Generation and human-expert evaluation of interesting research ideas using knowledge graphs and large language models

Xuemei Gu [1] **Mario Krenn** [1]

Abstract

Advanced artificial intelligence (AI) systems with access to millions of research papers could inspire new research ideas that may not be conceived by humans alone. However, how interesting are these AI-generated ideas, and how can we improve their quality? Here, we introduce SCIMUSE, a system that uses an evolving knowledge graph built from more than 58 million scientific papers to generate personalized research ideas via an interface to GPT-4. We conducted a large-scale human evaluation with over 100 research group leaders from the Max Planck Society, who ranked more than 4,000 personalized research ideas based on their level of interest. This evaluation allows us to understand the relationships between scientific interest and the core properties of the knowledge graph. We find that data-efficient machine learning can predict research interest with high precision, allowing us to optimize the interest-level of generated research ideas. This work represents a step towards an artificial scientific muse that could catalyze unforeseen collaborations and suggest interesting avenues for scientists.

1. Introduction

A compelling idea is often at the heart of successful research projects, crucial for their success and impact. However, with the enormous growth in the number of scientific papers published each year (Fortunato et al., 2018; Wang & Barabási, 2021; Bornmann et al., 2021), it becomes increasingly difficult for researchers to uncover novel and interesting ideas. This difficulty is even more pronounced for those looking for interdisciplinary research avenues or collaborations, as they face an overwhelming sea of literature.

Automated systems capable of extracting insights from the millions of scientific papers might offer a solution (Evans & Foster, 2011; Wang & Barabási, 2021). One promising approach involves the use of knowledge graphs, which map the relationships between different research concepts and domains. In a pioneering work, the authors of (Rzhetsky et al., 2015) demonstrate potentially more efficient research strategies in the field of biochemistry by compressing the content of millions of scientific papers into knowledge graphs. These graphs not only help in mapping existing knowledge but also enable the discovery of surprising and impactful ideas by linking previously unconnected concepts. For instance, researchers have utilized knowledge graphs to forecast future research directions in quantum physics (Krenn & Zeilinger, 2020), biomedicine (Sybrandt et al., 2020; Nadkarni et al., 2021), and artificial intelligence (Krenn et al., 2023). Beyond trend prediction and uncovering new links, these approaches have demonstrated that surprising combinations of research concepts are strongly associated with high-impact discoveries (Shi & Evans, 2023). Additionally, human-aware AI systems can generate scientifically promising 'alien' hypotheses that might otherwise be overlooked (Sourati & Evans, 2023), and knowledge graphs have been used to predict the impact of new research connections before a paper is written (Gu & Krenn, 2024).

Some recent efforts demonstrate how to generate research ideas in the form of text. One such example is PaperRobot, which starts with a knowledge graph and incrementally translates the idea into a draft of a paper (Wang et al., 2019). With the growing prominence of large language models (LLMs), various systems now leverage these models to generate research ideas. SciMON, for instance, generates novel scientific ideas by comparing them to prior work and continuously enhancing their novelty (Wang et al., 2024). Another system uses LLMs to mine large-scale scientific literature and generate hypotheses by finding unanticipated connections between research topics (Yang et al., 2023). Additionally, there are systems for scientific discovery that leverage user-specific goals to generate candidate hypotheses (Zhong et al., 2023) or employ either a single-LLM system or multi-agent collaboration for research hypothesis proposals (Qi

[1]Max Planck Institute for the Science of Light, Staudtstrasse 2, 91058 Erlangen, Germany. Correspondence to: Xuemei Gu <xuemei.gu@mpl.mpg.de>, Mario Krenn <mario.krenn@mpl.mpg.de>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024.

et al., 2023). Similarly, ResearchAgent develops new research ideas by analyzing scientific literature and refining them progressively to ensure both novelty and relevance (Baek et al., 2024).

While novelty and relevance of the generated ideas are crucial, a critical question arises: Are these AI-generated research ideas interesting for human scientists? The aforementioned works conducted small-scale human evaluations involving one biomedical domain expert (Wang et al., 2019), six natural language processing (NLP) PhD students (Wang et al., 2024), three social science PhD students (Yang et al., 2023) and ten PhD students in computer science and biomedicine (Baek et al., 2024).

However, it is often experienced researchers who define and evaluate research projects by writing and assessing research grant applications, as well as leading and shaping the research agenda of their groups. It would be particularly insightful to see how these experienced scientists evaluate AI-generated project ideas. With more evaluators and a greater number of evaluations, we could develop tools to predict which research ideas will be interesting in the future. This is precisely the goal of our paper, aiming to suggest interesting research projects and collaborations for scientists.

Here, we introduce SCIMUSE, a system designed to suggest new personalized research ideas for individual scientists or collaborations between researchers. To achieve this, we first generate a knowledge graph from more than 58 million papers, incorporating semantic and impact information. We then identify sub-graphs relevant to the research interests of individual scientists and use these sub-graphs to select research topics. Using GPT-4 (Achiam et al., 2023), we formulate these research topics into comprehensive research suggestions. To evaluate our approach, we conducted a large-scale survey with over 100 research group leaders from the Max Planck Society in natural sciences and technology (such as the Institutes for Biogeochemistry, Astrophysics, Quantum Optics, and Intelligent Systems) and social sciences and humanities (such as the Institutes for Geoanthropology, Demographic Research, and Human Development). These experienced researchers assessed the interest level of more than 4,000 personalized AI-generated project suggestions. We analyzed the evaluations and found clear correlations between the properties of the knowledge graph and the interest level of the research suggestions. Using these correlations together, we would then train a machine learning model to predict research interest based solely on the knowledge graph data. The model achieves high precision for the top-N predicted interesting suggestions, with precision exceeding 50% for $N \leq 15$. Our findings demonstrate the potential of SCIMUSE for suggesting highly interesting research ideas and collaborations, highlighting the role of artificial intelligence as a source of inspiration in science (Krenn et al., 2022; Hope et al., 2023; Wang et al., 2023; AI4Science & Quantum, 2023).

