Impact4Cast: Forecasting high-impact research topics via machine learning on evolving knowledge graphs

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

Abstract

The exponential growth in scientific publications poses a severe challenge for human researchers, forcing them to focus on narrower sub-fields and making it difficult to discover new, impactful research ideas and collaborations outside their own fields. Although there are methods to predict a scientific paper's future citation counts, these require the research to be completed and the paper to be written, often assessing impact long after the idea was first conceived. Here we show how to predict the impact of onsets of ideas that have never been published by researchers. For that, we developed a large evolving knowledge graph built from more than 21 million scientific papers, which combines a semantic network created from the content of the papers and an impact network created from the historic citations of papers. Using machine learning, we can predict the dynamic of the evolving network into the future with high accuracy, and thereby the impact of new research directions. We envision that the ability to predict the impact of new ideas will be a crucial component of future artificial muses that can inspire new impactful and interesting scientific ideas.

1. Introduction

As we see an explosion in the number of scientific articles (Fortunato et al., 2018; Wang & Barabási, 2021; Bornmann et al., 2021; Krenn et al., 2023a), it becomes increasingly challenging for researchers to find new impactful research directions beyond their own expertise. Consequently, researchers might have to focus on narrow subdisciplines. A tool that can read and intelligently act upon scientific literature could be an enormous aid to individual scientists in choosing their next new and high-impact research project, which – on a global scale – could significantly accelerate science itself.

These days, a natural first choice for an AI-assistant would be powerful large-language-models (LLMs) such as GPT-4 (OpenAI, 2023), Gemini (Google, 2023), LLaMA-2 (Touvron et al., 2023) or custom-made models (Wang et al., 2023b). However, these models often struggle in scientific reasoning, and it remains unclear how they can suggest new scientific ideas or evaluate their impact in a reliable way in the near term.

An alternative and complementary approach is to build scientific semantic knowledge graphs. Here, the nodes represent scientific concepts and the edges are formed when two concepts are researched together in a scientific paper (Wang & Barabási, 2021). While this approach extracts only small amounts of information from each paper, surprisingly nontrivial conclusions can be drawn if the underlying dataset of papers is large. An early example of this is a work in biochemistry (Rzhetsky et al., 2015). The authors use their semantic network, where nodes represent biomolecules, to find new potentially more efficient exploration strategies for the bio-chemistry community on a global scale. In these semantic networks, an edge between two concepts indicates that researchers have jointly investigated these research concepts. The edges are drawn from papers, thus they are created at a specific time when the paper was published. In this way, one creates an evolving semantic network that captures what researchers have investigated in the past. With such an evolving network, one can ask how the network might evolve in the future. In the scientific context, this question can be reformulated into what scientists will research in the future. For example, if two nodes do not share an edge, one can ask whether they will share an edge in the next three years – or, alternatively, whether scientists will investigate these two concepts jointly within three years. This question, denoted as a link-prediction problem in network theory (Martínez et al., 2016), has been successfully demonstrated with high prediction quality for semantic networks in the field of quantum physics (Krenn & Zeilinger, 2020) and artificial intelligence (Krenn et al., 2023a). These works fo-

[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. Copyright 2024 by the author(s).

cus on the question *what scientists will work on*, completely leaving out which of these topics will be impactful.

Impact in the scientific community is often approximated (for lack of better metrics (nat, 2017; 2024)) by citations (Fortunato et al., 2018; Wang & Barabási, 2021; Barabási, 2018; Frank et al., 2019), including exciting results that find interpretable mathematical models to describe citation evolution (Wang et al., 2013; Ke et al., 2015; Sinatra et al., 2016; Wu et al., 2019). Beside concrete mathematical modelling, impact of scientific papers has also been predicted using advanced statistical and machine-learning methods that use meta-data such as including authors and affiliations (Weis & Jacobson, 2021), the content and the references of the paper (Bai et al., 2017; Xia et al., 2023). One could analyze the structure of the citation network (Min et al., 2021) or gain insights from atypical knowledge (Uzzi et al., 2013; Fontana et al., 2020) for impact prediction. Techniques employed for the predictions of individual paper impact using a combination of characteristics include support-vector machines (Fu & Aliferis, 2010), regression (Yu et al., 2014; Stegehuis et al., 2015; Weihs & Etzioni, 2017), dense (Ruan et al., 2020) or graph neural networks (He et al., 2023).

The prediction of a paper's impact however is only possible after the research is completed, and long after its underlying idea is created. A true scientific assistant or muse however should contribute at the earliest stage of the scientific cycle, when the idea for the next impactful research project is born. One solution is the prediction at the concept level. Specifically, we can ask the question *Which scientific concepts, that have never been investigated jointly, will lead to the most impactful research?*.

In this work, we answer this question by combining semantic networks and citation networks that are purely based on the level of scientific concepts [1]. Specifically, we develop a large evolving knowledge graph using more than 21 million scientific papers, from 1709 (starting with a letter by Antoni van Leeuwenhoek (Leeuwenhoek, 1709)) to April 2023. The vertices of the knowledge graph are scientific concepts and the edges between two concepts contain information about when these topics have been investigated and how often they have been cited subsequently. We then train a machine learning model on the historic evolution of the knowledge graph. We find that the neural network can predict with high accuracy which concept pairs, that have never been jointly investigated before in any scientific paper, will be highly cited in the future. Being able to predict the potential impact of new research ideas – before the paper is written or the research is done or even started – could be a cornerstone in future scientific AI-assistants that help humans broadening their horizon of possible new research endeavours (Krenn et al., 2022).

[1] GitHub: Impact4Cast

2. Results

2.1. Creating a list of scientific concepts

At the heart of our knowledge graph are scientific concepts, as depicted in Fig. 1. We chose not to rely on existing concept lists, such as the APS or computer science ontology (Salatino et al., 2019), for several reasons. Firstly, our goal is to ultimately cover all natural sciences comprehensively, and a universal list encompassing this breadth doesn't currently exist. Secondly, we want to capture the most recent concepts that might be absent from existing lists. Lastly, generating our list ensures that we have a granular understanding and control over the concepts.

