditionally independent,

$$F(\mathcal{B}) = 1 - \prod_{x\in\mathcal{B}}\big(1 - h(x)\big). \tag{14}$$

Algorithm 1 ARC-SC at closed-loop iteration t

Require: Observed data $\mathcal{D}_t$, candidate pool $\mathcal{C}$, batch size B, anchors K, scenarios M, risk budget R
Ensure: Recommended batch $\mathcal{B}_t$
1: Fit target/validity models and compute $p_y, p_v, p_s, h, c, g_h, g_v, g_s$ on $\mathcal{C}$
2: $\mathcal{S} \leftarrow \varnothing$ ▷ Stage 1: marginal anchors
3: **for** $k = 1, \ldots, K$ **do**
4: Select the risk-admissible candidate maximizing $h(x)$ (fallback if none is feasible); add it to $\mathcal{S}$
5: **end for**
6: Generate M predictive scenarios and compute $r_m(x)$ for all m and $x \in \mathcal{C}$ ▷ Stage 2: scenario-coverage completion
7: **for** $\ell = 1, \ldots, B - K$ **do**
8: Select the risk-admissible candidate maximizing $\Delta_{\mathrm{SC}}(x \mid \mathcal{S})$ (fallback if none is feasible); add it to $\mathcal{S}$
9: **end for**
10: **return** $\mathcal{B}_t \leftarrow \mathcal{S}$

For fixed B, equation 14 is maximized by selecting the B largest $h(x)$ values, recovering POF.

ARC-SC targets the complementary regime, where high marginal scores may share the same predictive explanation and thus contribute redundant batch-level value. Because $r_m(x)$ includes a relaxed target response and rank-normalized gates, A_{SC} is not asserted to be a calibrated estimate of F; it is a bounded utility for covering multiple predictive scenarios.

3.2 Submodularity of Scenario Coverage

Proposition 1. *The set function*

$$A_{\mathrm{SC}} : 2^{\mathcal{C}} \to [0, 1]$$

defined in equation 11 is normalized, monotone, and submodular.

Proof. Normalization follows from $A_{\mathrm{SC}}(\varnothing) = 0$. For $\mathcal{S} \subseteq \mathcal{C}$ and $x \notin \mathcal{S}$, the marginal gain is

$$\Delta_{\mathrm{SC}}(x \mid \mathcal{S}) = A_{\mathrm{SC}}\big(\mathcal{S} \cup \{x\}\big) - A_{\mathrm{SC}}(\mathcal{S}) = \frac{1}{M} \sum_{m=1}^{M} r_m(x) \prod_{s \in \mathcal{S}} \big(1 - r_m(s)\big). \tag{15}$$

All factors in equation 15 are nonnegative, giving monotonicity.

For $\mathcal{S} \subseteq \mathcal{T} \subseteq \mathcal{C}$ and $x \notin \mathcal{T}$, since $1 - r_m(s) \in [0, 1]$,

$$\prod_{s \in \mathcal{S}} \big(1 - r_m(s)\big) \geq \prod_{s \in \mathcal{T}} \big(1 - r_m(s)\big).$$

Multiplying by $r_m(x) \geq 0$ and averaging over m gives

$$\Delta_{\mathrm{SC}}(x \mid \mathcal{S}) \geq \Delta_{\mathrm{SC}}(x \mid \mathcal{T}),$$

which is submodularity. □

Proposition 1 holds for any fixed response matrix with entries in $[0, 1]$; exact correlated posterior-function samples are not required for the submodular structure. Equation 15 also shows the redundancy discount: once a scenario is already well covered, the product term becomes small and similar candidates provide little additional gain.

3.3 Guarantee for Anchored Greedy Completion

Let $\mathcal{A} = \mathcal{B}_{\mathrm{anchor}}$ be fixed and let $k = B - K$. Conditional on $\mathcal{A}$, define

$$G_{\mathcal{A}}(\mathcal{S}) = A_{\mathrm{SC}}(\mathcal{A} \cup \mathcal{S}) - A_{\mathrm{SC}}(\mathcal{A}). \tag{16}$$

Table 1: Frozen-oracle benchmark environments.

DATASET	SAMPLES	FEATURES	THRESHOLD	ORACLE R^2	SPEARMAN ρ
Superconductivity	21,263	5	80.0	0.913	0.956
JARVIS CFID	55,723	15	1.539	0.838	0.912

Proposition 2. *For any fixed anchor set $\mathcal{A}$, the residual objective $G_{\mathcal{A}}$ is normalized, monotone, and submodular with respect to $\mathcal{S}$.*

Proof. Adding fixed anchors preserves monotonicity and diminishing returns; subtracting $A_{\mathrm{SC}}(\mathcal{A})$ establishes normalization. □

For a fixed admissible completion pool $\mathcal{F} \subseteq \mathcal{C} \setminus \mathcal{A}$ with $1 \leq k \leq |\mathcal{F}|$, let $\mathcal{S}_{\mathrm{greedy}}$ be the k-step greedy completion and let $\mathcal{S}^\star \in \arg\max_{\substack{\mathcal{S}\subseteq\mathcal{F} \\ |\mathcal{S}|\leq k}} G_{\mathcal{A}}(\mathcal{S})$ be the optimal completion over the same pool.

Corollary 1 (Greedy anchored completion)**.** *For fixed anchors and a fixed admissible completion pool,*

$$G_{\mathcal{A}}(\mathcal{S}_{\mathrm{greedy}}) \geq \left[1 - \left(1 - \frac{1}{k}\right)^k\right] G_{\mathcal{A}}(\mathcal{S}^\star) \geq \left(1 - \frac{1}{e}\right) G_{\mathcal{A}}(\mathcal{S}^\star). \tag{17}$$

Proof. By Proposition 2, $G_{\mathcal{A}}$ is normalized, monotone, and submodular, so the classical cardinality-constrained greedy guarantee applies (Nemhauser et al., 1978). Appendix A.2.1 gives the derivation. □

This guarantee concerns only fixed anchors and a fixed admissible completion pool. We need to note that the implemented ARC-SC procedure uses a dynamic average-risk constraint and a fallback when no risk-feasible extension exists, and Corollary 1 is not a global approximation guarantee for the full policy; see discussion in Appendix A.2.2.

4 Experiments and Results

4.1 Experimental Setup

Materials benchmarks and frozen oracles. We evaluate ARC-SC on UCI Superconductivity (Hamidieh, 2018) and JARVIS CFID (Choudhary et al., 2020). The tasks seek designs with high critical temperature and low formation energy per atom, respectively; for JARVIS we negate formation energy so larger response values are preferred in both benchmarks.

Because closed-loop optimization can recommend candidates outside the original rows, we construct a deterministic frozen response oracle for each benchmark; original labels are used only for oracle construction and validation. Table 1 summarizes the environments, with data preprocessing, descriptor selection, oracle training, and held-out validation detailed in in Appendix A.3.

The superconductivity target is the upper 15% of the oracle response distribution; JARVIS uses the upper 25% because stricter thresholds create an almost entirely no-hit regime (Appendix A.3.3).

Failure mechanisms. Each frozen oracle is augmented with four structured invalid-output mechanisms: feature-interaction failures model hidden incompatibilities among otherwise plausible variables; low-validity failures create nonlinear unreliable regions; off-manifold failures penalize extrapolation away from known designs; and hybrid failures combine these effects. Each mechanism is calibrated to nominal invalid rates of 0.2, 0.3, and 0.4 (Appendix A.4).

Closed-loop protocol. Each run starts from $n_0 = 50$ observations drawn from the bottom 65%, 75%, or 85% of the oracle response distribution, followed by five rounds with batch size $B = 5$. ARC-SC and POF use identical target and validity model classes. POF selects all five candidates by

Table 2: First-hit comparison against POF. W/T/L denotes earlier wins, ties, and later losses; conditional win fraction uses discordant pairs only. Intervals are 95% confidence intervals.

