Generation and Human-Expert Evaluation of Interesting Research Ideas using Knowledge Graphs and Large Language Models

Xuemei Gu
Max Planck Institute for the Science of Light
Staudtstrasse 2, 91058 Erlangen, Germany
xuemei.gu@mpl.mpg.de

Mario Krenn
Max Planck Institute for the Science of Light
Staudtstrasse 2, 91058 Erlangen, Germany
mario.krenn@mpl.mpg.de

Abstract

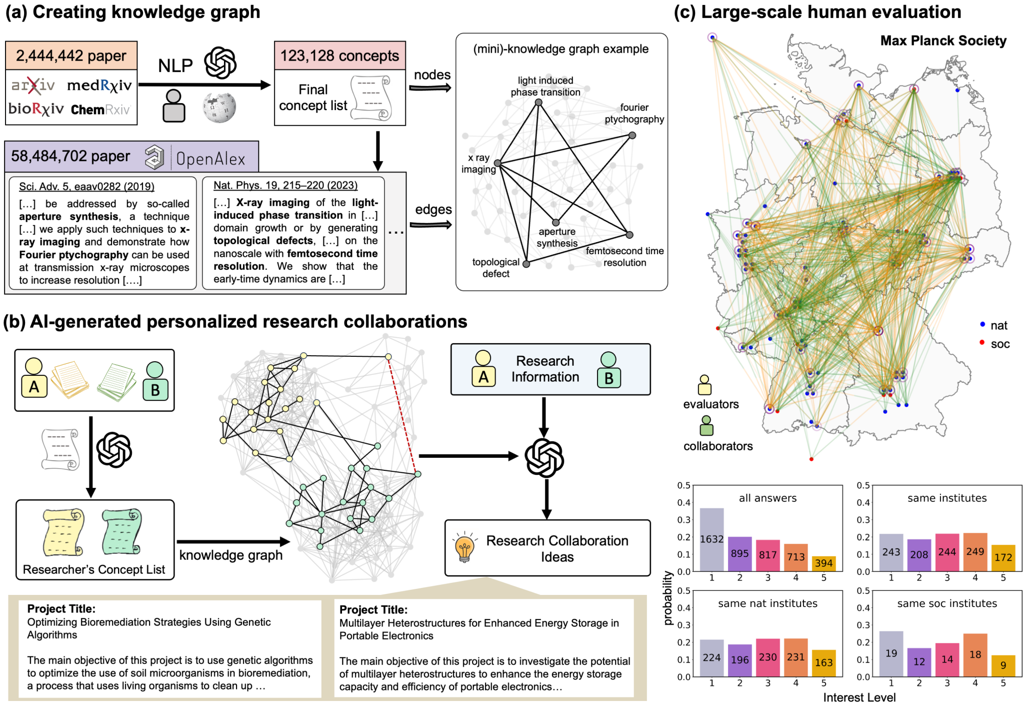
The rapid growth of scientific literature makes it challenging for researchers to identify novel and impactful ideas, especially across disciplines. Modern artificial intelligence (AI) systems offer new approaches, potentially inspiring ideas not conceived by humans alone. But how compelling are these AI-generated ideas, and how can we improve their quality? Here, we introduce SciMuse, which uses 58 million research papers and a large-language model to generate research ideas. We conduct a large-scale evaluation in which over 100 research group leaders – from natural sciences to humanities – ranked more than 4,400 personalized ideas based on their interest. This data allows us to predict research interest using (1) supervised neural networks trained on human evaluations, and (2) unsupervised zero-shot ranking with large-language models. Our results demonstrate how future systems can help generating compelling research ideas and foster unforeseen interdisciplinary collaborations.

1 Introduction

An interesting idea is often at the heart of successful research projects, crucial for their success and impact. However, with the accelerating growth in the number of scientific papers published each year [1, 2, 3], it becomes increasingly difficult for researchers to uncover novel and interesting ideas. This challenge is even more pronounced for those seeking interdisciplinary collaborations, who have to navigate an overwhelming volume of literature. Automated systems that extract insights from millions of scientific papers present a promising solution [4, 2, 5]. Recent advances have demonstrated that analyzing relationships between research topics across vast scientific literature can reliably predict future research directions [6, 7, 8, 9, 10], forecast the potential impact of emerging work [11, 12], and identify unconventional avenues for discovery [13]. With the advent of powerful large-language models (LLMs), it is now possible to leverage knowledge from millions of scientific papers to generate concrete research ideas [14, 15, 16].

Yet, a crucial question remains: Are AI-generated research ideas compelling to experienced scientists? Previous studies have only conducted small-scale evaluations with six natural language processing (NLP) PhD students [14], three social science PhD students [15] and ten PhD students in computer science and biomedicine [16]. However, perspectives from experienced researchers – who define and evaluate research projects through grant applications and shape their group's research agenda – are essential for assessing the value of new ideas. Involving a larger group of more experienced evaluators could offer deeper insights into what makes a research idea compelling, how to generate and predict them.

38th Conference on Neural Information Processing Systems (NeurIPS 2024).



Sci. Adv. 5, eaav0282 (2019)

[…] be addressed by so-called **aperture synthesis**, a technique […] we apply such techniques to **x-ray imaging** and demonstrate how **Fourier ptychography** can be used at transmission x-ray microscopes to increase resolution [….]