To build our concept list, we started with 2,444,442 papers from four publicly available preprint servers: arXiv, bioRxiv, medRxiv, and chemRxiv. The data cutoff is February 2023. From these, we extracted titles and abstracts of the papers. To single out concept candidates from this extensive collection, we applied the Rapid Automatic Keyword Extraction (RAKE) algorithm based on statistical text analysis to automatically detect important keywords (Rose et al., 2010). Concepts with two words, like *phase transition*, were retained if they appear in at least 9 papers, while longer concepts, such as *single molecule localization microscopy*, needed to appear in at least 6 papers. In this way, we can increase the fraction of high-quality concepts. We further developed a suite of natural language processing tools to refine the concepts, followed by manual inspection to remove any incorrectly identified ones. Finally, we got a list which contains over 368,000 concepts. We focus here on concepts specific to the sub-field of optics and quantum physics (representing roughly 10% of the entire concepts), but our method can immediately be translated to any other domain. This refined domain-specific concept list serves as the vertices of our knowledge graph.

2.2. Creating an evolving, citation-augmented knowledge graph

Now that we have the vertices, we can create edges that contain information from the scientific literature. We get the citation information from papers in OpenAlex (Priem et al., 2022), an open-source database containing detailed information on more than 92 million publications. Edges are drawn when two concepts co-occur in the title or abstract of a scientific paper. If a paper connects two vertices, the weight of the newly formed edge is the paper's annual citation numbers from 2012 to 2023 together with the total citation number since its publication. If more than one paper creates an edge, then the edge contains the sum of the annual citations (as well as the sum of the total citations) gained by all papers. As research papers appear over time, and their citations are created in time, we effectively build an evolving, citation-augmented knowledge graph that evolves

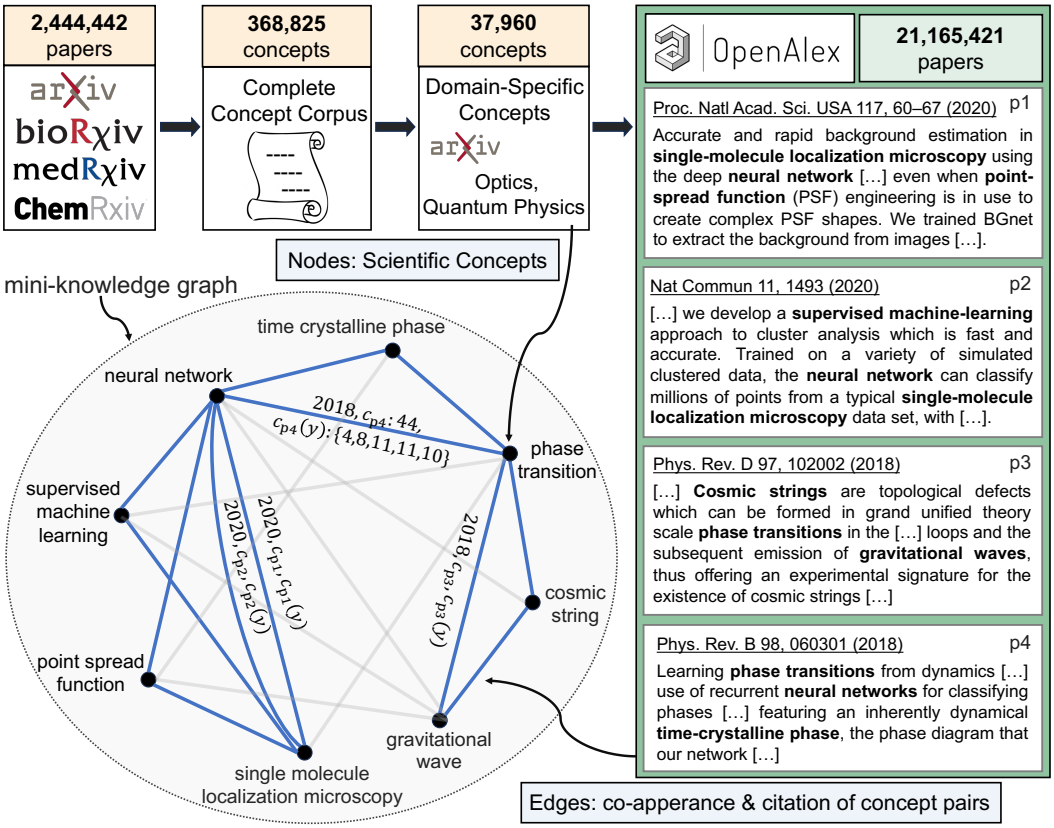

Figure 1. **Generation of the knowledge graph with time and citation information**. Vertices are formed by scientific concepts, which are extracted from scientific papers (titles and abstracts) from prominent academic preprint servers. Edges are formed when concepts are investigated jointly in a scientific publication. There are 21,165,421 out of 92,764,635 papers from OpenAlex which form at least one edge. The edges are augmented with citation information, acting as a proxy for impact in our work. A mini-knowledge graph (blue edges) is constructed from four random papers (p1-p4) (Möckl et al., 2020; Williamson et al., 2020; Abbott et al., 2018; van Nieuwenburg et al., 2018) from OpenAlex as an example. Here, c_{p4} represents the total citations of paper p4 since its publication, and $c_{p4}(y)$ is its annual citations from 2018 to 2022 (e.g., $c_{p4}(2018) = 4$). The citation value of the edge is the sum of the all papers creating the edge.

in time (see Fig. 1).

The final constructed knowledge graph has 37,960 vertices with more than 26 million edges (built from 190 million concept pairs, containing multi-edges when multiple papers create the same edge) from the OpenAlex dataset, with a data cutoff at April 2023. In Fig. 2, we see that the distribution of the 3-year citation increase for previously uncited concept pairs exhibits a heavy tail. This suggests that some concept pairs are cited significantly more than would be expected from an exponential decrease. In Fig. 3, we show the fastest growing (in terms of citation) concepts and concept pairs since 2012, where we can recognize many highly influential topics in quantum physics and optics research.