Dataset	W/T/L	Cond. win frac.	Sign p	$\Delta\widetilde{\tau}_{\mathrm{hit}}$
Superconductivity	70/251/39	$0.642\,[0.545, 0.732]$	0.004	$-0.269\,[-0.539, -0.025]$
JARVIS CFID	49/272/39	$0.557\,[0.447, 0.663]$	0.337	$-0.117\,[-0.297, 0.053]$

$h(x)$, whereas ARC-SC uses $K = 3$ anchors and two scenario-coverage positions. For each benchmark we evaluate four failure mechanisms, three invalid-rate levels, three initialization settings, and ten held-out seeds (5–14), giving 360 paired cases.

Candidate pools and JARVIS reachability. Uniform candidate pools of $N = 10{,}000$ are shared by both methods for all superconductivity cases and for JARVIS feature-interaction and low-validity failures. In the 15-dimensional JARVIS space, however, global box sampling combined with off-manifold or hybrid failure makes nearly every case a no-hit tie, so the experiment mainly measures candidate reachability rather than batch acquisition. For these two mechanisms we therefore sample locally around each method's currently observed valid designs using a closed-loop trust region with no future or held-out target information. These trajectory-dependent cases compare complete closed-loop policies rather than acquisitions on an identical pool; the all-uniform stress test and full construction are reported in Appendix A.6.1.

Evaluation and statistical analysis. The primary metric is censored first-hit round: $\widetilde{\tau}_{\mathrm{hit}} = \tau_{\mathrm{hit}}$ if a newly queried valid target hit is found, and 6 otherwise. We report paired differences $\Delta\widetilde{\tau}_{\mathrm{hit}} = \widetilde{\tau}_{\mathrm{hit}}^{\mathrm{ARC\text{-}SC}} - \widetilde{\tau}_{\mathrm{hit}}^{\mathrm{POF}}$, where negative values favor ARC-SC, and count wins/ties/losses by the sign of this difference. Uncertainty is quantified with an exact two-sided sign test on discordant pairs and 95% hierarchical-bootstrap intervals using 10,000 replicates. Secondary endpoints are cumulative valid target hits and invalid-evaluation rate.

4.2 Overall First-Hit Performance

Table 2 summarizes the aggregate paired results. On superconductivity, ARC-SC obtains 70 earlier first-hit outcomes, 251 ties, and 39 later outcomes. Among the 109 discordant pairs, ARC-SC wins 64.2% of the comparisons, with an exact 95% confidence interval of $[54.5\%, 73.2\%]$ and a two-sided sign-test value of $p = 0.004$. The mean paired censored first-hit effect is $\Delta\widetilde{\tau}_{\mathrm{hit}} = -0.269$, with hierarchical-bootstrap 95% confidence interval $[-0.539, -0.025]$. Thus, both the win–loss imbalance and the average first-hit effect favor ARC-SC.

ARC-SC's superconductivity advantage includes both exclusive discoveries, 41 cases versus 23 for POF, and earlier same-budget discoveries, 29 cases versus 16. It also reduces invalid evaluations by 5.9 percentage points on average and adds 102 valid target hits in aggregate, although the cumulative-hit interval includes zero. On mixed-pool JARVIS, ARC-SC is directionally favorable but inconclusive: 49/272/39 W/T/L, $\Delta\widetilde{\tau}_{\mathrm{hit}} = -0.117$ with 95% CI $= [-0.297, 0.053]$, 18 additional aggregate valid target hits, and a small mean reduction in invalid evaluations.

4.3 Discovery Probability Across Experimental Rounds

Figure 1 shows the probability of obtaining at least one newly queried valid target hit by each round.

For superconductivity, RC-SC separates from POF immediately: round-1 discovery probabilities are 15.3% versus 10.0%, and round-2 probabilities are 23.1% versus 16.4%, with paired bootstrap intervals excluding zero. By round 5, ARC-SC reaches 35.6% compared with 30.6% for POF. The JARVIS curves are closer, reaching 48.1% and 44.4% by round 5; their intervals include zero but remain consistent with competitive behavior in the higher-dimensional benchmark.

4.4 Behavior Across Failure Mechanisms

The experiments also reveal that the value of scenario coverage depends on the failure structure. On superconductivity, the clearest advantage occurs under feature-interaction failures, with 37/33/20

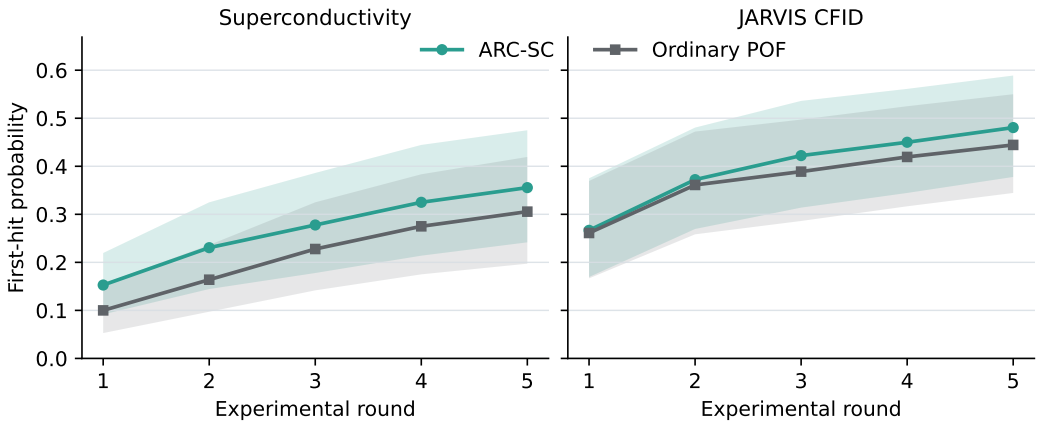


Figure 1: First-hit probability by round. Lines show ARC-SC and POF; shaded regions denote hierarchical-bootstrap 95% confidence intervals. Left: superconductivity with shared uniform pools. Right: JARVIS with the mixed candidate-pool protocol.

W/T/L, 98 additional valid target hits, and a 13.5 percentage-point invalid-rate reduction. Low-validity failures also favor ARC-SC in first-hit timing, while hybrid and off-manifold settings are tie-dominated. This pattern is consistent with scenario diversification being most useful when several high-scoring candidates can represent overlapping but uncertain opportunities, whereas in strongly off-manifold regimes the availability of reachable candidates can dominate the acquisition rule. On JARVIS, all four mechanisms remain nondegenerate under the mixed-pool protocol, but mechanism-level rows are descriptive rather than individually powered; full breakdowns are in Appendix A.6.

5 Conclusion

We introduced ARC-SC, an anchored risk-constrained scenario-coverage method for first-hit batch inverse design under experimental failure. ARC-SC preserves strong marginal valid-hit candidates as anchors, then uses submodular predictive-scenario coverage to allocate the remaining batch capacity to complementary target opportunities under validity/support risk control.

For fixed anchors and a fixed admissible completion pool, greedy scenario completion inherits the classical monotone-submodular approximation guarantee. Holdout-seed experiments support ARC-SC as a POF-anchored diversification strategy: it gives statistically supported first-hit gains on superconductivity and remains competitive on the higher-dimensional JARVIS benchmark.

Several directions follow naturally from these results. First, more faithful posterior-scenario modeling, together with adaptive control of the anchor count and risk budget, could allow the degree of diversification to respond to model confidence and estimated redundancy among high-scoring candidates. This direction also connects to a broader view of scientific modeling in which uncertainty should be propagated coherently across interacting components rather than treated locally within individual predictions (Wang et al., 2024). Second, the JARVIS experiments indicate that candidate generation and acquisition should not be viewed as independent components: when a high-dimensional proposal space is dominated by unsupported designs, candidate reachability can become more limiting than the acquisition rule itself. This observation connects scenario-aware selection with adaptive local and subspace-search methods (Eriksson et al., 2019; Papenmeier et al., 2022), and motivates policies that jointly adapt the reachable search region, risk allocation, and batch composition. Finally, the first-hit formulation points toward a broader objective for autonomous experimentation: optimizing the scientific discovery process under finite budgets, experimental failures, and parallel capacity, rather than only the terminal objective value (Garnett et al., 2012; Volk & Abolhasani, 2024; Tom et al., 2024). Broader baseline comparisons and real closed-loop studies will be important for determining when such adaptive discovery policies provide the greatest practical benefit.

