AI co-scientists for drug discovery

Accelerate Pharma R&D

NOVEMBER 2025

trail biomed

Table of content

A New Partner in Discovery 1

What “AI Co-Scientists” Really Do 3

The Emerging Stack of Scientific AI..... 3

The AI Co-Scientist Landscape..... 9

Developing Trail 1.0: A co-scientist for drug R&D..... 11

Limits and Lessons..... 13

Benchmarking and Validation..... 15

Outlook: Evolving the AI Co-Scientist..... 15

A New Partner in Discovery

If you've been watching the explosion of "AI co-scientists" with a mix of curiosity and skepticism, you're not alone [1]. In general terms, AI co-scientists are agents that reason, generate hypotheses, and support experimental design much like a collaborator would do.

Over the past year, several high-profile systems—from Google DeepMind's AI Co-Scientist [2] to Sakana AI's AI Scientist v2 [3], and most recently FutureHouse's Kosmos AI Scientist [4]—have claimed, and in some cases demonstrated, the ability to generate publishable research ideas or even predict experimental outcomes.

The question for many working in drug discovery is more specific: what could these systems actually do for preclinical drug R&D today? And, more practically, how should we navigate this rapidly growing ecosystem without losing time or rigour?

The purpose of this whitepaper is to make sense of this rapidly growing field—clarifying what these new tools can deliver and what responsible use might look like in modern drug discovery.

Drug development is among the most complex frontiers in science. AI co-scientists can accelerate this journey—but only when guided by careful, domain-aware design that ensures traceability and trust at every step.

AMALIO TELENTI, TRAIL BIOMED

What AI Co-Scientists Really Do

From ambition to execution

The current generation of AI scientific assistants are structured reasoning systems that learn from the scientific record. Given a problem statement, they read, debate internally, and propose hypotheses that seem plausible within existing knowledge.

Such systems have done impressive things:

- In collaboration with Stanford, an AI co-scientist correctly proposed two epigenetic drug classes that showed anti-fibrotic effects in *human hepatic organoids*. [5]
- An AI co-scientist replicated wet lab discovery of an unknown mechanism of bacterial gene transfer—essentially guessing a result that had taken human researchers nearly a decade to confirm. [6]
- Most recently, the Kosmos AI Scientist demonstrated coordinated, multi-agent reasoning across seven published research tasks—ranging from molecular property prediction to materials design—offering the clearest example yet of an integrated, cross-domain AI research framework. [4]

These examples show that AI can now assist with *hypothesis generation* and *experimental planning* in a broad range of scientific areas.

But they also illustrate the limits. Each success required **significant compute** (however most published studies do not disclose cost) and careful human oversight.

For every validated discovery, there are likely dozens of unreported or failed

runs—simply because these systems are expensive,opaque, and fragile outside of idealized settings.

That said, not all "AI co-scientists" are built alike. At one end of the spectrum are moonshot systems—technical demonstrations that show how far AI-driven reasoning can go when computation, coordination, and scale are pushed to their limits. Until recently, such systems were internal prototypes or enterprise research efforts—impressive, but largely out of reach for working scientists.

That changed with **FutureHouse's Kosmos**, announced in November 2025, the first AI scientist to demonstrate externally validated discoveries across biology, chemistry, and materials science—and to open its interface for public use. Built around long-running, multi-agent reasoning loops, Kosmos represents the first practical glimpse of an accessible, cross-domain AI scientist.

Even so, most researchers will continue to work with **more specialized, task-focused assistants**—systems that are narrower in scope but directly useful for everyday research. These tools don't autonomously run experiments or uncover new biology, but they meaningfully **augment how scientists think and decide**

Brainstorming and exploration — generating and organizing ideas, mapping connections across literature, and framing new research directions.

