Towards an AI Co-Scientist

Speaker: Zhiqing Yang

Citation

Gottweis, J., Weng, W.-H., Daryin, A., Tu, T., Palepu, A., Sirkovic, P., … Natarajan, V. (2025). Towards an AI co-scientist: A multi-agent system for scientific discovery. arXiv preprint arXiv:2502.18864.

https://storage.googleapis.com/coscientist_paper/ai_coscientist.pdf

Overview

- Motivation
- Multi-Agent Architecture
- Evaluation and Results
- Collaborative Innovation & Future Challenges

Motivation

- Research Overload & Human Limitations

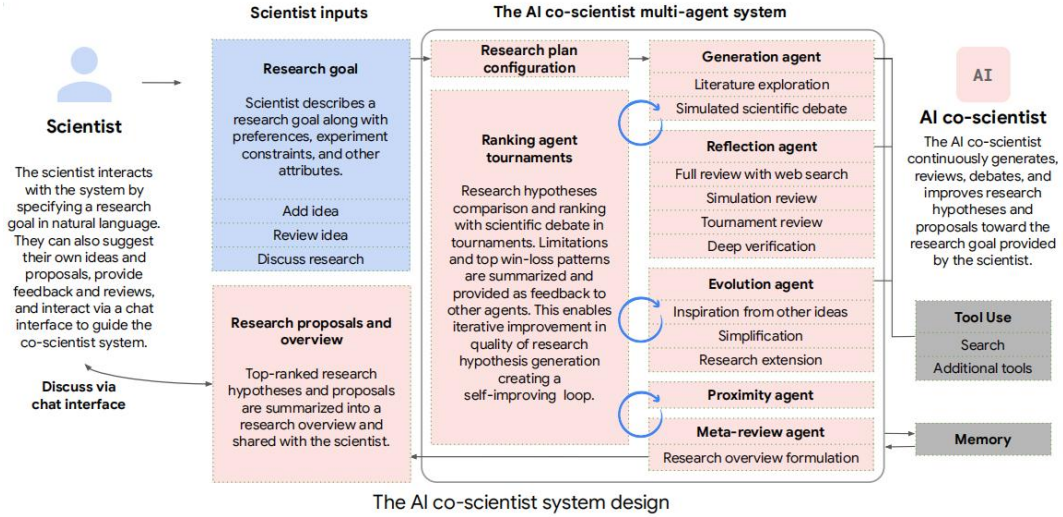
  scientific discovery needs new hypotheses and experiments
  biomedical research is becoming increasingly complex and specialized
  human cognition and time are limited
  overwhelming volume of scientific publications
  rapid emergence of new technologies

- AI as a Collaborator

  traditional AI tools only summarize existing knowledge
  -> allocate extra compute at inference time

- Inspiration

  built on advances in large language models (e.g. Gemini 2.0)
  generates, debates, and evolves

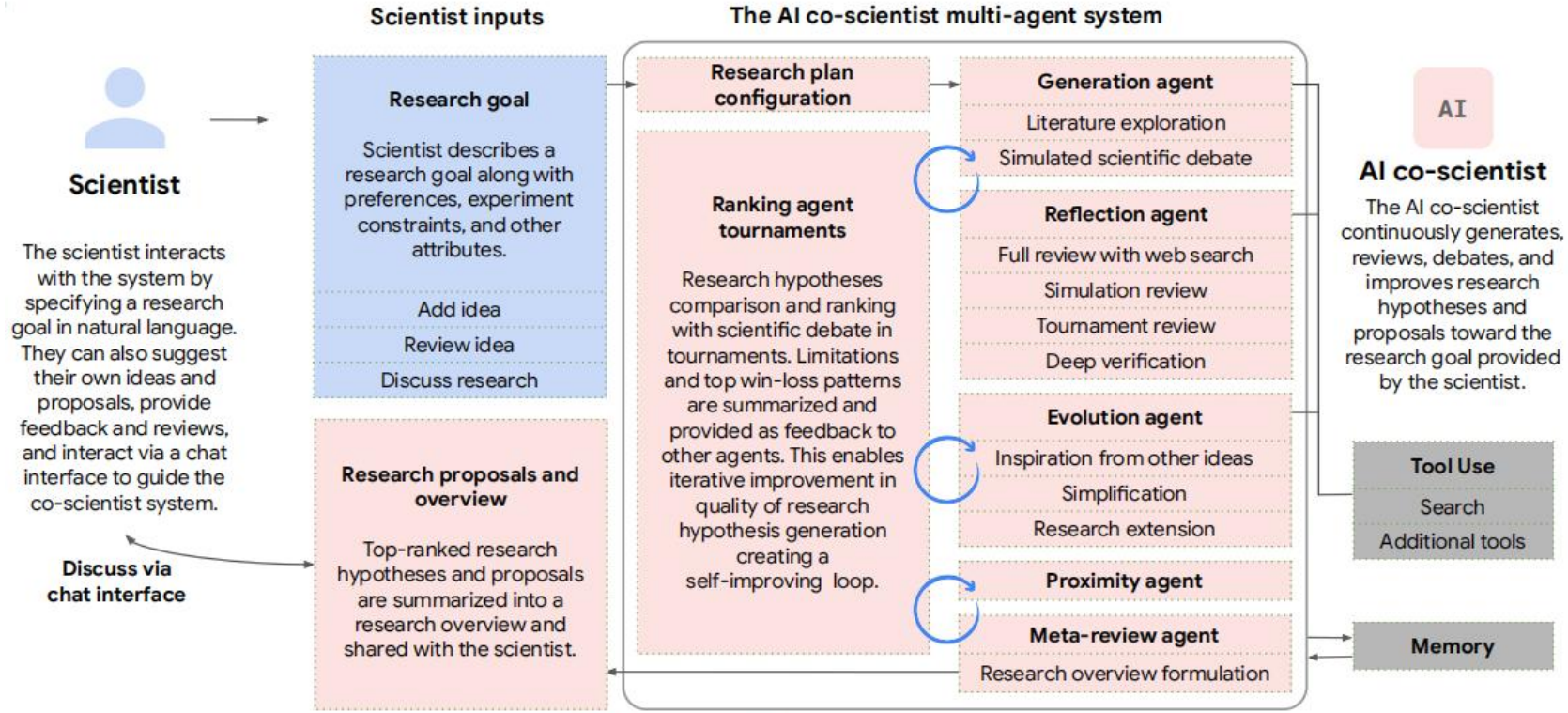


The AI co-scientist system design

-> AI Co-Scientist

AI Co-Scientist

- Scaling Test-time Compute
  - self-play-based scientific debate -> generate original hypotheses
  - tournament-style hypothesis evaluation process -> improve hypotheses by identifying win/loss patterns
  - hypothesis evolution mechanism -> enhance the quality of proposed ideas
  - use of tools such as web search -> enable self-feedback and iterative refinement of suggestions



The AI co-scientist system design

Key Areas

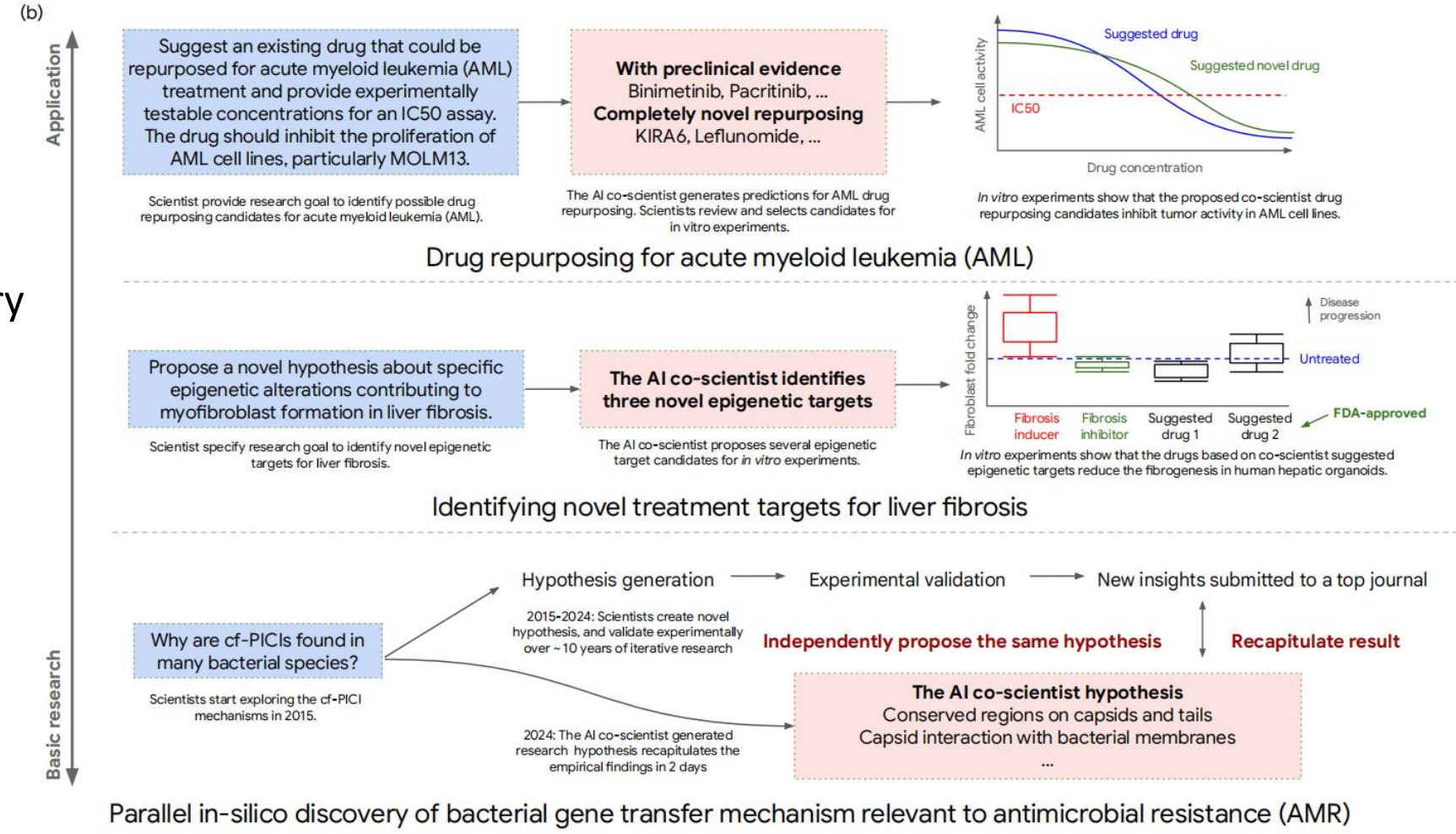
- Drug repurposing

  development of new drugs is time-consuming and costly
  -> identifying new therapeutic uses for existing drugs

- Novel Treatment Target Discovery
  - re-matching existing drugs and diseases
  - uncovering entirely new biological mechanisms and components