2. Results

2.1. Creating the knowledge graph

While we could directly use publicly available large language models such as GPT-4 (Achiam et al., 2023) or Gemini (Reid et al., 2024) or Claude (Anthropic, 2024) to suggest new research ideas and collaborations, our control over the generated ideas would be limited to the structure of the prompt. Therefore, we decided to build a large knowledge graph from the scientific literature to identify the personalized research interests of scientists.

The knowledge graph, depicted in Fig.1(a), consists of vertices, representing scientific concepts, and edges are drawn when two concepts jointly appear in a title or abstract of a scientific paper. The concept list is generated from the titles and abstracts of around 2.44 million papers from arXiv, bioRxiv, ChemRxiv, and medRxiv, with a data cutoff in February 2023. Rapid Automatic Key-word Extraction (RAKE) algorithm based on statistical text analysis is used to extract candidate concepts (Rose et al., 2010). Those candidates are further refined using GPT, Wikipedia, and human annotators, resulting in 123,128 concepts in the natural and social sciences. We then use more than 58 million scientific papers from the open-source database OpenAlex (Priem et al., 2022) to create edges. These edges contain information about the co-occurrence of concepts in scientific papers (in titles and abstracts) and their subsequent citation rates. This new knowledge graph representation was recently introduced in (Gu & Krenn, 2024) to predict the impact of future research topics. As a result, we have an evolving knowledge graph that captures part of the evolution of science from 1665 (a text by Robert Hooke on the observation of a great spot on Jupiter (Hooke, 1665)) to April 2023. Details of the knowledge graph generation are depicted in Fig.1(a) and the Appendix.

2.2. Personalized research suggestions

We focus on generating personalized research proposals for collaborations between two scientists, both group leaders from the Max Planck Society. One of these researchers will later evaluate the proposal.

To generate suggestions for pairs of researchers, as depicted in Fig.1(b), we begin by identifying the research interests of both Researcher A and Researcher B. This is done by analyzing all their published papers from the past two years. Specifically, we extract their concepts from the titles and abstracts of these papers using the full concept list shown in Fig.1(a). The personalized concept lists are further refined by GPT-4. Consequently, we are able to construct a

