AI for the Science of Science: From Robustness to Hypothesis Generation

Alexander Belikov
GrowGraph, 2025

Plan

1. Challenges in Science
2. Reproducibility crisis. Discovery of Facts. Consensus convergence.
3. Evaluation in Science.
4. Agentic AI. Hypothesis Generation. AI Scientist.

The Subject of Science

Science is a complex, non-linear, and deeply human process that we do not fully understand.

- Knowledge: the body of **Facts** vs the body of **Claims** (literature, subjective statements).
- Semantic Component: How is knowledge structured?
- Sociological Component: How does human interaction shape discovery? This includes the study of collaboration networks, funding mechanisms, prestige, and the social dynamics within research groups and institutions.

The Exponential Growth of Information

- Scale and Scope: This rapid growth challenges our ability to process, validate, and integrate new findings, making the identification of genuinely novel and high-impact work increasingly difficult.
- The "Matthew Effect": Established scientists and institutions often disproportionately benefit from new attention and resources.

Dynamics of Knowledge Accumulation

- Cumulative Nature: Much of science builds on previous work - a phenomenon known as cumulative knowledge accumulation.
- Disruptive Innovation: Disruptive or revolutionary breakthroughs, paradigm shifts (as described by Thomas Kuhn): exponential growth, relates to distant fields (percolates).

The Challenges of Modern Science

📚 Information Overload & Quality Control

- The Scale Problem: Too many research papers; too little time for proper review and synthesis.
- Truth Rendering: How do we effectively shift from publications (which can be subjective/incremental) to established textbook facts?
- The Measurement Problem: Are our current metrics (e.g., citation counts) truly reflecting genuine scientific progress and value?

🗺️ Direction & Focus

- Fashion Dominance: The field often focuses disproportionately on "hot" areas (e.g., String Theory, Large Language Models), potentially neglecting wide front development in less glamorous but crucial areas.
- Explore vs. Exploit: Striking the right balance between pursuing high-risk, high-reward exploration and capitalizing on current knowledge (exploitation).

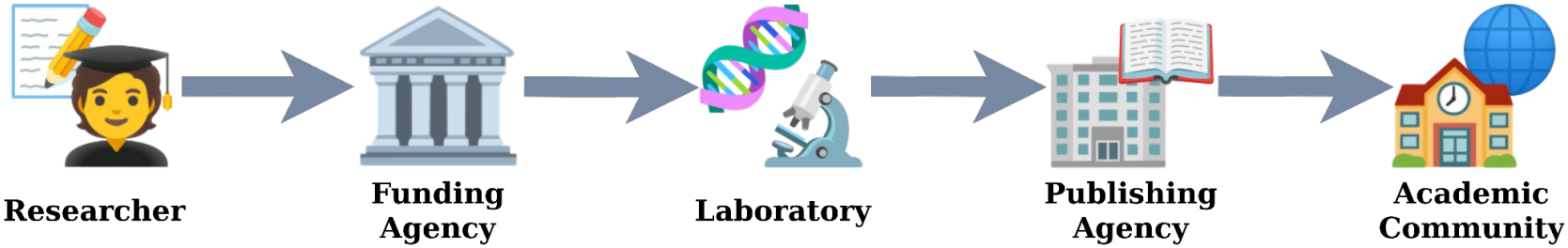
💼 Structural & Systemic Issues

- Career Incentives: Do the current pressures for publishing and grants translate into the greatest long-term scientific impact?
- Resource Allocation: Ensuring fair and strategic distribution of funding and infrastructure to maximize scientific benefit.

Knowledge Generation Loop

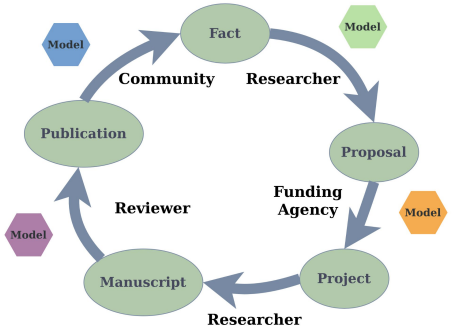
- Researchers struggle to identify non mainstream yet promising lines of research
- Funding agencies face difficulties deciding which research projects to fund
- Publishing agencies toil at manuscript triage and validation
- The challenge of academic community is to decide which pieces of published research are correct

The traditional scientific knowledge loop is hindered by subjective evaluation and systemic bottlenecks at every stage—from hypothesis proposal to community validation.



Where AI might help (is helping)

1. Derive new metrics of quality
   - how research affects the society; how the structure of knowledge changes
2. Improve interactions with **Claims,** facilitate **Claims** →**Facts** rendering
   - navigate Science in an informed manner
3. Improve academic processes
   - evaluate proposals, preprints, possible project
   - plan academic trajectories
4. Generative Science: AI Scientist
   - generate and verify hypotheses, rather than evaluate them



Publications to Facts

📈 The Problem with Consensus

- Insufficient Convergence: Empirical convergence to a general consensus is not sufficient for establishing robust truth.
- Non-Monotonic Progress: Consensus convergence is often not monotonic, meaning scientific 'understanding' can oscillate or revert, leading to confusion.
- The Reproducibility Crisis: A high-profile symptom of insufficient convergence, where reported results cannot be reliably replicated.

🔎 Systemic Flaws and Bias

- Bias Intrusion: Subjective biases (both conscious and unconscious) frequently steer research, funding, and interpretation.
- Lack of Transparency: Insufficient sharing of raw data, analysis code, and methods hinders verification and scrutiny.

🌊 Data Overload & Discovery

- Information Overwhelm: Given the sheer volume of modern data and literature, the discovery of pertinent, high-quality information is severely hindered.

* - “Nonreplicable publications are cited more than replicable ones.” Serra-Garcia et al., Science Advances (2021).

Examples of “deviant” empirical convergence

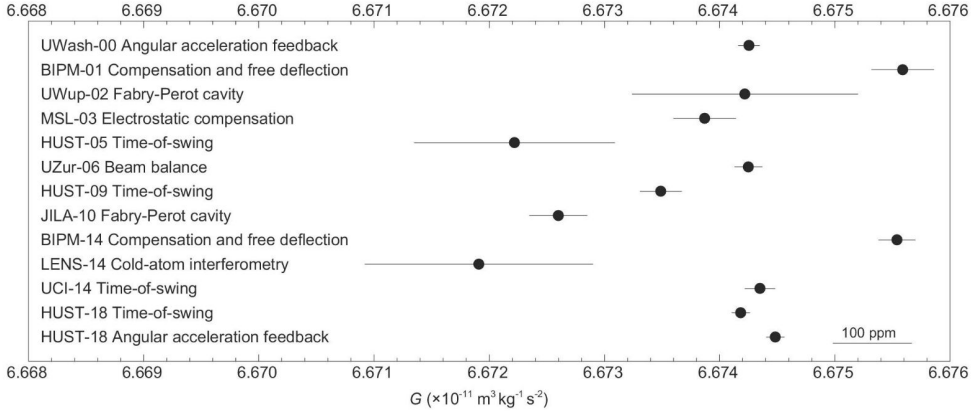
Social Sciences. Psychology.

An effect called ego depletion: willpower can be worn down over time was found by Baumestier et al., 1998 (more than 7K citations).

Hagger et al. (2016) tried to replicate these results in 24 labs. And failed.

Hard Sciences. Physics.

Gravitational constant remains the physical constant with the largest systematic error.



Xue et al, National Science Review, Vol. 7, Issue 12, 2020

Related Works

John P A Ioannidis, (2005): “Why Most Published Research Findings Are False”, PLoS Med. 2(8):e124

- The smaller the studies conducted in a scientific field, the less likely the research findings are to be true.
- The smaller the effect sizes in a scientific field, the less likely the research findings are to be true.
- The greater the number and the lesser the selection of tested relationships in a scientific field, the less likely the research findings are to be true
- The greater the flexibility in designs, definitions, outcomes, and analytical modes in a scientific field, the less likely the research findings are to be true
- The greater the financial and other interests and prejudices in a scientific field, the less likely the research findings are to be true
- The hotter a scientific field (with more scientific teams involved), the less likely the research findings are to be true.

