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

Xuemei Gu[1,*] and Mario Krenn[1,†]
[1]Max Planck Institute for the Science of Light, Staudtstrasse 2, 91058 Erlangen, Germany

The exponential growth in scientific publications poses a severe challenge for human researchers. It forces attention to more narrow sub-fields, which makes it challenging to discover new impactful research ideas and collaborations outside one's own field. While there are ways to predict a scientific paper's future citation counts, they need the research to be finished and the paper written, usually assessing impact long after the idea was 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. It 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.

Introduction

As we see an explosion in the number of scientific articles [1–4], 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 (LLM) such as GPT-4 [5], Gemini [6] or LLaMA-2 [7] or custom-made models [8]. 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 [2]. While this approach extracts only small amounts of information from each paper, surprisingly non-trivial conclusions can still be drawn if the underlying dataset of papers is large. An early example of this is a work in biochemistry [9]. 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 link-prediction problem in network theory [10], has been successfully demonstrated with high prediction quality for semantic networks in quantum theory [11] and artificial intelligence [4].

Novelty plays an essential role in scientific ideas, but being novel doesn't automatically mean that an idea will have a high impact. Impact in the scientific community is often approximated (for lack of better metrics [12, 13]) by citations [1, 2, 14, 15], including exciting results that find interpretable mathematical models to describe citation evolution [16–19]. 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 [20], the content and the references of the paper [21, 22]. Techniques employed for the predictions of individual paper impact using a combination of characteristics include support-vector machines [23], regression [24–26], dense [27] or graph neural networks [28].

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 assistent 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?*.