1. **Literature and evidence synthesis** — compressing weeks of reading into minutes by retrieving, summarizing, and cross-referencing sources with citations intact.
2. **Flexible data retrieval and review** — using natural-language search to access, filter, and summarize structured datasets for contextual analysis.
3. **Guided analysis and workflow planning** — recommending analytical methods, datasets, or experimental tools (as seen in open frameworks such as Biomni, ToolUniverse, or Superbio AI).
4. **Hypothesis generation** — proposing plausible causal links or mechanisms grounded in data, prior results, and mechanistic reasoning.
5. **Reasoning and critique** — evaluating hypotheses against evidence, highlighting contradictions, and suggesting refinements.
6. **Experimental planning** — proposing assays, prioritizing compounds, drafting protocols.

Together, these capabilities define what makes an AI assistant a co-scientist: it doesn't merely retrieve or predict, but **participates in the reasoning loop**—moving fluidly between evidence, interpretation, and design.

With that distinction in mind, the next sections examine the tools that researchers can access today—their design philosophies, levels of autonomy, and how they fit into the evolving landscape of AI-assisted preclinical discovery.

The Compute Cost of Intelligence

AI co-scientists demand far more computation than standard language models. Systems such as Google's Gemini-based agents run multiple reasoning paths in parallel, scaling inference costs **five- to six-fold than LLMs** even with optimized task scheduling. One run on Kosmos AI Scientist costs 200 credits (~$200). Sustained biomedical reasoning workflows can reach **tens of thousands of dollars per month** in cloud-compute expenses, driven by asynchronous task execution and inter-agent coordination.

The Emerging Stack of Scientific AI

Layers of intelligence

Understanding how AI participates in scientific discovery requires examining how autonomous and deeply they integrate reasoning, data, and experimentation.

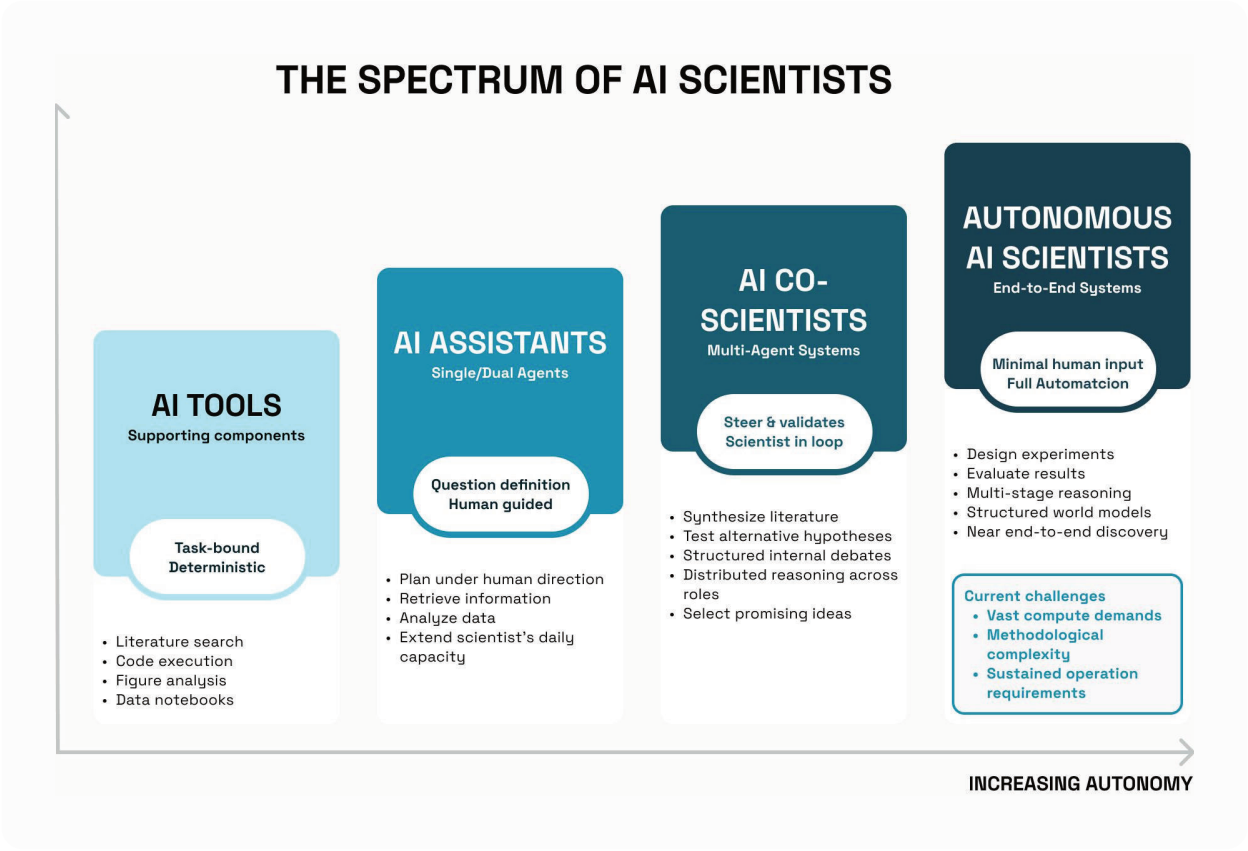
The simplest AI tools have **supporting roles**: search tools for literature access, code execution environments, vision-language models for figure critique, and data notebooks for analysis. These are deterministic and reliable—indispensable, but task-bound.