Devezer, B., et al. (2019): Scientific discovery in a model-centric framework: Reproducibility, innovation, and epistemic diversity. PLoS One, 14(5), e0216125.

- Simulation shedding light on the mechanism of convergence.

Devezer et al.: Framework

Data / Framework: a mathematical / computational model of the scientific process.

Core of their framework: scientists try to discover a “true model” that generates data, in a stochastic process.

Use agent-based modeling + Markov chains + Monte Carlo simulation to simulate research communities.

Define different "types" of scientists / research strategies:

- Mavericks (“Mave”): propose novel models, totally independent of consensus.
- Boundary testers (“Bo”): propose models by adding interactions to current consensus.
- Refiners (“Tess”): propose small modifications (refinements) to consensus.
- Others “epistemically diverse” which mixes strategies.

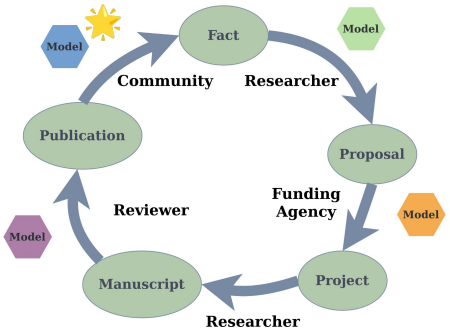
Model comparison criteria: they use standard statistical measures — Akaike Information Criterion (AIC) and Schwarz Criterion (SC / BIC) to evaluate which model “wins” in their simulated experiments.

Devezer et al.: Main Results

- Reproducibility ≠ Truth: High reproducibility (i.e., frequent confirmations) does not guarantee convergence to the true model.
- Irreproducible results are not necessarily false: Some scientifically “irreproducible” findings might still reflect aspects of truth.
- Epistemic diversity is beneficial: Communities that mix different research strategies (“epistemically diverse”) perform better overall across several desirable properties: they discover truth faster, spend more time on truth, and are more persistent.
- Innovation matters: Mavericks / novel / exploratory strategies accelerate discovery because they explore model space more broadly.

Trade-offs exist: Optimizing purely for reproducibility can harm other goals (like discovery speed or exploring truth), showing that scientific practice may need to balance multiple objectives.

Prediction of robust scientific facts from literature



Literature

Experiment

Rzhetsky, A. et al, 2004

Geneways

Poon,H. et al, 2014

Literome

Library of Integrated Network-Based Cellular Signatures
Subramanian, A. et al, et al, 2017

Lincs L1000

Belikov et al., *Nature Machine Intelligence* **4**, 445 (2022)

The goal of study: To use large-scale experimental data (LINCS L1000) to rigorously validate and predict the correctness of gene-interaction relationships extracted from the noisy biomedical literature."

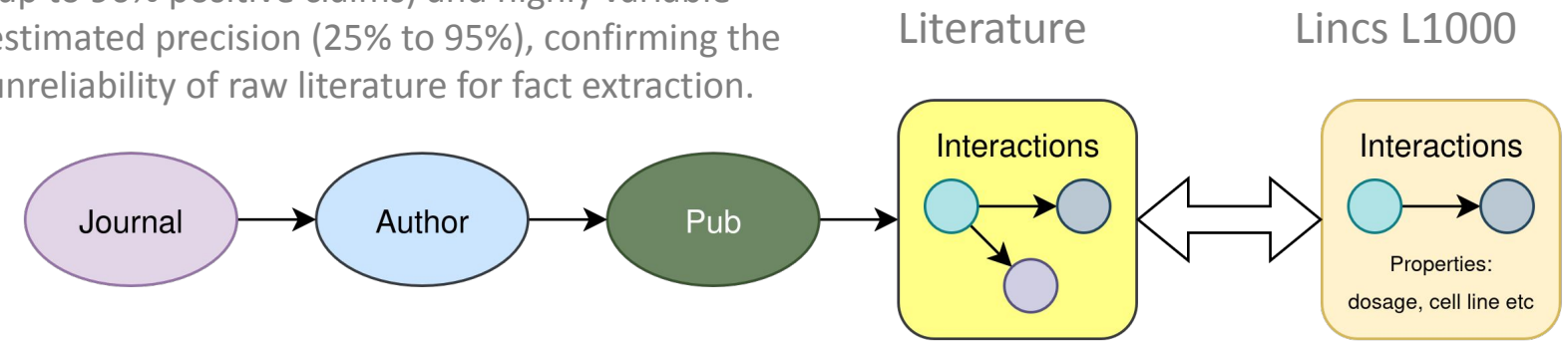
Data Schema

statement s: "Activation of [protein kinase C alpha] enhances human <growth hormone> binding protein release ." (pmid: 10022777)

(a, b, α) $\quad \pi_{(a,b)} : s \rightarrow \{T, F\}$

	GeneWays	Literome
#claims	612K	409K
# publications	197K	220K
# genes	5,141	10,703
# interactions	23,405	144,172
#positive claims	77%	96%
est. precision	95%	25%

The literature data sources show a strong inherent bias (up to 96% positive claims) and highly variable estimated precision (25% to 95%), confirming the unreliability of raw literature for fact extraction.



Lincs L1000: directed graph of interactions

measures genome-wide mRNA

1.3M gene profiles, for a total of 474K gene signatures

71 cell lines, from 19 primary sites

pert_iname	pert_type	cell_id	pert_idose	pert_itime	is_touchstone	up	dn	score	cdf
ADRB2	trt_oe	A375	2 L	96 h	1	154	153	-0.199	0.421
ADRB2	trt_oe	HA1E	1 L	96 h	1	154	153	1.139	0.873
ADRB2	trt_oe	HEPG2	2 L	96 h	1	154	153	-0.853	0.197
ADRB2	trt_oe	HT29	2 L	96 h	1	154	153	0.496	0.690
ADRB2	trt_oe	MCF7	2 L	96 h	1	154	153	0.157	0.562

Data Alignment

For each interaction in literature we have multiple claims, in the experimental dataset - multiple experiments.

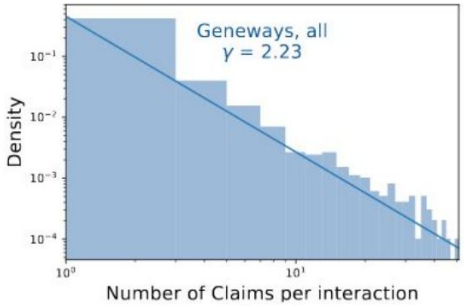
1. Aggregate claims per publication
2. Take only claims from abstracts
3. Keep claims for which features can be derived.
4. Keep interactions mappable to LINCS L1000

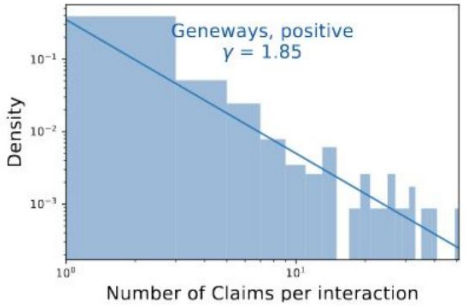
Overlap between Geneways/Literome: 2K interactions with a correlation ~ 0.38.