* xuemei.gu@mpl.mpg.de
† mario.krenn@mpl.mpg.de

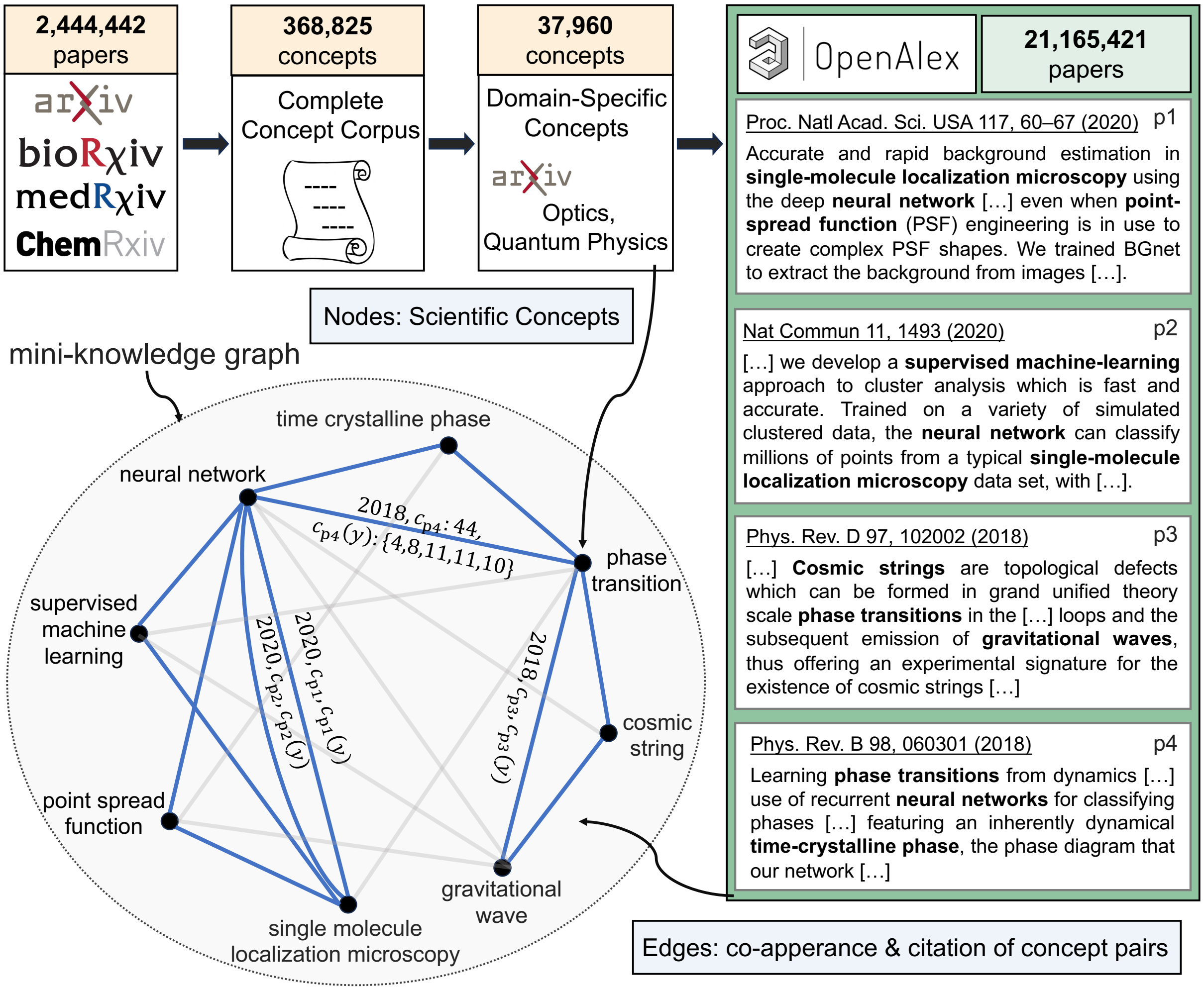

Fig. 1. **Generation of the knowledge graph with time and citation information**. Vertices are formed by scientific concepts, which are extracted from scientific papers (titles and abstracts) from prominent academic preprint servers. Edges are formed when concepts are investigated jointly in a scientific 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, which acts as a proxy for impact in our work. A mini-knowledge graph (blue edges) is constructed from four randomly selected papers (p1-p4) [29–32] from OpenAlex as an example. Here, c_{p4} represents the total citations of paper p4 since its publication, and $c_{p4}(y)$ denotes its annual citations from 2018 to 2022. The citation value of the edge is the sum of the all papers creating the edge.

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 [33]) 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 [34].

The Knowledge Graph

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 [40], for several reasons. Firstly, our goal is to ultimately cover

[1] GitHub: Impact4Cast

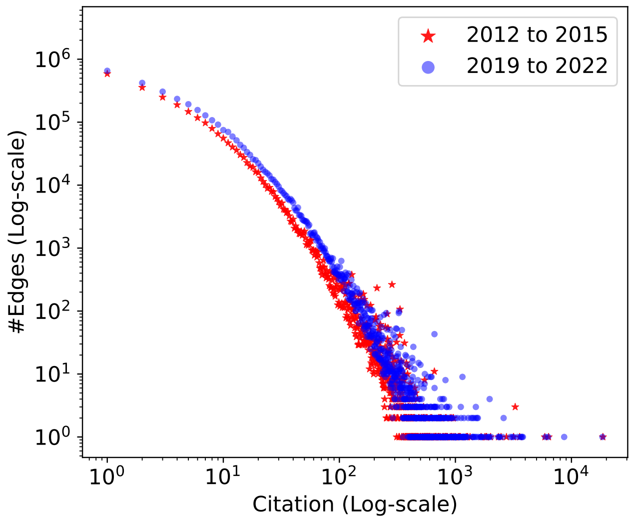

Fig. 2. **Heavy-tail distribution of edge-citation rates.** Histogram of citation growth over a three-year period for concept pairs initially cited zero times in 2012 (red, containing) and 2019 (blue).

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 included 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 [41]. 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 ratio of high-quality concepts. To refine our list further, we developed a suite of natural language processing tools. Finally, we got a list which contains over 368,000 concepts. While we focus here on concepts specific to the sub-field of optics and quantum physics (representing roughly 10% of the entire concepts), our method can immediately be translated to any other domain. This refined domain-specific concept list serves as the vertices for our knowledge graph.

Creating an evolving, citation-augmented knowledge graph – Now that we have the vertices, we can create edges that contain information from the scientific literature. To have citation information, we use the works from OpenAlex [42], an open-source database containing detailed information on more than 90 million scientific 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 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 database, 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, potentially due to the influence of concept hubs [4]. 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.

