THINK BEFORE YOU EXPERIMENT: EFFICIENT EXPERIMENT SELECTION FOR AI RESEARCH AGENTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Autonomous AI research agents improve solutions through repeated experimentation, but experiments are expensive while LLM inference is comparatively cheap. This asymmetry raises a simple question: which ideas are worth an experiment? Existing agents typically generate and execute candidate ideas sequentially, devoting little computation to deciding which candidate should be tested next. We introduce IdeaMAP, a lightweight plug-in that spends additional inference on experiment selection. IdeaMAP formulates selection as quality-diversity optimization over a semantic space of research ideas, maintaining a MAP-Elites-style archive of executed experiments and prioritizing candidates predicted either to improve existing elites or to fill promising unexplored niches. Across FML-bench-Lite, IdeaMAP improves multiple host research agents and yields agents that outperform existing systems including The AI Scientist v1 and v2, AIDE, and OpenEvolve. IdeaMAP reaches the host agent's final validation performance using 31% fewer validation experiments and, at the same validation budget, achieves 6% higher final validation performance. These results show that allocating additional inference to experiment selection can improve the efficiency and performance of autonomous research agents.

1 INTRODUCTION

Autonomous AI research agents improve machine learning solutions through repeated experimentation. They propose a change, implement it, evaluate the result, and use the outcome to decide what to try next (Lu et al., 2024; Yamada et al., 2025; Jiang et al., 2025). Each tested idea therefore consumes an experiment, often involving a costly training or validation run. As the search continues, the number of experiments an agent can afford constrains how much it can explore (Karpathy, 2026; Zou et al., 2026). At the same time, there is a substantial cost gap between reasoning about an experiment and actually running it. LLM inference can generate and compare several candidate ideas at relatively low cost, while evaluating each candidate may require much more wall-clock time and accelerator computation. Autoresearch, for example, spends a fixed five-minute GPU training run on every experiment (Karpathy, 2026). This asymmetry creates an opportunity to spend more inference before execution and use a limited experiment budget more selectively. It leads to the central question of this work: which ideas are worth an experiment?

Existing research agents use a variety of search strategies to generate and refine ideas (Yamada et al., 2025; Jiang et al., 2025; Novikov et al., 2025; Sharma, 2025). However, once an idea is selected for empirical evaluation, a portion of the limited experiment budget is spent even if the idea turns out to be weak or redundant. A natural alternative is to generate several ideas first and choose which one to test. One simple approach is to select the idea with the highest predicted quality. However, this can repeatedly favor similar directions that already appear promising (Si et al., 2025; Zhang et al., 2025). A slightly lower scoring idea may instead explore a different but still promising direction and make better use of the next experiment. Experiment selection should therefore consider not only how promising an idea appears, but also what has already been explored.

We introduce IdeaMAP, which casts experiment selection as quality-diversity optimization over a semantic space of research ideas (Mouret & Clune, 2015; Cully & Demiris, 2017). As shown in Figure 1(a), IdeaMAP groups executed ideas into semantic regions and keeps the best observed result in each region. Together, these regions form an archive that records both where the agent

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.

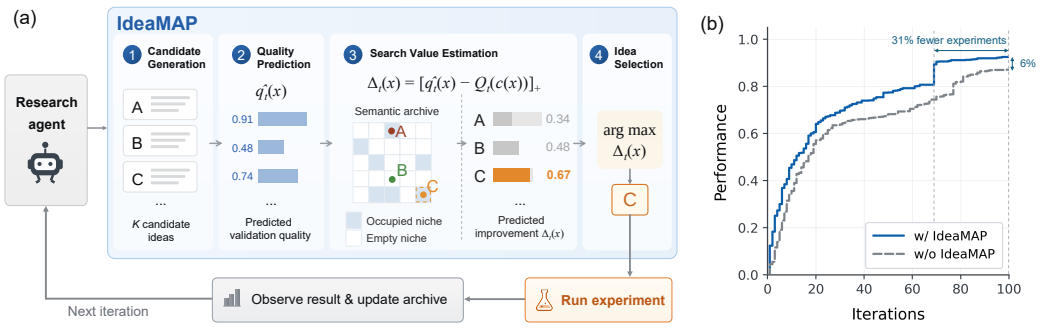


Figure 1: **IdeaMAP overview and budget efficiency. (a)** IdeaMAP selects among candidate ideas based on their predicted archive improvement in a semantic quality-diversity archive before committing the next experiment. **(b)** Task-averaged best-so-far validation performance over the search budget on FML-bench-Lite, comparing IdeaMAP (AR) with its Autoresearch host.

has explored and how well each region has performed. Given several new candidates, IdeaMAP predicts how much each one would improve the current archive and selects the candidate with the largest predicted archive improvement. A candidate can therefore be valuable either because it is expected to improve an explored region or because it opens a promising region that has not yet been explored. The archive guides which experiment to run next, while the agent still ultimately seeks the best final solution.

This formulation makes the value of an idea depend on the current state of the search. Instead of asking only which candidate looks best in isolation, IdeaMAP asks which candidate is expected to contribute most given what has already been discovered. This allows the agent to continue improving strong directions while reducing repeated experiments on semantically similar ideas. IdeaMAP only changes how the next experiment is selected, so it can be integrated into an existing research agent without changing how candidates are implemented or evaluated.

We evaluate IdeaMAP on FML-bench-Lite by integrating it with AdaptiveSearch and Autoresearch (Zou et al., 2026; Karpathy, 2026). Under the same validation budgets, IdeaMAP improves both host agents across aggregate evaluation metrics. When applied to Autoresearch, IdeaMAP reaches the host agent's final validation performance using 31% fewer validation experiments. With the same validation budget, it also achieves stronger final validation performance. These results show that additional inference before execution can make more effective use of a limited experiment budget.

