AI Agents for Accelerating Research and Discovery

James Zou

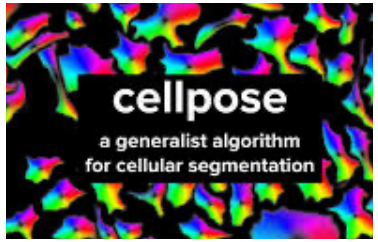
Stanford University

www.james-zou.com

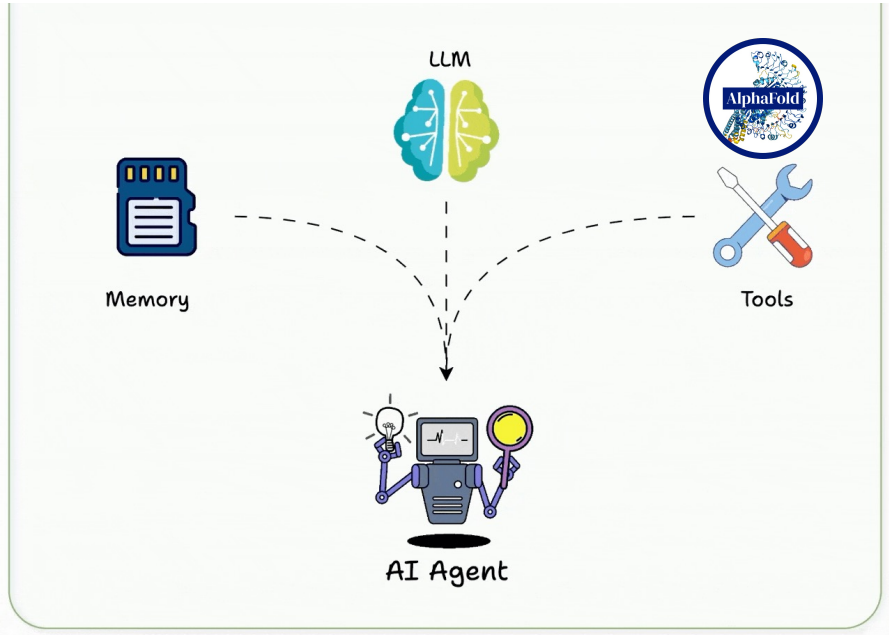
jamesz@stanford.edu

AI tool vs AI co-scientist

AI tools



AI agents

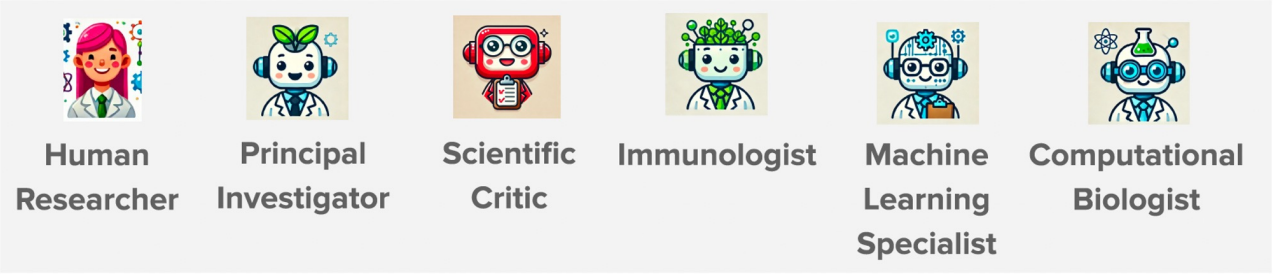


LLM systems that use tools

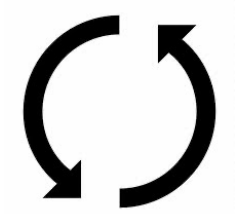
The Virtual Lab

The Virtual Lab is a team of interdisciplinary AI scientists that work with human scientist on challenging, open-ended research.