Results

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 3 years later this vertex 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 [43] over the past three years) and 21 edge features (including cosine, geometric, and Simpson similarities [44]). Additionally, we incorporated 100 impact features: 58 of these are node citation features, covering total citations and the count of papers mentioning the concept within the last three years. The other 42 features are about vertex pairs and include measures such as the citation ratio between them. The network feature is inspired by the winner of the *Science4Cast competition* [11, 45], 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

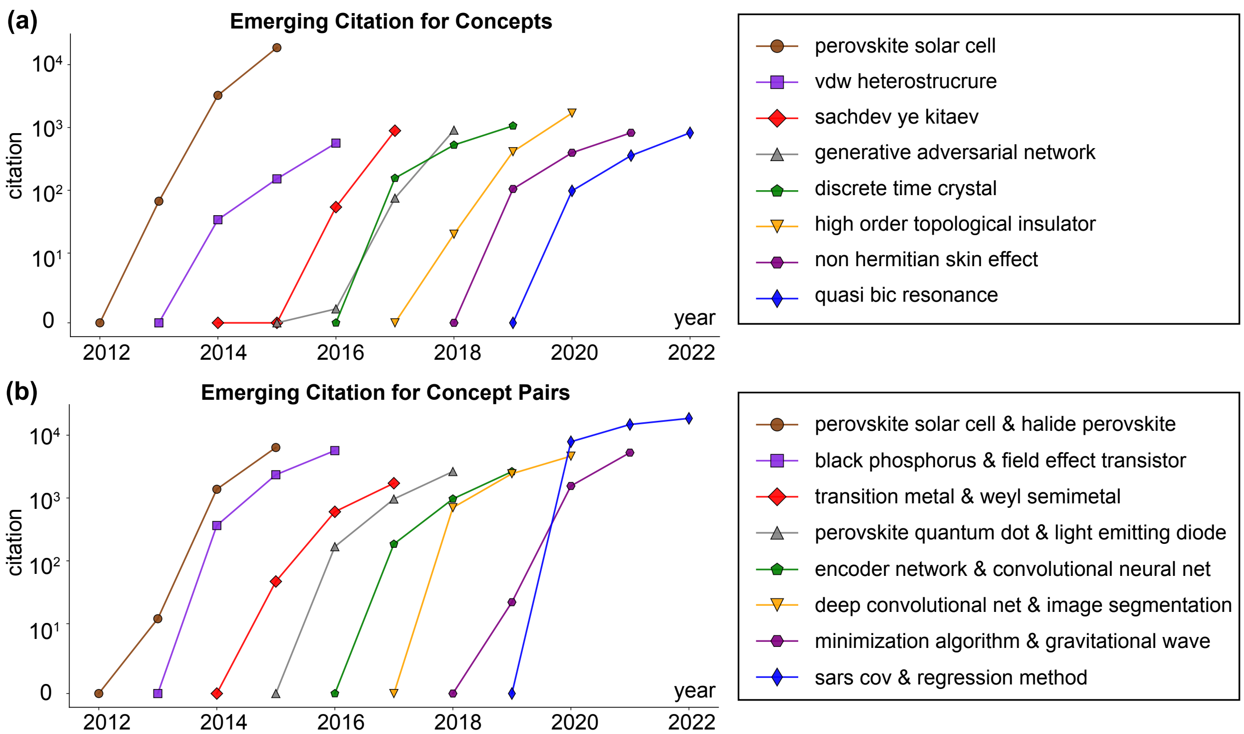

Fig. 3. **Fastest growing citations of concepts and concept pairs**: Evolution of citations over three years for the top-fastest growing, previously uncited concepts (a) and concept pairs (b). We find many revolutionary topics in the realm of quantum physics and optics research in the last decade, including Perovskite devices [35], the emergence of complex and non-hermitian topology [36], the introduction of advanced concepts of machine learning in physics [37–39] and quasi-BIC (bound state in continuum) resonances.

