AI Scholar: Learning the Scientific Taste of Elite Researchers to Predict Future Discovery Trajectories

OpenAGS.org

February 24, 2026

Abstract

As artificial intelligence transitions from an analytical tool to an autonomous agent of scientific discovery, a critical gap remains: current generative ideation systems produce generic, domain-level proposals that lack the idiosyncratic intuition, methodological preferences, and risk tolerance of human experts. We propose *AI Scholar*, a novel computational framework designed to operationalize and learn "scientific taste." By ingesting the lifetime bibliometric trajectories of 10,000 elite, independent principal investigators, our system constructs temporal latent representations of individual research styles. Unlike existing evaluation benchmarks, AI Scholar utilizes these taste embeddings to condition a multi-agent generative pipeline, autonomously predicting and synthesizing the precise future research proposals these scientists are likely to pursue. This paper outlines the theoretical foundations, the novel Trajectory-Conditioned Ideation methodology, and the time-partitioned evaluation design required to validate this massive-scale predictive framework, ultimately exploring the scaling laws of personalized scientific discovery.

1 Introduction and Motivation

1.1 The Evolution of Scientific Discovery and Collaboration

The landscape of scientific research has undergone a profound transformation over the past two decades. Seminal work by Wuchty et al. [23] documented that teams have increasingly dominated the production of knowledge, fundamentally altering how discoveries emerge. Complementing this, Boyack et al. [5] demonstrated that scientific fields evolve as complex ecosystems, with individual researchers playing distinct roles as pioneers, consolidators, or synthesizers [13]. Yet despite rising computational capabilities, the automation of scientific ideation remains rooted in a reductionist paradigm: these systems treat scientific discovery as a problem of *domain-level* optimization rather than *researcher-level* personalization.

1.2 The Current Epoch of AI-Driven Discovery

The pursuit of automating scientific discovery has entered a new epoch, driven by the reasoning and generative capabilities of Large Language Models (LLMs). Recent frameworks have successfully demonstrated that multi-agent systems can autonomously generate research ideas, write code, execute experiments, and even draft manuscripts [26, 3, 18]. Furthermore, empirical studies on the scaling laws of such AI and robot scientists suggest that computational discovery efficiency will soon outpace traditional human-driven paradigms [27].

However, a fundamental limitation persists in the current generation of AI ideation agents: they are inherently *domain-centric* rather than *researcher-centric*. When tasked with generating novel ideas, existing frameworks generally retrieve a set of recent papers and apply prompt-engineering techniques (e.g., the NOVA framework [8]) to extrapolate the next logical step. While this produces ideas that are statistically "novel," they are often disembodied from the reality of scientific practice. The mathematical abstraction of "novelty" (typically measured via semantic distance from existing literature) does not capture the idiosyncratic intuitions that guide great scientists.

1.3 Scientific Taste: The Missing Dimension

Real scientific breakthroughs are driven by *scientific taste*—a researcher's unique blend of methodological preferences, epistemic risk tolerance, interdisciplinary background, and aesthetic judgement regarding what constitutes an "elegant" solution. This concept, grounded in philosophy and empirical sociology of science, has been theoretically formalized by Acerbi [1], who demonstrated that taste constitutes a coherent cognitive framework that can be systematically analyzed. More concretely, Rzhetsky et al. [16] showed empirically that researchers *deliberately select* experiments and problems in ways that reflect deep, latent preferences—not via random sampling from the available possibility space.

Wang et al. [22] later quantified the notion of novelt within individual research trajectories, showing that what appears "novel" to one researcher may seem derivative to another, depending on their historical exposure and methodological background. This suggests that predicting future research requires not merely understanding the global frontier of a field, but rather acquiring a personalized model of each scholar's decision-making apparatus.

1.4 Temporal Dynamics and Trajectory-Level Prediction

Recent literature has begun to recognize the necessity of modeling temporal scientific trajectories with rigorous, time-aware evaluation protocols. The *Proof of Time* benchmark [25] leverages time-partitioned evaluation

to assess an LLM's ability to judge scientific ideas against future observable metrics, such as citation impact and peer-review awards. This framework establishes that scientific evaluation must be fundamentally *temporal*—hypotheses about future work can only be validated once that future arrives.

However, while *Proof of Time* focuses on the ex-post *evaluation* and *judging* of ideas, our work addresses the complementary ex-ante *generation* of ideas modeled on specific human trajectories. By predicting what *that individual researcher* will pursue next, we move beyond genre-level forecasting to the fine-grained modeling of research evolution.

1.5 The Vision: Operationalizing Scientific Taste at Scale

We propose the *AI Scholar* project. Our core motivation is to **bridge the gap between generic LLM ideation and personalized scientific evolution**. By scaling our analysis to 10,000 of the world's top scientists, we aim to extract mathematical representations of individual scientific taste. We will subsequently use these representations to predict their future research agendas, autonomously generating detailed proposals that reflect not just what the field might do next, but what *that specific scholar* will do next.

The significance of this endeavor is threefold:

1. **Scientific Impact:** By learning from elite researchers' decision-making patterns, we can discover structural regularities in how breakthrough research emerges—potentially uncovering universal principles of scientific taste.

2. **Methodological Contribution:** This work introduces novel techniques for fusing bibliometric trajectories with deep learning (heterogeneous graph neural networks, contrastive learning, and LLM-conditioned generation).

3. **Practical Application:** If successful, such a system could become a personalized research forecasting tool, augmenting human scientists' ability to anticipate emerging opportunities and emerging competitor research directions.

2 Related Work

2.1 LLM-Based Scientific Ideation and Generative Discovery

The application of Large Language Models to scientific ideation has rapidly evolved. Baek et al. [3] introduced ResearchAgent, a system that iteratively generates research ideas by mining scientific literature and applying LLM-based reasoning. Hu et al. [8] extended this with NOVA, which combines

iterative planning and diverse search strategies to enhance both novelty and diversity of generated ideas. More recently, Sinha et al. [18] demonstrated that LLMs can autonomously design rigorous experimental protocols, significantly advancing the feasibility of end-to-end automated research.