AI scientists



Self-learning agents

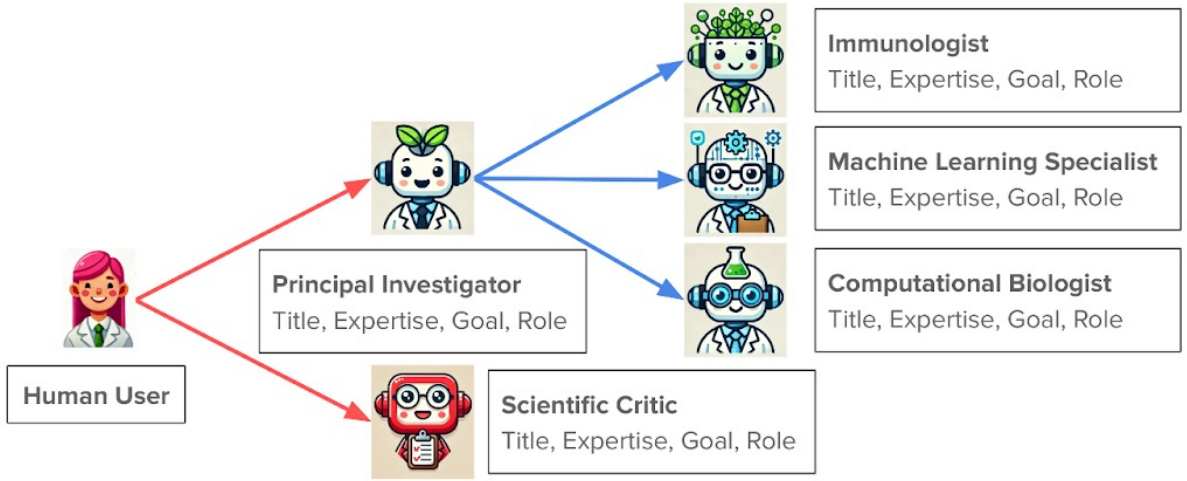


Virtual Lab School

Swanson et al. *Nature* 2025

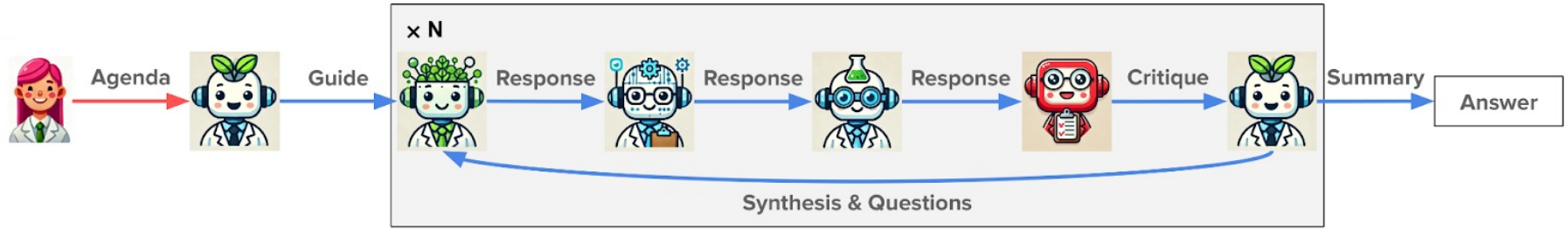
Virtual Lab Design

Agent creation

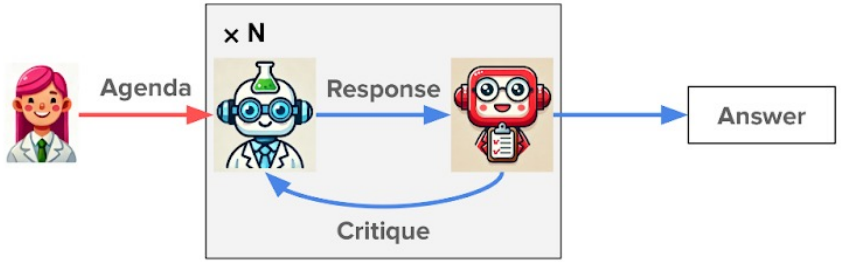


GPT-4o as the base model

Team meeting



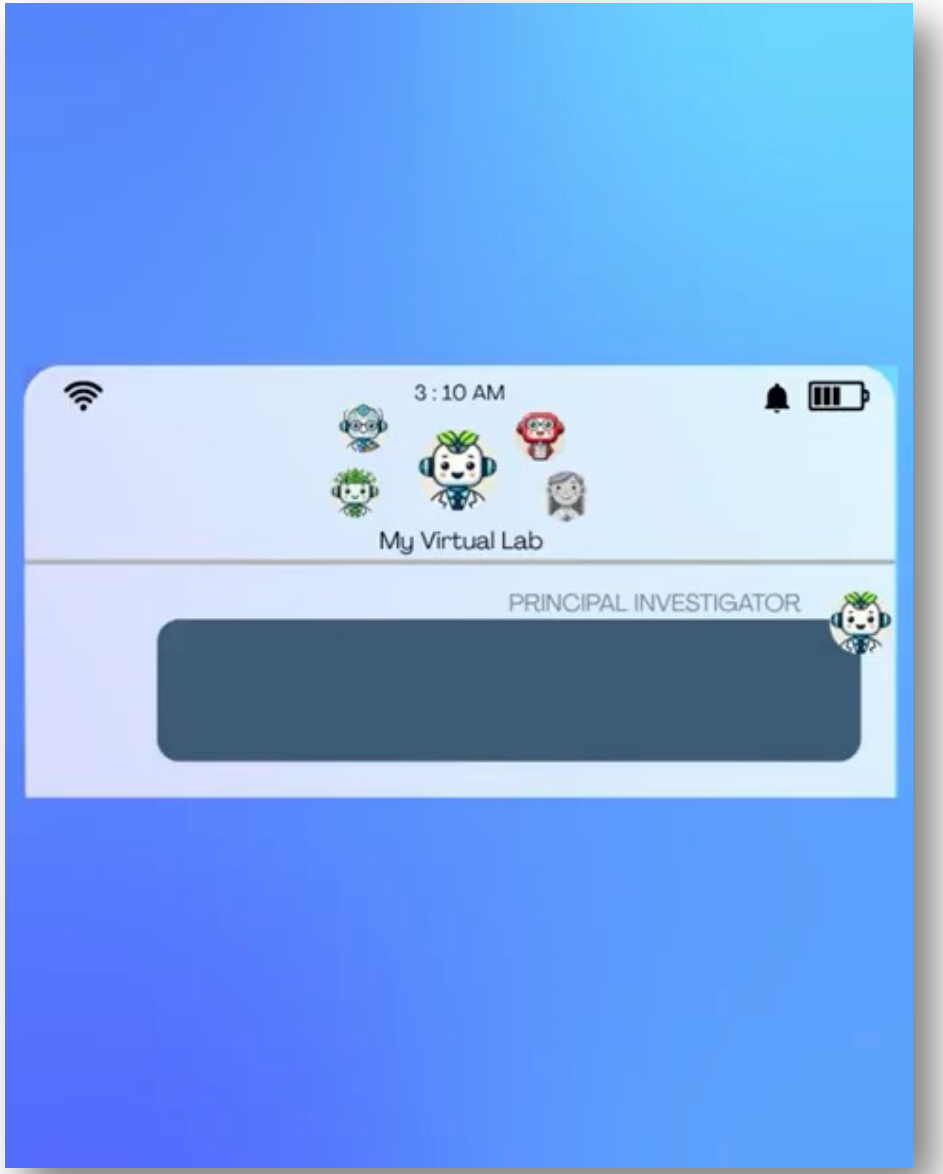
Individual meeting



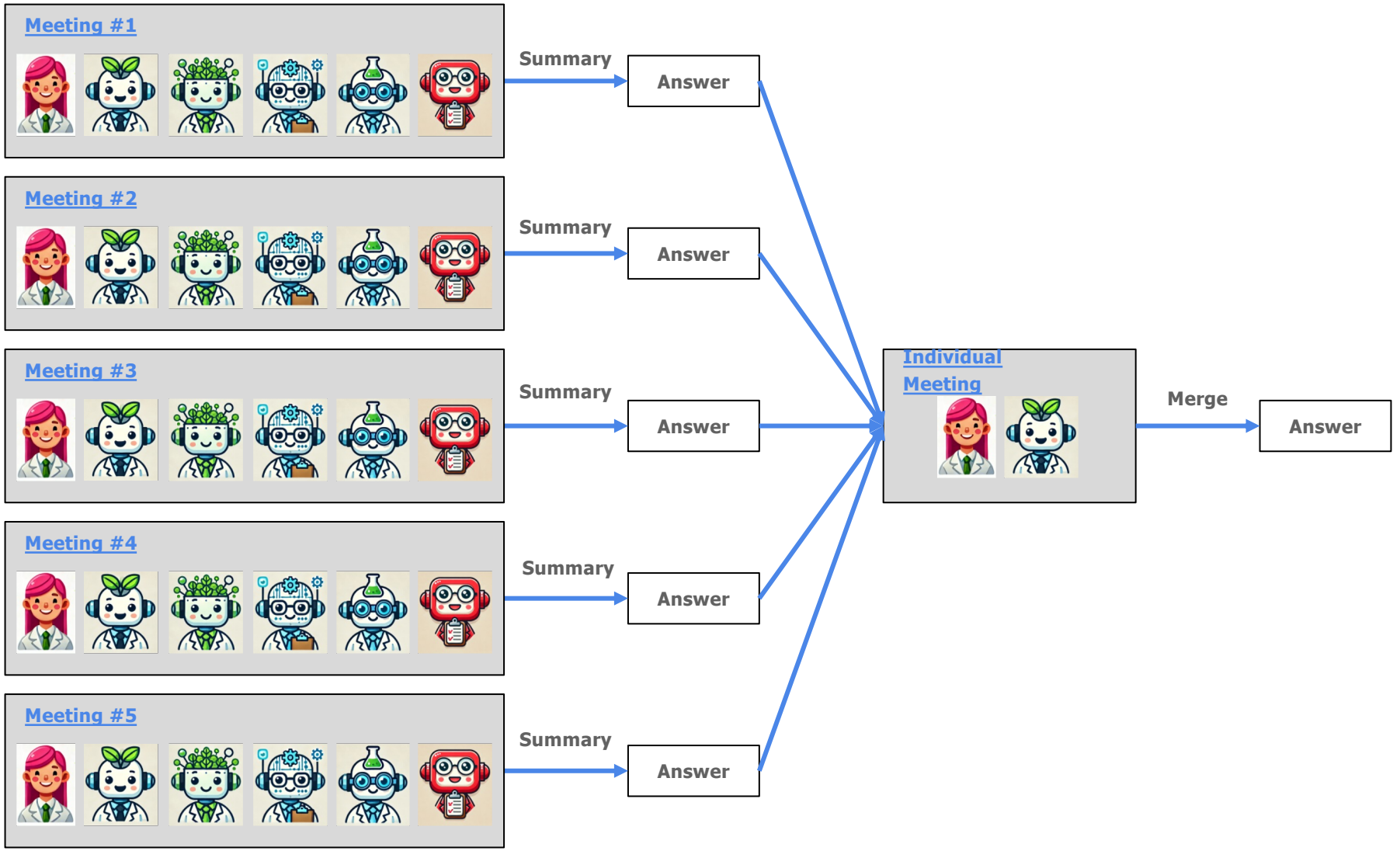
tools in sandbox environment

Swanson et al. *Nature* 2025

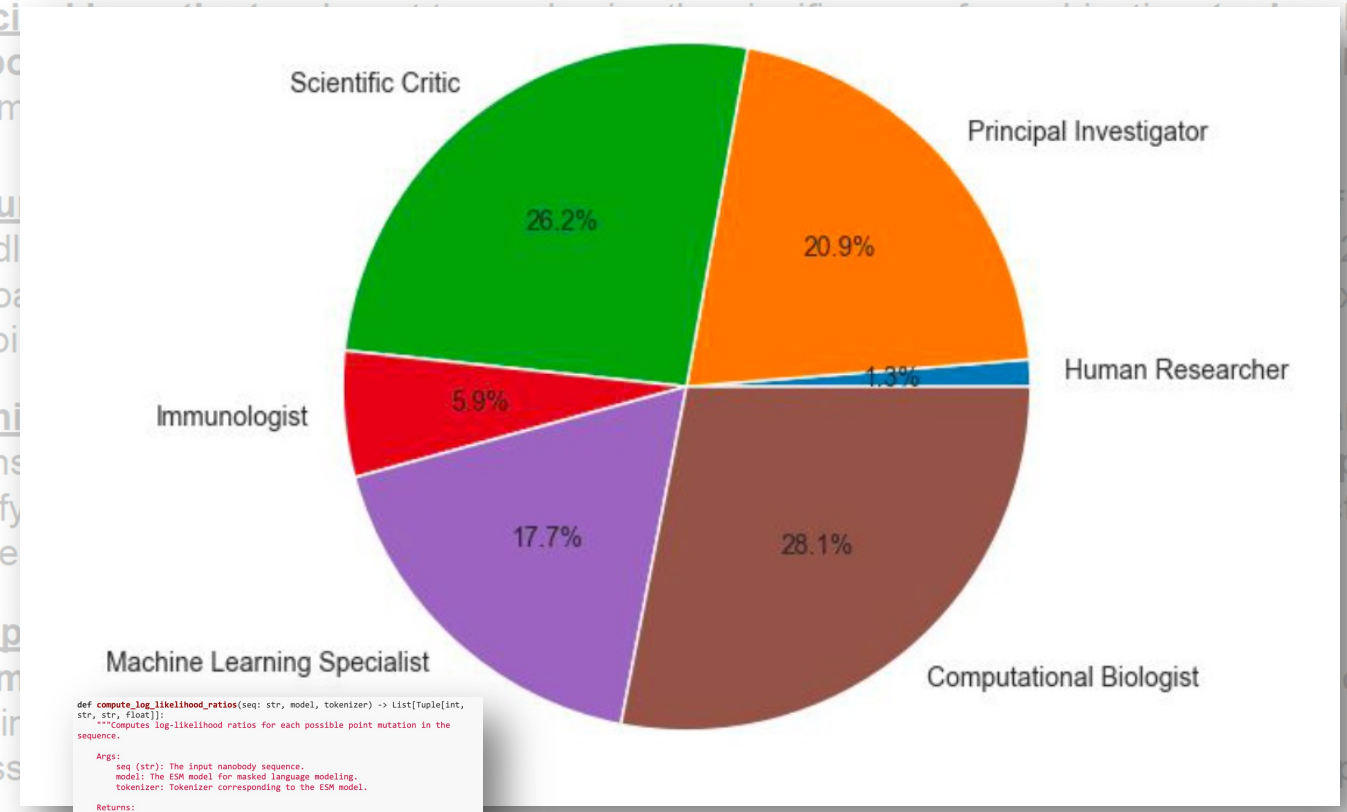
Example Virtual Lab team meeting



Virtual Lab: parallel meetings



Words written by different agents




```
def compute_log_likelihood_ratios(seq: str, model, tokenizer) -> List[Tuple[int, str, str, float]]:
    """Computes log-likelihood ratios for each possible point mutation in the sequence.

    Args:
        seq (str): The input nanobody sequence.
        model: The ESM model for masked language modeling.
        tokenizer: Tokenizer corresponding to the ESM model.

    Returns:
        List[Tuple[int, str, str, float]]: A list of tuples containing position, original amino acid, mutated amino acid, and log-likelihood ratio.
    """
    try:
        encoded_input = tokenizer(seq, return_tensors='pt', add_special_tokens=True)
        original_output = model(**encoded_input)

        log_likelihoods = []
        amino_acids = 'ACDEFGHIKLMNPQRSTVWY'

        for pos in range(1, len(seq) + 1):  # Skip [CLS] token which is at index 0
            for aa in amino_acids:
                if seq[pos - 1] == aa:
                    continue

                mutated_sequence = seq[:pos - 1] + aa + seq[pos:]
                mutated_input = tokenizer(mutated_sequence, return_tensors='pt', add_special_tokens=True)
                mutated_output = model(**mutated_input)

                original_ll = original_output.logits[0, pos,
```

Virtual Lab School of self-improving AI scientists

1. Choose Topics for Agents

nanobodies

SARS-CoV-2 spike protein

SARS-CoV-2 variants KP.3 and JN.1

ESM

AlphaFold-Multimer

Rosetta

2. Self Generate Queries

ESM →  →

1. What is ESM?
2. How is ESM used for nanobody design?
3. How are ESM embeddings used?
4. How are ESM log-likelihoods used?
5. How to design nanobodies with ESM?

3. Search PMC & Select Papers

What is ESM? → PMC →

Title: *ESM for Nanobodies*

Abstract: ESM is a tool for…

Title: *ESM for Peptides*

Abstract: ESM can be used to…

Title: *ESM for Antibodies*

Abstract: Here we use ESM…

→  →

Title: *ESM for Nanobodies*

Abstract: ESM is a tool for…

Title: *ESM for Antibodies*

Abstract: Here we use ESM…

4. Summarize Papers

Title: ESM for Nanobodies → PMC → Full Text: ESM is a tool for proteins that can be used to model nanobodies… →  → Summary: ESM is good for nanobody design.

5. Finetune on Summaries

Summary: ESM is good for nanobody design.

Summary: ESM can't be used for small molecules.

Summary: ESM log-likelihoods are useful.