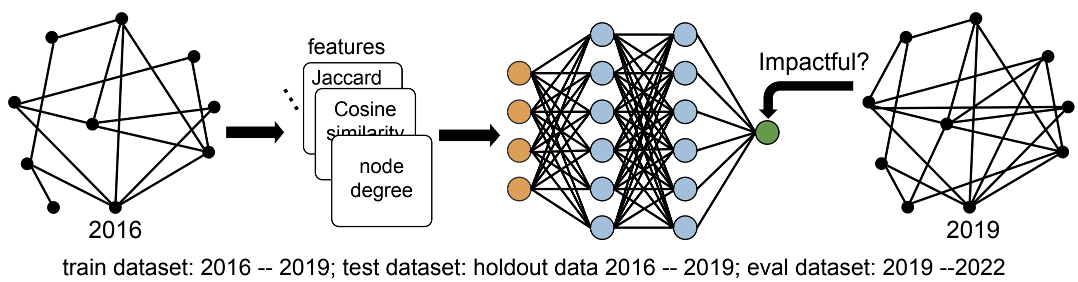

Fig. 4. **Forecasting the impact of new research connections.** Network and citation features from unconnected vertex pairs as of 2016 are used as input to a neural network. The citation information from 2019 is used as a supervision signal to train the neural network. After training, we evaluate the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

predict whether the unconnected vertex pair in 2019 will have more than IR citations (IR is impact range).

We perform the training from $IR = 1$ to $IR = 200$. We quantify the quality with the area under curve (AUC) of the receiver operating characteristic curve (ROC) [46]. 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 the year 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 the end of 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 values of the impact range IR. We can conclude that the neural network can forecast a high impact of previously