AI assistants build on these foundations. They are typically single or dual agents that plan, retrieve, or analyze under human direction. They extend what a scientist can do but still rely on human guidance for defining questions and interpreting results.

AI co-scientists are multi-agent systems that distribute reasoning across specialized roles. They synthesize literature, test alternative hypotheses, and even engage in structured internal debates to select the most promising ideas. Despite their sophistication, the human scientist remains in the loop, steering interpretation and validation. The *AI Co-Scientist* [2] exemplifies this collaborative model.

At the frontier are **autonomous AI scientists** such as *Kosmos* or *AI Scientist-v2*, which attempt end-to-end discovery: designing and evaluating experiments through multi-stage reasoning and structured world models. These systems offer a glimpse of what full automation could achieve—but also highlight the compute and methodological demands required to sustain it.

In practice, most platforms—including *Trail 1* , discussed below—blend elements across these layers: established analytical pipelines for reproducibility, assistant-level agents for synthesis, and collaborative reasoning for hypothesis generation. Progress lies not in removing the human, but in making the integration between human judgment and machine reasoning increasingly seamless.



The AI Co-Scientist Landscape

Current systems and directions

A growing list of companies now position themselves as "AI copilots" for discovery.

Commercial AI platforms are typically closed and proprietary. They promise acceleration, but they rarely disclose benchmarks or allow independent validation. Most operate under partnership or enterprise models—meaning that access, and thus reproducibility, remain limited. Within this commercial landscape, only a handful of platforms demonstrate features we would consider "co-scientist-like"—systems that reason, generate hypotheses, and support experimental design. Tools such as *Causaly*, *BenchSci ASCEND*, and *CytoReason* exemplify this emerging category.

At the opposite end are *research-grade* or *open frameworks*—for example **Biomni** [7], which attempts to map the entire biomedical "action space" by connecting models, databases, and protocols across disciplines. These are invaluable for experimentation, but they release code/benchmarks under open licenses with standard "as-is" disclaimers and no performance guarantees. In this space, beside **Biomni**, **FutureHouse's Kosmos** come closest to genuine research collaborators. They are both built as agentic frameworks that plan and execute workflows.

We can also differentiate AI systems based on how specific the tasks they are designed for. Generalist AI assistants, such as **Biomni** or **FutureHouse's Kosmos**, aim to reason broadly across scientific domains—they