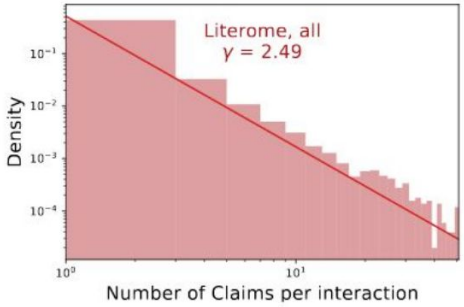
genes/interactions	Geneways	Literome
# feature merged	44K/23K	
# LINCS merged	16K/7K	51K/25K

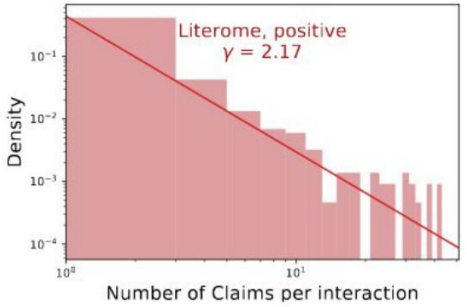
Claim number distribution

The distribution of claims is highly skewed (power-law or scale free), indicating a 'rich-get-richer' effect where a small number of interactions receive the vast majority of research attention and claims.





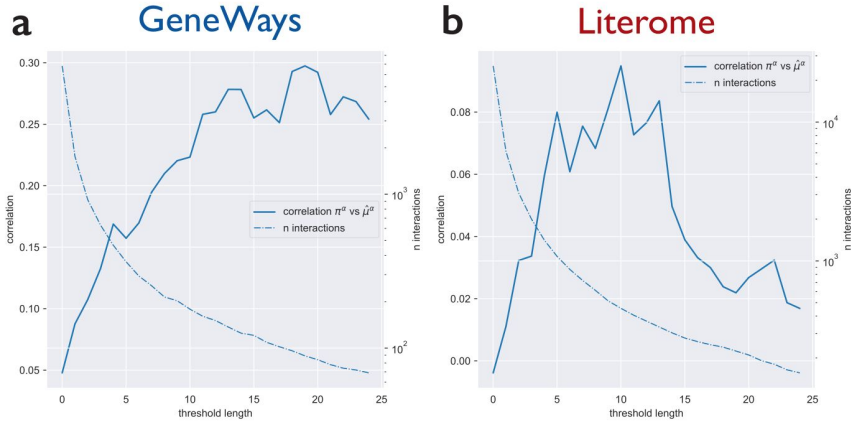


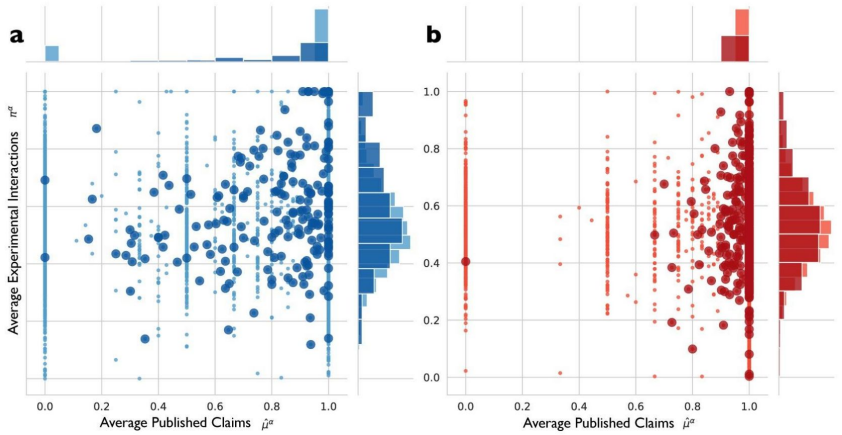


Correlation between claims and experimental strengths

CDF of experimental strength does not correlate well with the mean of claims' value, unless we start looking at more popular claims.

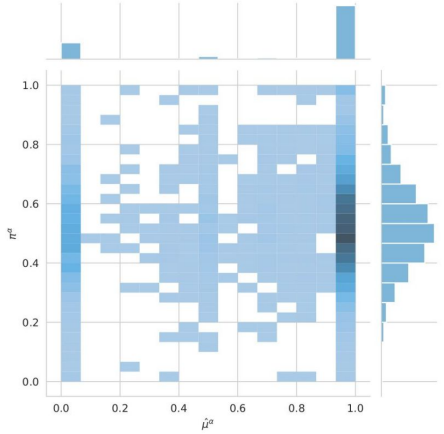
$$(\alpha, \beta) : \{(c_i, f_i)\} ; \boldsymbol{\pi}$$





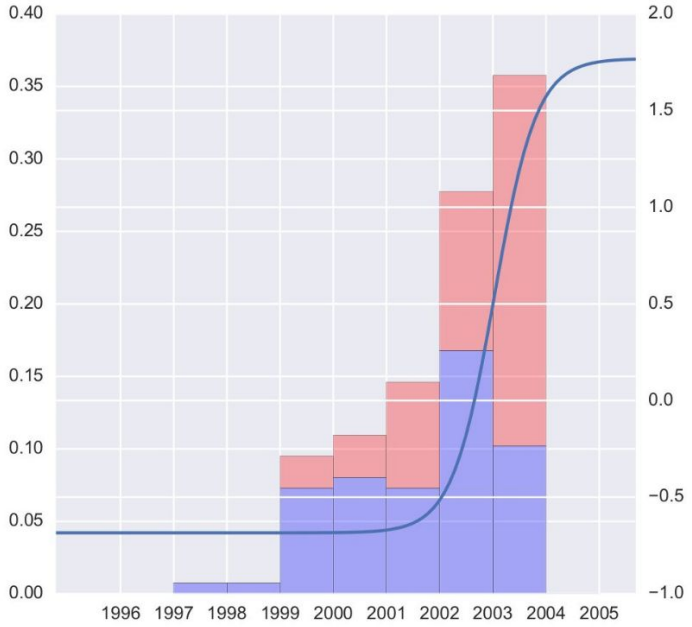
Preliminary Conclusions

- Literature contains two types of claims: positive and negative.
- The distribution of the claims in the claims corpora has a strong bias towards positive claims.
- The distribution of the experimental strength (of in the experiment does not have a strong bias.
- The positive claim bias varies between Geneways and Literome.
- The interaction strength can be discretized into at least 3 categories: neutral, positive and negative.
- The correlation between interaction strength and the mean claim increases as we consider more popular interactions (defined as having more claims per interaction).

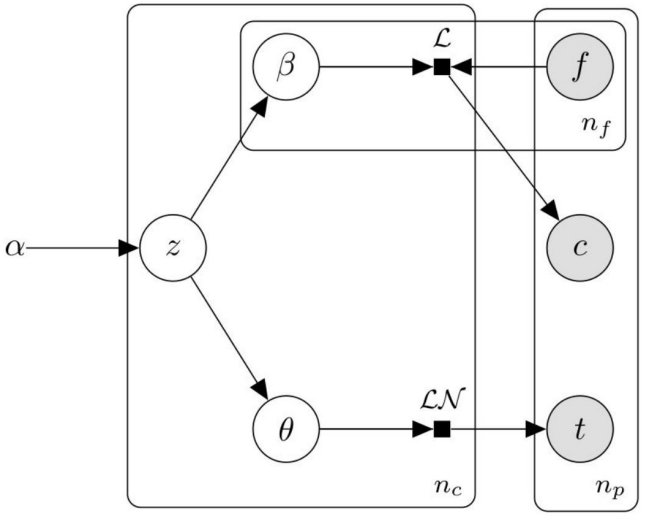


Generative Model of Claim Correctness: A Bayesian Approach to simultaneously model the underlying latent reality and the biases inherent in the publication process.

Mean window claims changes sign.



Bayesian approach, graphical models (pymc, pyro). Latent hyper-parameters α generate latent states θ and β, which generate observable publications at time t, features f and claim c.



Partition of interactions

Partition interactions into positive, neutral and negative: using Wasserstein distance between naive *Beta* posteriors derived from corresponding claims.

