



March 2026

42DeepThought: Knowledge Graph-Based Assessment of Scientific Novelty, Methodological Rigor, and Impact

James O'Neill^{1,2}, Parag Katira², Heather Maughan³, Michael Martinez⁴, Brett Rexius⁵, Forest Rohwer⁶

¹California Medical Innovations Institute, San Diego, CA

²Department of Mechanical Engineering, San Diego State University, San Diego, CA

³National Coalition of Independent Scholars

⁴Pineshore Ventures

⁵Software Architect

⁶Department of Biology, San Diego State University, San Diego, CA

Introduction

The global volume of scientific literature has grown to several million new articles per year (**Figure 1**). Researchers struggle to stay abreast of developments in their own fields, let alone adjacent or emerging areas. At the same time, the primary quantitative signals used to evaluate scientific work—citation counts, h-indices, and journal impact factors—are lagging indicators. They reflect influence only after it has accumulated, often years after publication, and are at best indirect proxies for novelty, methodological rigor, or substantive impact. In this ecosystem, researchers find it increasingly difficult to position their intellectual contributions.

42DeepThought, developed by 42Degrees, is an attempt to provide an earlier, structurally grounded signal. Rather than waiting for citations to accumulate, 42DeepThought (1) analyzes how a document (idea, preprint, publication) is situated within the existing body of knowledge and (2) uses the document's structural position to estimate three dimensions of quality: novelty, methodological rigor, and potential impact. The goal is not to replace peer review or expert judgment, but to provide interpretable, uncertainty-aware signals, or *scores*, that can help researchers, editors, funders, and institutions navigate an increasingly dense research landscape.

This document provides an overview of the 42DeepThought methodology used to create the knowledge graph and estimate the scores. Goodhart's Law states that once a measure becomes a target, it is no longer a good measure. Thus, to prevent gaming of 42DeepThought scores and to maintain their viability as useful measures, 42Degrees is not releasing detailed methodological descriptions of how knowledge graphs are generated or how scores are calculated.