Nat. Phys. 19, 215–220 (2023)

[…] **X-ray imaging** of the **light-induced phase transition** in […] domain growth or by generating **topological defects**, […] on the nanoscale with **femtosecond time resolution**. We show that the early-time dynamics are […]

Project Title:
Optimizing Bioremediation Strategies Using Genetic Algorithms

The main objective of this project is to use genetic algorithms to optimize the use of soil microorganisms in bioremediation, a process that uses living organisms to clean up …

Project Title:
Multilayer Heterostructures for Enhanced Energy Storage in Portable Electronics

The main objective of this project is to investigate the potential of multilayer heterostructures to enhance the energy storage capacity and efficiency of portable electronics…

Figure 1: **Research ideas using knowledge graphs and GPT-4, with large-scale evaluation.** (a), Generation of a knowledge graph. Nodes are scientific concepts extracted from 2.44 million papers' titles and abstracts across four preprint servers, and edges indicate concept co-occurrences across over 58 million papers from OpenAlex [21], augmented with citation data as a proxy for impact. A mini-knowledge graph is shown for two random papers [22, 23]. (b), AI-generated research collaborations. Publications of two researchers are processed through a concept list, refined by GPT-4 to identify subnetworks in our knowledge graph that correspond to each researcher's interests. Relevant concept pairs are combined and fed into GPT-4, along with research information, to generate personalized research suggestions. (c), 4,451 AI-generated suggestions were rated by 110 research group leaders from the Max Planck Society in Germany, visualized as bi-colored edges on a graph, with edge transparency showing the number of evaluations and purple circles indicating intra-institute collaborations. Fields are marked by blue (natural sciences) or red (social sciences) dots. Suggestions were rated on a scale from 1 ('not interesting') to 5 ('very interesting'), and the summary shows the rating distribution. The Germany map is based on GISCO statistical unit dataset from Eurostat [24].

We introduce SciMuse, a system designed to suggest personalized research ideas or collaborations for scientists. To do so, we first create a knowledge graph from over 58 million papers, and identify sub-graphs related to researchers' interests. Concept pairs from the sub-graphs are combined with research information and fed into GPT-4, to generate personalized research suggestions. Then, we conducted a large-scale survey with over 100 research group leaders from the Max Planck Society, across diverse fields, who evaluated over 4,400 AI-generated ideas. We predict the level of interest of new ideas with two different methods: (1) training supervised neural networks and (2) using LLMs for zero-shot prediction without access to human evaluations, which will be important when expensive human-expert data is unavailable. This shows SciMuse's potential to suggest interesting research ideas, emphasizing AI's role as a source of inspiration in science [17, 18, 19, 20].

2 Research ideas using Knowledge Graphs and GPT-4

Generation of our knowledge graph While language models like GPT-4 [25], Gemini [26], and Claude [27] can directly suggest research ideas, our control over these suggestions would be limited to prompt structuring. To better identify personalized research interests, we built a large knowledge graph from scientific literature. This graph, shown in Fig. 1(a), has vertices for scientific concepts and edges for co-occurrences in paper titles or abstracts. Concepts were extracted from 2.44 million papers from arXiv, bioRxiv, ChemRxiv, and medRxiv (data cutoff: February 2023) using natural language processing (NLP) tools like RAKE [28], refined through customized NLP techniques, manual review, and GPT to remove non-conceptual phrases. Wikipedia was used to restore any mistakenly removed concepts. The final list included 123,128 concepts. Edges were drawn from over 58 million papers in the OpenAlex database [21], capturing concept co-occurrences and citation rates. This evolving graph, introduced in [12], spans scientific developments from 1665 (Robert Hooke's observation of a spot on Jupiter [29]) to April 2023 (see Fig. 1(a) and the Appendix A-Bfor details).

Personalized research suggestions We aim to generate personalized research proposals for collaborations between two scientists, with one evaluating the proposal later. As shown in Fig. 1(b), we first identify the research interests of Researchers A and B by analyzing their publications from the past two years. Concepts are extracted and refined with GPT-4 to create subgraphs for each researcher within the knowledge graph. Using these subgraphs, we prompt GPT-4 to create a research project, including up to seven paper titles from each researcher (details see the Appendix C-F). Concepts are selected randomly, by pairs with the highest predicted impact [12], or without specific pairs, rely solely on GPT-4. The prompt employs self-reflection [30], where GPT-4 generates three ideas, refines them twice, and selects the most suitable project idea as the final result.

3 Large-scale human evaluation

To assess how interesting these AI-generated ideas are, we asked research group leaders at scientific institutes, who regularly deal with and act upon research ideas, to participate in the evaluation. Specifically, 110 research group leaders from 54 Max Planck Institutes within the Max Planck Society in Germany participated (Fig. 1 (c)). They were tasked with evaluating up to 48 personalized research projects for their interest level, ranging from 1 ('not interesting') to 5 ('very interesting'). Of the 110 researchers, 104 are from natural science institutes, and 6 are from social science institutes. In total, we received 4,451 responses. Notably, 1,107 ideas received an interest level of 4 or 5 (nearly 25% of all suggestions), with 394 of these being ranked as *very interesting* (Fig. 1(c) and Appendix G).

4 Interest versus knowledge graph features

We found no significant difference in interest levels between research projects generated by random, high-impact, or no concept pairs, allowing us to analyze which knowledge graph features influence the project *interest*. Identifying these features can help us suggest ideas of higher interest in the future. We computed 144 features for used each concept pair (A, B) in the suggestions, including 141 from [12] (e.g., node characteristics like degree and PageRank [31], edge features like Simpson similarity, Sørensen–Dice coefficient [32], and impact-based features like citation rates) and three additional

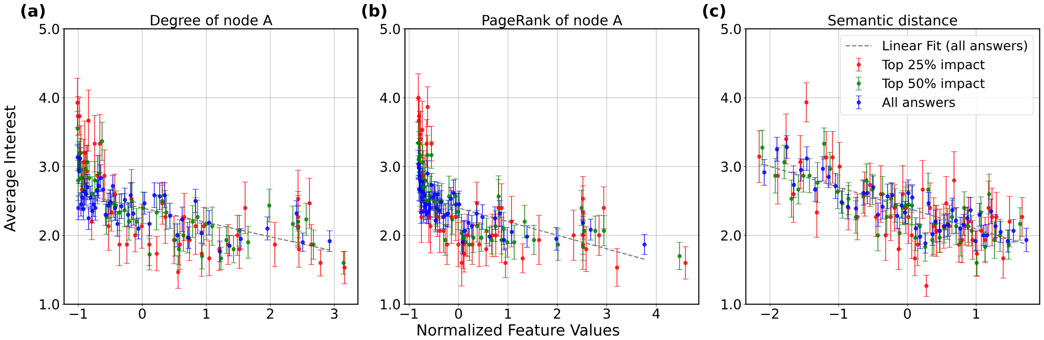


Figure 2: **Analysis of Interest Levels vs. Knowledge Graph Features.** Features (a) and (b) are node features, while (c) represents the semantic distance between the researchers' sub-networks (higher values indicate more distant fields). Data points are color-coded: blue for all 2,996 responses, green for the top 50% by predicted impact, and red for the top 25%.