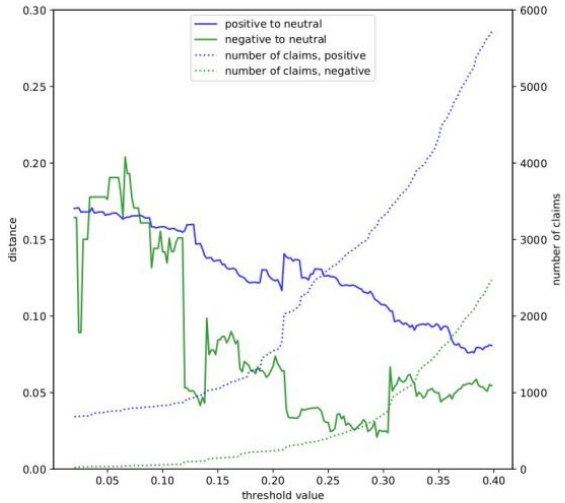
$$g_x(\mu) = Beta(a_0 + \sum_{\alpha \in \mathcal{C}_x} \sum_{i=1}^{n_\alpha} y_i^\alpha, \quad b_0 + \sum_{\alpha \in \mathcal{C}_x} \left(n_\alpha - \sum_{i=1}^{n_\alpha} y_i^\alpha \right))$$

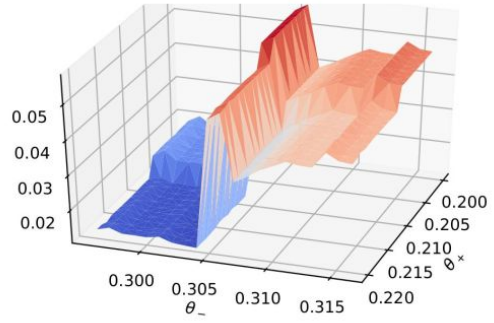
$$W(g_+, g_0) = \inf_{\gamma \in \Gamma(g_+, g_0)} \int d(x,y) d\gamma(x,y)$$

$$\theta_-^* = \arg\min_{\theta_-} \delta^L W(g_-, g_0, \theta_-, \theta_+)$$

$$\theta_+^* = \arg\min_{\theta_+} \delta^R W(g_+, g_0, \theta_-, \theta_+)$$

GeneWays	0.305	0.218
Literome	0.256	0.157





Target variables

interaction (α,β) : π_0 - interaction neutral? interaction neutrality

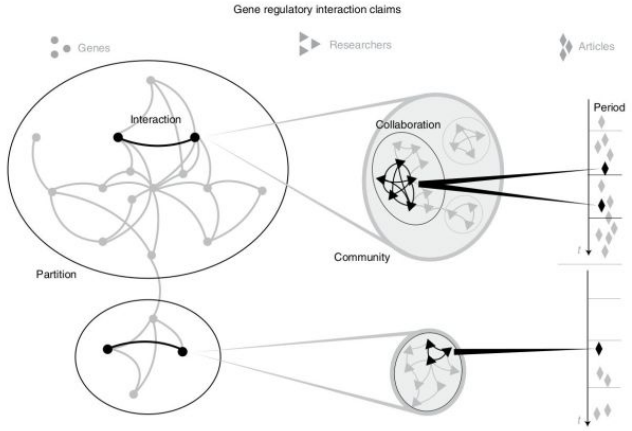
interaction (α,β) : π_+ - non-neutral interaction positive? interaction positivity

interaction (α,β), claim C_i : y_i - is this a correct claim? claim validity

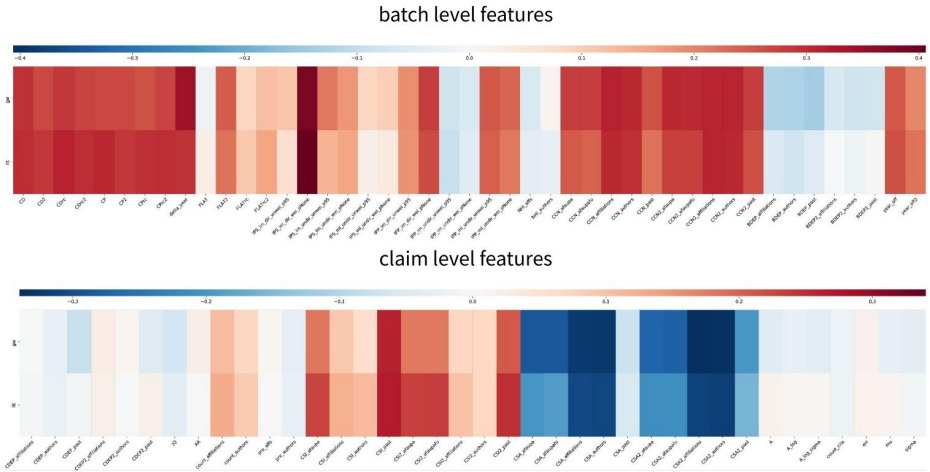
Types of features

claim level, batch level, interaction level

- features are defined with respect to an time interval
- infomap used for community detection



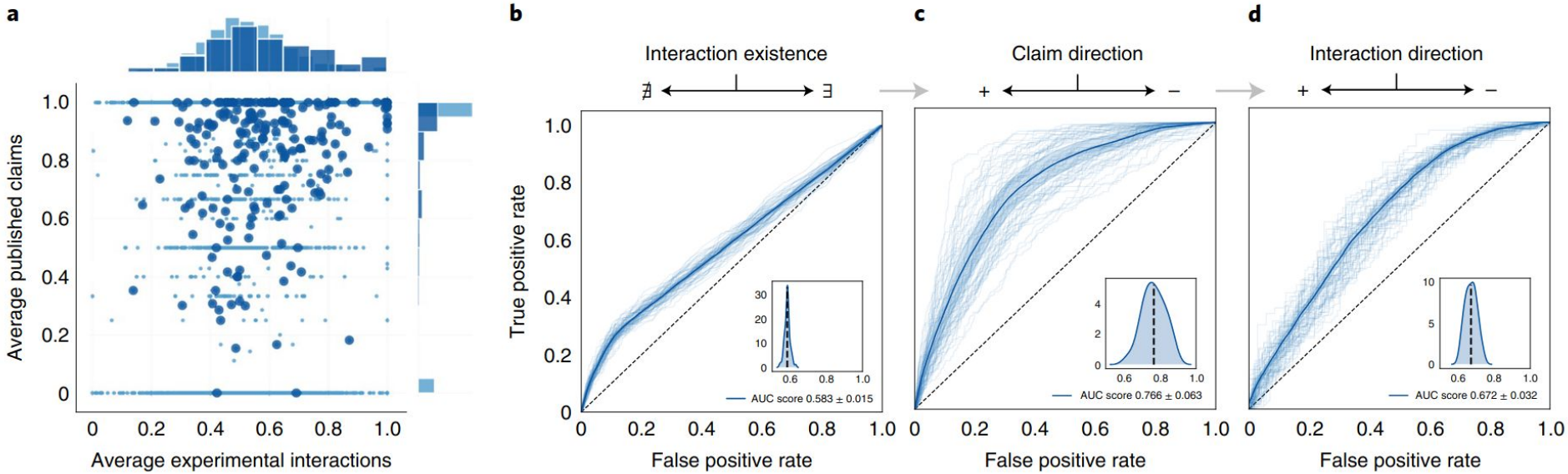
Model of Claim Correctness



20 threefold samples of interactions. 1 out of 3 for validation : 60 training–validation pairs.