2.3. Forecasting impact of newly created concept connections

With an evolving knowledge graph from the past, we can formulate the prediction of impact for new concept pairs as a supervised learning task, as illustrated in Fig. 4. For a vertex pair that has not had any connection in the year 2016, we predict whether three years later this pair accumulated more than a certain number of citations. Using the historical knowledge graph, we possess an ideal supervision signal for our binary classification task. During the training phase, we selected pairs of vertices that were not connected and calculated 141 features for each pair. These features include 41 network features, divided into 20 node features (such as the number of neighbors and PageRank (Page et al., 1999) over the past three years) and 21 edge features (including cosine, geometric, and Simpson similarities (Barabási, 2016)). Additionally, we incorporated 100 impact features: 58 of these are node citation features, covering total citations and yearly citations within the last three years. The other 42 features are about vertex pairs and include measures such as the citation ratio between them. Detailed feature description are available in Impact4Cast link. The network features are inspired by the winner of the *Science4Cast competition*

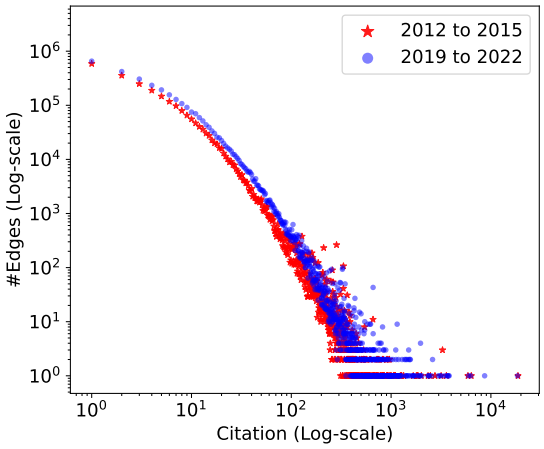

Figure 2. **Heavy-tail distribution of edge-citation increase rates.** Histogram of citation growth over a three-year period for concept pairs initially cited zero times in 2012 (red star, containing) and 2019 (blue dot).

(Lu, 2021; Krenn & Zeilinger, 2020), and the citation features are developed empirically and could potentially be improved by careful feature importance analysis. Our neural network is a fully connected feed-forward network with four hidden layers of 600 neurons each. The exploration of more advanced architectures might improve the prediction qualities further. The neural network has to predict whether the unconnected vertex pair in 2019 will have at least IR citations (IR stands for the impact range).

We perform the training for different values of the impact range IR from $IR = 1$ to $IR = 200$. We quantify the quality with the area under curve (AUC) of the receiver operating characteristic curve (ROC) (Fawcett, 2004). The AUC gives a measure of classification quality and stands for the probability that a randomly chosen true example is ranked higher than a randomly chosen false example. A random classifier has $AUC = 0.5$. We measure the AUC for a test set (which contains unconnected pairs not in the training set) for a prediction from 2016 to 2019, and for an evaluation dataset, with 10 million random data from 2019 to 2022 (while keeping the training data of the neural network from 2016 to 2019). The evaluation dataset shows how well the neural network performs on future, never-seen datasets. This is motivated by our goal that ultimately we want to train a neural network with all available data (let's say, until January 2023) and predict what happens until the future in 2026. In Fig. 5(a), we find that the AUC scores for both the test set and the evaluation set are beyond 0.8, in most of the cases beyond 0.9, for different IR. We can conclude that the neural network can forecast a high impact of previously never-investigated concept connections to a high degree. In Fig. 5(b), we sort the concept pairs of the evaluation dataset with the neural network ($IR = 100$), and plot their true citation counts. We further divide the 10 million evaluation dataset into 20 equal parts and plot their average citation count (represented by green bars) for each 5% segment. This clearly demonstrates good predictions at the individual concept pair level. As seen in Fig. 5(c), the highest predicted concept pairs indeed get more than 3 orders of magnitude more citations than the average citation of all 10 million pairs.

2.4. Forecasting genuine impact beyond link prediction

Next, we perform an even more challenging, genuine impact prediction task that goes beyond link prediction (i.e., predicting which concept pairs will be investigated in the future by a scientific paper). Concretely, in this task training data is conditioned on unconnected vertex pairs in 2016 which are actually connected in 2019. The neural network only gets citation information from 2016 and has to predict whether the newly generated concept pair will be highly impactful or not in the future. For that, our classification task asks whether the newly generated edge will receive citations within 0-5 or above 100 (Fig. 5(d)) in 2019. We see that the AUC score is beyond 0.7 (for the test set) and beyond 0.67 for the evaluation set, clearly indicating that the neural network can predict impact properties that go beyond the simple link-prediction task.

2.5. Highly predicted impact pair and potential applications

We can now investigate the largest predicted pairs of concepts, by taking all unconnected vertex pairs (∼694 million pairs) until 2023, and let the neural network (trained with all unconnected pairs in 2019 with supervision signal in 2022) sort them by impact predictions. We find that the highest predicted pair is *renewable energy* and *cancer cell*. This prediction is a very high-risk bet. For more practical, personalized suggestions, one can restrict the unconnected concept pairs to those related to specific scientists or research groups, aiming for high-impact collaboration suggestions. By examining the published works of scientists to identify their research interests, it becomes possible to identify concept pairs where one aligns with one scientist's specialty and the other with another scientist's. Thereby, one can suggest potential collaborations of high impact. As an example, by constraining the personalized research interests of scientists in experimental quantum optics and one researcher in biophysics, the highest predicted impact concepts pairs are '*microfluidic channel*' with '*Kerr resonator*', '*SARS CoV*' with '*quantum enhanced sensitivity*' or '*electron microscopy*' with '*quantum vacuum field*'. These suggestions can be further refined based on their similarity (e.g., represented by the cosine similarity) or the prominence of the concepts (indicated by the node degree), as we show in