Complementary to these works, Zhang et al. [26] presented aiXiv, a comprehensive ecosystem for AI-driven scientific discovery, integrating paper generation, code implementation, and experimental execution. The latter showcases sophisticated multi-agent coordination, in which specialized agents assume distinct roles (architect, reviewer, synthesizer) to collaboratively produce coherent research outputs. Our work builds upon these architectural insights but introduces a fundamentally new constraint: *taste-conditioning*. Rather than optimizing for generic novelty or field-level plausibility, we constrain generation to align with the historical decision patterns of specific researchers.

Valmeekam et al. [20] have also examined the planning and reasoning capabilities of LLMs in complex, multi-step problem-solving scenarios. This work informs our multi-agent design, particularly the role of the Taste-Critic Agent in filtering proposals for authorial alignment. However, existing work largely treats LLMs as general-purpose planners; our contribution is to *personalize* the planning objective itself.

2.2 Computational Modeling of Research Trajectories and Scholar Behavior

The quantitative analysis of individual research trajectories has deep roots in scientometrics. Rzhetsky et al. [16] pioneered empirical methods to model how researchers select problems, demonstrating that choice patterns reflect systematic preferences rather than uniform randomness. Their framework for characterizing researcher decision-making provides crucial theoretical support for our hypothesis that taste is a learnable, predictive signal.

Boyack et al. [5] advanced the field by characterizing how research fields and their leaders co-evolve over time, using bibliometric data to trace the emergence and decline of research directions led by individual scientists. This work motivates our focus on elite, independent researchers whose decisions genuinely shape disciplinary trajectories.

Wang et al. [22] developed quantitative measures for research novelty within individual publication histories, showing that novelty is not absolute but *relative to each researcher's prior knowledge and trajectory*. This finding is essential: it implies that predicting what appears "novel" to a specific researcher requires personalized calibration.

More conceptually, Acerbi [1] provided philosophical and empirical foundations for understanding taste as a structured, learnable construct. His work demonstrates that aesthetic judgment in science (as in art) follows coherent principles and can be systematically analyzed. Additionally, Wuchty

et al. [23] and Newman [13] established that research collaboration patterns encode deep information about research organization and diffusion, providing theoretical grounding for our use of co-authorship networks to identify independent PIs.

2.3 Temporal Graph Learning and Dynamic Embeddings

Our approach to operationalizing taste relies on heterogeneous, dynamic graph neural networks. Hamilton et al. [7] introduced GraphSAGE, a scalable framework for learning inductive node representations on large graphs—crucial for handling 10,000 researchers. This foundational work enables us to generalize taste embeddings to unseen scholars.

Pareja et al. [14] proposed EvolveGCN for modeling evolving graph structures, allowing nodes' representations to adapt over time. Xu et al. [24] extended this with Temporal Graph Networks for continuous-time dynamic graphs, precisely what we require to capture the temporal evolution of each researcher's taste. Song et al. [19] introduced DynGEM, emphasizing deep embedding preservation during graph evolution—important for maintaining consistency in taste representations as researchers age and shift focus.

Kipf and Welling [11] provided the foundational GCN architecture, while Veličković et al. [21] introduced attention mechanisms for graphs, enabling weighted heterogeneous interactions. Manessi and Rozza [12] advanced heterogeneous aggregation techniques, directly supporting our fusion of author-paper-field networks into a unified taste representation.

2.4 Evaluation Methodologies and Temporal Validation

The evaluation of AI systems for scientific tasks requires novel methodologies. Ye et al. [25] introduced the Proof of Time benchmark, which rigorously enforces temporal separation between training data (pre-2023) and evaluation targets (2024-2025 papers), preventing data contamination. This work deeply influences our experimental design and provides a gold-standard protocol for validating scientific prediction tasks.

For semantic similarity assessment, Devlin et al. [6] established BERT as a robust foundation for computing dense embeddings of scientific text, enabling precise measurement of semantic alignment between predicted and actual research proposals. Karpukhin et al. [10] advanced this with Dense Passage Retrieval, a technique we adapt for retrieving methodologically similar papers when evaluating methodological overlap.

2.5 Privacy, Ethics, and Responsible AI in Scientific Predictive Analytics

Our scale (10,000 living researchers) and predictive focus raise critical ethical concerns. Abuattieh et al. [2] provide a comprehensive framework for

pseudonymization in research contexts, directly informing our Reversible Pseudonymization Protocol. Shmueli [17] discusses the societal implications of predictive analytics applied to human behavior, cautioning against "locking-in" effects where external predictions subtly influence subjects' future choices.

Modern work on privacy-preserving machine learning and differential privacy in language models (though not single-cited here) forms the conceptual backdrop for our design of federated or isolated sandbox environments for taste learning and prediction.

2.6 Positioning AI Scholar: Key Distinctions

While our work draws substantially from these literatures, AI Scholar introduces several novel dimensions:

1. **Scale:** Simultaneously modeling 10,000 elite researchers, rather than analyzing individual trajectories or field-level aggregates.
2. **Personalization:** Conditioning generative ideation directly on learned taste embeddings, rather than treating taste as post-hoc evaluation criteria.
3. **Integrated pipeline:** Combining bibliometrics (trajectory extraction), graph learning (taste embedding), and generative LLMs (ideation) into a single, end-to-end framework.
4. **Rigorous temporal validation:** Enforcing strict data separation and evaluating against ground-truth future publications, not synthetic baselines.
5. **Ethical operationalization:** Embedding privacy and consent mechanisms from the outset, rather than retrofitting them post-hoc.