- Samples are drawn randomly per interaction (using the claim number distribution function).
- Claim correctness model is trained and then validated on sets containing disjoint genetic interactions.
- Tests for over-fitting revealed the regime of high variance and therefore it is desirable to use models of low complexity

Integral Modeling Results

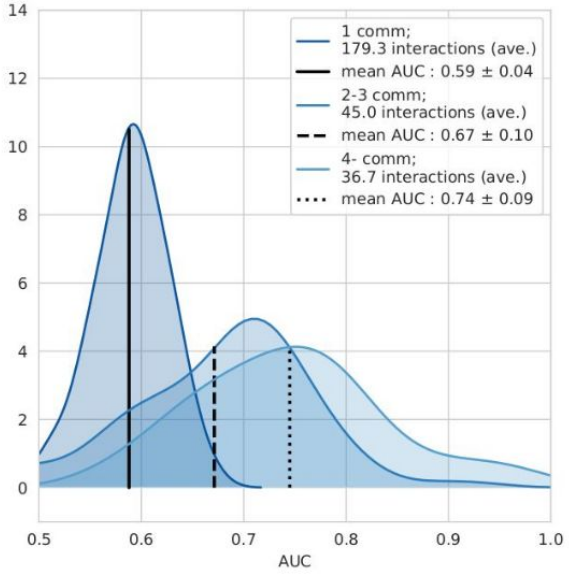


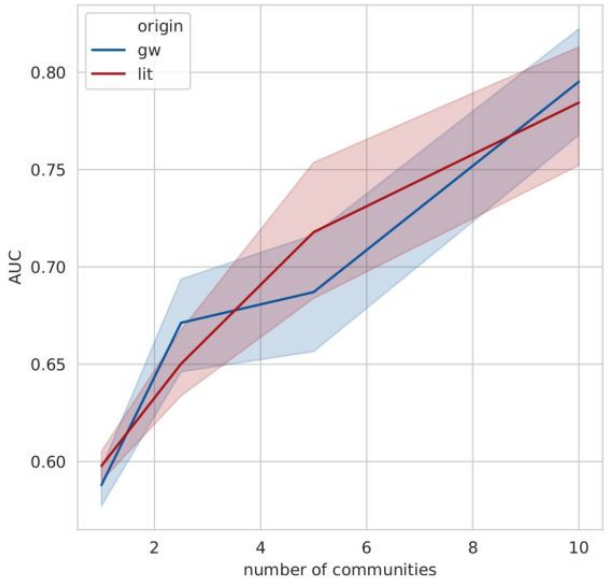
$$P(\pi_+^\alpha \mid \{(c_i^\alpha, f_i^\alpha)\}) \propto \prod_i P(c_i^\alpha, f_i^\alpha \mid \pi_+^\alpha) P(\pi_+^\alpha) \propto$$

$$\prod_i P(\pi_+^\alpha) \sum_{y_i^\alpha} P(c_i^\alpha \mid y_i^\alpha \pi_+^\alpha) P(y_i^\alpha \mid f_i^\alpha) \quad .$$

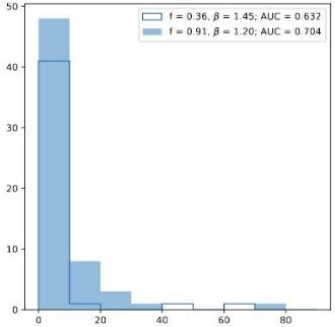
Policy A: promote independence

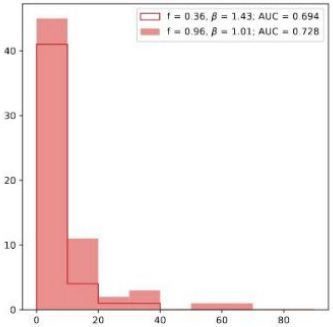
Selecting subsamples with more communities improves AUCs of interaction prediction model



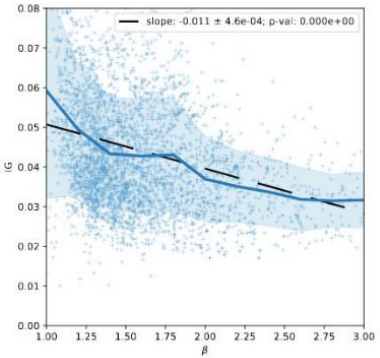


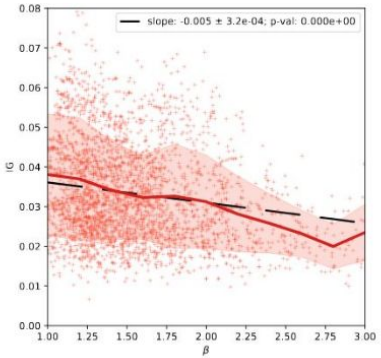
Policy B: altering the attention





Flatter distributions result in higher information gain

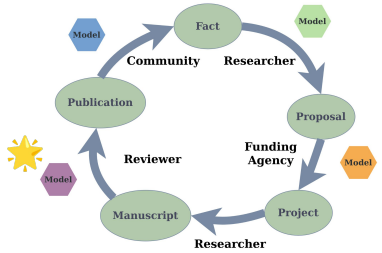




$$IG = \mathrm{ent}(p^{(0)}) - \frac{1}{k}\sum_{\alpha=1}^{\alpha=k}\mathrm{ent}(p^{\alpha})$$

$$\mathrm{ent}(p^{\alpha}) = -\sum p_i^{\alpha}\log p_i^{\alpha}$$

Classical Publication Metrics



Classical Citation Metrics

- Citation Count: Total references by other works.
- Field-Weighted Impact: Normalizes citations for cross-field comparison.
- Percentile Ranking: Ranks article performance within its cohort (e.g., top 1%, 10%).

Early Impact & Usage

- Usage Metrics: Downloads, views, and reads provide early impact signals before citations accrue.

Alternative Metrics (Altmetrics)

- Attention: Mentions on social media, blogs, and news.
- Engagement: Saves in reference managers like Mendeley.
- Altmetric Score: A composite measure of online attention.

Patent Citations

- Innovation Link: Tracks citations in patents, linking academic research to commercial technology.

Advanced Measures

- Disruptiveness (CD Index): Distinguishes consolidating vs. paradigm-shifting research.
- Network Centrality: Identifies influential hub papers that bridge research fields.

Shortcomings

- citations are subject to bias (institutional)
- societal impact is difficult to track
- impact might change sign
- “disruptive” research might have a better definition

Funding Metrics: Funding the Frontier

An interconnected data collection of 7M research grants, 140M scientific publications, 160M patents, 10.9M policy documents, 800K clinical trials, and 5.8M newsfeeds, with 1.8B citation linkages among these entities.

The effectiveness and usability of the system is evaluated using case studies and expert interviews.

Wang et al., arXiv:2509.16323

XSI: Knowledge Graph Generation

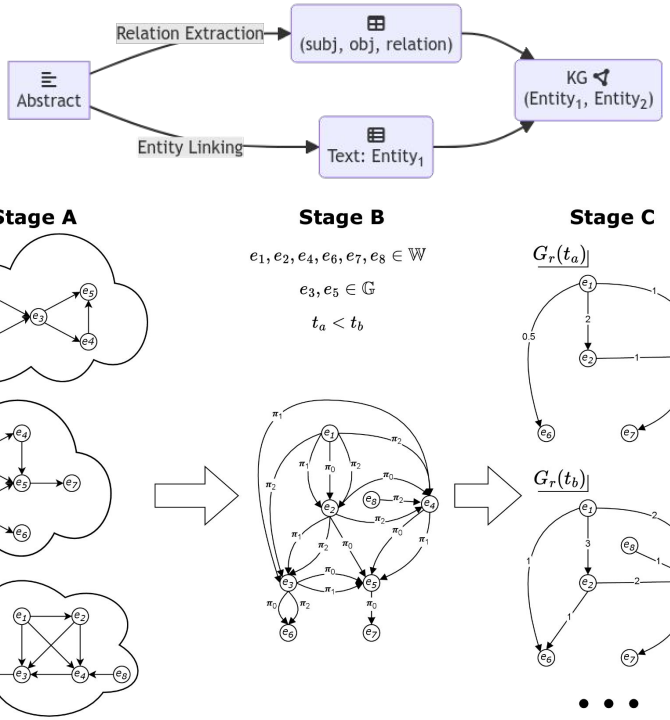
Input Data: **arXiv**, **biorXiv**, **medrXiv**

Relation Extraction: **Syntactic Features**
Entity Linking: **wiki, CHEBI, NCBI, omim**

Graph Representations:
Raw KG -> Redux KG -> Reference KG

Reference KG: KG of entities with weighted edges

arXiv:2502.13912



What is Semantic Impact?

