HypoEvolve: GENETIC ALGORITHMS ENABLE MULTI-AGENT LLMS TO DISCOVER SCIENTIFIC HYPOTHESES

Anonymous authors
Paper under double-blind review

ABSTRACT

Scientific agents increasingly contribute to hypothesis discovery by synthesizing evidence, assessing proposals, and developing new explanations. Recent systems bring scientific agents and evolutionary search together to develop hypotheses through cycles of critique, comparison, and revision. However, how different forms of agent collaboration affect hypothesis quality remains an open question. Answering this question requires separating the effects of agents' scientific capabilities from those of their collaboration. A suitable framework must therefore preserve the agents' scientific roles and support different rules for combining, revising, and retaining hypotheses. Building on this perspective, we introduce HypoEvolve, which makes collaboration explicit through successive updates to a hypothesis population. Specifically, we propose to use a generational genetic algorithm to coordinate specialized large language model (LLM) agents that integrate mechanistic arguments, reconsider assumptions, and assess evidence and testability. Each generation specifies how scientific judgments and new proposals reshape the population, which makes the effects of collaboration on hypothesis quality directly testable. Moreover, we design our evaluation around scientifically meaningful hypotheses that explain how a proposed intervention could work. Drug repurposing connects these explanations to target-level biological claims that can be assessed against external evidence. Specifically, we adapt DepMap and Open Targets into complementary external measures grounded in experimental, genetic, and clinical evidence. The evaluation spans 34 cancer types, with HypoEvolve achieving the highest scores against six baselines on both measures. DepMap selectivity reaches 0.171, compared with 0.115 for the strongest baseline. Gains over single-pass generation also generalize to held-out cancer types. HypoEvolve advances a vision of autonomous science in which AI research teams achieve a capacity for discovery beyond that of individual models.

1 INTRODUCTION

Large language models (LLMs) are enabling scientific agents to formulate hypotheses that connect existing evidence to new research directions (Wang et al., 2024; Yang et al., 2025b; Gottweis et al., 2026). They can synthesize findings across studies into explicit scientific claims and supporting rationales that connect proposed relationships to the available evidence (Baek et al., 2025; Ghafarollahi & Buehler, 2025). These capabilities open a path to systems that develop scientific ideas through repeated examination of hypotheses and their supporting evidence (Gottweis et al., 2026; Ghareeb et al., 2026; Gao et al., 2025).

Recent progress in automated discovery spans scientific-agent workflows and evolutionary search. One line of research develops agents that ground proposals in the literature and refine them through critical feedback (Wang et al., 2024; Baek et al., 2025; Ghafarollahi & Buehler, 2025). Another uses evolutionary search to develop LLM-generated programs, equations, and molecules, with evaluation and selection guiding subsequent exploration (Romera-Paredes et al., 2024; Novikov et al., 2025; Wang et al., 2026a). Recent systems bring these directions together for scientific hypotheses through tournament-based evolution and hierarchical refinement (Gottweis et al., 2026; Yang et al., 2025a). Yet it remains unclear how the design of agent collaboration affects the hypotheses a research team

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

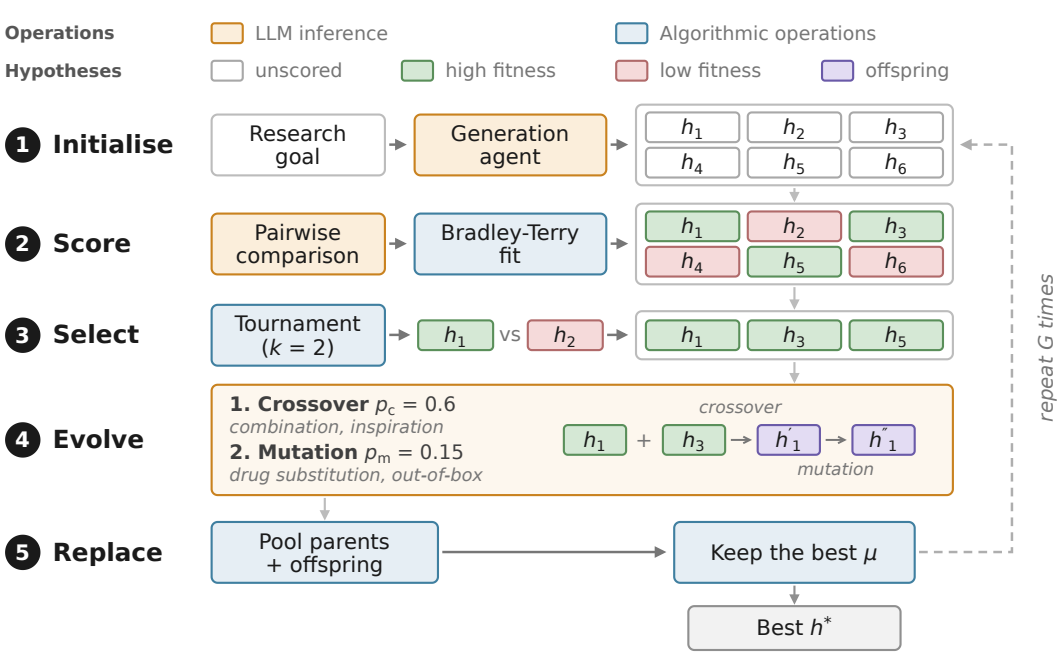

Figure 1: **Overview of HypoEvolve.** The genetic algorithm connects agents' scientific judgments to the hypotheses developed in the next generation, coordinating evaluation, semantic variation, and population replacement. Orange and blue nodes denote LLM calls and algorithmic operations.

develops. A system's performance reflects both the agents' scientific capabilities and the decisions that direct their work. Isolating the contribution of collaboration would provide a basis for designing teams with scientific capabilities beyond those of their individual members.

Controlled comparisons require scientific roles and search decisions to be specified separately (Hao et al., 2023; 2024). Scientific-agent systems assign generation, critique, and synthesis to specialized roles (Ghafarollahi & Buehler, 2025; Gottweis et al., 2026). Evolutionary algorithms provide explicit rules for selecting and varying candidate solutions (Eiben & Smith, 2015; Novikov et al., 2025). Our formulation makes these rules govern how agents develop a population of hypotheses. Each population update determines which proposals agents receive and which outputs enter the next round. We can then vary the search rules with scientific roles, prompts, and evaluation criteria held fixed, making collaboration an experimental variable and hypothesis quality the outcome.

To realize this formulation, we propose HypoEvolve (Figure 1), a generational genetic framework in which specialized LLM agents provide both scientific variation and comparative fitness. We use pairwise judgments of evidence and testability to direct exploration toward promising hypotheses. To develop substantive scientific alternatives, we formulate crossover and mutation as reasoning over claims and rationales. Agents can combine mechanistic arguments across hypotheses or reconsider the assumptions behind an explanation. We evaluate parents and offspring together and retain a fixed-size population, so new proposals compete directly with the ideas they build on. This replacement rule connects comparative judgment to the direction of subsequent search. We record parentage and operator choices to make each hypothesis's development inspectable across generations. The resulting framework makes the coordination of scientific agents explicit and supports controlled changes to the search without redefining their scientific roles.

We evaluate hypothesis discovery through the biological implications of proposed scientific explanations. Because prospective experiments are costly (Gottweis et al., 2026; Wang et al., 2026b), we construct a drug repurposing evaluation that connects candidate interventions and mechanistic rationales to external evidence (Ashburn & Thor, 2004; Pushpakom et al., 2019; Yue et al., 2020). Each rationale implies that the drug's targets are relevant to the specified cancer, providing a concrete biological claim for assessment. Across 34 cancer types (The Cancer Genome Atlas Research Network et al., 2013), we assess this claim with DepMap selectivity (Tsherniak et al., 2017; Meyers

et al., 2017) and Open Targets association (Ochoa et al., 2023), reserving both measures for use after the search. Under a shared task and retrieval protocol, HypoEvolve achieves the highest mean scores among six baselines on both measures. DepMap selectivity reaches 0.171 and Open Targets association reaches 0.426, compared with 0.115 and 0.329 for Tree of Thoughts, the strongest baseline (Yao et al., 2023). The advantage over single-pass generation generalizes to held-out cancer types. With scientific operations and hypothesis count held fixed, fitness-guided parent selection improves the population's mean and minimum scores. This connection between collaboration design and hypothesis quality offers a foundation for building autonomous AI research teams.

2 Related Work

Scientific Hypothesis Discovery. Literature-based discovery generates hypotheses by connecting findings across scientific studies (Swanson, 1986; Spangler et al., 2014; Sybrandt et al., 2017). LLM systems now make hypotheses explicit natural-language artifacts that can be generated, evaluated, and revised. HypoGeniC iteratively updates hypotheses from labeled examples (Zhou et al., 2024), SciMON optimizes literature-grounded scientific directions for novelty (Wang et al., 2024), and ResearchAgent uses reviewing agents to refine research proposals (Baek et al., 2025). A large-scale expert study further shows that novelty, feasibility, and self-evaluation capture different dimensions of research-idea quality (Si et al., 2025). At the level of complete research workflows, the AI Scientist automates idea generation, experimentation, analysis, and manuscript writing (Lu et al., 2026); its template-free variant uses agentic tree search to develop experimental implementations (Yamada et al., 2025). Agent Laboratory carries a researcher-provided idea through literature review, experimentation, and report generation (Schmidgall et al., 2025). HypoEvolve targets the upstream problem of developing scientific hypotheses, using generational search to refine claims whose biological implications are assessed against external evidence.

Multi-Agent Systems for Scientific Hypothesis Discovery. Multi-agent scientific systems distribute generation, criticism, synthesis, and prioritization across specialized roles (Ghafarollahi & Buehler, 2025; Wang et al., 2026c; Hu et al., 2026). MOOSE-Chem retrieves scientific inspirations and composes chemistry hypotheses (Yang et al., 2025b), SciAgents combines ontological knowledge graphs with collaborating agents for materials research (Ghafarollahi & Buehler, 2025), and multi-agent LLMs have generated drug-combination hypotheses (Xu et al., 2025). Robin connects hypothesis formation to experimental feedback (Ghareeb et al., 2026). Co-Scientist is the closest multi-agent reference point. It uses generation, debate, ranking, and evolution agents, and ranks hypotheses through an Elo-based tournament within an expanding pool (Gottweis et al., 2026). The Hypothesis Evolution Protocol separately records hypothesis generation, testing, evidence, and belief updates in an auditable registry (Takahara & Mizoguchi, 2026). HypoEvolve defines collaboration through a fixed-size generational genetic search, with explicit parent selection, controlled semantic variation, joint parent-offspring replacement, and recorded lineages.

Evolutionary Search over Language Artifacts. Evolutionary methods increasingly treat language-model artifacts as members of a population. EvoPrompt and Promptbreeder evolve prompts (Guo et al., 2024; Fernando et al., 2024), while Evolution through Large Models and FunSearch evolve executable programs (Lehman et al., 2024; Romera-Paredes et al., 2024). Quality-Diversity through AI Feedback extends population search to diverse text (Bradley et al., 2024). Language Model Crossover provides a general crossover operator for text-representable artifacts, including sentences, equations, prompts, and code (Meyerson et al., 2024). Related approaches evolve agent teams (Yuan et al., 2025) or train agents jointly through co-evolution (Chen et al., 2025). EvoDiverse brings population-based exploration to scientific-hypothesis search (Wang et al., 2026a). It uses multiple temperature-controlled populations and swap rules to optimize quality and diversity under a fixed validation budget, with experiments over molecules, equations, and algorithms scored by domain-specific automated oracles that drive selection. HypoEvolve evolves structured scientific claims under agent-derived fitness, separating generational genetic search from subsequent assessment against external biological evidence.