Our contributions are:

- We formulate experiment selection as a distinct decision problem for autonomous research agents under limited empirical budgets, motivated by the cost of empirical evaluation and the finite number of experiments an agent can perform.
- We introduce IdeaMAP, a plug-in experiment-selection method that casts experiment selection as quality-diversity optimization over a semantic idea space and selects candidates by predicted archive improvement.
- We evaluate IdeaMAP on FML-bench-Lite, where its integration with Autoresearch and AdaptiveSearch improves both hosts across aggregate metrics and, with Autoresearch, reaches the host's final validation performance with 31% fewer validation experiments.

2 Related Work

Recent LLM-based systems automate substantial parts of scientific and algorithmic discovery, including idea generation, code modification, experimentation, and iterative refinement (Lu et al., 2024; Schmidgall et al., 2025; Romera-Paredes et al., 2024; Novikov et al., 2025). These systems show that LLMs can participate not only in proposing solutions, but also in repeatedly interacting with executable environments and empirical feedback. Research agents further employ sequential and tree-based search to organize candidate exploration and refinement (Yamada et al., 2025; Jiang

et al., 2025; Karpathy, 2026). Such search mechanisms determine how previous experimental outcomes influence subsequent proposals and which directions are explored further. Existing systems therefore provide different strategies for navigating the broader research process. Our work studies a distinct decision point within this process: how to allocate the next empirical execution when multiple candidate research ideas are available.

Quality-diversity research studies the discovery of diverse, high-performing solutions and has developed a range of archive-based formulations and optimization frameworks (Mouret & Clune, 2015; Pugh et al., 2016; Cully & Demiris, 2017; Chatzilygeroudis et al., 2021). A central idea in this literature is to preserve strong solutions across different regions of a descriptor space rather than representing search only through a single incumbent. Subsequent surrogate-assisted methods extend archive-based search to settings with costly evaluations (Gaier et al., 2017; Kent & Branke, 2020; Kent et al., 2023). This line of work provides a natural foundation for reasoning about which evaluations are most valuable under a limited budget. We build on this perspective to formulate experiment selection within autonomous research agents, using the archive to represent both explored research directions and their empirical quality.

LLMs have also been widely studied for scientific ideation, hypothesis generation, and research planning. Existing systems use literature grounding, critique, planning, retrieval, debate, and search to generate and refine research ideas and hypotheses (Baek et al., 2025; Hu et al., 2024; Gottweis et al., 2025), while empirical studies examine the novelty, diversity, and self-evaluation of LLM-generated ideas (Si et al., 2025). This literature primarily focuses on generating and refining research ideas. Our work instead focuses on selecting which proposed idea should receive an empirical evaluation under a limited experiment budget.

3 IDEAMAP

IdeaMAP decides which candidate research idea receives the next validation execution. When multiple candidate ideas are available, selecting the next experiment should consider not only predicted quality but also what has already been explored. Given a candidate set proposed by a surrounding research agent, IdeaMAP maps ideas to semantic cells and scores each candidate by its predicted marginal contribution to a quality-diversity archive. The surrounding agent remains responsible for parent selection, candidate generation, implementation, validation, and final solution selection.

3.1 SEMANTIC QUALITY-DIVERSITY SEARCH OBJECTIVE

Given a budget of T validation executions, we formulate the sequential exploration of research ideas as quality-diversity optimization over a semantic space. We use the QD-score to reward both exploring promising new directions and improving strong previously explored directions. Let A_t denote the archive after t validation executions, $\mathcal{C}$ the finite set of semantic cells, and $Q_t(c)$ the quality of the current elite in cell c. The archive value at step t is measured by the standard QD-score:

$$\mathrm{QD}(A_t) = \sum_{c \in \mathcal{C}} Q_t(c). \tag{1}$$

The QD-score sums the elite qualities across semantic cells, thereby rewarding both the discovery of previously empty cells and quality improvements within occupied cells. It is a standard archive objective in MAP-Elites and quality-diversity optimization (Mouret & Clune, 2015; Cully & Demiris, 2017; Chatzilygeroudis et al., 2021). After the fixed budget of T validation executions, the search objective is to maximize

$$\mathbb{E}[\mathrm{QD}(A_T)]\,, \tag{2}$$

where the expectation is taken over the stochasticity in candidate generation, solution implementation, and validation outcomes.

We use the textual semantics of a research idea as its behavior descriptor and the validation quality obtained after executing the idea as its solution quality. An idea describes the scientific hypothesis, modeling principle, or algorithmic modification adopted by a candidate solution. Its textual semantics can therefore group candidates with similar research content. Semantic representations are widely used to measure textual similarity (Reimers & Gurevych, 2019), and prior work shows

that LLM-generated research ideas may repeat similar concepts despite surface-level differences (Si et al., 2025).

Let $\mathcal{X}$ denote the space of natural-language research ideas. After t validation executions, the evaluated history is

$$\mathcal{D}_t = \{(x_i, q_i)\}_{i=1}^{t}, \tag{3}$$

where $x_i \in \mathcal{X}$ is the i-th executed idea and $q_i \in \mathbb{R}_{\geq 0}$ is its realized archive quality derived from the validation outcome, with larger values indicating greater improvement over the task baseline. At step t, the agent generates a finite candidate set

$$S_t = \left\{x^{(1)}, \ldots, x^{(K)}\right\} \subset \mathcal{X}, \tag{4}$$

from which one candidate is selected for the next validation execution.

Let ϕ be a text embedding function and π a discrete partition of the semantic representation space. The semantic cell assigned to a candidate x is

$$\phi : \mathcal{X} \to \mathcal{Z}, \qquad \pi : \mathcal{Z} \to \mathcal{C}, \qquad c(x) = \pi(\phi(x)). \tag{5}$$

For notational simplicity, we suppress the time dependence of the semantic mapping; in our implementation, the projection is updated as evaluated ideas accumulate, and selection at each step uses the current mapping. Each cell groups semantically similar research ideas, and the archive retains the evaluated idea with the highest validation quality in each cell. Assigning a quality of zero to an empty cell, the elite quality of cell c at step t is

$$Q_t(c) = \max\left(\{0\} \cup \{q_i \mid 1 \leq i \leq t,\ c(x_i) = c\}\right). \tag{6}$$

For an empty cell, $Q_t(c) = 0$. For an occupied cell, only an evaluated result that outperforms the incumbent elite increases $Q_t(c)$. Consequently, the first positive-quality result in an empty cell increases both archive coverage and the QD-score, whereas a result that fails to outperform the elite in the same cell contributes no additional archive value.