Total entity graph: *Γ(t)* at time *t*.
Publication subraph: *γ*
*Semantic impact (*SI):

$$J_\gamma = \frac{1}{|\gamma|} \sum_{e \in \gamma} \left(\frac{w_e(\gamma \bigcup \Gamma(t+\Delta))}{w_e(\gamma \bigcup \Gamma(t))} - 1 \right)$$

Low impact case: publication subgraph repeats the edges already present in the reference graph at time *t* and these edges do not have higher weight at time *t+Δ*.
High impact case: publication subgraph introduces new edges in the reference graph at time *t* and these edges acquire high weight by time *t+Δ* (extra mentions).

Less biased than citations, correlates with disruptive research

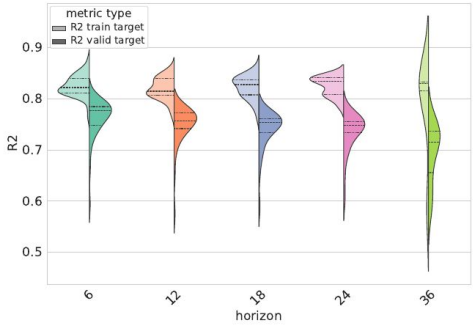
Models, Features

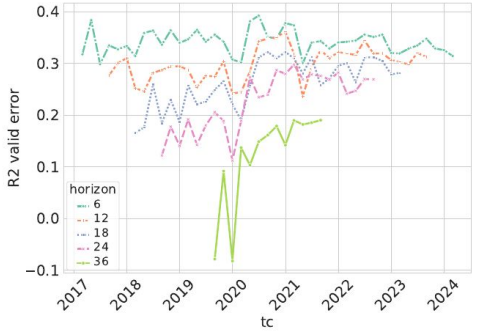
Models: add the model for the prediction error.
Features: network and diffusion features for publication subgraph γ (info flow through γ)

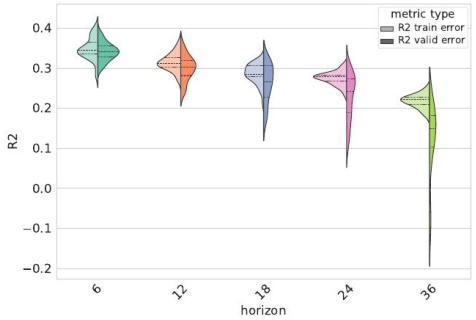
Plots: models for prediction of $r = \log J_{\gamma}$ and **prediction error** $\varepsilon = |r - r^{*}|$, with a horizon of 3 years, where r^{*} is the ground truth value.

Out-of-sample model performance: (R-squared) is >0.7 for **SI** and 0.2 for **SI error.**

NB: Covid trough.







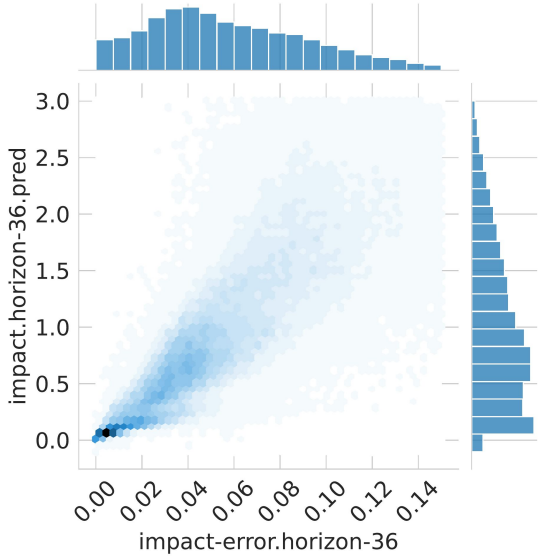
Efficient Frontier

Key Question: "How can we optimize academic research portfolios for maximum Semantic Impact ('Return') while minimizing Prediction Error ('Volatility')?

No time series: each publication is a single observation.
Semantic Impact: $r = \log J_y$
Prediction Error $\varepsilon = |r - r^*|$, with a prediction horizon of ***t*** years, where r^* is the ground truth value

NB: not at all Mean-Variance problem but an interesting analogy.

Financial mathematics analogy
impact : return, *error* : volatility



Portfolio Optimization For Academic Publications

For each period we take available publications, fix the fraction α of publications to choose, predict out of sample (future)

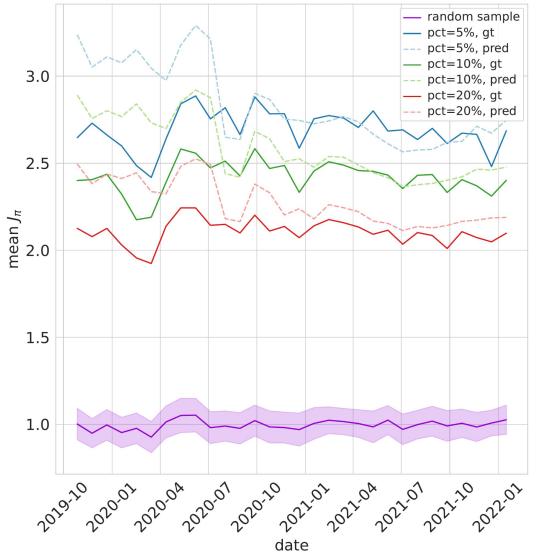
Integer Linear programming (0-1 knapsack)

$$\max \sum_{i \in I} (r_i - \epsilon_i) x_i$$

$$s.t. \frac{1}{|I|} \sum_{i \in I} x_i \leq \alpha$$

$$x_i \in \{0, 1\}, \quad \forall i \in I$$

Prediction of at *t+3* years



XSI Model online

Results

- SI has low but statistically significant correlation with citation counts (~ 0.2)

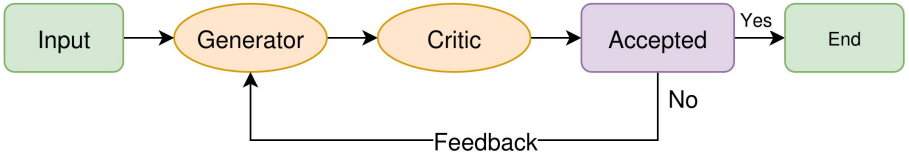
Shortcomings

- Quality of research execution is not evaluated

Future Steps

- Extend the model beyond biomedical domain
- Add the quality of research execution

Agentic AI (for Science)



Agents - LLMs with Initiative: It's a Large Language Model that doesn't just answer questions — it sets goals, makes plans, and takes action.

The Go-Getter Flow: Operates on an iterative loop: Plan → Execute → Critique → Refine.

Knowledge-Grounded Reasoning: Uses Retrieval-Augmented Generation (RAG) to find, verify, and incorporate real-time data from knowledge bases.

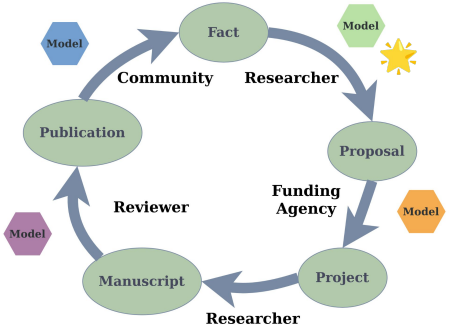
Self-Correction is Key: Includes an internal "Critique" mechanism to evaluate its own output and fix mistakes autonomously.