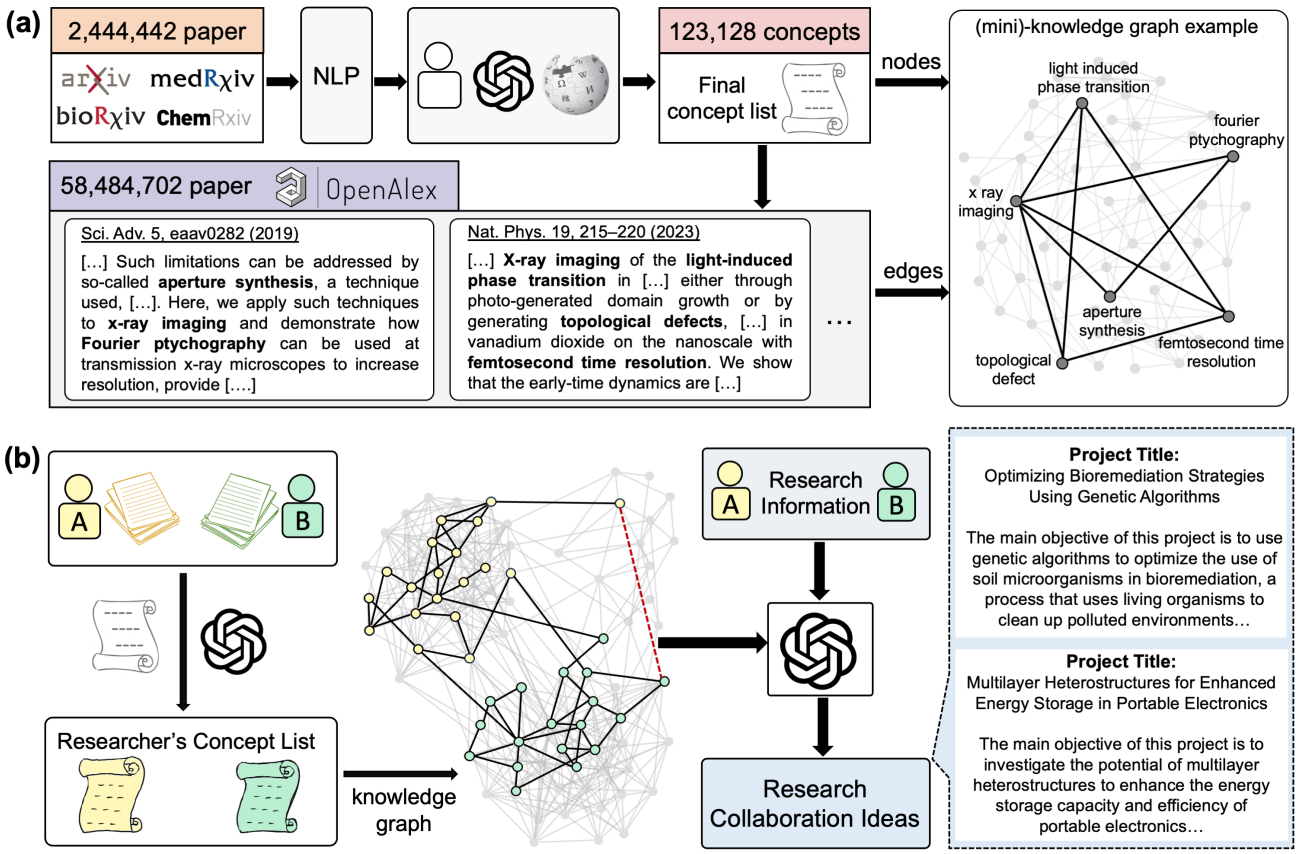

Figure 1. **SciMuse suggests research ideas or collaborations using knowledge graph and GPT-4**. **(a)**, Generation of a knowledge graph. Nodes represent scientific concepts extracted from about 2.44 million paper titles and abstracts from four academic preprint servers. Using natural language processing (NLP) tools such as RAKE (Rose et al., 2010) to create a concept list, we then refined it with customized NLP techniques, manual review, and GPT, removing non-conceptual phrases like verbs, ordinal numbers, conjunctions, and adverbials. Wikipedia was used to restore any mistakenly removed concepts. In the end, we obtained a final list of 123,128 concepts. Edges are created when two concepts co-occur in the title or abstract of more than 58 million scientific papers in the OpenAlex database (Priem et al., 2022). These edges are augmented with citation information, which can serve as a proxy for impact. A mini-knowledge graph as an example is shown for two randomly selected papers (Wakonig et al., 2019; Johnson et al., 2023) in OpenAlex. **(b)**, AI-generated research collaborations. We first process the publications of Researcher A and Researcher B through our refined concept list from (a), generating individual concept lists for each researcher. We then use GPT-4 to enhance these lists to create high-quality concept representations. These refined lists identify distinct subnetworks within our knowledge graph that correspond to each researcher's interests. To propose research collaborations or ideas, we identify and combine relevant concept pairs between the two researchers along with their research information. This combined input is then fed into GPT-4, which generates personalized research ideas or collaboration projects.

subgraph in the knowledge graph for each researcher based on theirs personalized concepts.

With the researchers' subgraphs, we generate a prompt for GPT-4 to create a research project (details in the Appendix). In the prompt, we provide the titles of up to seven papers from each researcher and ask GPT-4 to create a research project based on two selected scientific concepts. We choose these concepts in three different ways. In one-third of the cases, we use a randomly sampled concept pair, with one concept from each researcher. In another third, we select the concept pair with the highest predicted future impact, using an adaptation of (Gu & Krenn, 2024). In the final third, we do not specify concept pairs, instead asking GPT-4 to create the project using only the paper titles. Although we cannot directly relate these pure GPT-4 suggestions to knowledge graph features and interest levels from human evaluation, they serve as an important sanity check for our method (see Appendix). The prompt itself employs self-reflection, as described in (Madaan et al., 2024), to improve the response. Specifically, we ask GPT-4 to generate three ideas, reflect upon them, and improve them twice. In the end, GPT-4 selects the most suitable project idea as the final result.