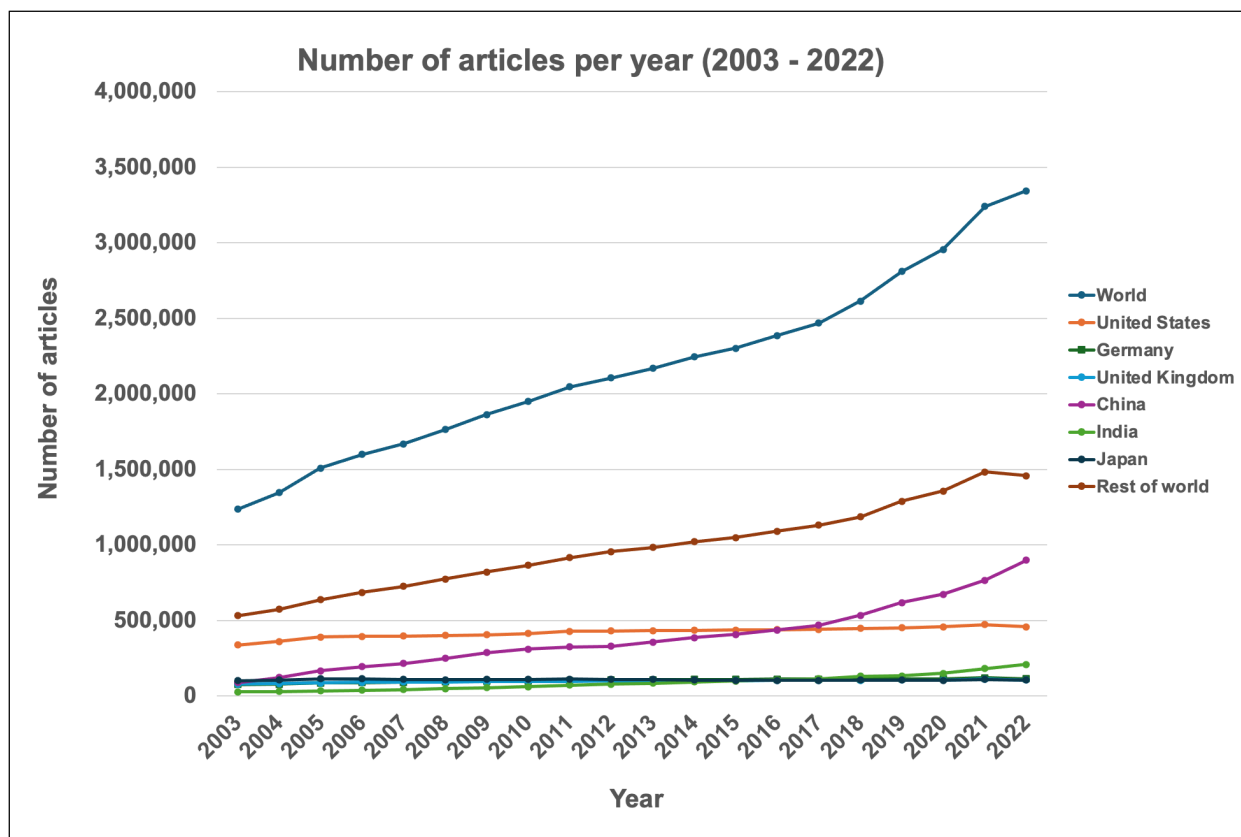


Figure 1. From Science & Engineering Indicators [replotted from (National Center for Science & Engineering Statistics (NCSES), 2023)]. “Article counts refer to publications from a selection of conference proceedings and peer-reviewed journals in S&E fields from Scopus. Articles are classified by their year of publication and are assigned to a region, country, or economy on the basis of the institutional address(es) of the author(s) listed in the article. Articles are credited on a fractional count basis (i.e., for articles produced by authors from different countries, each country receives fractional credit on the basis of the proportion of its participating authors).”

Overview of 42Degrees and 42DeepThought

42Degrees is building a platform for validated research intelligence. A core component of this platform is 42DeepThought, an application that (1) accepts research ideas, manuscript drafts, preprints, or published articles as input; (2) embeds them into a large, field-specific knowledge graph of the literature; (3) derives structural features and learned representations of their position in that graph; and (4) produces calibrated, ordinal scores for novelty, methodological rigor, and potential impact.

Scores are reported on a 1–99 scale within each dimension, grouped into Low (1–33), Medium (34–66), and High (67–99) categories. These scores are designed to support a first-pass triage of large document pools, facilitate the discovery of promising yet under-noticed research, and systematically map problems, methods, and impacts within a field.



Alongside scoring, the 42Degrees platform provides (1) a Similar Works feature that allows users to visualize and eventually analyze where a given work sits in relation to related problems, methods, and impacts and (2) a multimodal journal prototype for improved research translation and dissemination (fortytwodegrees.org).

The Knowledge Graph

At the core of 42DeepThought is a heterogeneous knowledge graph that represents the structure of a scientific field (**Figure 2A**). The graph includes four types of nodes (sub-themes are in development):

- **Publications** — Individual research articles or preprints.
- **Methods** — Techniques, assays, experimental protocols, or computational approaches.
- **Problems** — Research questions or domain-level challenges being addressed.
- **Impacts** — Downstream effects, findings, or applications reported by the work.

Edges in the graph capture semantic relationships:

- A publication *uses* particular methods.
- A publication *addresses* particular problems.
- A publication *produces* or is associated with particular impacts.

Publications are linked in the knowledge graph through shared problems, methods, and impacts. Two papers that use similar techniques to address related questions and report comparable outcomes will occupy similar neighborhoods in this graph. In contrast, work that combines disparate methods, crosses community boundaries, or connects previously unlinked problem/impact combinations will occupy more unusual structural positions.

From a network-science perspective, the graph encodes a structural prior over publications: it asserts that a written work's position in the research landscape carries information about the nature and quality of that work's contribution to the research landscape. 42DeepThought operationalizes and empirically tests this prior.

The knowledge graph is interactive; users can click on a node to see publications, and their authors, within the cluster comprising that node (**Figure 2B**). PubMed links provide a quick path to these related publications (see the Similar Works section below).



NETWORK DEPTH

+ - 182% 1° (21 nodes)

Figure 2A. A One-Degree heterogeneous knowledge graph that shows the Problems (orange diamonds), Methods (blue hexagons), and Impacts (green triangles) one degree (one edge) away from the focal written work (red star). The user can click on each node to see papers that are similar to the focal written work. Thus, within each node is a cluster of papers that are related by the problem they solve, the methods they use, or their impact. Features used to predict scores rely on the structures of these cluster and node relationships and their connecting nodes/clusters. Toggling Network Depth increases node degrees by two or three connections.