3 METHOD

Problem Formulation. Given a natural-language research goal g, we seek hypotheses that address the goal with scientifically grounded explanations and potentially new insights. Each candidate h is a structured document containing a title, summary, hypothesis statement, and supporting rationale. We formulate discovery as a finite population search with μ retained candidates, λ offspring per generation, and a horizon of G generations. Let P_t denote the population at generation t and f_t the fitness inferred from task-specific comparisons in that generation. The search returns the highest-fitness hypothesis in the final population,

$$h^* = \arg\max_{h \in P_G} f_G(h). \tag{1}$$

Fitness summarizes the agents' assessments under the specified scientific criteria and directs parent selection and population replacement. External biological evidence is applied only after search to assess the resulting drug repurposing hypotheses (Section 4.1).

Algorithm Overview. HypoEvolve separates reasoning over scientific content from the population update that coordinates it. Agents supply hypothesis generation, semantic variation, and comparative fitness; the genetic algorithm specifies how these outputs change the population (Holland, 1992; Eiben & Smith, 2015). A generation agent initializes P_0 from retrieved literature. At generation t, selected parents produce λ offspring O_t through LLM-based crossover and mutation. A pairwise scorer evaluates parents and offspring together, and a deterministic supervisor retains the top μ candidates,

$$Q_t = P_{t-1} \uplus O_t, \qquad f_t = \mathrm{Score}(Q_t), \qquad P_t = \mathrm{Top}_\mu(Q_t; f_t). \tag{2}$$

Here $\uplus$ pools candidate records, preserving distinct identities even when their text is unchanged. Lineage records support traceability, while comparative fitness guides selection. Search decisions alter the hypotheses supplied to the comparison and evolution agents while their role definitions, prompts, and scientific criteria remain fixed. The parent-selection study in Section 4.4 uses this separation to change a search rule while preserving the scientific operators and hypothesis count. Figure 1 depicts the agent calls, and Appendix B formalizes the full search in Algorithm 1.

3.1 LLM AGENTS AS SEMANTIC SEARCH OPERATORS

Three specialized agents implement generation, comparison, and evolution. They operate on the claims and rationales within each hypothesis, allowing genetic operations to act on scientific content. Appendix D provides the prompts for each role in our drug-repurposing instantiation.

Literature-Grounded Initialization. The generation agent derives literature queries from g, retrieves relevant papers, and synthesizes their findings. It uses this evidence to propose μ hypotheses spanning different mechanisms, pathways, and interventions. Each proposal follows the same structured format, so later agents receive both a scientific claim and the rationale supporting it.

Comparative Scientific Judgment. The pairwise scorer compares two hypotheses under task-specific criteria and selects the stronger candidate or declares a tie. Pairwise judgments offer a practical basis for ranking open-ended language outputs (Zheng et al., 2023; Liusie et al., 2024; Yin et al., 2026). For drug repurposing, the scorer considers specificity to the named cancer, evidence implicating the proposed target, and whether the hypothesis makes a concrete, falsifiable prediction. These criteria direct attention to cancer-specific dependencies and the scientific argument for each drug repurposing hypothesis. DepMap and Open Targets data are reserved for external assessment and do not enter the scorer.

Semantic Crossover. The evolution agent develops offspring from two selected parents using language-model crossover (Meyerson et al., 2024). The *combination* operator integrates compatible mechanisms or evidence from both parents into a coherent explanation. The *inspiration* operator uses their ideas as starting points for a different explanation aligned with the research goal.

Semantic Mutation. Mutation develops a single hypothesis by revising its proposed intervention or reconsidering its explanation. In the drug-repurposing instantiation, *drug substitution* changes the proposed compound within the allowed vocabulary while retaining the mechanistic argument. The *out-of-box* operator revisits the hypothesis's assumptions and explores alternative explanations.

3.2 Generational Search with Comparative Fitness

The supervisor turns these agent operations into an explicit generational search. It determines which hypotheses reproduce, which variation operators act on them, and which candidates remain in the population. The same procedure repeats at every generation, with parent and operator records tracing the origin of each offspring.

Population-Level Fitness. Population fitness aggregates pairwise scientific judgments into a ranking. The scorer compares every unordered pair in P_0 at initialization and in the parent-offspring pool Q_t at each subsequent generation. A Bradley-Terry model (Bradley & Terry, 1952) converts the comparison outcomes into positive latent strengths $\pi_{t,h}$ and the resulting search fitness,

$$P(h_i \succ h_j) = \frac{\pi_{t,h_i}}{\pi_{t,h_i} + \pi_{t,h_j}}, \qquad f_t(h) = a \log \pi_{t,h} + b_t, \quad a > 0. \tag{3}$$

The initial fit sets the population mean to 50 and the spread to 60 points, fixing the multiplier a for the run. Later fits retain this multiplier and adjust only the offset b_t to align with the previous scores of surviving candidates. This anchoring supplies a common within-run reference for fitness trajectories; selection uses the ordering within each comparison pool.

Fitness-Guided Reproduction. Each parent selection samples two distinct candidates uniformly from P_{t-1} and chooses the one with higher fitness (Miller & Goldberg, 1995). Crossover draws two parents through separate tournaments, resampling the second if it matches the first. With six strictly ranked hypotheses, the strongest wins a third of tournaments and the fifth-ranked wins one in fifteen. The tournament therefore favors stronger candidates while allowing every member except the weakest to reproduce.

For each offspring, crossover is applied with probability $p_c = 0.6$. Mutation then acts on the result with probability $p_m = 0.15$, or with probability 1 if crossover was skipped. Each operation selects uniformly between its two variants. An empty operator return triggers an unchanged parent copy, preserving the offspring count. Every offspring receives a separate record with its parentage and operator provenance, including these fallback copies.

Joint Parent-Offspring Replacement. The supervisor scores the μ parents and λ offspring together and retains the top μ, implementing $(\mu + \lambda)$ truncation (Beyer & Schwefel, 2002). Parents remain eligible alongside their descendants, so a new proposal enters the retained population by ranking among the strongest candidates in the combined pool. This joint comparison links semantic variation to population change and supplies the parents for the next generation of hypothesis development.

4 Experiments

4.1 Experimental Setup

Task Definition. We assess whether hypothesis development identifies interventions supported by independent biological evidence. Drug repurposing makes this question concrete by asking whether an existing compound could act on a disease-specific vulnerability (Ashburn & Thor, 2004; Pushpakom et al., 2019). Each hypothesis proposes a drug candidate and explains how its targets or pathways could affect the specified cancer. This explanation entails an assessable biological implication, namely that the implicated targets are relevant to that cancer. We test this implication through CRISPR perturbations and curated target-disease associations. These scores measure biological support for the proposed drug repurposing opportunity; prospective experiments are needed to establish the full mechanism and therapeutic benefit.

Dataset. Our evaluation spans 34 cancer types, covering the 33 represented in The Cancer Genome Atlas (TCGA) (The Cancer Genome Atlas Research Network et al., 2013) and chronic myelogenous leukemia. Paired comparisons use the 29 cancer types for which every method produced an answer. Three types, kidney chromophobe (KICH), pheochromocytoma and paraganglioma (PCPG), and thymoma (THYM), have no matching DepMap cell lines, leaving 26 types for DepMap selectivity and 29 for Open Targets association.

Held-Out Protocol. Four cancer types informed protocol development, namely acute myeloid leukemia, breast invasive carcinoma, pancreatic adenocarcinoma, and skin cutaneous melanoma.

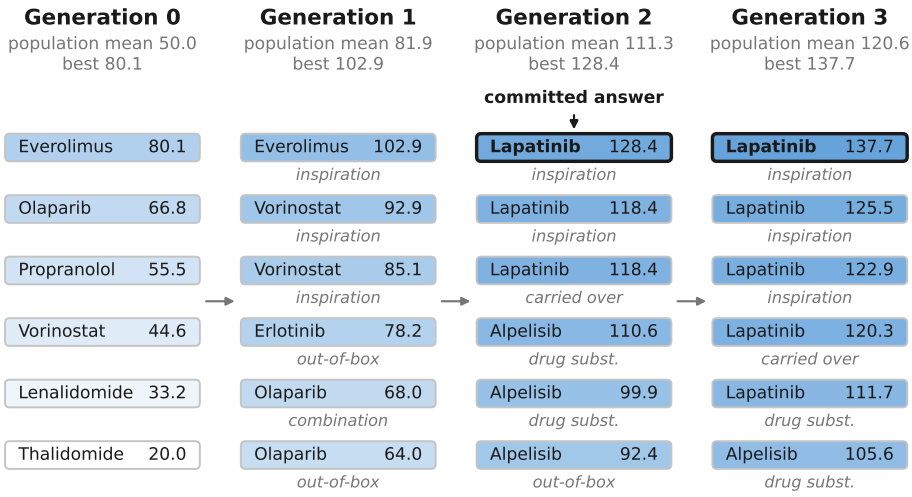

Figure 2: **A hypothesis population across generations.** A LAPATINIB hypothesis introduced by inspiration leads the population in generations 2 and 3, illustrating how new proposals redirect the search. Columns show six retained hypotheses in a uterine corpus endometrial carcinoma run; numbers are fitness scores and labels identify generating operators or unchanged carryovers.

Three more appeared in an interim inspection of the frozen batch, namely adrenocortical carcinoma, bladder urothelial carcinoma, and brain lower grade glioma. We exclude all seven from the held-out analysis. The remaining 27 types were evaluated with no further configuration changes, providing a test beyond the cancer contexts used during protocol development and inspection.

Evaluation Metrics. DepMap CRISPR screens measure how strongly cancer cell lines depend on individual genes for survival (Tsherniak et al., 2017; Meyers et al., 2017). Raw target dependency can reward genes that are essential across many cancers. A constant THALIDOMIDE answer ranks first, with ties, in 30 of the 31 cancer types with matched cell lines under this score. We therefore measure selectivity relative to each target's pan-cancer dependency. THALIDOMIDE in acute myeloid leukemia falls from 1.0000 to $+0.0040$, while IMATINIB in chronic myelogenous leukemia retains $+0.9674$ and VEMURAFENIB in melanoma $+0.9378$. The two external metrics are:

- **DepMap selectivity**: For each drug, we subtract each target's pan-cancer median dependency from its median in the matched cancer and take the maximum across annotated targets.
- **Open Targets association** (Ochoa et al., 2023): The association score between the drug's annotated targets and the matched cancer, providing evidence independent of CRISPR screens.

Evaluation Protocol. Each run selects one drug repurposing hypothesis before external scoring. For HypoEvolve, this is the highest-fitness hypothesis in the final population; every baseline likewise returns one hypothesis and its proposed drug. Scores are averaged within each cancer type before paired comparisons, giving cancer types equal weight. No method is evaluated by taking an externally selected maximum over its candidate pool. DepMap and Open Targets scores are computed after candidate selection and never enter search fitness. All methods use the same curated vocabulary of 61 drugs with annotated targets covered by DepMap (Appendix D).