To move beyond the direct adoption of existing frameworks, we formulate an original methodology consisting of three core pillars: Taste-Driven Cohort Extraction, Scientific Taste Embedding (STE), and Trajectory-Conditioned Generative Ideation.

2.7 Taste-Driven Cohort Extraction at Scale

To map the highest echelon of scientific taste, we will construct a dataset of 10,000 leading researchers. We utilize the Stanford–Elsevier "Top 2% Scientists" database [9], which employs the composite c-score to normalize impact across 174 subfields, mitigating biases inherent in raw citation counts. The c-score accounts for subfield size, career stage, and nonlinear citation

distributions, providing robust comparability across diverse disciplines from particle physics to social psychology.

However, naively selecting the top 10,000 scientists risks biasing our cohort toward hierarchical laboratory structures (e.g., a dominant PI plus subordinate postdocs), which would confound our taste embeddings with institutional artifacts rather than individual cognition. To ensure the 10,000 selected scientists represent *distinct evolutionary branches of thought*, we introduce a **Graph-based Epistemic Disambiguation step**, drawing inspiration from community detection methodologies in complex networks [13, 5].

Step 1: Global Co-authorship Network Construction. We query the OpenAlex API [15] to construct a global co-authorship network $\mathcal{G} = (\mathcal{A}, \mathcal{E})$, where $\mathcal{A}$ is the set of authors and $\mathcal{E}$ represents co-authorships. Edge weights w_{ij} are computed as:

$$w_{ij} = \sum_{\text{paper } p} \delta(i \in p) \cdot \delta(j \in p) \cdot e^{-\lambda \Delta t_{ijp}} \tag{1}$$

where Δt_{ijp} is the temporal distance between the focal year and co-publication, weighted by decay factor λ to emphasize recent collaboration. This temporal weighting ensures we identify *active* collaborators, not merely historical ones.

Step 2: Leiden Community Detection. We apply the Leiden algorithm [5], a state-of-the-art method for discovering hierarchical community structures in weighted networks, to identify distinct academic "laboratories" or research groups. Leiden improves upon Louvain by maintaining coherence at all levels and preventing small-cluster fragmentation. Communities $\mathcal{L}_1, \mathcal{L}_2, \ldots, \mathcal{L}_K$ are formally identified as locally maximal partitions of $\mathcal{G}$.

Step 3: Principal Investigator Apex Selection. Within each detected cluster $\mathcal{L}_k$, we identify the apex PI using two complementary metrics:

1. **Eigenvector Centrality:** $\text{EC}_i = \frac{1}{\lambda} \sum_j w_{ij} \text{EC}_j$, capturing influence within the local collaboration subgraph.

2. **Composite c-score:** The researcher's standardized citation impact across their field.

We retain only the researcher with the highest combined rank across these two metrics, eliminating subordinate nodes (postdocs, graduate students). Formally, for cluster $\mathcal{L}_k$:

$$\text{PI}^* = \arg\max_{i \in \mathcal{L}_k} [\text{rank}(\text{EC}_i) + \text{rank}(c\text{-score}_i)] \tag{2}$$

This yields 10,000 highly independent elite researchers, whose bibliometric trajectories represent *individual* taste rather than laboratory consensus. The validation of this procedure involves manually auditing 200 randomly selected researchers to ensure that selected PIs are indeed recognized as independent leaders in their fields (target: ¿95% precision).

3 Scientific Taste Embedding (STE)

Once the longitudinal full-text corpora (acquired via OpenAlex and Semantic Scholar APIs and parsed via Vision-Transformers like Nougat [4]) are established for each scholar, we move to quantify "taste." We hypothesize that a scientist's taste can be represented as a trajectory in a *dynamic heterogeneous latent space*, explicitly capturing their evolving preferences across three orthogonal dimensions.

3.1 Three-Dimensional Taste Framework

For a scholar S, their publication history up to time T is a sequence of works $W_S = (w_1, w_2, \ldots, w_n)$, each represented as a rich feature vector including abstract, keywords, methodology tags, and field tags. We define the Scientific Taste Embedding (STE) to explicitly model:

1. **Methodological Inertia vs. Exploration:** Denoted $\theta_{\text{method}}(S,T) \in [0,1]$, capturing the rate at which the scholar abandons old techniques (e.g., linear regression) for emerging paradigms (e.g., Vision-Language Models). Operationally:

$$\theta_{\text{method}}(S,T) = 1 - \text{Jaccard}\left(\text{Methods}(w_{t-2}), \text{Methods}(w_t)\right) \tag{3}$$

   where $\text{Methods}(\cdot)$ extracts methodological keywords from a paper, and averaging over recent papers yields a smoothed trajectory.

2. **Problem Selection Heuristics:** Denoted $\theta_{\text{problem}}(S,T)$, capturing the semantic distance between the scholar's chosen problems and the mainstream consensus of their field at publication time. Formally:

$$\theta_{\text{problem}}(S,T) = \text{cosine}\left(\text{emb}(\text{abstract}_S), \text{centroid}(\text{emb}(\text{abstracts}_{\text{field}}))\right) \tag{4}$$

   where embeddings are derived from BERT [6]. High values indicate alignment with mainstream; low values indicate contrarian or emerging-frontier focus.

3. **Interdisciplinary Bridging:** Denoted $\theta_{\text{interdisc}}(S,T)$, measuring the propensity to introduce concepts from distal subfields into their primary domain:

$$\theta_{\text{interdisc}}(S,T) = \text{avg}_{w \in W_S} \text{Shannon}(\text{field distribution of cited papers}) \tag{5}$$

   Higher values indicate greater interdisciplinary reach.

3.2 Continuous-Time Dynamic Graph Neural Network (CT-DGNN) Formulation

To operationalize these dimensions into a unified embedding $\mathcal{H}_S(T) \in R^d$, we employ a Continuous-Time Dynamic Graph Neural Network (CT-DGNN), inspired by recent advances in temporal graph learning [14, 24].