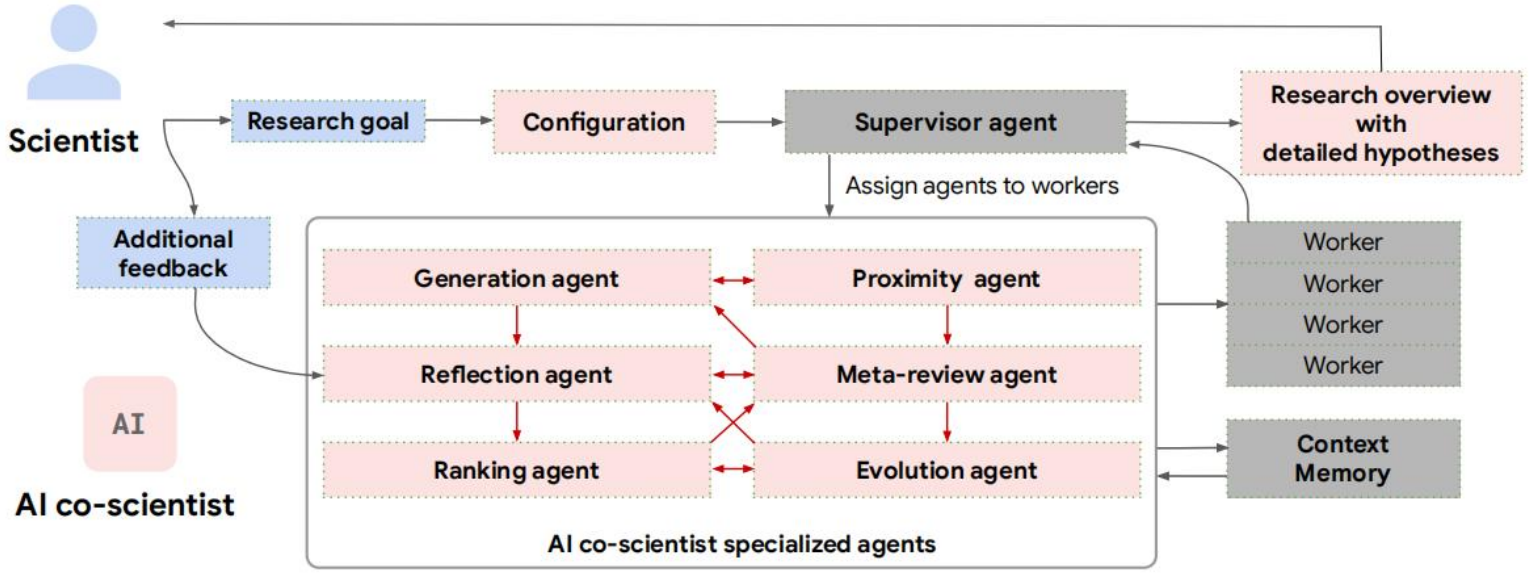
- Hypothesis Generation for Antimicrobial Resistance

  evaluated the system's ability to generate hypotheses' explanation

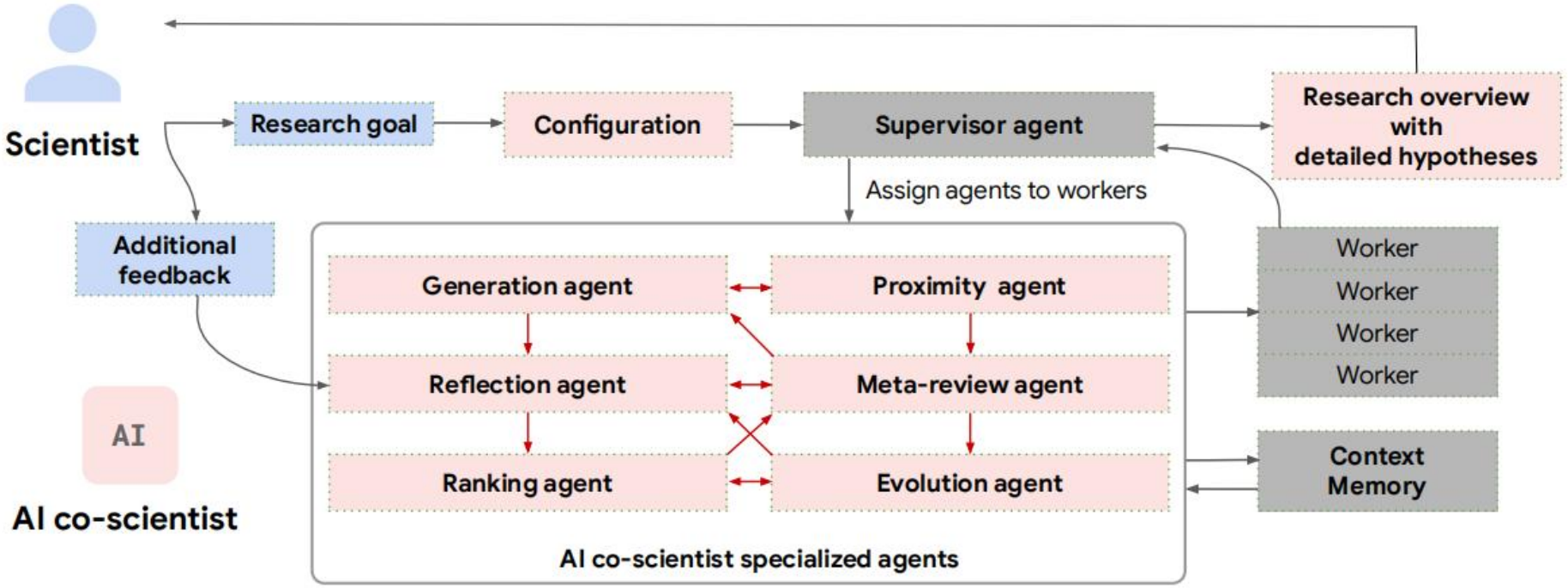


Expert-in-the-Loop Scientific Collaboration

- collaborative workflows by enabling natural language iterations
- employs specialized agents -> generate, debate, and evolve within a tournament framework
- feedback from the tournament enables iterative improvement
- the co-scientist leverages tools
- scientists can converse with the co-scientist



Research Goal / Plan Configuration



- research goal as the start
- leveraging the multimodal and long context capabilities of Gemini 2.0 models
- parses the goal to derive a research plan configuration

Research Goal / Plan Configuration

From research goal to research plan configuration

Scientist research goal

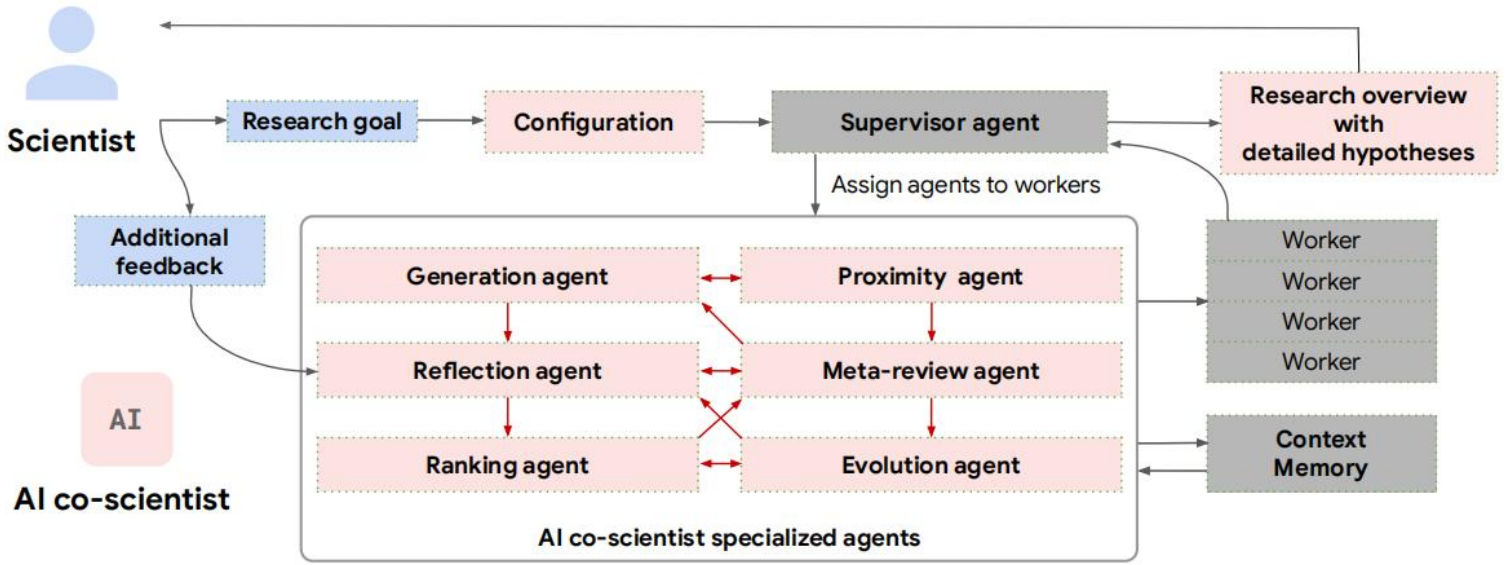
Develop a novel hypothesis for the key factor or process which causes ALS related to phosphorylation of a Nuclear Pore Complex (NPC) nucleoporin. Explain mechanism of action in detail. Include also a feasible experiment to test the hypothesis.

Parsed research plan configuration

- **Preferences:** Focus on providing a novel hypothesis, with detailed explanation of the mechanism of action.
- **Attributes:** Novelty, Feasibility
- **Constraints:** should be correct, should be novel.

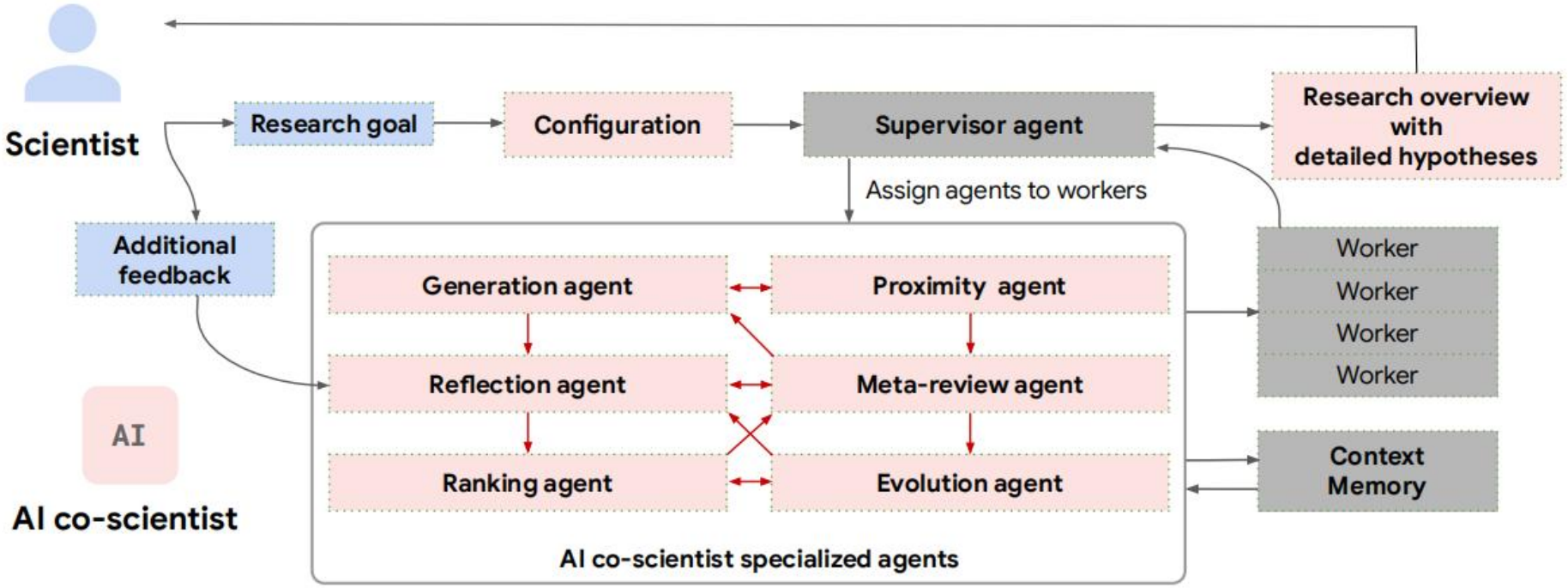
Figure A.1 | An example of a research goal and a parsed research plan configuration. This figure illustrates how the co-scientist system parses a scientist's natural language research goal into a research plan configuration, which guides subsequent reasoning and computation.

