nature machine intelligence

Article

https://doi.org/10.1038/s42256-026-01206-y

Predicting new research directions in materials science using large language models and concept graphs

Received: 19 June 2025

Accepted: 16 February 2026

Published online: 01 April 2026

Check for updates

Thomas Marwitz[1], Alexander Colsmann[2], Ben Breitung[3], Christoph Brabec[4,5], Christoph Kirchlechner[6], Eva Blasco[7], Gabriel Cadilha Marques[3], Horst Hahn[3,8], Michael Hirtz[3,9], Pavel A. Levkin[10], Yolita M. Eggeler[11], Tobias Schlöder[3] & Pascal Friederich[1,3]

Due to an exponential increase in published research articles, it is impossible for individual scientists to read all publications, even within their own research field. Here we investigate the use of large language models to extract the main concepts and semantic information from scientific abstracts in the domain of materials science to identify links that were not noticed by humans and to suggest inspiring near and/or mid-term future research directions. We show that large language models can extract concepts more efficiently than automated keyword extraction methods to build a concept graph as an abstraction of the scientific literature. A machine learning model is trained to predict emerging combinations of concepts, that is, new research ideas, based on historical data. We demonstrate that integrating semantic concept information leads to increased prediction performance. The applicability of our model is demonstrated in qualitative interviews with domain experts based on individualized model suggestions. We show that the model can inspire materials scientists in their creative thinking process by predicting innovative combinations of concepts that have not yet been investigated.

Promising new research directions often arise from combining concepts that have not previously been investigated together[1]. While experienced scientists possess vast domain knowledge enabling them to thoroughly explore research topics within (and adjacent to) their area(s) of expertise, finding new connections between their research topics and other yet unfamiliar topics to foster new ideas and findings is inherently challenging. Machine learning (ML) methods can help to look beyond the personal area of expertise by identifying previously unthought-of combinations of research topics and thus enable the exploration of a vast hypothesis space beyond human intuition[2,3].

Scientific information is contained in a plethora of research publications in a rich but unstructured manner, and this lack of structured information poses challenges for automated analysis[4,5]. Focusing on the extensive domain of material science, we first examine how to systematically extract the main concepts of scientific articles, namely keywords or key phrases. Recent breakthroughs in natural language

[1]Institute of Theoretical Informatics, Karlsruhe Institute of Technology, Karlsruhe, Germany. [2]Material Research Center for Energy Systems, Karlsruhe Institute of Technology, Karlsruhe, Germany. [3]Institute of Nanotechnology, Karlsruhe Institute of Technology, Karlsruhe, Germany. [4]Department of High Throughput Methods in Photovoltaics, Forschungszentrum Jülich GmbH, Helmholtz-Institute Erlangen-Nürnberg (HI ERN), Erlangen, Germany. [5]Department of Materials Science and Engineering, Institute of Materials for Electronics and Energy Technology (i-MEET), Friedrich-Alexander-Universität Erlangen-Nürnberg, Erlangen, Germany. [6]Institute for Applied Materials, Karlsruhe Institute of Technology, Karlsruhe, Germany. [7]Institute for Molecular Systems Engineering and Advanced Materials, Heidelberg University, Heidelberg, Germany. [8]Department of Materials Science and Engineering, University of Arizona, Tucson, AZ, USA. [9]Karlsruhe Nano Micro Facility, Karlsruhe Institute of Technology, Karlsruhe, Germany. [10]Institute of Biological and Chemical Systems - Functional Molecular Systems, Karlsruhe Institute of Technology, Karlsruhe, Germany. [11]Laboratory for Electron Microscopy, Karlsruhe Institute of Technology, Karlsruhe, Germany. e-mail: pascal.friederich@kit.edu