The network operates on a heterogeneous graph $\mathcal{G}_{\text{het}} = (\mathcal{V}, \mathcal{E})$, where:

- $\mathcal{V} = \mathcal{A} \cup \mathcal{P} \cup \mathcal{F}$ comprises author nodes, paper nodes, and field nodes.
- $\mathcal{E}$ includes author–paper (writes), paper–field (belongs-to), author–author (co-authors), and paper–paper (cites) edges.
- Each edge has an associated timestamp t.

Graph updates proceed in discrete time windows, but the embedding function $\mathcal{H}_S(t)$ is interpolated continuously via neural ordinary differential equations (Neural ODEs), similar to [24]. For author node S at time T:

$$\mathcal{H}_S(T) = \Phi_{\text{aggregate}}\left(\text{GNN}_{\text{het}}\left(\mathcal{G}_{\text{het}}[0:T], S\right)\right) \oplus \text{Encode}(\theta_{\text{method}}, \theta_{\text{problem}}, \theta_{\text{interdisc}}) \tag{6}$$

where $\oplus$ denotes concatenation, and $\Phi_{\text{aggregate}}$ is a learned aggregation function that weights the contributions of paper nodes, co-authors, and field contexts according to their relevance.

3.3 Contrastive Learning Objective

To train the STE such that it truly captures individualized "style," we employ **contrastive learning**. The intuition is: papers by the same author should have similar taste embeddings, while papers by different authors in the same subfield should be distinguishable.

For a scholar S, we define the contrastive loss:

$$\mathcal{L}_{\text{cont}}(S) = -\log \frac{\exp(\text{sim}(\mathcal{H}_{S,t}, \mathcal{H}_{S,t+1}))/\tau}{\sum_{S' \neq S} \exp(\text{sim}(\mathcal{H}_{S,t}, \mathcal{H}_{S',t+1}))/\tau} \tag{7}$$

where $\text{sim}(\cdot,\cdot)$ is cosine similarity, τ is a temperature parameter, and S' ranges over other researchers in the same subfield. The loss encourages consecutive papers by S to remain proximate in embedding space, while maintaining separation from peers. Negative sampling uses hard negatives (researchers with similar c-scores, same field, but different trajectories) to maximize discriminative power.

Training proceeds via standard mini-batch SGD over 10,000 researchers' trajectories, using all papers published before the 2023 cutoff. Embedding dimension is set to $d = 512$, and we use a 4-layer heterogeneous GCN with attention heads as the underlying architecture [21].

3.4 Trajectory-Conditioned Generative Ideation

With the STE established, we redesign the autonomous ideation phase to move beyond generic field-level forecasting. Instead of a standard Retrieval-Augmented Generation (RAG) approach that queries the entire literature agnostically, we utilize **Trajectory-Conditioned Generation**—a novel framework that constrains the generation space to align with each researcher's learned taste profile.

Core Architecture. We deploy a *taste-aware multi-agent ecosystem*, extending the collaborative dynamics of frameworks like aiXiv [26] with taste-conditioning at every stage. The generation pipeline comprises three specialized agents, each with distinct roles and constraints:

1. **The Architect Agent** generates an initial pool of seed ideas by finding the intersection between (a) the current global technological frontier (SOTA), identified via recent arXiv preprints and arxiv-mined SOTA benchmarks, and (b) the specific historical trajectory encoded in $\mathcal{H}_S(T)$. Formally:

$$\text{SeedPool}_S = \{\text{idea} : \text{cosine}(\text{emb}(\text{idea}), \mathcal{H}_S(T)) > \theta_1\} \tag{8}$$

   where θ_1 is a tuned threshold (typically 0.6–0.7). The Architect Agent queries a literature database (papers 2023–2024) and performs multi-hop reasoning over method combinations that extend the scholar's prior work. We use an LLM prompt that explicitly mentions the scholar's recent papers (without revealing identity) to ground ideation: "Given a researcher who has recently worked on [Topic A, Topic B], what natural extensions combining [Emerging Method X] and their prior interests might they pursue?"

2. **The Taste-Critic Agent** filters these ideas not just for generic novelty or feasibility, but for *authorial alignment*. For each idea in SeedPool_S, the Taste-Critic computes:

$$\text{align}_{\text{score}}(\text{idea}, S) = \alpha \cdot \text{sim}(\text{idea}, \mathcal{H}_S(T)) + \beta \cdot \text{RiskProfile}(\text{idea}, S) + \gamma \cdot \text{MethodFit}(\text{idea}, S) \tag{9}$$

   where:

   - $\text{sim}(\cdot)$ captures semantic alignment (cosine distance in embedding space).
   - $\text{RiskProfile}(\text{idea}, S)$ evaluates whether the idea matches the scholar's historical epistemic risk tolerance—high-risk ideas are filtered if S has a conservative track record; conversely, conservative ideas are deprioritized for risk-seeking researchers.
   - $\text{MethodFit}(\text{idea}, S)$ assesses whether the required methodologies are within the scholar's known toolkit or represent a natural extension.

Ideas failing to meet the alignment threshold are discarded. This step prevents the absurd scenario wherein an AI system predicts that a classical theoretical computer scientist will suddenly pivot to wet-lab neurobiology.

3. **The Synthesis Agent** fleshes out the surviving ideas into highly rigorous, scaffolded research proposals. Starting from each surviving seed, it generates:

   - A 250-word research motivation specifically written from the perspective of a researcher with the scholar's taste.
   - Formal hypotheses, stated as testable predictions.
   - A detailed experimental design or theoretical roadmap, including proposed datasets, baselines, computational budgets.
   - A 3-5 year research roadmap showing how this initial direction could evolve, consistent with the scholar's trajectory.