Substituting Equations equation 5 and equation 6 into Equation equation 1, the terminal search objective in Equation equation 2 can be written explicitly as

$$\mathbb{E}[\mathrm{QD}(A_T)] = \mathbb{E}\left[\sum_{c \in \mathcal{C}} \max\left(\{0\} \cup \{q_i \mid 1 \leq i \leq T,\ \pi(\phi(x_i)) = c\}\right)\right]. \tag{7}$$

The terminal QD-score is therefore the sum of the highest validation quality attained in each semantic cell. The finite validation budget can be allocated either to establishing a new elite in an empty cell or to improving the incumbent elite in an occupied cell. After search, the agent returns the evaluated solution with the highest validation quality.

3.2 Greedy Optimization through Predicted Archive Improvement

The terminal QD-score in Equation equation 7 decomposes into the archive improvements accumulated over individual search steps:

$$\mathrm{QD}(A_T) = \mathrm{QD}(A_t) + \sum_{s=t}^{T-1} \left[\mathrm{QD}(A_{s+1}) - \mathrm{QD}(A_s)\right]. \tag{8}$$

Taking expectations preserves this decomposition. Under a fixed semantic partition, selecting the candidate with the largest predicted one-step archive improvement greedily approximates the terminal objective. IdeaMAP uses this quantity as a local acquisition score under the current partition, without explicitly modeling how the current choice affects future candidate generation or subsequent archive improvements.

Let $\hat{q}_t(x) \in \mathbb{R}_{\geq 0}$ denote the LLM-predicted archive-quality score for candidate x. The predictor conditions on previously evaluated ideas and their observed archive qualities and provides a heuristic estimate of candidate quality. Under the current semantic partition, the predicted contribution of x depends on its predicted quality $\hat{q}_t(x)$ relative to the incumbent elite of its assigned cell. This yields its predicted archive improvement:

$$\begin{aligned} \Delta_t(x) &= \max\left\{Q_t(c(x)), \hat{q}_t(x)\right\} - Q_t(c(x)) \\ &= \left[\hat{q}_t(x) - Q_t(c(x))\right]_+, \end{aligned} \tag{9}$$

Algorithm 1 IdeaMAP greedy optimization through predicted archive improvement

Require: History $\mathcal{D}_t$; archive A_t; candidate set S_t
Require: Predicted qualities $\{\hat{q}_t(x) : x \in S_t\}$; semantic cell map $c(\cdot)$
Ensure: Selected idea x_{t+1}; updated history $\mathcal{D}_{t+1}$ and archive A_{t+1}
1: **for** each candidate $x \in S_t$ **do**
2: $\quad c_x \leftarrow c(x)$
3: $\quad \Delta_t(x) \leftarrow [\hat{q}_t(x) - Q_t(c_x)]_+$
4: **end for**
5: $x_{t+1} \leftarrow \arg\max_{x \in S_t} \Delta_t(x)$
6: Execute x_{t+1} with the surrounding agent and observe q_{t+1}.
7: $\mathcal{D}_{t+1} \leftarrow \mathcal{D}_t \cup \{(x_{t+1}, q_{t+1})\}$
8: $A_{t+1} \leftarrow \textsc{UpdateArchive}(A_t, x_{t+1}, q_{t+1}, c(\cdot))$
9: **return** $x_{t+1}, A_{t+1}, \mathcal{D}_{t+1}$

where $[a]_+ = \max\{0, a\}$. For an empty cell, $Q_t(c(x)) = 0$, and the candidate is assigned its full predicted quality $\hat{q}_t(x)$. For an occupied cell, the candidate receives only its predicted margin over the incumbent elite.

Under fixed candidate sets, qualities, and semantic partition, the resulting archive utility has the standard diminishing-returns structure of QD objectives: as the elite quality of a cell increases, the marginal gain obtainable from additional candidates assigned to that cell cannot increase. IdeaMAP uses predicted marginal archive improvement as a practical greedy approximation to this objective.

IdeaMAP executes the candidate with the largest predicted archive improvement:

$$x_{t+1} = \arg\max_{x \in S_t} \Delta_t(x). \tag{10}$$

After executing x_{t+1} and observing its realized quality q_{t+1}, the record (x_{t+1}, q_{t+1}) is added to the evaluated history. The archive is then updated according to Equation equation 6, retaining the highest observed quality in each semantic cell.

Algorithm 1 summarizes one IdeaMAP selection step and archive update.

4 Experiments

4.1 Setup

We evaluate on FML-bench-Lite, the official eight-task subset of FML-bench (Zou et al., 2026). We compare against The AI Scientist v1/v2 (TASv1/TASv2) (Lu et al., 2024; Yamada et al., 2025), AIDE (Jiang et al., 2025), AIRA-MCTS (AIRA) (Toledo et al., 2026), OpenEvolve (OpenEvol) (Sharma, 2025), AdaptiveSearch (AdaptiveS) (Zou et al., 2026), and Autoresearch (AutoRes) (Karpathy, 2026). We integrate IdeaMAP with AdaptiveSearch and Autoresearch, denoted IdeaMAP (AS) and IdeaMAP (AR), respectively. In tables, AdaptiveS$_{\text{IM}}$ and AutoRes$_{\text{IM}}$ denote the corresponding IdeaMAP-augmented systems.

Each agent runs three independent 100-step searches per task using GPT-5.4 (OpenAI, 2026). During search, agents observe only validation feedback; after the budget is exhausted, the solution with the best validation performance is evaluated on the held-out test split. We report task-averaged FML-bench normalized improvement (NI), pairwise win rate, and Friedman mean rank, using both Mean-of-3 and Best-of-3 views. Metric definitions and reporting details are provided in Appendix A.

