Georgia Tech | Wallace H. Coulter Department of Biomedical Engineering
Georgia Tech College of Engineering and Emory School of Medicine

EMORY UNIVERSITY

NEURAL INFORMATION PROCESSING SYSTEMS

SAGE: Evidence-First Biomarker Discovery Through Multi-Agent Reasoning



Dr. Sahar A. Nasser
Emory University



Juan P. Borja
MS, Georgia Tech



Jincheng Liu
MS, Georgia Tech



Dr. Sandeep Manandhar
Emory University



Shikhar Shiromani
MS, Georgia Tech



Tanvir Hasan
MS, UALR



Zenghan Wang
MS, Georgia Tech



Suman Ghosh
PhD Scholar, Georgia Tech



Jinchu Li
MS, Georgia Tech



Xuejian Xu
MS, Georgia Tech



Aniket Iyer
MS, Georgia Tech



Dr. Naoto Tokuyama
Emory University



Twisha Shah
MS, Georgia Tech



Tilak Bahadur Pathak
MD, Emory



Soundharya Kumaresan
PhD Scholar, Emory



Yohei Abe
MD, Emory



Himanshu Maurya
Researcher, Emory



Dr. Anant Madabhushi
PI, Emory

Biomarkers guide cancer treatment. Discovering them remains a bottleneck.