   The Synthesis Agent is prompted with the full scholar context (history of methods, problems, collaborators) to ensure coherence and depth. Each proposal undergoes internal consistency checks (e.g., proposed methods actually exist; computational requirements are feasible given announced resources).

Prompt Engineering and Model Specifications. We employ a state-of-the-art LLM (GPT-4 or equivalent proprietary LLMs in the 2025 era) for both Architect and Synthesis agents, parameterized with learned system prompts that encode field-specific conventions. For the Taste-Critic, we employ both LLM-based scoring and learned neural modules (small regression heads trained on ground-truth papers by the scholar). This hybrid approach balances flexibility with explainability.

The generation temperature is set to $\tau = 0.8$ (higher than standard) to encourage diversity, while nucleus sampling ($p = 0.92$) filters low-probability tokens to maintain coherence. All generated proposals undergo deduplication to remove near-identical ideas.

Computational Efficiency. To scale to 10,000 researchers, we cache the STE embeddings $\mathcal{H}_S(T)$ and pre-compute SOTA literature embeddings once. At generation time, Architect and Synthesis agents operate in a streaming fashion, taking $\sim$ 30-60 seconds per researcher. Total inference time for 10,000 researchers is parallelizable and expected to complete in $\sim$ 2-4 days on a 100-GPU cluster.

4 Experimental Design: Time-Partitioned Evaluation

Validating the ability to predict a specific human's future research requires a rigorous, contamination-free evaluation protocol. We adopt and extend the methodological insights from the *Proof of Time* benchmark [25], which established that rigorous temporal separation is paramount in scientific prediction tasks.

4.1 Retrospective Backcasting Design and Temporal Cutoff

We employ a **Retrospective Backcasting** protocol to simulate real-world performance. All data used for developing embeddings, training neural networks, and conditioning prompts are drawn from publications *strictly before January 1, 2023*. The system is then tasked with predicting the research proposals that the 10,000 scholars *would* pursue in 2024 and 2025, without any access to papers published after the cutoff.

Formally, we partition the publication datasets recursively:

1. **Training Corpus:** All researcher publications, field trends, and methods known before 2023-01-01.
2. **Holdout Corpus:** All publications by the 10,000 focal researchers in 2024 and 2025, kept completely hidden during all model development, training, and inference.
3. **Validation/Test Split:** We withhold papers from 50 randomly selected researchers for internal validation (tuning thresholds, hyperparameters), and reserve the remaining 9,950 for final evaluation (to be reported in a companion full paper).

This separation prevents both data leakage and implicit contamination (e.g., via papers discussing predictions or future directions found in arxiv).

4.2 Evaluation Metrics and Formal Definitions

We evaluate along three complementary dimensions, each capturing a different aspect of taste prediction:

4.2.1 1. Semantic Trajectory Alignment (STA)

This metric measures the similarity between the generated proposal abstract and the abstracts of the scholar's *actual* papers published in 2024-2025.

For scholar S, let PropAbstract_S denote the concatenated abstract of our generated proposal(s), and let $\text{ActualAbstracts}_S = \{a_1, a_2, \ldots, a_k\}$ be the set of abstracts from their k actual papers in the holdout period.

$$\text{STA}(S) = \max_j \text{cosine}\left(\text{BERT-emb}(\text{PropAbstract}_S), \text{BERT-emb}(a_j)\right) \quad (10)$$

where BERT embeddings are computed using a fine-tuned BERT model trained on scientific abstracts [6]. The "max" operation reflects that even a single high-alignment predicted paper counts as success; partial credit is earned for near-matches. We report both the mean and median STA across all 10,000 researchers, and stratify by field and research seniority to detect potential systematic biases.

Interpretation: STA measures whether the predicted research direction matches the semantic character of the scholar's actual future work. A STA score > 0.70 is considered strong alignment.

4.2.2 2. Methodological Overlap Score (MOS)

Generated proposals are only useful if they are methodologically feasible for the target researcher. We extract the *explicit methods* mentioned in both the generated proposal and the scholar's actual holdout papers, then measure overlap.

Let Methods(text) denote a set of methodological keywords extracted via a learned NER model trained on scientific papers (e.g., "Vision Transformer," "molecular dynamics simulation," "causal inference"). Then:

$$\text{MOS}(S) = \frac{|\text{Methods}(\text{GeneratedProposal}_S) \cap \text{Methods}(\text{ActualPapers}_S)|}{|\text{Methods}(\text{GeneratedProposal}_S) \cup \text{Methods}(\text{ActualPapers}_S)|} \quad (11)$$

This is the Jaccard similarity of method sets. It can range from 0 (no overlap) to 1 (perfect match). We separately report precision and recall:

$$\text{MOS-Precision} = \frac{|\text{Predicted Methods} \cap \text{Actual Methods}|}{|\text{Predicted Methods}|} \quad (12)$$

$$\text{MOS-Recall} = \frac{|\text{Predicted Methods} \cap \text{Actual Methods}|}{|\text{Actual Methods}|} \quad (13)$$

High precision indicates the AI avoided proposing techniques outside the scholar's toolkit; high recall indicates the AI captured the main methods the scholar actually used.

4.2.3 3. Expert Turing Test

To supplement automated metrics, we conduct a blinded human evaluation. Domain experts (recruited from the same field as the focal researcher, but not collaborators) are presented with:

1. An anonymized abstract of one actual paper published by Scholar S in 2024-2025.
2. An anonymized abstract of one AI Scholar-generated proposal for Scholar S.

Experts are asked: "Which of these two did a human researcher actually write, and which was generated by an AI system trained on historical publications?" If evaluators can reliably distinguish the AI proposal from the human paper, the system has failed the Turing test. Conversely, if evaluators are confused or guess at random, the system has succeeded in generating research directions indistinguishable from human insight.