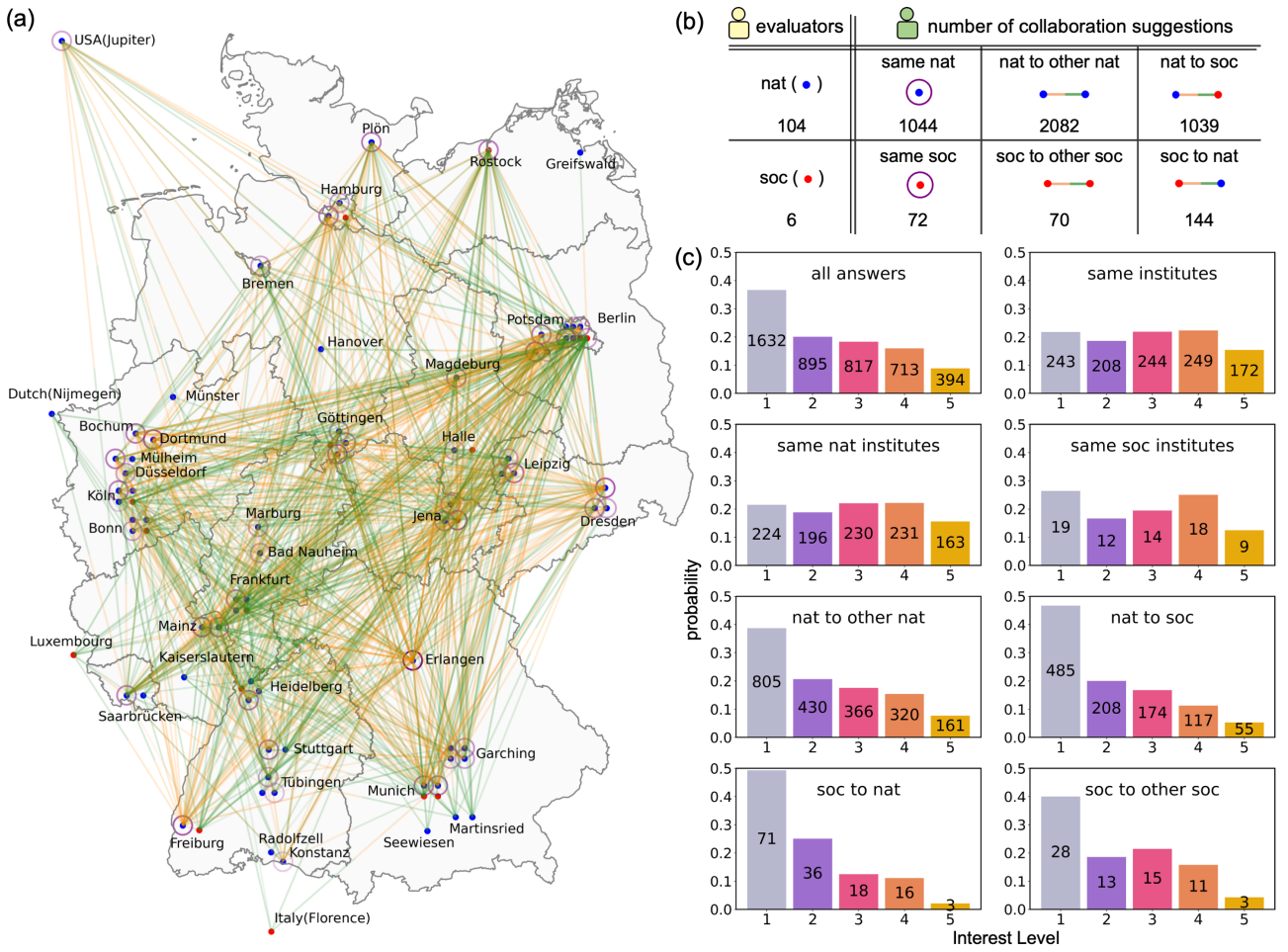

Figure 2. **Large-scale human evaluation within the Max Planck Society. (a)-(b)**, A total of 4,451 AI-generated personalized research suggestions were evaluated by 110 research group leaders. Each suggestion proposes a collaboration between the evaluating researcher (Researcher A) and another researcher (Researcher B) from the Max Planck Society. These proposed collaborations are visualized as edges on a graph, where an edge is bi-colored from orange (representing Researcher A) to green (representing Researcher B). If Researchers A and B are from the same institute, this is indicated by a purple circle around that institute. The transparency of the edge is proportional to the number of evaluated suggestions. Additionally, the research fields of the researchers are categorized into natural science (denoted by a blue dot, labeled as *nat*) and social science (denoted by a red dot, labeled as *soc*). The map of Germany is based on GISCO statistical unit dataset from Eurostat (Commission, 2024). **(c)**, For each research suggestion, participants were asked to rate their interest on a scale from 1 ('not interesting') to 5 ('very interesting'). The summary figure displays the distribution of these ratings. In total, 394 research suggestions were rated as *very interesting*, and 713 ideas received a rating of 4. The figure includes separate sections for responses where both researchers are from the same institute, as well as for those from different institutes, further categorized by their affiliation with either the natural science or social science faculties. For example, 'nat to other nat' means researchers A and B are from different natural science institute, 'nat to soc' means that researcher A is from natural science institute and research B is from social science institute.

2.3. 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 (one of the largest research societies worldwide) participated (see Fig.2(a) and (b)). 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. The full statistics are shown in Fig.2(c). Notably, 1,107 projects received an interest level of 4 or 5 (nearly 25% of the projects), with 394 of these being ranked as *very interesting*.

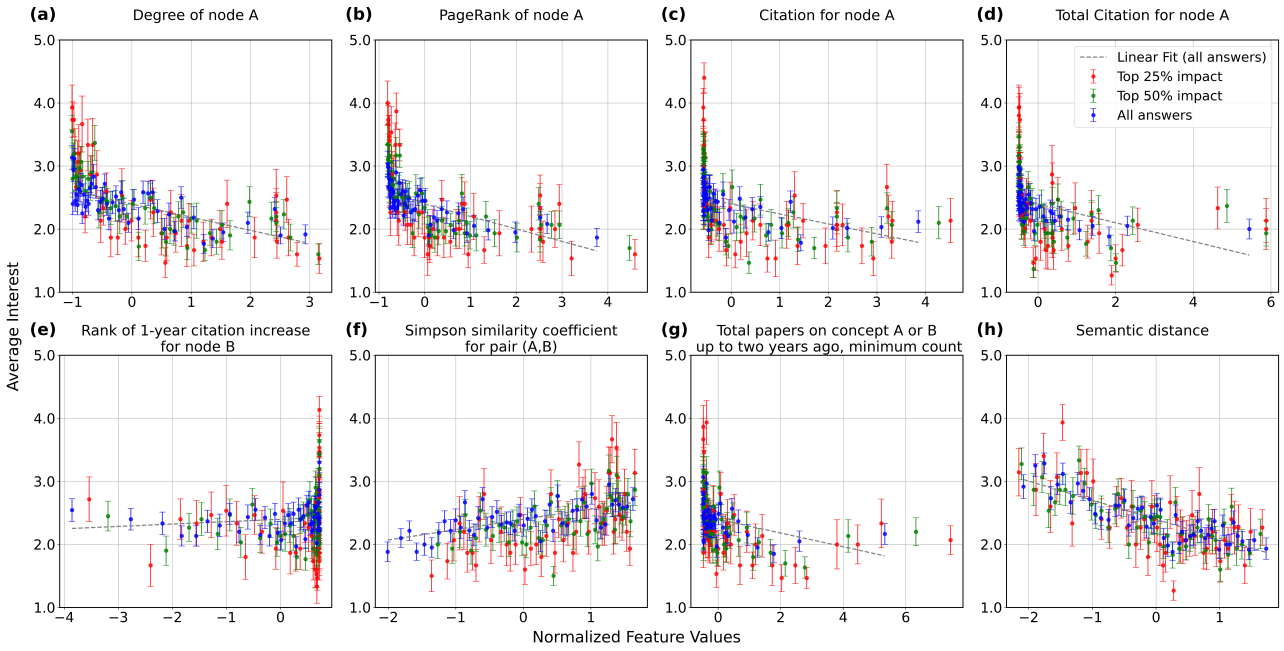

Figure 3. **Analysis of interest levels versus knowledge graph features**. We analyzed how eight individual features of a knowledge graph relate to researchers' interest levels. After normalizing these features using z-scores, we arranged them from lowest to highest and segmented the data into 50 equal groups. For each group, we plotted the average normalized feature value (x-axis) alongside the corresponding average interest value (y-axis), including the standard deviation for each point, to identify trends in how different graph features influence researchers' preferences. Features (a) and (b) relate to node features, (c) to (e) to node citation metrics, (f) is an edge feature, (g) is an edge citation metric, and (h) represents the semantic distance of the two researchers' sub-networks (larger values mean that the researcher's scientific fields are further apart). The plot includes data points in blue representing all 2,996 responses, green for the top 50% of research questions by predicted impact, and red for the top 25%.

2.4. Interest versus knowledge graph features

We find that, on average, there is no significant difference in the interest value between projects generated by the three different methods: random concept pairs, high-impact concept pairs, and without concept pairs. The fact that the sanity test (a project generated without a concept pair in the prompt) and the cases where we provide concept pairs yield very similar results allows us to further analyze which knowledge graph features strongly influence the *interest*. If we can determine which features affect the interestingness of a research project, we can use this insight in the future to suggest research projects with higher research interest.