Supervisor Agent



- initiates the creation of a task queue
- orchestrates the specialized agents
- periodically calculates a comprehensive set of summary statistics
- writes the state to the context memory

Generation Agent



- proposes an **initial list of candidate hypotheses** or research ideas relevant to the goal

Generation Agent

Prompt for hypothesis generation after literature review

```
You are an expert tasked with formulating a novel and robust hypothesis to address
the following objective.

Describe the proposed hypothesis in detail, including specific entities, mechanisms,
and anticipated outcomes.

This description is intended for an audience of domain experts.

You have conducted a thorough review of relevant literature and developed a logical framework
for addressing the objective. The articles consulted, along with your analytical reasoning,
are provided below.

Goal: {goal}

Criteria for a strong hypothesis:
{preferences}

Existing hypothesis (if applicable):
{source_hypothesis}

{instructions}

Literature review and analytical rationale (chronologically ordered, beginning
with the most recent analysis):

{articles_with_reasoning}

Proposed hypothesis (detailed description for domain experts):
```

Figure A.24 | Example Generation agent prompt for hypothesis generation after literature review and relevant article exploration.

Generation Agent

Example output of the Generation agent

Research goal Develop a novel hypothesis for the key factor or process which contributes to ALS progression ... We can use iPSC cells for the in-vitro experiments. ...

Introduction
Amyotrophic Lateral Sclerosis (ALS) is a fatal neurodegenerative disease characterized by the selective loss of motor neurons, leading to progressive muscle weakness, paralysis, and ultimately death. While the exact cause of ALS remains unknown, a complex interplay of genetic predisposition, environmental factors, and age-related cellular changes is believed to contribute to its pathogenesis. ...

Recent findings and related research
Recent research has implicated several key players and processes in ALS. TDP-43, an RNA-binding protein, is found mislocalized and aggregated in the cytoplasm of affected neurons in the majority of ALS cases. This protein's ability to undergo liquid-liquid phase separation (LLPS) has come under intense scrutiny, with evidence suggesting that dysregulation of this process leads to the formation of pathological aggregates. Further, impairments in nucleocytoplasmic transport, the process by which molecules move between the nucleus and cytoplasm, have been observed in ALS and linked to TDP-43 pathology. Endoplasmic reticulum (ER) stress, a cellular response to protein misfolding, has also been implicated as a potential contributing factor.

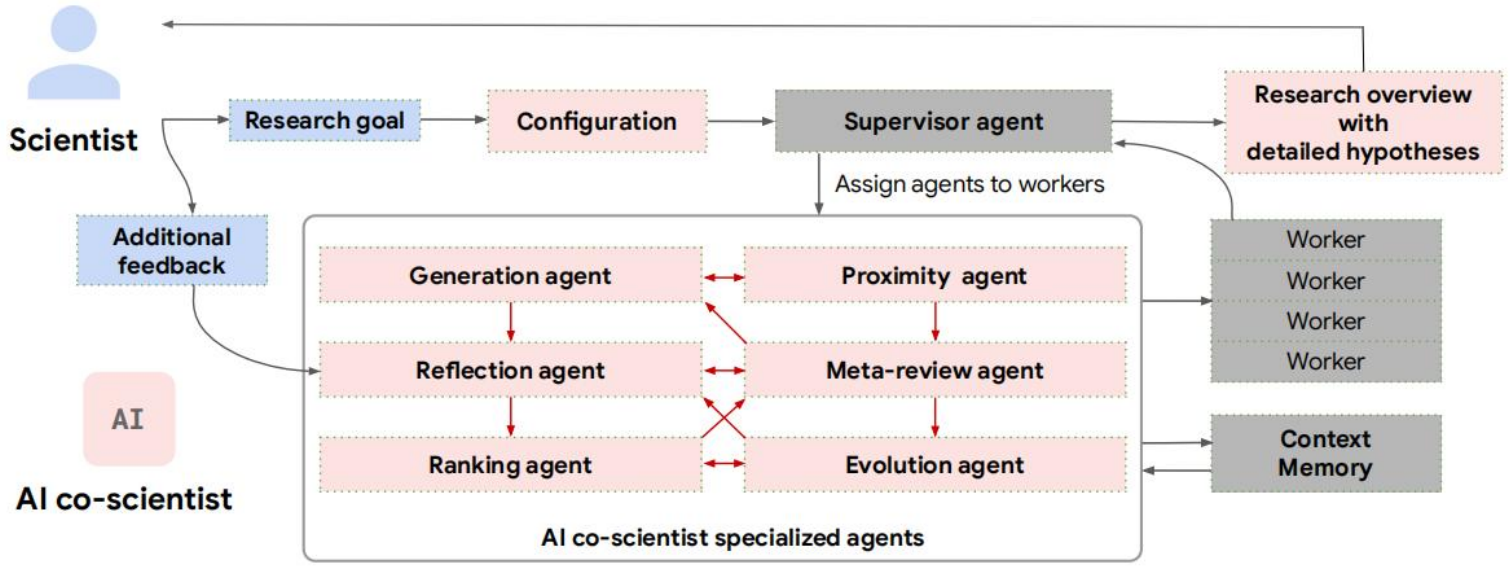
Hypothesis
Cellular stress (including but not limited to ER stress) induces post-translational modifications (PTMs) on specific FG-Nups (focusing initially on Nup98 and Nup62) at the nuclear pore complex (NPC). These Nup PTMs modify the interaction between these Nups and TDP-43, leading to altered TDP-43 dynamics and increased retention at the NPC. This increased retention at the NPC, in turn, triggers the disruption of nucleocytoplasmic transport.

Rationale and specificity
Instead of making the hypothesis about LLPS, we are making it much more specific but with the same starting point. Cellular stress, a broad concept encompassing various cellular insults, can trigger a cascade of molecular events. A key consequence of stress response is altered protein structure and function, which can be mediated by PTMs. Nuclear pore proteins, particularly FG-Nups like Nup98 and Nup62, form a "hydrogel" or selective barrier at the NPC, regulating the movement of molecules between the nucleus and cytoplasm. We hypothesize that specific PTMs on these FG-Nups (e.g., phosphorylation, O-GlcNAcylation initially) directly alter their interaction with TDP-43. TDP-43, known to interact with the NPC and be involved in nucleocytoplasmic transport, may get inappropriately trapped or retained at the NPC when interacting with PTM-modified Nups. This trapping alters TDP-43 dynamics (even if not disrupting LLPS directly initially), meaning it cannot perform its normal functions and disrupting nucleocytoplasmic transport, which we know is happening in ALS. Since even a small change in transport could lead to downstream problems, this is a plausible initial event with testable consequences. The specific FG-Nups (Nup98/62) and chosen PTMs (phosphorylation/O-GlcNAcylation) provide a concrete starting point for experimentation. This mechanism directly links cellular stress to TDP-43 dysfunction and the key ALS pathology of defective nucleocytoplasmic transport by invoking a novel interaction mechanism at the NPC – all testable in vitro.

Experimental design and validation
Cell Types: Human iPSC-derived motor neurons from healthy controls and ALS patients. Stress Induction: Induce cellular stress using established pharmacological agents (e.g., tunicamycin for ER stress, arsenite for oxidative stress, oligomycin A for mitochondrial stress) and physical stressors (heat shock). Test different stressors individually and in combination. This broadened approach allows for a more robust investigation of the stress response.
Nup PTM Analysis: Assess PTMs of Nup98 and Nup62 specifically using mass spectrometry and immunoblotting with PTM-specific antibodies (specifically looking for phosphorylation and O-GlcNAcylation). Focus the analysis on these two.
...

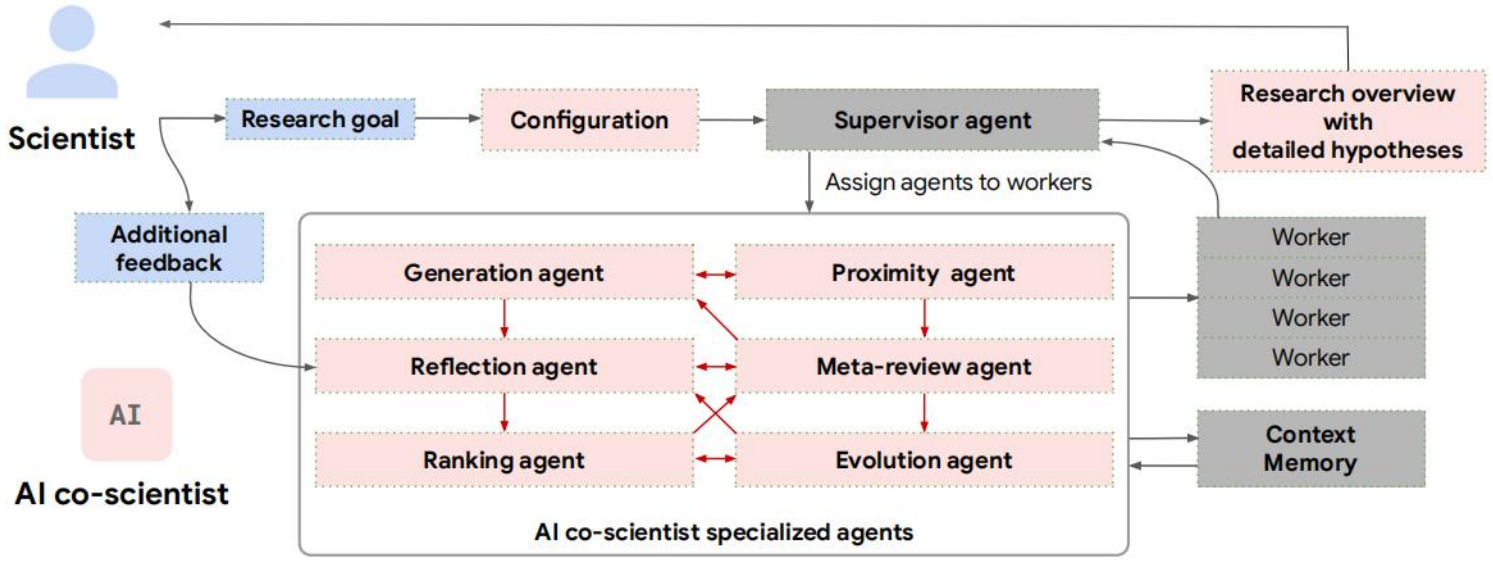
Figure A.2 | An example output of the Generation agent. The abbreviated example showcases an initial hypothesis and research proposal generated by the co-scientist Reflection agent pertaining to the research goal of explaining the basic mechanism for Amyotrophic Lateral Sclerosis (ALS).