Baselines. Six task-matched baselines cover independent generation, sampling, reranking, agentic revision, and tree search. They share the base model, drug vocabulary, retrieval protocol, and single-hypothesis output format. Table 1 reports computational costs; Appendix C.5 details the scoring protocol and tests alternative retrieval and answer-selection settings.

(1) Single-pass generation produces one hypothesis per run. **(2) Self-consistency** tests agreement across 40 independent samples (Wang et al., 2023). **(3) Static reranking** selects from a fixed pool of 15 candidates without iterative refinement (Snell et al., 2025). **(4) Multi-agent debate** refines hypotheses through critique and revision (Du et al., 2024). **(5) Co-scientist scaffold** uses generation,

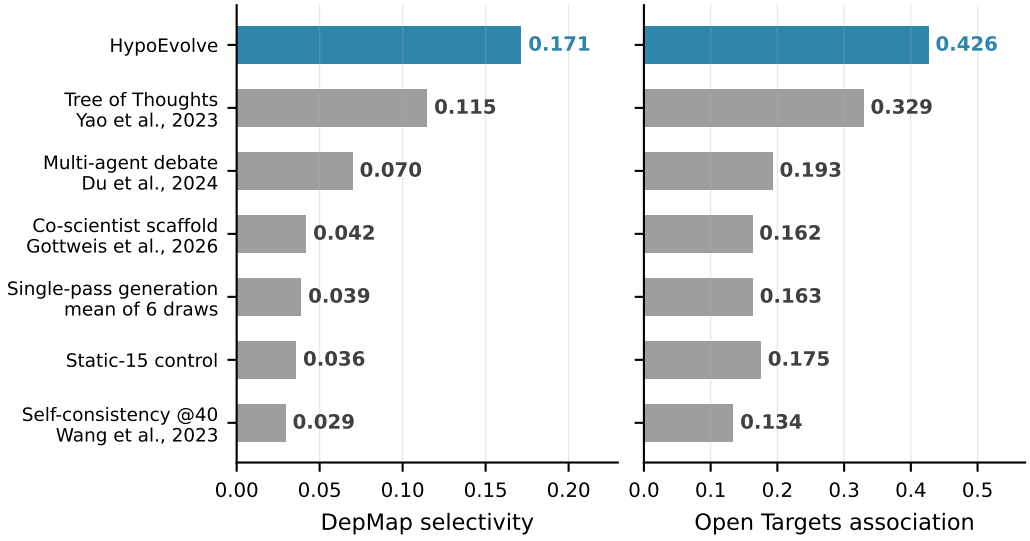

Figure 3: **Comparison with six hypothesis-discovery baselines.** HypoEvolve achieves the highest mean on both biological metrics, with Tree of Thoughts the strongest baseline. Means cover 26 cancer types for DepMap selectivity and 29 for Open Targets. Each run contributes one hypothesis selected by the method before external scoring.

ranking, and meta-review in an expanding pool (Gottweis et al., 2026). Feedback guides subsequent proposals, without crossover, mutation, or population replacement. **(6) Tree of Thoughts** uses beam search to develop and select hypotheses (Yao et al., 2023).

Implementation Details. All agents use gpt-5.4-mini, with structured prompts and Tavily retrieval of literature relevant to each research goal. We use a population of $\mu = 6$, produce $\lambda = 6$ offspring per generation, and run for $G = 3$ generations, with crossover probability $p_c = 0.6$ and mutation probability $p_m = 0.15$. We run HypoEvolve two or three times per cancer type, yielding 94 runs, and average six independent draws for single-pass generation. The sensitivity analysis also tests a population of ten and a horizon of five generations. We fixed the configuration independently of the sensitivity tests and ablations in Section 4.4.

4.2 Main Results

HypoEvolve achieves the highest mean scores. On the 26 cancer types covered by every method and DepMap (Figure 3), selectivity reaches 0.171 against 0.039 for single-pass generation, a paired margin of $+0.133$ with 19 wins and 7 losses. On the 29 types shared by all methods and Open Targets, association reaches 0.426 against 0.163, a margin of $+0.263$ with 26 wins and 3 losses. Tree of Thoughts is the strongest baseline on both metrics, scoring 0.115 and 0.329, with a DepMap margin of $+0.057$ in favor of HypoEvolve. Rankings differ only among the closely grouped co-scientist scaffold, single-pass generation, and static reranking, whose scores lie within 0.007 on DepMap and 0.013 on Open Targets. The ordering otherwise agrees across the two sources of biological evidence.

The gains generalize to held-out cancer types. Across the 27 held-out cancer types, HypoEvolve exceeds single-pass generation by $+0.280$ on Open Targets and by $+0.111$ on the 24 types with matching DepMap cell lines. All seven development or interim-inspection types are excluded. These margins assess the same frozen configuration beyond the cancer contexts used to develop and inspect the protocol, supporting transfer within the evaluated application domain. Appendix C.1 reports the paired comparisons and held-out statistics.

The advantage is strongest on genetic evidence. We separate Open Targets evidence sources to examine whether the advantage is concentrated in channels that directly document drug-disease pairs. Across the 34 cancer types where both methods produced an answer, HypoEvolve exceeds single-pass generation by $+0.247$ on known drugs and clinical trials, with 26 wins and 5 losses; by

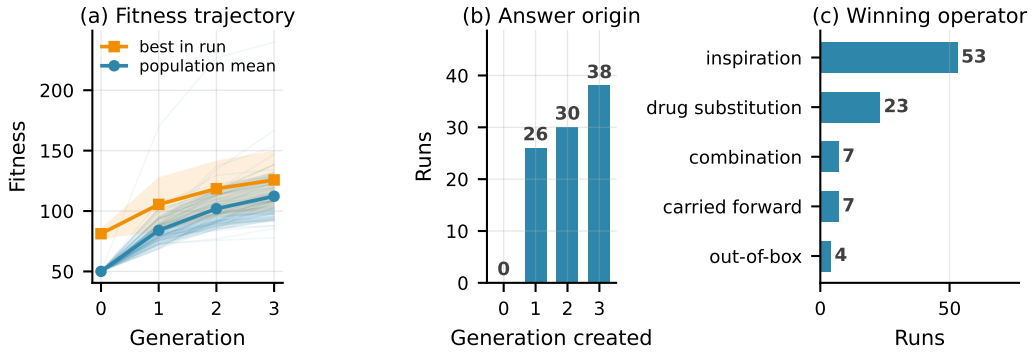

Figure 4: **Fitness trajectories and origins of final hypotheses.** Most final hypotheses are produced during search (87 of 94 through crossover or mutation), and fitness increases in every run. Panels show (a) mean and best fitness with shading for one standard deviation, (b) the generation of each final record, and (c) its generating operator; *carried forward* denotes seven unchanged parent copies.

$+0.313$ on literature, with 30 wins and 4 losses; and by $+0.334$ on genetic association, with 25 wins and 4 losses. The strongest margin in genetics extends the advantage beyond directly documented drug-disease evidence. Prior exposure may still contribute to these gains.

4.3 Hypothesis Evolution

Evolved hypotheses better match drugs to cancer types. We test whether evolution produces drug candidates whose biological evidence is more specific to the proposed cancer. For each external metric, we compare a drug's score in that cancer with its mean score across other cancers. The resulting residual measures how well the proposed cancer matches the drug's biological evidence. Appendix Table 5 reports the full trajectory over 31 cancer types on DepMap and 34 on Open Targets. Both metrics show their largest increase after the first round. The DepMap residual increases through generation 3, while the Open Targets residual peaks at generation 2. From initialization to the final generation, matching improves in 21 of 31 cancer types on DepMap and 25 of 34 on Open Targets. Appendix C.4 gives the statistical comparisons and controls for drugs that score highly across cancers.

Fitness improves in every run. Fitness scores reflect the agents' pairwise judgments of hypotheses (Figure 4a). Population-mean fitness rises from 50.0 in generation 0 to 112.2 in generation 3, and best-member fitness rises from 81.2 to 125.7, with both increasing in all 94 runs. The initial population fixes the reference mean at 50 and the spread at 60 points. The largest gains occur in generations 1 and 2 under the agents' comparative judgments.

Crossover and mutation produce most final hypotheses. Figure 2 illustrates one search trajectory, and Figure 4b,c summarizes the origins of the final hypotheses across runs. Of 94 final hypotheses, crossover produced 60, including 53 from inspiration; mutation produced 27, including 23 from drug substitution. The remaining 7 were unchanged parent copies created after an empty operator return. The final-output records were created in generation 1 for 26 runs, generation 2 for 30, and generation 3 for 38, including the seven unchanged copies.

Search uses all four variation operators. Across all 1,692 offspring records, inspiration accounts for 27.6%, out-of-box mutation for 24.8%, combination for 22.9%, and drug substitution for 17.1%. The remaining 7.5% are unchanged parent copies after empty operator returns.

Evolution refines rationales and changes drug choices. In pancreatic adenocarcinoma, the leading hypothesis retains OLAPARIB while narrowing a generic DNA-damage rationale to stratification by *BRCA1*, *BRCA2*, and *PALB2*. In melanoma, the leading candidate changes from TRAMETINIB to VEMURAFENIB. Both patterns occur across all three seeds. The first makes the conditions for a proposed intervention more specific; the second changes the intervention under consideration. These are generated hypotheses requiring experimental validation. The case studies in Appendix C.6 show how crossover combines parent explanations and mutation proposes an alternative mechanism.

4.4 Ablations and Computational Cost

Larger populations and longer searches give mixed results. Across eight cancer types, we test larger populations, longer horizons, and an alternative operator setting (Appendix Figure 5, left). Raising the population to $\mu = 10$ changes DepMap selectivity by -0.020 and Open Targets association by $+0.058$. Extending the horizon to $G = 5$ gives $+0.022$ and -0.044, while an operator setting with crossover probability 0.3 and mutation probability 0.4 gives $+0.007$ and $+0.019$. No setting differs significantly from the predefined configuration on either metric.

Removing crossover or mutation has mixed effects. We disable crossover, mutation, and both together on the same eight cancer types (Figure 5, center). Removing both reduces each generation to parent cloning and reranking, changing DepMap selectivity by $+0.034$ and Open Targets association by -0.121. The three conditions yield six comparisons across the two metrics, none of which survives Holm correction for multiple testing across the eight cancer types.

Feedback has small effects in the co-scientist baseline. We compare scaffold versions that provide or withhold feedback from the generator (Figure 5, right). Across 31 cancer types on DepMap selectivity and 34 on Open Targets, feedback changes the scores by -0.017 and -0.010, respectively. Repeating the withheld-feedback condition changes them by $+0.012$ and $+0.028$. The scaffold and static reranking also achieve similar scores in the main comparison (Figure 3).

Fitness-guided selection raises average and minimum scores. We replace fitness-based parent selection with uniform random selection, holding generation, crossover, mutation, retrieval, evaluation, and the hypothesis count fixed. Across 31 cancer types on DepMap selectivity and 34 on Open Targets, fitness-guided selection raises the weakest member's scores by 0.088 and 0.218, and the population mean by 0.075 and 0.128. Maximum external scores show no statistically detectable change. For the final hypotheses, the margins are $+0.090$ on DepMap selectivity and $+0.156$ on Open Targets. The clearest effect is stronger biological support for the average and weakest hypotheses. Appendix C.1 provides selection, configuration, and scaffold statistics.