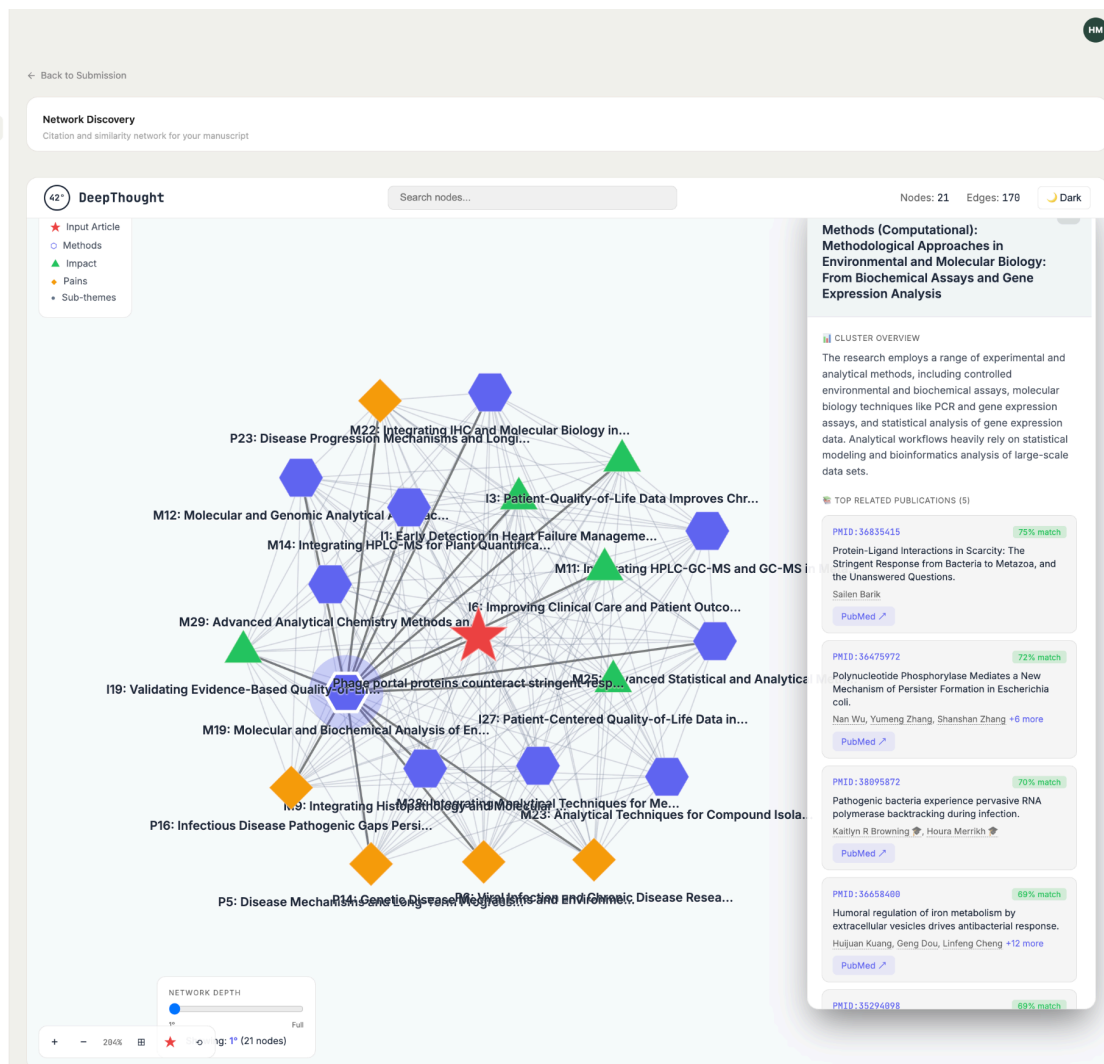


Figure 2B. Selecting a node results in a list of publications (with external PubMed links) within that node's cluster. The focal written work for this figure and Figure 3: (Kronborg et al., 2026).

From Graph Structure to Quality Scores

42DeepThought converts a paper's position in the knowledge graph into quality scores via a multi-stage pipeline comprising the following high-level steps.

Structural Feature Extraction

For each publication node, the system computes a suite of graph-theoretic features that together form a structural fingerprint. These features are interpretable in familiar graph-theoretic terms and show how a paper is structurally situated in the literature. These features are drawn from several broad families:

1. Centrality features capture different notions of "importance" or influence in the graph and include features such as connectivity to well-connected nodes, how central a publication



is within the field's network, and whether it lies on many shortest paths linking different regions of the graph.

2. Neighborhood and context statistics characterize the immediate and extended neighborhood of a publication, including how dense or sparse its local subgraph is, how similar or diverse its neighbors are in terms of degree, connectivity, and type, and whether it sits in a cohesive cluster or a more diffuse, interdisciplinary region.
3. Community-structure features describe how a publication fits into larger modules or subfields, such as the size of the community it belongs to and whether it is primarily a "local hub" within a subfield or a "connector" spanning multiple areas.
4. Bridging and structural-hole measures quantify whether a publication connects otherwise weakly linked parts of the graph, consistent with classical notions of structural holes and brokerage in network theory.
5. Type-composition ratios summarize the balance of connections to methods, problems, impacts, and other publications, distinguishing, for example, method-heavy from impact-heavy works.

Learned Graph Representations

The above features capture known, hypothesized relationships between structure and quality. However, there may be patterns of connectivity that are difficult to anticipate or specify manually. To complement the interpretable features, 42DeepThought also learns compact vector representations (embeddings) of each publication directly from the graph structure.

These embeddings are produced by a graph neural network trained in an unsupervised manner on the full knowledge graph. The training objective encourages connected or structurally similar nodes to have similar embeddings while pushing apart randomly paired nodes. Once trained, the model provides each publication with a learned representation that captures higher-order structural patterns beyond the feature set listed above.