We additionally evaluate on seven CPU-compatible BioML-bench tasks with released public leaderboards (Miller et al., 2025), spanning drug discovery, protein engineering, and single-cell omics. Each agent runs three independent 100-step searches using Gemini-3.5-Flash-Lite (Google DeepMind, 2026). We report Leaderboard Percentile, Above Median, and Any Medal, averaged across the three repeats with SEM. Full task and metric details are provided in Appendix A.

Table 1: Results on FML-bench-Lite. Green deltas below each IdeaMAP-equipped row show the change relative to the corresponding host.

Agent	Mean-of-3			Best-of-3		
	NI Score ↑	Win Rate ↑	Mean Rank ↓	NI Score ↑	Win Rate ↑	Mean Rank ↓
TASv1	0.0765±.0039	0.297	7.00	0.0945	0.312	6.62
TASv2	0.1602±.0288	0.547	4.62	0.1924	0.438	5.31
AIDE	0.1653±.0172	0.495	5.25	0.2047	0.531	4.56
AIRA	0.1143±.0633	0.250	7.25	0.1841	0.234	7.00
OpenEvol	0.1426±.0076	0.531	4.62	0.1913	0.500	4.81
AdaptiveS	0.1811±.0222	0.562	4.00	0.2118	0.547	4.62
w/ IdeaMAP (Ours)	**0.1844**±.0378	0.604	3.88	**0.2356**	0.594	4.25
(Δ vs. host)	(+.0033)	(+.042)	(-.12)	(+.0238)	(+.047)	(-.37)
AutoRes	0.1598±.0090	0.531	4.62	0.1835	0.594	4.25
w/ IdeaMAP (Ours)	0.1720±.0249	**0.635**	**3.75**	0.2132	**0.656**	**3.56**
(Δ vs. host)	(+.0122)	(+.104)	(-.87)	(+.0297)	(+.062)	(-.69)

4.2 Implementation Details

At each IdeaMAP selection step, the host agent generates $K=2$ candidate ideas and IdeaMAP selects one for execution. Candidates are generated using the host's proposal mechanism, with an additional instruction encouraging the second candidate to propose a different idea. For FML-bench-Lite, realized archive quality is the executed idea's validation NI, while candidate quality is predicted by the backbone LLM using previously evaluated ideas and their observed outcomes as reference points.

We construct the semantic archive from text embeddings projected into a discretized low-dimensional semantic space. As evaluated ideas accumulate, the projection is refit and the archive is rebuilt under the updated mapping. Full embedding, projection, normalization, and grid configurations are provided in Appendix A.

For Autoresearch, IdeaMAP is applied whenever the agent proposes a new idea. For AdaptiveSearch, IdeaMAP is applied to improvement proposals in its second search phase; the remainder of each host's search procedure is unchanged.

4.3 Evaluation on FML-bench-Lite

Table 1 compares nine agents on FML-bench-Lite in aggregate. Both IdeaMAP variants improve over their respective hosts on all aggregate metrics, and first place across the six statistics is held by an IdeaMAP variant in every case. Among the two hosts, IdeaMAP (AR) shows the larger host-relative improvement across all six aggregate statistics; its mean-of-3 pairwise win rate rises from 0.531 to 0.635, while its Friedman mean rank improves from 4.62 to 3.75. IdeaMAP (AS) achieves the highest task-averaged NI overall (best-of-3: 0.2356). The two variants show complementary strengths: IdeaMAP (AS) leads in task-averaged NI, while IdeaMAP (AR) leads in pairwise win rate and Friedman mean rank.

Detailed task-level results are reported in Appendix Table 6. The largest gains over the corresponding host appear on DomainBed-CM and Opacus: on DomainBed-CM, IdeaMAP (AS) reaches a mean NI of 0.3441, up from 0.3003 for AdaptiveSearch; on Opacus, IdeaMAP (AR) reaches a mean NI of 0.0840, up from 0.0540 for Autoresearch. Both variants score slightly below their hosts on USB and, in mean NI, on PrivacyMeter.

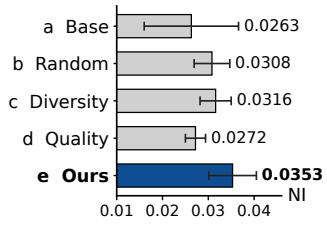
4.4 Budget Efficiency

Figure 1(b) compares the task-averaged best-so-far validation performance of IdeaMAP (AR) and its Autoresearch host over the 100-step search budget on FML-bench-Lite. Before averaging across tasks, the two trajectories for each task are normalized by the maximum final validation NI across the two agents. The terminal performance that Autoresearch reaches only with the full 100-step budget is reached by IdeaMAP (AR) at step 69, corresponding to 31% fewer validation experiments. At equal budget (step 100), IdeaMAP (AR) also achieves a 6% higher terminal value than its host. This

Table 2: Results on the evaluated BioML-bench tasks.

Agent	Leaderboard Percentile ↑	Above Median (%) ↑	Any Medal (%) ↑
AdaptiveS	38.21±5.96	28.6±8.2	28.6±8.2
AdaptiveS$_{IM}$	35.40±1.56	28.6±0.0	28.6±0.0
AutoRes	34.67±6.14	33.3±12.6	28.6±8.2
AutoRes$_{IM}$	**41.31**±3.82	**44.4**±6.9	**34.9**±4.2

Table 3: **Selection-rule ablation.** Left: NI performance across selection rules. Right: implementation and archive diagnostics. "Idea Uniq." denotes Idea Uniqueness.



Selection	Bug Rate	Fill Rate	Idea Uniq.	Q-D Score
a Base (no selection)	0.227±.034	0.0629±.0040	0.3743±.0258	0.6170±.1894
b Random	0.337±.034	0.0709±.0049	0.4877±.0475	0.6163±.0459
c Diversity	0.367±.046	0.0661±.0098	0.4945±.0426	0.6164±.0520
d Quality (LLM-as-Judge)	0.327±.090	0.0661±.0067	0.4473±.0607	0.6478±.1295
e IdeaMAP Q-D Selection	0.283±.095	0.0629±.0098	0.3976±.0328	0.6519±.0742

pattern aligns with the design motivation of IdeaMAP, which invests additional LLM inference at each step to generate and filter candidate ideas before committing the next validation execution.