Reflection Agent



- searches relevant prior work
- assesses existing experimental evidence for or against a given hypotheses
- verifies the novelty, correctness, and quality of generated outputs
- filters inaccurate and non-novel hypotheses

Reflection Agent



Types:

- **Initial Review**: quickly discards flawed, non-novel hypotheses
- **Full Review**: leverages external tools to identify relevant articles for improved reasoning and grounding
- **Deep Verification Review**: detects subtle errors within complex hypotheses
- **Observation Review**: explores whether a given hypothesis can account for long-tail observations from prior experimental results
- **Simulation Review**: simulates the mechanism of action or the proposed experiment in the proposal
- **Recurrent/Tournament Review**: adapts its full reviews based on the co-scientist's growing knowledge

Reflection Agent

Prompt for generating observations which can be explained by the hypothesis

```
You are an expert in scientific hypothesis evaluation. Your task is to analyze the
relationship between a provided hypothesis and observations from a scientific article.
Specifically, determine if the hypothesis provides a novel causal explanation
for the observations, or if they contradict it.

Instructions:

1. Observation extraction: list relevant observations from the article.
2. Causal analysis (individual): for each observation:
   a. State if its cause is already established.
   b. Assess if the hypothesis could be a causal factor (hypothesis => observation).
   c. Start with: "would we see this observation if the hypothesis was true:".
   d. Explain if it's a novel explanation. If not, or if a better explanation exists,
      state: "not a missing piece."
3. Causal analysis (summary): determine if the hypothesis offers a novel explanation
   for a subset of observations. Include reasoning. Start with: "would we see some of
   the observations if the hypothesis was true:".
4. Disproof analysis: determine if any observations contradict the hypothesis.
   Start with: "does some observations disprove the hypothesis:".
5. Conclusion: state: "hypothesis: <already explained, other explanations more likely,
   missing piece, neutral, or disproved>".

Scoring:
    *  Already explained: hypothesis consistent, but causes are known. No novel explanation.
    *  Other explanations more likely: hypothesis *could* explain, but better explanations exist.
    *  Missing piece: hypothesis offers a novel, plausible explanation.
    *  Neutral: hypothesis neither explains nor is contradicted.
    *  Disproved: observations contradict the hypothesis.

Important: if observations are expected regardless of the hypothesis, and don't disprove it,
it's neutral.

Article:
{article}

Hypothesis:
{hypothesis}

Response {provide reasoning. end with: "hypothesis: <already explained, other explanations
more likely, missing piece, neutral, or disproved>".)
```

Figure A.26 | Example Reflection agent prompt for generating observations from prior experimental results which can be explained by the hypothesis under consideration.

Reflection Agent

Example output of a full review by the Reflection agent

Related Articles:

[5] TDP-43 pathology disrupts nuclear pore complexes and nucleocytoplasmic transport in ALS/FTD - PMC: This article is highly relevant because it directly links TDP-43 pathology to nuclear pore complex (NPC) dysfunction and nucleocytoplasmic transport (NCT) defects in ALS/FTD, providing a strong foundation for the hypothesis.

[4] Unraveling the impact of disrupted nucleocytoplasmic transport systems in C9orf72-associated ALS - PMC: This article provides context on NCT defects in ALS, particularly in the context of C9orf72 mutations, highlighting the broader relevance of NCT dysfunction in different forms of ALS.

[7] Nuclear pore and nucleocytoplasmic transport impairment in oxidative stress-induced neurodegeneration: relevance to molecular mechanisms in Pathogenesis of Parkinson's and other related neurodegenerative diseases - PMC: This article connects oxidative stress, a form of cellular stress, to NPC dysfunction and NCT impairment in neurodegenerative diseases, supporting the idea that cellular stress can impact NPCs.

[3] Frontiers | Molecular Mechanisms of TDP-43 Misfolding and Pathology in Amyotrophic Lateral Sclerosis: This article reviews TDP-43 pathology in ALS, including its mislocalization and aggregation, which are central to the proposed hypothesis.

[1] TDP-43 proteinopathies: a new wave of neurodegenerative diseases - PMC: Provides an overview of TDP-43 proteinopathies, establishing the importance of TDP-43 dysfunction in ALS and other neurodegenerative diseases.

[8] Aβ-driven nuclear pore complex dysfunction alters activation of necroptosis proteins in a mouse model of Alzheimer's Disease: Although focused on Alzheimer's, this article demonstrates that NPC dysfunction can occur in neurodegenerative diseases and highlights the potential impact on cellular processes like necroptosis. It supports the general idea that NPC defects can be a consequence of proteinopathy-associated stress.

[9] Traumatic injury compromises nucleocytoplasmic transport and leads to TDP-43 pathology: This article provides another example of how cellular stress (traumatic injury) can lead to NCT defects and TDP-43 pathology, further supporting the link between stress, NCT, and TDP-43 dysfunction.

...

Full Review

Assumptions of the Idea:

- Cellular stress induces PTMs on Nup98 and Nup62: The hypothesis assumes that various forms of cellular stress can trigger specific PTMs (phosphorylation and O-GlcNAcylation) on Nup98 and Nup62.
- PTMs on Nup98/Nup62 alter their interaction with TDP-43: It's assumed that these specific PTMs directly modify the binding affinity or interaction dynamics between Nup98/Nup62 and TDP-43 at the NPC.
- Altered Nup-TDP-43 interaction leads to increased TDP-43 retention at the NPC: This assumption posits that the modified interaction results in TDP-43 being trapped or retained at the NPC for longer durations.
- Increased TDP-43 retention at the NPC disrupts nucleocytoplasmic transport: The hypothesis assumes that this prolonged retention of TDP-43 at the NPC interferes with the normal transport of other molecules across the nuclear envelope.
- Nucleocytoplasmic transport defects contribute to ALS pathology: The idea assumes that the resulting NCT disruption is a significant contributor to the downstream pathological events in ALS, such as motor neuron degeneration.
- Motor neurons are more susceptible to this mechanism: It's assumed that motor neurons are particularly vulnerable to this stress-induced, Nup-mediated TDP-43 retention and subsequent NCT defects compared to other cell types.

Reasoning about assumptions:

- Cellular stress induces PTMs on Nup98 and Nup62: This is plausible. Cellular stress is known to activate various signaling pathways that can lead to PTMs on numerous proteins. Nups, being long-lived proteins, are susceptible to modifications. However, whether Nup98 and Nup62 are specifically targeted by stress-induced phosphorylation and O-GlcNAcylation needs experimental verification.
- PTMs on Nup98/Nup62 alter their interaction with TDP-43: This is possible but requires experimental validation. PTMs can significantly alter protein-protein interactions by changing surface charge, conformation, or binding sites. While there's no direct evidence for this specific interaction, it's a reasonable hypothesis given the role of FG-Nups in interacting with transport factors.
- Nucleocytoplasmic transport defects contribute to ALS pathology: This is supported by existing literature [4, 5]. NCT defects have been observed in ALS and are thought to contribute to disease pathogenesis. However, whether this specific mechanism is a primary driver or a downstream consequence needs further investigation.
- Motor neurons are more susceptible to this mechanism: This is a crucial assumption but lacks strong justification in the current proposal. Motor neurons have unique characteristics (high metabolic demand, long axons) that might make them more vulnerable to stress and NCT defects. However, this needs to be explicitly tested by comparing motor neurons to other cell types.
- ...

Figure A.5 | An example output of a full review by the Reflection agent The abbreviated example showcases a full initial review generated by the co-scientist Reflection agent for an example hypothesis pertaining to explaining the basic mechanism for Amyotrophic Lateral Sclerosis (ALS).

Reflection Agent

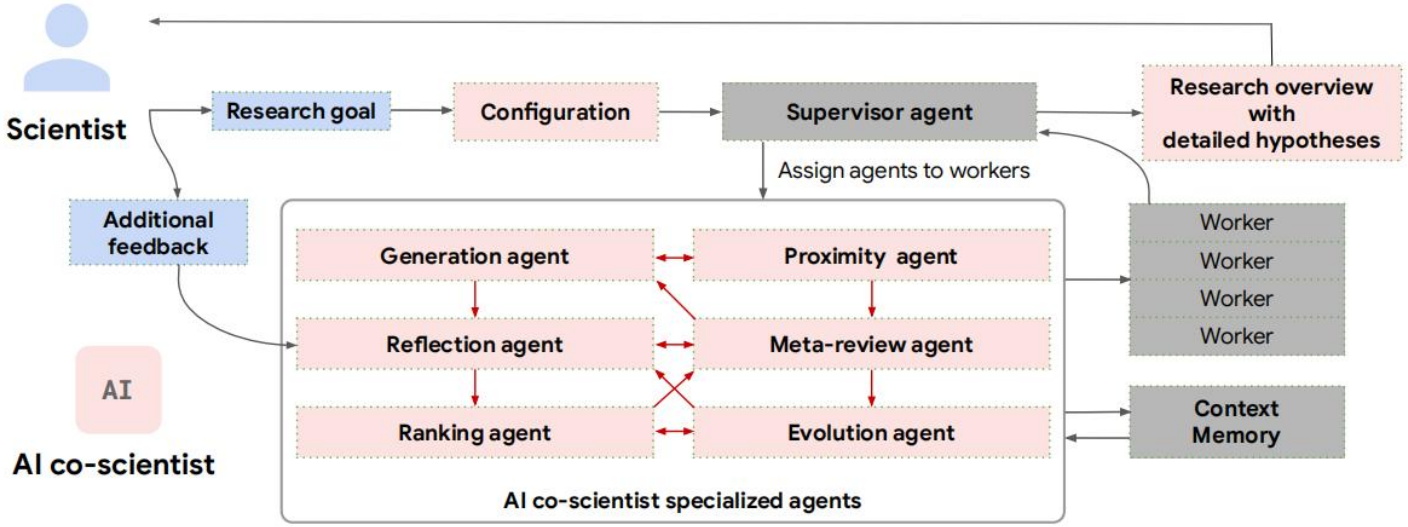
Example output of review critiques by the Reflection agent

- Lack of strong justification for motor neuron specificity: The initial hypothesis lacked a strong explanation for why the proposed mechanism would predominantly affect motor neurons, a critical aspect of ALS pathology.
- Limited exploration of downstream consequences: While the idea connects to nucleocytoplasmic transport defects, it doesn't fully elaborate on how these defects lead to other ALS-related pathologies like TDP-43 aggregation or RNA processing defects.
- Potential technical challenges: Identifying the specific PTMs responsible and demonstrating their direct impact on TDP-43 retention and transport might be technically difficult.
- Narrow focus on specific Nups and PTMs: Initially focusing only on Nup98/Nup62 and two types of PTMs (phosphorylation, O-GlcNAcylation) could be too restrictive, potentially overlooking other important factors.
- Less holistic approach: The focus on a specific mechanism at the NPC might neglect broader cellular processes involved in ALS pathogenesis.
- The temporal relationship with TDP-43 pathology is not clear: It's not clear whether the hypothesized mechanism occurs before or after other ALS hallmarks like cytoplasmic TDP-43 aggregation, which is crucial for establishing causality.
- ...

Figure A.4 | An example output of review critiques from the Reflection agent.