Tool Masters: Connects the LLM to specialized tools (like code execution, APIs, or database queries) to interact with the world.

From **Answerer** to **Investigator**: Shifts the AI role from a reactive chat partner to a proactive, independent researcher.

Dynamic Goal Pursuers: Capable of breaking down a complex, high-level goal into a sequence of smaller, manageable tasks.

Hypothesis Generation



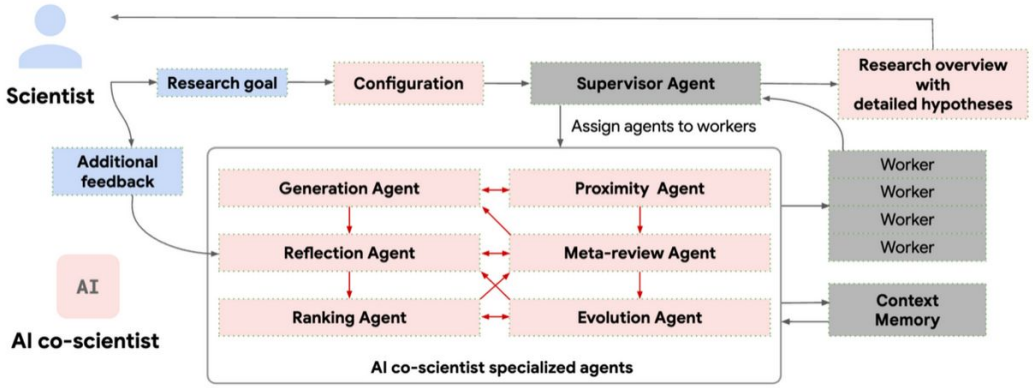
Hypothesis Space Reduction: e.g. from the perspective of hypothesis validation, AlphaFold's primary power is its ability to dramatically constrain the solution space for protein structures.

Agentic Discovery represents a paradigm shift, where AI agents proactively pursue complex goals.

These systems leverage Retrieval Augmented Generation to ground their actions in real-time, external data, moving beyond static knowledge. Crucially, they operate through iterative **Critique feedback loops**, continuously refining their approach based on self-evaluation or external input. This "Agentic AI" is distinct from sociological Agentic Modeling, as it refers to LLM-driven agents exhibiting purposeful intent and task-oriented autonomy. The result is a dynamic cycle of planning, action, and learning that accelerates scientific and intellectual discovery.

Research on AI Scientist

1. AI Co-Scientist (Google)
   Multi-agent system on Gemini 2.0
2. The AI Scientist-v2
3. Robin: Multi-Agent Scientific Discovery System
4. Iterative Hypothesis Generation with MC-NEST
5. DeepScientist
6. AstroAgents
7. Empirical Software Scientist



*AI co-scientist system overview. Specialized agents (**red boxes**, with unique roles and logic); scientist input and feedback (**blue boxes**); system information flow (**dark gray arrows**); inter-agent feedback (**red arrows** within the agent section).*

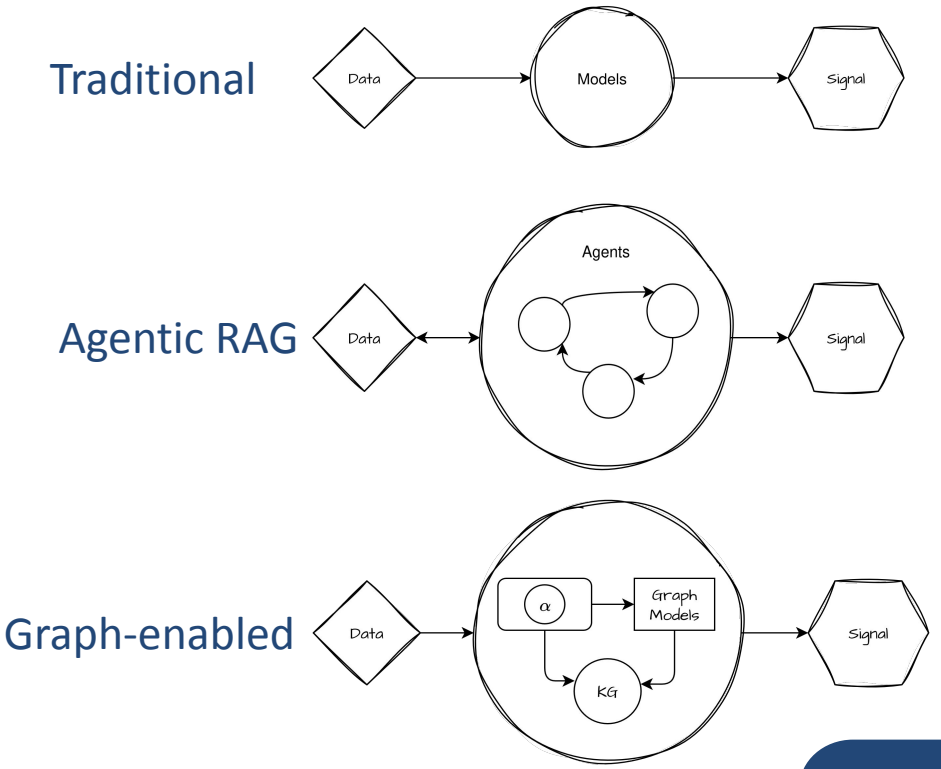
AI Scientist Initiative | Funding Calls | Startups

1. ARIA AI Scientist
2. Perplexity
3. Zilliz Deep Researcher
4. Edison Scientific's Kosmos

Traditional vs Deep Research vs Graph Optimization

Traditional scientific models excel at evaluating research quality and impact, offering statistically rigorous predictions. In a different dimension, LLM-enabled hypothesis generation models aka **Deep Research/Deep Research** produce novel research proposals based on generative experience, though they often lack a robust statistical foundation and a holistic view of the data.

An **agentic graph optimization approach** is as a synthesis, aiming to capture the statistical rigor of the former and the creative, generative potential of the latter.



Shortcomings of Vanilla Deep Research Frameworks

Lack of Deeper Grounding & Structure (The Core Problem)

- Generated hypotheses lack verifiable statistical support because they rely on text embedding similarity (RAG) rather than structured, multi-hop reasoning over facts in a Knowledge Graph (KG).
- They struggle to incorporate complex, structured domain knowledge (e.g., hierarchical ontologies, causal links) crucial for scientific fields.

Inability for Statistical Interpretation and Optimization

- Without a world/twin graph, these systems cannot run statistically interpretable models (like Graph Neural Networks) or perform quantitative optimization (like portfolio selection or risk modeling) over the suggested hypotheses.

Inefficient Context Management

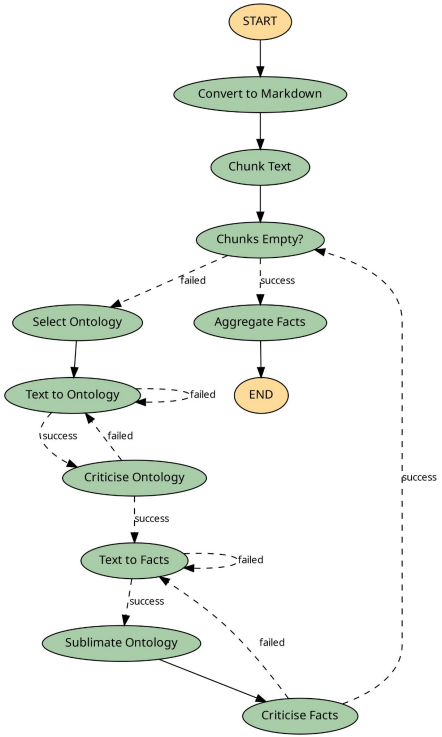
- Basic RAG systems hit context window limits when synthesizing information from hundreds of documents, leading to an inability to maintain a complete, high-fidelity 'world state' needed for long-horizon planning.