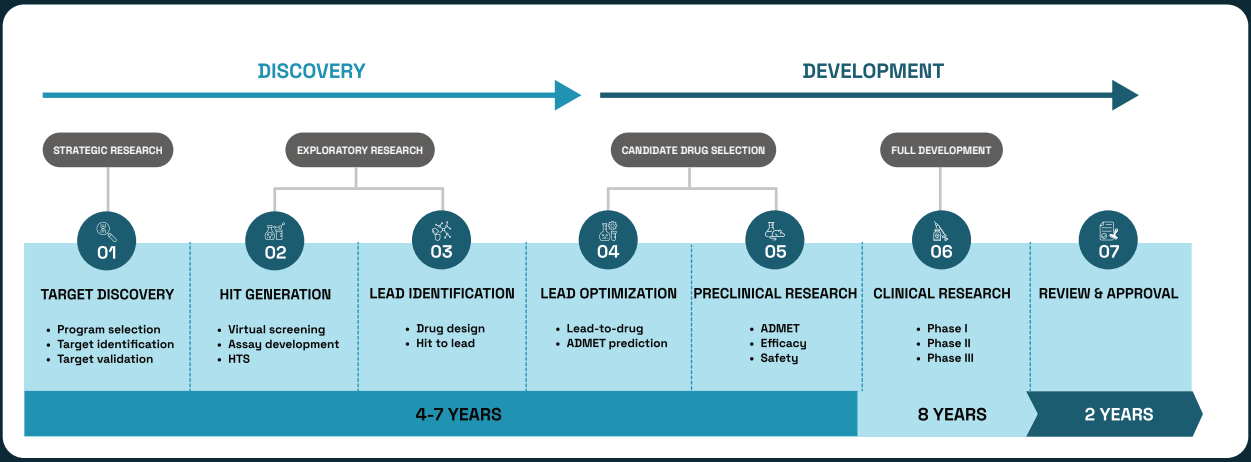
can read literature, compose analyses, and even assemble workflows spanning biology, chemistry, and physics. These systems strength lies in flexibility and exploration: they are ideal for open-ended brainstorming, early hypothesis scoping, or connecting distant concepts. Recent advances, like Biomni's Know-How Library, show how generalist systems are beginning to capture procedural expertise—bridging the gap between broad reasoning and domain-specific practice. Task-specific systems like **Trail 1** , by contrast, are built for depth within a defined workflow. They integrate domain data, established statistical methods, and validation logic to produce reproducible, actionable outputs in a specific context—for example, target identification for drug development using functional genomics data. In practice, the two are complementary: generalist systems help frame the right questions; specialized ones ensure those questions are answered rigorously and transparently within the boundaries of experimental science.

The autonomy of the agentic infrastructure can be constrained by specifying tools or sources. However, higher-order workflows—such as those of drug development—call for an explicit design of the steps, requirements and outputs, and stringent quality control.

Amalio Telenti, Trail Biomed

CASE STUDY

Developing Trail 1.0: An AI co-scientist for drug R&D



Drug R&D provides a clear example of how task-specific AI systems can integrate into structured scientific processes. **Trail 1** was built precisely for this environment.

Drug development follows a defined sequence of stages that transition from preclinical discovery to clinical validation. The **preclinical phase** encompasses target identification, druggability assessment (e.g., target tractability), optimization from early chemical matter to a lead compound, and evaluation of ADMET properties (absorption, distribution, metabolism, excretion, and toxicity). This stage also includes a comprehensive review of scientific precedent, market potential, intellectual property landscape, and the intended clinical population (**Table below**).

The **clinical phase**spans Phases I through III clinical trials and continues with post-marketing surveillance to ensure long-term safety and efficacy.

TRAIL 1 integrates AI co-scientist technologies into this standardized framework for preclinical R&D. The system directs AI research tools into defined, task-specific programs rather than broad generalist models, allowing precise execution of each step in the drug development process. This **task specificity**limits AI autonomy, validates data sources, and enhances scrutiny of AI-generated hypotheses. Furthermore, the task-specific AI co-scientist oversees and curates collective research data across a consortium or pharmaceutical organization, maintaining a consistent and growing scientific context. In essence, it establishes an**institutional memory**that preserves experimental data, results, and validation—past and present.