The resulting publication representation is thus a combination of interpretable graph-theoretic features and a learned embedding that summarizes more complex patterns of connectivity.

Ordinal Quality Modeling

Using these representations, 42DeepThought estimates the probability that each publication falls into low, medium, or high categories for novelty, methodological rigor, and potential impact.

Crucially, the system treats these categories as ordered rather than as unrelated labels. This means the model explicitly encodes the assumption that "high" represents a higher level on the same underlying continuum as "medium" and "low," rather than a separate class.



March 2026

Internally, a small pool of candidate models with different inductive biases (e.g., favoring recall of high-quality papers versus conservative classification) is evaluated using resampling-based procedures. A final model is selected for each dimension based on its performance and stability under these perturbations.

The model outputs full probability distributions over the ordered categories. These probabilities are then transformed into a discrete category (Low, Medium, or High), and a continuous score in the 1–99 range within that category, calibrated to respect the ordering (e.g., strong “Medium” papers receive scores near the top of the Medium band).

Users can view their scores by clicking on the relevant submission within the Submissions tab (**Figure 3A**). Below the dashboard the user can find detailed feedback on the focal work, grouped by novelty, methodology, and impact with priority indicators to help organize efforts to improve the focal written work (**Figure 3B**).

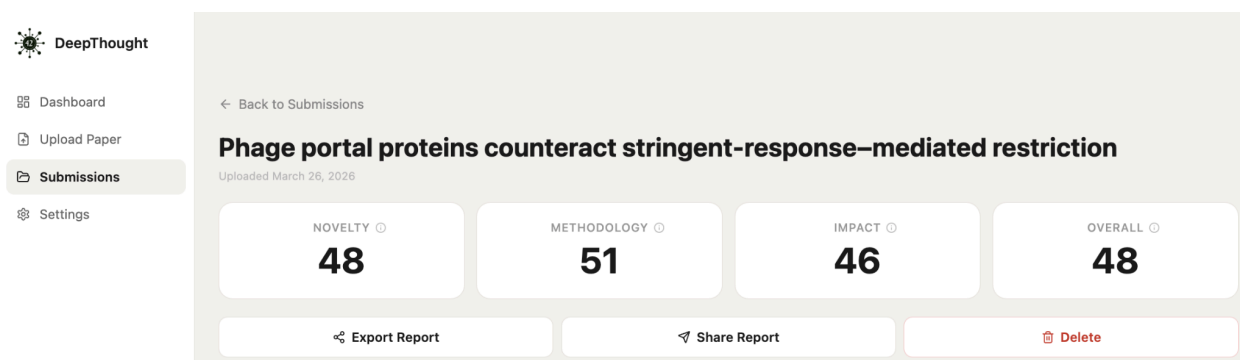


Figure 3A. Scoring dashboard showing Novelty, Methodology, and Impact scores along with overall score. Users can also export a report containing these scores, share the report (via email), or delete the submission. Note that 42DeepThought automatically deletes the uploaded written work after scoring - uploads are never stored.



DeepThought

Dashboard

Upload Paper

Submissions

Settings

Novelty Analysis1 item

Medium Priority

Claim of broad conservation without direct evidence from diverse phage species or structural data

"Portal proteins from diverse coliphages share similar structural features, interact with stringent response enzymes, and display SR-linked phenotypes, indicating a broadly conserved viral strategy."

Provide specific examples of phage species tested (e.g., lambda, phi80, SPP1) and include structural or functional data (e.g., binding assays, mutational analysis) for at least two non-T7 phages to support the claim of conserved interaction and phenotype.

Methodology Analysis2 items

Medium Priority

Lack of detailed characterization of phage concentration and viability post-PEG precipitation

"The phage stock was concentrated, and the buffer was exchanged following the protocol described here. Briefly, a 5 PEG solution (5 M NaCl, 20% (w/v) PEG800) was added to the cleared lysate and incubated overnight at 4C. The precipitated phages were pelleted (1 h, 4C, 19000 g), resuspended in 0.5-1 mL phage buffer (150 mM NaCl, 40 mM Tris, 10 mM MgSO4), and PFU was determined via plaque assay."

Add quantitative validation of phage titer after PEG precipitation (e.g., PFU/mL) and include a control for non-specific precipitation (e.g., PEG-only treatment). Specify whether the buffer exchange step affects phage stability or activity, and justify the choice of PEG concentration and incubation time.

High Priority

Reliance on figure legends for critical experimental conditions undermines reproducibility

"Growth and induction conditions are specified in the corresponding figure legends."

Explicitly list all growth and induction conditions in the main text, including temperature, shaking speed, media composition, and time points for each experiment. Avoid placing essential parameters in figure legends only.