https://doi.org/10.1038/s42256-026-01206-y

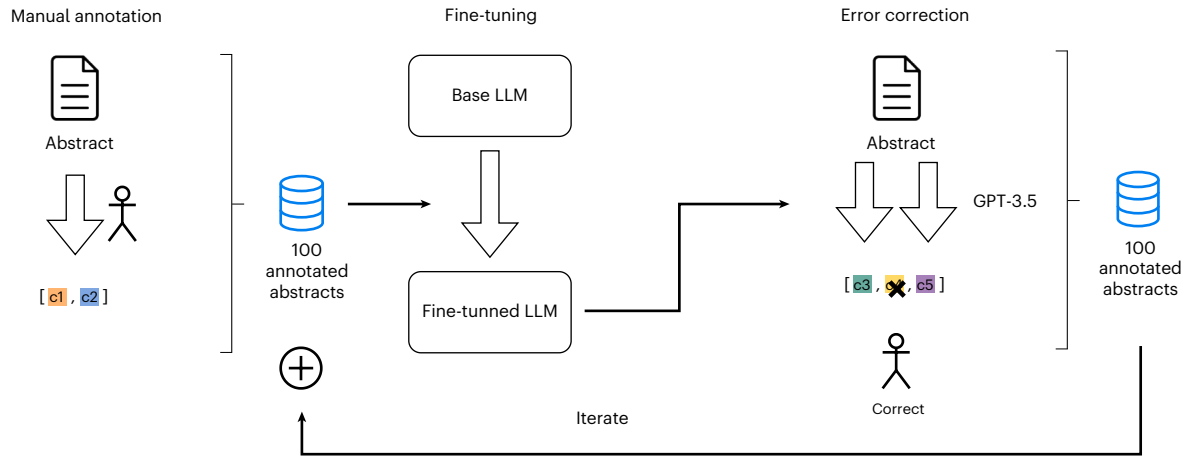


Fig. 1 | Generation of labelled data. Manual labelling (concept extraction) of 100 abstracts, fine-tuning of an LLM-base model on the annotated data, automatic concept extraction from 100 further abstracts with human correction, and then repeat fine-tuning of the base LLM with the new extended labelled dataset.

processing now allow us to extract structured data from text and process it automatically[6–11]. Here we investigate whether large language models (LLMs) can offer improvements over traditional algorithmic methods in this extraction process.

After identifying and extracting the concepts and their connections (that is, the co-occurrence in the same article), we investigate how to use this information to predict new combinations of concepts. In a previous study, Krenn et al. proposed SemNet, a graph that tracks the evolution of scientific literature in the domain of quantum physics[12]. The nodes of the SemNet graph are concepts, that is, keywords extracted from text, using an algorithm called rapid automatic keyword extraction (RAKE) in conjunction with some predefined rules[13]. Apart from analysing emerging trends within SemNet, the authors use changes in the graph to make predictions. To this end, they derive topological properties, such as node degree, and use them as input for a neural network (NN) to predict future connections. In a Kaggle challenge, the participants predicted changes in a SemNet built from the artificial intelligence (AI) literature[14]. While the most successful models combined specific hand-selected network features with ML techniques, such as NNs or graph NNs (GNNs), other participants employed purely theoretical or end-to-end ML approaches. However, the participants' models could only use the structure of SemNet because the real meaning behind the nodes was not revealed in the challenge.

In this study, the information on materials science concepts is similarly compressed into a concept graph. Given the advances in language encoder models[15,16], we use the MatSciBERT model[17] to provide additional information on the concepts in the form of semantic embeddings to enrich the topological information of the nodes. Then, we explore how ML methods can use the time evolution of this representation of the literature to perform link predictions.