→ 

6. Evaluate via Q&A

Summary: ESM is good for nanobody design. →  →

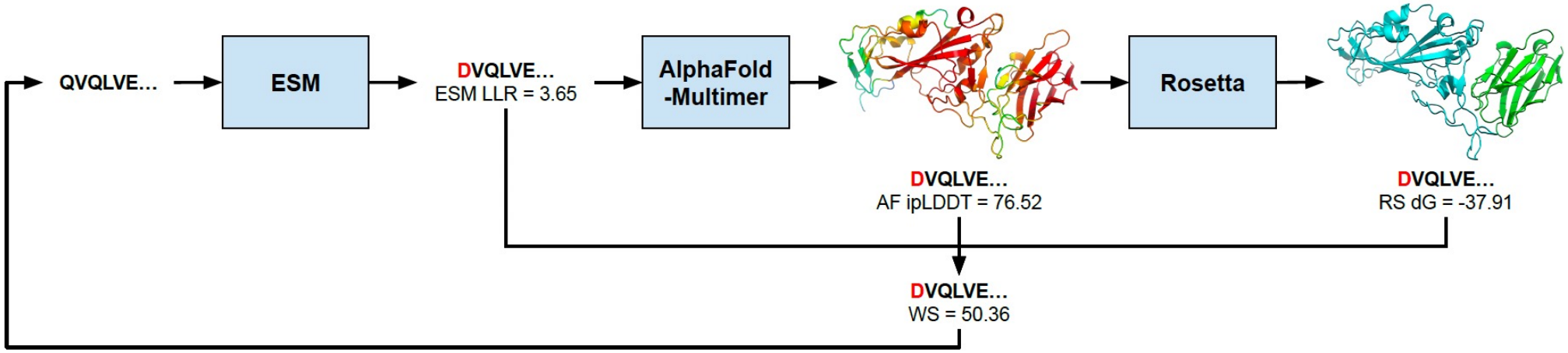
Question: What is ESM good for?

1. Small molecules
2. Nanobodies
3. DNA

→  → Nanobodies

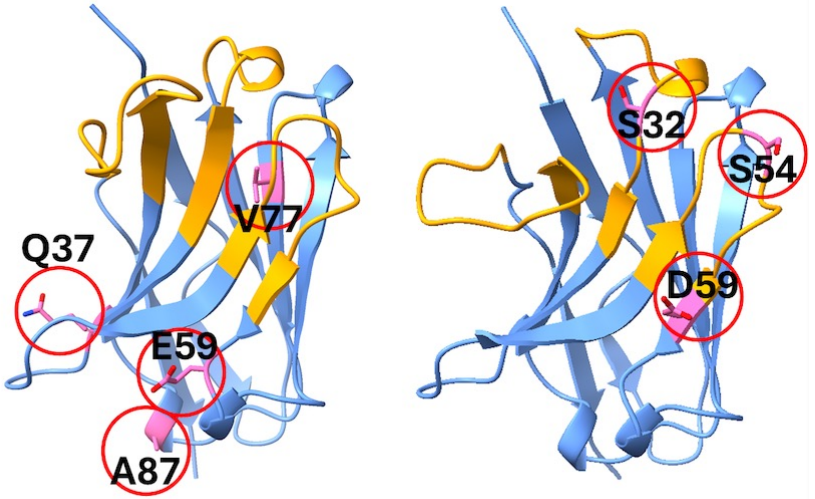
Example: Virtual Lab designs nanobodies for recent COVID variants

Virtual lab agents created novel computational workflow to design nanobodies.

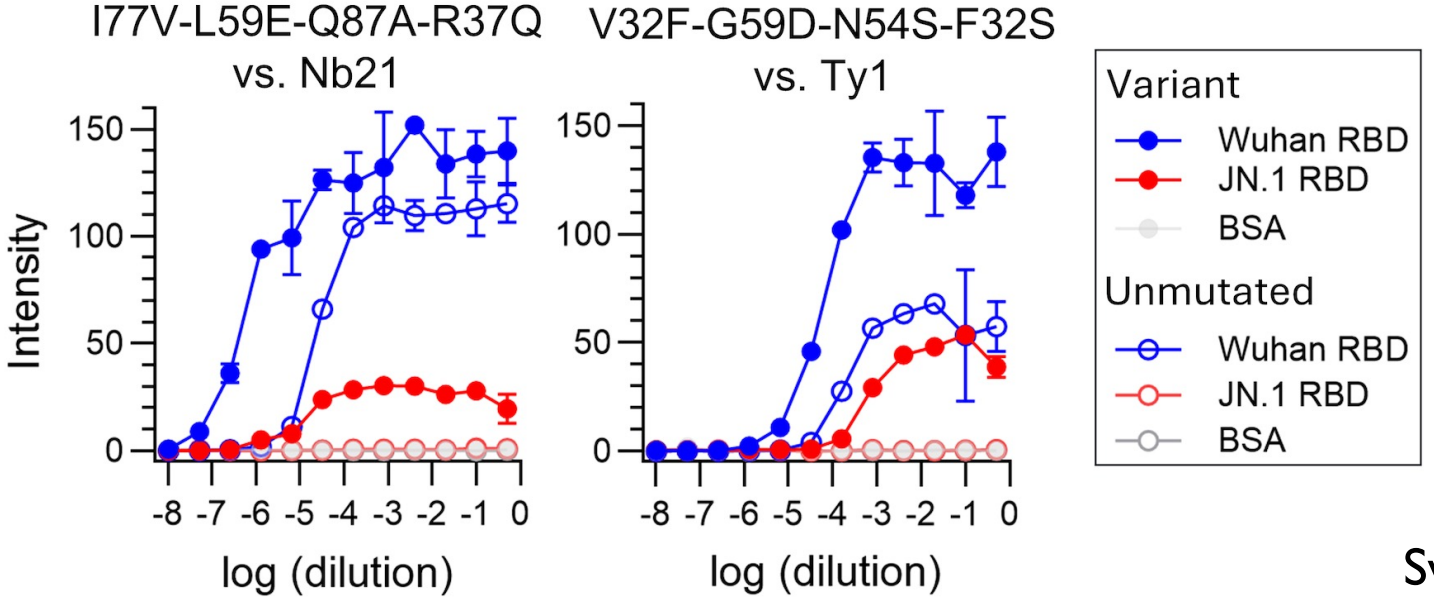


ESM → selected stable nanobodies
AlphaFold → predicts nanobody + spike structure
Rosetta → estimates binding energy
3 models combined to select candidates for next round.

Virtual Lab designed nanobodies experimentally validated



Promising candidates showing enhanced binding to recent JN.1 variant and the original Wuhan variant.



Swanson et al. *Nature* 2025

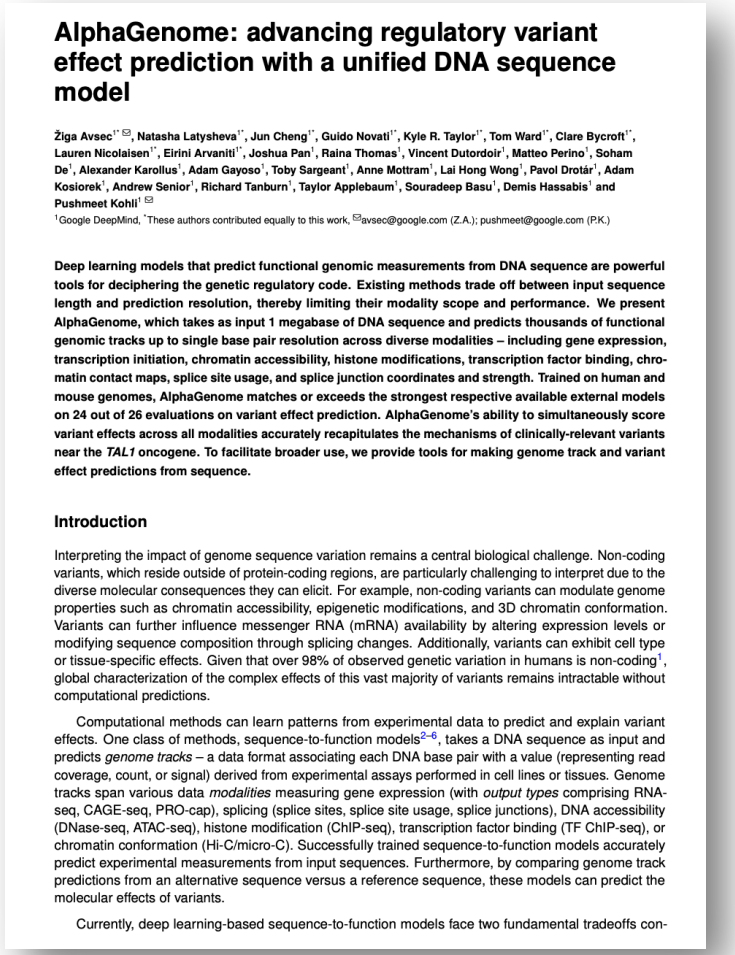
Part 1 Discussion

- Virtual Lab of AI scientists can tackle versatile research.
- Multiple agents w/ diverse expertise better than single agent.
- Helpful to provide agent with memory and sandbox to safely execute tools.
- Agents have their own social dynamics.

Swanson et al. *Nature* 2025

2. Paper2Agent

Passive paper



AlphaGenome: advancing regulatory variant effect prediction with a unified DNA sequence model

Žiga Avsec[1*], Natasha Latysheva[1*], Jun Cheng[1*], Guido Novati[1*], Kyle R. Taylor[1*], Tom Ward[1*], Clare Bycroft[1*], Lauren Nicolaisen[1*], Eirini Arvaniti[1*], Joshua Pan[1], Raina Thomas[1], Vincent Dutordoir[1], Matteo Perino[1], Soham De[1], Alexander Karollus[1], Adam Gayoso[1], Toby Sargeant[1], Anne Mottram[1], Lai Hong Wong[1], Pavol Drotár[1], Adam Kosiorek[1], Andrew Senior[1], Richard Tanburn[1], Taylor Applebaum[1], Souradeep Basu[1], Demis Hassabis[1] and Pushmeet Kohli[1]

[1]Google DeepMind, [*]These authors contributed equally to this work, avsec@google.com (Z.A.); pushmeet@google.com (P.K.)

Deep learning models that predict functional genomic measurements from DNA sequence are powerful tools for deciphering the genetic regulatory code. Existing methods trade off between input sequence length and prediction resolution, thereby limiting their modality scope and performance. We present AlphaGenome, which takes as input 1 megabase of DNA sequence and predicts thousands of functional genomic tracks up to single base pair resolution across diverse modalities – including gene expression, transcription initiation, chromatin accessibility, histone modifications, transcription factor binding, chromatin contact maps, splice site usage, and splice junction coordinates and strength. Trained on human and mouse genomes, AlphaGenome matches or exceeds the strongest respective available external models on 24 out of 26 evaluations on variant effect prediction. AlphaGenome's ability to simultaneously score variant effects across all modalities accurately recapitulates the mechanisms of clinically-relevant variants near the *TAL1* oncogene. To facilitate broader use, we provide tools for making genome track and variant effect predictions from sequence.