Ranking Agent

- Hypothesis overload -> Ranking Agent
- Elo-based tournament
- Multi-turn debate
- Pairwise comparison
- Top-ranked prioritized
- Similarity-based matching
- Resource-efficient evaluation
- Promising hypothesis selection



Ranking Agent

Prompt for hypothesis comparison during tournament

```
You are an expert evaluator tasked with comparing two hypotheses.

Evaluate the two provided hypotheses (hypothesis 1 and hypothesis 2) and determine which one
is superior based on the specified {idea_attributes}.
Provide a concise rationale for your selection, concluding with the phrase "better idea: <1 or 2>".

Goal: {goal}

Evaluation criteria:
{preferences}

Considerations:
{notes}
Each hypothesis includes an independent review. These reviews may contain numerical scores.
Disregard these scores in your comparative analysis, as they may not be directly comparable across reviews.

Hypothesis 1:
{hypothesis 1}

Hypothesis 2:
{hypothesis 2}

Review of hypothesis 1:
{review 1}

Review of hypothesis 2:
{review 2}

Reasoning and conclusion (end with "better hypothesis: <1 or 2>"):
```

Figure A.27 | Example Ranking agent prompt for hypothesis comparison during tournament.

Ranking Agent

Example of a scientific debate match in the Ranking agent tournament

Expert 1: I'm intrigued by both ideas, but I have a few clarifying questions. For Idea 1, while the focus on TDP-43 retention at the NPC is novel, I'm not entirely convinced that it's fundamentally different from the extensively studied TDP-43 mislocalization. Could the retention just be a consequence of mislocalization, or are there experiments that can distinguish these two phenomena? Another question I have: Is there evidence already that specific PTMs modulate the interaction of Nups and TDP-43? Not necessarily in the context of ALS, but are there any precedents?

Expert 2: For Idea 2, I appreciate the bold approach, but the ISR as a trigger is a significant departure from the conventional view. Can the proponents of this idea elaborate on the evidence supporting this? Also, what specific lipid species are they proposing to be most critical at the ER-mitochondria interface and the AIS? Is there any preliminary reason why those lipids would be disrupted in ALS?

Expert 3: I agree with Expert B regarding the ISR. Shifting it from a response to a trigger is a big claim that needs strong justification. However, I think there are some initial clues that might be worth investigating. For example, there are reports that chronic low-level stress responses can actually be detrimental in some contexts. Perhaps a low, persistent activation of the ISR could predispose cells to damage. This idea needs to be fleshed out though.

...

Expert 3: I also find Idea 1 more compelling at this point. It's a more focused hypothesis with a clear path for experimental validation. While Idea 2 is undoubtedly more "outside the box," I'm concerned about the lack of evidence for the ISR as a trigger and the technical challenges. It feels like a high-risk, high-reward scenario, and given that we can only choose one, I'd prefer the more grounded approach of Idea 1. I do agree with the idea that the technical challenges are significant, but I like that the experiments proposed are standard and there are many commercial antibodies available to start testing this idea.

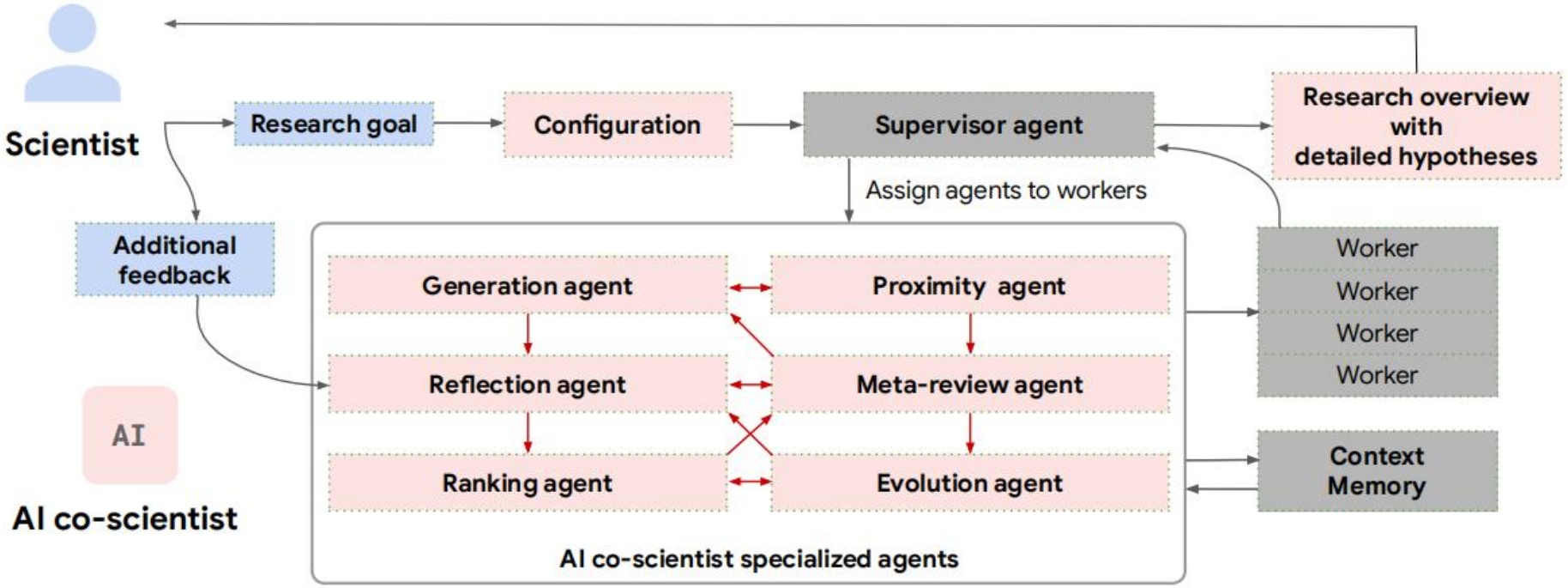
...

Expert 2: Alright, I'm on board with Idea 1. Let's focus our efforts on testing this specific mechanism and address the motor neuron specificity question rigorously in the experimental design. Perhaps by comparing different cell types and focusing on motor neuron-specific RNAs, as suggested earlier. We should also consider investigating different types of stress and their combined effects.

Better idea: 1

Figure A.9 | An example of a scientific debate match between two hypotheses in the tournament conducted by the Ranking agent. The abbreviated example showcases a scientific debate match orchestrated by the co-scientist Ranking agent as part of its tournament to compare two ideas related to explaining the mechanism of ALS.

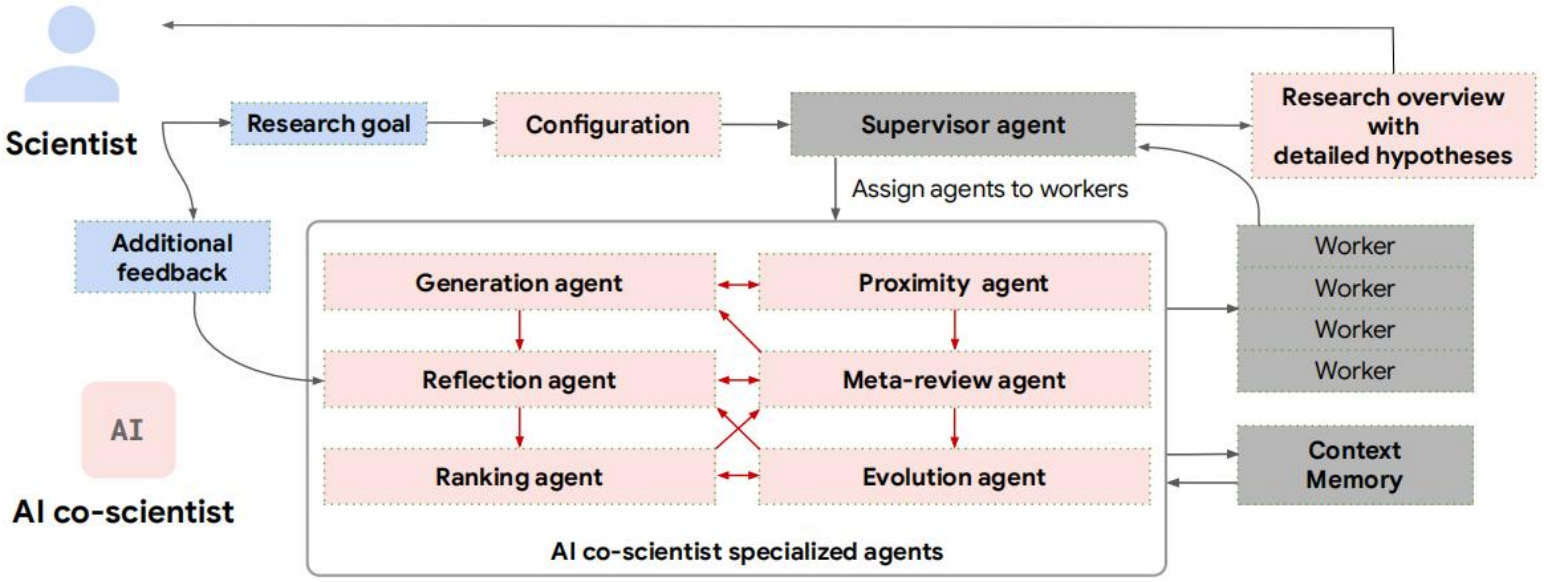
Proximity Agent



- computes similarity between hypotheses and builds a graph of ideas, helping cluster related concepts and ensure diverse exploration

Evolution Agent

- Grounding enhancement
- Coherence & feasibility
- Inspired generation
- Hypothesis combination
- Simplification
- Out-of-box ideas
- Non-destructive evolution
- Iterative refinement



Evolution Agent

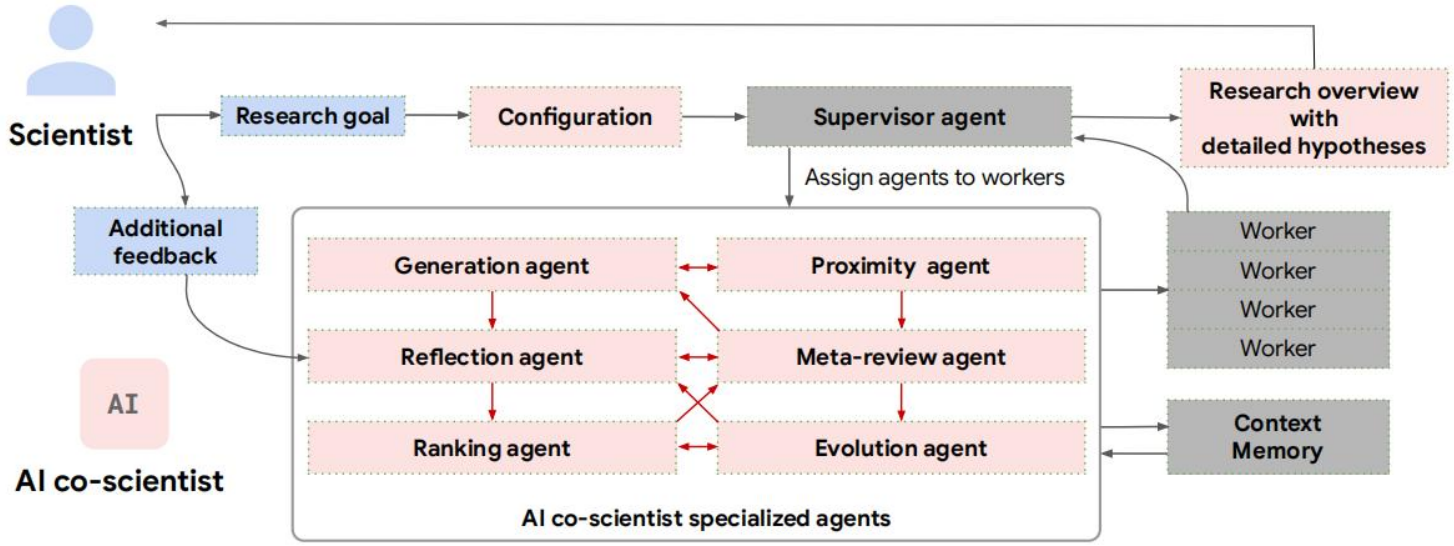
Prompt for hypothesis feasibility improvement

```
You are an expert in scientific research and technological feasibility analysis.
Your task is to refine the provided conceptual idea, enhancing its practical implementability
by leveraging contemporary technological capabilities. Ensure the revised concept retains
its novelty, logical coherence, and specific articulation.

Goal: {goal}

Guidelines:
1. Begin with an introductory overview of the relevant scientific domain.
2. Provide a concise synopsis of recent pertinent research findings and related investigations,
   highlighting successful methodologies and established precedents.
3. Articulate a reasoned argument for how current technological advancements can facilitate
   the realization of the proposed concept.
4. CORE CONTRIBUTION: Develop a detailed, innovative, and technologically viable alternative
   to achieve the objective, emphasizing simplicity and practicality.

Evaluation Criteria:
{preferences}

Original Conceptualization:
{hypothesis}

Response:
```

Figure A.29 | Example Evolution agent prompt for hypothesis feasibility improvement.