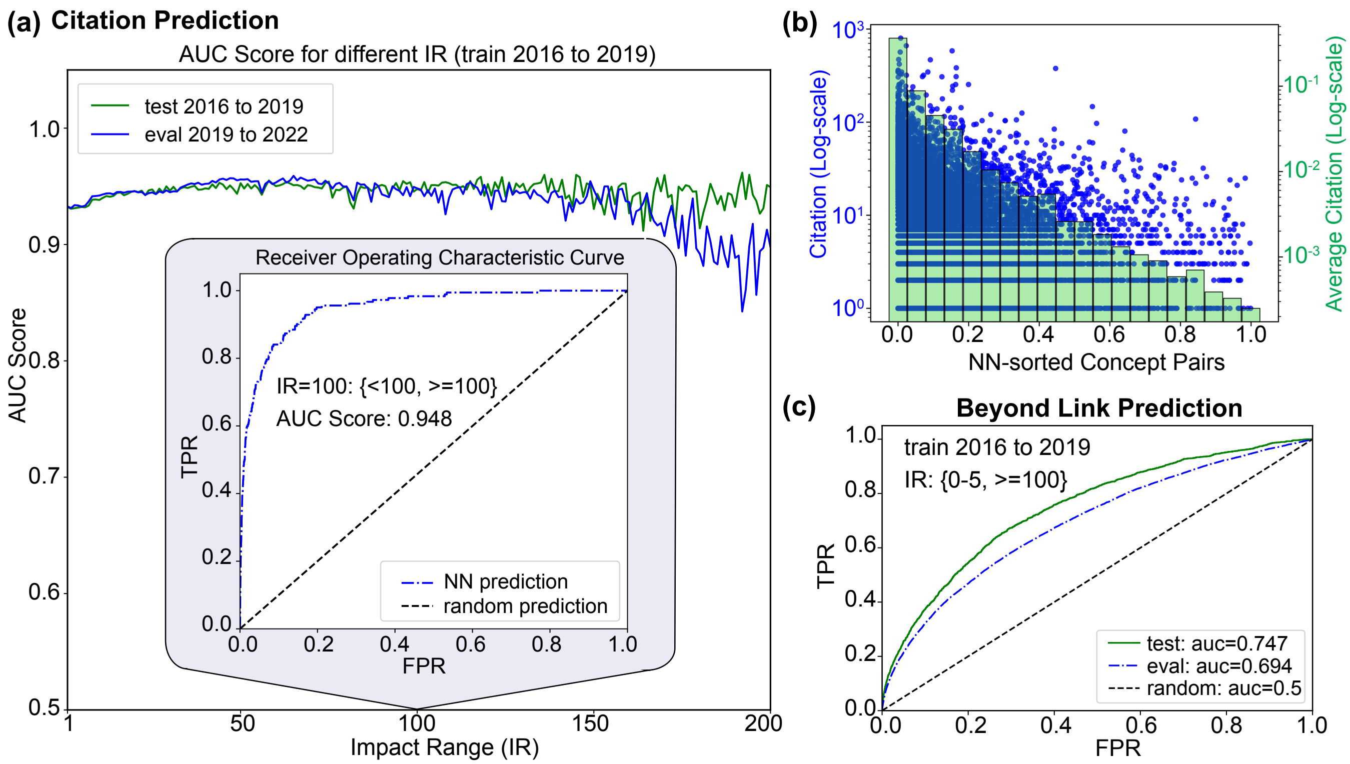

Fig. 5. **Evaluating the machine-learning-based impact forecast.** (a): Classification of unconnected concept pairs, whether they will have more than certain citations 3 years later. We quantify the quality using the area under the receiver operating characteristic curve (labeled as AUC) for different impact ranges (IR). For example, $IR = 100$, i.e, $\{< 100, >= 100\}$, refers to whether the citation count accumulated in three years after 2016 (test) or after 2019 (eval) is at least 100. (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 parts. (c): This significantly 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 links, but is learning intrinsic citation features. Training data contains unconnected vertex pairs from 2016 and the supervision signal of its IR 3 years later. Test data also contains unconnected vertex pairs from 2016, but only those not contained in the training dataset. The evaluation dataset (which is a more challenging test set) contains unconnected vertex pairs from the year 2019, with ground truth from 2022 (which is, importantly, not used during the training). In (c), $IR = [5, 100]$, i.e, $\{0 - 5, >= 100\}$, meaning whether the citation count accumulated in the three years after 2016 (test) or after 2019 (eval) is at most 5 or at least 100.

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 very good predictions at the individual concept pair level.

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 second task training data is conditioned on unconnected vertex pairs from 2016 which are actually connected in 2019. The neural network only gets information from 2016 and has to predict whether the newly generated concept pair will be highly impactful or not. For that, our classification task asks whether the newly generated edge will receive between 0-5 citations or above 100 (Fig. 5(c)). We see that the AUC 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 simpler link-prediction task.

Extracting highly impactful cliques of concepts – In what follows, we show one way how our results can be applied to inspire new directions that potentially human researchers would not have thought about. So far, we have limited our predictions to pairs of concepts. However, incorporating a larger number of concepts could more directly indicate a research direction. Directly predicting high-impact concept triples or quadruples is either highly computationally expensive or necessitates en-