AI USE STATEMENT

In this work, we used generative AI tools only to edit the manuscript for language, clarity, and readability. We did not use generative AI tools for any tasks requiring mandatory disclosure under the ICLR 2027 AI Policy. We have reviewed all AI-assisted work. We take responsibility for the final content of this work.

REFERENCES

Javad Azimi, Alan Fern, and Xiaoli Fern. Batch bayesian optimization via simulation matching. *Advances in neural information processing systems*, 23, 2010.

Daniel Bergmann and Knut Graichen. Safe bayesian optimization under unknown constraints. In *2020 59th IEEE Conference on Decision and Control (CDC)*, pp. 3592–3597. IEEE, 2020.

Yuxin Chen and Andreas Krause. Near-optimal batch mode active learning and adaptive submodular optimization. In *International Conference on Machine Learning*, pp. 160–168. PMLR, 2013.

Kamal Choudhary, Kevin F Garrity, Andrew CE Reid, Brian DeCost, Adam J Biacchi, Angela R Hight Walker, Zachary Trautt, Jason Hattrick-Simpers, A Gilad Kusne, Andrea Centrone, et al. The joint automated repository for various integrated simulations (jarvis) for data-driven materials design. *npj computational materials*, 6(1):173, 2020.

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

Jonathan Foldager, Mikkel Jordahn, Lars K Hansen, and Michael R Andersen. On the role of model uncertainties in bayesian optimisation. In *Uncertainty in Artificial Intelligence*, pp. 592–601. PMLR, 2023.

Jacob R Gardner, Matt J Kusner, Zhixiang Eddie Xu, Kilian Q Weinberger, and John P Cunningham. Bayesian optimization with inequality constraints. In *ICML*, volume 2014, pp. 937–945, 2014.

Roman Garnett, Yamuna Krishnamurthy, Xuehan Xiong, Jeff Schneider, and Richard Mann. Bayesian optimal active search and surveying. In *Proceedings of the 29th International Coference on International Conference on Machine Learning*, pp. 843–850, 2012.

Javier González, Zhenwen Dai, Philipp Hennig, and Neil Lawrence. Batch bayesian optimization via local penalization. In *Artificial intelligence and statistics*, pp. 648–657. PMLR, 2016.

Kam Hamidieh. Superconductivty Data. UCI Machine Learning Repository, 2018. DOI: https://doi.org/10.24432/C53P47.

José Miguel Hernández-Lobato, Michael Gelbart, Matthew Hoffman, Ryan Adams, and Zoubin Ghahramani. Predictive entropy search for bayesian optimization with unknown constraints. In *International conference on machine learning*, pp. 1699–1707. PMLR, 2015.

Riley J Hickman, Gary Tom, Yunheng Zou, Matteo Aldeghi, and Alán Aspuru-Guzik. Anubis: Bayesian optimization with unknown feasibility constraints for scientific experimentation. *Digital Discovery*, 4(8):2104–2122, 2025.

Shogo Iwazaki, Tomohiko Tanabe, Mitsuru Irie, Shion Takeno, Kota Matsui, and Yu Inatsu. No-regret bayesian optimization with stochastic observation failures. In *International Conference on Artificial Intelligence and Statistics*, pp. 415–423. PMLR, 2025.

Shali Jiang, Gustavo Malkomes, Geoff Converse, Alyssa Shofner, Benjamin Moseley, and Roman Garnett. Efficient nonmyopic active search. In *International Conference on Machine Learning*, pp. 1714–1723. PMLR, 2017.

Tarun Kathuria, Amit Deshpande, and Pushmeet Kohli. Batched gaussian process bandit optimization via determinantal point processes. *Advances in neural information processing systems*, 29, 2016.

A Gilad Kusne, Heshan Yu, Changming Wu, Huairuo Zhang, Jason Hattrick-Simpers, Brian DeCost, Suchismita Sarker, Corey Oses, Cormac Toher, Stefano Curtarolo, et al. On-the-fly closed-loop materials discovery via bayesian active learning. *Nature communications*, 11(1):5966, 2020.

Turab Lookman, Prasanna V Balachandran, Dezhen Xue, and Ruihao Yuan. Active learning in materials science with emphasis on adaptive sampling using uncertainties for targeted design. *npj Computational Materials*, 5(1):21, 2019.

Willie Neiswanger and Aaditya Ramdas. Uncertainty quantification using martingales for misspecified gaussian processes. In *Algorithmic learning theory*, pp. 963–982. PMLR, 2021.

George L Nemhauser, Laurence A Wolsey, and Marshall L Fisher. An analysis of approximations for maximizing submodular set functions—i. *Mathematical programming*, 14(1):265–294, 1978.

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

Anjali Parashar, Rachel Luo, Apoorva Sharma, Sushant Veer, Edward Schmerling, Carson Sobolewski, Mingxin Yu, Chuchu Fan, and Marco Pavone. Coverage aware active evaluation for failure discovery with paired systems. *arXiv preprint arXiv:2608.13719*, 2026.

Amar Shah and Zoubin Ghahramani. Parallel predictive entropy search for batch global optimization of expensive objective functions. *Advances in neural information processing systems*, 28, 2015.

Yunsheng Tian, Ane Zuniga, Xinwei Zhang, Johannes P Dürholt, Payel Das, Jie Chen, Wojciech Matusik, and Mina Konaković Luković. Boundary exploration for bayesian optimization with unknown physical constraints. In *Proceedings of the 41st International Conference on Machine Learning*, pp. 48295–48320, 2024.

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

Amanda A Volk and Milad Abolhasani. Performance metrics to unleash the power of self-driving labs in chemistry and materials science. *Nature communications*, 15(1):1378, 2024.

Yuki K Wakabayashi, Takuma Otsuka, Yoshiharu Krockenberger, Hiroshi Sawada, Yoshitaka Taniyasu, and Hideki Yamamoto. Bayesian optimization with experimental failure for high-throughput materials growth. *npj Computational Materials*, 8(1):180, 2022.

Chenxi Wang, Jihui Zhao, Jingjing Zheng, Barak Raveh, Xuming He, and Liping Sun. Bayesian occam's razor to optimize metamodeling for complex biological systems. *bioRxiv*, pp. 2024–05, 2024.

Jian Wu and Peter Frazier. The parallel knowledge gradient method for batch bayesian optimization. *Advances in neural information processing systems*, 29, 2016.

Linhan Wu, Chenxi Wang, Chuhan Yang, Zhengwei Yang, and Yuyang Liu. Matformbench: A benchmarking evaluation framework for target-driven materials formulation. *arXiv preprint arXiv:2605.26741*, 2026.

Dezhen Xue, Prasanna V Balachandran, John Hogden, James Theiler, Deqing Xue, and Turab Lookman. Accelerated search for materials with targeted properties by adaptive design. *Nature communications*, 7(1):11241, 2016.

A Appendix

A.1 Implementation and Reproducibility

This section gives the implementation details needed to reproduce the acquisition rules used in Section 4. We distinguish the shared predictive infrastructure from the batch-selection rules. ARC-SC