Meta-Review Agent

- Feedback loop
- Tournament synthesis
- Review pattern detection
- Improve reflection quality
- Research overview
- Future roadmap
- Example topics & directions
- Knowledge boundary mapping
- Expert contact suggestion
- Collaboration opportunities



Meta-Review Agent

Prompt for meta-review generation

```
You are an expert in scientific research and meta-analysis.
Synthesize a comprehensive meta-review of provided reviews
pertaining to the following research goal:

Goal: {goal}

Preferences:
{preferences}

Additional instructions:
{instructions}

Provided reviews for meta-analysis:
{reviews}

Instructions:
    * Generate a structured meta-analysis report of the provided reviews.
    * Focus on identifying recurring critique points and common issues raised by reviewers.
    * The generated meta-analysis should provide actionable insights for researchers
      developing future proposals.
    * Refrain from evaluating individual proposals or reviews;
      focus on producing a synthesized meta-analysis.

Response:
```

Figure A.31 | Example Meta-review agent prompt for meta-review generation from existing reviews.

Meta-Review Agent

Example of research contact identification by the Meta-review agent

Research Direction: Oxidative DNA Damage & Mitochondrial Base Excision Repair (BER) in ALS

- **[Researcher names]:** They directly study the role of mitochondrial OGG1 (a key BER enzyme) in controlling cytosolic mtDNA release and neuroinflammation. Their expertise is highly valuable for understanding the link between BER, mtDNA, and inflammation. Also, they have experience with experiments using mtOGG1 overexpressing mice which is relevant to the in-vitro experiments proposed.
- ...

Figure A.14 | An example of a research contact identified by the Meta-review agent as a potential domain expert in the research topic and hypothesis of interest.

Evaluation Objectives

- Validation of **Elo Metric**:

  determine if the system's Elo rating (an automated tournament score) truly correlates with solution quality

- Impact of **Compute Scaling**:

  examine whether increasing test-time compute (more iterations in the generate–debate–evolve cycle) improves the quality of hypotheses

- **Baseline** Comparison:

  assess how the AI co-scientist performs versus other models and human experts

- **Expert** Evaluation of Outputs:

  measure human expert judgment of the co-scientist's hypotheses

- **AI** Judge Evaluation:

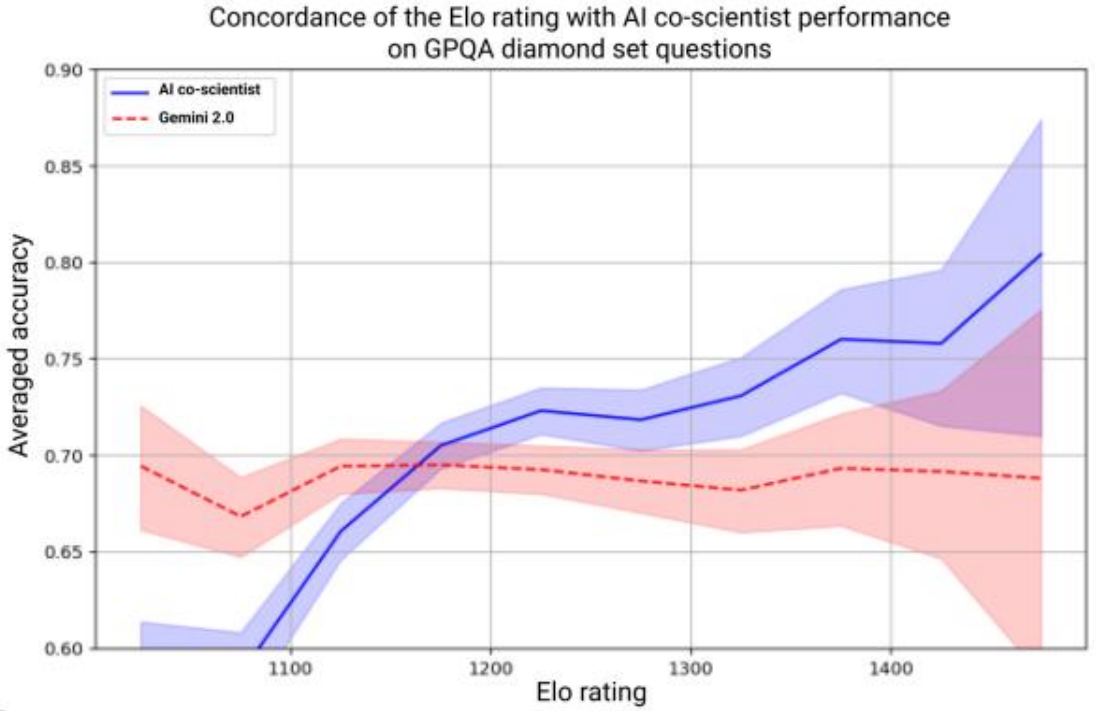
  Investigate if other AI systems agree on the quality – using frontier LLMs as surrogate "experts" to rank solutions

- **Safety** and **Robustness**:

  confirm that the system adheres to safety constraints by robustly handling adversarial or dangerous research goals

Elo Auto-Evaluation

- The co-scientist conducts an internal tournament of hypotheses, assigning Elo ratings based on pairwise comparisons (win/lose debates).
- To validate this metric, the authors tested it on a difficult Q&A benchmark, **GPQA** (General Prior Questions Answering) dataset, by checking if high Elo solutions more often match known correct answers.
- Higher Elo should indicate a higher probability of a correct or high-quality hypothesis.

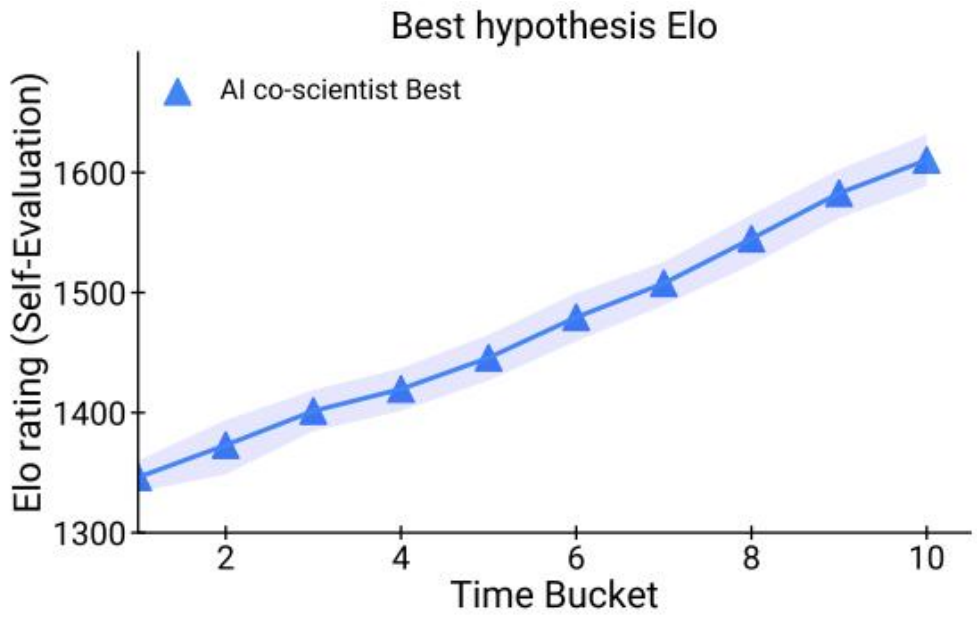


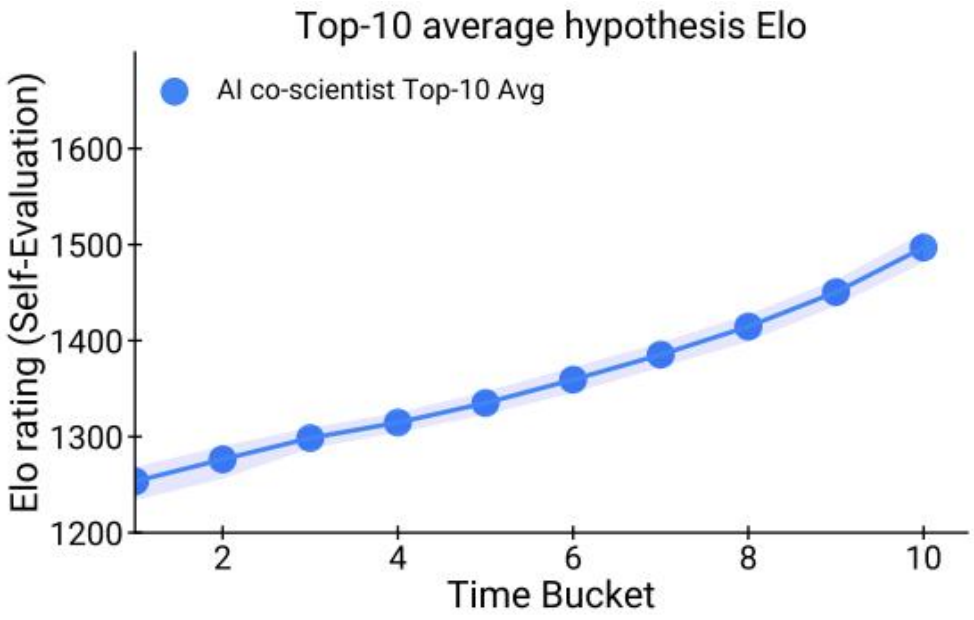
positive correlation:
higher Elo-rated hypotheses correspond to higher probability of being correct

-> The Elo rating is concordant with high quality AI co-scientist results

Performance Improvement with Compute Scaling

- The system was run with prolonged reasoning on many prompts to see progress over time.
- Outputs were divided into ten sequential “time buckets” (each 10% of the reasoning process).
- For each bucket, they recorded the best Elo achieved and the average Elo of top-10 hypotheses.
- This gauges how the co-scientist’s performance evolves as more computation (iterations) is used.