Introduction

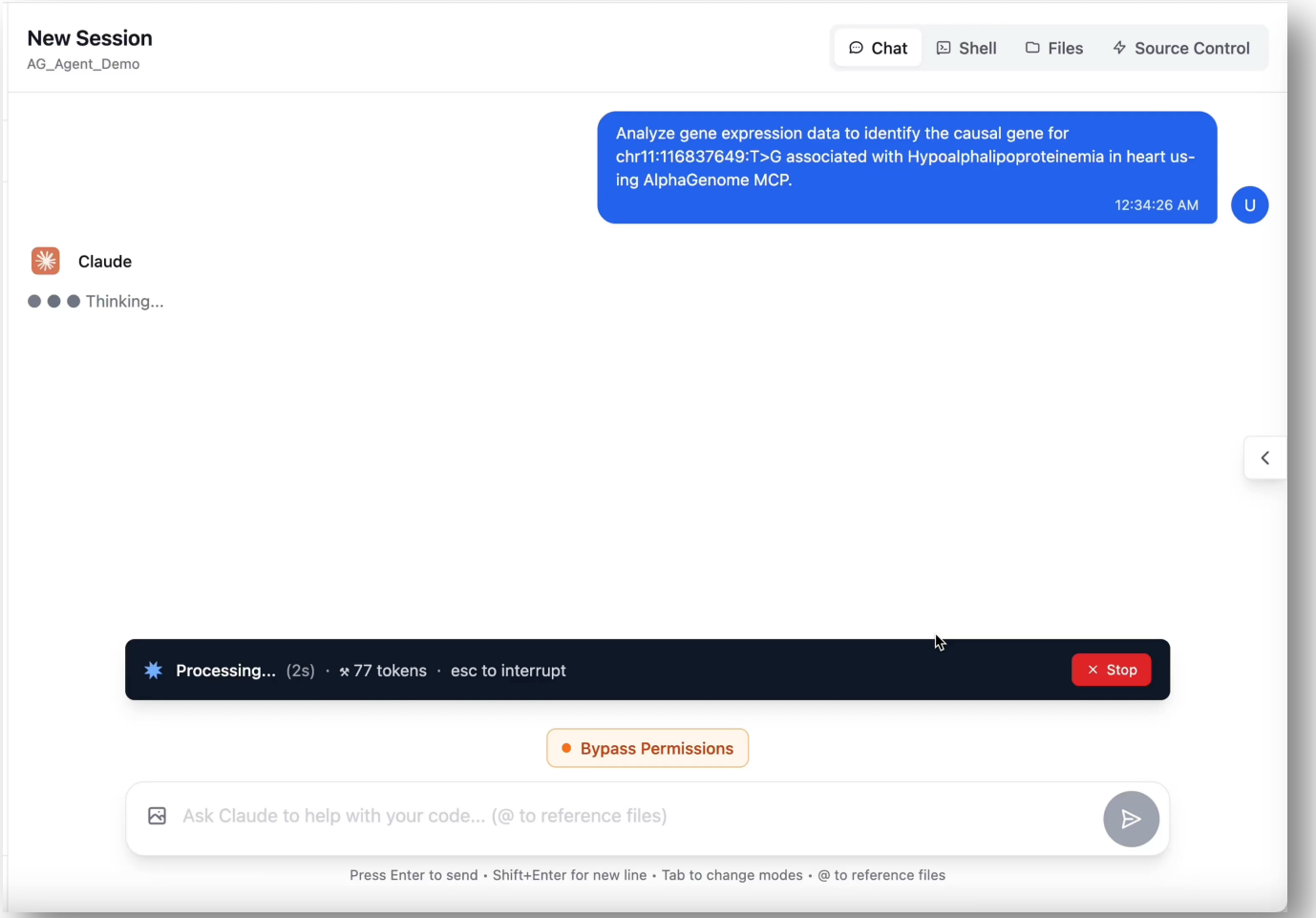
Interpreting the impact of genome sequence variation remains a central biological challenge. Non-coding variants, which reside outside of protein-coding regions, are particularly challenging to interpret due to the diverse molecular consequences they can elicit. For example, non-coding variants can modulate genome properties such as chromatin accessibility, epigenetic modifications, and 3D chromatin conformation. Variants can further influence messenger RNA (mRNA) availability by altering expression levels or modifying sequence composition through splicing changes. Additionally, variants can exhibit cell type or tissue-specific effects. Given that over 98% of observed genetic variation in humans is non-coding[1], global characterization of the complex effects of this vast majority of variants remains intractable without computational predictions.

Computational methods can learn patterns from experimental data to predict and explain variant effects. One class of methods, sequence-to-function models[2–6], takes a DNA sequence as input and predicts *genome tracks* – a data format associating each DNA base pair with a value (representing read coverage, count, or signal) derived from experimental assays performed in cell lines or tissues. Genome tracks span various data *modalities* measuring gene expression (with *output types* comprising RNA-seq, CAGE-seq, PRO-cap), splicing (splice sites, splice site usage, splice junctions), DNA accessibility (DNase-seq, ATAC-seq), histone modification (ChIP-seq), transcription factor binding (TF ChIP-seq), or chromatin conformation (Hi-C/micro-C). Successfully trained sequence-to-function models accurately predict experimental measurements from input sequences. Furthermore, by comparing genome track predictions from an alternative sequence versus a reference sequence, these models can predict the molecular effects of variants.

Currently, deep learning-based sequence-to-function models face two fundamental tradeoffs con-

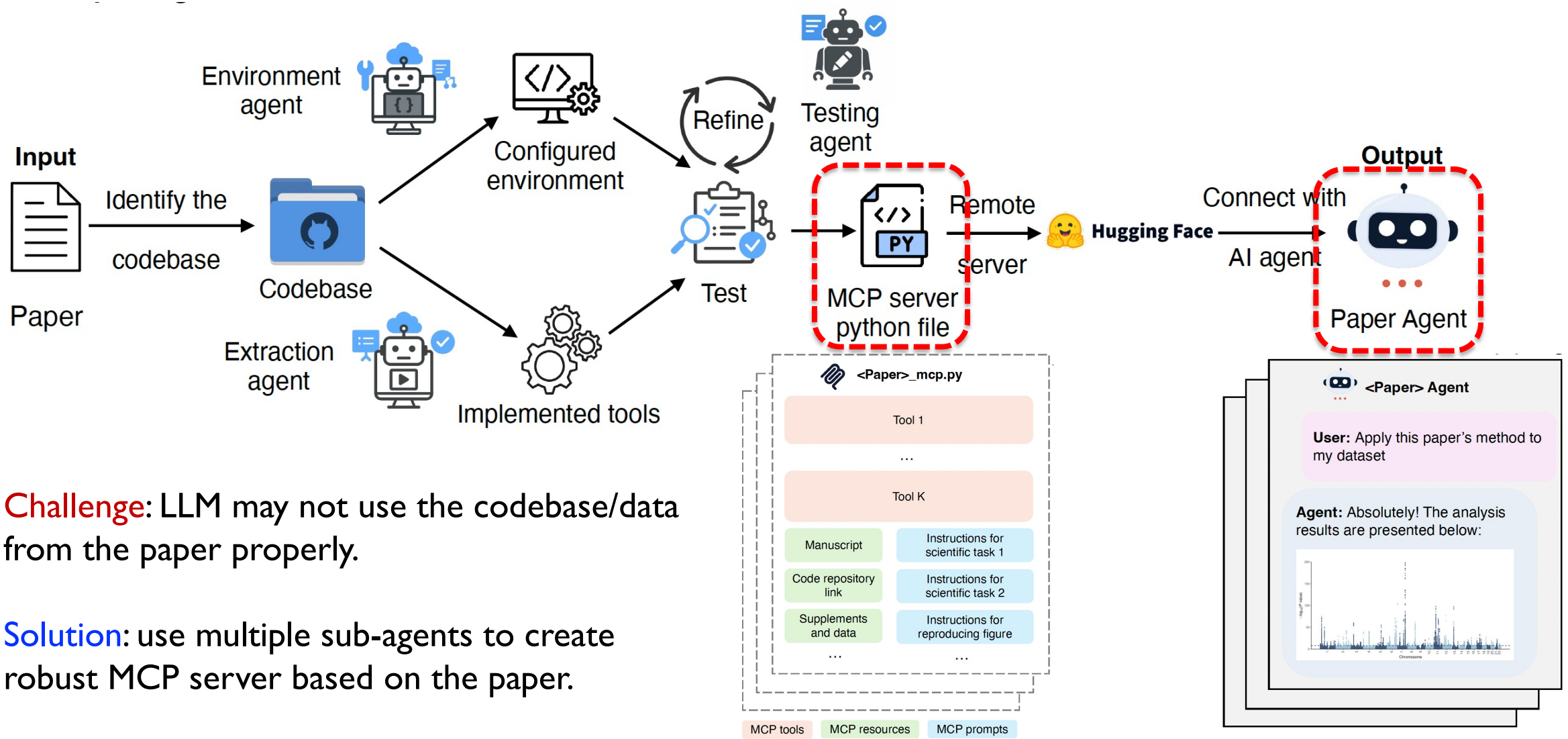
Hard to adopt code/data to new project.

Interactive agent



Miao et al. *arXiv* 2025

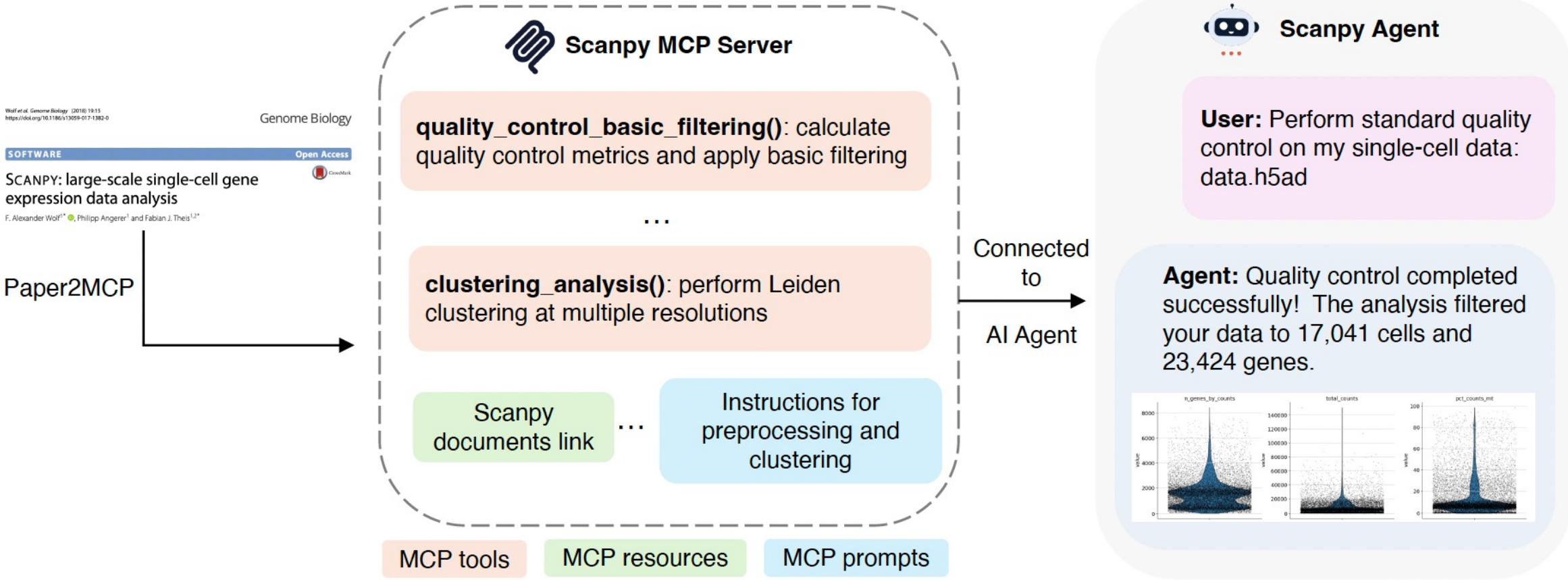
Paper2Agent automated workflow



Challenge: LLM may not use the codebase/data from the paper properly.