We first compute 144 knowledge graph features for each concept pair used in a research project. The first 141 features are the same features as those used to predict the future impact of concept pairs, as described in (Gu & Krenn, 2024). The features include node characteristics of the first and second concepts, such as node degree and PageRank (Page et al., 1999), as well as edge features, such as the Simpson similarity, and the Sørensen–Dice coefficient (Barabási, 2016). Additionally, several features are based on impact information, such as citations within the last year. The final three features are the predicted impact and two different distance metrics of the researchers' subgraphs (see 1(b)). The first distance metric considers only the subgraphs, using the concepts from Researcher A's and Researcher B's concept lists to determine the distance between these subgraphs. The second metric accounts for the entire neighborhood of the subgraphs, defined as semantic distances. For this, we collect the neighbors of all concepts in both researchers' concept lists and determine the distance between these expanded subgraphs built from neighboring concepts.

We then split the 2,996 suggested research projects, created using concept pairs from the knowledge graph, into 50 equally sized bins. For each bin, we compute the mean interest and its standard deviation.

In Fig.3, we display these correlations and identify several notable properties. For instance, the degree and PageRank of the first concept, selected from the evaluating researcher's concept list, is strongly negatively correlated with human-evaluated interest-level. This means that the more widely connected a concept is within the knowledge graph, the less appealing the research projects are. A similar effect is ob-

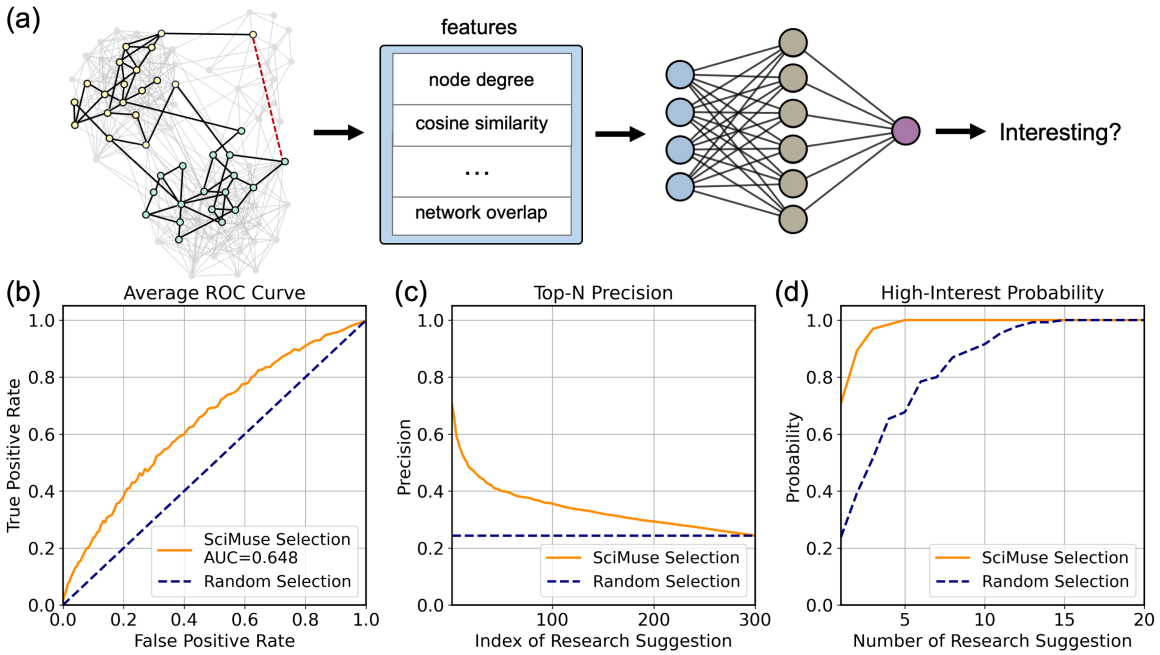

Figure 4. **Learning Scientific Interest**. **(a)**, We use the evaluations from research group leaders to train a neural network. This model predicts whether research suggestions are assigned an interest level of 4 or 5 (on a 5-point scale) or below 4, thereby setting up a binary classification task. The input to the neural network consists of 25 features from the knowledge graph of a concept pair, and its output is a single number indicating whether the concept pair is highly interesting (i.e., interest level is $\geq$4) or not. Given the small size of our training dataset, which comprises a total of 2,996 evaluated research suggestions generated through our knowledge graph, we employ Monte Carlo cross-validation to determine the accuracy of our learning process. **(b)**, The ROC curve indicates that we can correctly predict a randomly chosen highly interesting concept pair over a randomly chosen not-highly interesting concept pair in nearly 65% of cases. **(c)**, The precision of our model for the top-N highest-interest research suggestions is significantly higher than for a random selection of suggestions. Especially for the Top-1 suggestion, the precision is larger than 70%, and Top-5, the precision is still above 60%. **(d)**, The probability of having at least one high-interest suggestion among N research suggestions is significantly higher than with a random selection. This indicates that our machine learning model, which has access to the knowledge graph in conjunction with GPT-4, is able to produce more high-interest research suggestions than GPT-4 itself.

served for the citation rate: the more frequently a concept has been cited in the past (in the last year, and sum over all years), the less interesting the research projects are evaluated. Some features, such as the rank of concept B's citation growth rate or the minimum count of the total number of papers containing concept A or B from the paper's publication time up to 2020 (i.e., two years before the current year 'y'; 'y=2022' means 2022-12-31), show peculiar behaviour for very large or small values. This behaviour could be exploited to predict the interest level. On the other hand, we find a strong positive correlation between the Simpson similarity coefficient of the two concepts and the evaluated interest-level. Additionally, using semantic distance feature, we find a negative correlation in Fig.3(h), indicating that research proposals from researchers in similar fields are considered more interesting than those from distant fields. This finding is consistent with Fig.2(c), where research proposals from the same institutes are generally considered more interesting than those from other institutes (with different research focus).