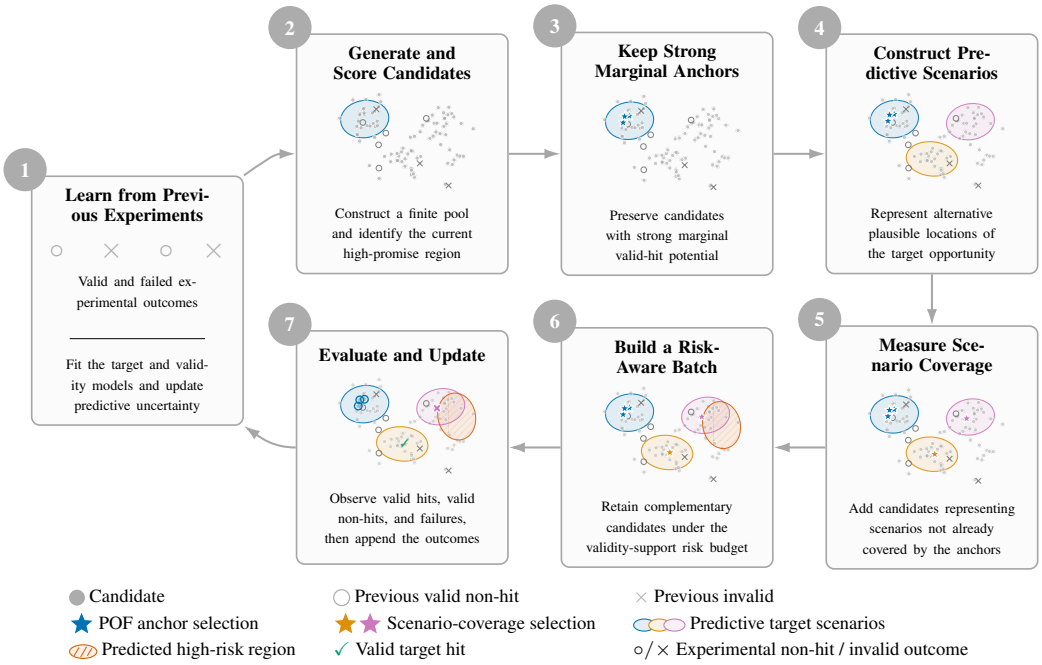


Figure 2: Integrated schematic overview of ARC-SC. At each closed-loop iteration, target and validity models are updated from previous valid and failed experiments and used to score a finite candidate pool. Strong marginal valid-hit candidates are retained through POF-style risk-admissible anchoring. Multiple predictive target scenarios then expose alternative plausible opportunities, and the remaining batch positions are assigned to complementary candidates under a validity-support risk budget. Experimental evaluation reveals valid target hits, valid non-hits, and invalid outcomes, which are appended to the observed data for the next iteration. The embedded design-space illustrations are schematic and do not represent a particular experimental run.

and the POF baseline use the same target surrogate, validity model, observed data, target threshold, and—for the uniform-pool experiments—the same finite candidate pool. The POF baseline is the method defined in equation 4: it selects the B largest marginal valid-hit scores $h(x) = p_y(x)p_v(x)$ and does not apply ARC-SC's support-based risk filter. ARC-SC instead uses $h(x)$ for risk-admissible anchoring and uses scenario coverage for the remaining positions.

A.1.1 Uniform Candidate-Pool Construction

For the uniform-pool experiments, bounds are inferred independently for every selected descriptor from the complete benchmark reference feature matrix. If ℓ_j and u_j are the observed minimum and maximum of descriptor j, the padded search interval is

$$[\ell_j - \rho_b(u_j - \ell_j),\ u_j + \rho_b(u_j - \ell_j)]\,, \qquad \rho_b = 0.05. \tag{18}$$

At every round, $N = 10{,}000$ candidates are sampled independently and uniformly from this axis-aligned box. In a paired uniform-pool case, the pool is sampled once using the condition- and round-specific seed and is supplied unchanged to ARC-SC and POF. Thus, the two methods are compared on exactly the same candidates. This uniform protocol is used for all superconductivity experiments and for the JARVIS feature-interaction and low-validity mechanisms. We also retain an all-uniform JARVIS stress test in Appendix A.6.1.

A.1.2 Closed-Loop JARVIS Trust-Region Candidate Pool

In the 15-dimensional JARVIS descriptor space, global uniform sampling becomes strongly dominated by unsupported candidates under the off-manifold and hybrid failure mechanisms. For these two mechanisms in the main mixed-pool JARVIS experiment, we therefore generate candidates from a closed-loop empirical trust region.

At the start of each round, let $\mathcal{O}_t$ denote the rows observed by the current method and let $\mathcal{O}_t^v \subseteq \mathcal{O}_t$ denote its valid rows. Candidate anchors are sampled uniformly with replacement from $\mathcal{O}_t^v$ whenever at least one valid row is available; otherwise all observed rows are used. For descriptor j, let $\mathrm{IQR}_{t,j}$ be the interquartile range among the currently observed designs. The local perturbation scale is

$$s_{t,j} = 0.05\,\widetilde{\mathrm{IQR}}_{t,j}, \tag{19}$$

where a degenerate observed IQR is replaced first by the corresponding IQR of the benchmark reference features and, if necessary, by 10^{-3} times the full reference range. Given a sampled anchor a, a candidate is generated as

$$x_j = a_j + \epsilon_j, \qquad \epsilon_j \sim \mathcal{N}(0, s_{t,j}^2), \tag{20}$$

then clipped to the unpadded minimum and maximum of the benchmark reference features. Duplicate candidate rows are removed.

The trust-region pool uses only feature values and observed validity outcomes available to that method at the start of the round. It does not use held-out target labels, oracle values for unqueried candidates, future observations, or observations generated by the competing method. Consequently, the trust-region pools are trajectory dependent and need not be identical between ARC-SC and POF. These cases therefore compare complete closed-loop policies rather than two acquisition rules on a fixed common pool. This distinction is explicit in our interpretation of the JARVIS results.

A.1.3 Target and Validity Models

At every round, the target surrogate is refitted using only valid observations. The implemented experiments use a Gaussian-process regressor after standardizing the input descriptors. Its kernel is

$$k(x, x') = C\, k_{\text{Matérn},\nu=5/2}(x, x') + k_{\text{white}}(x, x'), \tag{21}$$

where the constant-kernel amplitude has bounds $[10^{-3}, 10^3]$, the initial Matérn length scale is one, and the white-noise level is initialized at 10^{-3}. The regressor normalizes the target internally and uses two optimizer restarts. Predictive means and standard deviations are denoted by $\mu(x)$ and $\sigma_y(x)$.

The validity model is fitted to all observed valid/invalid labels. The experiments use a random-forest classifier with 300 trees, square-root feature subsampling, balanced-subsample class weighting, and the condition-specific random seed. If only one validity class has been observed, the implementation uses a constant-probability model until both classes are available. ARC-SC and POF use the same model classes and model seeds within each paired case.

A.1.4 Empirical-Support Score

Let $\widetilde{x}$ denote a candidate after standardizing each descriptor using the mean and standard deviation of the currently observed designs. Let $\mathcal{N}_k(x)$ be its k nearest observed designs in Euclidean distance. With $k = 5$,

$$d_{\mathrm{kNN}}(x) = \frac{1}{k} \sum_{x_i \in \mathcal{N}_k(x)} \|\widetilde{x} - \widetilde{x}_i\|_2. \tag{22}$$

The scale τ_s is recomputed at each round as the empirical 75th percentile of the corresponding five-neighbor distances evaluated on the observed designs, with a numerical lower bound of 10^{-6}. The support score is

$$p_s(x) = \exp\left[-\frac{d_{\mathrm{kNN}}(x)^2}{2\tau_s^2}\right]. \tag{23}$$

This is a relative empirical-support score rather than a calibrated probability.

A.1.5 Predictive Scenarios, Rank Gates, and Risk Fallback

The implementation can draw correlated Gaussian-process samples for small candidate pools. Automatic correlated sampling is enabled only when the candidate pool contains at most 256 candidates. Since all implemented experiments use $N = 10{,}000$, they use independent marginal draws

$$f^{(m)}(x_j) = \mu(x_j) + \sigma_y(x_j) z_{m,j}, \qquad z_{m,j} \overset{\text{iid}}{\sim} \mathcal{N}(0, 1), \tag{24}$$

Table 3: Algorithm and evaluation settings. Dataset-specific entries are shown explicitly.

Quantity	Superconductivity	JARVIS CFID
Initial observations n_0	50	50
Batch size B / rounds	5 / 5	5 / 5
POF anchors K	3	3
Candidate-pool size N	10,000	10,000
Predictive scenarios M	64	64
Target quantile	0.85	0.75
Risk budget R	0.40	0.80
Uniform bounds padding ρ_b	0.05	0.05
Support neighbors k	5	5
Support-scale quantile	0.75	0.75
Relaxation fraction π	0.05	0.05
Effective temperature τ	0.05	0.05
Rank floors $(\epsilon_h, \epsilon_v, \epsilon_s)$	$(.05, .05, .05)$	$(.05, .05, .05)$
Rank exponents $(\alpha_h, \alpha_v, \alpha_s)$	$(1, .5, .5)$	$(1, .5, .5)$
Fallback coefficient λ_R	10	10
Correlated-sampling limit	256	256
Seed IDs	5–14	5–14
Main candidate pools	uniform	mixed; see Appendix A.1.2
Trust-region noise scale	–	$0.05\times$ observed IQR

for $m = 1, \ldots, M$ and $j = 1, \ldots, N$. These samples preserve candidate-wise predictive means and variances but not cross-candidate posterior covariance.