Solution: use multiple sub-agents to create robust MCP server based on the paper.

Case study: Scanpy agent

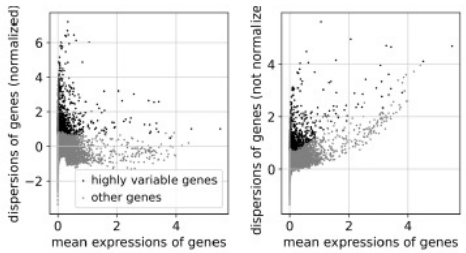


Scanpy agent reproduces human expert analyses

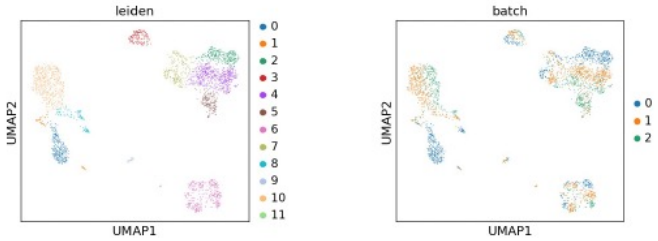
User: Perform standard single-cell preprocessing and clustering pipeline on this single cell data: data.h5ad

Scanpy agent produced results

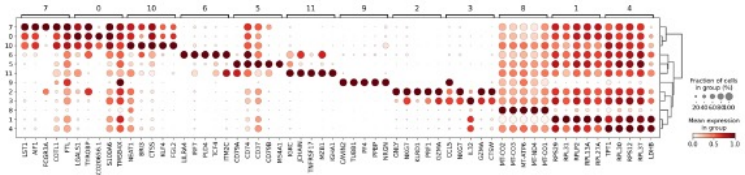
Highly variable gene



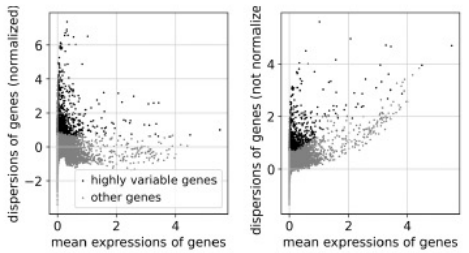
UMAP

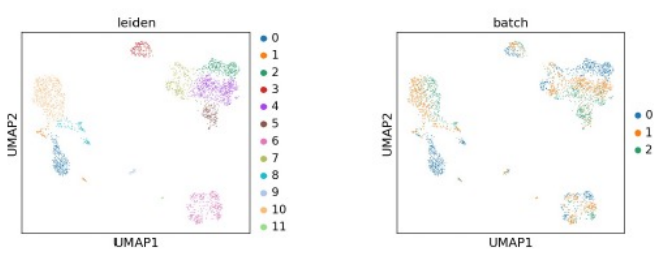


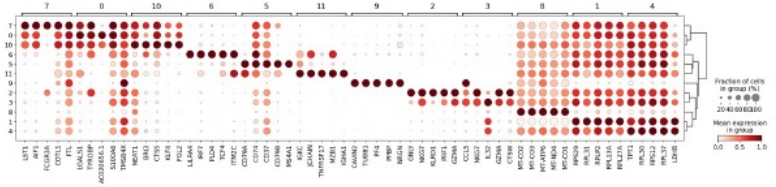
Cell type annotation using differentially expressed genes