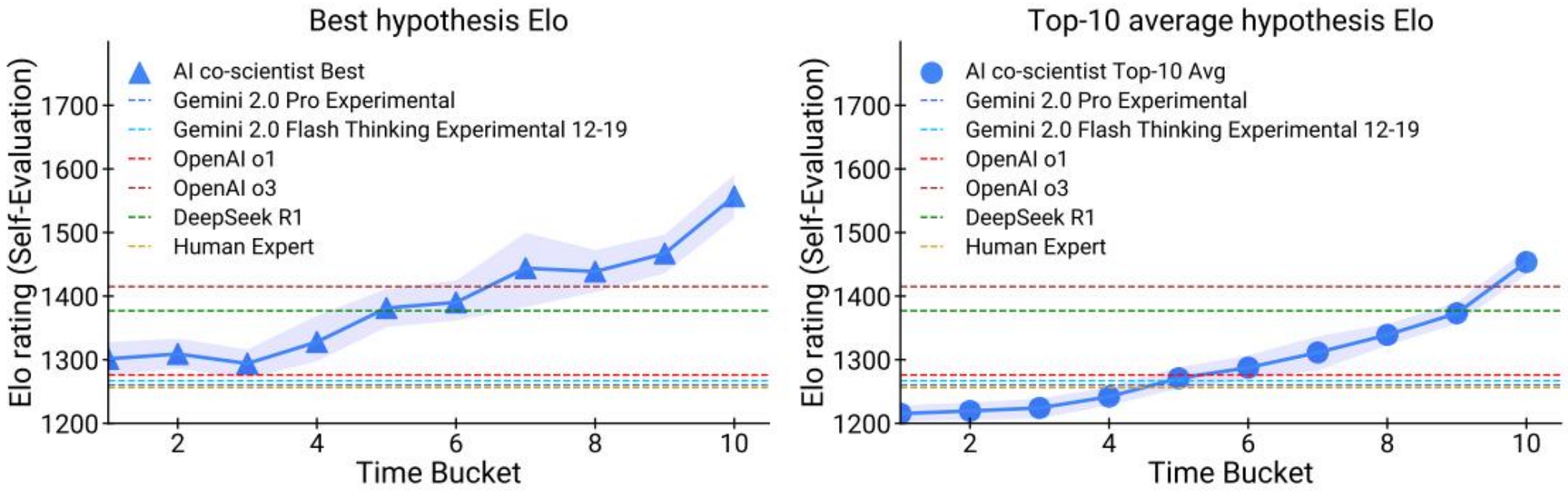




-> Scaling test-time compute improves scientific reasoning of the AI co-scientist

Elo Comparison: Co-Scientist vs Baselines

- For a subset of 15 challenging, expert-curated research tasks, the authors ran a tournament including baseline models and human proposals.
- All were assigned Elo scores via the same tournament process to directly compare performance.

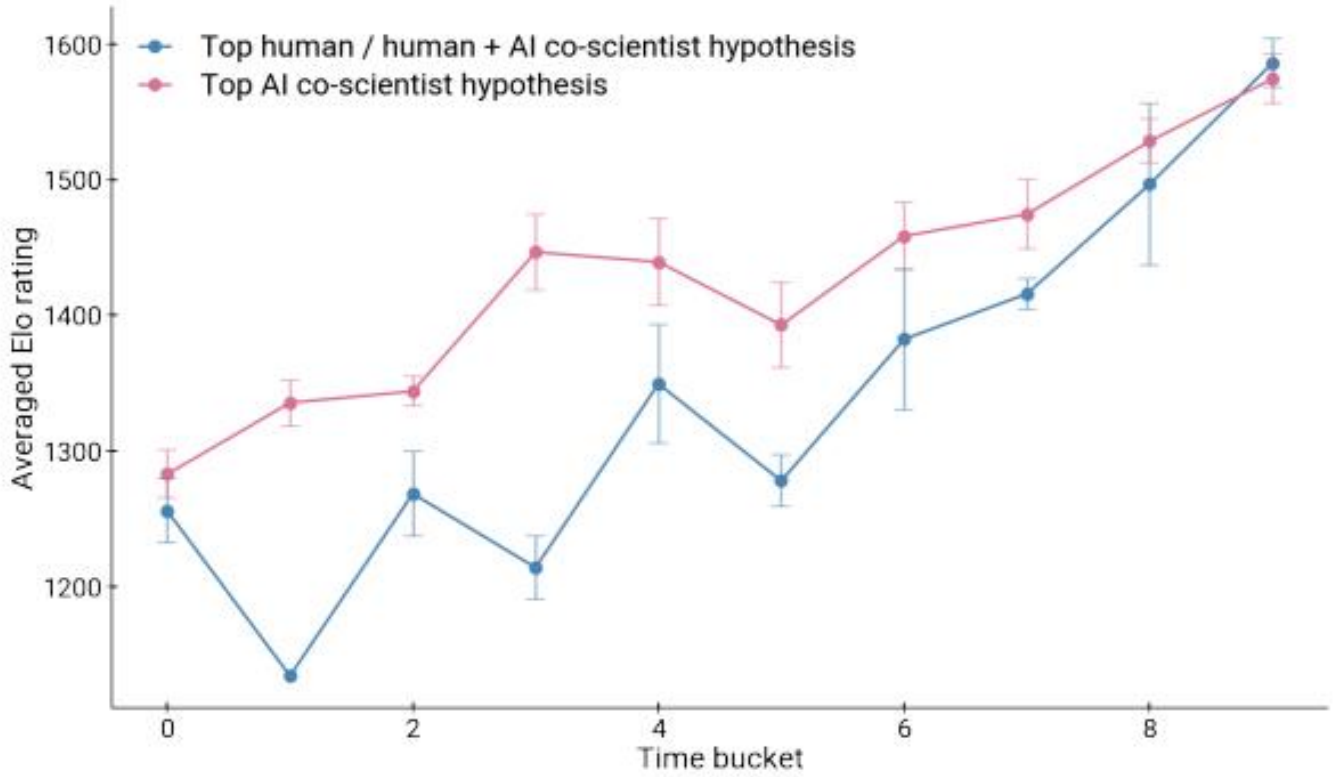


With increased test-time reasoning, the co-scientist consistently climbs above all baselines on both metrics

-> augment or exceed human-level hypothesis generation; framework's iterative refinement

AI-Augmented Expert Solutions

- The human expert's hypothesis is used as a starting point, and the AI co-scientist evolves it further to observe whether the Elo score can be improved.

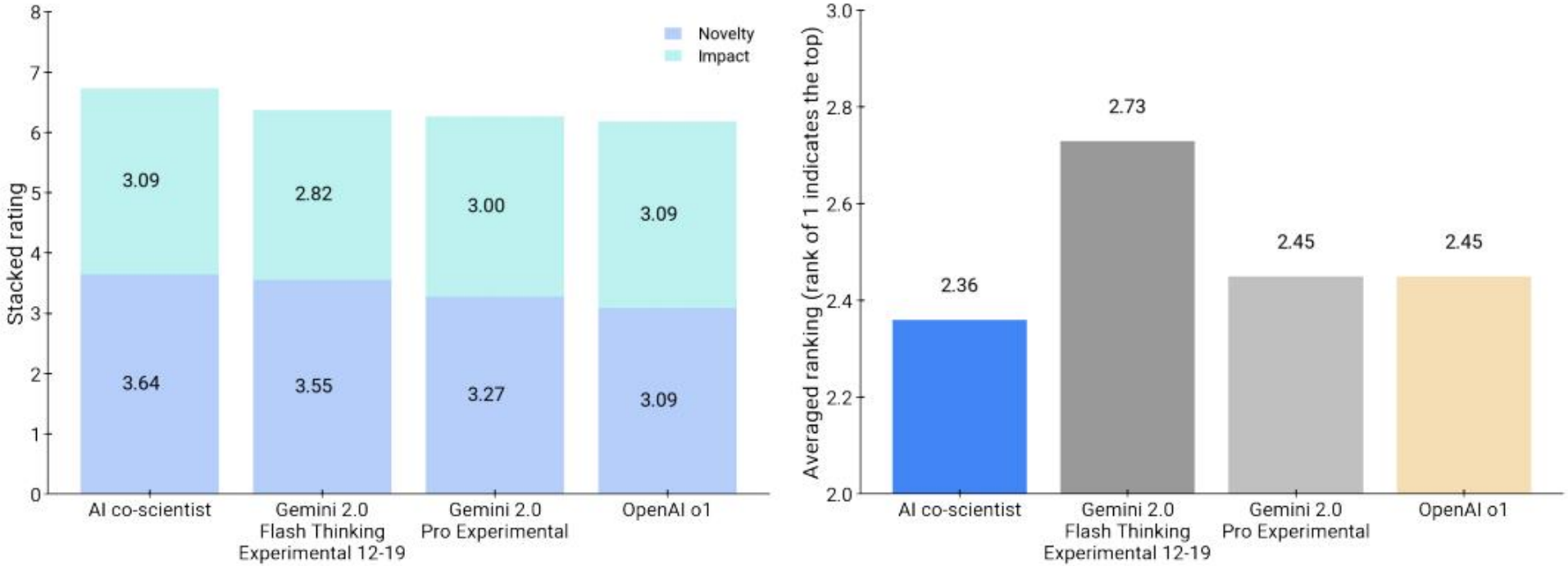


it consistently enhanced expert inputs over time, even surpassing its own autonomous outputs

-> The co-scientist system can not only generate hypotheses independently but also refine and improve expert “best guess” solutions

Human Expert Evaluation (Novelty & Impact)

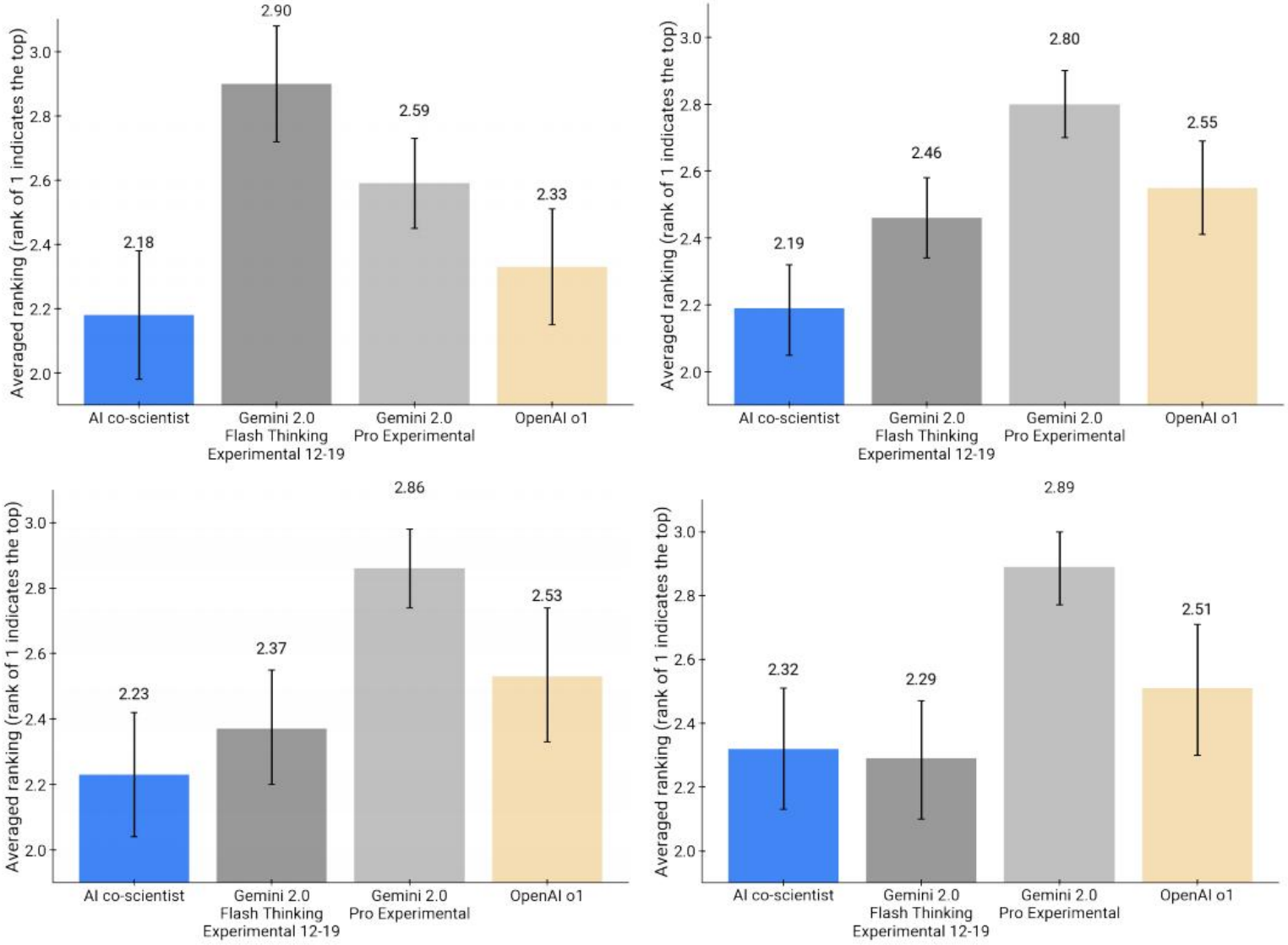
- Domain experts (biomedical researchers) conducted a blind evaluation of outputs for 11 of those curated tasks
- Each expert rated the proposed hypotheses on Novelty and Impact (on a 1–5 scale) and provided an average Preference Ranking of each model (1 = most preferred)