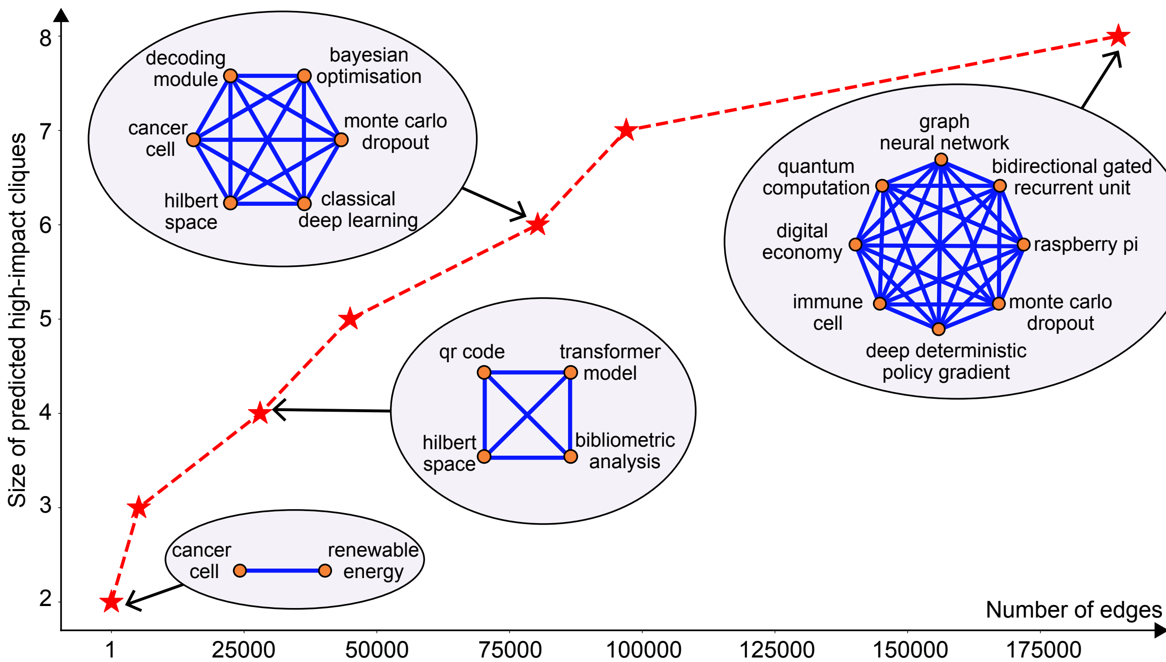

Fig. 6. **Clusters of concepts with high predicted impact.** We demonstrate the smallest cliques (fully connected graphs) of concepts that are pairwise highly predicted impactful. The x-axis is the size of the subgraph with the N highest predicted edges, while the y-axis is the size of the clique (for impact range $IR = 10$). For clarity, we show the cliques for $N = 2, 4, 6$, and 8, while cliques for $N = 3, 5$, and 7 are (*sars cov, hibert space, dilated convolution*), (*deep deterministic policy gradient, quantum circuit, cancer cell, digital economy, convolutional operation*), and (*bidirectional recurrent unit, monte carlo dropout, classical deep learning, hibert space, cancer cell, decoding module*), respectively.

tirely different data structures, such as hypergraphs. Alternatively, we can approximate multi-concept combinations in our knowledge graph by searching high-impact cliques. Cliques are fully connected subgraphs with N vertices. If $N = 2$, we have concept pairs, but for higher N, we get a larger number of concepts.

To get the cliques, we train a neural network to predict high-impact (for example, $IR = 10$) with all available data and predict what currently unconnected pairs will be highly impactful until 2026. We then apply the neural network to all unconnected vertex pairs (694 million pairs) and sort the result from highly likely to be of high impact, to least likely. With this sorted edge list, we create a subgraph, starting from the highest predicted edge (*cancer cell* with *renewable energy*) and add edges one by one. In Fig. 6, we collect the first cliques of size N ranging from 2 to 8. We eagerly await the year 2026 to see whether a paper investigating the concepts *Hilbert space, QR code, transformer model* and *bibliometric analysis* will indeed be highly impactful. In a real application scenario, one could personalize the list of concepts to fit the research interest of individual scientists.

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 paper will be crucial, for instance by employing hyper-graph structures that carry more information from each paper [47, 48], which has already been demonstrated to lead to exciting results in other domains[2, 49, 50]. This might also allow for the forecast of new concepts [51, 52] and their impact. 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 [1, 2]. Additionally, introducing metrics of surprise, as discussed in [53, 54], could serve as a complementary metric to citation prediction for ranking suggestions. Finally, while the suggestion of *impactful* new ideas might be a crucial component of future AI assistants, it will be crucial to study its relation to the *scientific interest* of working researchers.

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

[2] D. Wang and A.-L. Barabási, *The science of science* (Cambridge University Press, 2021).

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

[4] M. Krenn, L. Buffoni, B. Coutinho, S. Eppel, J. G. Foster, A. Gritsevskiy, H. Lee, Y. Lu, J. P. Moutinho, N. Sanjabi, *et al.*, Forecasting the future of artificial intelligence with machine learning-based link prediction in an exponentially growing knowledge network, Nature Machine Intelligence **5**, 1326 (2023).

[5] OpenAI, Gpt-4 technical report, arXiv:2303.08774 (2023).