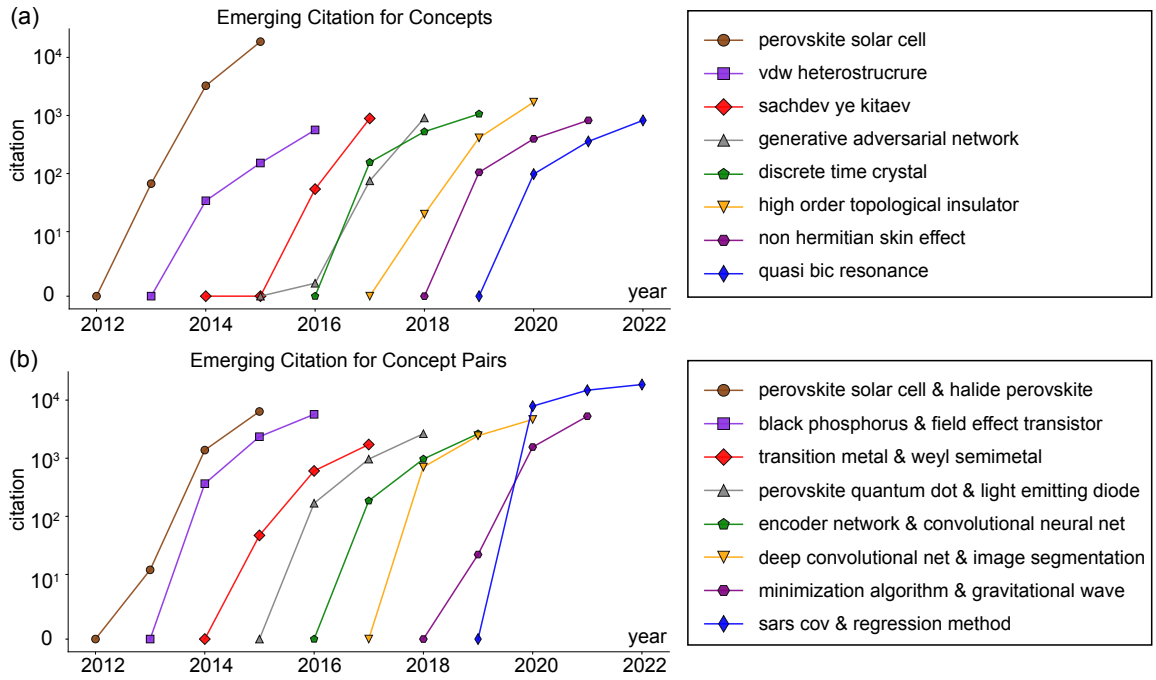

Figure 3. **Fastest growing citations of concepts and concept pairs**: Evolution of citations over three years for the top-fastest growing, previously uncited concepts (a) and concept pairs (b). We find many revolutionary topics in the realm of quantum physics and optics research in the last decade, including Perovskite devices (Rong et al., 2018), the emergence of complex and non-hermitian topology (Bergholtz et al., 2021), the introduction of advanced concepts of machine learning in physics (Carleo et al., 2019; Krenn et al., 2023b; Wang et al., 2023a) and quasi-BIC (bound state in continuum) resonances.

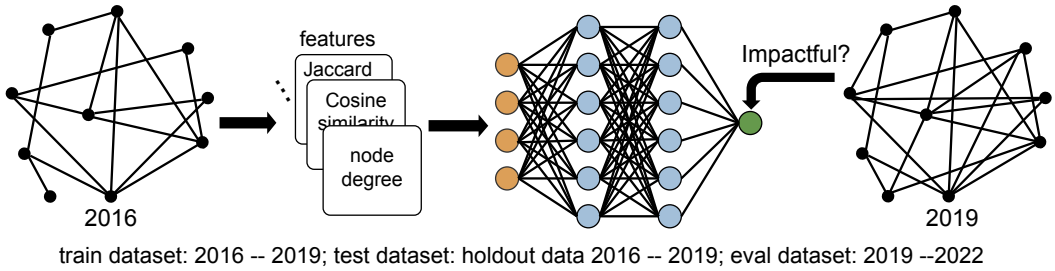

Figure 4. **Forecasting the impact of new research connections.** Network and citation features from unconnected vertex pairs from 2016 are used as input to a neural network. The citation information from 2019 is used as a supervision signal to train the neural network. After training, we evaluate the neural network's abilities by applying it to unconnected vertex pairs from 2019, aiming to predict the developments in 2022 – a task involving data the network has never encountered before.

Fig. 6. Here, we plot 100,000 concept pairs that have not been studied together until January 2023 and use the neural network trained on 2019 dataset to predict their impact. The points are plotted based on various properties, such as the similarity between concepts, their prominence within the network, their growth rate in the network (reflected in newly acquired neighbors), and how often the concepts have been cited previously. Plotting in this way allows us to identify rare outliers – concept pairs with high predicted impact that have unique properties, such as the bright yellow spots highlighted in the insets of Fig. 6. These methods help us narrow down the enormously large number of possibilities into a small number of personalized and targeted suggestions, which could inspire new ideas.

3. Discussion and Outlook

We show how to forecast the impact of future research topics. Although we view this as a significant step towards developing truly useful AI-driven assistants, achieving this goal requires numerous further advancements. Firstly, developing methods to extract more complex information from each