Recent advances have shown that graph-based approaches can accelerate discovery in materials science: SciAgents employs multi-agent graph reasoning, Graph-PRefLexOR integrates symbolic graph abstractions with LLMs and generative knowledge extraction with graph representations further supports hypothesis generation[18,19]. Complementary efforts on AI-driven ideation include SciMuse and SciMON, which use enriched co-occurrence and temporal knowledge graphs for idea generation, ResearchAgent, which iteratively refines literature-grounded ideas with knowledge-augmented LLMs, and SCI-IDEA, which applies context-aware embeddings for systematic ideation[20–23]. In contrast to approaches that analyse understanding, intelligence and creativity in general and try to evoke these in machines[24], we aim to foster human creativity by using AI to help materials scientists propose new research directions by combining previously uncombined concepts. To explore the real-world applicability of our model and its suggestions, we conducted interviews with materials sciences researchers to assess how well the concepts generated and suggested by our model align with concepts from their own research.

Table 1 | Selected examples of abstracts and concepts extracted by our fine-tuned Llama-2-13B model

Abstract excerpt	Extracted concept	Note
'Strengthening mechanisms in short carbon fibre reinforced Nb/Nb5Si3 composites'	carbon fibre reinforcement	Nominalization
'Removal of Metal Impurities from the'	metal impurity removal	Removal of 'of', singular normalization
'Resistance of Al2O3Coatings on Functional Structure'	al2o3 coating	Singular normalization, dealing with wrong formatting
'Both fully and partially amorphous ribbons have been obtained'	fully amorphous ribbon, partially amorphous ribbon	Normalization of 'and' in concepts
'Successive ionic layer adsorption and reaction SILAR trend for'	successive ionic layer adsorption and reaction	Extraction of long form without abbreviation in same concept

Results

Concept extraction and concept graph

Using an LLM-based approach (Methods), approximately 510,000 chemical formulae and 3,600,000 concepts were extracted from the 221,000 abstracts in our database, which corresponds to an average of 2.3 chemical formulae and 16.3 concepts per abstract. The extracted concepts were then condensed into approximately 52,000 unique formulae and 1,241,000 unique concepts by removing duplicates. In general, our method resulted in more precise concept extraction than rule-based approaches (see Supplementary Note 1 for details). Due to the extraction capabilities of LLMs, the amount of manual annotation work needed to generate the initially required data is negligible, especially as our iterative approach (Fig. 1; Methods) reduces the manual effort to a minimum. Notably, the fine-tuned LLMs were able to extract concepts that were not present verbatim in the text. Table 1 shows selected examples to demonstrate the capabilities of the fine-tuned LLMs for nominalization, the removal of fill words such as 'of', plural-to-singular conversion and formatting corrections.