WHY BIOMARKERS MATTER

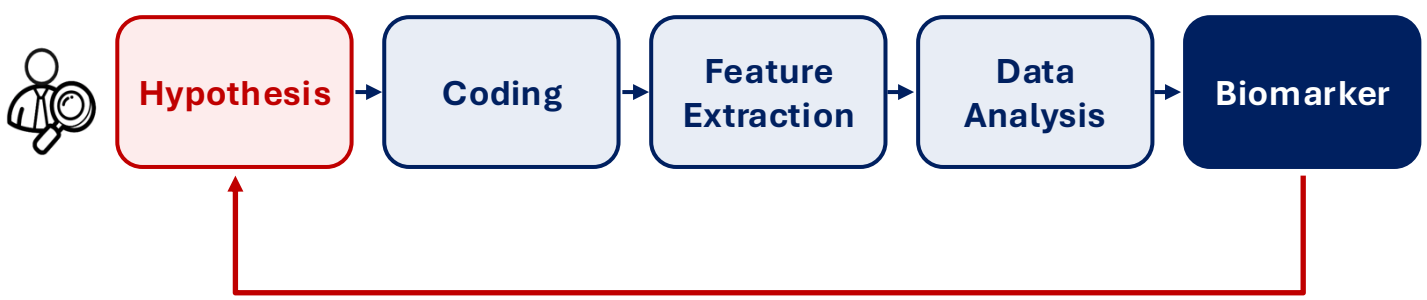


Prognostic
who is at risk



Predictive
who responds to a treatment

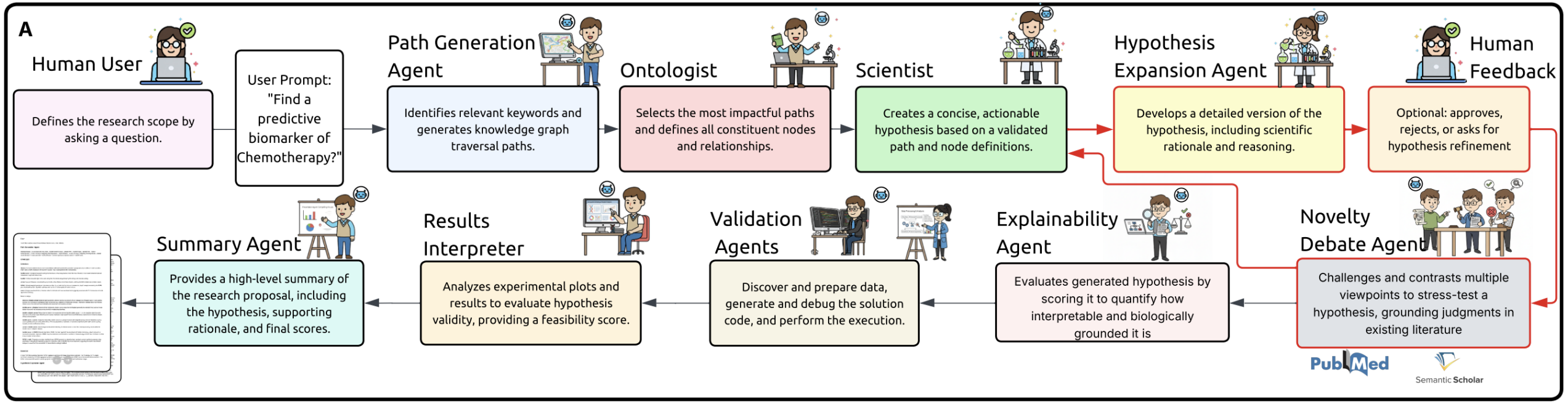
WHY DISCOVERY IS A BOTTLENECK



- **Knowledge overload**
- **Limited exploration**
- **Costly validation**

SAGE automates and accelerates biomarker discovery

SAGE agents generate, debate and validate each hypothesis



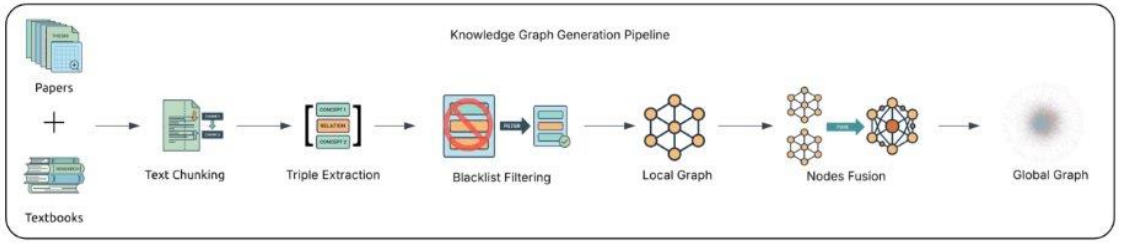
1 End-to-end pipeline
Hypothesis to validation on patient cohorts

2 Multi-path reasoning
Over a pathology knowledge graph

3 Debate novelty
Stress-tests ideas against literature

4 Grounded results
Interpretable, evidence-linked results

A knowledge graph from the literature grounds every hypothesis



41,053
nodes

56,338
edges

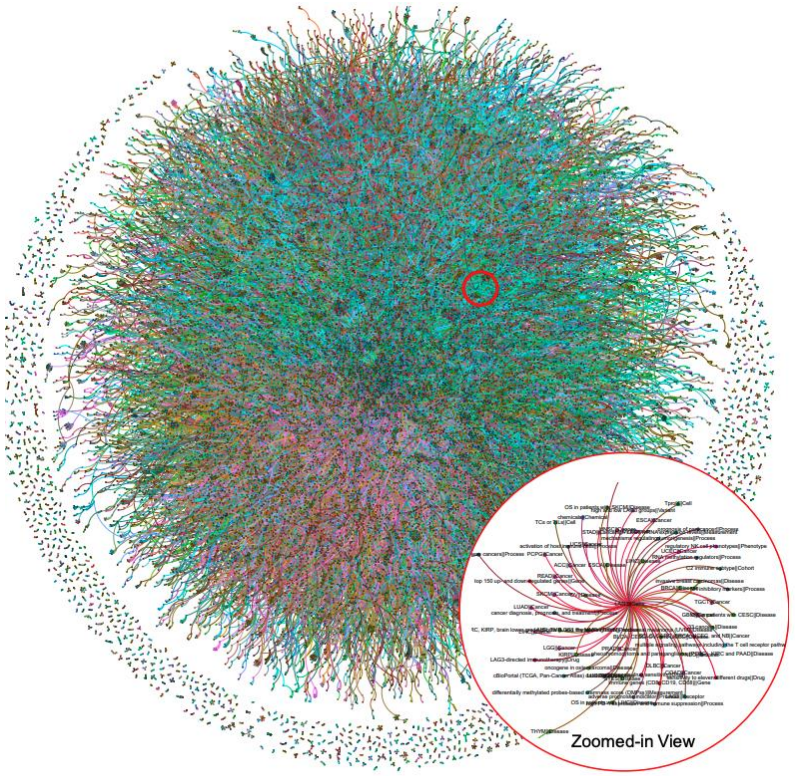
1,650
sources

99.0%
factual grounding

99.0% factual grounding: 198 of 200 triples, sampled at random across relation types, were supported by the source passage they cite.