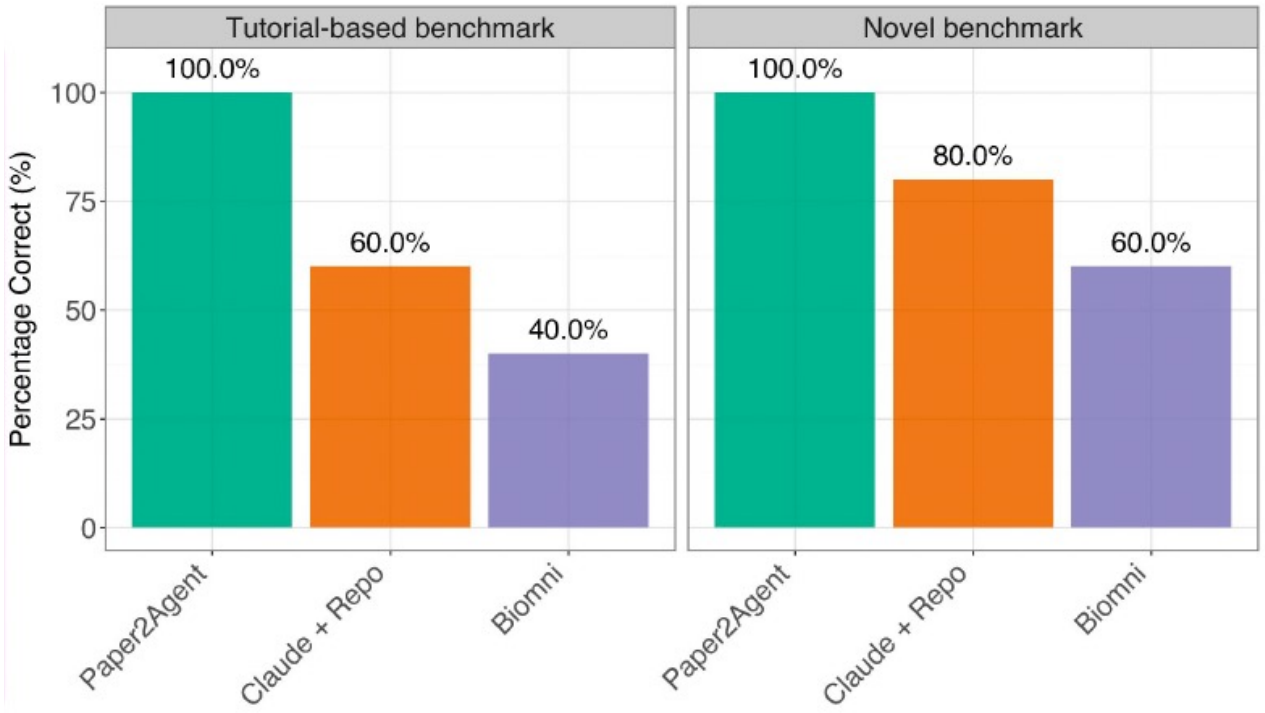
Human researcher produced results

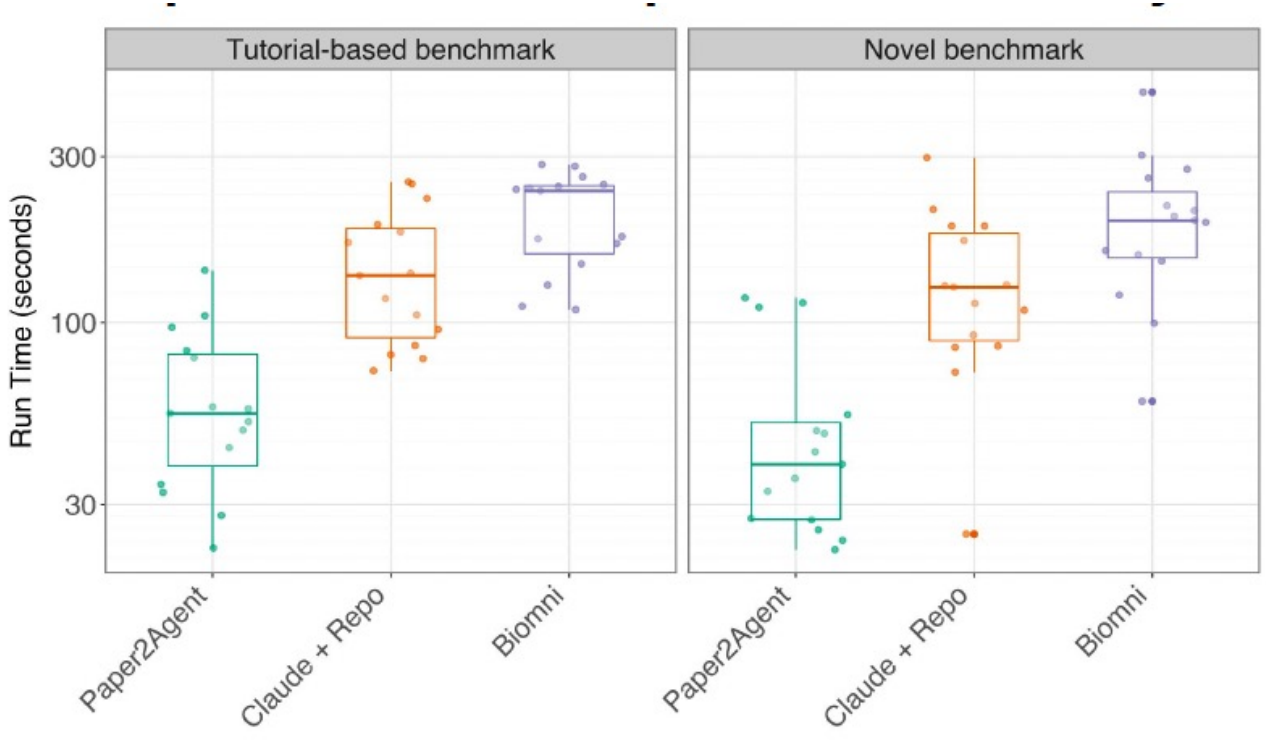




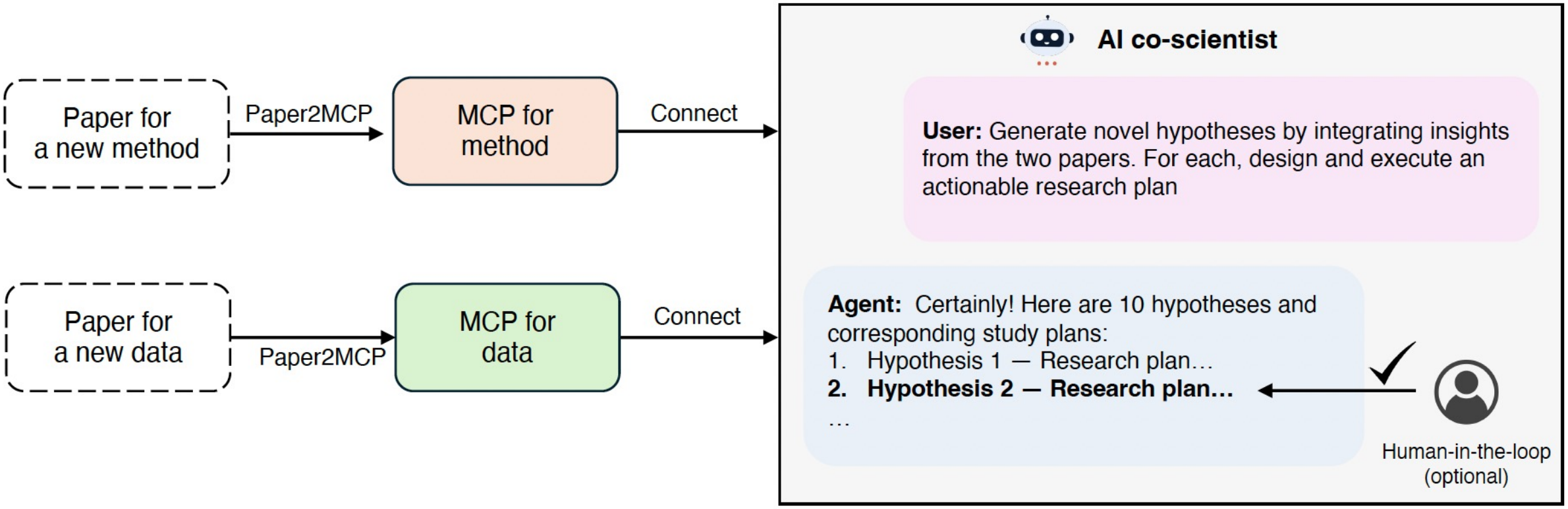


Paper2Agent correctly performs complex analyses





Agent-agent collaboration → new discoveries



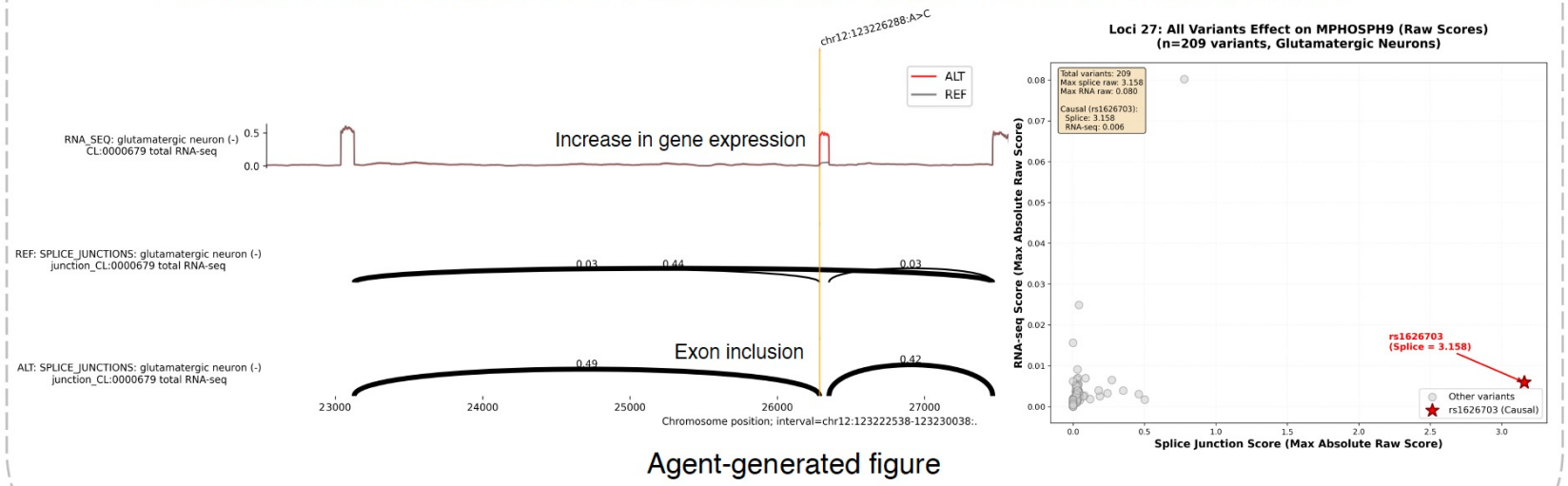
Agent-agent collaboration → new discoveries

AlphaGenome Agent collaborating with the ADHD GWAS paper agent

↓

Findings

- Among 209 candidate variants in locus 27, **rs1626703** alters ***MPHOSPH9*** **splicing and expression in glutamatergic neurons** (AlphaGenome quantile scores: splice = 1.000, RNA-seq = 0.963).
- *MPHOSPH9* is an M-phase phosphoprotein involved in cell cycle regulation and mitotic spindle organization, which may influence neuronal cell division, differentiation, and synaptic remodeling in excitatory neurons.



Agent-generated figure

Full results for causal variants and mechanisms across 39 loci are provided in Supplementary Table.

New splicing error associated with ADHD risk

Part 2 Discussion

- Paper2Agent converts passive papers into interactive agents.
- Easier for readers to use the methods + data from the paper.
- New way to disseminate knowledge.
- Agents can collaborate! New method agent + new data agent.

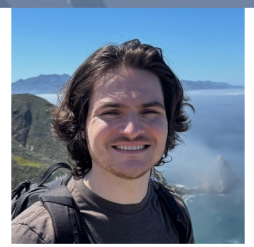
Miao et al. *arXiv* 2025

Open Conference of AI Agents for Science 2025

The 1st open conference where AI serves as both primary authors and reviewers of research papers

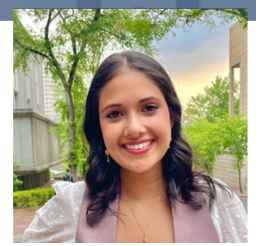
Exploring the future of AI-driven scientific discovery through transparent AI-authored research and AI-driven peer review.

Federico Bianchi

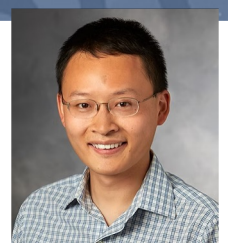


Owen Queen

Eric Sun



Nitya Thakkar



James Zou

How creative are AI scientist agents?

How should human researchers collaborate with AI agents?

How good are LLMs at reviewing papers?

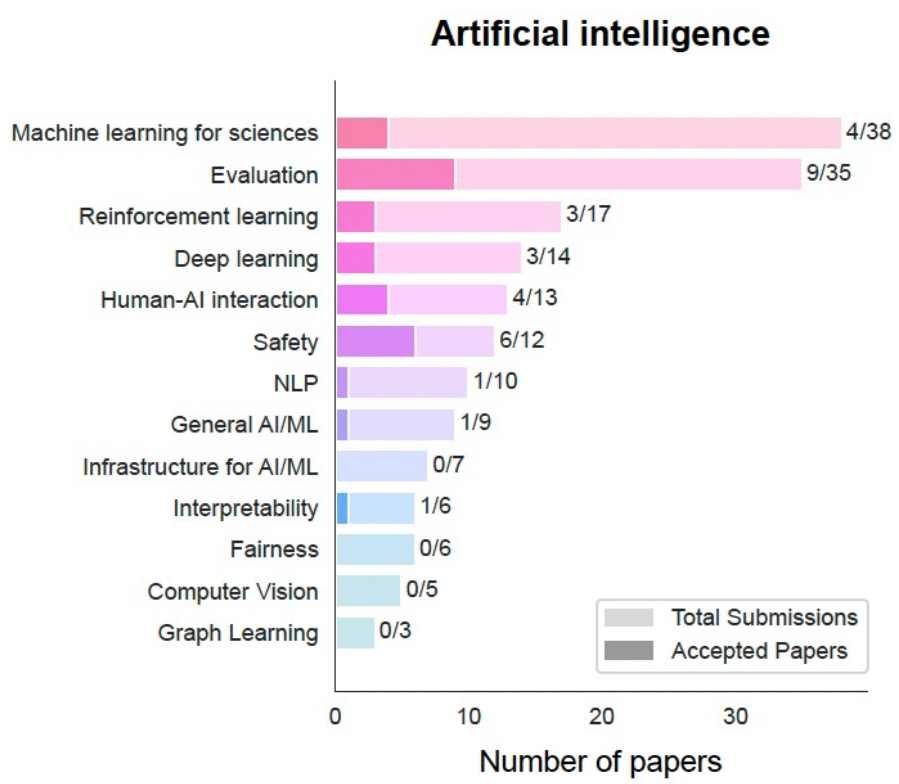
...

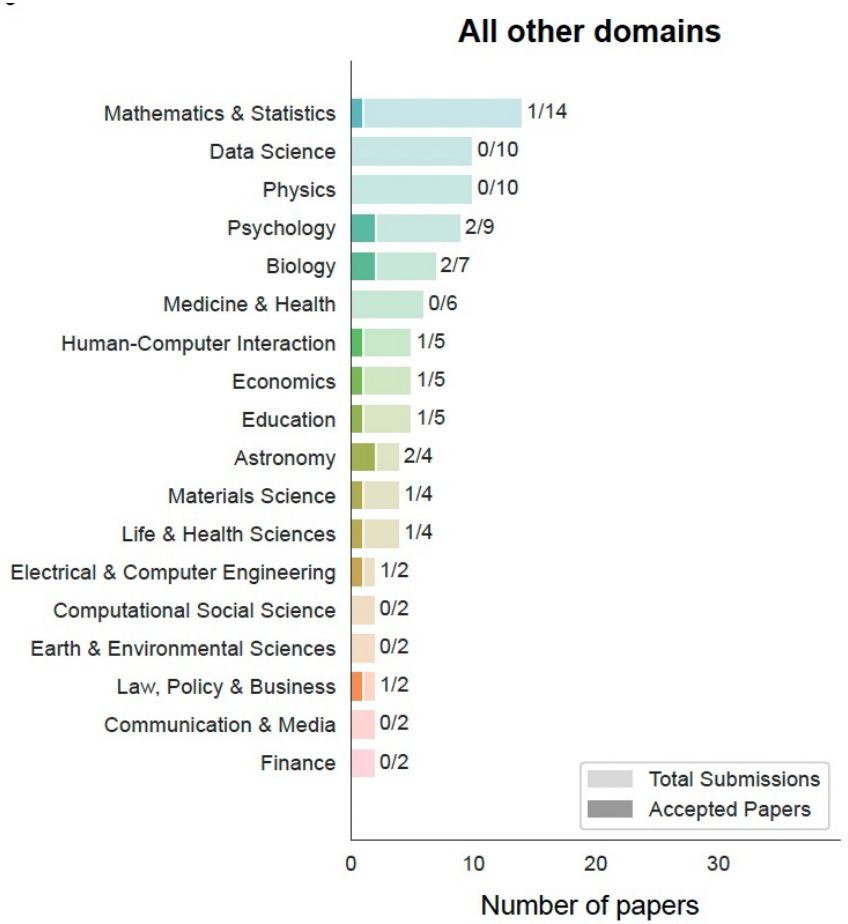
To answer these questions

- AI agents as both authors and reviewers.
- Submissions and reviews are public.
- Document AI-human collaboration.
- Additional human expert assessments.