More model calls do not consistently improve results. HypoEvolve uses 206 model calls per run, including 167 pairwise comparisons (Appendix Table 1). The co-scientist scaffold uses 288 calls and the static reranking control uses 227, reflecting the cost of comparing every pair in a 15-candidate pool. Both score within 0.007 of single-pass generation on DepMap selectivity. Tree of Thoughts offers the strongest lower-cost baseline, reaching 0.115 with 31 calls. With 40 samples, self-consistency scores below single-pass generation on both metrics. Costs use 94 HypoEvolve run logs and 204 single-pass run logs; other estimates use observed per-call prices. Token budgets are not matched across methods.

5 Conclusion

HypoEvolve formulates hypothesis development as a population-search problem for scientific agents. We couple reasoning over scientific claims and rationales with a generational genetic algorithm that directs selection, variation, and replacement. This formulation gives each agent contribution a defined role in the search and makes the rules of collaboration available for controlled study. Our drug repurposing evaluation connects the generated explanations to target-level biological evidence from two complementary sources. Across 34 cancer types, HypoEvolve achieves the highest mean scores among six baselines on both DepMap selectivity and Open Targets association. On the shared 26-type DepMap panel, selectivity is 0.171, compared with 0.039 for single-pass generation. Crossover or mutation produces 87 of the 94 final hypotheses.

The parent-selection study directly tests how a coordination decision affects hypothesis quality. Fitness-guided selection improves the population's mean and minimum scores on both external measures with scientific operations and hypothesis count held fixed. The result establishes a contribution of search design to the biological support for the hypotheses agents develop. It motivates a broader research agenda in which the organization of scientific teams is designed and evaluated alongside the capabilities of individual agents. By making collaboration itself a subject of algorithmic design, HypoEvolve creates a foundation for autonomous research teams with scientific capabilities beyond those of individual models.

AI USE STATEMENT

Language model agents are the object of study in this work. The agents inside HypoEvolve and the baselines generate, compare, and revise the drug repurposing hypotheses described in Section 3 and evaluated in Section 4. Separately, we used generative AI assistants to implement and debug the search pipeline, baselines, and analysis scripts, to run experiments and statistical analyses, and to give feedback on experimental design. We also used them to identify related literature, regenerate figures from analysis scripts, format references, draft parts of the paper, and edit the paper for readability and for consistency between the text, figures, and results. Generating synthetic data, formulating or proving mathematical claims, and translation are not applicable to this work. We reviewed all AI-assisted work. We checked code against its outputs, verified reported numbers against run logs, and checked suggested references against the original papers. We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

Section 3 and Appendix B specify the search procedure, and Section 4.1 gives the task, data, metrics, baselines, and configuration. Appendix D lists the prompts for every agent role, and Appendices C.1 and C.5 describe the statistical procedures and the comparison and scoring details. All methods share one base model and a fixed vocabulary of 61 drugs, and both evaluation sources, DepMap and Open Targets, are public.

REFERENCES

Ted T. Ashburn and Karl B. Thor. Drug repositioning: Identifying and developing new uses for existing drugs. *Nature Reviews Drug Discovery*, 3(8):673–683, 2004. doi: 10.1038/nrd1468. URL https://doi.org/10.1038/nrd1468.

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. doi: 10.18653/v1/2025.naacl-long.342. URL https://doi.org/10.18653/v1/2025.naacl-long.342.

Hans-Georg Beyer and Hans-Paul Schwefel. Evolution strategies – a comprehensive introduction. *Natural Computing*, 1(1):3–52, 2002. doi: 10.1023/A:1015059928466. URL https://doi.org/10.1023/A:1015059928466.

Herbie Bradley, Andrew Dai, Hannah Teufel, Jenny Zhang, Koen Oostermeijer, Marco Bellagente, Jeff Clune, Kenneth Stanley, Grégory Schott, and Joel Lehman. Quality-diversity through AI feedback. In *The Twelfth International Conference on Learning Representations*, pp. 21036–21147, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/5b9bef4eae0f574cedbf9f4bf29d8ae7-Abstract-Conference.html.

Ralph Allan Bradley and Milton E. Terry. Rank analysis of incomplete block designs: I. The method of paired comparisons. *Biometrika*, 39(3–4):324–345, 1952. doi: 10.1093/biomet/39.3-4.324. URL https://doi.org/10.1093/biomet/39.3-4.324.

Yixing Chen, Yiding Wang, Siqi Zhu, Haofei Yu, Tao Feng, Muhan Zhang, Mostofa Patwary, and Jiaxuan You. Multi-Agent Evolve: LLM self-improve through co-evolution. *arXiv preprint arXiv:2510.23595*, 2025. doi: 10.48550/arXiv.2510.23595. URL https://arxiv.org/abs/2510.23595.

Yilun Du, Shuang Li, Antonio Torralba, Joshua B. Tenenbaum, and Igor Mordatch. Improving factuality and reasoning in language models through multiagent debate. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 11733–11763, 2024. URL https://proceedings.mlr.press/v235/du24e.html.

A. E. Eiben and J. E. Smith. *Introduction to Evolutionary Computing*. Natural Computing Series. Springer, Berlin, Heidelberg, second edition, 2015. doi: 10.1007/978-3-662-44874-8. URL https://doi.org/10.1007/978-3-662-44874-8.

Chrisantha Fernando, Dylan Sunil Banarse, Henryk Michalewski, Simon Osindero, and Tim Rocktäschel. Promptbreeder: Self-referential self-improvement via prompt evolution. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 13481–13544, 2024. URL https://proceedings.mlr.press/v235/fernando24a.html.

Yiming Gao, Zhen Wang, Jefferson Chen, Mark Antkowiak, Mengzhou Hu, JungHo Kong, Dexter Pratt, Jieyuan Liu, Enze Ma, Zhiting Hu, and Eric P. Xing. scPilot: Large language model reasoning toward automated single-cell analysis and discovery. In *Advances in Neural Information Processing Systems*, volume 38, pp. 1172–1209, 2025. doi: 10.52202/085713-0043. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/01dde7941bce255c2a061eef4fbb7fad-Abstract-Conference.html.

Alireza Ghafarollahi and Markus J. Buehler. SciAgents: Automating scientific discovery through bioinspired multi-agent intelligent graph reasoning. *Advanced Materials*, 37(22):2413523, 2025. doi: 10.1002/adma.202413523. URL https://doi.org/10.1002/adma.202413523.

Ali E. Ghareeb, Benjamin Chang, Ludovico Mitchener, Angela Yiu, Caralyn J. Szostkiewicz, Dmytro Shved, Gavin J. Gyimesi, Jon M. Laurent, Samantha M. Wright, Muhammed T. Razzak, Andrew D. White, Silvia C. Finnemann, Michaela M. Hinks, and Samuel G. Rodriques. A multi-agent system for automating scientific discovery. *Nature*, 655(8122):497–505, 2026. doi: 10.1038/s41586-026-10652-y. URL https://doi.org/10.1038/s41586-026-10652-y.

Juraj Gottweis, Wei-Hung Weng, Alexander Daryin, Tao Tu, Petar Sirkovic, Artiom Myaskovsky, Grzegorz Glowaty, Felix Weissenberger, Alessio Orlandi, Dan Popovici, Anil Palepu, Keran Rong, Ryutaro Tanno, Khaled Saab, Fan Zhang, Jacob Blum, Andrew Carroll, Kavita Kulkarni, Nenad Tomašev, Dina Zverinski, Ivor Rendulic, Elahe Vedadi, Florian Hasler, Luka Rimanic, Marina Boia, Ivan Budiselic, Ben Feinstein, Mathias Bellaiche, Tom Sheffer, Jan Freyberg, Jeremy Ratcliff, Ottavia Bertolli, Katherine Chou, Avinatan Hassidim, Burak Gokturk, Amin Vahdat, Yuan Guan, Vikram Dhillon, Eeshit Dhaval Vaishnav, Byron Lee, Tiago R. D. Costa, José R. Penadés, Gary Peltz, Yossi Matias, James Manyika, Demis Hassabis, Yunhan Xu, Pushmeet Kohli, Annalisa Pawlosky, Alan Karthikesalingam, and Vivek Natarajan. Accelerating scientific discovery with Co-Scientist. *Nature*, 655(8122):487–496, 2026. doi: 10.1038/s41586-026-10644-y. URL https://doi.org/10.1038/s41586-026-10644-y.

Qingyan Guo, Rui Wang, Junliang Guo, Bei Li, Kaitao Song, Xu Tan, Guoqing Liu, Jiang Bian, and Yujiu Yang. Connecting large language models with evolutionary algorithms yields powerful prompt optimizers. In *The Twelfth International Conference on Learning Representations*, pp. 34133–34156, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/9156b0f6dfa9bbd18c79cc459ef5d61c-Abstract-Conference.html.

Weitang Guo, Xin Wang, Bing Lu, Jiaming Yu, Mingxian Xu, Renxuan Huang, Mingzhe Cheng, Meiling Yang, Wei Zhao, and Changye Zou. Super-enhancer-driven MLX mediates redox balance maintenance via SLC7A11 in osteosarcoma. *Cell Death & Disease*, 14(7):439, 2023. doi: 10.1038/s41419-023-05966-y. URL https://www.nature.com/articles/s41419-023-05966-y.

Shibo Hao, Yi Gu, Haodi Ma, Joshua Jiahua Hong, Zhen Wang, Daisy Zhe Wang, and Zhiting Hu. Reasoning with language model is planning with world model. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 8154–8173, 2023. doi: 10.18653/v1/2023.emnlp-main.507. URL https://aclanthology.org/2023.emnlp-main.507/.

Shibo Hao, Yi Gu, Haotian Luo, Tianyang Liu, Xiyan Shao, Xinyuan Wang, Shuhua Xie, Haodi Ma, Adithya Samavedhi, Qiyue Gao, Zhen Wang, and Zhiting Hu. LLM Reasoners: New evaluation, library, and analysis of step-by-step reasoning with large language models. In *First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=b0y6fbSUG0.

John H. Holland. *Adaptation in Natural and Artificial Systems: An Introductory Analysis with Applications to Biology, Control, and Artificial Intelligence*. MIT Press, Cambridge, Massachusetts, 1992. doi: 10.7551/mitpress/1090.001.0001. URL https://doi.org/10.7551/mitpress/1090.001.0001.

Zhengding Hu, Kuntal Talit, Zhen Wang, Haseeb Ahmad, Yichen Lin, Prabhleen Kaur, Christopher Lane, Elizabeth A. Peterson, Zhiting Hu, Elizabeth A. Nowadnick, and Yufei Ding. TritonDFT: Automating DFT with a multi-agent framework. *arXiv preprint arXiv:2603.03372*, 2026. doi: 10.48550/arXiv.2603.03372. URL https://arxiv.org/abs/2603.03372.

Kexin Huang, Ying Jin, Ryan Li, Michael Y. Li, Emmanuel Candès, and Jure Leskovec. Automated hypothesis validation with agentic sequential falsifications. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 25372–25437, 2025. URL https://proceedings.mlr.press/v267/huang25n.html.

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

Joel Lehman, Jonathan Gordon, Shawn Jain, Kamal Ndousse, Cathy Yeh, and Kenneth O. Stanley. Evolution through large models. In *Handbook of Evolutionary Machine Learning*, pp. 331–366. Springer Nature Singapore, 2024. doi: 10.1007/978-981-99-3814-8_11. URL https://doi.org/10.1007/978-981-99-3814-8_11.