We recruit 300 expert evaluators, each assessing 30 anonymized pairs from diverse fields. Inter-rater agreement (Fleiss' kappa) and overall accuracy (fraction correctly identifying human vs. AI) are reported.

4.3 Cohort Stratification and Bias Analysis

To detect potential biases or domainspecific failures, we stratify results by:

- **Research Field:** Does the system perform better on computational fields (where SOTA is fast-moving) vs. theoretical fields (where breakthroughs are rarer)?
- **Research Seniority:** Do senior (30+ year career) researchers have more predictable tastes than mid-career researchers?
- **Interdisciplinarity:** Do researchers with high interdisciplinary bridges have taste profiles harder to predict?
- **Publication Velocity:** Do prolific researchers (¿50 papers from 2015-2023) have more stable, predictable taste profiles?

4.4 Baseline Comparisons

To contextualize our results, we compare against two baselines:

Baseline 1: Field-Level Forecasting. We use ResearchAgent [3] and NOVA [8] to generate ideas at the field level (without taste conditioning). The resulting pool of ideas is randomly sampled to match the number of proposals generated by AI Scholar. This baseline demonstrates the value of personalization.

Baseline 2: Trivial Extension. We extract the scholar's top recent papers and naively propose incremental extensions (e.g., extending dataset size, combining two prior methods). This baseline tests whether the value comes from sophisticated taste modeling or simply from exploiting obvious continuities in research.

We expect AI Scholar to substantially outperform both baselines, with large effect sizes.

4.5 Statistical Significance and Power Analysis

With 9,950 researchers, we are adequately powered to detect modest effect sizes ($d > 0.15$) with high confidence ($p < 0.001$). We employ one-sample t-tests and paired t-tests (comparing AI Scholar to baselines) with Bonferroni correction for multiple comparisons across fields. Confidence intervals (95%) are reported for all point estimates.

5 Ethical and Privacy Considerations

Scaling predictive profiling to 10,000 living scientists introduces substantial ethical complexities regarding privacy, consent, intellectual property, and the potential "locking-in" of research trajectories. A system that publicly predicts an elite scientist's next move creates multiple risks: (1) it violates the intellectual privacy of ongoing, unpublished laboratory work; (2) it may incentivize adversarial behavior by competitors; (3) it creates a permanent record of predictability that could be used manipulatively.

5.1 The Prediction Paradox and Reactivity

Shmueli [17] documents the "prediction paradox": when individuals are aware they are being predicted, their behavior often changes, either to conform to or defy the prediction. In academic contexts, publishing AI predictions of a researcher's future work risks subtly conditioning that researcher's choices. They may feel pressure to either (a) follow the system's suggestion, or (b) deliberately diverge to assert intellectual autonomy. Either outcome distorts the authenticity of the scientific process.

To mitigate this, we commit to: (1) *never* publishing individual researcher-level predictions before validating that they have consented; (2) publishing only aggregate, de-identified analyses at the field level; (3) conducting the evaluation study using historical data without the consent of studied researchers, which is permissible under IRB guidelines (45 CFR 46.104) as long as predictions are kept confidential.

5.2 Strict Reversible Pseudonymization Protocol

To prevent implicit contamination from the tastes and decisions of the AI itself, and to protect researcher privacy throughout the study, we implement a comprehensive **Strict Reversible Pseudonymization Protocol** [2].

Data Ingestion and Anonymization. During the data ingestion phase, all personally identifiable information is stripped:

- Author names → Cryptographic hash IDs (e.g., Scholar `X7-Alpha`).
- Institutional affiliations → Coarse-grained region codes (e.g., "North America", "Europe", "Asia").

- Unique identifying details (e.g., "the pioneering researcher who first applied X to Y in 1998") → Removed or abstracted to reduce re-identification risk.

Publication abstracts and metadata are retained in full, as these are already de facto public. The mapping key (Scholar `X7-Alpha` $\leftarrow\!\!\shortmid\!\!\rightarrow$ "Jane Smith") is held in a secure offline vault, accessible only to a designated ethics officer and lead researcher, and never provided to data scientists during model development.

LLM Sandbox Isolation. LLM agents (Architect, Taste-Critic, Synthesis) operate entirely within this decoupled sandbox. They receive pseudonymized publication histories, methodology tags, and embedding vectors, but *never* see real names, author photos, or institutional information. This design prevents LLMs from relying on parametric memory (i.e., "I recognize this author from my training data and know their current projects") and forces the system to rely purely on bibliometric latent structures.

Evaluation and Results Reporting. The offline mapping key is used *only* for the final automated metric evaluations—computing ground-truth alignment between predictions and actual papers, which requires knowing who wrote what. Reports are generated in two tiers:

1. **Publishing Tier:** Aggregate statistics only (mean STA across all 10,000 researchers; field-level stratified results). No individual researcher predictions are disclosed.

2. **Internal Research Tier:** The team may jointly review de-identified case studies (e.g., "Researcher X-42's predicted vs. actual trajectory showed 75% semantic alignment") for qualitative insights, but these remain internal.

5.3 Data Governance, Retention, and Right to Deletion

We commit to:

- **Minimal Data Collection:** Only publication metadata and full-text abstracts are collected; no personal communications, grant proposals, or unpublished manuscripts.

- **Ethical Review:** An Institutional Review Board (IRB) reviews the study design. Given that we predict on historical, public data without individual consent, the study likely qualifies as IRB-exempt; however, we proactively seek ethics board approval rather than claiming exemption.

- **Right to Deletion:** Any researcher may request deletion of their data from our training corpus (though retraining the full model would

be computationally expensive; new researchers can opt out of future studies entirely).

- **Data Retention:** Raw mapping keys are destroyed 5 years post-publication. Pseudonymized embeddings and model weights are retained indefinitely for reproducibility.