[6] Google, Gemini: a family of highly capable multimodal models, arXiv:2312.11805 (2023).

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

[8] Q. Wang, D. Downey, H. Ji, and T. Hope, Learning to generate novel scientific directions with contextualized literature-based discovery, arXiv:2305.14259 (2023).

[9] A. Rzhetsky, J. G. Foster, I. T. Foster, and J. A. Evans, Choosing experiments to accelerate collective discovery, Proceedings of the National Academy of Sciences **112**, 14569 (2015).

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

[11] M. Krenn and A. Zeilinger, Predicting research trends with semantic and neural networks with an application in quantum physics, Proceedings of the National Academy of Sciences **117**, 1910 (2020).

[12] Challenge the impact factor, Nature Biomedical Engineering **1**, 0103 (2017).

[13] The many facets of impact, Nature Reviews Physics **6**, 71 (2024).

[14] A.-L. Barabási, *The formula: The universal laws of success* (Hachette UK, 2018).

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

[16] D. Wang, C. Song, and A.-L. Barabási, Quantifying long-term scientific impact, Science **342**, 127 (2013).

[17] Q. Ke, E. Ferrara, F. Radicchi, and A. Flammini, Defining and identifying sleeping beauties in science, Proceedings of the National Academy of Sciences **112**, 7426 (2015).

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

[19] L. Wu, D. Wang, and J. A. Evans, Large teams develop and small teams disrupt science and technology, Nature **566**, 378 (2019).

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

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

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

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

[24] T. Yu, G. Yu, P.-Y. Li, and L. Wang, Citation impact prediction for scientific papers using stepwise regression analysis, Scientometrics **101**, 1233 (2014).

[25] C. Stegehuis, N. Litvak, and L. Waltman, Predicting the long-term citation impact of recent publications, Journal of informetrics **9**, 642 (2015).

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

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

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

[29] L. Möckl, A. R. Roy, P. N. Petrov, and W. Moerner, Accurate and rapid background estimation in single-molecule localization microscopy using the deep neural network bgnet, Proceedings of the National Academy of Sciences **117**, 60 (2020).

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

[31] B. P. Abbott, R. Abbott, T. D. Abbott, F. Acernese, K. Ackley, C. Adams, T. Adams, P. Addesso, R. X. Adhikari, V. B. Adya, *et al.* (LIGO Scientific Collaboration and Virgo Collaboration), Constraints on cosmic strings using data from the first advanced ligo observing run, Phys. Rev. D **97**, 102002 (2018).

[32] E. van Nieuwenburg, E. Bairey, and G. Refael, Learning phase transitions from dynamics, Phys. Rev. B **98**, 060301 (2018).

[33] A. V. Leeuwenhoek, 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**, 53 (1709).

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

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

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

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

[38] M. Krenn, J. Landgraf, T. Foesel, and F. Marquardt, Artificial intelligence and machine learning for quantum technologies, Phys. Rev. A **107**, 010101 (2023).

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

[40] A. A. Salatino, F. Osborne, T. Thanapalasingam, and E. Motta, 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* (Springer, 2019) pp. 296–311.

[41] S. Rose, D. Engel, N. Cramer, and W. Cowley, Automatic keyword extraction from individual documents, Text mining: applications and theory , 1 (2010).

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

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

[44] A.-L. Barabási, *Network Science* (Cambridge University Press, 2016).

[45] Y. Lu, 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)* (IEEE, 2021) pp. 5809–5814.

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

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

[48] A. V. Belikov, A. Rzhetsky, and J. Evans, Prediction of robust scientific facts from literature, Nature Machine Intelligence **4**, 445 (2022).

[49] J. G. Foster, A. Rzhetsky, and J. A. Evans, Tradition and innovation in scientists' research strategies, American Sociological Review **80**, 875 (2015).

[50] J. Sourati and J. A. Evans, Accelerating science with human-aware artificial intelligence, Nature Human Behaviour **7**, 1682 (2023).

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

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

[53] J. G. Foster, F. Shi, and J. Evans, Surprise! measuring novelty as expectation violation, SocArXiv 2t46f (2021).

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