Druggability Assessment	Competitive Intelligence
Target Validation and Biological Relevance: Confirm that the target plays a critical, disease-relevant role and is non-redundant within its biological pathway.	**Patent Landscape:**Identify existing patents covering the target, related compounds, or mechanisms of action; assess novelty and freedom-to-operate.
Pathological Relevance: Demonstrate association with disease progression and therapeutic benefit from modulation (e.g., inhibition, activation).	**Key Patent Holders:**Map leading academic groups, biotech firms, or pharmaceutical companies holding key intellectual property.
Homology and Precedence: Evaluate conservation across species and precedent for druggability in related protein families.	**Patent Status and Lifespan:** Determine patent maturity, expiration timelines, and implications for exclusivity and entry barriers.
Structural and Sequence-Based Assessment: Assess presence of druggable domains, defined binding pockets, and available high-resolution structural data.	**Development Stage:** Track competitors by preclinical and clinical phase; identify first-in-class versus fast-follower positioning.
Ligandability: Identify known ligands, tool compounds, or fragments demonstrating binding feasibility.	**Approved or Pipeline Drugs:**Note approved drugs or pipeline candidates targeting similar mechanisms or pathways.
Chemical Tractability: Evaluate feasibility of small-molecule or biologic targeting; assess structure-based design potential.	**Companies and Players:**Identify organizations active in this target space and their strategic focus areas.
Experimental Validation: Confirm druggability through biochemical, cellular, and in vivo assays demonstrating efficacy and selectivity.	**Market Potential:**Estimate commercial opportunity based on disease prevalence, unmet need, and therapeutic positioning.
Selectivity and Toxicity: Evaluate on-target and off-target risks; determine therapeutic window and safety profile.	**Competition and Differentiation:**Assess level of innovation, risk of redundancy, and potential for differentiation.
Risk Assessment: Continuously monitor for compensatory pathways, resistance mechanisms, and safety liabilities.	**Investment and Partnerships:** Track funding trends, collaborations, and licensing activity to gauge field momentum.
AI co-scientist Summary: Combine biological validation, structural tractability, and experimental proof-of-concept to determine overall druggability confidence.	**AI co-scientist Outlook:**Integrate patent, market, and pipeline data to identify strategic opportunities and risks for target development.

Limits and Lessons

Evaluating AI scientists

Evaluating AI scientists is a hot topic and actively researched. Several cited works below are preprints; findings should be interpreted accordingly.

One of the most comprehensive assessments examined **Sakana AI's "AI Scientist"** [8], a platform that aimed to automate the entire research cycle. The review found that nearly half of its experiments failed because of coding or logic errors, that novelty assessments consistently misclassified known methods as new, and that more than half of its manuscripts contained fabricated or inconsistent numerical results. Even its built-in peer-review agent misjudged human-written papers and overlooked fundamental methodological flaws.

Beyond individual systems, several **recurrent weaknesses** appear across AI-researcher workflows. Independent analyses identify four major failure modes [9]:

- **Inappropriate benchmark selection**, inflating performance through poorly matched evaluation tasks.
- **Data leakage**, where information from test sets contaminates training data, producing non-generalizable results.
- **Metric misuse**, optimizing quantities that do not correspond to genuine scientific quality or predictive value.
- **Post-hoc selection bias**, the digital equivalent of p-hacking—reporting only favorable outcomes after many hidden trials.

Such issues are difficult to detect once only the final manuscript remains visible. Studies show that **access to complete workflow logs**—including code, prompts, and intermediate reasoning steps—is essential for auditing AI-generated research.

A further complication in biomedical contexts is **hallucination**: large language models often invent citations, numerical values, or mechanistic links that appear plausible to non-experts but

are factually incorrect. In domains like drug discovery or disease biology, such errors can redirect effort and resources toward false leads.

These cautionary findings do not negate progress but highlight the conditions under which AI co-scientists can be trusted. Reliable discovery demands transparency, reproducible workflows, and human oversight—principles that should guide the next generation of AI-driven research systems.

These failures underscore why systematic benchmarking is not peripheral but central to credible AI discovery.