Adian Liusie, Vatsal Raina, Yassir Fathullah, and Mark Gales. Efficient LLM comparative assessment: A product of experts framework for pairwise comparisons. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 6835–6855, 2024. doi: 10.18653/v1/2024.emnlp-main.389. URL https://doi.org/10.18653/v1/2024.emnlp-main.389.

Chris Lu, Cong Lu, Robert Tjarko Lange, Yutaro Yamada, Shengran Hu, Jakob Foerster, David Ha, and Jeff Clune. Towards end-to-end automation of AI research. *Nature*, 651(8107):914–919, 2026. doi: 10.1038/s41586-026-10265-5. URL https://doi.org/10.1038/s41586-026-10265-5.

Robin M. Meyers, Jordan G. Bryan, James M. McFarland, Barbara A. Weir, Ann E. Sizemore, Han Xu, Neekesh V Dharia, Phillip G Montgomery, Glenn S Cowley, Sasha Pantel, Amy Goodale, Yenarae Lee, Levi D Ali, Guozhi Jiang, Rakela Lubonja, William F Harrington, Matthew Strickland, Ting Wu, Derek C Hawes, Victor A Zhivich, Meghan R Wyatt, Zohra Kalani, Jaime J Chang, Michael Okamoto, Kimberly Stegmaier, Todd R Golub, Jesse S Boehm, Francisca Vazquez, David E Root, William C Hahn, and Aviad Tsherniak. Computational correction of copy number effect improves specificity of CRISPR-Cas9 essentiality screens in cancer cells. *Nature Genetics*, 49(12):1779–1784, 2017. doi: 10.1038/ng.3984. URL https://doi.org/10.1038/ng.3984.

Elliot Meyerson, Mark J. Nelson, Herbie Bradley, Adam Gaier, Arash Moradi, Amy K. Hoover, and Joel Lehman. Language model crossover: Variation through few-shot prompting. *ACM Transactions on Evolutionary Learning and Optimization*, 4(4):27:1–27:40, 2024. doi: 10.1145/3694791. URL https://doi.org/10.1145/3694791.

Brad L. Miller and David E. Goldberg. Genetic algorithms, tournament selection, and the effects of noise. *Complex Systems*, 9(3):193–212, 1995. URL https://www.complex-systems.com/abstracts/v09_i03_a02/.

Alexander Novikov, Ngân Vũ, Marvin Eisenberger, Emilien Dupont, Po-Sen Huang, Adam Zsolt Wagner, Sergey Shirobokov, Borislav Kozlovskii, Francisco J. R. Ruiz, Abbas Mehrabian, M. Pawan Kumar, Abigail See, Swarat Chaudhuri, George Holland, Alex Davies, Sebastian Nowozin, Pushmeet Kohli, and Matej Balog. AlphaEvolve: A coding agent for scientific and algorithmic discovery. *arXiv preprint arXiv:2506.13131*, 2025. doi: 10.48550/arXiv.2506.13131. URL https://arxiv.org/abs/2506.13131.

David Ochoa, Andrew Hercules, Miguel Carmona, Daniel Suveges, Jarrod Baker, Cinzia Malangone, Irene Lopez, Alfredo Miranda, Carlos Cruz-Castillo, Luca Fumis, Manuel Bernal-Llinares, Kirill Tsukanov, Helena Cornu, Konstantinos Tsirigos, Olesya Razuvayevskaya, Annalisa Buniello, Jeremy Schwartzentruber, Mohd Karim, Bruno Ariano, Ricardo Esteban Martinez Osorio, Javier Ferrer, Xiangyu Ge, Sandra Machlitt-Northen, Asier Gonzalez-Uriarte, Shyamasree Saha, Santosh Tirunagari, Chintan Mehta, Juan María Roldán-Romero, Stuart Horswell, Sarah Young, Maya Ghoussaini, David G Hulcoop, Ian Dunham, and Ellen M McDonagh. The next-generation Open Targets Platform: Reimagined, redesigned, rebuilt. *Nucleic Acids Research*, 51(D1):D1353–D1359, 2023. doi: 10.1093/nar/gkac1046. URL https://doi.org/10.1093/nar/gkac1046.

Sudeep Pushpakom, Francesco Iorio, Patrick A. Eyers, K. Jane Escott, Shirley Hopper, Andrew Wells, Andrew Doig, Tim Guilliams, Joanna Latimer, Christine McNamee, Alan Norris, Philippe Sanseau, David Cavalla, and Munir Pirmohamed. Drug repurposing: Progress, challenges and recommendations. *Nature Reviews Drug Discovery*, 18(1):41–58, 2019. doi: 10.1038/nrd.2018.168. URL https://doi.org/10.1038/nrd.2018.168.

Bernardino Romera-Paredes, Mohammadamin Barekatain, Alexander Novikov, Matej Balog, M. Pawan Kumar, Emilien Dupont, Francisco J. R. Ruiz, Jordan S. Ellenberg, Pengming Wang, Omar Fawzi, Pushmeet Kohli, and Alhussein Fawzi. Mathematical discoveries from program search with large language models. *Nature*, 625(7995):468–475, 2024. doi: 10.1038/s41586-023-06924-6. URL https://doi.org/10.1038/s41586-023-06924-6.

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. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 5977–6043, 2025. doi: 10.18653/v1/2025.findings-emnlp.320. URL https://doi.org/10.18653/v1/2025.findings-emnlp.320.

Chenglei Si, Diyi Yang, and Tatsunori Hashimoto. Can LLMs generate novel research ideas? A large-scale human study with 100+ NLP researchers. In *The Thirteenth International Conference on Learning Representations*, pp. 94003–94092, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/ea94957d81b1c1caf87ef5319fa6b467-Abstract-Conference.html.

Charlie Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. Scaling LLM test-time compute optimally can be more effective than scaling parameters for reasoning. In *The Thirteenth International Conference on Learning Representations*, pp. 10131–10165, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/1b623663fd9b874366f3ce019fdfdd44-Abstract-Conference.html.

Scott Spangler, Angela D. Wilkins, Benjamin J. Bachman, Meena Nagarajan, Tajhal Dayaram, Peter Haas, Sam Regenbogen, Curtis R. Pickering, Austin Comer, Jeffrey N. Myers, Ioana Stanoi, Linda Kato, Ana Lelescu, Jacques J. Labrie, Neha Parikh, Andreas Martin Lisewski, Lawrence Donehower, Ying Chen, and Olivier Lichtarge. Automated hypothesis generation based on mining scientific literature. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1877–1886, 2014. doi: 10.1145/2623330.2623667. URL https://doi.org/10.1145/2623330.2623667.

Don R. Swanson. Fish oil, Raynaud's syndrome, and undiscovered public knowledge. *Perspectives in Biology and Medicine*, 30(1):7–18, 1986. doi: 10.1353/PBM.1986.0087. URL https://doi.org/10.1353/PBM.1986.0087.

Justin Sybrandt, Michael Shtutman, and Ilya Safro. MOLIERE: Automatic biomedical hypothesis generation system. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1633–1642, 2017. doi: 10.1145/3097983.3098057. URL https://doi.org/10.1145/3097983.3098057.

Izumi Takahara and Teruyasu Mizoguchi. Toward auditable AI scientists: A hypothesis evolution protocol for LLM agents. *arXiv preprint arXiv:2607.09195*, 2026. doi: 10.48550/arXiv.2607.09195. URL https://arxiv.org/abs/2607.09195.

The Cancer Genome Atlas Research Network, John N. Weinstein, Eric A. Collisson, Gordon B. Mills, Kenna R. Mills Shaw, Brad A. Ozenberger, Kyle Ellrott, Ilya Shmulevich, Chris Sander, and Joshua M. Stuart. The Cancer Genome Atlas Pan-Cancer analysis project. *Nature Genetics*, 45(10):1113–1120, 2013. doi: 10.1038/ng.2764. URL https://doi.org/10.1038/ng.2764.

Aviad Tsherniak, Francisca Vazquez, Phil G. Montgomery, Barbara A. Weir, Gregory Kryukov, Glenn S. Cowley, Stanley Gill, William F. Harrington, Sasha Pantel, John M. Krill-Burger, Robin M. Meyers, Levi Ali, Amy Goodale, Yenarae Lee, Guozhi Jiang, Jessica Hsiao, William F.J. Gerath, Sara Howell, Erin Merkel, Mahmoud Ghandi, Levi A. Garraway, David E. Root, Todd R. Golub, Jesse S. Boehm, and William C. Hahn. Defining a cancer dependency map. *Cell*, 170(3):564–576.e16, 2017. doi: 10.1016/j.cell.2017.06.010. URL https://doi.org/10.1016/j.cell.2017.06.010.

Haorui Wang, Parshin Shojaee, Kazem Meidani, Kunyang Sun, José Miguel Hernández-Lobato, Teresa Head-Gordon, Jiajun He, Chandan K. Reddy, Chao Zhang, and Yuanqi Du. Towards diverse scientific hypothesis search with large language models. In *Proceedings of the 43rd International Conference on Machine Learning*, volume 306 of *Proceedings of Machine Learning Research*, 2026a. URL https://icml.cc/virtual/2026/poster/66259.

Qingyun Wang, Doug Downey, Heng Ji, and Tom Hope. SciMON: Scientific inspiration machines optimized for novelty. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 279–299, 2024. doi: 10.18653/v1/2024.acl-long.18. URL https://doi.org/10.18653/v1/2024.acl-long.18.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=1PL1NIMMrw.

Zhen Wang, Fan Bai, Zhongyan Luo, Jinyan Su, Kaiser Sun, Xinle Yu, Jieyuan Liu, Kun Zhou, Claire Cardie, Mark Dredze, Zhiting Hu, and Eric P. Xing. FIRE-Bench: Evaluating AI agents on the rediscovery of scientific insights. In *Proceedings of the 43rd International Conference on Machine Learning*, volume 306 of *Proceedings of Machine Learning Research*, 2026b. URL https://icml.cc/virtual/2026/poster/61975.

Zhen Wang, Yiming Gao, Jieyuan Liu, Enze Ma, Jefferson Chen, Mark Antkowiak, Mengzhou Hu, JungHo Kong, Dexter Pratt, Zhiting Hu, Wei Wang, Trey Ideker, and Eric P. Xing. CellMaster: Collaborative cell type annotation in single-cell analysis. *arXiv preprint arXiv:2602.13346*, 2026c. doi: 10.48550/arXiv.2602.13346. URL https://arxiv.org/abs/2602.13346.

Qidi Xu, Claudio Soto, Mohammad Shahnawaz, Xiaozhong Liu, Xiaoqian Jiang, and Yejin Kim. Multi agent large language models for biomedical hypothesis generation in drug combination discovery. *iScience*, 28(12):113984, 2025. doi: 10.1016/j.isci.2025.113984. URL https://doi.org/10.1016/j.isci.2025.113984.

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. doi: 10.48550/arXiv.2504.08066. URL https://arxiv.org/abs/2504.08066.

Zonglin Yang, Wanhao Liu, Ben Gao, Yujie Liu, Wei Li, Tong Xie, Lidong Bing, Wanli Ouyang, Erik Cambria, and Dongzhan Zhou. MOOSE-Chem2: Exploring LLM limits in fine-grained scientific hypothesis discovery via hierarchical search. In *Advances in Neural Information Processing Systems*, volume 38, pp. 89045–89076, 2025a. doi: 10.52202/085713-2979. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/8066ae1446b2bbccb5159587cc3b3bcc-Abstract-Conference.html.