4.5 Evaluation on BioML-bench

Table 2 reports results on the evaluated BioML-bench tasks. When applied to Autoresearch, IdeaMAP improves all three metrics, with Above Median rising from 33.3% to 44.4%, consistent with the gains on FML-bench-Lite. AdaptiveSearch with and without IdeaMAP perform comparably. One possible explanation is that AdaptiveSearch's phase-transition thresholds were calibrated on FML-bench tasks and may trigger less reliably on BioML-bench, limiting the opportunities for IdeaMAP to act in the second-phase search.

4.6 Ablation Studies

Selection rule. Table 3 compares five matched idea selection rules on PyCIL. Base (AdaptiveSearch) generates a single candidate idea per step and executes it directly. Rows b through e generate two candidate ideas per step and select one for execution, with all other experimental conditions unchanged. Random Selection chooses uniformly at random, Diversity Selection chooses the candidate farthest from its nearest archive elite in the semantic behavior space, Quality Selection chooses the candidate with the highest LLM-predicted quality, and IdeaMAP Q-D Selection uses the same quality estimate but ranks candidates by predicted archive improvement (Section 3.2). Beyond NI, Table 3 reports bug rate (fraction of steps without a valid result) and three metrics computed on a shared 625-cell semantic grid, namely fill rate (fraction of occupied cells), idea uniqueness (the number of semantic clusters normalized by the number of executed ideas), and Q-D score (sum of elite quality in each occupied cell, with quality measured on the validation split). All reported ablation statistics are mean ± standard deviation across the three repeats.

IdeaMAP Q-D Selection reaches the highest NI (0.0353) while matching Base for the lowest fill rate and achieving the highest Q-D score, a combination broadly consistent with concentrating budget on high-value semantic niches rather than pursuing coverage. Quality Selection achieves a comparable Q-D score but does not appear to translate this archive-level advantage into higher test NI. Diversity Selection produces the most semantically diverse set of executed ideas and also exhibits the highest bug rate (0.367). Random Selection reaches the highest fill rate but the lowest Q-D score. All four selection methods show higher bug rates than Base, with IdeaMAP having the lowest bug rate among the four. The NI ordering and the diagnostic patterns are consistent with the benefit of archive-guided selection on this task.

Table 4: **Semantic-archive sensitivity.** (a) Grid resolution. (b) Idea embedding model.

Res.	NI	Res.	NI
1	0.0272	7	**0.0887**
3	0.0236	9	0.0328
5	0.0353	10	0.0179

(a) Grid resolution.

Model	NI
text-embedding-3-large	**0.0353**
text-embedding-3-small	0.0337
text-embedding-ada-002	0.0293

(b) Embedding model.

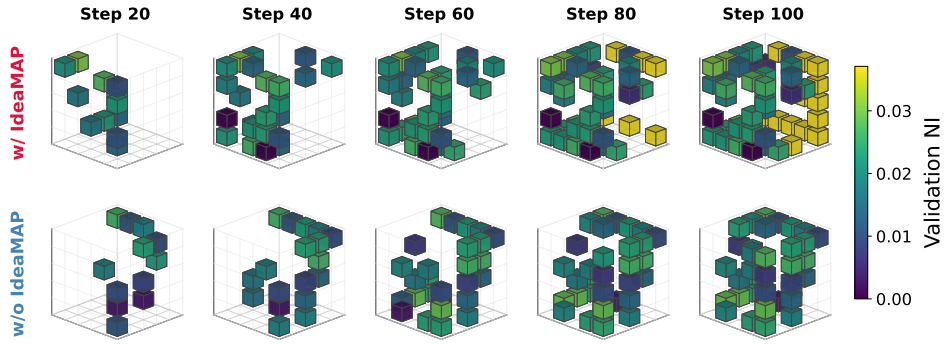


Figure 2: **Qualitative Q-D trajectory.** Archive snapshots across the search compare IdeaMAP (top) with its Autoresearch host (bottom). Color denotes the elite validation NI of each occupied cell.

Grid resolution and embedding model. Table 4 varies the two hyperparameters of the semantic archive on PyCIL, the grid resolution and the idea embedding model. Each entry reports NI averaged over the three repeats. Across grid resolutions, NI broadly follows an inverted-U pattern in which intermediate resolutions (5, 7, 9) exceed both extremes, with the peak at resolution 7. The two extremes degenerate by construction. At resolution 1, the entire idea space collapses into a single cell and the archive ranking reduces to Quality Selection. At resolution 10 (10^4 cells), nearly every candidate maps to an empty cell, so the occupied-cell discount rarely engages. This pattern is consistent with the expected behavior of the niche discount mechanism, which requires enough cells to provide semantic discrimination without so many that the discount becomes vacuous. Across three embedding models, NI ranges from 0.0293 to 0.0353 and all three values exceed the archive-blind baselines in Table 3 (Base at 0.0263, Quality Selection at 0.0272), indicating that the archive mechanism remains effective across the three tested embedding models.

LLM backbone. To examine whether the gains from IdeaMAP hold with other LLMs, we tested Qwen3.8-Flash (Alibaba Cloud, 2026) and Gemini-3.5-Flash-Lite (Google DeepMind, 2026) on DomainBed-CM, which exhibits variation across agents in the task-level results (Appendix Table 6). IdeaMAP (AS) obtains the highest mean NI on all three backbones, and IdeaMAP (AR) also improves over Autoresearch on GPT-5.4 and Gemini-3.5-Flash-Lite, with mean NI rising from 0.0466 to 0.1661 on the latter. On Qwen3.8-Flash, IdeaMAP (AR) scores slightly below Autoresearch in mean NI, though the host agents and their IdeaMAP counterparts show smaller performance differences on this backbone. These results show that the effect is observed across multiple backbones but varies by host and model on this task.