Agents4Science statistics

315 submissions; 48 accepted papers;
Covers broad domains AI, Physics, Medicine, Econ





What LLMs and agents did people use?

Claude

GPT-4/5

Gemini

Liner

DeepResearch (OpenAI)

Cosmetics Discovery Agent

CMB Agent

Denario

Specialized and Proprietary Agents

ClaudeCode

Cursor

General Coding Agents

Grok-4

Qwen

Kimi-K2

GLM-4.5

Deepseek-V3

Other Models

Submissions: AI Checklist

Authors need to **disclose AI involvement** (hypothesis generation, writing, experiments).

Authors share the **limitations** they found when using AI in their work.

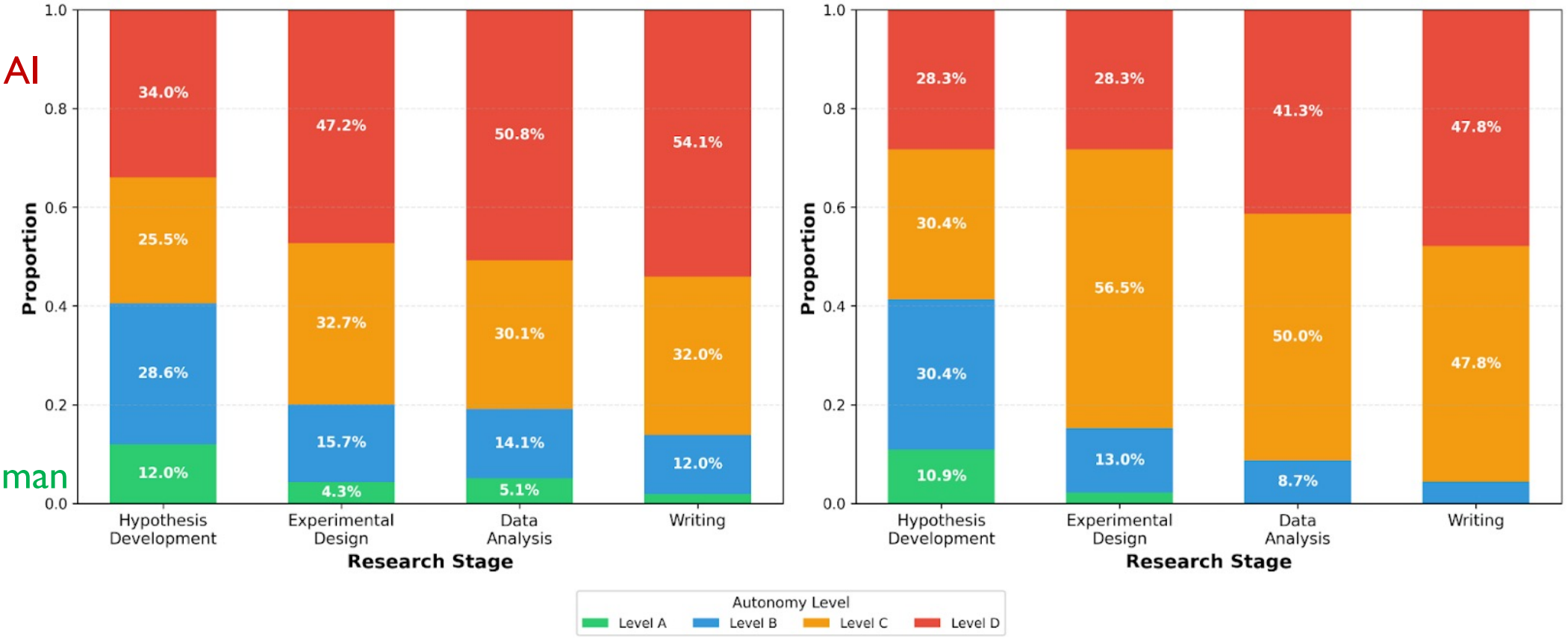
Ratings

A: 95% human	**C**: >50% AI
B: >50% human	**D**: >95% AI

Human-AI collaboration patterns

mostly AI

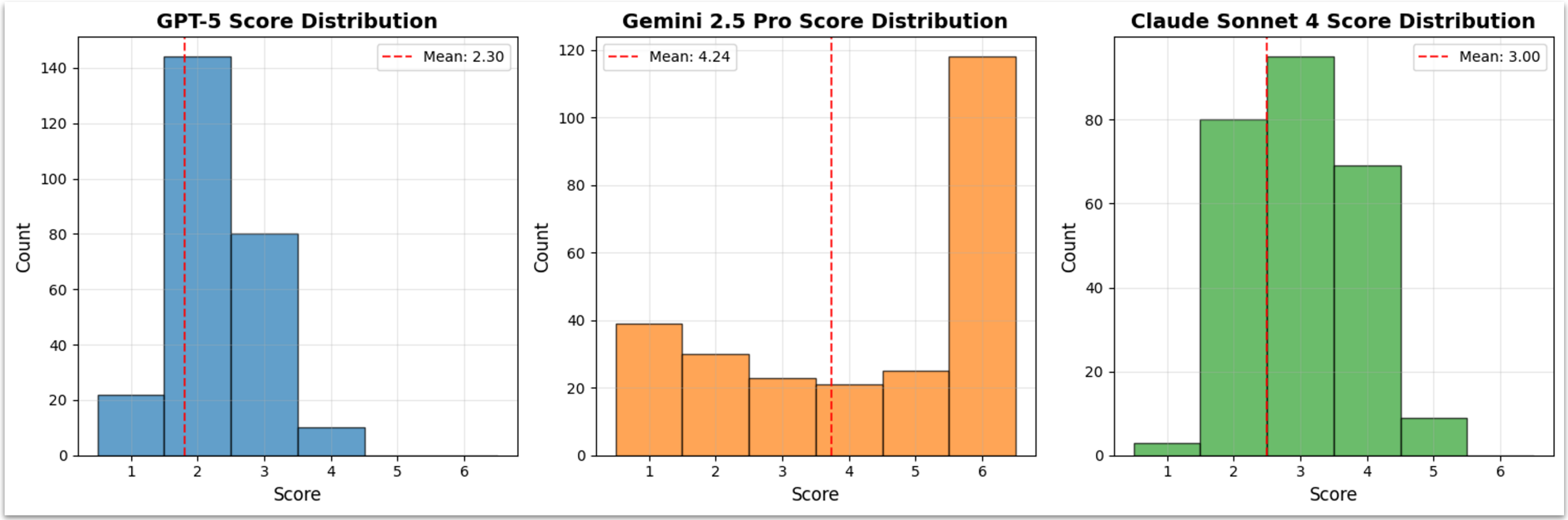
mostly human



Bianchi et al. *Nature Biotech 2025*

Reviewers: GPT5, Gemini2.5 Pro, Claude Sonnet 4

- Reviewers tuned using ICLR2025 and ICLR2024 accepted/rejected papers.
- Scores from 1-6 following NeurIPS guidelines.
  - (1 strong reject, 3 borderline reject, 4 borderline accept, …).



Bianchi et al. *Nature Biotech 2025*

Examples of LLM Reviews

AI reviewer finds mismatches in the papers:

- **AI Feedback:** *"The text states $R^2 = 0.0148$ whereas Table 1 … reports $R^2 = 0.005$"*

AI reviewer sycophancy

- **AI Feedback:** *"Overall, this is a groundbreaking and highly recommended paper for a forward-thinking conference."*

Automated Reference Verification

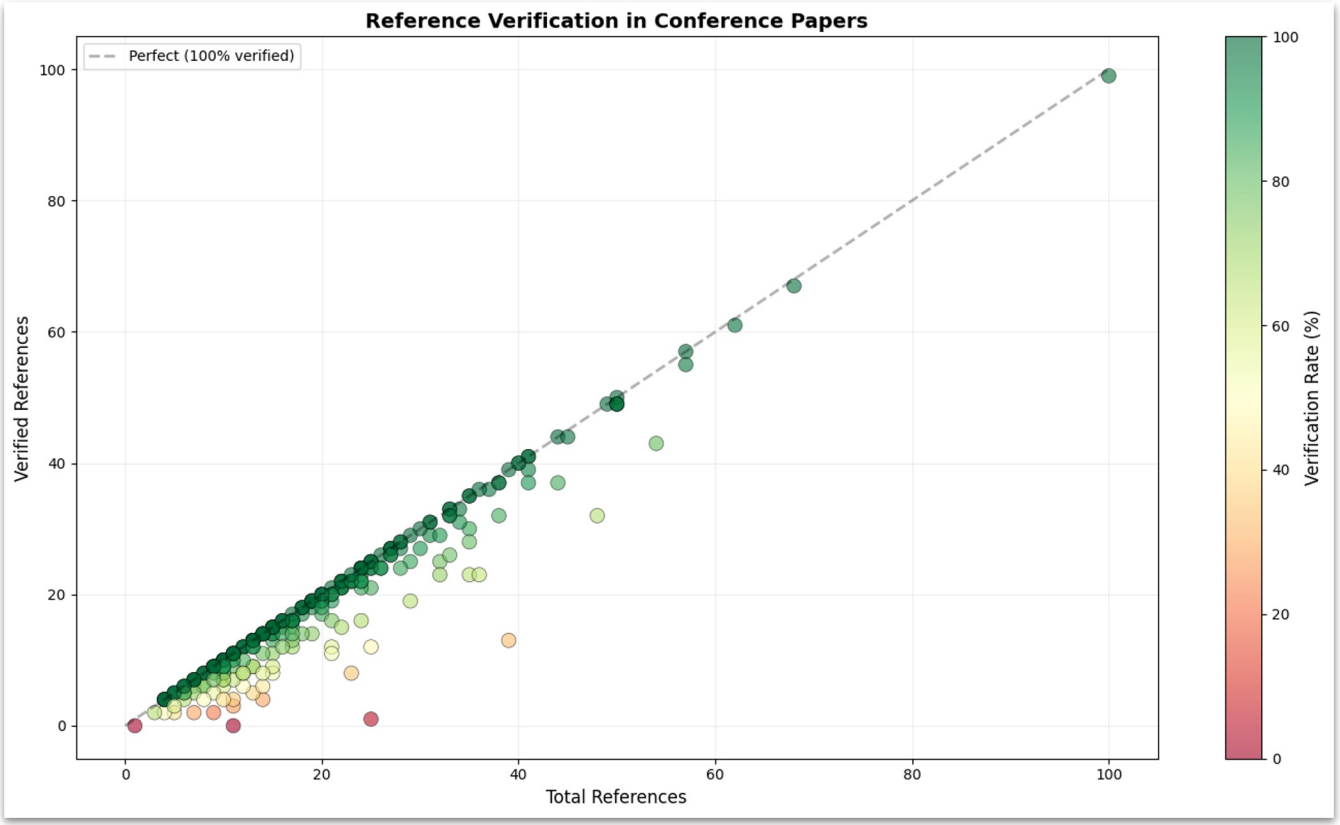
- **AI Agent searches online** if a reference exists.
- If references **cannot be verified**, **users will see a warning** on OpenReview.