5.4 Dual-Use Concerns

A robust taste predictor could be misused:

1. **Competitive Intelligence:** Rival labs might use predictions to preempt competitors' research directions.
2. **Scooping:** Unethical researchers might rapidly publish findings predicted for a target scholar, claiming priority.
3. **Manipulation:** Journals or funding agencies might preferentially fund research matching AI predictions, creating self-fulfilling prophecies.

We mitigate these risks by:

- Publishing the system's results under open-science principles (full code, model weights, data splits released on GitHub), so researchers can inspect and critique the methodology rather than trusting a black box.
- Engaging with the scientific community transparently; soliciting feedback from research ethics committees and funders before deploying any consumer-facing prediction tool.
- Implementing usage terms specifying that AI Scholar predictions are for *research and educational purposes only*, not commercial forecasting.

6 Conclusion and Future Directions

The *AI Scholar* framework represents a paradigm shift in the intersection of bibliometrics, machine learning, and AI-driven scientific discovery. By transitioning from generic, domain-level idea generation to the fine-grained modeling of individual researcher "taste," we address a fundamental gap in current automation efforts.

6.1 Summary of Contributions

This work makes five key contributions:

1. **Theoretical Foundation:** We operationalize the abstract notion of "scientific taste" as a learnable, high-dimensional latent construct, grounded in recent philosophical and empirical work on aesthetic judgment in science.

2. **Methodological Innovation:** We introduce the Taste-Driven Cohort Extraction procedure, ensuring that 10,000 elite researchers are genuinely independent rather than clustered in hierarchical laboratory structures.

3. **Technical Advances:** The Scientific Taste Embedding (STE) leverages continuous-time dynamic graph neural networks to simultaneously model three dimensions of taste (methodological fluidity, problem-selection heuristics, interdisciplinary breadth), trained via contrastive learning.

4. **Taste-Conditioned Ideation:** We redesign the generative ideation pipeline to directly condition on learned taste embeddings, introducing a multi-agent framework (Architect, Taste-Critic, Synthesis) that mitigates the risk of predicting absurdly misaligned proposals.

5. **Rigorous Evaluation Protocol:** We establish a time-partitioned, retrospective evaluation design inspired by the Proof of Time benchmark, enabling validation against ground-truth future publications rather than synthetic benchmarks. Combined with semantic trajectory alignment metrics, methodological overlap scoring, and human Turing tests, this protocol provides unprecedented rigor in evaluating scientific prediction systems.

6. **Principled Ethical Framework:** We implement comprehensive privacy protections—strict pseudonymization, LLM sandbox isolation, and careful result reporting—ensuring that the system respects researcher autonomy and intellectual privacy while advancing scientific knowledge.

6.2 Implications and Vision

If successful, AI Scholar will yield several scientific insights:

- **Universality of Taste:** Are taste profiles predictable across fields, or are they discipline-specific? This illuminates whether scientific taste reflects universal cognitive principles or field-embedded conventions.

- **Scaling Laws of Predictability:** How does prediction accuracy scale with researcher seniority, field size, publication velocity, and interdisciplinarity? Such scaling laws could guide investment in foundational research forecasting.

- **Innovation Dynamics:** By analyzing which researchers successfully deviate from their predicted trajectories (and which follow them predictably), we can study the mechanisms of scientific innovation—do breakthroughs emerge from contrarians who defy their own taste, or from those who explore new frontiers within a coherent taste framework?

- **Augmentation, Not Replacement:** Ultimately, such a system should be deployed as a *research augmentation tool*. Young scientists could use AI Scholar to benchmark their emerging taste alongside peers; senior scientists could use it to identify emerging trends their taste naturally positions them to exploit; funding agencies could use field-level analyses to forecast where human creativity and capital should be invested.

6.3 Limitations and Future Work

We acknowledge several limitations that future work should address:

1. **Data Bias:** Our cohort of top 2% scientists is skewed toward well-established fields and Western institutions. Scaling to underrepresented research communities is essential.

2. **Temporal Generalization:** The 2023-2025 evaluation window is brief. Longer holdout periods (predicting 2030 from 2020 data) would reveal whether taste embeddings remain stable or drift on longer timescales.

3. **Causal Interpretation:** Our STE embeddings capture correlation, not causation. Are the learned dimensions causally driving choices, or are they spurious proxies for confounding factors (funding, institutional norms, field trends)?

4. **Interpretability:** While three-dimensional taste frames are interpretable, high-dimensional embeddings learned by neural networks remain opaque. Future work should invest in attention visualization and local linear approximations to make taste representations human-understandable.

5. **Interventional Validation:** Ideally, we would test predictions by informing researchers of AI forecasts and observing whether they resist, embrace, or ignore them. Such interventional studies raise complex ethical questions but could provide stronger evidence of predictive validity.

6.4 Closing Remarks

Scientific discovery has long been portrayed as the domain of individual genius and serendipity. Yet our work suggests that at least some fraction of great science emerges from coherent, learnable preferences—taste. By developing systems that capture this taste at scale, we do not diminish human creativity; rather, we illuminate its structure, making it a subject of rigorous, computational study. In doing so, we hope to unlock new partnerships between human intuition and machine intelligence, accelerating the pace of discovery while respecting the autonomy and dignity of the scientific community.

References

[1] Vittorio Acerbi. *The Semantics of Taste*. Oxford University Press, 2012.

[2] Hammam Abu Attieh, Armin Müller, Felix Nikolaus Wirth, and Fabian Prasser. Pseudonymization tools for medical research: a systematic review. *BMC Medical Informatics and Decision Making*, 25:128, 2025.

[3] Jinheon Baek, Sujay Kumar Jauhar, Silviu Cucerzan, and Sung Ju Hwang. Researchagent: Iterative research idea generation over scientific literature with large language models. arXiv preprint arXiv:2404.07738, 2024. `https://doi.org/10.48550/arXiv.2404.07738`.