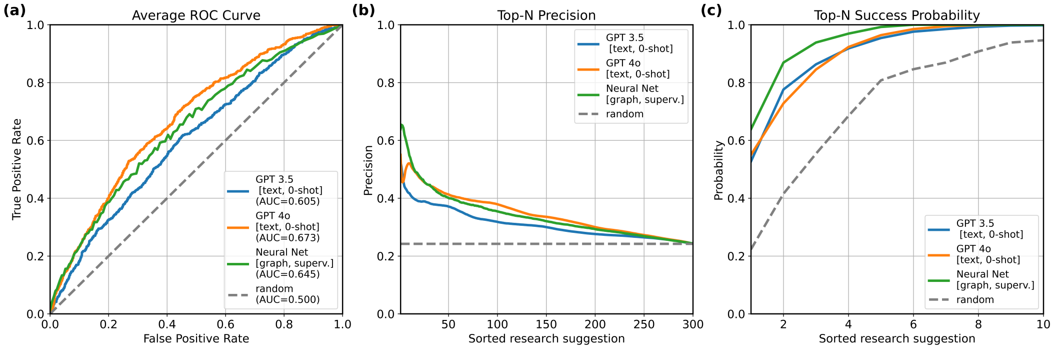


Figure 3: **Predicting scientific interest**. (a), The ROC curve shows prediction accuracy, and (b), demonstrates that both methods achieve higher precision for the top-N suggestions than random selection. (c) The probability of finding a high-interest suggestion among the top N.

features: predicted impact and two subgraph distance metrics (one based on subgraph distance and another on semantic distances in their extended neighborhoods). Features were normalized using z-scores, sorted them from lowest to highest, and divided into 50 groups. For each group, we plotted the average normalized feature value (x-axis) alongside the average interest value (y-axis) with standard deviations to observe trends in how graph features affect researcher's preferences (Fig. 2).

5 Predicting interest

We aimed to predict which research suggestions would receive high interest ratings (4 or 5 on a 5-point scale) or below 4 using two methods: a neural network trained on knowledge graph features linked to ratings and GPT in a zero-shot manner (without human evaluation feedback) to rank 2,996 suggestions. The neural network used 25 high-performing features in a small architecture (25 input neurons, 50 neurons in a hidden layer, and one output neuron, with dropout to enhance training [33]) and Monte Carlo cross-validation due to limited data (see Appendix H). In the second approach, GPT-3.5 and GPT-4o ranked suggestions by comparing pairs and updating rankings using an ELO system, starting with an initial score of 1400 (see Appendix I). Both methods achieved an average AUC of nearly 2/3 for the binary classification task (Fig. 3 (a)), indicating that concept pair selection alone effectively influences interest rankings while GPT's zero-shot ranking can be valuable when human evaluations are unavailable. The supervised approach reached 66.4% precision for the top-3 suggestions, while GPT-4o and GPT-3.5 achieved 45.0% and 47.2%, respectively, significantly outperforming the random selection rate of 23% (Fig. 3 (b)). Additionally, our methods showed a higher probability of finding at least one highly interesting suggestion within the top N compared to random sampling (Fig. 3 (c)).

6 Outlook

Our results show that it is possible to predict which research suggestions will interest scientists by analyzing the knowledge graph properties of concept pairs used in prompts to GPT-4, without relying on the text generated by GPT-4. This enables SciMuse to identify novel, high-interest research topics and translate them into full-fledged proposals using modern language models. As LLMs like GPT-4 [25], Gemini 1.5 [26], LLaMa3 [34], and Claude [27] continue to improve, the generated research ideas are expected to become more targeted and reasonable. The methodologies employed by SciMuse have the potential to inspire unexpected cross-disciplinary research on a large scale by providing a comprehensive analysis of millions of scientific papers. This approach facilitates the discovery of interesting collaborations between scientists from different domains, which could lead to impactful, award-winning results [35, 6, 1, 2]. Large scientific societies, national funding agencies, and other stakeholders might adopt such methodologies in the line of SciMuse to foster new interdisciplinary collaborations and ideas that might otherwise remain untapped, advancing scientific progress and impact on a broad scale.

Acknowledgements

The authors wholeheartedly thank all the researchers who spent the time participating in our study. The authors also thank the organizers of OpenAlex, arXiv, bioRxiv, medRxiv, and chemRxiv for making scientific resources freely accessible. X.G. acknowledges support from the Alexander von Humboldt Foundation.

Ethics Statement

The research was reviewed and approved by the Ethics Council of the Max Planck Society.

Data and code availability

Data for the knowledge graph is accessible on Zenodo at https://doi.org/10.5281/zenodo.13900962 [36]. Codes and evaluation data for this work are available on GitHub at https://github.com/artificial-scientist-lab/SciMuse.

References

[1] Santo Fortunato, Carl T Bergstrom, Katy Börner, James A Evans, Dirk Helbing, Staša Milojević, Alexander M Petersen, Filippo Radicchi, Roberta Sinatra, Brian Uzzi, et al. Science of science. *Science*, 359(6379):eaao0185, 2018.

[2] Dashun Wang and Albert-László Barabási. *The science of science*. Cambridge University Press, 2021.