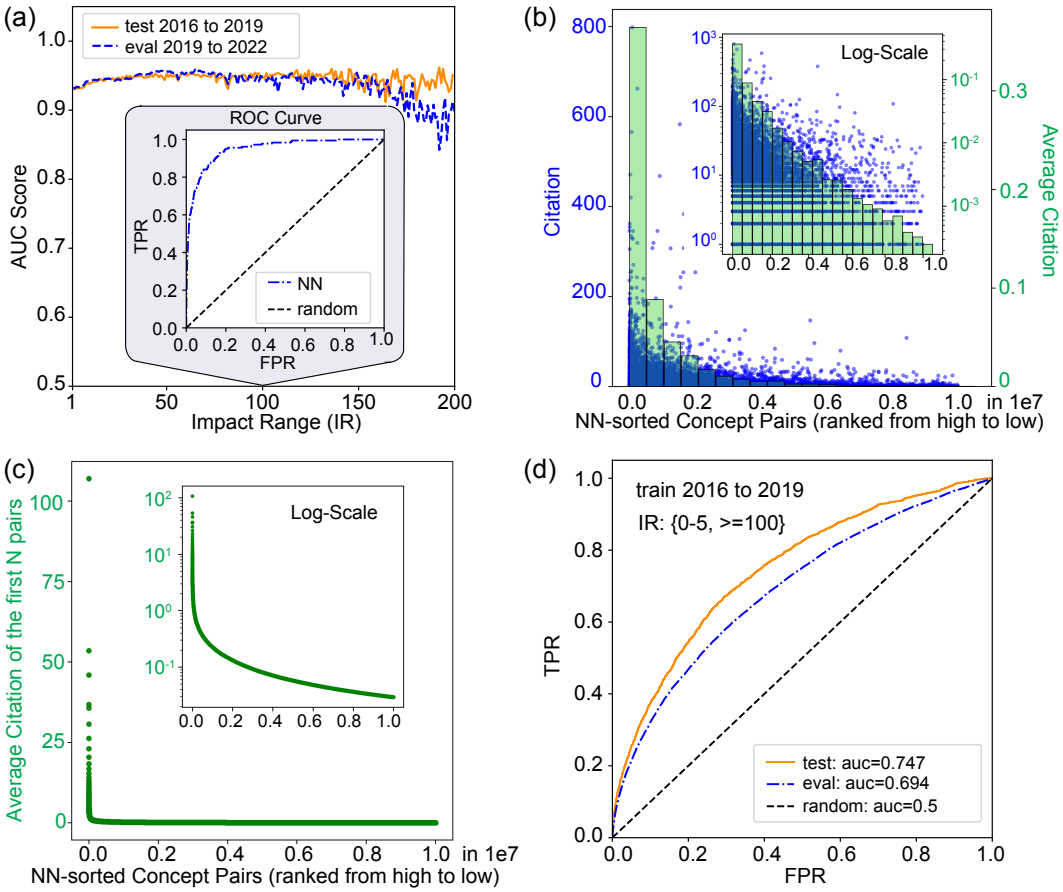

Figure 5. **Evaluating the machine-learning-based impact forecast.** (a): Classification of unconnected pairs, whether they will exceed a certain threshold three years later. Training data contains unconnected vertex pairs from 2016 and the supervision signal according to their impact range IR 3 years later. The test dataset also includes pairs from 2016, but excludes those in the training set. A more challenging evaluation set contains unconnected pairs from 2019, with outcomes verified in 2022, importantly noting that the neural network was only trained on data from 2016 to 2019, not 2019 to 2022. We quantify the quality using the area under curve of the receiver operating characteristic curve (labeled as AUC). For example, $IR = 100$, i.e, $(< 100, >= 100)$, refers to whether the 3-year citation counts after 2016 (test) or after 2019 (eval) is at least 100. TPR (true positive rate) measures how often a test correctly identifies a true positive, while FPR (false positive rate) measures how often it correctly identifies a true negative. (b): Sorted predictions of the neural network on the evaluation set (blue curve in (a)) shows the very high quality prediction at the level of individual concept pairs. The y-axis stands for the respective fraction of the evaluation dataset (10^7 data points). The histogram is separated into 20 equal bins. In (c), we show the average citation of the first N highest predicted concept pairs. This plot shows impressively that the highest predicted concept pairs indeed have very high citation, more than 3 orders of magnitude higher than the average citation of all 10^7 pairs (0.029 citations). (d): This more challenging step shows that citation prediction goes beyond link predictions. Here we take unconnected vertex pairs, conditioned on a connection 3 years later. The neural network is tasked to classify these concept pairs in low or high citations, revealing that it is not just predicting new links, but is learning intrinsic citation features. Here $IR = [5, 100]$, i.e, $(0 - 5, >= 100)$, means whether the 3-year citation count after 2016 (test) or after 2019 (eval) is at most 5 or at least 100.

paper will be crucial, for instance by employing hyper-graph structures that carry more information from each paper (Battiston et al., 2021; Belikov et al., 2022), which has already been demonstrated to lead to exciting results in other domains (Foster et al., 2015; Wang & Barabási, 2021; Sourati & Evans, 2023). This might also allow for the forecast of new concepts (Salatino et al., 2017; 2018) and their impact. Incorporating the recent dataset (Lin et al., 2023; Li et al., 2019) into our research could also allow us to explore more complex data structures than those used in our paper. Secondly, it will be interesting to approximate impact with metrics that go beyond citations – which is a crucial topic in computational sociology and the study of the science of science (Fortunato et al., 2018; Wang & Barabási, 2021). Additionally, introducing metrics of surprise, as discussed in (Foster et al., 2021; Shi & Evans, 2023), could serve as a complementary metric to citation prediction for ranking suggestions. Finally, while the suggestion of *impactful* new