4.7 Acquisition Analysis

To examine the signal quality of the acquisition score at the step level, we analyzed the execution logs from the IdeaMAP (AR) main experiments across all eight tasks. On steps that produced a valid execution outcome, the mean within-run Spearman correlation between the LLM quality prediction $\hat{q}$ and the realized validation quality q is 0.205 across the 24 task-by-repeat runs. For predicted versus reconstructed realized local archive improvement, the mean within-run correlation is 0.380. The paired difference is 0.175 on average (95% CI: 0.108–0.242; paired t-test $p < 0.001$) and is positive in 22 of 24 runs and 7 of 8 tasks. This pattern is consistent with the design of the archive

Table 5: LLM backbone sensitivity.

LLM	Statistic	AutoRes		AdaptiveS	
		Host	w/ IdeaMAP	Host	w/ IdeaMAP
GPT-5.4	Mean	0.0672	0.1575	0.3003	**0.3441**
	Best	0.1289	0.3765	0.4395	**0.6140**
Qwen3.8-Flash	Mean	0.4561	0.4520	0.4587	**0.5099**
	Best	0.4593	0.4646	0.4634	**0.6141**
Gemini-3.5-Flash-Lite	Mean	0.0466	0.1661	0.1546	**0.2168**
	Best	0.0534	0.2095	**0.2168**	**0.2168**

term. By referencing the elite quality of the candidate's cell, the acquisition score tends to rank candidates by their local marginal archive contribution more reliably than the raw quality prediction alone.

Beyond signal quality, we also examined how often the archive term affects selection decisions in practice. Across all acquisition-enabled steps in the IdeaMAP (AR) runs on eight tasks, Q-D Selection and Quality Selection chose different candidates in 21.9% of steps, where the archive term adjusted the quality-only ranking. Using the PyCIL ablation (Table 3, rows d and e) as a case study, we compared the outcomes on their respective disagreement steps. Among the 24 disagreement steps in the Q-D Selection runs, 63% resulted in an archive update, where the realized quality exceeded the current elite in that cell. For the 21 disagreement steps in the Quality Selection runs, this rate was 33%. On steps where the two rules diverged, archive-guided selection more frequently produced local archive improvements, consistent with the Q-D Selection ablation results.

4.8 Quality–Diversity Trajectory Analysis

We visualize the archive evolution as a case study to examine how archive-guided selection affects exploration behavior. Figure 2 compares one run of IdeaMAP (AR) against its Autoresearch host on DomainBed-OH over a matched 100-step validation budget. For visualization, we fit a fixed semantic space separately to each complete run and replay the evaluated ideas over the search. Within these run-specific spaces, by the end of the search, IdeaMAP (AR) has filled 70 out of 625 semantic cells compared to 48 for the host, and the corresponding QD-scores are 1.57 and 0.85 respectively. The snapshot sequence shows increasing coverage over the course of both runs, with IdeaMAP ending with more occupied cells. Later snapshots of IdeaMAP also exhibit more bright-yellow regions, corresponding to cells with higher elite validation NI. In this run, IdeaMAP ends with more high-quality occupied cells in its run-specific semantic space. The test NI of this run is consistent with the archive-level trend (0.0153 for IdeaMAP (AR) vs 0.0047 for the host).

4.9 Token Usage and API Cost

Averaged over the eight FML-bench-Lite tasks, IdeaMAP (AS) consumes approximately 2.80M input tokens and 0.31M output tokens per 100-step run, compared with 1.90M and 0.29M for AdaptiveSearch. IdeaMAP (AR) consumes approximately 3.70M input tokens and 0.36M output tokens, compared with 1.89M and 0.31M for Autoresearch. Under GPT-5.4 pricing, the API costs of IdeaMAP (AS) and IdeaMAP (AR) are approximately 1.22× and 1.50× those of their respective hosts.

5 Conclusion

We study how autonomous research agents should select experiments under limited empirical budgets. We introduce IdeaMAP, a plug-in method that uses semantic quality-diversity optimization to prioritize candidate ideas before empirical evaluation. Across FML-bench-Lite, IdeaMAP improves both Autoresearch and AdaptiveSearch, and with Autoresearch reaches the host's final validation performance using 31% fewer validation experiments. These results suggest that autonomous research can benefit not only from generating better ideas, but also from allocating scarce empirical evaluations more selectively.

AI USE STATEMENT

In this work, we used generative AI tools for manuscript editing and literature search. AI-assisted manuscript content was reviewed by the authors; references identified with AI assistance were verified against the original sources, and reported results were checked against the underlying experimental records. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

To support reproducibility, we provide detailed descriptions of the IdeaMAP method, experimental setup, and evaluation protocol, together with the implementation details and configurations used in our experiments.

REFERENCES

Alibaba Cloud. Qwen3.8-Flash, 2026. https://www.alibabacloud.com/help/en/model-studio/qwen3-8-flash.

Jinheon Baek, Sujay Kumar Jauhar, Silviu Cucerzan, and Sung Ju Hwang. Researchagent: Iterative research idea generation over scientific literature with large language models. In *Proceedings of the 2025 conference of the nations of the Americas chapter of the association for computational linguistics: human language technologies (volume 1: long papers)*, pp. 6709–6738, 2025.

Konstantinos Chatzilygeroudis, Antoine Cully, Vassilis Vassiliades, and Jean-Baptiste Mouret. Quality-diversity optimization: a novel branch of stochastic optimization. In *Black box optimization, machine learning, and no-free lunch theorems*, pp. 109–135. Springer, 2021.

Antoine Cully and Yiannis Demiris. Quality and diversity optimization: A unifying modular framework. *IEEE Transactions on Evolutionary Computation*, 22(2):245–259, 2017.

Adam Gaier, Alexander Asteroth, and Jean-Baptiste Mouret. Data-efficient exploration, optimization, and modeling of diverse designs through surrogate-assisted illumination. In *Proceedings of the Genetic and Evolutionary Computation Conference*, pp. 99–106, 2017.

Google DeepMind. Gemini 3.5 flash-lite, July 2026. https://deepmind.google/models/model-cards/gemini-3-5-flash-lite/.

Juraj Gottweis, Wei-Hung Weng, Alexander Daryin, Tao Tu, Anil Palepu, Petar Sirkovic, Artiom Myaskovsky, Felix Weissenberger, Keran Rong, Ryutaro Tanno, et al. Towards an ai co-scientist. *arXiv preprint arXiv:2502.18864*, 2, 2025.