[3] Lutz Bornmann, Robin Haunschild, and Rüdiger Mutz. Growth rates of modern science: a latent piecewise growth curve approach to model publication numbers from established and new literature databases. *Humanities and Social Sciences Communications*, 8(1):1–15, 2021.

[4] James A Evans and Jacob G Foster. Metaknowledge. *Science*, 331(6018):721–725, 2011.

[5] Zihang Lin, Yian Yin, Lu Liu, and Dashun Wang. Sciscinet: A large-scale open data lake for the science of science research. *Scientific Data*, 10(1):315, 2023.

[6] Andrey Rzhetsky, Jacob G Foster, Ian T Foster, and James A Evans. Choosing experiments to accelerate collective discovery. *Proc. Natl. Acad. Sci. USA*, 112(47):14569–14574, 2015.

[7] Mario Krenn and Anton Zeilinger. Predicting research trends with semantic and neural networks with an application in quantum physics. *Proc. Natl. Acad. Sci. USA*, 117(4):1910–1916, 2020.

[8] Justin Sybrandt, Ilya Tyagin, Michael Shtutman, and Ilya Safro. Agatha: automatic graph mining and transformer based hypothesis generation approach. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, pages 2757–2764, 2020.

[9] Rahul Nadkarni, David Wadden, Iz Beltagy, Noah A Smith, Hannaneh Hajishirzi, and Tom Hope. Scientific language models for biomedical knowledge base completion: an empirical study. *arXiv:2106.09700*, 2021.

[10] Mario Krenn, Lorenzo Buffoni, Bruno Coutinho, Sagi Eppel, Jacob Gates Foster, Andrew Gritsevskiy, Harlin Lee, Yichao Lu, João P Moutinho, Nima Sanjabi, et al. Forecasting the future of artificial intelligence with machine learning-based link prediction in an exponentially growing knowledge network. *Nature Machine Intelligence*, 5(11):1326–1335, 2023.

[11] Feng Shi and James Evans. Surprising combinations of research contents and contexts are related to impact and emerge with scientific outsiders from distant disciplines. *Nature Communications*, 14(1):1641, 2023.

[12] Xuemei Gu and Mario Krenn. Forecasting high-impact research topics via machine learning on evolving knowledge graphs. *arXiv:2402.08640*, 2024.

[13] Jamshid Sourati and James A Evans. Accelerating science with human-aware artificial intelligence. *Nature Human Behaviour*, 7(10):1682–1696, 2023.

[14] Qingyun Wang, Doug Downey, Heng Ji, and Tom Hope. Scimon: Scientific inspiration machines optimized for novelty. *Annual Meeting of the Association for Computational Linguistics*, 2024.

[15] Zonglin Yang, Xinya Du, Junxian Li, Jie Zheng, Soujanya Poria, and Erik Cambria. Large language models for automated open-domain scientific hypotheses discovery. *arXiv:2309.02726*, 2023.

[16] Jinheon Baek, Sujay Kumar Jauhar, Silviu Cucerzan, and Sung Ju Hwang. Researchagent: Iterative research idea generation over scientific literature with large language models. *arXiv:2404.07738*, 2024.

[17] Mario Krenn, Robert Pollice, Si Yue Guo, Matteo Aldeghi, Alba Cervera-Lierta, Pascal Friederich, Gabriel dos Passos Gomes, Florian Häse, Adrian Jinich, AkshatKumar Nigam, et al. On scientific understanding with artificial intelligence. *Nature Reviews Physics*, 4(12):761–769, 2022.

[18] Tom Hope, Doug Downey, Daniel S Weld, Oren Etzioni, and Eric Horvitz. A computational inflection for scientific discovery. *Communications of the ACM*, 66(8):62–73, 2023.

[19] Hanchen Wang, Tianfan Fu, Yuanqi Du, Wenhao Gao, Kexin Huang, Ziming Liu, Payal Chandak, Shengchao Liu, Peter Van Katwyk, Andreea Deac, et al. Scientific discovery in the age of artificial intelligence. *Nature*, 620(7972):47–60, 2023.

[20] Microsoft Research AI4Science and Microsoft Azure Quantum. The impact of large language models on scientific discovery: a preliminary study using gpt-4. *arXiv:2311.07361*, 2023.

[21] Jason Priem, Heather Piwowar, and Richard Orr. Openalex: A fully-open index of scholarly works, authors, venues, institutions, and concepts. *arXiv:2205.01833*, 2022.

[22] Klaus Wakonig, Ana Diaz, Anne Bonnin, Marco Stampanoni, Anna Bergamaschi, Johannes Ihli, Manuel Guizar-Sicairos, and Andreas Menzel. X-ray fourier ptychography. *Science advances*, 5(2):eaav0282, 2019.

[23] Allan S Johnson, Daniel Perez-Salinas, Khalid M Siddiqui, Sungwon Kim, Sungwook Choi, Klara Volckaert, Paulina E Majchrzak, Søren Ulstrup, Naman Agarwal, Kent Hallman, et al. Ultrafast x-ray imaging of the light-induced phase transition in vo2. *Nature Physics*, 19(2):215–220, 2023.

[24] European Commission. Eurostat gisco - nuts geodata, 2024.

[25] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv:2303.08774*, 2023.

[26] Machel Reid, Nikolay Savinov, Denis Teplyashin, Dmitry Lepikhin, Timothy Lillicrap, Jean-baptiste Alayrac, Radu Soricut, Angeliki Lazaridou, Orhan Firat, Julian Schrittwieser, et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv:2403.05530*, 2024.

[27] Anthropic. The claude 3 model family: Opus, sonnet, haiku. Papers with Code, 2024. Available at `https://paperswithcode.com/paper/the-claude-3-model-family-opus-sonnet-haiku`.

[28] Stuart Rose, Dave Engel, Nick Cramer, and Wendy Cowley. Automatic keyword extraction from individual documents. *Text mining: applications and theory*, pages 1–20, 2010.