Entity type (10 of 35)

- Process
- Disease
- Gene
- Measurement
- Phenotype
- Symptom
- Protein
- Drug
- Cohort
- Chemical



Knowledge graph, one node per entity, colored by type

Removing Surprise changed the hypothesis ranking in 100% of test cases

Logic plausibility — Path matches the question asked

Relevance plausibility — Path stays close to the target disease domain

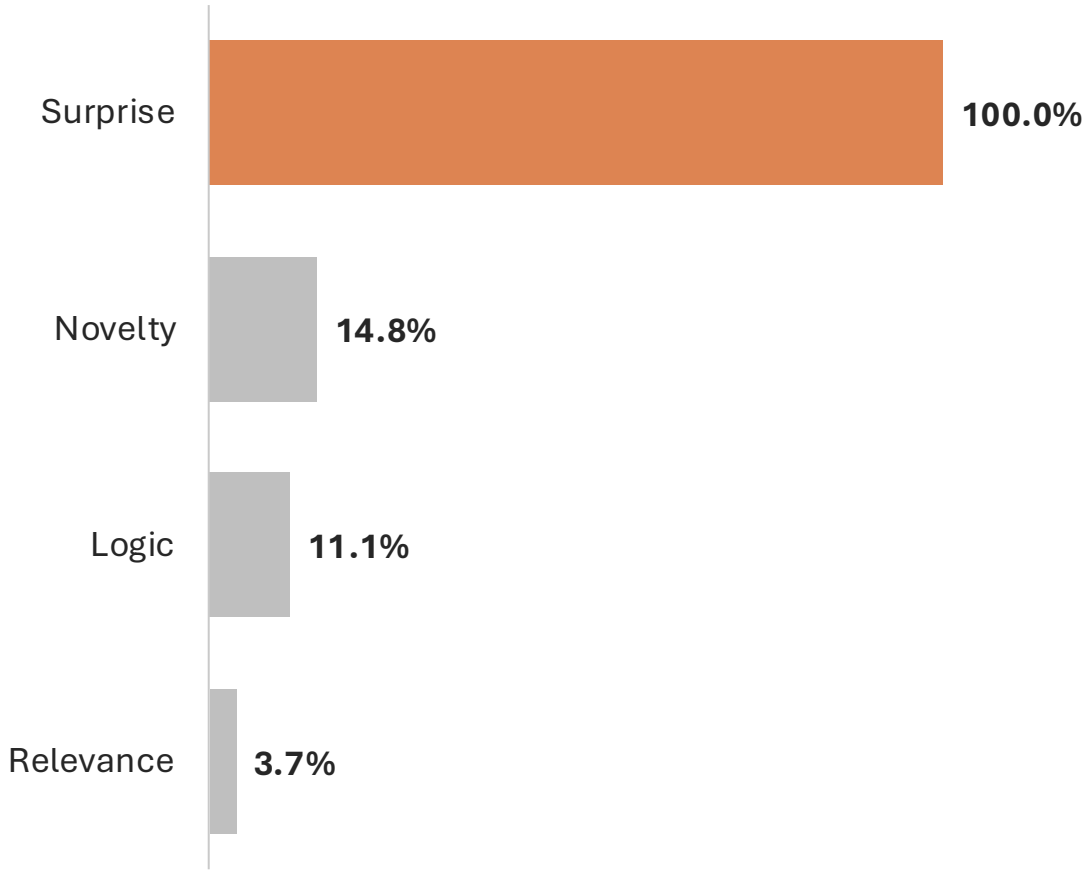
Novelty discovery — Path uses rare, underexplored entities