Impact Analysis2 items

Medium Priority

Impact statement is broad and lacks concrete implications for phage therapy or bacterial pathogenesis

"Together, these findings identify phage portal proteins as a previously unrecognized class of viral counter-defense factors that target a central bacterial stress signaling pathway, revealing that essential structural virion components can moonlight as regulators of host stress physiology."

Refine the impact statement to highlight specific applications: "These findings suggest that phage portal proteins could be engineered as targeted inhibitors of bacterial stress responses, offering a new strategy for phage-based antimicrobials in clinical settings where bacterial stress pathways are exploited."

High Priority

Paper lacks explicit linkage to a high-impact prior work in phage-bacteria interactions or stress signaling

"Ensure at least one strong connection to a highly influential paper."

Explicitly cite and discuss a highly influential paper (e.g., one with >100 citations in phage biology or bacterial stress signaling) that established a foundational concept in the field, and show how the current work extends or challenges that concept.

Figure 3B: Along with numerical scores, written feedback is given with suggestions to improve the written work. Feedback is grouped by high, medium, and low priority.

Similar Works

In addition to scores, 42DeepThought provides a Similar Works feature that visualizes a focal written work's local neighborhood in the knowledge graph (**Figure 2, bottom panel**). A typical Similar Works view shows the focal work as a central node, surrounding nodes representing problems, methods, and impacts, with edges indicating one-hop relationships in the graph. Within each problem, method, or impact node, users can access clusters of papers that share that attribute with the focal work.

8



March 2026

This visualization allows researchers to (1) see which problems, methods, or impacts their work is structurally closest to; (2) identify closely related or partially overlapping lines of work; and (3) explore neighboring subfields that are connected via shared methods or problem formulations. The same structural patterns that feed into scoring—such as bridging across communities or sitting in dense methodological clusters—are thus made visually accessible to users, improving interpretability.

An additional page listing Similar Works facilitates in-depth investigation of similar publications. (**Figure 4**). 42Degrees is currently developing and optimizing additional features to maximize the utility of Similar Works so that users can search for clusters of related papers using author names, DOIs, or keywords. Comparative analysis of multiple publications, regardless of whether the submitted written work is included, will also be available.

The screenshot displays the 'Similar Works' page within the 42Degrees application. On the left is a sidebar with navigation links: Dashboard, Upload Paper, Submissions (highlighted), and Settings. The main content area has a header with 'Similar Works' and '79 papers show similarity'. Below this is a 'Section: All' dropdown menu, which is open, showing options: All (selected), Problems, Experimental, Impact, and Computational. To the right of the dropdown is a search bar. Below the search bar is a pagination control showing '1 2 ... 8' with 'Previous' and 'Next' buttons. The main list of similar works is shown below, with a table structure. The first column contains the title and author of the similar work, and the second column shows the '% OVERLAP' percentage. The list includes titles like 'Interactions in Scarcity: The Stringent Response from Bacteria to Metazoa, and the Unanswered Questions.' and 'From magic spot ppGpp to MESH1: Stringent response from bacteria to metazoa.' with authors like Sallen Barik and Jen-Tsan Chi, Pei Zhou. The overlap percentages range from 92% to 89%.

	% OVERLAP
Interactions in Scarcity: The Stringent Response from Bacteria to Metazoa, and the Unanswered Questions. ↗ Sallen Barik	92%
From magic spot ppGpp to MESH1: Stringent response from bacteria to metazoa. ↗ Problems Jen-Tsan Chi, Pei Zhou	92%
Microbial Primer: what is the stringent response and how does it allow bacteria to survive stress? ↗ Problems Lucy Urwin, Orestis Savva, Rebecca M Corrigan	90%
Physiology and Transcriptional Analysis of ppGpp-Related Regulatory Effects in Streptomyces diastatochromogenes 1628. ↗ Problems Yang Song, Xiangli Zhang, Zixuan Zhang, Xuping Shentu, Xiaoping Yu	89%
How to save a bacterial ribosome in times of stress. ↗ Problems Victor Zegarra, Patricia Bedrunka, Gert Bange, Laura Czech	89%
Adaptive strategies and ecological roles of phages in habitats under physicochemical stress. ↗ Problems Dan Huang, Rong Xia, Chengyi Chen, Jingqiu Liao, Linxing Chen, Dongsheng Wang, Pedro J J Alvarez, Pingfeng Yu	89%
Clinical rel mutations in Staphylococcus aureus prime pathogen expansion under nutrient stress. ↗ Experimental Edwin Chen, Marla G Shaffer, Robert E Blodeau, Raymond E West, Patrick J Oberly, Thomas D Nolin, Matthew J Culyba	89%

Figure 4. Similar Works feature within 42DeepThought. A list of publications identified as most similar to the submitted written work. The user can use the Category toggle to filter similar works.