Comment: **Related Work Check**

Please look at your references to confirm they are good.

Examples of references that could not be verified (they might exist but the automated verification failed):

- Towards autonomous scientific research agents by Chen, T., et al.
- Working with machines: Impact of algorithmic management by Lee, M. K., et al.
- Science as a multi-agent system by Zou, J. Y., et al.



56% of submission has ≥1 hallucinated reference

Simulating Two-Sided Job Marketplaces with AI Agents

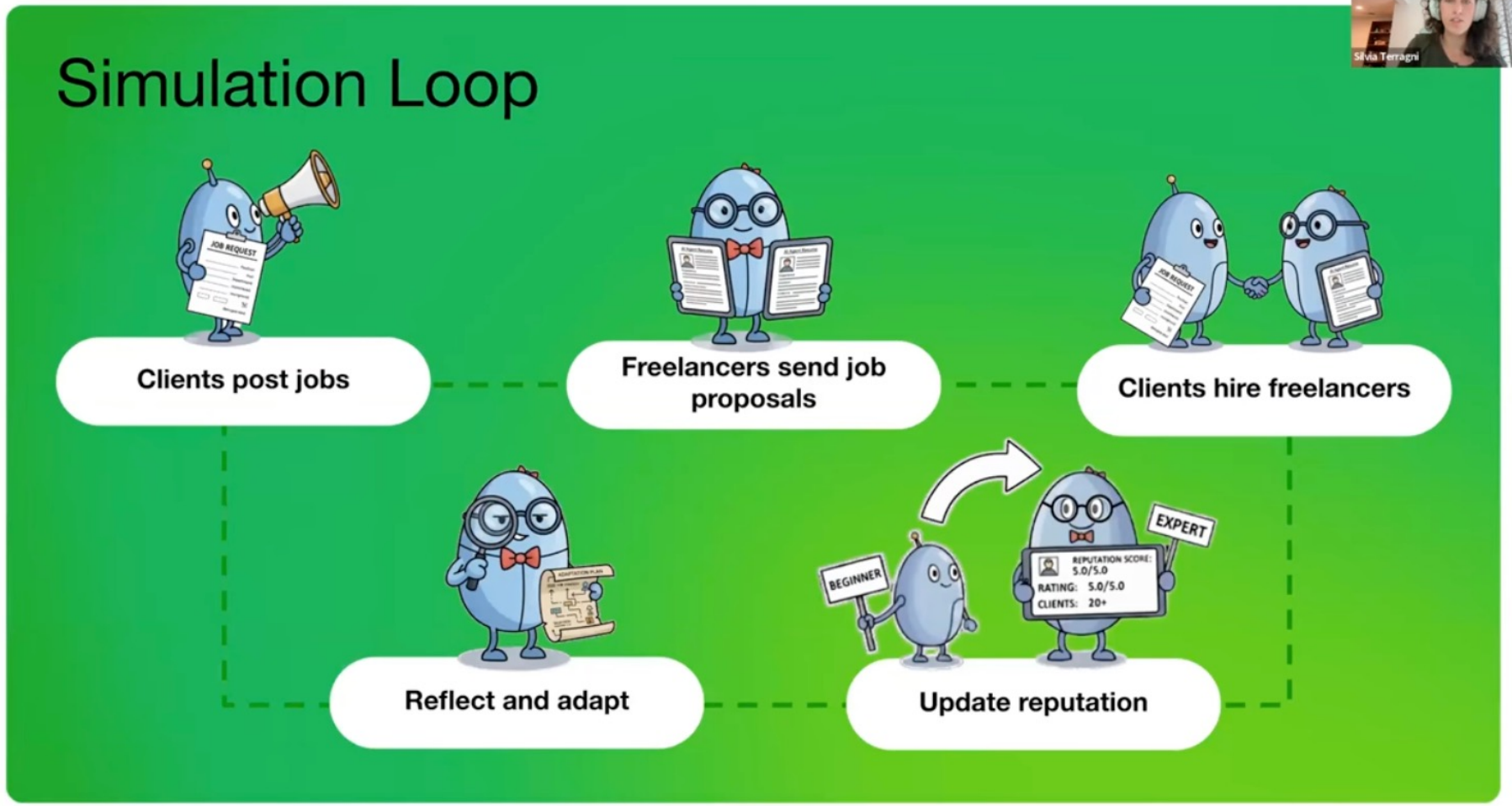
AI Agent

Silvia Terragni
Upwork Inc
San Francisco, CA
silviaterragni@upwork.com

Behnaz Nojavanasghari
Upwork Inc.
San Francisco, CA

Frank Yang
Upwork Inc.
San Francisco, CA

Andrew Rabinovich
Upwork Inc.
San Francisco, CA



Simulating Two-Sided Job Marketplaces with AI Agents

AI Agent

Silvia Terragni
Upwork Inc
San Francisco, CA
silviaterragni@upwork.com

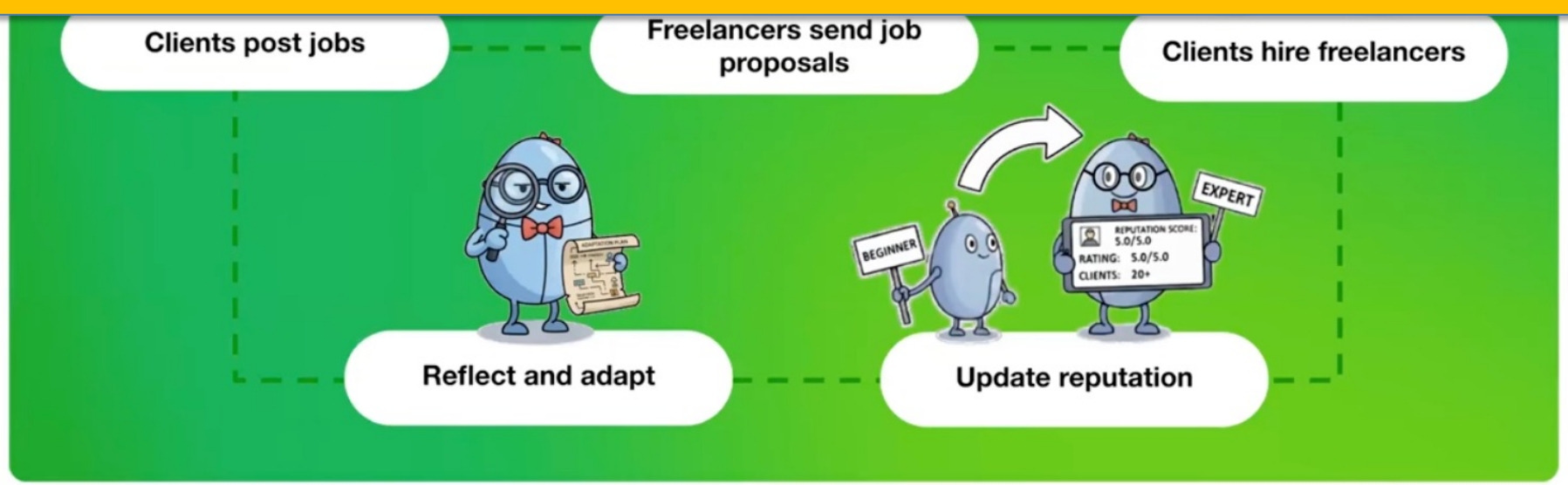
Behnaz Nojavanasghari
Upwork Inc.
San Francisco, CA

Frank Yang
Upwork Inc.
San Francisco, CA

Andrew Rabinovich
Upwork Inc.
San Francisco, CA

"This is a really interesting paper in a fast developing area. Both the agent based modelling the the two-sided market places are interesting topics and the combination is fascinating."

Expert human reviewer (Nobel Laureate in economics)



Multi-target Parallel Drug Discovery with Multi-agent Orchestration

Using an AI framework to automate and accelerate early-stage drug discovery

The Challenge: Traditional Drug Discovery is Broken

Over 10 Years
from lab to patient

$2.5+ Billion
to develop one drug

>90% Failure Rate
of drugs in clinical trials

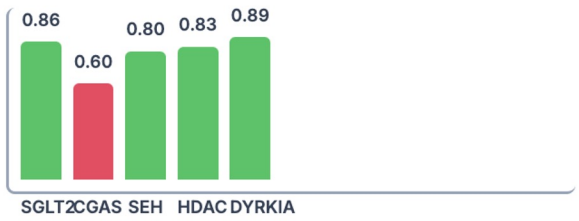
The Solution: A Multi-Agent AI Framework

An automated system where specialized AI "Agents" collaborate to perform the entire early discovery pipeline, from identifying targets to creating new drug candidates.

Case Study: Alzheimer's Disease

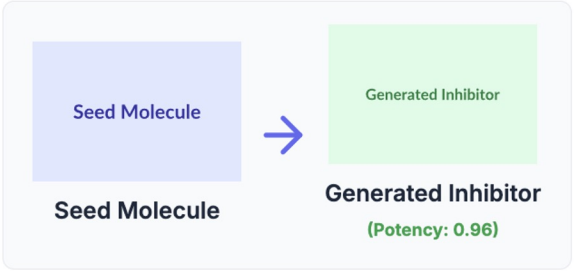
Model Performance (AUPRC)

(AUPRC = Area Under Precision-Recall Curve. Higher is better.)



Example: "Scaffold Hopping"

The AI generated entirely new molecular structures (right) from a known seed molecule (left).



Seed Molecule

Generated Inhibitor
(Potency: 0.96)

The Automated Workflow

1. Abstract Mining Agent
Scans PubMed literature to identify potential protein targets for a disease.

2. Target Evaluator Agent
Scores and ranks targets based on novelty, evidence, and confidence.

3. Medicinal Chemist Agent
Selects known active "seed" molecules for the validated targets from ChEMBL database.

4. Molecule Generator Engine
Uses Generative AI (NVIDIA MolMIM) to design thousands of new, novel molecules.

5. Molecule Evaluator Agent
Uses ML models to filter molecules for high potency and good drug-like properties (Safety, Solubility, etc.).

Optimized Hit Candidates
Novel, potent, and safe molecules ready for further testing.

Key Implications & The Big Picture

SUCCESS: A Powerful Discovery Engine
The AI framework successfully generated novel, potent, and drug-like inhibitors for 4 out of 5 targets (SGLT2, SEH, HDAC, DYRKIA).

CRITICAL LIMITATION: Data is King
The framework failed on the **CGAS** target and on predicting **microsomal half-life**. Why? **Not enough high-quality data.** The AI's performance is "fundamentally tethered" to the quality and availability of its training data.

THE FUTURE: Human-in-the-Loop
Autonomous AI is powerful, but it works best when combined with human expertise. The future is a partnership where experts guide data curation and validate the AI's results, creating a "Human-in-the-Loop" paradigm.

Jake Chen