For the greater-than targets used here,

$$T_m^{\text{relax}} = \min\Big\{T,\ Q_{1-\pi}\big(\{f^{(m)}(x) : x \in \mathcal{C}\}\big)\Big\}, \qquad \pi = 0.05, \tag{25}$$

and

$$\psi_m(x) = \text{sigm}\left(\frac{f^{(m)}(x) - T_m^{\text{relax}}}{\tau}\right). \tag{26}$$

The effective soft-response temperature in all runs is $\tau = 0.05$.

For a score vector $q(x)$ over a pool of size N, average ranks are used for ties, the best candidate receives rank percentile one, and the worst receives $1/N$. The gates are

$$g_q(x) = [\epsilon_q + (1 - \epsilon_q)R_q(x)]^{\alpha_q}, \qquad q \in \{h, v, s\}, \tag{27}$$

with

$$\epsilon_h = \epsilon_v = \epsilon_s = 0.05, \qquad \alpha_h = 1, \quad \alpha_v = 0.5, \quad \alpha_s = 0.5. \tag{28}$$

For a partial batch $\mathcal{S}$ and candidate x, define the risk violation

$$V_R(x \mid \mathcal{S}) = [C(\mathcal{S} \cup \{x\}) - R]_+. \tag{29}$$

If a risk-feasible extension exists, ARC-SC maximizes the stage-specific desirability over the feasible extensions: $h(x)$ for POF anchoring and $\Delta_{\text{SC}}(x \mid \mathcal{S})$ for scenario completion. If none exists, it selects

$$x^\star \in \arg\max_{x \in \mathcal{C} \setminus \mathcal{S}} \{D(x \mid \mathcal{S}) - \lambda_R V_R(x \mid \mathcal{S})\}, \qquad \lambda_R = 10. \tag{30}$$

The fallback ensures that a full batch can always be returned, but it means that R is enforced whenever a feasible extension exists rather than as an unconditional terminal-batch guarantee.

A.1.6 Hyperparameters Used in the Experiments

The JARVIS risk budget $R = 0.80$ is a dataset-specific operating setting identified in a JARVIS sensitivity run and then held fixed across the JARVIS grids. It should therefore be interpreted as a fixed JARVIS operating point rather than as independent evidence about the optimal value of R.

A.1.7 Computational Complexity

Let n_t be the number of observations at the current round, $n_v \le n_t$ the number of valid observations, d the descriptor dimension, N the candidate-pool size, M the number of predictive scenarios, and $B - K$ the number of scenario-coverage positions. Target-surrogate and validity-model fitting are shared by ARC-SC and the POF baseline. For the Gaussian-process target surrogate, exact fitting requires $O(n_v^3)$ time and $O(n_v^2)$ memory; predictive variances on N candidates can require $O(Nn_v^2)$ time with a dense exact GP implementation. These costs are common to both acquisition rules. A brute-force support-distance evaluation is $O(Nn_td)$, although nearest-neighbor acceleration can reduce this in favorable low-dimensional regimes.

Conditional on the shared candidate scores, the ARC-SC-specific acquisition cost is dominated by the scenario matrix and greedy completion. Computing percentile ranks is $O(N \log N)$ using sorting. Generating M marginal predictive scenarios and their response matrix requires $O(MN)$ arithmetic and $O(MN)$ memory. The scenario-specific relaxed quantiles can be computed in expected $O(MN)$ time using selection-based quantile routines (or $O(MN \log N)$ under full sorting). The K anchor scans require $O(KN)$ time. For each of the $B - K$ coverage positions, all remaining candidates are scored against the current uncovered scenario mass, which costs $O(MN)$; therefore greedy scenario completion requires $O((B - K)MN)$ time. Risk-feasibility checks add $O(BN)$, which is lower order for the M. Thus, excluding model fitting shared with the baseline, the principal ARC-SC acquisition cost is $O(N \log N + (B - K + 1)MN + BN)$, with memory $O(MN + N)$.

A.2 Additional Theoretical Results

Throughout this section, the observed dataset, finite candidate pool, predictive scenarios, and response matrix are conditioned upon and treated as fixed.

A.2.1 Complete Greedy-Approximation Derivation

Let $\mathcal{A}$ be a fixed anchor set and let $k = B - K \ge 1$. Over a fixed admissible completion pool $\mathcal{F} \subseteq \mathcal{C} \setminus \mathcal{A}$, define

$$G_{\mathcal{A}}(\mathcal{S}) = A_{\mathrm{SC}}(\mathcal{A} \cup \mathcal{S}) - A_{\mathrm{SC}}(\mathcal{A}). \tag{31}$$

As established in Proposition 2, $G_{\mathcal{A}}$ is normalized, monotone, and submodular. Let $\mathcal{S}_i$ be the greedy completion after i selections, with $\mathcal{S}_0 = \varnothing$, and let

$$\mathcal{S}^\star \in \arg\max_{\substack{\mathcal{S} \subseteq \mathcal{F} \\ |\mathcal{S}| \le k}} G_{\mathcal{A}}(\mathcal{S}) \tag{32}$$

be an optimal completion.

At step i, monotonicity and submodularity imply

$$G_{\mathcal{A}}(\mathcal{S}^\star) - G_{\mathcal{A}}(\mathcal{S}_{i-1}) \le G_{\mathcal{A}}(\mathcal{S}^\star \cup \mathcal{S}_{i-1}) - G_{\mathcal{A}}(\mathcal{S}_{i-1}) \tag{33}$$

$$\le \sum_{x \in \mathcal{S}^\star \setminus \mathcal{S}_{i-1}} \Delta_{\mathcal{A}}(x \mid \mathcal{S}_{i-1}), \tag{34}$$

where $\Delta_{\mathcal{A}}$ is the marginal gain of the residual objective. Since $|\mathcal{S}^\star| \le k$, at least one candidate in the sum has marginal gain no smaller than

$$\frac{G_{\mathcal{A}}(\mathcal{S}^\star) - G_{\mathcal{A}}(\mathcal{S}_{i-1})}{k}. \tag{35}$$

The exact greedy candidate has at least this marginal gain, and hence

$$G_{\mathcal{A}}(\mathcal{S}_i) - G_{\mathcal{A}}(\mathcal{S}_{i-1}) \ge \frac{G_{\mathcal{A}}(\mathcal{S}^\star) - G_{\mathcal{A}}(\mathcal{S}_{i-1})}{k}. \tag{36}$$

Define the remaining gap

$$E_i = G_{\mathcal{A}}(\mathcal{S}^\star) - G_{\mathcal{A}}(\mathcal{S}_i). \tag{37}$$

Equation 36 gives

$$E_i \le \left(1 - \frac{1}{k}\right) E_{i-1}. \tag{38}$$

Recursing for k steps and using $G_{\mathcal{A}}(\varnothing) = 0$ yields

$$E_k \leq \left(1 - \frac{1}{k}\right)^k G_{\mathcal{A}}(\mathcal{S}^\star). \tag{39}$$

Therefore,

$$G_{\mathcal{A}}(\mathcal{S}_k) \geq \left[1 - \left(1 - \frac{1}{k}\right)^k\right] G_{\mathcal{A}}(\mathcal{S}^\star) \geq \left(1 - \frac{1}{e}\right) G_{\mathcal{A}}(\mathcal{S}^\star), \tag{40}$$

which is the guarantee stated in Corollary 1. This result compares greedy completion only with the optimal completion of the same fixed anchors over the same fixed admissible pool.