[29] Robert Hooke. A spot in one of the belts of jupiter. *Philosophical Transactions of the Royal Society of London*, 1(1):3–3, 1665.

[30] Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, et al. Self-refine: Iterative refinement with self-feedback. *Advances in Neural Information Processing Systems*, 36, 2024.

[31] Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. The pagerank citation ranking : Bringing order to the web. *Stanford InfoLab*, 1999.

[32] Albert-László Barabási. *Network Science*. Cambridge University Press, 2016.

[33] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(56):1929–1958, 2014.

[34] Meta AI. Llama 3: Open foundation and fine-tuned chat models. `https://github.com/meta-llama/llama3`, 2024.

[35] Brian Uzzi, Satyam Mukherjee, Michael Stringer, and Ben Jones. Atypical combinations and scientific impact. *Science*, 342(6157):468–472, 2013.

[36] Xuemei Gu. Interesting scientific idea generation using knowledge graphs and llms: Evaluations with 100 research group leaders [data set]. zenodo. *https://doi.org/10.5281/zenodo.13900962*, 2024.

[37] Leo Breiman. *Classification and regression trees*. Routledge, 2017.

[38] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

A Datasets for creating knowledge graph

We compiled a list of scientific concepts using metadata from arXiv, bioRxiv, medRxiv, and chemRxiv. The arXiv data is available on Kaggle, while metadata for other preprint sources can be accessed through their respective APIs. Our dataset includes ∼2.44 million prapers, with a cutoff date of February 2023. For edge generation, we used the OpenAlex database snapshot (available on the OpenAlex bucket) with a cutoff date of April 2023. For more details, refer to the OpenAlex website [21]. The original dataset was filtered to entries of journal papers that contain titles, abstracts, and citation data, resulting roughly 92 million papers. From these 92 million papers, 58 million contain at least two concepts of our concept list and can therefore for an edge in the knowledge graph.

B Creating the concept list

We analyzed the titles and abstracts of ∼2.44 million papers from four preprint datasets using the RAKE algorithm, enhanced with additional stopwords, to extract potential concept candidates. Initial filtering retained two-word concepts (e.g. *gouy phase*) appearing in at least nine articles and concepts with more than three words (e.g. *recurrent neural network*) appearing in six or more, reducing the list to 726,439 concepts. To further improve the quality of the identified concepts, we developed a suite of automated tools to eliminate domain-independent errors commonly associated with RAKE and performed a manual review to remove inaccuracies like non-conceptual phrases, verbs, and conjunctions. This step refined the list to 368,825 concepts. Next, we used GPT-3.5 to further refine the concepts, which resulted in the removal of 286,311 concepts. We then employed Wikipedia to restore 40,614 mistakenly removed entries, resulting in a final refined list of 123,128 concepts.

C Prompt to GPT-4 for concept refinement

The prompt to refine the researchers' concept list is shown below:

> **Prompt to Refine the Researchers' Concept List**
>
> A scientist has written the following papers:
> 0) title1
> 1) title2
> 2) title3
> ...
>
> I have a noisy list of the researchers' topics of interest, and I would like your help in filtering them. Please look at the list below and return all concepts that are relevant to the scientist's research (based on their paper titles) and meaningful in the context of their research direction. The concepts can be detailed; I mainly want you to filter out concepts that are not meaningful, words that are not concepts, or concepts that are too general for the direction of the scientist (e.g., "artificial intelligence" might be a meaningful concept for a geologist, but not for a machine learning researcher). Do not change or add any concepts – only remove or keep them.
>
> concept list=[c1, c2, c3, c4, c5, c6, ...]

D Classification of Max Planck Institutes

We classified all 87 Max Planck Institutes into two categories: Class 1, abbreviated as *nat*, includes natural sciences, technology, mathematics, and medicine (68 institutes), while Class 2, abbreviated as *soc*, includes social sciences and humanities (19 institutes). The initial classification was done manually based on each institute's title and research field. To validate this, we further used GPT-4o for automatic classification, which perfectly matched with our manual classification.

E Researcher Statistics

Over 100 highly experienced researchers, spanning fields from the natural sciences to the humanities, participated in evaluating the personalized research ideas. Table. 1 summarizes the researchers' publication and citation statistics as of January 1, 2024, when the evaluations were conducted. On average, the researchers had published 59 papers and received over 3,750 citations.

Table 1: **Summary statistics of researchers' publications and citations.**

	Mean	**Median**	**Min**	**Max**
Number of papers	59.7	36.0	9	402
Number of citations	3759.7	1630.0	20	85778

F Prompt to GPT-4 for project idea generation

Prompt to GPT-4 for Project Idea Generation

Two researchers A and B, with expertise in “concept1" and “concept2" respectively, are eager to collaborate on a novel interdisciplinary project that leverages their unique strengths and creates synergy between their fields.

To better understand their backgrounds, here are the titles of recent publications from each researcher:
Researcher A:
1: title1
2: title2
3: title3
...
Researcher B:
1: title1
2: title2
3: title3
...

Please suggest a creative and surprising scientific project that combines “concept1" and “concept2". In your response, follow this outline:

First, explain “concept1" and “concept2" in one short sentence each.

Then, do the following three steps 3 times, improving in each time the response:
A) Describe 4 interesting and new scientific contexts, in which those two concepts might appear together in a natural and useful way.
B) Criticize the 4 contexts (one short sentence each), based on how well the contexts merge the idea of the two concepts.
C) Give a 2 sentence summary of your reflections above, on how well one can combine these concepts naturally and interestingly.

Then, start finding a project. Taking your reflections from (A-C) into account, define in your response a project title, followed by a brief explanation of the project's main objective.

Finally, address the following questions (Take the full reflections (A-C) into account):
What specific interesting research questions will this project address, that will lead to innovative novel results? [2 bullet points, one sentence each]

Rather than relying on a knowledge graph to supply “concept1" and “concept2", it is also possible to direct GPT-4 to extract these concepts from the research paper titles of Researchers A and B, respectively. GPT-4 can then use these identified concepts within the same prompting context to generate innovative research ideas.