We show the correlations for all 2,996 answers containing concepts from the knowledge graph (blue), as well as for the top 50% and top 25% of concept pairs with the highest predicted impact (green and red, respectively) in Fig.3, indicating that some correlations are stronger for suggestions using high-impact concept pairs.

2.5. Predicting interest

Given that the features of the knowledge graph significantly influence the interest in suggested research projects, we can take this analysis a step further by training a machine learning model to predict the level of interest based solely on these properties. If successful, this approach would allow us to suggest research projects that are more likely to be considered highly interesting in the future for scientists.

We start with a concept pair, compute the relevant features in the knowledge graph, and use these features to predict whether a research proposal will receive an interest rating

of 4 or 5 (on a scale from 1 to 5: *not interesting* to *very interesting*) or below 4, as illustrated in Fig.4(a). Due to the scarcity of training data – each data point representing the evaluation of a proposed research project's interestingness by a research group leader – we employ a low-data machine learning technique. Specifically, we use a small neural network configured with 25 individually high-performing features, 50 neurons in a single hidden layer, and one output neuron, incorporating dropout to train the neural network (Srivastava et al., 2014). To ensure robust evaluation and maximize the utility of our limited data, we utilize Monte Carlo cross-validation, also known as repeated random subsampling validation (see Appendix).

For our binary classification task, we achieve an average Area Under the Curve (AUC) of the receiver operating characteristic (ROC) curve (Fawcett, 2004) of nearly 2/3, as shown in Fig.4(b). More relevant for our task is achieving high precision, as we want SCIMUSE to suggest highly interesting projects within a very small number of overall suggestions. For this, we compute the precision of the top-N highest predicted concept pairs. For small N, we find a precision higher than 65%. This indicates that within the highest predicted suggested concept pairs, roughly 65% are evaluated with high interest level, as illustrated in Fig.4(c). This precision is significantly higher than random selection, which achieves only 23%. Additionally, we can ask what is the probability of obtaining at least one highly interesting suggestion within the first N suggestions. Fig.4(d) shows that our machine learning method provides a significantly higher probability of finding interesting suggestions within the first few suggestions compared to random sampling.

3. Discussion

Our results show that one could predict which project suggestions experienced researchers will find interesting by analyzing the knowledge-graph properties of the concept pairs used for the prompts to GPT-4, without considering the detailed text produced by GPT-4. This finding allows us to enhance SCIMUSE such that it can select novel, and high-interest research topics from knowledge graphs and translate them into full-fledged proposals using modern large language models. As publicly available large language models like GPT-4 (Achiam et al., 2023), Gemini 1.5 (Reid et al., 2024), LLaMa3 (AI, 2024), and Claude (Anthropic, 2024) become increasingly powerful, with improvements occurring nearly monthly (Chiang et al., 2024), we anticipate that the generated personalized research ideas will become more targeted and reasonable.

The methodologies demonstrated in our work, employed by SCIMUSE, have the potential to inspire novel, unexpected cross-disciplinary research on a large scale. By providing a big-picture view through the analysis of millions of scientific papers, SCIMUSE allows the discovery of interesting research projects between scientists in different domains, which might otherwise be very challenging to find. Research projects in distant fields are known to have great potential for impactful, award-winning results (Uzzi et al., 2013; Rzhetsky et al., 2015; Fortunato et al., 2018; Wang & Barabási, 2021). Therefore, large scientific societies, national funding agencies, and other stakeholders might be motivated to implement methodologies in the line of SCIMUSE, which could foster new highly interdisciplinary and interesting collaborations and ideas that might otherwise remain untapped. This, hopefully, could advance the progress and impact of science at a large 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, and medRxiv 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.

Impact Statement

This paper presents work whose goal is to suggest interesting research ideas for scientists and accelerate the progress of scientific discovery. There are no specific ethical or societal impacts that needs to be considered.

References

Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al. Gpt-4 technical report. *arXiv:2303.08774*, 2023.

AI, M. Llama 3: Open foundation and fine-tuned chat models. https://github.com/meta-llama/llama3, 2024.

AI4Science, M. R. and Quantum, M. A. The impact of large language models on scientific discovery: a preliminary study using gpt-4. *arXiv:2311.07361*, 2023.

Anthropic. The claude 3 model family: Opus, sonnet, haiku. Papers with Code, 2024.

Baek, J., Jauhar, S. K., Cucerzan, S., and Hwang, S. J. Researchagent: Iterative research idea genera-

tion over scientific literature with large language models. *arXiv:2404.07738*, 2024.

Barabási, A.-L. *Network Science*. Cambridge University Press, 2016.

Bornmann, L., Haunschild, R., and Mutz, R. 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.

Chiang, W.-L., Zheng, L., Sheng, Y., Angelopoulos, A. N., Li, T., Li, D., Zhang, H., Zhu, B., Jordan, M., Gonzalez, J. E., et al. Chatbot arena: An open platform for evaluating llms by human preference. *arXiv:2403.04132*, 2024.

Commission, E. Eurostat gisco - nuts geodata. https://ec.europa.eu/eurostat/web/gisco/geodata/reference-data/administrative-units-statistical-units/nuts, 2024.

Evans, J. A. and Foster, J. G. Metaknowledge. *Science*, 331 (6018):721–725, 2011.

Fawcett, T. Roc graphs: Notes and practical considerations for researchers. *Machine learning*, 31(1):1–38, 2004.

Fortunato, S., Bergstrom, C. T., Börner, K., Evans, J. A., Helbing, D., Milojević, S., Petersen, A. M., Radicchi, F., Sinatra, R., Uzzi, B., et al. Science of science. *Science*, 359(6379):eaao0185, 2018.

Gu, X. and Krenn, M. Forecasting high-impact research topics via machine learning on evolving knowledge graphs. *arXiv:2402.08640*, 2024.

Hooke, R. A spot in one of the belts of jupiter. *Philosophical Transactions of the Royal Society of London*, 1 (1):3–3, 1665. URL https://doi.org/10.1098/rstl.1665.0005.