A.2.2 Why the Dynamic Risk Rule Does Not Inherit the Cardinality Guarantee

For a nonempty set $\mathcal{S}$, the implemented average risk-support cost is

$$C(\mathcal{S}) = \frac{1}{|\mathcal{S}|} \sum_{x \in \mathcal{S}} c(x). \tag{41}$$

Consider the family

$$\mathfrak{F}_R = \{\mathcal{S} \subseteq \mathcal{C} : \mathcal{S} \neq \varnothing,\ C(\mathcal{S}) \leq R\} \cup \{\varnothing\}. \tag{42}$$

This family is not necessarily downward closed. For example, let $c(x_1) = 0.9$, $c(x_2) = 0.1$, and $R = 0.5$. Then

$$C(\{x_1, x_2\}) = 0.5, \tag{43}$$

so the pair is feasible, whereas $C(\{x_1\}) = 0.9 > 0.5$, so one of its subsets is infeasible.

Moreover, the admissibility of a candidate depends on the current prefix. The condition

$$C(\mathcal{S} \cup \{x\}) \leq R \tag{44}$$

is equivalent to

$$c(x) \leq (|\mathcal{S}| + 1)R - \sum_{s \in \mathcal{S}} c(s). \tag{45}$$

The right-hand side changes as low-cost candidates are added, so a candidate can be inadmissible at one step and admissible at a later step. The implemented procedure therefore does not optimize over a fixed hereditary feasible family. In addition, the fallback in equation 30 can deliberately select an infeasible extension when no feasible candidate remains. These features prevent direct application of the fixed-cardinality $1 - 1/e$ guarantee to the complete ARC-SC implementation.

A precise local statement still holds. Define

$$\mathcal{F}_R(\mathcal{S}) = \{x \in \mathcal{C} \setminus \mathcal{S} : C(\mathcal{S} \cup \{x\}) \leq R\}. \tag{46}$$

Whenever $\mathcal{F}_R(\mathcal{S}) \neq \varnothing$, the scenario-completion step selects

$$x^\star \in \arg\max_{x \in \mathcal{F}_R(\mathcal{S})} \Delta_{\mathrm{SC}}(x \mid \mathcal{S}). \tag{47}$$

Thus, each feasible completion step is exactly optimal among the currently risk-admissible one-step extensions, although no global approximation ratio is claimed for the dynamically filtered final batch.

A.2.3 Finite-Scenario Approximation

For a fixed batch $\mathcal{B}$, define the contribution of predictive scenario m as

$$Z_m(\mathcal{B}) = 1 - \prod_{x \in \mathcal{B}} (1 - r_m(x)), \qquad 0 \leq Z_m(\mathcal{B}) \leq 1. \tag{48}$$

Under the implemented independent sampling across scenario index m,

$$A_{\mathrm{SC}}(\mathcal{B}) = \frac{1}{M} \sum_{m=1}^{M} Z_m(\mathcal{B}). \tag{49}$$

Table 4: Held-out superconductivity oracle performance.

Model	R^2	RMSE	MAE	Spearman ρ
Extra Trees	0.9128	10.0184	5.6564	0.9556
Random forest	0.9095	10.2057	5.8667	0.9535
Hist. gradient boosting	0.8932	11.0899	6.8611	0.9352

Hence, for any fixed $\mathcal{B}$, Hoeffding's inequality gives

$$\Pr(|A_{\mathrm{SC}}(\mathcal{B}) - \mathbb{E}[A_{\mathrm{SC}}(\mathcal{B})]| \geq \varepsilon) \leq 2\exp(-2M\varepsilon^2). \tag{50}$$

This statement concerns Monte Carlo error for a *fixed* batch under the implemented scenario generator. It does not by itself provide a uniform guarantee over all candidate batches or a post-selection guarantee for the adaptively chosen ARC-SC batch. Its role is therefore to characterize the sampling error of the scenario-coverage estimator, not the statistical calibration of the resulting acquisition value.

A.3 Frozen-Oracle Benchmark Construction

The frozen-oracle protocol separates response-surface construction from closed-loop acquisition evaluation. Original dataset labels are used to select descriptors, train candidate regression models, and assess held-out quality. Once an oracle is selected, its predictions are frozen and used as the benchmark response for both existing reference rows and newly generated candidates. The closed-loop algorithms never alternate between original labels and oracle predictions.

A.3.1 UCI Superconductivity Benchmark

The source table contains 21,263 rows, 81 numerical composition descriptors, and the critical-temperature target. Rows containing missing numerical values are removed; no rows were removed in the downloaded table. An 80/20 split with random seed 42 is formed before descriptor selection.

Each descriptor is scored using three training-only statistics: absolute Spearman correlation with the target, mutual information, and Extra-Trees feature importance. Each statistic is converted to a percentile rank and the three ranks are averaged. The selector prioritizes a prespecified list of physically interpretable descriptors and excludes a candidate if its absolute Spearman correlation with any already selected descriptor exceeds 0.95. The selected five descriptors are

1. `wtd_mean_Valence`,
2. `wtd_std_ThermalConductivity`,
3. `wtd_std_atomic_radius`,
4. `wtd_entropy_atomic_mass`, and
5. `wtd_std_fie`.

Three oracle candidates are trained on the selected descriptors: a random forest with 800 trees, an Extra-Trees regressor with 800 trees, and histogram gradient boosting with 800 iterations, learning rate 0.04, and 31 leaf nodes. Models are ordered primarily by held-out Spearman correlation and secondarily by R^2. Table 4 reports the held-out metrics. Extra Trees is selected as the frozen oracle.

The superconductivity discovery threshold is the 0.85 quantile of the frozen-oracle predictions over the reference dataset, which equals $T = 80.0$ in the benchmark.

A.3.2 JARVIS CFID Benchmark

The JARVIS CFID table contains 55,723 rows. Expanding the stored CFID vectors produces 1,543 candidate numerical descriptor columns. The original formation energy per atom is transformed as

$$y = -\mathrm{formation_energy_per_atom}, \tag{51}$$

Table 5: Held-out JARVIS oracle performance.

Model	R^2	RMSE	MAE	Spearman ρ
Random forest	0.8375	0.4320	0.2518	0.9121
Extra Trees	0.8311	0.4404	0.2531	0.9088
Hist. gradient boosting	0.8239	0.4498	0.2828	0.9002

Table 6: JARVIS target-threshold calibration. W/T/L compares ARC-SC with the POF baseline using the round-level first-hit metric.

Target quantile	Threshold	Cases	W/T/L	ΔHits	ΔInvalid
0.75	1.5391	36	8/21/7	-3	$+0.079$
0.80	1.8224	36	2/30/4	-2	$+0.011$
0.85	2.0941	36	1/35/0	$+4$	$+0.060$

so that both benchmarks use a greater-than target. The data are divided into an 80/20 train/test split with random seed 42. Missing descriptor values are handled by median imputation inside the oracle pipelines.

Descriptor scores again average percentile ranks of absolute Spearman correlation, mutual information, and Extra-Trees importance. To control feature-selection cost, scoring uses at most 20,000 training rows. The 250 highest-scoring descriptors are considered sequentially, and candidates with absolute Spearman correlation greater than 0.98 with an already selected descriptor are excluded. The selected descriptors are

`cfid_0048`, `cfid_0124`, `cfid_0240`, `cfid_0168`, `cfid_0075`, `cfid_0384`, `cfid_0405`, `cfid_0324`, `cfid_0236`, `cfid_0189`, `cfid_0345`, `cfid_0134`, `cfid_0361`, `cfid_0505`, and `cfid_0183`.

The candidate oracles are a 700-tree Extra-Trees regressor, a 500-tree random forest, and histogram gradient boosting with 700 iterations, learning rate 0.04, and 31 leaf nodes. Each pipeline uses median imputation. Table 5 reports the held-out metrics. The random forest is selected as the frozen JARVIS oracle.

A.3.3 JARVIS Target-Threshold Calibration

The initial JARVIS configuration used the 0.85 quantile of the frozen-oracle response distribution, corresponding to the upper 15%. This threshold created a nearly degenerate no-hit benchmark. We therefore compared response quantiles 0.75, 0.80, and 0.85 on a calibration grid containing the feature-interaction and low-validity mechanisms, nominal invalid rates 0.2, 0.3, and 0.4, initialization quantiles 0.65 and 0.75, and three random seeds, for 36 paired cases per threshold.