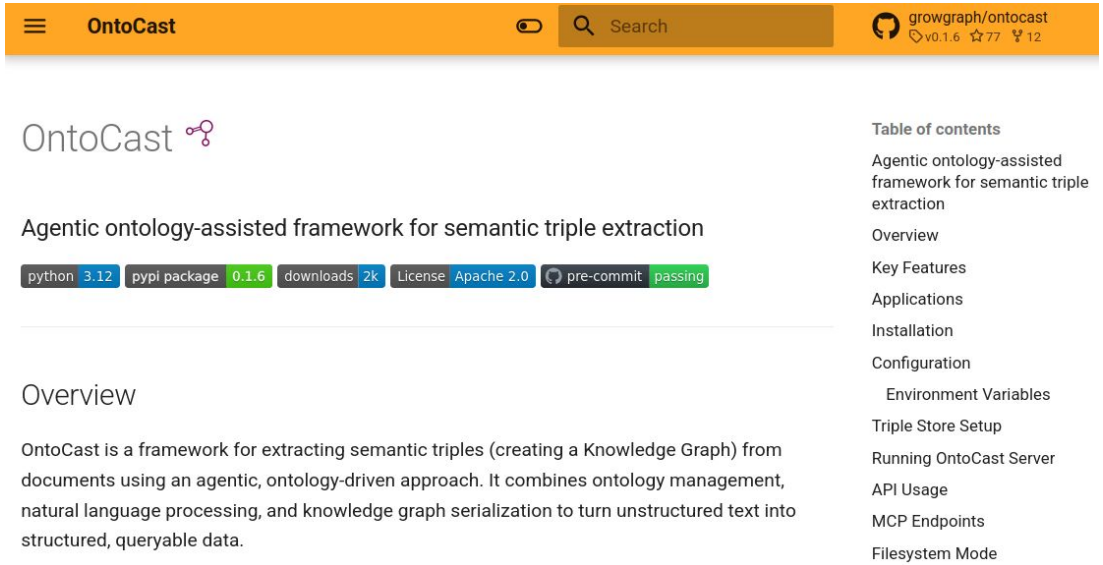
Bias Reproduction at Scale

- They reproduce the biases of the text corpus (e.g., confirmation bias, over-reliance on popular claims) and fail to use the structure of a KG to quantitatively debias knowledge and promote epistemic diversity.

Ontocast

Ontology Assisted Agentic Transformation to Semantic Triples (RDF)





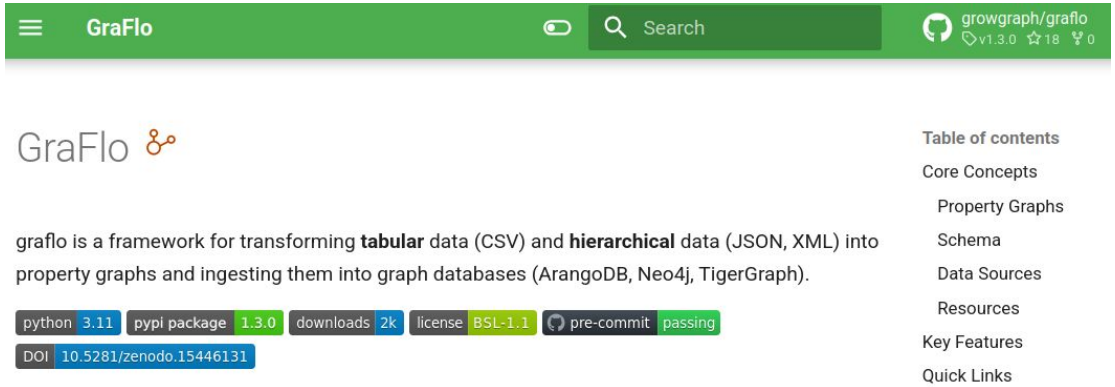
https://growgraph.github.io/ontocast

GraFlo

Property graph DBs are the workhorses of graph modeling and analytics, they facilitate data prep for complex models.

Graflo is a multi-adapter that speeds up data ingestion in Property graphs. It features:

- Declarative transformations vs. custom ETL coding
- Adapters for Neo4j, ArangoDB and TigerGraph: Multi-database adapter eliminates vendor lock-in
- Tested on graphs with billions of edges



https://github.com/growgraph/graflo

Ontocast + Graph Optimization

🔧 **Ontocast: Ontology-Assisted KG Construction**

Automated Knowledge Graph Generation from Unstructured Data

- **Semantic Triple Extraction**: LLM-powered entity and relation extraction with domain-specific ontologies
- **Ontology Integration**: Pre-built schemas for research domains
- **Multi-Source Ingestion**: Pre-prints, Grant Proposals, Patent, Economic and Financial Data
- **Quality Assurance**: Automated entity linking, co-reference resolution, and fact verification

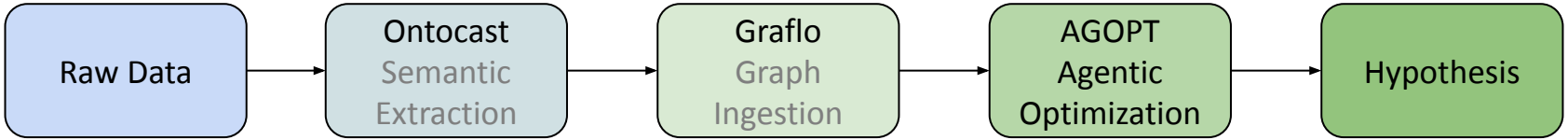
Key Benefits:

- Reduces KG construction time from months to days
- Ensures semantic consistency across heterogeneous data sources
- Maintains provenance and confidence scores for all extracted facts

📊 **Graph Optimization Engine: Signal Generation**

- **Resource Portfolio Optimization**: Integer programming over knowledge graphs for asset selection
- **Novel Hypothesis Generation:** The graph structure facilitates LLM-driven multi-hop reasoning to discover latent, non-obvious connections in the data, generating truly novel hypotheses that link previously isolated concepts.
- **Hypothesis Validity/Risk Modeling: Performs** Multi-hop Dependency Analysis and Graph Neural Network (GNN) inference to assess the systemic validity of a hypothesis, identifying claims that are structurally unsound or rely on known biased literature.

KG generation + Agentic Optimization

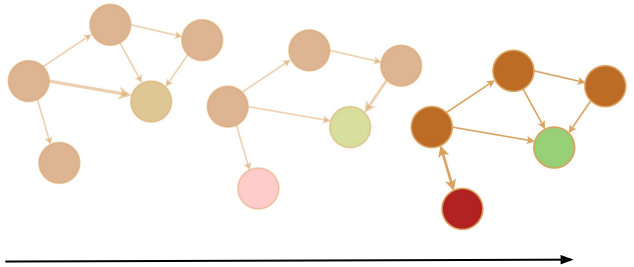


Ontocast extracts semantic triples from unstructured text

Graflo ingests these triples plus structured data into property graphs

AGOPT runs GNNs and agentic optimization over the unified knowledge graph [Symbolic AI]

KG provides content; LLM provides shape



Conclusions

- Science is Under Strain: Exponential information growth and systemic reproducibility issues challenge the current scientific process. The literature is biased toward positive claims over established facts.
- Epistemic Diversity is Key: Communities mixing exploratory Mavericks and refining Refiners discover truth faster and more persistently than those focused solely on consensus.
- AI for Evaluation: AI is necessary to derive less-biased metrics, such as Semantic Impact (SI), which is a stronger predictor of a publication's future influence than traditional citations.
- The Future is Agentic AI: The next paradigm is moving from static evaluation to Agentic AI - a system that actively generates, critiques, and refines complex scientific hypotheses.
- Agentic Graph Optimization Avenue: True discovery requires an agentic approach that unifies the creativity of LLMs with the structure and statistical rigor of Knowledge Graphs (KG) (e.g., Ontocast/Graflo/AGOPT).