Benchmarking and Validation

From ideas to biological truth

After seeing where AI co-scientists fail, the question is not whether they can be improved, but how we can evaluate progress meaningfully. The problem is that current assessments often stop at surface metrics — accuracy in text retrieval or correlation with known pathways — which say little about whether an idea would hold up in a living system. What we need instead are frameworks that connect **reasoning quality** to **biological validity**, so that we can tell which systems genuinely contribute to discovery rather than just generating plausible results.

A more practical path forward begins with **multi-tiered validation**. At the first tier, we should assess how internally consistent an AI's reasoning is — whether its logic, citations, and calculations hold together under scrutiny. Frameworks such as *SPOT* [10] for automated manuscript verification are an early step, but even state-of-the-art models detect only a fraction of significant scientific errors. The second tier involves testing whether hypotheses align with established principles of biology and chemistry. Benchmarks like *BioKGBench* [11], which evaluate how well agents distinguish verified facts from hallucinated ones, provide useful direction here. The third tier moves toward **surrogate validation**, where hypotheses are tested computationally or in controlled models — for example, running molecular-dynamics simulations on predicted ligands. The highest tier remains **experimental confirmation**, the point at which a prediction is shown to reproduce a real, measurable effect. Given cost and time constraints, not every idea can be tested this way, but systematic sampling — validating a representative subset of predictions — would already improve the field's credibility.

Beyond the hierarchy itself, what matters is transparency. Each system should be evaluated on the same problem sets and held to the same reporting standards: what data were used, what failed, and why. Emerging initiatives such as *IdeaBench*

[12] and *ResearchBench* [13] move in this direction by measuring the **novelty** and **feasibility** of AI-generated hypotheses relative to human ones. Their results are instructive — language models can produce ideas as novel as those in published papers, yet these ideas are consistently less feasible. That gap between novelty and practicality is exactly what future benchmarks must expose and help narrow.

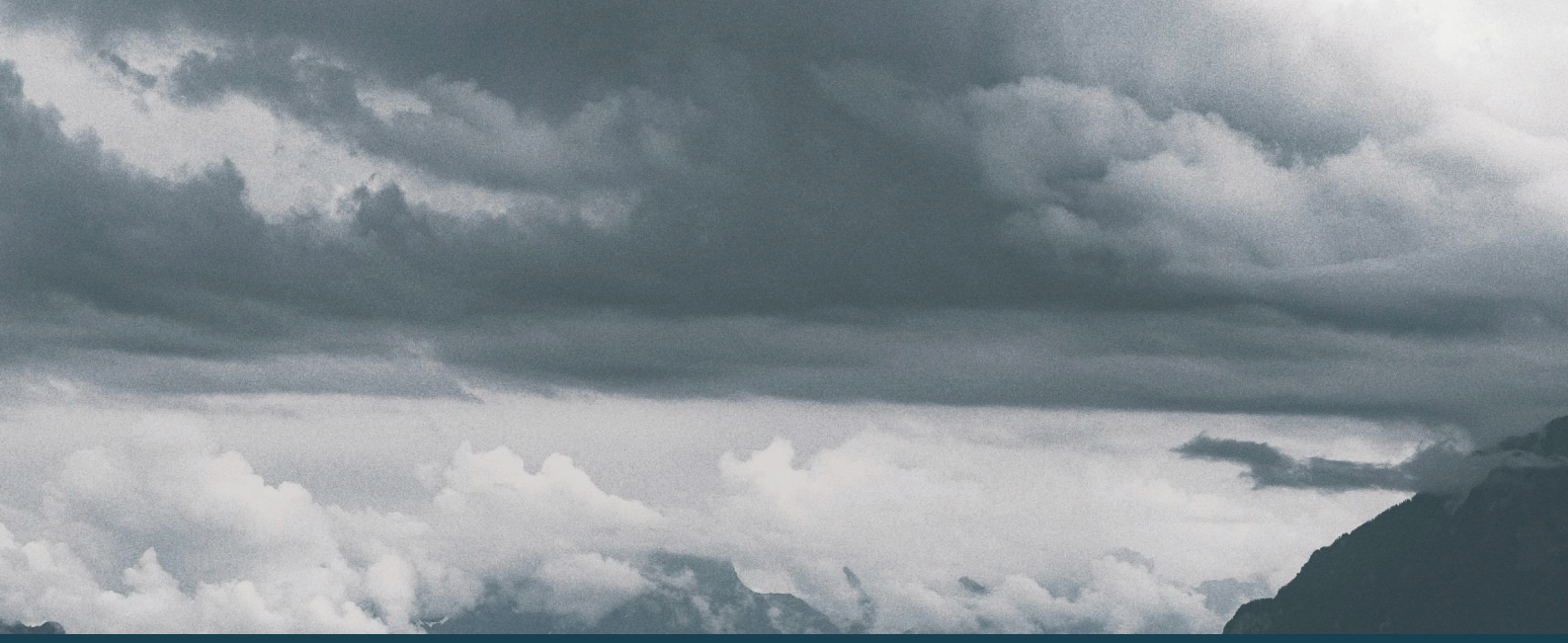
Equally important are the practices that prevent self-deception: guarding against data leakage, pre-registering experiments, and documenting negative outcomes. Too many AI researchers still publish only the successful runs. If we want AI co-scientists to earn trust, they must be tested under the same standards as human science — repeatability, transparency, and critical self-review.