We select the 0.75 quantile because it produces a materially nonzero discovery regime for both methods and therefore supports meaningful comparison of first-hit behavior and invalid evaluations. The choice is not the threshold at which ARC-SC has the largest relative cumulative-hit advantage; indeed, the calibration grid at 0.75 gives a cumulative-hit difference of -3. The JARVIS experiments use the resulting threshold $T = 1.5391$.

A.4 Synthetic Failure Mechanisms

The source datasets provide target properties but not controlled experimental-failure labels. We therefore apply a deterministic synthetic failure mask to the frozen-oracle response. All mechanisms are target independent: they depend on descriptor-space geometry and hidden feature interactions, while the frozen oracle separately determines the raw target value.

Let x be robustly standardized using the component-wise median and interquartile range of the full reference feature matrix. If an interquartile range is numerically zero, the standard deviation is used, followed by a fallback scale of one. Denote the standardized vector by $\bar{x} \in \mathbb{R}^d$. For every

experimental condition, three random unit vectors w_1, w_2, w_3 are generated from its deterministic seed, and

$$z_j = \bar{x}^\top w_j, \qquad j = 1, 2, 3. \tag{52}$$

A.4.1 Off-Manifold Failure

A ten-nearest-neighbor model is fitted to the robustly standardized reference data, and $d_{\mathrm{man}}(x)$ is the mean neighbor distance. Let $d_{0.8}$ be the 80th percentile and s_d the standard deviation of reference distances. The off-manifold risk is

$$r_{\mathrm{dist}}(x) = \mathrm{sigm}\left(\frac{d_{\mathrm{man}}(x) - d_{0.8}}{s_d}\right). \tag{53}$$

A.4.2 Low-Validity-Probability Failure

The smoothly varying nonlinear risk is

$$r_{\mathrm{nonlin}}(x) = \mathrm{sigm}\left(0.7 \sin z_1 + 0.3 \cos z_2 + 0.15 \frac{z_1 z_2}{\sqrt{d}}\right). \tag{54}$$

This mechanism creates a hidden validity landscape that is not reducible to distance from the reference manifold.

A.4.3 Feature-Interaction Failure

The interaction mechanism uses

$$r_{\mathrm{int}}(x) = \mathrm{sigm}\left(0.70 \frac{z_1 z_2}{\sqrt{d}} + 0.20|z_3| + 0.10 \max\{z_1 - z_2, 0\}\right). \tag{55}$$

It represents a hidden process window governed by combinations of descriptor directions rather than by a single coordinate.

A.4.4 Hybrid Failure

The hybrid risk is the weighted average

$$r_{\mathrm{hyb}}(x) = 0.60 r_{\mathrm{dist}}(x) + 0.25 r_{\mathrm{nonlin}}(x) + 0.15 r_{\mathrm{int}}(x). \tag{56}$$

The weights sum to one.

A.4.5 Calibration to a Nominal Invalid Rate

For any selected risk function $r(x)$ and nominal invalid rate $\rho \in \{0.2, 0.3, 0.4\}$, the deterministic threshold is

$$\theta_\rho = Q_{1-\rho}\big(\{r(x) : x \in \mathcal{X}_{\mathrm{ref}}\}\big). \tag{57}$$

A queried candidate is invalid when $r(x) \geq \theta_\rho$. Its raw frozen-oracle response is retained for diagnostics, but the observed target is set to missing and the validity label is set to zero. Calibration is performed separately for every dataset, mechanism, nominal invalid rate, and seed. Because empirical quantiles can contain ties, the achieved reference invalid rate is approximately, rather than necessarily exactly, ρ.

The four synthetic failure mechanisms represent qualitatively different sources of experimental invalidity. Low-validity failure creates nonlinear regions of design space in which experiments are intrinsically less likely to succeed. Feature-interaction failure represents hidden process constraints that arise from combinations of otherwise individually plausible descriptors. Off-manifold failure penalizes extrapolative candidates that lie far from the empirical distribution of known designs. Finally, the hybrid mechanism combines these effects to represent experimental environments in which invalidity may arise from multiple causes. Together, these mechanisms test whether an acquisition strategy remains robust to local feasibility structure, variable interactions, extrapolation risk, and their combination.

A.5 Experimental Protocol and Statistical Analysis

A.5.1 Holdout-Seed Evaluation Grid

The results in the main paper use the seed IDs $5, \ldots, 14$; Each dataset combines four failure mechanisms, three nominal invalid rates, three initialization quantiles, and ten seed IDs, $4 \times 3 \times 3 \times 10 = 360$ paired evaluation cases. Each run begins with $n_0 = 50$ reference designs sampled without replacement from rows whose frozen-oracle response lies below the specified initialization quantile (0.65, 0.75, or 0.85), followed by five closed-loop rounds of batch size five.

For the shared uniform-pool protocol, ARC-SC and POF receive the same initial data, frozen oracle, failure mechanism, target threshold, candidate pool, and model seeds. The candidate pool at round t is generated once from the common condition seed and then passed to both methods. For the JARVIS trust-region mechanisms, the initial condition and failure environment remain paired, but candidate pools are generated from each method's own observed trajectory as described in Appendix A.1.2.

The first-hit metric is computed only over newly queried rounds 1–5. Initial target-level observations, when present, are recorded diagnostically but do not count as a closed-loop first hit. This matters for JARVIS when the initialization quantile is 0.85 while the target quantile is 0.75.

A.5.2 Primary and Secondary Endpoints

Let $H = 5$ denote the closed-loop horizon. The censored first-hit round is

$$\widetilde{\tau}_{\mathrm{hit}} = \begin{cases} \tau_{\mathrm{hit}}, & \text{if a newly queried hit is found,} \\ H + 1 = 6, & \text{otherwise.} \end{cases} \tag{58}$$

The paired first-hit effect is

$$\Delta\widetilde{\tau}_{\mathrm{hit}} = \widetilde{\tau}_{\mathrm{hit}}^{\mathrm{ARC\text{-}SC}} - \widetilde{\tau}_{\mathrm{hit}}^{\mathrm{POF}}, \tag{59}$$

so negative values favor ARC-SC. A paired run is an earlier-hit win, tie, or later-hit loss according to whether this difference is negative, zero, or positive.

Secondary effects are

$$\Delta N_{\mathrm{hit}} = N_{\mathrm{hit}}^{\mathrm{ARC\text{-}SC}} - N_{\mathrm{hit}}^{\mathrm{POF}}, \tag{60}$$

and

$$\Delta r_{\mathrm{invalid}} = r_{\mathrm{invalid}}^{\mathrm{ARC\text{-}SC}} - r_{\mathrm{invalid}}^{\mathrm{POF}}, \tag{61}$$

where positive ΔN_{hit} and negative $\Delta r_{\mathrm{invalid}}$ favor ARC-SC.

A.5.3 Exact Sign Test and Hierarchical Bootstrap

For the sign test, ties are removed and the number of ARC-SC wins among the discordant pairs is tested against a $\mathrm{Binomial}(W + L, 0.5)$ null model. The main paper reports a two-sided exact p-value and an exact 95% confidence interval for the conditional win fraction $W/(W + L)$.

To quantify effect-size uncertainty while respecting the factorial design, we use a two-level hierarchical bootstrap with 10,000 replicates. The 36 experimental-condition cells are defined by failure mechanism, nominal invalid rate, and initialization quantile. Each bootstrap replicate (i) samples 36 condition cells with replacement and then (ii) resamples the ten paired seed-level observations with replacement within each sampled cell. The statistic is computed within each resampled cell and then averaged across the sampled cells. This procedure is used for the paired first-hit, cumulative-hit, invalid-rate, and discovery-by-round intervals. It therefore reflects variation over both the prespecified condition grid and stochastic repeats rather than treating all 360 rows as homogeneous IID observations.