42DeepThinkers

The 42DeepThinkers feature is like the Similar Works in that it uses the knowledge graph to identify authors of papers that are structurally similar to the focal work. When available, author contact information is included in order to promote collaboration (**Figure 5**). This feature is currently under development.

Your DeepThinkers						
Researchers whose work intersects with yours through the knowledge graph — potential collaborators, reviewers, and contacts.						
NAME	AFFILIATION	ROLE	ORCID	EMAIL	PAPERS	
Soumila Mondal	—	—	—	—	2	
Connected via 2 papers through clusters: M8: Research on nucleotide poly... M9: Research on nuttallii, copro...						
Computational studies on photolyase (Phr) proteins of cyanobacteria.						
83.5% 34587467 M9: Research on nuttallii, copro...						
Photoautotrophic black-colored cyanobacterial soil crust biosynthesizes photoprotective compounds and is capable of usin						
83.4% 36576619 M8: Research on nucleotide pol...						
Haidai Hu	NNF-CPR, University of Copenhagen	Researcher, Ph.D.	0000-0001-7774-9157	huhaidai@gmail.com (95%)	1	
Philipp F Popp	Humboldt-Universität zu Berlin	Postdoctoral Researcher	0000-0002-7825-1550	—	1	
Thomas C D Hughes	—	—	0009-0005-6057-4852	—	1	

Figure 5. DeepThinkers feature lists authors, and when available, their contact email address, of papers structurally related to the uploaded focal work in the knowledge graph.

Validation and Preliminary Performance

42DeepThought is calibrated and evaluated against a curated set of publications that have been manually assessed by domain experts along the three target dimensions. Each labeled paper is assigned an ordinal category (low, medium, high) for novelty, methodological rigor, and potential impact, based on a rubric that emphasizes:

- For novelty: conceptual or methodological distinctiveness relative to the field,
- For methodological rigor: appropriateness and robustness of study design, analysis, and reporting,
- For impact: potential significance of the findings or anticipated downstream influence.

Because the expert-labeled dataset is small, robust performance estimates were obtained using (1) cross-validation to assess how well models generalize to unseen labeled papers, and (2) bootstrap-style resampling to estimate variability in model performance under plausible changes to the training set.

Across these experiments, the models show high agreement with expert labels in distinguishing low, medium, and high categories across all three dimensions. These results are best



interpreted as preliminary but encouraging: further validation is ongoing as more labeled examples are collected and additional training sets are pursued.

In addition, 42Degrees is conducting “back-in-time” studies in which historically influential papers (e.g., those later associated with high citation counts, major awards, or key patents) are evaluated against only the literature that existed at the time of their original publication. These experiments probe whether 42DeepThought would have identified such work as methodologically strong, novel, or potentially impactful *before* its influence became apparent in traditional bibliometric indicators.

Uncertainty and Interpretation of Scores

No single number can fully capture the quality of a scientific contribution, and model predictions are inherently uncertain. 42DeepThought therefore provides not only point scores but also uncertainty information for each dimension.

42DeepThought assesses uncertainty using two approaches:

1. **Stability to data perturbations**
The model is re-trained and re-evaluated on slightly perturbed versions of the training data and input features. If a paper’s predicted scores remain stable across these perturbations, its score is considered more robust; if they vary widely, the score is considered less certain.
2. **Probability calibration**
The raw outputs of the classifier are adjusted so that predicted probabilities better match observed frequencies on held-out data (for example, ensuring that items assigned an 80% probability of being “high novelty” actually fall into the high-novelty category about 80% of the time).

In the user interface, these results are shown as:

- A primary score (e.g., Novelty = 74),
- An associated category (e.g., High novelty), and
- An uncertainty band or confidence interval indicating the range within which the score is expected to fall under reasonable perturbations.

Users should interpret narrow intervals as indicating higher model confidence and wide intervals as a signal to apply greater caution. *Note that even highly confident scores are intended as decision-support, not as definitive judgments.*



Scope, Assumptions, and Limitations

Like any model of scientific quality, 42DeepThought rests on explicit assumptions and has a defined scope of applicability.

Key Assumptions

1. Structural homophily. Papers occupying similar positions in the knowledge graph tend to share similar levels of novelty, rigor, and impact. This will likely hold more strongly for some subfields and dimensions than for others. 42Degrees is currently testing this hypothesis.
2. Label reliability. Expert labels, while imperfect and potentially heterogeneous across experts and fields, provide a sufficiently reliable signal for calibration. Ongoing tests with multiple training sets will address this assumption.
3. Feature sufficiency. The graph-structural representation (problems, methods, impacts, and their relationships) captures a substantial fraction of the quality-relevant information, even though it omits, for example, detailed textual content, fine-grained experimental details, and author-level factors.
4. Stationarity. The relationship between graph structure and quality learned from current data is assumed to generalize to new publications added to the graph. Rapid shifts in research practice or the emergence of new subfields can challenge this assumption.