Hope, T., Downey, D., Weld, D. S., Etzioni, O., and Horvitz, E. A computational inflection for scientific discovery. *Communications of the ACM*, 66(8):62–73, 2023.

Johnson, A. S., Perez-Salinas, D., Siddiqui, K. M., Kim, S., Choi, S., Volckaert, K., Majchrzak, P. E., Ulstrup, S., Agarwal, N., Hallman, K., et al. Ultrafast x-ray imaging of the light-induced phase transition in vo2. *Nature Physics*, 19(2):215–220, 2023.

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

Krenn, M., Pollice, R., Guo, S. Y., Aldeghi, M., Cervera-Lierta, A., Friederich, P., dos Passos Gomes, G., Häse, F., Jinich, A., Nigam, A., et al. On scientific understanding with artificial intelligence. *Nature Reviews Physics*, 4 (12):761–769, 2022.

Krenn, M., Buffoni, L., Coutinho, B., Eppel, S., Foster, J. G., Gritsevskiy, A., Lee, H., Lu, Y., Moutinho, J. P., Sanjabi, N., 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.

Madaan, A., Tandon, N., Gupta, P., Hallinan, S., Gao, L., Wiegreffe, S., Alon, U., Dziri, N., Prabhumoye, S., Yang, Y., et al. Self-refine: Iterative refinement with self-feedback. *Advances in Neural Information Processing Systems*, 36, 2024.

Nadkarni, R., Wadden, D., Beltagy, I., Smith, N. A., Hajishirzi, H., and Hope, T. Scientific language models for biomedical knowledge base completion: an empirical study. *arXiv:2106.09700*, 2021.

Page, L., Brin, S., Motwani, R., and Winograd, T. The pagerank citation ranking : Bringing order to the web. *Stanford InfoLab*, 1999.

Priem, J., Piwowar, H., and Orr, R. Openalex: A fully-open index of scholarly works, authors, venues, institutions, and concepts. *arXiv:2205.01833*, 2022.

Qi, B., Zhang, K., Li, H., Tian, K., Zeng, S., Chen, Z.-R., and Zhou, B. Large language models are zero shot hypothesis proposers. *arXiv preprint arXiv:2311.05965*, 2023.

Reid, M., Savinov, N., Teplyashin, D., Lepikhin, D., Lillicrap, T., Alayrac, J.-b., Soricut, R., Lazaridou, A., Firat, O., Schrittwieser, J., et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv:2403.05530*, 2024.

Rose, S., Engel, D., Cramer, N., and Cowley, W. Automatic keyword extraction from individual documents. *Text mining: applications and theory*, pp. 1–20, 2010.

Rzhetsky, A., Foster, J. G., Foster, I. T., and Evans, J. A. Choosing experiments to accelerate collective discovery. *Proc. Natl. Acad. Sci. USA*, 112(47):14569–14574, 2015.

Shi, F. and Evans, J. 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.

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

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(56):1929–1958, 2014.

Sybrandt, J., Tyagin, I., Shtutman, M., and Safro, I. Agatha: automatic graph mining and transformer based hypothesis generation approach. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, pp. 2757–2764, 2020.

Uzzi, B., Mukherjee, S., Stringer, M., and Jones, B. Atypical combinations and scientific impact. *Science*, 342(6157): 468–472, 2013.

Wakonig, K., Diaz, A., Bonnin, A., Stampanoni, M., Bergamaschi, A., Ihli, J., Guizar-Sicairos, M., and Menzel, A. X-ray fourier ptychography. *Science advances*, 5(2): eaav0282, 2019.

Wang, D. and Barabási, A.-L. *The science of science*. Cambridge University Press, 2021.

Wang, H., Fu, T., Du, Y., Gao, W., Huang, K., Liu, Z., Chandak, P., Liu, S., Van Katwyk, P., Deac, A., et al. Scientific discovery in the age of artificial intelligence. *Nature*, 620(7972):47–60, 2023.

Wang, Q., Huang, L., Jiang, Z., Knight, K., Ji, H., Bansal, M., and Luan, Y. Paperrobot: Incremental draft generation of scientific ideas. *arXiv:1905.07870*, 2019.

Wang, Q., Downey, D., Ji, H., and Hope, T. Scimon: Scientific inspiration machines optimized for novelty. *Annual Meeting of the Association for Computational Linguistics*, 2024.

Yang, Z., Du, X., Li, J., Zheng, J., Poria, S., and Cambria, E. Large language models for automated open-domain scientific hypotheses discovery. *arXiv:2309.02726*, 2023.

Zhong, R., Zhang, P., Li, S., Ahn, J., Klein, D., and Steinhardt, J. Goal driven discovery of distributional differences via language descriptions. *Advances in Neural Information Processing Systems*, 36:40204–40237, 2023.

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 bioRxiv, medRxiv, and chemRxiv metadata can be accessed through their APIs. Our dataset includes ∼2.44 million papers, with a data cutoff in February 2023.

For edge generation, we used the OpenAlex database snapshot, available for download in OpenAlex bucket, with a data cutoff in April 2023. For more details, refer to the OpenAlex website (Priem et al., 2022). The complete dataset is around 330 GB, expanding to 1.6 TB when decompressed. We focused on scientific journal papers with publication time, title, abstract, and citation information, reducing the dataset to a more manageable 68 GB gzip-compressed file, comprising about 92 million papers.

B. Creating the concept list

From four preprint dataset of approximately 2.44 million papers, we analyzed each article's title and abstract using the RAKE algorithm, enhanced with additional stopwords, to extract potential concept candidates. These candidates were stored for subsequent analysis. We filtered out concepts to retain only those with two words that appeared in nine or more articles, and those with three or more words that appeared in six or more articles. This step significantly reduced the noise from the RAKE-generated concepts, yielding a refined list of 726,439 relevant concepts.