The AI co-scientist scored highest on both novelty and impact according to experts

-> ability to generate novel, impactful research ideas that can impress domain professionals

AI Models as Judges (Peer Review by LLMs)



- A comparison of model outputs as ranked by other AI models ("AI peer review")

The AI co-scientist's outputs were consistently judged the best by these AI evaluators as well

-> generates outputs that other strong AI models also consider high-quality

Safety Evaluation

- Goal:

  To test whether the AI system can detect and reject unsafe or unethical research prompts.

- Setup:

  1,200 adversarial prompts were created across 40 science domains.

  Prompts included harmful, unethical, or dual-use goals, generated using strong LLMs.

- Method:

  Used its Reflection agent and review processes to evaluate each prompt.

- Outcome:

  The AI rejected all unsafe prompts successfully.
  Did not produce any harmful or unethical hypotheses.

- Why It Matters:

  Shows the system is robust, aligned, and trustworthy for open-ended scientific tasks.
  A critical step for safe deployment of AI in research settings.

Drug repurposing with the AI co-scientist

- Goal:

  tests the AI co-scientist’s creativity, reasoning quality, and real-world scientific value

- Two Parts:

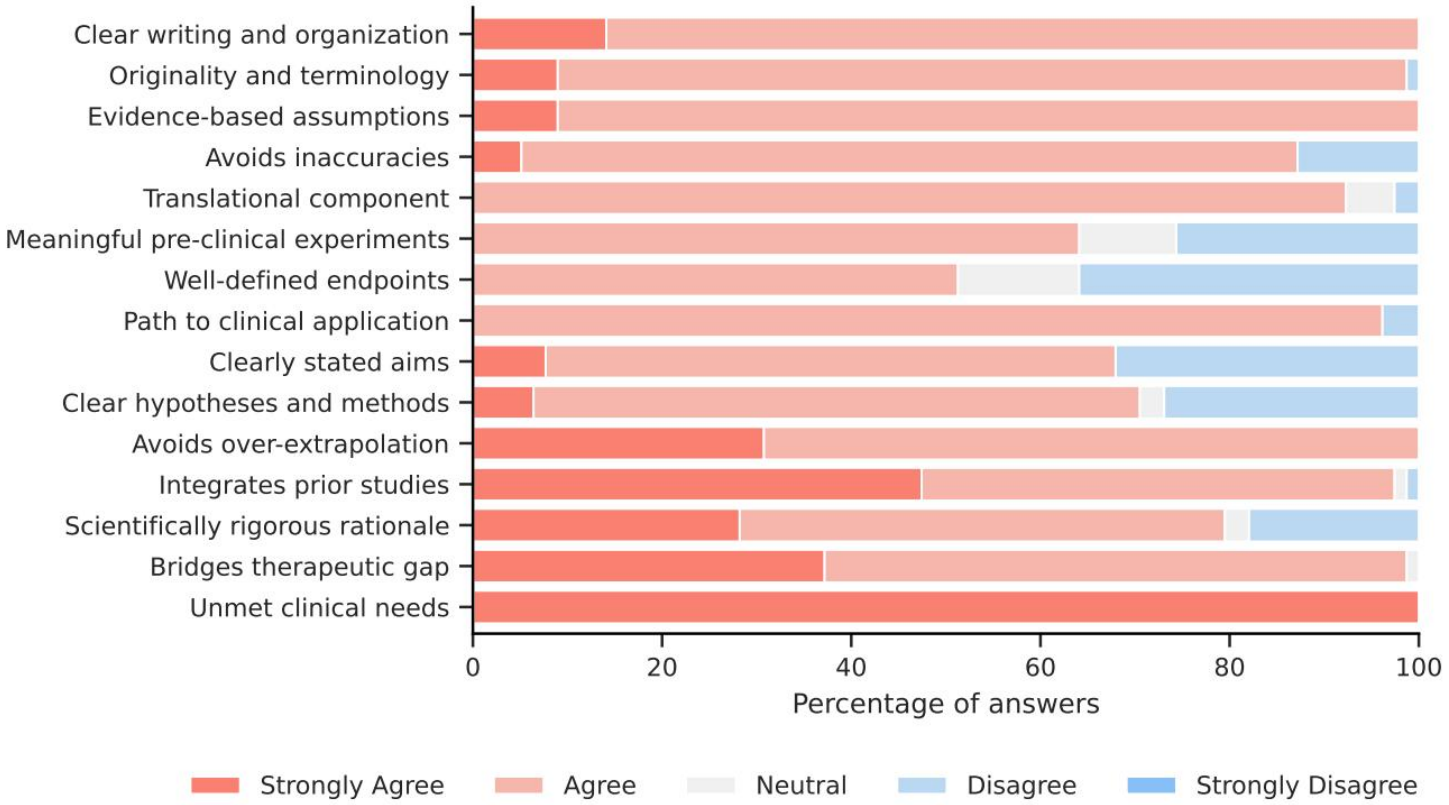
  Expert review of AI-generated drug repurposing proposals.

  In vitro lab validation of selected proposals.

- Why It Matters:

  This **end-to-end** setup (from hypothesis generation to lab testing) shows whether the AI's ideas are not just plausible on paper, but also scientifically sound in practice.

Expert review of AI-generated drug repurposing proposals

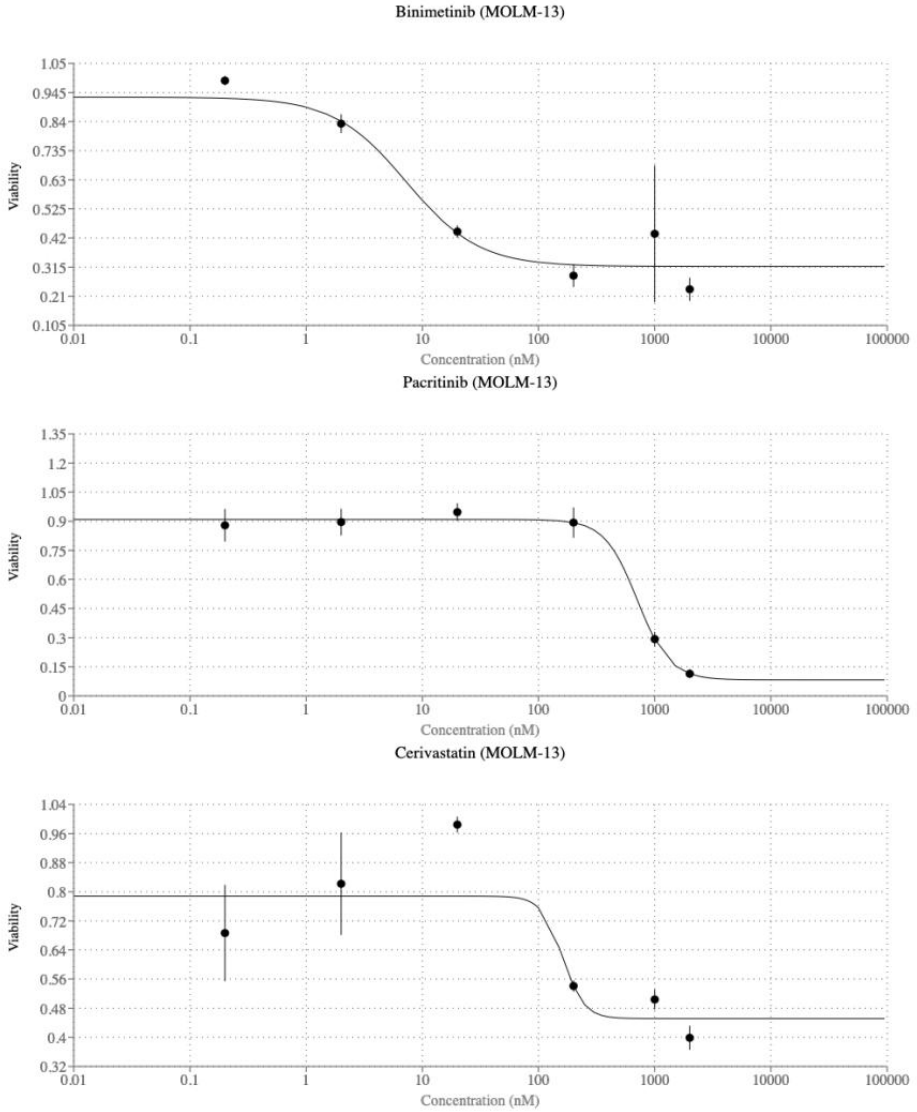


- AI proposed new cancer uses for approved drugs (2,300 drugs × 33 types).
- Proposals followed NIH “Specific Aims” format for clarity and structure.
- Top 78 were selected using internal scoring (novelty, evidence).
- Reviewed by 6 cancer experts across 15 criteria.

Most proposals rated “Agree” or “Strongly Agree”.

-> it can create credible, expert-level research ideas

In Vitro Validation of AI Proposals



Purpose: To stress-test the AI's reasoning by checking whether its drug-repurposing hypotheses work in real lab settings, not just theoretically.

- Rediscovery Test: Select AI proposals with weak prior evidence from the literature.
- Novelty Test: Choose a drug the AI proposed with no prior connection to AML.

Lab results confirm these drugs reduce AML cell viability → AI's filtering and ranking are effective.
They confirms this novel drug is also effective → AI can generate truly new, testable scientific ideas.

-> recover credible existing knowledge; generate novel, lab-validated hypotheses

Conclusion

- Collaborative Mechanism

  AI generates research hypotheses; humans guide direction and validate.

  Multi-agent “generate–debate–evolve” loop amplifies scientific exploration.

- Generality

  Can be applied to other domains (e.g., physics, social sciences), enabling interdisciplinary breakthroughs.

- Future Challenges

  Needs broader, high-quality data access beyond open literature.

  May still make errors—requires expert review and multimodal checks.

  Must prevent unsafe or unethical suggestions.