Current Limitations

1. The labeled dataset used for calibration is intentionally small, reflecting the cost and complexity of expert annotation. As a result, performance estimates have nontrivial uncertainty, and rare edge cases (e.g., genuinely disruptive but initially isolated work) may be underrepresented.
2. The current models focus on static graph structure. Temporal dynamics—such as how quickly methods diffuse or how rapidly new problem–method–impact combinations emerge—are not yet explicitly modeled.
3. Some forms of scientific value, such as careful confirmatory studies or negative results, may be underemphasized in available labels and historical signals, and thus undercaptured by the model.

Ongoing work at 42Degrees includes expanding labeled training datasets, incorporating temporal information, exploring richer graph architectures, and using active learning to prioritize the most informative new examples for expert review.



Ethical Use and Recommended Practices

Given its potential role in research evaluation, the responsible use of 42DeepThought is a central design consideration. Scores are intended as one signal among many, to be interpreted alongside peer review, expert judgment, and domain-specific considerations.

42DeepThought is designed to support efforts such as (1) helping editors, preprint moderators, or funding panels prioritize and triage subsets of submissions for in-depth human examination; (2) enabling the discovery of research that may otherwise be overlooked, including papers that link different subfields or use novel combinations of methods; and (3) mapping one's own or others' research to the landscape comprising problems, methods, and impacts within a field to help inform decisions regarding most fruitful future research avenues, hiring and career advancement, or funding allocation.

42Degrees does not recommend using scores as the sole basis for manuscript accept/reject decisions or hiring and career advancement decisions. In addition small differences in scores (e.g., 71 vs 73) should not be considered meaningful, and applying thresholds based on scores (e.g., we only accept papers above 70 in novelty) are not advised without context.

42Degrees is committed to monitoring the platform for systematic biases, such as under-scoring work from particular regions, institutions, or subfields. Models and guidelines will be updated as new evidence emerges. Results from validation experiments will be published by 42Degrees, along with recommended practices as they evolve; the goal is to promote transparency and informed use.

Changing terminology and synonyms in science

Scientific language changes over time, and different communities often describe the same concept with different terms. This raises a natural concern: if the knowledge graph is built from semantic relationships in text, could it mistakenly treat the same biological property or concept as multiple separate things, and thereby distort scores?

42DeepThought is designed to reduce—but not completely eliminate—this problem in a few ways. It builds on a dedicated disambiguation layer (for synonym and entity resolution) and on modern semantic embeddings that group lexically different but conceptually similar phrases close together. Instead of relying on raw keywords, it uses context-aware representations and then looks at *relational patterns* (who a paper is connected to, and how) rather than just exact wording. That said, the system does not assume the graph is perfect or static: it treats the knowledge graph as an evolving, approximate model of the literature. As terminology shifts or new fields emerge, the graph and models must be refreshed and revalidated.

Practically, this means two things for users. First, within the domains and time window where the embeddings and disambiguation have been trained and checked, the network representation is “good enough” to support accurate prediction of methods, novelty, and impact scores. Second, the framework is explicit that it can drift if language or research landscapes shift; monitoring performance and periodically retraining on newer corpora are part of keeping the scores



trustworthy. In short, the method doesn't claim to perfectly encode "true" scientific concepts, but it does provide a robust, empirically validated approximation that works well in practice while remaining honest about its limits.

Caveat: a time-limited graph

42DeepThought measures novelty relative to the *graph it can see*, not the full history of science. If an older concept doesn't appear (or hardly appears) in the current 2020–2025 Virology & Microbiology semantic network, then a new paper that reintroduces it may occupy a sparse, unusual position in the current graph and be treated as structurally novel. In that sense, 42DeepThought is "local-in-time": it estimates how unusual a paper is in the *contemporary* landscape, not whether the idea was ever new in absolute historical terms.

This characteristic has upsides and downsides. The upside is that the model aligns with how experts often talk about novelty in practice—"novel relative to current work in this field"—and it is validated precisely on expert labels drawn from that recent corpus. The downside is that it can overstate novelty for rediscoveries or "forgotten" ideas if the relevant older literature is outside the graph. The framework acknowledges this as "distribution shift": it assumes that the relationship between graph structure and quality in the training period generalizes, and that missing parts of the literature (by time or domain) can bias structure-based inferences.

Thus, to summarize, the intended interpretation is novel within the modeled window and domains, not "never seen in the history of science." Expanding the corpus backward in time and periodically rebuilding the graph will reduce this artifact, especially for mature fields. For edge cases where historical novelty vs. contemporary rarity matters (e.g., historiographic work, rediscovery narratives), scores should be read alongside expert judgment and citation tracing rather than used in isolation.