Once such validation becomes routine, the next challenge is to see how far these systems can go when connected directly to experimental data. That's the frontier now emerging around biological foundation models and lab-in-the-loop discovery.

At **Trail Biomed**, we approach hypothesis validation pragmatically. Supporting context from experimental data, literature review, and bioinformatics analyses converge in the hypothesis generation step. To ensure that AI-generated hypotheses are scientifically sound and actionable, Trail 1 evaluates three key dimensions:

1. **Meaning Consistency**: AI-generated hypothesis statements are evaluated for semantic coherence across repeated retrieval and synthesis cycles. Stable interpretation and reasoning over provided context guards against likely hallucinations.
2. **Information Consistency**: When integrating data from multiple sources and lines of evidence, Trail 1 tracks whether key claims in the rationale supporting a hypothesis statement are not variable across repeat cycles of hypothesis generation.
3. **Faithfulness (Claim-Context Alignment)**: Every claim in the rationale is traced back to the supporting context such as experimental data, bioinformatics analysis, or literature review Unsupported claims (or claims not explicitly in the supporting context) are flagged for review.

As a final step in prioritizing actionable hypotheses, those meeting the robustness criteria are re-evaluated through a more targeted cycle of AI-assisted literature and evidence retrieval to clarify remaining gaps, strengthen support, and re-assess for novelty.



Outlook: Evolving the AI Co-Scientist

The next generation of AI co-scientists will be built on **biological foundation models**—large, pre-trained systems that learn directly from multimodal experimental data to capture how molecules, pathways, and cells behave in context. The most ambitious expression of this idea is the AI Virtual Cell initiative, which seeks to model these processes across scales in a single, data-driven computational framework [14].

As these models mature, their usefulness will increase with **lab-in-the-loop** cycles that connect computation and experiment in real time. Here, each experiment informs the model, and each model update proposes the next experiment—a continuous reasoning loop between AI and wet lab. This is where the idea of a co-scientist becomes tangible: not a static assistant, but an adaptive partner whose hypotheses evolve with evidence.

Recent studies already hint at this convergence. In *The Virtual Lab*, Swanson et al. (2024) demonstrated an AI system that autonomously designed new SARS-CoV-2 nanobodies and iteratively refined them through wet-lab validation—a practical glimpse of what foundation-model-driven, lab-in-the-loop discovery can achieve [15].

Foundation models and lab-in-the-loop discovery together define the future trajectory of AI co-scientists: reasoning agents grounded in biology, guided by data, and accountable to experiment. Our aim in this process is in the successful and trusted implementation of AI co-scientist in the standard and regulated processes of drug R&D.