G Interest evaluation for three different generation methods

Fig. 4 presents the interest-level distributions for research suggestions generated using three different methods. The interest levels are notably similar between suggestions generated with and without concepts from the knowledge graph. This similarity enables us to analyze the correlations between knowledge graph properties and interest levels, and to use these properties for predicting the interest level of generated research proposals.

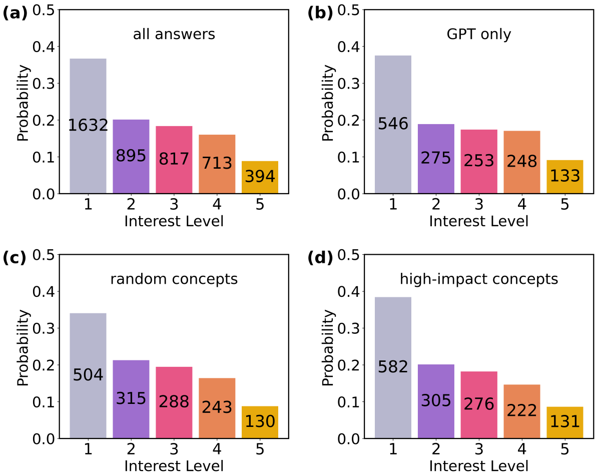


Figure 4: **Interest levels across different generation methods**. Research ideas are generated using three methods: (1) no concepts provided by the knowledge graph, (2) random concepts from the researchers' subnetwork, and (3) predicted high-impact concept pairs from the researchers' subnetwork. The figures displays: **(a)** overall interest levels (numbers within bars show the number of responses for that evaluation), **(b)** interest levels for ideas without using concepts from the knowledge graph, **(c)** interest levels with random concept pairs, and **(d)** interest levels using high-impact concept pairs (predicted by adapting the computational methods from [12], and applying them to a different and much larger knowledge graph).

H Predicting high interest from knowledge graph features

In Fig. 3 (main text), we aim to predict whether a research proposal will be rated with high interest. Specifically, using only data from the knowledge graph (excluding the final text generated by GPT), we predict if a proposal will receive an interest rating of 4 or 5 (on a scale of 1 to 5: *not interesting* to *very interesting*) or below 4. This is formulated as a binary classification task.

The input to the neural networks consists of network-theoretical features extracted from the knowledge graph. For each concept pair in a research project, we compute 144 features. The first 141 features are derived from those used to predict the future impact of concept pairs, as described in [12]. These features include node properties (e.g., node degree and PageRank [31]) and edge properties (e.g., Simpson similarity and Sørensen–Dice coefficient [32]). Several features also account for impact information, such as recent citation counts. The remaining three features include the predicted impact and two distance metrics between the researchers' subgraphs (Fig. 1(b)). The first distance metric measures the distance based solely on the concepts present in Researcher A and Researcher B's concept lists. In contrast, the second metric takes into account the entire neighborhood of these subgraphs by calculating semantic distances between all neighboring concepts and the concepts from the subgraphs. These features serve as the input to the neural network for predicting whether a proposal will achieve a high interest rating.

Given the small dataset size (2,996 answers with properties from the knowledge graph), we use a data-efficient learning method – a small neural network with dropout. The input layer consists of the 25 best-performing features (see Table. 2), selected from the total 144 by independently analyzing

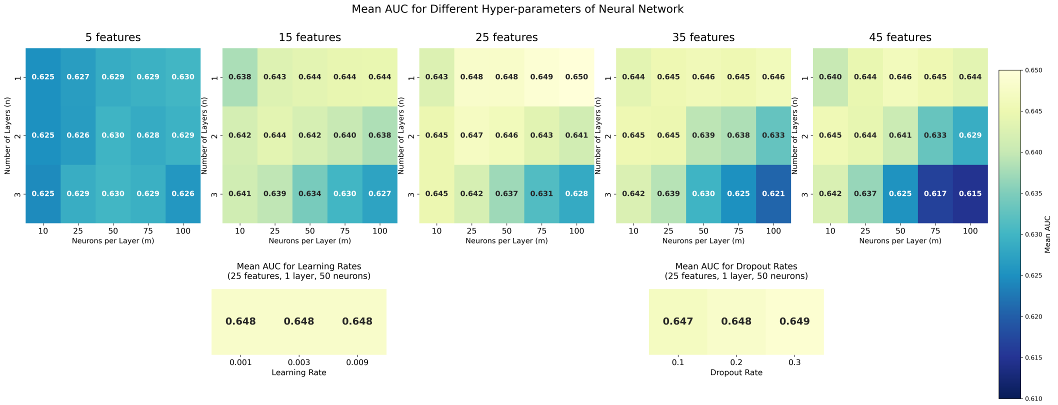


Figure 5: **Choice of alternative hyper-parameters for training of neural network.** We analyse the prediction of interest level quality (in terms of AUC) with different parameters of the neural network, such as different number of features, different number of layers and neurons, learning rate and drop-out rate. We see that the final results are robust under variations of the hyper-parameters.

the feature importance of each and choosing the top 25. The neural network has one hidden layer with 50 neurons and a single output neuron. Mean square error is used as the loss function.

To ensure robust performance estimation for the small dataset, we use Monte Carlo cross-validation. The dataset is repeatedly split into training and validation sets, and the model is trained and evaluated on each split. This approach ensures that the performance metrics are robust and not dependent on a particular split of the data. This iterative process continues until the standard deviation of the mean AUC is less than $\frac{10^{-2}}{3}$, achieved after 130 iterations. This method provides a reliable estimate of the model's performance, which is crucial for small datasets where individual splits may lead to high variance in the evaluation metrics.