Xiang Hu, Hongyu Fu, Jinge Wang, Yifeng Wang, Zhikun Li, Renjun Xu, Yu Lu, Yaochu Jin, Lili Pan, and Zhenzhong Lan. Nova: An iterative planning and search approach to enhance novelty and diversity of llm generated ideas. *arXiv preprint arXiv:2410.14255*, 2024.

Zhengyao Jiang, Dominik Schmidt, Dhruv Srikanth, Dixing Xu, Ian Kaplan, Deniss Jacenko, and Yuxiang Wu. Aide: Ai-driven exploration in the space of code. *arXiv preprint arXiv:2502.13138*, 2025.

Andrej Karpathy. autoresearch: Ai agents running research on single-gpu nanochat training automatically, 2026. URL https://github.com/karpathy/autoresearch.

Paul Kent and Juergen Branke. Bop-elites, a bayesian optimisation algorithm for quality-diversity search. *arXiv preprint arXiv:2005.04320*, 2020.

Paul Kent, Adam Gaier, Jean-Baptiste Mouret, and Juergen Branke. Bop-elites, a bayesian optimisation approach to quality diversity search with black-box descriptor functions. *arXiv preprint arXiv:2307.09326*, 2023.

Chris Lu, Cong Lu, Robert Tjarko Lange, Jakob Foerster, Jeff Clune, and David Ha. The ai scientist: Towards fully automated open-ended scientific discovery. *arXiv preprint arXiv:2408.06292*, 2024.

Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018.

Henry E Miller, Matthew Greenig, Benjamin Tenmann, and Bo Wang. Bioml-bench: Evaluation of ai agents for end-to-end biomedical ml. *bioRxiv*, pp. 2025–09, 2025.

Jean-Baptiste Mouret and Jeff Clune. Illuminating search spaces by mapping elites. *arXiv preprint arXiv:1504.04909*, 2015.

Alexander Novikov, Ngân Vũ, Marvin Eisenberger, Emilien Dupont, Po-Sen Huang, Adam Zsolt Wagner, Sergey Shirobokov, Borislav Kozlovskii, Francisco JR Ruiz, Abbas Mehrabian, et al. Alphaevolve: A coding agent for scientific and algorithmic discovery. *arXiv preprint arXiv:2506.13131*, 2025.

OpenAI. New embedding models and API updates. `https://openai.com/index/new-embedding-models-and-api-updates/`, January 2024.

OpenAI. Introducing GPT-5.4. `https://openai.com/index/introducing-gpt-5-4/`, March 2026.

Justin K Pugh, Lisa B Soros, and Kenneth O Stanley. Quality diversity: A new frontier for evolutionary computation. *Frontiers in Robotics and AI*, 3:40, 2016.

Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pp. 3982–3992, 2019.

Bernardino Romera-Paredes, Mohammadamin Barekatain, Alexander Novikov, Matej Balog, M Pawan Kumar, Emilien Dupont, Francisco JR Ruiz, Jordan S Ellenberg, Pengming Wang, Omar Fawzi, et al. Mathematical discoveries from program search with large language models. *Nature*, 625(7995):468–475, 2024.

Samuel Schmidgall, Yusheng Su, Ze Wang, Ximeng Sun, Jialian Wu, Xiaodong Yu, Jiang Liu, Michael Moor, Zicheng Liu, and Emad Barsoum. Agent laboratory: Using llm agents as research assistants. *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 5977–6043, 2025.

Asankhaya Sharma. Openevolve: an open-source evolutionary coding agent, 2025. URL `https://github.com/algorithmicsuperintelligence/openevolve`.

Chenglei Si, Diyi Yang, and Tatsunori Hashimoto. Can llms generate novel research ideas? a large-scale human study with 100+ nlp researchers. In *International Conference on Learning Representations*, volume 2025, pp. 94003–94092, 2025.

Edan Toledo, Karen Hambardzumyan, Martin Josifoski, Rishi Hazra, Nicolas Baldwin, Alexis Audran-Reiss, Michael Kuchnik, Despoina Magka, Minqi Jiang, Alisia Lupidi, et al. Ai research agents for machine learning: Search, exploration, and generalization in mle-bench. *Advances in Neural Information Processing Systems*, 38:35309–35348, 2026.

Yutaro Yamada, Robert Tjarko Lange, Cong Lu, Shengran Hu, Chris Lu, Jakob Foerster, Jeff Clune, and David Ha. The ai scientist-v2: Workshop-level automated scientific discovery via agentic tree search. *arXiv preprint arXiv:2504.08066*, 2025.

Yiming Zhang, Harshita Diddee, Susan Holm, Hanchen Liu, Xinyue Liu, Vinay Samuel, Barry Wang, and Daphne Ippolito. Noveltybench: Evaluating creativity and diversity in language models. *arXiv preprint arXiv:2504.05228*, 2025.

Qiran Zou, Hou Hei Lam, Wenhao Zhao, Tingting Chen, Yiming Tang, Samson Yu, Yingtao Zhu, Srinivas Anumasa, Zufeng Zhang, Tianyi Zhang, et al. Fml-bench: A controlled study of ai research agent strategies from the perspective of search dynamics. *arXiv preprint arXiv:2605.17373*, 2026.

A Experimental and Implementation Details

For FML-bench-Lite, the official eight-task subset of FML-bench (Zou et al., 2026), we evaluate on PyCIL, USB, DomainBed-CM, DomainBed-OH, OpenOOD, Opacus, PrivacyMeter, and ART. We use FML-bench normalized improvement (NI) as the primary performance metric. NI measures improvement over the task baseline, normalized by the range between the task-specific best and worst reference values, with the metric direction accounted for and negative improvement clipped to zero. For aggregate evaluation, NI is averaged equally across the eight tasks. Win Rate is the strict pairwise win rate across agent pairs. Mean Rank is the Friedman mean rank, where lower values indicate stronger performance. Mean-of-3 averages the three repeats; Best-of-3 takes the best repeat. In tables, these two views are denoted Mean and Best, respectively. For Mean-of-3 NI in the main FML-bench-Lite results, we additionally report the standard deviation across the three repeats. Each agent runs three repeats of 100 validation steps per task. During search, agents observe only validation feedback. After the 100-step budget is exhausted, the solution with the best validation performance is selected and evaluated on the held-out test split. Unless stated otherwise, reported NI is computed on this test evaluation. GPT-5.4 is the LLM backbone used in the main experiments (OpenAI, 2026).