To further enhance the quality of the identified concepts, we developed a suite of automatic tools designed to identify and eliminate common, domain-independent errors often associated with RAKE. Additionally, we conducted a manual review to remove inaccuracies in the concepts, such as non-conceptual phrases, verbs, ordinal numbers, conjunctions, and adverbials, reducing 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. To address potential incorrect removals, we used Wikipedia to recover mistakenly removed concepts, successfully restoring 40,614 concepts. This process ultimately produced a final list of 123,128 concepts.

C. Classification of Max Planck Institutes

We classify all 87 Max Planck Institutes into two classes: 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). We did manual classification based on institute titles and research fields, and we also used GPT-4o for automatic classification. The two approaches perfectly matched each other.

D. Prompt to GPT-4 for concept refinement

The prompt to refine the researchers' concept list is:

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 that you help me filtering them. Please look at the list below, and return all concepts in that list, which are relevant to the scientists research (based on their paper titles), and that are meaningful in the context of their research direction. The concepts can be detailed, I mainly want that you filter out not meaningful concepts, words which are not concepts, or concepts that are too general for the direction of the scientist (for example, artificial intelligence might be a meaningful concept for a geologist, but not for a machine learning researcher). Do not change or add any of the concepts. only remove them or keep them.

concept_list=[c1, c2, c3, c4, c5, c6, ...]

E. Prompt to GPT-4 for project idea generation

The prompt to suggest research ideas using concept pair from knowledge graph is:
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" for GPT-4, it is possible to direct GPT-4 to extract these concepts from the titles of research papers authored by Researcher A and Researcher B, respectively. Subsequently, GPT-4 can use these identified concepts within the same prompting context to generate innovative research ideas.

F. Interest-Evaluation for three different generation methods

In Fig.5, we show the three different generation methods for the research suggestions. The interest-level distributions are very similar, particularly between those with and without concepts from the knowledge graph. This similarity allows us to analyze the correlations between the properties of knowledge graph and interest level, and to use these properties to predict the interest level of proposals.

G. Predicting high interest from knowledge graph features

In Fig.4 in the main text, our goal is to predict whether a certain research proposal will be evaluated with high interest. Specifically, using only data from the knowledge graph (and not the final text of the research proposal generated with GPT), we want to predict whether the proposal receives an interest value of 4 or 5 (on a scale from 1 to 5: *not interesting* to *very*

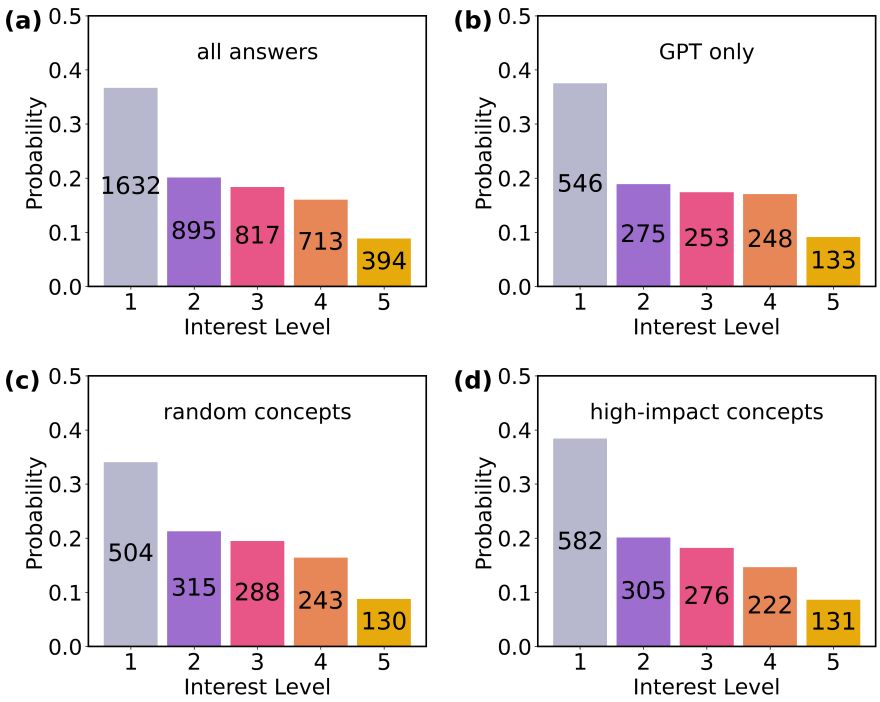

Figure 5. **Interest levels depending on generation method**. We use three different ways to generate research ideas: (1) no concepts provided by the knowledge graph, (2) random concepts from the researchers' subnetwork, and (3) high-impact concept pairs from the researchers' subnetwork. The figures show: **(a)** the interest level of all answers (numbers inside the bars indicate the number of answers with that evaluation), **(b)** answers without using concepts from the knowledge graph, **(c)** answers with random concept pairs, and **(d)** high-impact predicted concept pairs (using the neural network from (Gu & Krenn, 2024)).

interesting) or below 4, which constitutes a binary classification task.

Due to the small dataset size (2,996 answers with properties from the knowledge graph), we use a data-efficient learning method for the prediction task, specifically a small neural network with dropout. The input to the neural network consists of the 25 best-performing features from the knowledge graph. The neural network has only one hidden layer with 50 neurons and a single output neuron. We use mean square error as the loss function.

To get a consistent evaluation of the neural network performance for this small dataset, we perform Monte Carlo cross-validation. In this method, the dataset is randomly split into training and validation sets multiple times, and the model is trained and evaluated on each split. This process ensures that the performance metrics are robust and not dependent on a particular split of the data. We continue splitting and evaluating until the standard deviation of the mean AUC is less than $\frac{10^{-2}}{3}$, which is achieved after 130 iterations. This approach 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 choice: *learning rate=0.003*, *dropout=20%*, *weight decay=0.0007*, *training dataset=75%*, *validation dataset=15%*, *test dataset=10%*.

We experimented with other data-efficient learning methods, such as decision trees, but they did not significantly outperform the neural network.