The neural network performance is not specifically sensitive to hyperparameter choices, thus we refrained from hyperparameter optimization, and instead used a reasonable defaults: learning rate=0.003, dropout=20%, weight decay=0.0007, training dataset=75%, validation dataset=15%, test dataset=10%. In Fig.5, we investigate alternative hyper-parameters of the training process, and find that the results are robust under variations of the hyper-parameters.

I Zero-shot ranking of research suggestions by GPT

We ranked 2,996 research suggestions – previously evaluated by human experts – using GPT-3.5, GPT-4o, and GPT-4o-mini. For each pair of randomly selected suggestions, we asked the LLMs to rank which one was more interesting, considering the personalized research interests of the evaluating human expert. This pairwise comparison was repeated between 22,000 and 45,000 times (for GPT4o and GPT4o-mini, respectively). We treated this task as a tournament where all 2,996 suggestions compete pairwise against each other. Using the ELO ranking system, each suggestion started with an initial ELO score of 1400. Each comparison by GPT updated the ELO rankings based on the outcome, producing a final sorted list of suggestions from highest to lowest ELO score. We evaluated the ranking quality by calculating the AUC to determine how well the ranked list aligns with the human-expert evaluations of interest levels (Fig.6).

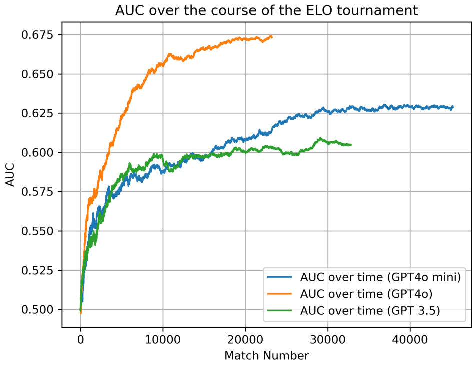


Figure 6: **Zero-Shot ranking of research suggestions by LLMs.** The research suggestions are generated using the knowledge graph together with GPT4. They are then ranked using GPT4o, GPT4o-mini and GPT3.5, without feedback from the human evaluation. The human evaluation is used to compute the final quality of the ranking. the ranking is performed in a pair-wise choice where we ask the LLM to select the more interesting one given the research background of the researchers. One match is one pairwise selection. The LLMs perform 10,000 of these pairwise selections, which allows us to compute ELO scores for each generated research idea.

Prompt for Zero-Shot Ranking of Research Ideas

I will present two research ideas. The first idea is for Researchers A1 and B1, and the second idea is for Researchers A2 and B2.
Researchers A1 and A2 will evaluate how interesting they find the respective ideas.
You will determine which of the two suggestions will be considered more interesting.

The suggestions are randomly ordered, and you should evaluate each suggestion independently and without bias.

Researcher A1 Context and Suggestion 1:
Here are a few papers of Researcher A1:
(papersA1)

Suggestion 1: [suggestion1]

*Summary for Researcher A1**: Provide a one-sentence summary of Suggestion 1 in the context of Researcher A1.

Researcher A2 Context and Suggestion 2:
Here are a few papers of Researcher A2:
(papersA2)

Suggestion 2: [suggestion2]

*Summary for Researcher A2**: Provide a one-sentence summary of Suggestion 2 in the context of Researcher A2.

Evaluation:
Based on the summaries and the research interests of A1 and A2, evaluate which suggestion is more likely to be ranked higher in terms of interest.
*Result**: If Suggestion 1 is ranked higher by Researcher A1 than Suggestion 2 is by Researcher A2, write 'RESULT: SUGGESTION 1'. Otherwise, write 'RESULT: SUGGESTION 2'.
Remember, the suggestions are randomly ordered, and your evaluation should be impartial and based solely on the research interests of A1 and A2.