Relation to Prior Work

Existing metrics that signal novelty or impact consider keywords (Arts et al., 2025; Luo et al., 2022), MeSH term combinations (Mishra & Torvik, 2016), unusual co-occurrences of topics or references (Bornmann et al., 2019; Shibayama et al., 2021; Shi & Evans, 2023), or shifts in semantic usage over time (He & Chen, 2018; Shi & Evans, 2023), usually implemented with PageRank variants, topic models, or LLM-based text models (Wu et al., 2025). Of note is DeSci Labs, which introduces new publishing workflows and a Surprisal Score, though it has not improved researchers' efficiency and suffers from fundamental limitations (Shi & Evans, 2023). DeSci Labs' impact and novelty measures rely on surface-level proxies—lists of cited journals, curated keywords, and ontologies like MeSH—that collapse true scientific context and miss genuinely new concepts. Their validation approach relies on correlation with citations, which invites Goodhart's Law (Fire & Guestrin, 2019) and appears to favor incremental, discipline-bound work over transformative, interdisciplinary advances. Finally, DeSci's reliance on just two graph metrics (novelty and bridging centrality) is too narrow to capture the multidimensional nature of scientific quality and impact.



In contrast, 42DeepThought treats each publication as a node in a large semantic-based knowledge graph and determines whether the structural position of that node within the graph, e.g., its centrality, bridging role, neighbor patterns, and learned graph embeddings, can predict expert-assessed novelty, methodological rigor, and impact. Thus, rather than rely on rare combinations of terms or citations in a paper, 42DeepThought infers novelty, methodological rigor, and impact from a paper's connections within the knowledge graph. 42DeepThought adds (1) rich graph-theoretic features, (2) a GNN-based embedding of position in the full network, and (3) an uncertainty-aware scoring pipeline calibrated on expert labels, yielding a probabilistic assessment of novelty rather than a purely term- or citation-combination score.

Summary

42DeepThought is a multi-stage system that uses a heterogeneous knowledge graph of a research field's scientific literature, interpretable graph features, and learned structural representations to estimate scores for novelty, methodological rigor, and potential impact. It emphasizes uncertainty quantification, reproducibility, and responsible use, recognizing both the promise and the limitations of automated assessment. As labeled datasets grow and the system is further validated across domains, 42Degrees will continue to refine both the underlying models and the guidance it provides to the scientific community on how best to interpret and apply these scores.

Disclosure

Artificial Intelligence (AI) tools were used to suggest revisions for clarity and flow and generate text based on internal 42Degrees documents. All content within this document has been edited by 42Degrees team members to ensure accuracy.

References Cited

- Arts, S., Melluso, N., & Veugelers, R. (2025). Beyond citations: Measuring novel scientific ideas and their impact in publication text. *The Review of Economics and Statistics*, 1–33.
- Bornmann, L., Tekles, A., Zhang, H. H., & Ye, F. Y. (2019). Do we measure novelty when we analyze unusual combinations of cited references? A validation study of bibliometric novelty indicators based on F1000Prime data. In *arXiv [cs.DL]*.
<https://doi.org/10.48550/ARXIV.1910.03233>
- Fire, M., & Guestrin, C. (2019). Over-optimization of academic publishing metrics: observing Goodhart's Law in action. *GigaScience*, 8(6). <https://doi.org/10.1093/gigascience/giz053>
- He, J., & Chen, C. (2018). Predictive effects of novelty measured by temporal embeddings on



the growth of scientific literature. *Frontiers in Research Metrics and Analytics*, 3.

<https://doi.org/10.3389/frma.2018.00009>

Kronborg, K., Wang, L., Schicketanz, M. L., Gerdes, K., & Zhang, Y. E. (2026). Phage portal proteins counteract stringent-response-mediated restriction. In *bioRxiv*. bioRxiv.

<https://doi.org/10.64898/2026.01.27.700999>

Luo, Z., Lu, W., He, J., & Wang, Y. (2022). Combination of research questions and methods: A new measurement of scientific novelty. *Journal of Informetrics*, 16(2), 101282.

Mishra, S., & Torvik, V. I. (2016). Quantifying conceptual novelty in the biomedical literature.

D-Lib Magazine: The Magazine of the Digital Library Forum, 22(9-10).

<https://doi.org/10.1045/september2016-mishra>

National Center for Science, & Engineering Statistics (NCSES). (2023). *Publications Output: U.S. Trends and International Comparisons* (No. NSB-2023-33).

<https://nces.nsf.gov/pubs/nsb202333/acknowledgments-and-citation>

Shibayama, S., Yin, D., & Matsumoto, K. (2021). Measuring novelty in science with word embedding. *PloS One*, 16(7), e0254034.

Shi, F., & Evans, J. (2023). Surprising combinations of research contents and contexts are related to impact and emerge with scientific outsiders from distant disciplines. *Nature Communications*, 14(1), 1641.

Wu, W., Zhang, C., & Zhao, Y. (2025). Automated novelty evaluation of academic paper: A collaborative approach integrating human and large language model knowledge. *Journal of the Association for Information Science and Technology*, 76(11), 1452–1469.