[4] Lukas Blecher, Guillem Cucurull, Thomas Scialom, and Robert Stojnic. Nougat: Neural optical understanding for academic documents. arXiv preprint arXiv:2308.13418, 2023. `https://doi.org/10.48550/arXiv.2308.13418`.

[5] Kevin W. Boyack, Katy Börner, and Jan Klavans. Characterizing the evolution of scientific research fields and their leaders. *Scientometrics*, 114(3):921–936, 2018.

[6] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2019. `https://arxiv.org/abs/1810.04805`.

[7] Will Hamilton, Zhitao Ying, and Jure Leskovec. Inductive representation learning on large graphs. *Advances in Neural Information Processing Systems*, pages 1024–1034, 2017. NeurIPS 2017.

[8] Xiang Hu, Hongyu Fu, Jinge Wang, Yifeng Wang, Zhikun Li, Renjun Xu, Yu Lu, Yaochu Jin, Lili Pan, and Zhenzhong Lan. Nova: An

iterative planning and search approach to enhance novelty and diversity of llm generated ideas. arXiv preprint arXiv:2410.14255, 2024. https://doi.org/10.48550/arXiv.2410.14255.

[9] John P. A. Ioannidis, Kevin W. Boyack, and Jan Baas. Updated science-wide author databases of standardized citation indicators. *PLOS Biology*, 18(10):e3000918, 2020.

[10] Vladimir Karpukhin, Barlas Oğuz, Sewon Min, Patrick Lewis, Ledell Wu, et al. Dense passage retrieval for open-domain question answering. *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6469–6481, 2021.

[11] Thomas N. Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. *International Conference on Learning Representations*, pages 3837–3845, 2017. ICLR 2017.

[12] Florian Manessi and Alessio Rozza. Learning to rank graph collections by heterogeneous aggregation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(3):3284–3294, 2023.

[13] M. E. J. Newman. Coauthorship networks and patterns of scientific collaboration. *Proceedings of the National Academy of Sciences*, 101(15):5200–5205, 2004.

[14] Aldo Pareja, Giacomo Domeniconi, Johannes Garten, Charlotta Lindqvist, Rafeal Maltoni, et al. Evolvegcn: Evolving graph convolutional networks for dynamic graphs. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(4):5363–5370, 2020.

[15] Jason Priem, Heather Piwowar, and Richard Orr. Openalex: A fully-open index of scholarly works, authors, venues, institutions, and concepts. arXiv preprint arXiv:2205.01833, 2022. https://arxiv.org/abs/2205.01833.

[16] Andrey Rzhetsky, David E. Foster, Ian T. Foster, Shannon M. Galloway, and James A. Evans. Choosing experiments to accelerate collective discovery. *Proceedings of the National Academy of Sciences*, 112(49):14569–14574, 2015.

[17] Galit Shmueli. To explain or to predict? *Statistical Science*, 25(3):289–310, 2010.

[18] Karthik Sinha, Shen Wang, Michael Zhang, Robert Baldauf, and Danqi Chen. Autonomous experimental design for scientific discovery with large language models. *Nature Machine Intelligence*, 6:42–52, 2024.

[19] Letìtia Song, Yozen Liu, Xiaoming Sun, Abolfazl Asudeh, et al. Dyngem: Deep embedding method for dynamic graphs. arXiv preprint arXiv:1805.11273, 2021. https://arxiv.org/abs/1805.11273.

[20] Karthik Valmeekam, Matthew Marques, Alberto Abadie, and Siddhartha Srinivasa. On the planning and reasoning capabilities of large language models. *arXiv preprint arXiv:2401.14733*, 2024. https://arxiv.org/abs/2401.14733.

[21] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. Graph attention networks. *International Conference on Learning Representations*, pages 4170–4180, 2018. ICLR 2018.

[22] Mengdi Wang, Luoyi Fu, Luming Liang, and Guangyu Sun. Quantifying research novelty in academic papers. *Information Processing & Management*, 59(3):102923, 2022.

[23] Stefan Wuchty, Benjamin F. Jones, and Brian Uzzi. The increasing dominance of teams in production of knowledge. *Science*, 316(5827):1036–1039, 2007.

[24] Da Xu, Chuanwei Ruan, Evrenkörüprülü, Emre Kiciman, and Srinivasan Parthasarathy. Temporal graph networks for deep learning on dynamic graphs. arXiv preprint arXiv:2006.10637, 2023. https://arxiv.org/abs/2006.10637.

[25] Bingyang Ye, Shan Chen, Jingxuan Tu, Chen Liu, Zidi Xiong, Samuel Schmidgall, and Danielle S. Bitterman. Proof of time: A benchmark for evaluating scientific idea judgments. arXiv preprint arXiv:2601.07606, 2026. https://doi.org/10.48550/arXiv.2601.07606.

[26] Pengsong Zhang, Xiang Hu, Guowei Huang, Yang Qi, Heng Zhang, Xiuxu Li, Jiaxing Song, Jiabin Luo, Yijiang Li, Shuo Yin, Chengxiao Dai, Eric Hanchen Jiang, Xiaoyan Zhou, Zhenfei Yin, Boqin Yuan, Jing Dong, Guinan Su, Guanren Qiao, Haiming Tang, Anghong Du, Lili Pan, Zhenzhong Lan, and Xinyu Liu. aixiv: A next-generation open access ecosystem for scientific discovery generated by ai scientists. arXiv preprint arXiv:2508.15126, 2025. https://arxiv.org/abs/2508.15126.

[27] Pengsong Zhang, Heng Zhang, Huazhe Xu, Renjun Xu, Zhenting Wang, Cong Wang, Animesh Garg, Zhibin Li, Arash Ajoudani, and Xinyu Liu. Scaling laws in scientific discovery with ai and robot scientists. arXiv preprint arXiv:2503.22444, 2025. https://doi.org/10.48550/arXiv.2503.22444.