References

1. Van Noorden R, Perkel JM. AI and science: what 1,600 researchers think. Nature. 2023;621(7980):672-675. doi:https://doi.org/10.1038/d41586-023-02980-0
2. Gottweis J, Weng WH, Daryin A, et al. Towards an AI co-scientist. arXiv.org. Published 2025. https://arxiv.org/abs/2502.18864
3. Yamada Y, Lange RT, Lu C, et al. The AI Scientist-v2: Workshop-Level Automated Scientific Discovery via Agentic Tree Search. arXiv.org. Published 2025. Accessed April 18, 2025. https://arxiv.org/abs/2504.08066
4. Mitchener L, Yiu A, Chang B, et al. Kosmos: An AI Scientist for Autonomous Discovery. arXiv.org. Published 2025. Accessed November 6, 2025. https://arxiv.org/abs/2511.02824
5. Guan Y, Cui L, Jakkapong Inchai, et al. AI-Assisted Drug Re-Purposing for Human Liver Fibrosis. Advanced Science. Published online September 14, 2025. doi:https://doi.org/10.1002/advs.202508751
6. Penadés JR, Gottweis J, He L, et al. AI mirrors experimental science to uncover a mechanism of gene transfer crucial to bacterial evolution. Cell. Published online September 9, 2025. doi:https://doi.org/10.1016/j.cell.2025.08.018
7. Huang K, Zhang S, Wang H, et al. Biomni: A General-Purpose Biomedical AI Agent. bioRxiv (Cold Spring Harbor Laboratory). Published online June 2, 2025. doi:https://doi.org/10.1101/2025.05.30.656746
8. Beel J, Kan MY, Baumgart M. Evaluating Sakana's AI Scientist: Bold Claims, Mixed Results, and a Promising Future? ACM SIGIR Forum. 2025;59(1):1-20. doi:https://doi.org/10.1145/3769733.3769747
9. Luo Z, Kasirzadeh A, Shah NB. The More You Automate, the Less You See: Hidden Pitfalls of AI Scientist Systems. arXiv.org. Published 2025. Accessed November 4, 2025. https://arxiv.org/abs/2509.08713
10. Son G, Hong J, Fan H, et al. When AI Co-Scientists Fail: SPOT-a Benchmark for Automated Verification of Scientific Research. arXiv.org. Published 2025. Accessed November 6, 2025. https://arxiv.org/abs/2505.11855
11. Lin X, Ma S, Shan J, et al. BioKGBench: A Knowledge Graph Checking Benchmark of AI Agent for Biomedical Science. arXiv.org. Published 2024. Accessed November 6, 2025. https://arxiv.org/abs/2407.00466
12. Guo S, Hassan SA, Xiong G, et al. IdeaBench: Benchmarking Large Language Models for Research Idea Generation. arXiv.org. Published 2024. Accessed November 6, 2025. https://arxiv.org/abs/2411.02429
13. Liu Y, Yang Z, Xie T, Ni J, Gao B. ResearchBench: Benchmarking LLMs in Scientific Discovery via Inspiration-Based Task Decomposition. Arxiv.org. Published 2024. Accessed November 6, 2025. https://arxiv.org/html/2503.21248v1

1. Bunne C, Roohani Y, Rosen Y, et al. How to build the virtual cell with artificial intelligence: Priorities and opportunities. Cell. 2024;187(25):7045-7063. doi:https://doi.org/10.1016/j.cell.2024.11.015
2. Swanson K, Wu W, Bulaong NL, Pak JE, Zou J. The Virtual Lab of AI agents designs new SARS-CoV-2 nanobodies. Nature. Published online July 29, 2025. doi:https://doi.org/10.1038/s41586-025-09442-9

About us

Trail Biomed was founded with the mission of putting the power of artificial intelligence (AI) at the service of biomedical innovation.

Our approach integrates bioinformatics and biomedical data with large language models (LLMs) and scientific foundation models, delivering tailored data science and enterprise solutions to tackle complex challenges across the biomedical landscape.

trailbiomed.com

trail biomed