A.6 Additional Results and Robustness Analyses

A.6.1 Why JARVIS Uses a Mixed Candidate-Pool Protocol

The main JARVIS evaluation intentionally separates two failure regimes. Feature-interaction and low-validity failures are evaluated on the same shared uniform box used in the strict paired comparison. For off-manifold and hybrid failures, however, uniform sampling in the 15-dimensional CFID

Table 7: JARVIS candidate-pool protocol comparison on the 360 cases. The all-uniform protocol uses a shared global box for all four mechanisms. The mixed protocol replaces only off-manifold and hybrid cases with method-specific closed-loop trust-region pools. Negative $\Delta\widetilde{\tau}_{\mathrm{hit}}$ and $\Delta r_{\mathrm{invalid}}$ favor ARC-SC.

Protocol	First-hit W/T/L	$\Delta\widetilde{\tau}_{\mathrm{hit}}$ [95% CI]	ΔHits	$\Delta r_{\mathrm{invalid}}$ [95% CI]	Both no hit
All uniform	32/304/24	$-0.067\,[-0.219, 0.075]$	-21	$+0.029\,[0.015, 0.044]$	275 (76.4%)
Mixed	49/272/39	$-0.117\,[-0.297, 0.053]$	$+18$	$-0.009\,[-0.042, 0.021]$	165 (45.8%)

Table 8: JARVIS results by failure mechanism and candidate-pool construction. Feature-interaction and low-validity rows are identical in the all-uniform and mixed designs.

Failure mechanism	Candidate pool	Cases	First-hit W/T/L	ΔHits	Mean ΔInvalid
Feature interaction	shared uniform	90	15/62/13	$+16$	$+0.059$
Low validity	shared uniform	90	17/62/11	-37	$+0.057$
Off manifold	shared uniform	90	0/90/0	0	0.000
Off manifold	trust region	90	10/73/7	$+7$	-0.101
Hybrid	shared uniform	90	0/90/0	0	0.000
Hybrid	trust region	90	7/75/8	$+32$	-0.052

box produces a candidate domain dominated by unsupported points. In the all-uniform stress test, both of these mechanisms yield $0/90/0$ first-hit W/T/L and zero aggregate hit difference. Under this regime, the experiment provides little information about the relative acquisition rules because candidate reachability dominates the optimization problem.

We therefore replace the candidate generator only for these two mechanisms with the closed-loop trust-region protocol of Appendix A.1.2. Table 7 compares the resulting mixed design with the all-uniform JARVIS stress test. The mixed protocol substantially reduces the fraction of cases in which neither method obtains a hit within five rounds, from $275/360 = 76.4\%$ to $165/360 = 45.8\%$, while leaving the feature-interaction and low-validity experiments unchanged.

The mixed protocol should not be interpreted as making the JARVIS result statistically significant; its first-hit interval still includes zero. Its purpose is instead to avoid an almost uninformative candidate-generation regime and to evaluate the complete closed-loop policies in a reachable local domain. Table 8 shows that the change affects only the off-manifold and hybrid rows.

A.6.2 Complete Effect Sizes

Table 9 reports the paired effects and hierarchical-bootstrap intervals for the two paper-facing designs. The superconductivity first-hit and invalid-rate intervals exclude zero. The cumulative-hit interval does not, consistent with the fact that ARC-SC is designed around early first discovery rather than maximizing total hit count. All three mixed-JARVIS intervals include zero.

A.6.3 First-Hit Tie Decomposition

A first-hit tie can correspond either to both methods discovering in the same round or to neither method discovering within the five-round horizon. Table 10 separates these cases. On superconductivity, ARC-SC's 70 wins consist of 41 ARC-SC-only discoveries and 29 cases in which both methods eventually hit but ARC-SC is earlier. On mixed JARVIS, the corresponding counts are 35 and 14.

A.6.4 Robustness Across Nominal Invalid Rates

Table 11 aggregates the grid by nominal invalid rate. On superconductivity, ARC-SC records more first-hit wins than losses at all three invalid-rate levels and reduces the invalid-evaluation rate by approximately 5.6–6.3 percentage points. Mixed JARVIS is balanced at the two lower invalid-rate levels and becomes more favorable in first-hit W/T/L at the highest level; these rows are descriptive and are not treated as separately powered hypothesis tests.

Table 9: Paired effects with hierarchical-bootstrap 95% confidence intervals. Negative first-hit and invalid-rate effects favor ARC-SC; positive cumulative-hit effects favor ARC-SC.

Dataset	$\Delta\widetilde{\tau}_{\mathrm{hit}}$	Mean ΔN_{hit}	Mean $\Delta r_{\mathrm{invalid}}$
Superconductivity	$-0.269\,[-0.539, -0.025]$	$+0.283\,[-0.064, 0.667]$	$-0.059\,[-0.082, -0.038]$
JARVIS CFID (mixed)	$-0.117\,[-0.297, 0.053]$	$+0.050\,[-0.386, 0.481]$	$-0.009\,[-0.042, 0.021]$

Table 10: Decomposition of the first-hit relation for the 360 cases per paper-facing dataset design.

Dataset	ARC only	ARC earlier	Same round	Both no hit	POF earlier	POF only
Superconductivity	41	29	42	209	16	23
JARVIS CFID (mixed)	35	14	107	165	17	22

Table 11: Results by nominal invalid rate. Each row aggregates four failure mechanisms, three initialization levels, and ten seeds (120 paired cases).

Dataset	Invalid rate	First-hit W/T/L	Hit W/T/L	ΔHits	Mean ΔInvalid
Superconductivity	0.2	32/77/11	39/67/14	$+53$	-0.063
Superconductivity	0.3	22/85/13	26/76/18	$+24$	-0.059
Superconductivity	0.4	16/89/15	22/81/17	$+25$	-0.056
JARVIS CFID (mixed)	0.2	14/92/14	27/67/26	-24	-0.008
JARVIS CFID (mixed)	0.3	15/90/15	26/65/29	-1	-0.011
JARVIS CFID (mixed)	0.4	20/90/10	35/65/20	$+43$	-0.009

A.6.5 Development-Stage Anchor-Count Ablation

The number of POF anchors was selected before the holdout-seed evaluation. A development-stage superconductivity ablation compares $K \in \{1, 2, 3\}$ with total batch size fixed at five. The 15 paired development configurations use feature-interaction and low-validity failures and development seed IDs rather than the seeds 5–14.

Table 12: Development-stage anchor-count ablation. These results were used to choose $K = 3$ and are not pooled into the statistics.

POF anchors K	Coverage slots	First-hit W/T/L	ΔHits
1	4	5/8/2	-7
2	3	4/10/1	$+11$
3	2	5/9/1	$+24$

The one-anchor variant increases diversification but loses cumulative hits relative to POF; the two- and three-anchor variants avoid this degradation, and $K = 3$ gives the largest cumulative-hit improvement in the development grid while retaining a low first-hit loss count. This supports the intended role of scenario coverage as a complement to strong marginal anchors rather than a complete replacement for marginal exploitation. Because the same development results informed the choice of K, Table 12 is a tuning/ablation result rather than independent evidence of generalization.

A.6.6 Development-Stage Candidate-Generation Robustness

A separate superconductivity development study tests whether favorable behavior depends entirely on access to global benchmark ranges. It uses only feature-interaction and low-validity failures, nominal invalid rates 0.2, 0.3, and 0.4, initialization quantiles 0.65, 0.75, and 0.85, and five development seeds, giving 90 paired cases per generator.

For the *observed-box* generator, the uniform bounds of equation 18 are recomputed from the currently observed designs. For the development-stage *observed trust-region* generator, candidates are perturbed around observed high-quality anchors and clipped to the observed padded bounds. This

development trust region is distinct from the final JARVIS generator in Appendix A.1.2; the latter uses valid observed anchors and an IQR-scaled 0.05 perturbation.

Table 13: Development-stage observed-only candidate-generation robustness on superconductivity. These runs are not pooled with the holdout-seed results.

Generator	Cases	First-hit W/T/L	ΔHits	Mean ΔInvalid
Observed box	90	31/42/17	$+105$	-0.094
Observed trust region	90	32/44/14	$+115$	-0.088

Although auxiliary, these results indicate that the superconductivity advantage is not unique to candidate pools constructed from full benchmark-level bounds.