Zonglin Yang, Wanhao Liu, Ben Gao, Tong Xie, Yuqiang Li, Wanli Ouyang, Soujanya Poria, Erik Cambria, and Dongzhan Zhou. MOOSE-Chem: Large language models for rediscovering unseen chemistry scientific hypotheses. In *The Thirteenth International Conference on Learning Representations*, pp. 33251–33277, 2025b. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/51fd9a7d1706023cb9f8210cc6ac357c-Abstract-Conference.html.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L. Griffiths, Yuan Cao, and Karthik Narasimhan. Tree of Thoughts: Deliberate problem solving with large language models. In *Advances in Neural Information Processing Systems*, volume 36, pp. 11809–11822, 2023. doi: 10.52202/075280-0517. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/271db9922b8d1f4dd7aaef84ed5ac703-Abstract-Conference.html.

Yanbin Yin, Kun Zhou, Zhen Wang, Xiangdong Zhang, Yifei Shao, Shibo Hao, Yi Gu, Jieyuan Liu, Somanshu Singla, Tianyang Liu, Eric P. Xing, Zhengzhong Liu, Haojian Jin, and Zhiting Hu. Decentralized Arena: Towards democratic and scalable automatic evaluation of language models. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 35453–35469, 2026. doi: 10.18653/v1/2026.acl-long.1639. URL https://aclanthology.org/2026.acl-long.1639/.

Siyu Yuan, Kaitao Song, Jiangjie Chen, Xu Tan, Dongsheng Li, and Deqing Yang. EvoAgent: Towards automatic multi-agent generation via evolutionary algorithms. 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. 6192–6217, 2025. doi: 10.18653/v1/2025.naacl-long.315. URL https://doi.org/10.18653/v1/2025.naacl-long.315.

Xiang Yue, Zhen Wang, Jingong Huang, Srinivasan Parthasarathy, Soheil Moosavinasab, Yungui Huang, Simon M. Lin, Wen Zhang, Ping Zhang, and Huan Sun. Graph embedding on biomedical networks: Methods, applications and evaluations. *Bioinformatics*, 36(4):1241–1251, 2020. doi: 10.1093/bioinformatics/btz718. URL https://doi.org/10.1093/bioinformatics/btz718.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. In *Advances in Neural Information Processing Systems*, volume 36, pp. 46595–46623, 2023. doi: 10.52202/075280-2020. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/91f18a1287b398d378ef22505bf41832-Abstract-Datasets_and_Benchmarks.html.

Yangqiaoyu Zhou, Haokun Liu, Tejes Srivastava, Hongyuan Mei, and Chenhao Tan. Hypothesis generation with large language models. In *Proceedings of the 1st Workshop on NLP for Science (NLP4Science)*, pp. 117–139, 2024. doi: 10.18653/v1/2024.nlp4science-1.10. URL https://doi.org/10.18653/v1/2024.nlp4science-1.10.

A Discussion

A.1 Implications and Scope

HypoEvolve separates reasoning about scientific claims from decisions about which hypotheses develop across generations. LLM agents supply scientific judgments and hypothesis transformations; the genetic algorithm coordinates their contributions through population updates. Because these updates are explicit, researchers can compare parent-selection and replacement rules with the scientific roles of the agents held fixed.

The drug repurposing study instantiates this principle with domain-specific retrieval, fitness criteria, semantic operators, and external evidence. The reusable component is the population-search architecture that coordinates those elements. Scientific agents already operate over single-cell data, cell-type annotation, and materials simulation (Gao et al., 2025; Wang et al., 2026c; Hu et al., 2026). These domains illustrate how the same genetic control flow could be paired with task-specific representations and evidence sources. Evaluating this transfer across tasks and model families is an important direction for future work.

A.2 Limitations and Future Directions

The main practical constraint is computational cost. Pairwise comparison accounts for 167 of the 206 model calls in a run and grows quadratically with the evaluated pool. Sparse comparison schedules can draw on work in efficient comparative assessment (Liusie et al., 2024; Yin et al., 2026) and redirect this budget toward larger populations or longer horizons. Fitness also depends on LLM judgment, so future versions can incorporate expert preferences or domain evidence directly into the comparison process. Agentic sequential falsification offers a route from free-form hypotheses to external tests (Huang et al., 2025).

The current operator study covers eight cancer types, and broader replicated analyses can clarify the individual contributions of crossover and mutation. Adaptive operator rates and diversity-aware replacement may also help allocate search effort across hypothesis transformations (Lehman & Stanley, 2011; Bradley et al., 2024).

A.3 Impact Statement

HypoEvolve could help researchers compare competing explanations and identify drug repurposing hypotheses for experimental follow-up. The accompanying rationales and parent-offspring records make it possible to inspect which claims were retained, revised, or combined. LLM-based selection can also propagate unsupported assumptions across generations, directing attention toward persuasive hypotheses with weak evidence. Expert review of the supporting literature and independent experiments remain necessary before these proposals inform therapeutic decisions.

B ALGORITHM

Algorithm 1 formalizes the population update in Equation 2. The generation agent supplies P_0, and SCORE compares every unordered pair in its input pool before fitting Bradley-Terry fitness as in Equation 3. Each size-2 tournament samples distinct candidates uniformly and returns the higher-fitness candidate. Crossover applies with probability $p_c = 0.6$ and mutation with probability $p_m = 0.15$, with mutation forced when crossover is skipped. Both operators choose uniformly between their two variants; an empty return invokes the unchanged-parent fallback described in Section 3.2. Each offspring receives a separate record with parentage and operator provenance, making both semantic transformations and fallback copies traceable across generations.

Algorithm 1 HypoEvolve

Require: Research goal g; population size μ; offspring count λ; horizon G
Require: Crossover rate p_c; mutation rate p_m
Ensure: Highest-fitness hypothesis h^* from the final population

$P_0 \leftarrow \textsc{Generate}(g, \mu)$
$f_0 \leftarrow \textsc{Score}(P_0)$ {Initialize the fitness scale}
for $t = 1$ to G **do**
  $O_t \leftarrow \emptyset$
  while $|O_t| < \lambda$ **do**
    $p_1 \leftarrow \textsc{Tournament}(P_{t-1}, f_{t-1}, 2)$
    $h_{\text{new}} \leftarrow p_1$; $c \leftarrow$ FALSE
    if $\text{rand}() < p_c$ **then**
      repeat
        $p_2 \leftarrow \textsc{Tournament}(P_{t-1}, f_{t-1}, 2)$
      until $p_2 \neq p_1$
      $h_{\text{new}} \leftarrow \textsc{Crossover}(p_1, p_2)$; $c \leftarrow$ TRUE
    end if
    if $c =$ FALSE or $\text{rand}() < p_m$ **then**
      $h_{\text{new}} \leftarrow \textsc{Mutation}(h_{\text{new}})$
    end if
    $O_t \leftarrow O_t \uplus \{h_{\text{new}}\}$ {Record offspring identity, parents, and operators}
  end while
  $Q_t \leftarrow P_{t-1} \uplus O_t$
  $f_t \leftarrow \textsc{Score}(Q_t)$ {Compare parents and offspring together}
  $P_t \leftarrow \text{Top}_\mu(Q_t; f_t)$
end for
return $\arg\max_{h \in P_G} f_G(h)$

C ADDITIONAL RESULTS

Figure 5 and Table 1 show the ablation and cost results summarized in Section 4.4.

C.1 STATISTICAL DETAILS FOR MAIN RESULTS

Table 2 details five baseline comparisons in Figure 3 and the held-out analysis in Section 4.2. Tables 3 and 4 provide the selection-ablation statistics and configuration scores underlying Section 4.4. Each analysis uses the cancer panel stated in its table.

Evidence-channel tests use the 34-cancer panel in Section 4.2. Comparisons with single-pass generation give $p = 1.4 \times 10^{-5}$ for known drugs and clinical trials, $p = 1.3 \times 10^{-8}$ for literature, and $p = 9.8 \times 10^{-6}$ for genetic association.

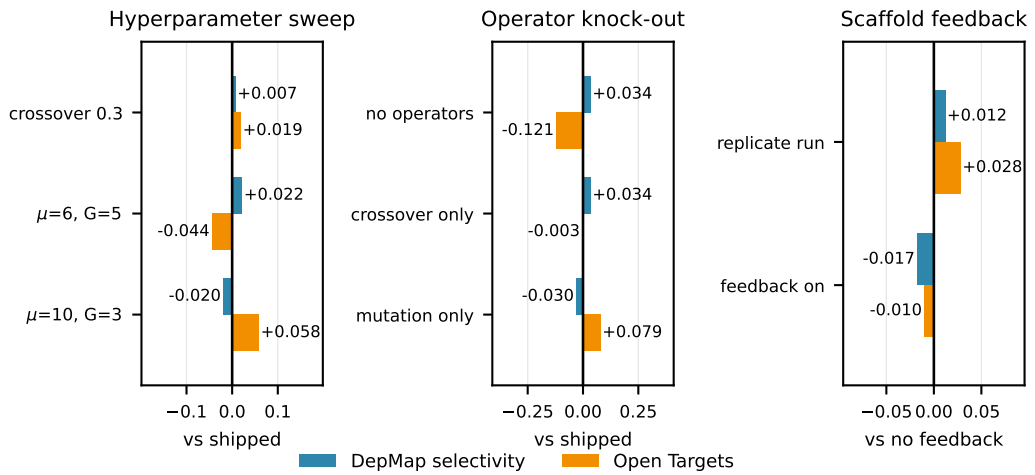

Figure 5: **Search settings, variation operators, and scaffold feedback.** Configuration changes (left) and operator removal (center) produce mixed shifts across the two metrics on eight cancer types; no comparison survives multiple-testing correction. For the co-scientist scaffold, feedback changes scores by amounts comparable to repeating the run (right). These scaffold comparisons cover 31 cancer types on DepMap and 34 on Open Targets.

Table 1: **Computational cost of hypothesis discovery.** HypoEvolve uses fewer model calls than static reranking and the co-scientist scaffold, with pairwise scoring accounting for 167 of 206 calls. *Model calls/run* and *Cost/run* exclude retrieval; costs use run logs for HypoEvolve and single-pass generation and estimates from model-call counts for other methods.

Method	Candidates/run	Model calls/run	Cost/run
Co-scientist scaffold	15.0	288	$0.71
Static reranking (15 candidates)	15.0	227	$0.55
HypoEvolve	**24.0**	**206**	**$0.56**
Self-consistency (40 samples)	39.1	42	$0.21
Tree of Thoughts	9.0	31	$0.09
Multi-agent debate	N/A	17	$0.08
Single-pass generation	1.0	1	$0.004

Table 2: **Statistical details for main and held-out comparisons.** DepMap comparisons favor HypoEvolve; the Open Targets difference from Tree of Thoughts remains statistically unresolved. *Margin* reports mean score differences in favor of HypoEvolve, *W/L* counts cancer-type wins and losses excluding ties, and p denotes paired Wilcoxon tests.