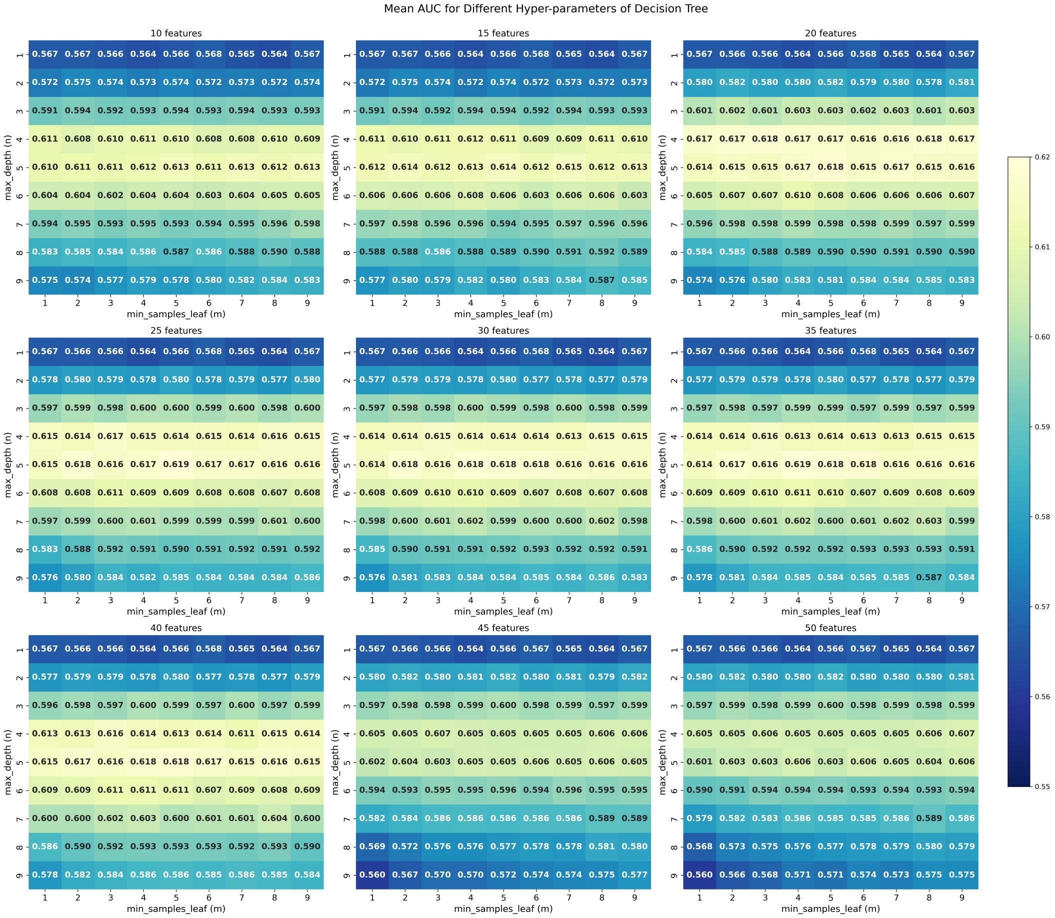


Figure 7: **Choice of hyper-parameters for training of decision tree.** The model is trained using Monte Carlo cross-validation until a statistical uncertainty of σ=0.001 is reached. We find that no setting of number of features, maximum depth and minimal sample leaf can reach the performance of the data-efficient neural network.

J Predicting high interest from knowledge graph features with decision trees

We experimented with other data-efficient learning methods, specifically with decision trees [37] using[38]. However, decision trees did not outperform the neural network predictions, as can be seen in Fig.7. These values confirm that neural networks are advantageous for the task of ranking research ideas by their interest value in a supervised way, which can also be confirmed in Fig.8.

K Prediction of Interest with different methods

We show the full data of all five methods (supervised training with neural networks and decision trees, as well as unsupervised zero-shot prediction with GPT3.5, GPT4o and GPT4o-mini), with their corresponding AUC, top-N precision and top-N success probability, in Fig.8. We see that the neural network outperforms decision trees when trained in a supervised way, and that GPT4o is better than the other tested models, when the ranking is performed in a zero-shot manner without giving any information about the evaluations of humans.

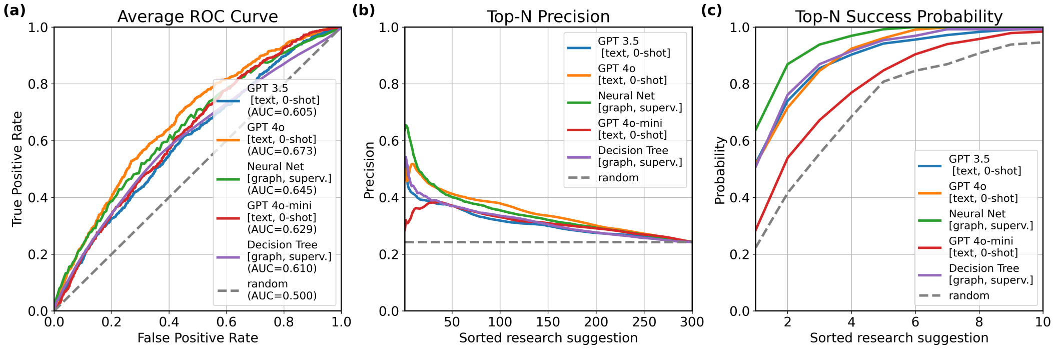


Figure 8: **Interest predictions with five methods.** We reproduce Fig.3 (main text), and add results from a supervised decision tree training, and an unsupervised GPT4o-mini.

L Prompt engineering

We have explored manual and automated improvements of the prompts, both for the research question design and the zero-shot prediction. Specifically, we attempted to improve the prompts for the idea generation using GPT-4o. While the prompts were more structured, a small-scale evaluation did not show any improvement in terms of more interesting results.

M GPT-4o and GPT-o1 for idea generation

We conducted two small-scale tests where GPT-4 and GPT-4o generated ideas using the exact same settings described above.

In the first test, three research group leaders evaluated 180 pairs of questions (one generated by GPT-4 and the other by GPT-4o using the same prompt). They found 31.1% of GPT-4 answers to be more interesting, while 60.56% favored GPT-4o answers (8.3% were draw). In the second test, a research group leader evaluated 11 pairs of questions (one generated by GPT-4o and the other by GPT-o1 with the same prompt), and all 11 ideas generated by GPT-o1 were ranked as more interesting.

These additional small-scale tests suggest that improved models can enhance idea generation, thus directly improving the results of SciMuse.

Table 2: **Top-25 Best-Performing Features of Each Concept Pair (c_A, c_B) to the Neural Network**

	Feature
1	Semantic distance between Researchers A and B (using all neighboring concepts and all concepts from the subgraphs)
2	Number of new neighbors gained by c_A from the years 2022 to 2023
3	Rank of the number of new citations for c_A from the years 2022 to 2023
4	Rank of the number of new papers mentioning c_A from the years 2021 to 2023
5	Number of papers mentioning either concept c_A or c_B until the year 2022
6	Annual citations for c_A during the year 2020
7	Total citations for c_A from its first publication until the year 2021
8	PageRank score for c_B until the year 2023
9	Number of neighbours for c_A until the year 2022
10	Number of new papers mentioning c_A from the years 2021 to 2023
11	Rank of the number of new neighbors gained by c_A from the years 2021 to 2023
12	Total citations for c_A from the years 2020 to 2023
13	PageRank score for c_B until the year 2022
14	Rank of the number of new neighbors for c_A from the years 2022 to 2023
15	Annual citations for c_A during the year 2022
16	Total citations for c_A from the years 2019 to 2022
17	Number of neighbours for c_A until the year 2023
18	Number of neighbours for c_A until the year 2021
19	Number of new neighbors gained by c_B from the years 2022 to 2023
20	PageRank score for c_A until the year 2023
21	Total citations for c_A from its first publication until the year 2023
22	PageRank score for c_A until the year 2022
23	Number of papers mentioning either concept c_A or c_B until the year 2023
24	Number of neighbours for c_B until the year 2021
25	Total citations for c_A from its first publication until the year 2022