Surprise discovery — Path mixes entity types in unexpected ways

Weighted total weights set by user

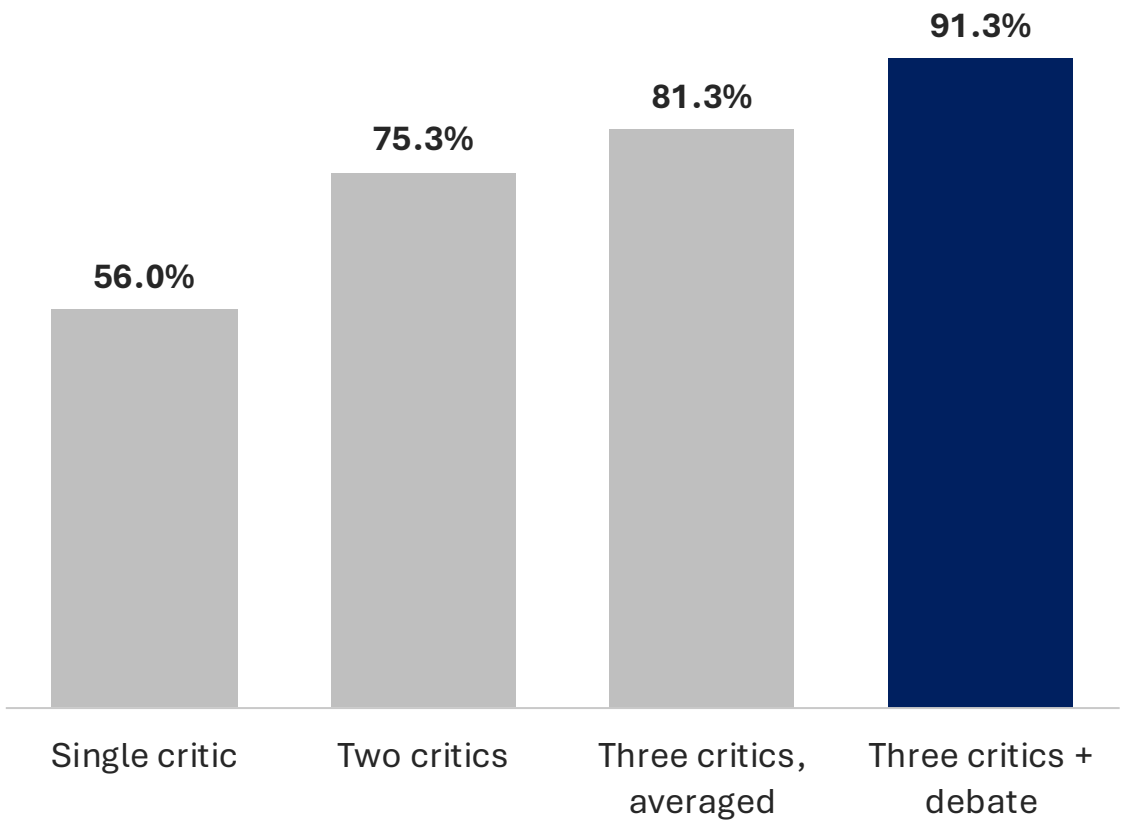
$$S_{\text{total}} = \sum_i w_i S_i$$

Ranking changed after removing each score, % of tested examples



Multi-critic debate lifted novelty-scoring accuracy from 56% to 91%

Novelty-assessment accuracy by critic setup



Score gap = mean Tier A (breakthrough) score minus mean Tier E (specious) score.

SCORE GAP

5.02 → 6.78

Mean Tier A minus Tier E score, without vs with debate: +1.76 ($p < 0.05$)



HOW IT WAS TESTED

150 historical proposals, 30 per tier, from Nobel-level breakthroughs (A) to retracted claims (E). The authors set the tiers from the historical record.

COULD THE MODELS ALREADY KNOW THE OUTCOMES?

Each proposal was judged only against literature from before its publication or retraction date. The models' own memory of famous cases is not ruled out and could inflate accuracy.

Multi-path reasoning produced the most novel hypotheses

	Debate agent scores				Human evaluation (blinded)			
System	Novelty	Conf.	Cons.	Vendi	Novelty	Feasibility	Conf.	Cons.
ChatGPT 5.2	6.91*	6.70*	7.06	5.00	6.74	**10.00***	7.07	6.83
Gemini 3.1 Pro	6.90*	6.30	7.23	**6.10**	4.53*	8.56*	7.00	6.92
AgentFlow	6.82*	**6.76***	6.76	5.22	4.83*	8.27*	7.67	6.67
SAGE (Text)	7.00*	6.18	7.21	4.51	5.41*	7.26	**7.93**	6.50
SAGE (Single-Path)	7.33	6.42*	7.18	4.11	6.36	7.86	7.13	**7.25**
SAGE (Full)	**7.67**	6.00	**7.85**	5.07	**7.10**	7.03	6.40	6.83

Experts rated SAGE's hypotheses less feasible because the associations were less familiar to them.

Bold: best in column. *: differs significantly from SAGE (Full), non-overlapping 95% CIs. Conf.: confidence. Cons.: consensus. Vendi: diversity.