	DepMap selectivity ($n = 26$)			Open Targets ($n = 29$)		
Baseline	Margin	W/L	p	Margin	W/L	p
Single-pass	$+0.1325$	19/7	3.2×10^{-4}	$+0.2633$	26/3	1.1×10^{-6}
Self-consistency	$+0.1420$	18/7	0.0027	$+0.2926$	23/5	4.6×10^{-5}
Static reranking	$+0.1356$	17/8	0.0030	$+0.2513$	23/4	9.9×10^{-5}
Multi-agent debate	$+0.1015$	16/8	0.0056	$+0.2337$	24/3	8.1×10^{-5}
Tree of Thoughts	$+0.0567$	15/7	0.0113	$+0.0974$	14/10	0.126
	Held-out DepMap ($n = 24$)			Held-out Open Targets ($n = 27$)		
Single-pass	$+0.1114$	18/6	2.8×10^{-4}	$+0.2798$	25/2	8.2×10^{-7}

Table 3: **Fitness-guided versus uniform parent selection.** Fitness-guided selection raises mean and minimum population scores on both external measures; population maxima show no statistically detectable change. Positive *Margin* values favor fitness-guided selection with all other search components and hypothesis count fixed; *Final hypothesis* reports the answer selected by the method.

	DepMap selectivity ($n = 31$)		Open Targets ($n = 34$)	
Statistic	Margin	p	Margin	p
Population minimum	$+0.0879$	6.6×10^{-4}	$+0.2181$	1.2×10^{-6}
Population mean	$+0.0753$	8.3×10^{-4}	$+0.1283$	1.0×10^{-6}
Population maximum	$+0.0154$	0.987	$+0.0017$	0.927
Final hypothesis	$+0.0904$	0.064	$+0.1555$	0.0017

The configuration study uses eight cancer types selected before the runs, with one seed per type and setting. Four or five types retain the same drug across settings; increasing the population yields four wins and no losses on Open Targets.

Table 4: **Search configuration sensitivity on eight cancer types.** No setting differs significantly from the default on either metric. Unspecified parameters retain $\mu = 6$, $G = 3$, $p_c = 0.6$, and $p_m = 0.15$; p values compare each setting with the default.

	DepMap selectivity		Open Targets	
Configuration	Mean	p	Mean	p
Default	0.0939	N/A	0.4495	N/A
Population $\mu = 10$	0.0743	0.144	0.5079	0.068
Horizon $G = 5$	0.1157	0.715	0.4052	0.273
$p_c = 0.3$, $p_m = 0.4$	0.1013	0.593	0.4685	0.285

The scaffold feedback test uses three rounds of five candidates and one seed per cancer type. Compared with withholding feedback, providing it yields 3 wins and 10 losses on DepMap ($n = 31$, $p = 0.173$), and 6 wins and 8 losses on Open Targets ($n = 34$, $p = 0.778$). Approximately 60% of cancer types receive the same drug under both conditions.

C.2 Fitness Improves Across All Cancer Types

Fitness improves for every evaluated cancer type (Figure 6), consistent with the aggregate trajectory in Figure 4. Generation 0 is scaled the same way in every run, with the population mean centred at 50 and the spread set to 60 points, so the percentages below are measured against a common baseline. Mean fitness rises in 94 of 94 runs, by 124.2% on average ($p = 1.7 \times 10^{-49}$, paired t-test), and best fitness rises in 94 of 94 runs, by 54.9% ($p = 1.3 \times 10^{-32}$).

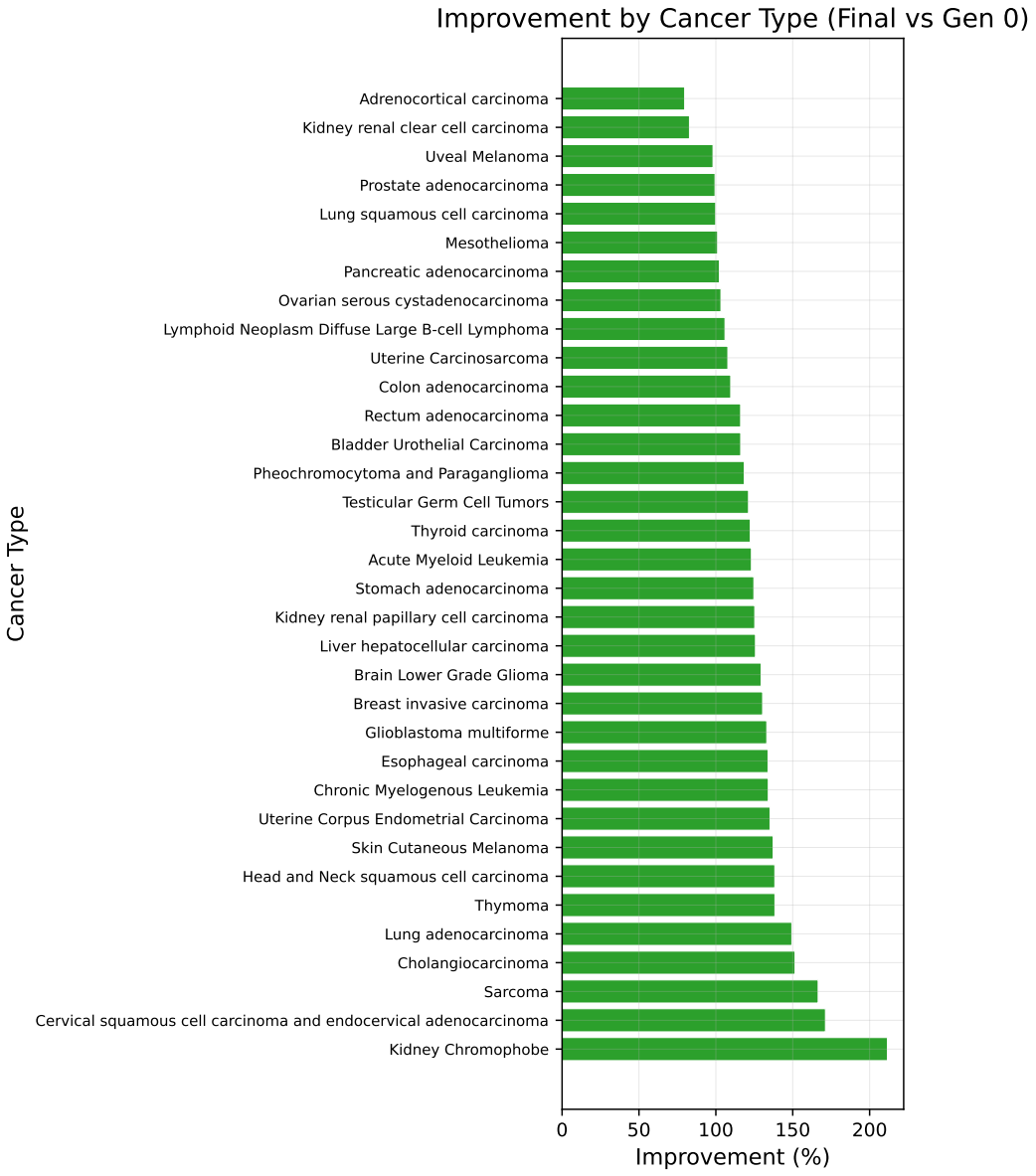

Figure 6: **Fitness improvement by cancer type.** Mean fitness increases in all 34 cancer types, showing consistent progress under the agents' comparative assessments. Bars show percentage changes from generation 0 to generation 3 relative to the common initial fitness scale.

C.3 DepMap Selectivity Improves in Most Cancer Types

HypoEvolve exceeds single-pass generation on DepMap selectivity in 19 of the 26 cancer types shared by all methods and DepMap; single-pass generation leads in 7 (Figure 7). Scores average the hypotheses selected by each method across its runs for each cancer type, following Section 4.1. This comparison shows how the aggregate advantage in Figure 3 varies across cancer contexts.

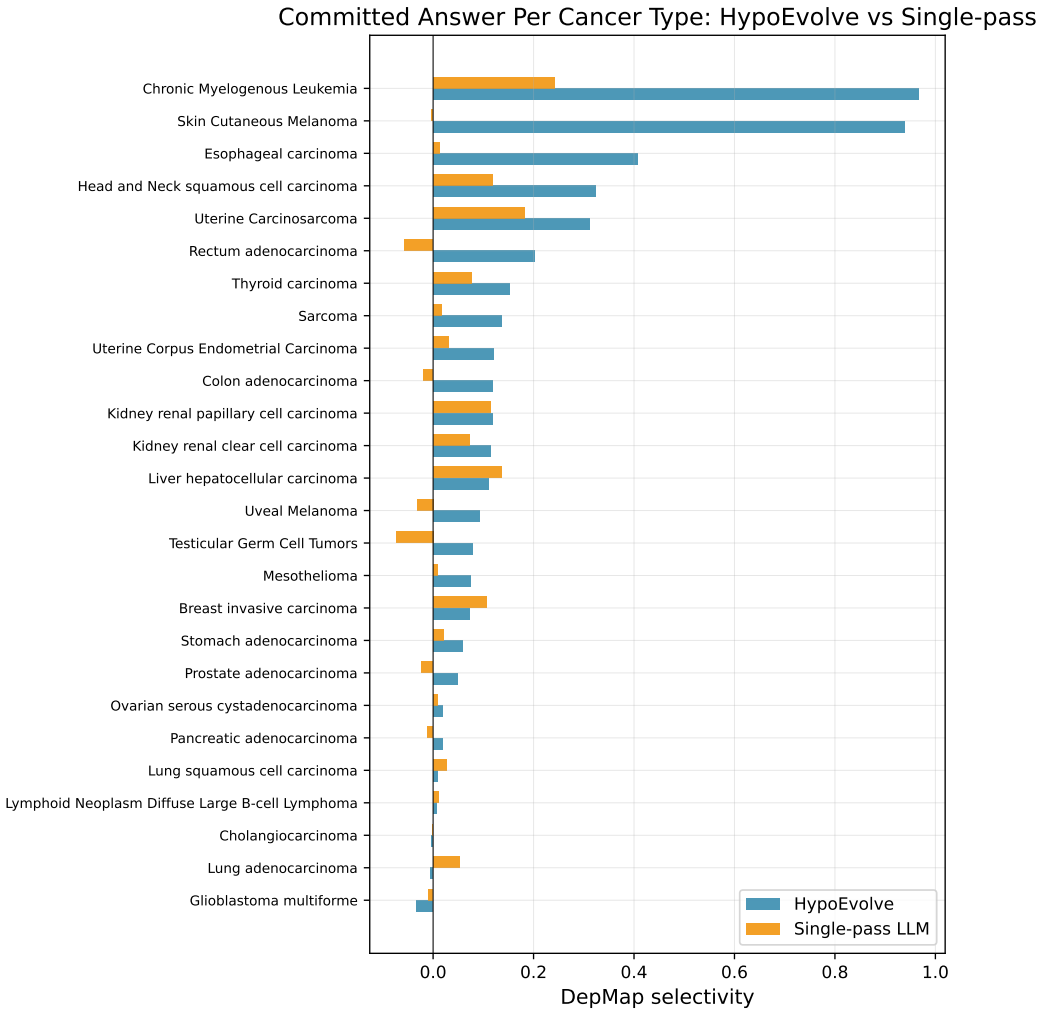

Figure 7: **DepMap selectivity by cancer type.** HypoEvolve exceeds single-pass generation in 19 of 26 cancer types, showing that the aggregate advantage extends to most evaluated contexts. Bars compare per-cancer scores for HypoEvolve (blue) and single-pass generation (orange). Positive selectivity indicates greater target dependency in the matched cancer than in the pan-cancer reference.

Table 5: **Drug-cancer matching across generations.** Mean residuals compare a proposed drug's score in the matched cancer with its average across other cancer types. Both metrics increase from the initial to the final generation, indicating better drug-cancer matching after accounting for drugs that score highly across many cancers.