For BioML-bench (Miller et al., 2025), we evaluate on seven tasks from the CPU-compatible subset of representative tasks that provide released public leaderboards: `drug_bbb`, `drug_cyp2d6`, `drug_solu`, `protein_sbi`, `protein_spike`, `sc_modality`, and `sc_perturbation`. These tasks span drug discovery, protein engineering, and single-cell omics. Each agent runs three independent repeats of 100 validation steps with Gemini-3.5-Flash-Lite as the backbone (Google DeepMind, 2026). Performance is measured against BioML-bench's released leaderboards by Leaderboard Percentile, defined as the fraction of entries strictly below the submission; Above Median, defined as the fraction of tasks exceeding the median; and Any Medal, defined as the fraction of tasks reaching a medal threshold. Following the reporting convention used for BioML-bench evaluations (Miller et al., 2025), we report each metric as the mean across the three independent repeats, with uncertainty given by the standard error of the mean (SEM) across repeats.

For IdeaMAP implementation, at each selection step we generate K candidate ideas using the surrounding agent's proposal mechanism and select one for execution. The first candidate follows the host agent's original proposal mechanism; subsequent candidates use the same mechanism with an additional instruction to propose a different idea. Unless otherwise noted, $K{=}2$ in all experiments. For FML-bench-Lite, the realized archive quality q is the single-step validation NI of the executed idea: the direction-corrected improvement over the task baseline, normalized by the benchmark reference range and clipped below at zero. Candidate quality estimates $\hat{q}$ are produced by the backbone LLM using previously evaluated ideas together with their observed validation metrics and corresponding archive qualities as reference points. For the semantic archive, we embed each idea with text-embedding-3-large and use UMAP (McInnes et al., 2018) to project the embeddings into a four-dimensional behavior space. As evaluated ideas accumulate, we refit the projection and rebuild the archive under the updated mapping, allowing the representation to adapt to the growing set of observed ideas. The projected coordinates are robustly normalized and discretized into a uniform grid with five bins per dimension, yielding $5^4 = 625$ cells. The hyperparameter sensitivity study in Table 4 additionally evaluates text-embedding-3-small and text-embedding-ada-002; all three embedding models are from OpenAI (OpenAI, 2024). For Autoresearch, IdeaMAP is applied whenever the agent proposes a new idea. For AdaptiveSearch, IdeaMAP is applied to improvement proposals in its second search phase; the remainder of each host's search procedure is unchanged.

B Task-level Results on FML-bench-Lite

Table 6 reports test NI for each of the eight FML-bench-Lite tasks. Mean and Best denote the Mean-of-3 and Best-of-3 views, respectively; the final two rows average these quantities equally across tasks. AdaptiveS$_{\text{IM}}$ and AutoRes$_{\text{IM}}$ denote AdaptiveSearch and Autoresearch augmented with IdeaMAP, respectively.

Table 6: Task-level NI on FML-bench-Lite.

Task	Statistic	AdaptiveS	AIDE	AIRA	AutoRes	OpenEvol	TASv1	TASv2	AdaptiveS$_{IM}$	AutoRes$_{IM}$
PyCIL	Mean	0.0263	0.0564	0.0201	0.0261	0.0370	0.0302	**0.0639**	0.0353	0.0253
	Best	0.0402	0.0617	0.0275	0.0342	0.0609	0.0597	**0.0676**	0.0417	0.0302
USB	Mean	0.0443	0.0293	0.0188	0.0478	**0.0671**	0.0398	0.0386	0.0341	0.0469
	Best	0.0484	0.0379	0.0239	**0.0807**	0.0794	0.0474	0.0479	0.0366	0.0670
DomainBed-CM	Mean	0.3003	0.2306	0.1650	0.0672	0.2425	0.0713	0.2448	**0.3441**	0.1575
	Best	0.4395	0.4379	0.4080	0.1289	0.4544	0.1007	0.3924	**0.6140**	0.3765
DomainBed-OH	Mean	0.0084	0.0108	0.0116	0.0047	0.0126	0.0065	**0.0137**	0.0088	0.0112
	Best	0.0141	**0.0153**	0.0135	0.0082	**0.0153**	0.0094	**0.0153**	0.0118	**0.0153**
OpenOOD	Mean	0.0488	0.0119	0.0213	**0.0571**	0.0064	0.0230	0.0392	0.0496	0.0568
	Best	0.0548	0.0347	0.0236	**0.0698**	0.0191	0.0610	0.0562	0.0578	0.0588
Opacus	Mean	0.0461	0.0059	0.0000	0.0540	0.0031	0.0000	0.0000	0.0573	**0.0840**
	Best	0.0636	0.0177	0.0000	0.0704	0.0051	0.0000	0.0000	0.0820	**0.1006**
PrivacyMeter	Mean	0.5130	0.5311	0.3749	**0.5734**	0.3130	0.0123	0.4871	0.4845	0.5339
	Best	0.5492	0.5556	0.5745	**0.5977**	0.4306	0.0239	0.5439	0.5530	0.5684
ART	Mean	**0.4614**	0.4465	0.3028	0.4481	0.4590	0.4289	0.3945	0.4612	0.4601
	Best	0.4846	0.4765	0.4019	0.4783	0.4657	0.4539	0.4160	0.4883	**0.4885**
Average Mean		0.1811	0.1653	0.1143	0.1598	0.1426	0.0765	0.1602	**0.1844**	0.1720
Average Best		0.2118	0.2047	0.1841	0.1835	0.1913	0.0945	0.1924	**0.2356**	0.2132