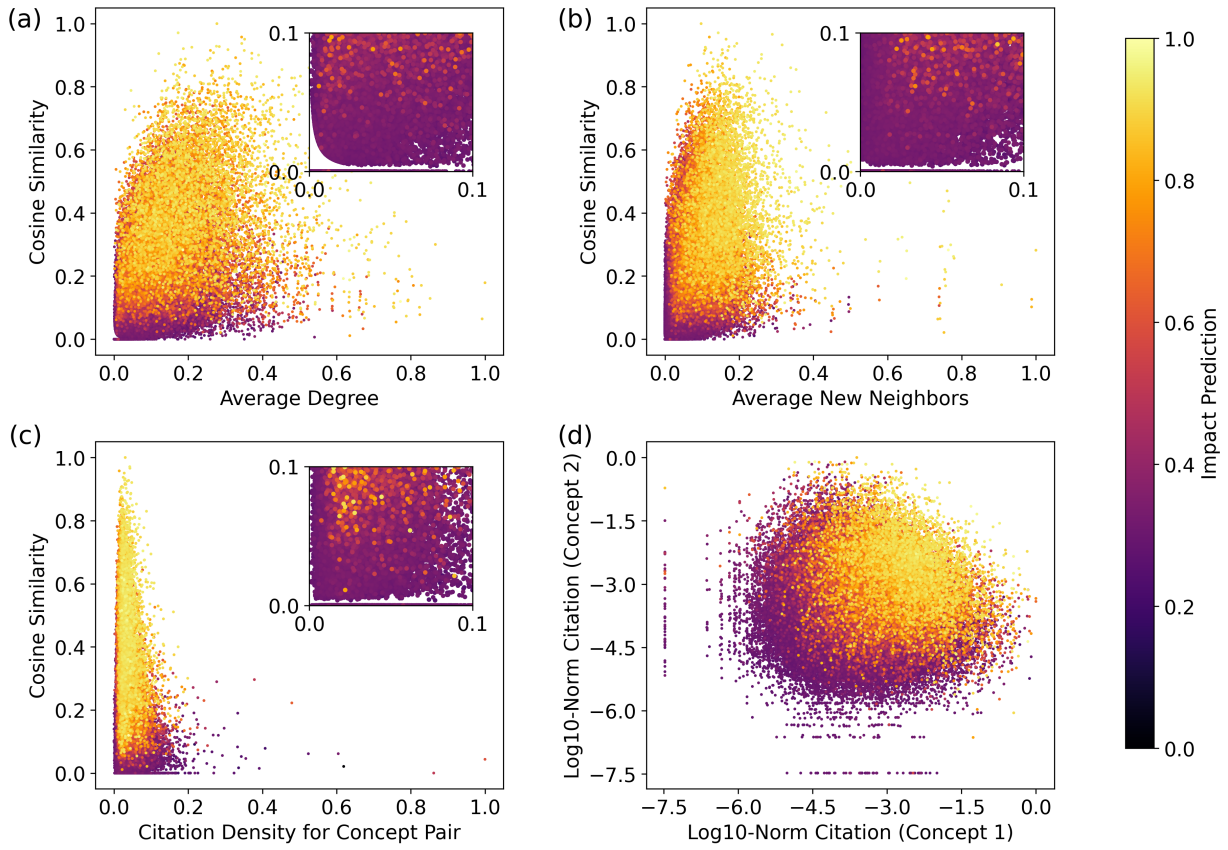

Figure 6. **Network features vs. predicted impact.** A randomly selected set of 100,000 unconnected concept pairs until January 2023 is used. The color represents the neural network's prediction of each concept pair's impact. (a): The y-axis shows cosine similarity, indicating the semantic similarity between concepts; lower values represent concept pairs that are semantically distinct. The x-axis is the average vertex degree of the two concepts in the knowledge graph, reflecting their overall prominence. Concepts with low similarity and low degree yet predicted to have high impact could be surprising and offer interesting suggestions. (b): The x-axis represents the average number of new neighbors each concept gained over the last three years. Concept pairs with low similarity and few new neighbors but high impact predictions might highlight potentially overlooked but intriguing ideas. (c): The x-axis denotes *citation density* (average citations per paper mentioning the concepts). Pairs with low similarity and citation density but high predicted impact could again indicate overlooked potential ideas. (d): Citation counts for concept 1 (x-axis) and concept 2 (y-axis) over last three years are plotted on a logarithmic scale. We can easily identify concept pairs predicted to have high impact in the future, even though they have individually received few citations in the past.

ideas might be a key component of future AI assistants, it will be crucial to study its relation to the *scientific interest* of working researchers.

Software and Data

Codes are available at https://github.com/artificial-scientist-lab/Impact4Cast. Data are accessible at https://doi.org/10.5281/zenodo.10692137 (Gu, 2024).

Acknowledgements

The authors thank Burak Gurlek for interesting discussions at the start of this project, and the organizers of OpenAlex, arXiv, bioRxiv, and medRxiv for making scientific resources freely and readily accessible. X.G acknowledges the support from the Alexander von Humboldt Foundation.

Impact Statement

This paper presents work whose goal is to accelerate the progress of scientific discovery. There are no specific ethical or societal impacts that needs to be considered.

References

Challenge the impact factor. *Nature Biomedical Engineering*, 1:0103, 2017.

The many facets of impact. *Nature Reviews Physics*, 6:71, 2024.

Abbott, B. P., Abbott, R., Abbott, T. D., Acernese, F., Ackley, K., Adams, C., Adams, T., Addesso, P., Adhikari, R. X., Adya, V. B., et al. Constraints on cosmic strings using data from the first advanced ligo observing run. *Phys. Rev. D*, 97:102002, May 2018. doi: 10.1103/PhysRevD.97.102002.

Bai, X., Liu, H., Zhang, F., Ning, Z., Kong, X., Lee, I., and Xia, F. An overview on evaluating and predicting scholarly article impact. *Information*, 8(3):73, 2017.

Barabási, A.-L. *The formula: The universal laws of success*. Hachette UK, 2018.

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

Battiston, F., Amico, E., Barrat, A., Bianconi, G., Ferraz de Arruda, G., Franceschiello, B., Iacopini, I., Kéfi, S., Latora, V., Moreno, Y., et al. The physics of higher-order interactions in complex systems. *Nature Physics*, 17(10): 1093–1098, 2021.

Belikov, A. V., Rzhetsky, A., and Evans, J. Prediction of robust scientific facts from literature. *Nature Machine Intelligence*, 4(5):445–454, 2022.

Bergholtz, E. J., Budich, J. C., and Kunst, F. K. Exceptional topology of non-hermitian systems. *Rev. Mod. Phys.*, 93 (1):015005, 2021.

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.

Carleo, G., Cirac, I., Cranmer, K., Daudet, L., Schuld, M., Tishby, N., Vogt-Maranto, L., and Zdeborová, L. Machine learning and the physical sciences. *Rev. Mod. Phys.*, 91 (4):045002, 2019.

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

Fontana, M., Iori, M., Montobbio, F., and Sinatra, R. New and atypical combinations: An assessment of novelty and interdisciplinarity. *Research Policy*, 49(7):104063, 2020.

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.

Foster, J. G., Rzhetsky, A., and Evans, J. A. Tradition and innovation in scientists' research strategies. *American Sociological Review*, 80(5):875–908, 2015.

Foster, J. G., Shi, F., and Evans, J. Surprise! measuring novelty as expectation violation. *SocArXiv 2t46f*, 2021.

Frank, M. R., Wang, D., Cebrian, M., and Rahwan, I. The evolution of citation graphs in artificial intelligence research. *Nature Machine Intelligence*, 1(2):79–85, 2019.

Fu, L. and Aliferis, C. Using content-based and bibliometric features for machine learning models to predict citation counts in the biomedical literature. *Scientometrics*, 85(1): 257–270, 2010.

Google. Gemini: a family of highly capable multimodal models. *arXiv:2312.11805*, 2023.

Gu, X. Impact4cast: Forecasting high-impact research topics via machine learning on evolving knowledge graphs [data set], zenodo. *https://doi.org/10.5281/zenodo.10692137*, 2024.

He, G., Xue, Z., Jiang, Z., Kang, Y., Zhao, S., and Lu, W. H2cgl: Modeling dynamics of citation network for impact prediction. *arXiv:2305.01572*, 2023.

Ke, Q., Ferrara, E., Radicchi, F., and Flammini, A. Defining and identifying sleeping beauties in science. *Proc. Natl. Acad. Sci. USA*, 112(24):7426–7431, 2015.

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, 2023a.

Krenn, M., Landgraf, J., Foesel, T., and Marquardt, F. Artificial intelligence and machine learning for quantum technologies. *Phys. Rev. A*, 107(1):010101, 2023b. doi: 10.1103/PhysRevA.107.010101.

Leeuwenhoek, A. V. Ii. microscopical observations on the blood vessels and membranes of the intestines. in a letter to the royal society from mr. anthony van leeuwenhoek,

frs. *Philosophical Transactions of the Royal Society of London*, 26(314):53–58, 1709.

Li, J., Yin, Y., Fortunato, S., and Wang, D. A dataset of publication records for nobel laureates. *Scientific Data*, 6 (1):33, 2019.

Lin, Z., Yin, Y., Liu, L., and Wang, D. Sciscinet: A large-scale open data lake for the science of science research. *Scientific Data*, 10(1):315, 2023.