Generation	DepMap selectivity ($n = 31$)	Open Targets ($n = 34$)
0	$+0.0034$	$+0.0133$
1	$+0.0413$	$+0.0542$
2	$+0.0497$	$+0.0719$
3	$+0.0612$	$+0.0658$

C.4 Evolution Improves Drug-Cancer Matching

We examine whether hypothesis development improves the match between a drug and a cancer context after accounting for the drug's overall score. For an external metric s, let $\mathcal{C}_s$ be the cancer types with available scores. The cancer-specificity residual for a proposed drug d and cancer c is

$$r_s(d, c) = s(d, c) - \frac{1}{|\mathcal{C}_s| - 1} \sum_{\substack{c' \in \mathcal{C}_s \\ c' \neq c}} s(d, c'). \tag{4}$$

This adjustment removes a drug's average advantage across other cancer contexts. The analysis covers 31 types on DepMap selectivity and 34 on Open Targets, using the available HypoEvolve results. The main comparison uses the smaller common set covered by all methods.

From generation 0 to generation 3, the residual increases in 21 of 31 cancer types on DepMap selectivity ($p = 0.0071$) and 25 of 34 on Open Targets ($p = 0.0017$). The final-generation means are $+0.0612$ and $+0.0658$, respectively (Table 5). These results support improvement within HypoEvolve. Comparisons with four multi-candidate baselines favor HypoEvolve directionally on both residuals, but none survives correction for multiple testing.

A constant-drug control illustrates why cancer specificity requires separate assessment. We select one drug using the seven development cancer types, freeze that choice, and evaluate it on the 27 held-out types. Its Open Targets mean is 0.4827, compared with 0.4290 for HypoEvolve, with 8 wins and 17 losses for HypoEvolve ($p = 0.045$). DepMap selectivity favors HypoEvolve by $+0.0917$, with 14 wins and 8 losses ($p = 0.017$). On Open Targets, the constant drug exceeds Tree of Thoughts by 0.1735 and self-consistency by 0.3596.

C.5 Comparison and Scoring Details

The task-matched comparison shares the backbone, drug vocabulary, three literature-search queries, and output format. Each method commits to its answer before external scoring. For single-pass generation, scores average six independent draws within each cancer type.

The main Tree of Thoughts configuration uses the shared retrieval protocol. A second configuration without retrieval obtains mean DepMap selectivity of 0.1183 and Open Targets association of 0.3556. Paired comparisons with HypoEvolve on this variant's evaluated cancer set give margins of $+0.0577$ on DepMap ($p = 0.0061$) and $+0.0651$ on Open Targets ($p = 0.489$). Tree of Thoughts remains the strongest alternative, and the Open Targets margin for this configuration is not statistically resolved.

A separate diagnostic selects the externally best of the six single-pass draws. The mean margins in favor of HypoEvolve are $+0.032$ on DepMap selectivity ($p = 0.87$) and $+0.120$ on Open Targets ($p = 0.019$). This comparison assesses the sampled pool using external evidence unavailable to the methods during answer selection; the DepMap difference remains statistically unresolved.

Drugs map to curated target genes, and cancer types map to DepMap cell lines through `OncotreePrimaryDisease`. The LGG/GBM, COAD/READ, and KIRC/KIRP pairs resolve to identical cell-line sets. DepMap selectivity is computed within the resulting matched sets, with each target's pan-cancer median subtracted before taking the maximum over targets. These mappings specify each comparison's biological coverage and target-level interpretation. Under this selectivity

metric, drug-by-cancer interaction accounts for 79.34% of score variance, and ten distinct drugs attain the maximum across cancer contexts.

C.6 Qualitative Examples of Hypothesis Development

(a) VORINOSTAT in uveal melanoma

Parent 1: Epigenetic reprogramming
HDAC inhibition may promote differentiation in BAP1-deficient tumors.

Parent 2: Survival signaling
HDAC inhibition may disrupt YAP-driven survival programs sustained by GNAQ/GNA11 signaling.

↓ Combination crossover

Final hypothesis (excerpt). Vorinostat will be particularly effective in uveal melanoma cells that are BAP1-deficient because HDAC inhibition can reverse pathogenic chromatin states that support tumor growth while also dampening YAP-driven survival and adaptive resistance pathways.

Testable prediction (excerpt). It predicts that vorinostat should preferentially suppress viability, clonogenic growth, and YAP target gene expression in BAP1-deficient uveal melanoma compared with BAP1-intact cells, and that restoring BAP1 or blocking YAP should reduce the incremental benefit of vorinostat.

(b) EVEROLIMUS in bladder urothelial carcinoma

Parent 1: Pathway dependence
Everolimus may suppress mTORC1-dependent growth in tumors with altered pathway signaling.

Parent 2: Molecular subgroup
Tumors with FGFR3 activation or TSC1 loss may be particularly sensitive to mTORC1 inhibition.

↓ Combination crossover

Final hypothesis (excerpt). Bladder urothelial carcinoma contains a therapeutically relevant subgroup characterized by FGFR3 pathway activation and/or TSC1 inactivation that creates heightened reliance on mTORC1 signaling for growth and survival. In this setting, everolimus will preferentially suppress tumor proliferation and downstream anabolic signaling, with the strongest responses occurring in tumors harboring these alterations and the weakest responses in mTOR-independent disease.

Testable prediction (excerpt). In tumor models and patient-derived samples, FGFR3/TSC1-altered cases should show elevated phospho-S6 and phospho-4EBP1, and these readouts should be suppressed by everolimus alongside reduced proliferation and increased growth arrest.

(c) SULFASALAZINE in sarcoma

Parent hypothesis (summary). The parent proposes inhibiting mTORC1 with SIROLIMUS in sarcomas with activated growth signaling, predicting reduced protein synthesis and proliferation.

↓ Out-of-box mutation

Final hypothesis (verbatim).
I hypothesize that a biologically defined subset of sarcomas is "redox addicted" rather than growth-factor addicted, relying on the cystine transporter system xCT (SLC7A11/SLC3A2) to prevent lethal lipid peroxidation under oxidative and metabolic stress. In these tumors, repurposing SULFASALAZINE will inhibit cystine uptake, collapse glutathione defenses, and selectively induce ferroptosis, with the strongest responses in sarcomas showing high SLC7A11 expression, mesenchymal plasticity, or resistance to conventional therapy.

Testable prediction (excerpt). This hypothesis is testable and feasible in sarcoma cell lines, patient-derived organoids, or ex vivo tumor slices by measuring viability, lipid peroxidation, glutathione depletion, and rescue with ferroptosis inhibitors such as ferrostatin-1 or liproxstatin-1.

Figure 8: **Qualitative examples of hypothesis development.** Crossover connects epigenetic and survival-signaling arguments in (a) and links pathway dependence to molecular conditions for drug response in (b). Mutation shifts the proposed vulnerability from growth signaling to antioxidant defense in (c), with a mechanistic rescue test. Parent arguments are summarized; final hypotheses and testable predictions are reproduced as verbatim excerpts.

Literature context. Published experiments in osteosarcoma report that SULFASALAZINE lowers glutathione, increases lipid peroxidation, and induces cell death that can be rescued by ferroptosis inhibitors (Guo et al., 2023). These findings support the proposed mechanism in osteosarcoma; extending the prediction to other sarcoma subtypes requires direct testing.

D PROMPT TEMPLATES

This section presents the core prompts used by HypoEvolve agents for the drug repurposing task.

D.1 GENERATION AGENT PROMPT

The generation prompt supplies the research goal, supporting literature, and required output format.

Generation Prompt

You are an expert tasked with formulating a novel and robust hypothesis to address the following objective. You have conducted a thorough review of relevant literature and developed a logical framework for addressing the objective.
Goal: {goal}
Criteria for a strong hypothesis: {preferences}
Literature review and analytical rationale: {articles_with_reasoning}
Required Output Format:
`TITLE`: [A concise, descriptive title]
`SUMMARY`: [Single-sentence summary]
`HYPOTHESIS`: [Clear statement in 2-3 sentences]
`RATIONALE`: [Detailed explanation including key mechanisms, evidence from literature, and testability]
`FINAL DRUG`: [Drug name]
`CANCER TYPE`: [TCGA cancer type]

D.2 PAIRWISE COMPARISON PROMPT

The comparison prompt ranks two hypotheses for the same cancer and allows a tie. Its judgments determine the fitness used for parent selection and population replacement.

Pairwise Comparison Prompt

Compare two drug repurposing hypotheses for the SAME cancer type and pick the stronger one.
A stronger hypothesis is one whose proposed drug acts on a dependency that is SPECIFIC to this cancer type, a lineage-defining oncogene, a mutated or amplified driver, or a pathway this tumour type is selectively addicted to.
Judge on:
1. **Specificity**. Would this drug plausibly work better in THIS cancer than in an arbitrary other cancer? A mechanism that applies equally to every tumour type is WEAKER, not stronger, because it does not explain why this cancer was chosen.
2. **Target evidence**. Is the named target actually implicated in this cancer type?
3. **Testability**. Does the hypothesis make a concrete, falsifiable prediction?

Explicitly DO NOT reward: generic cytotoxicity, broadly pleiotropic agents, or mechanisms that reduce to "this pathway matters in cancer generally".
Respond in exactly this format:
`REASONING`: <one to three sentences explaining your choice>
`CONFIDENCE`: <HIGH | MEDIUM | LOW>
`WINNER`=<A | B | TIE>

D.3 Evolution Agent Prompts

Combination Crossover. The agent integrates scientific content from selected parent hypotheses.

Combination Crossover Prompt

You are synthesizing a unified hypothesis from multiple parent hypotheses.
Goal: {goal}
Parent hypotheses: {hypotheses}
Review feedback: {reviews}
Instructions: Integrate the strongest aspects from each parent into a coherent unified hypothesis. Preserve beneficial mechanisms while addressing identified weaknesses. The offspring should be superior to any individual parent.

Out-of-Box Mutation. The agent reconsiders assumptions to develop an alternative explanation.

Out-of-Box Mutation Prompt

You are generating a novel hypothesis inspired by but distinct from provided concepts.
Goal: {goal}
Inspiration (use analogy, not replication): {hypotheses}
Instructions:
1. Identify promising avenues for exploration
2. Develop a detailed, original hypothesis leveraging analogous principles
3. This should not be a mere aggregation of existing methods; think out-of-the-box

D.4 Drug Constraint

Each drug repurposing hypothesis must select from 61 FDA-approved drugs with known targets covered by DepMap CRISPR data, enabling consistent external assessment across methods.

Drug Constraint

You MUST select your drug repurposing candidate ONLY from this approved list:
SIMVASTATIN, ATORVASTATIN, METFORMIN, HYDROXYCHLOROQUINE, PROPRANOLOL, SERTRALINE, OMEPRAZOLE, ASPIRIN, CELECOXIB, DOXYCYCLINE, DISULFIRAM, THALIDOMIDE, SIROLIMUS, EVEROLIMUS, IMATINIB, DASATINIB, SORAFENIB, ERLOTINIB, VEMURAFENIB, OLAPARIB, VENETOCLAX, IBRUTINIB, PALBOCICLIB, RUXOLITINIB, ... [61 drugs total]
These drugs have been verified to have: (1) FDA approval, (2) known target genes in Open Targets Platform, (3) target genes present in DepMap CRISPR data.