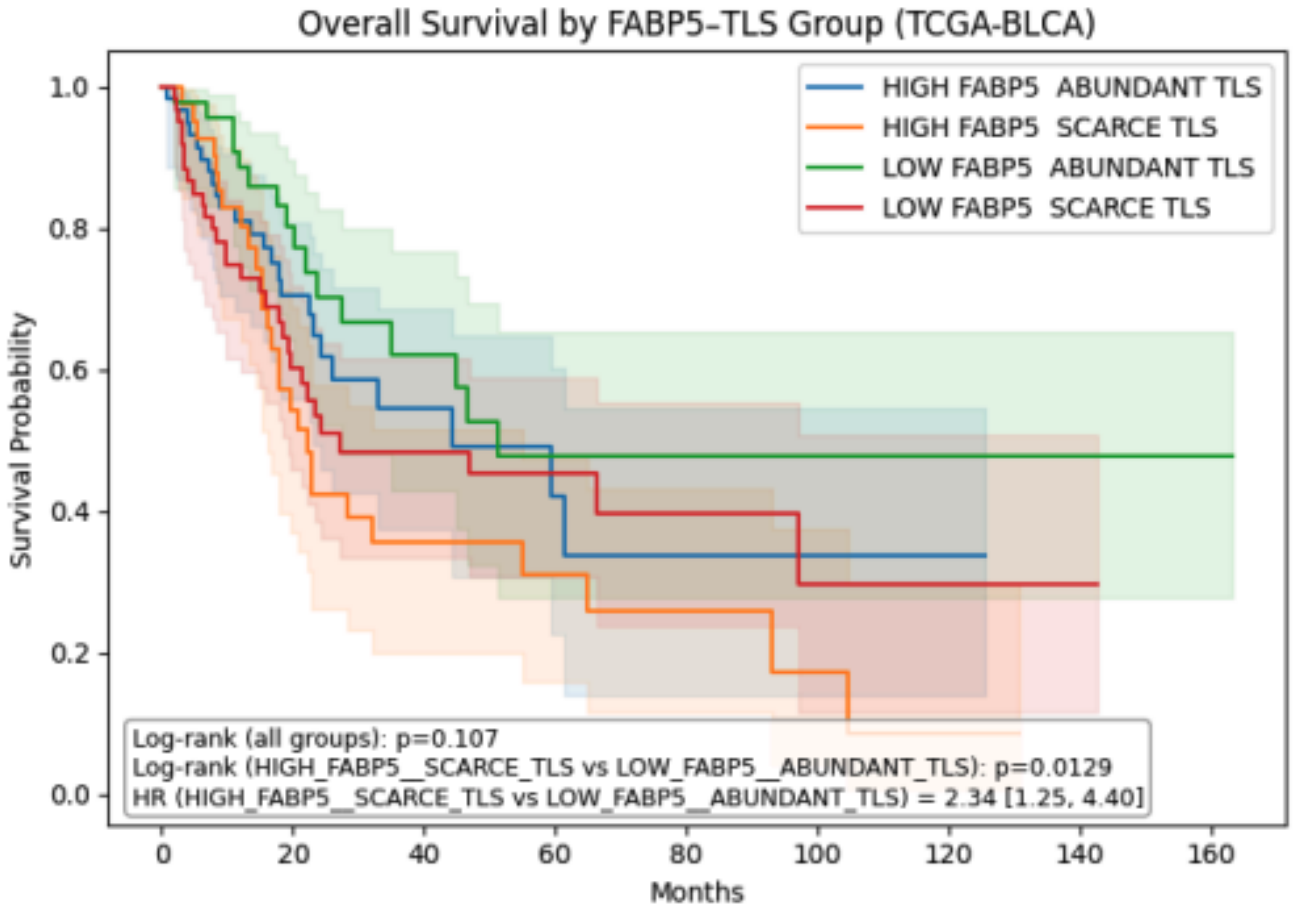
SAGE found a multimodal biomarker of poor bladder cancer survival

High **FABP5** expression with scarce **tertiary lymphoid structures (TLS)** on whole-slide images predicts worse overall survival in bladder cancer.

2.34
hazard ratio of death

0.013
log-rank p-value

412
patients, TCGA-BLCA



A biomarker absent from prior literature, validated on patient data without manual intervention.

Takeaways

Generate → Debate → Validate → Discover

5.07

Hypothesis diversity (Vendi score) with multi-path graph reasoning, vs 4.51 for text-based retrieval

91.3%

Novelty-scoring accuracy with three-critic debate, vs 56.0% for a single critic

+40%

Validation success with a dedicated dataset discovery agent

HR 2.34

Novel FABP5 and scarce-TLS biomarker of poor survival, validated on 412 patients, $p = 0.013$

SAGE treats biomarker discovery as a structured reasoning problem: grounded in evidence, stress-tested for novelty, and validated on patient data.