Lu, Y. Predicting research trends in artificial intelligence with gradient boosting decision trees and time-aware graph neural networks. In *2021 IEEE International Conference on Big Data (Big Data)*, pp. 5809–5814. IEEE, 2021.

Martínez, V., Berzal, F., and Cubero, J.-C. A survey of link prediction in complex networks. *ACM computing surveys (CSUR)*, 49(4):1–33, 2016.

Min, C., Bu, Y., and Sun, J. Predicting scientific breakthroughs based on knowledge structure variations. *Technological Forecasting and Social Change*, 164:120502, 2021.

Möckl, L., Roy, A. R., Petrov, P. N., and Moerner, W. Accurate and rapid background estimation in single-molecule localization microscopy using the deep neural network bgnet. *Proc. Natl. Acad. Sci. USA*, 117(1):60–67, 2020.

OpenAI. Gpt-4 technical report. *arXiv:2303.08774*, 2023.

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.

Rong, Y., Hu, Y., Mei, A., Tan, H., Saidaminov, M. I., Seok, S. I., McGehee, M. D., Sargent, E. H., and Han, H. Challenges for commercializing perovskite solar cells. *Science*, 361(6408):eaat8235, 2018.

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

Ruan, X., Zhu, Y., Li, J., and Cheng, Y. Predicting the citation counts of individual papers via a bp neural network. *Journal of Informetrics*, 14(3):101039, 2020.

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.

Salatino, A. A., Osborne, F., and Motta, E. How are topics born? understanding the research dynamics preceding the emergence of new areas. *PeerJ Computer Science*, 3: e119, 2017.

Salatino, A. A., Osborne, F., and Motta, E. Augur: forecasting the emergence of new research topics. In *Proceedings of the 18th ACM/IEEE on joint conference on digital libraries*, pp. 303–312, 2018.

Salatino, A. A., Osborne, F., Thanapalasingam, T., and Motta, E. The cso classifier: Ontology-driven detection of research topics in scholarly articles. In *Digital Libraries for Open Knowledge: 23rd International Conference on Theory and Practice of Digital Libraries, TPDL 2019, Oslo, Norway, September 9-12, 2019, Proceedings 23*, pp. 296–311. Springer, 2019.

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.

Sinatra, R., Wang, D., Deville, P., Song, C., and Barabási, A.-L. Quantifying the evolution of individual scientific impact. *Science*, 354(6312):aaf5239, 2016.

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

Stegehuis, C., Litvak, N., and Waltman, L. Predicting the long-term citation impact of recent publications. *Journal of informetrics*, 9(3):642–657, 2015.

Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv:2307.09288*, 2023.

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

van Nieuwenburg, E., Bairey, E., and Refael, G. Learning phase transitions from dynamics. *Phys. Rev. B*, 98: 060301, Aug 2018. doi: 10.1103/PhysRevB.98.060301.

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

Wang, D., Song, C., and Barabási, A.-L. Quantifying long-term scientific impact. *Science*, 342(6154):127–132, 2013.

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, 2023a.

Wang, Q., Downey, D., Ji, H., and Hope, T. Learning to generate novel scientific directions with contextualized literature-based discovery. *arXiv:2305.14259*, 2023b.

Weihs, L. and Etzioni, O. Learning to predict citation-based impact measures. In *2017 ACM/IEEE joint conference on digital libraries (JCDL)*, pp. 1–10. IEEE, 2017.

Weis, J. W. and Jacobson, J. M. Learning on knowledge graph dynamics provides an early warning of impactful research. *Nature Biotechnology*, 39(10):1300–1307, 2021.

Williamson, D. J., Burn, G. L., Simoncelli, S., Griffié, J., Peters, R., Davis, D. M., and Owen, D. M. Machine learning for cluster analysis of localization microscopy data. *Nature communications*, 11(1):1493, 2020.

Wu, L., Wang, D., and Evans, J. A. Large teams develop and small teams disrupt science and technology. *Nature*, 566(7744):378–382, 2019.

Xia, W., Li, T., and Li, C. A review of scientific impact prediction: tasks, features and methods. *Scientometrics*, 128(1):543–585, 2023.

Yu, T., Yu, G., Li, P.-Y., and Wang, L. Citation impact prediction for scientific papers using stepwise regression analysis. *Scientometrics*, 101:1233–1252, 2014.

A. Datasets for knowledge graph construction

To compile a list of scientific concepts in natural science, we used metadata from four major preprint servers: arXiv, bioRxiv, medRxiv, and chemRxiv. The arXiv dataset can be directly downloaded from Kaggle, while metadata from bioRxiv, medRxiv, and chemRxiv are accessible through their APIs. The full methodology and codes are available on the Supplementary Material. Our comprehensive dataset encompasses ∼2.44 million papers, including 78,084 from arXiv's physics.optics and quant-ph categories, which were specifically used for identifying domain concepts.

For edge generation, we used the OpenAlex database snapshot, available for download in OpenAlex bucket. More details can be found at the OpenAlex documentation site. The complete dataset occupies around 330 GB, expanding to approximately 1.6 TB when decompressed. Our interest was specifically in scientific journal papers that include publication time, title, abstract, and citation information. By focusing on these criteria, we managed to reduce the dataset to a more manageable gzip-compressed size of 68 GB, comprising around 92 million scientific papers.

B. Details on concept and edge generation

From the preprint dataset of ∼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. In addition, we conduct a manual review to identify and eliminate any inaccuracies in the concepts. The entire process, which included eliminating non-conceptual phrases, verbs, ordinal numbers, conjunctions, and adverbials, resulted in a full list of 368,825 concepts.

We then specifically focused on articles within the physics.optics and quant-ph categories from arXiv to extract domain-specific concepts. Iterating this entire list of concepts to these domain-specific articles, we identified 87,741 relevant concepts. Employing our specially designed automated filtering tool for initial refinement and then conducting a thorough manual review to remove inaccuracies, we narrowed the list down to 37,960 high-quality, domain-specific concepts.

As an example, we show the extraction of concepts for the four papers used in Fig. 1:

1. Accurate and rapid background estimation in single-molecule localization microscopy using the deep neural network BGnet (Möckl et al., 2020): 'super resolution reconstruction', 'neural network', 'single molecule tracking', 'deep neural net', 'deep neural network', 'localization microscopy', 'biological structure', 'point source', 'point spread function', 'single molecule localization microscopy', 'optical microscopy', and 'neural net'.

2. Machine learning for cluster analysis of localization microscopy data (Williamson et al., 2020): 'neural network', 'supervised machine learning', 'spatial relation', 'localization microscopy', 'single molecule localization microscopy', 'neural net', and 'machine learning'.

3. Constraints on cosmic strings using data from the first Advanced LIGO observing run (Abbott et al., 2018): 'phase transition', 'cosmic string', 'gravitational wave', 'cosmic microwave', 'cosmic microwave background', 'topological defect', 'string theory', and 'ring theory'.

4. Learning phase transitions from dynamics (van Nieuwenburg et al., 2018): 'neural network', 'recurrent network', 'time crystalline phase', 'phase transition', 'localization transition', 'spin chain', 'recurrent neural net', 'ct model', 'recurrent neural network', 'phase diagram', 'crystalline phase', 'neural net', and 'recurrent net'.

We created concept pairs, or edges, from the OpenAlex dataset, by detecting when domain-specific concepts co-occurred in paper titles or abstracts. This yielded 193,977,096 concept pairs (including multi-edges) across about 21 million papers. Each edge receives a time-stamp based on its paper's publication date, converted to the number of days since January 1, 1990. The final full knowledge graph comprises 26,010,946 unique edges after merging multiple edges between the same concept pairs. The citation information for an edge includes the paper's yearly citations from 2012 to 2023, alongside its total citation since publication. The OpenAlex dataset excludes yearly citations older than ten years, hence the focus on this specific ten-year time frame due to the absence of data prior to 2012. For edges formed by multiple papers, the edge weight combines the annual and total citations from all contributing papers.

Consider the edge formed by the concepts 'single molecule localization microscopy' and 'neural net', generated from paper $p1$ (Möckl et al., 2020) published on January 7, 2020. The time-stamp for this edge is derived from the days elapsed since January 1, 1990. The citation metrics for this edge includes the total and yearly citations

from each contributing paper. Paper $p1$ with 38 citations (c_{p1}=38), contributes yearly citations represented as $c_{p1}(y_i)$ for i=2023,2022,..., 2012, with actual values $\{5, 8, 16, 9\}$ for 2023 to 2020, and zeros for previous years, culminating in a citation sequence $\{5, 8, 16, 9, 0, 0, 0, 0, 0, 0, 0, 0\}$. Similarly, paper $p2$ (Williamson et al., 2020), published on March 20, 2020, with 43 citations (c_{p2}=43), adds its yearly citations $\{6, 16, 14, 7\}$ for the same period (2023 to 2020). The aggregated citation data for this edge, combining c_{p1} and c_{p2}, yields a total of 81 citations, with an annual citation sequence of $\{11, 24, 30, 16, 0, 0, 0, 0, 0, 0, 0, 0\}$.

C. Training

Our neural network consists of six fully connected layers, which include four hidden layers with 600 neurons each. The network inputs are 141 features for each unconnected concept pair (v_i, v_j), denoted as $p_{i,j} = (p^1_{i,j}, p^2_{i,j}, ..., p^{141}_{i,j})$, where each $p_{i,j} \in \mathbb{R}$. For instance, $p^1_{i,j}$ and $p^2_{i,j}$ represent the vertex degree of concepts v_i and v_j for the current year, y. Detailed feature description and feature generation code are available in GitHub: Impact4Cast.

In our training process for the year 2016 to predict impact in 2019, we prepared a dataset comprising approximately 689 million unconnected concept pairs. The goal was to evaluate these pairs to determine whether their 3-year citation counts would have at least IR citations (IR is impact range) or not. From this extensive collection, we selected all positive samples (the 3-year citation counts are at least IR). An equivalent number of negative samples were then randomly chosen to match the size of the positive set. The refined dataset was subsequently divided, allocating 85% for training and 15% for testing purposes. For the evaluation dataset in 2019, which aims to predict the impact in 2022, we randomly selected 10 million unconnected pairs. Our neural network was trained using the Adam optimizer with a learning rate of 3×10^{-5} and a mini-batch size of 1000. In every training batch, we randomly chose an equal number of positive and negative samples from the training set. This approach was also applied to our 2019 training process for predictions into 2022, where the trained neural network is used for future forecasting.

The full dynamic knowledge graph, alongside the data required for feature preparation and evaluation, was processed using an Intel Xeon Gold 6130 CPU with 1 TiB RAM. The network training was performed on a NVIDIA Quadro RTX 6000 GPU, with 187 GiB RAM.

D. Individual feature's predictive ability

In Fig. 5 (a) (main text), we observe an AUC score of 0.948 for the 2019 evaluation dataset with the neural network that uses all 141 features, trained on 2016 dataset and impact range $IR = 100$. To explore the predictive ability of individual features, we trained separate neural networks on each feature using the same 2016 dataset, and then applied the 2019 evaluation dataset to these models. This resulted in 141 individual predictions, each from a network trained on a single feature. The features were ranked by their impact predictions, shown in the Fig. 7.

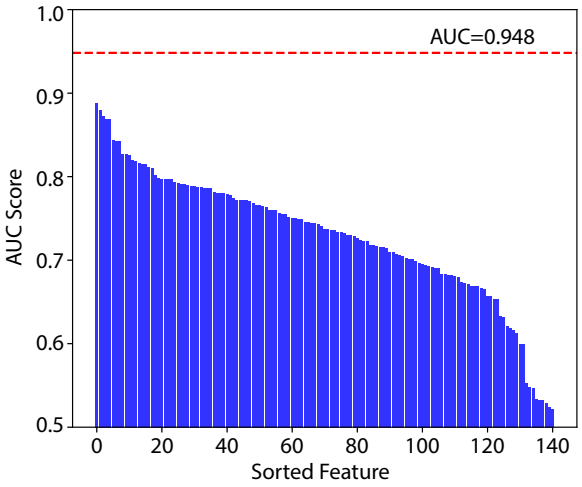

Figure 7. **Neural network performance across individual features.** The highest-performing four features are the Simpson similarity coefficient for the unconnected pair (u, v) across the years y, $y-1$, and $y-2$, and the cosine similarity coefficient for unconnected pairs (u, v) in year y (i.e., y=2016), with AUC scores of 0.8880, 0.8795, 0.8720, and 0.8683, respectively. In contrast, the lowest predictive three features are the average total citation count up to year y for vertex v, and the total citation count for the pair (u, v) up to years $y-1$ and $y-2$, with AUC scores of 0.5219, 0.5234, and 0.5285. Using all 141 features together leads to a significant improvement in the AUC score to 0.948, showing that the combination of all features works better.