https://doi.org/10.1038/s42256-026-01206-y

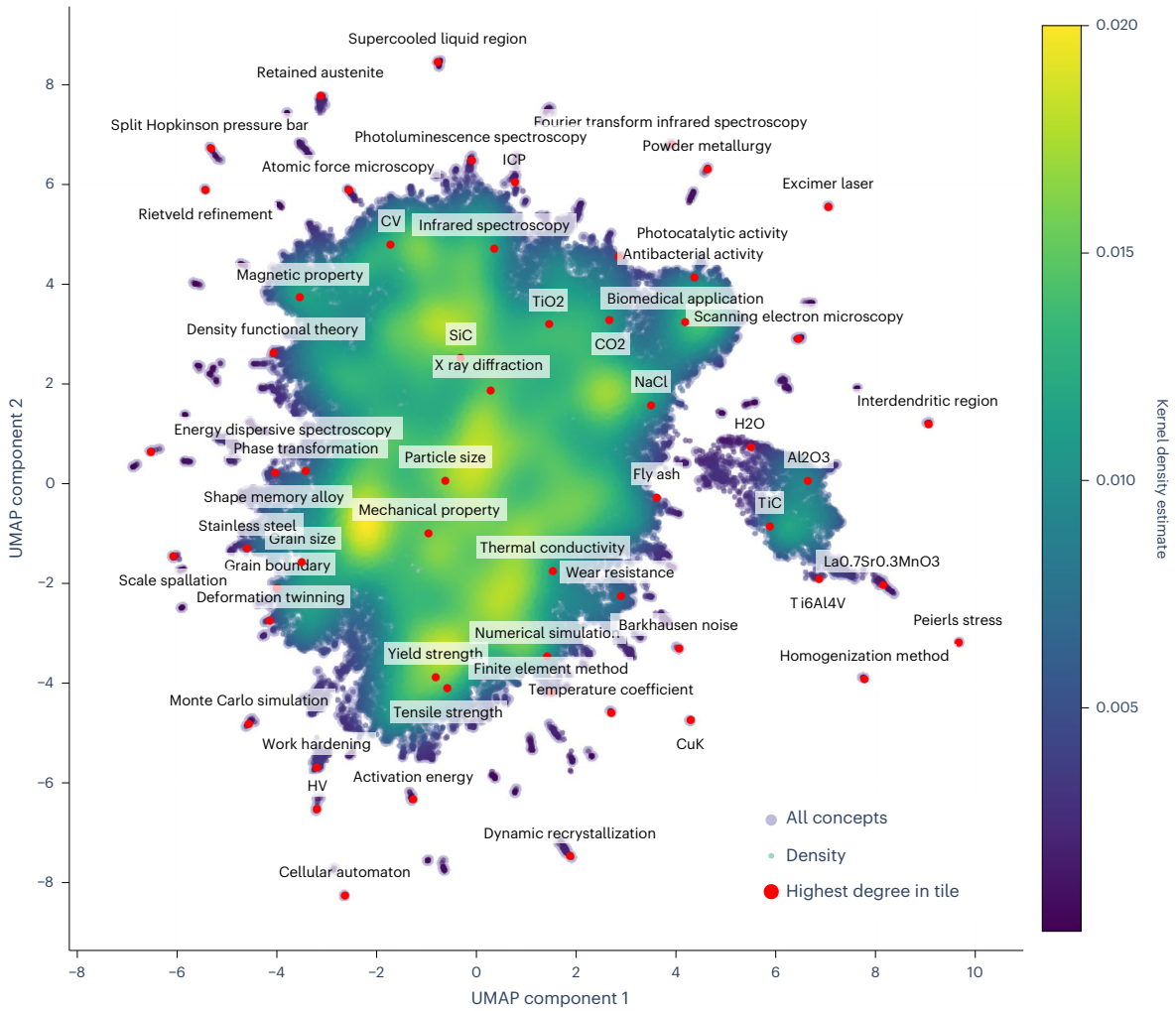


Fig. 2 | Map of materials science. Two-dimensional UMAP[25] projection of all extracted concepts with the highest-degree concepts in each square of length 2 highlighted and annotated ('Highest degree in tile'). Yellow and purple background colours respectively indicate high and low concept densities calculated using kernel density estimation.

To construct a concept graph, we only include concepts that appeared at least three times and consist of at least two words. The resulting graph comprises approximately 137,000 nodes and 13,000,000 edges, making the calculation of topological features that require the squaring of the adjacency matrix possible. Supplementary Table 1 presents an overview of the 25 most frequently encountered concepts and formulas.

An analysis of the node degree distribution in the resulting graph shows that the majority of the nodes have a degree between 30 and 1,000 (Supplementary Fig. 8). While a few concept nodes act as hubs—having a notably larger number of connections than others—most of the concepts in the graph are directly linked to only a few others, making the resulting graph sparse. The evolution of the concept graph over time shows that connectivity further increases as more papers are being published using already existing concepts. We observe an increase in concept centralization, which is that fewer and fewer nodes account for a larger share of the total connections (Supplementary Fig. 9).

We visualize all concepts by projecting their high-dimensional concept embeddings to two dimensions using the uniform manifold approximation and projection (UMAP)[25] technique with default settings. Figure 2 displays the result, which we call the 'Map of materials science' (an interactive version that can be explored on inspire.aimat.science). We then run nearest neighbour queries[26] on the concept embeddings to explore whether these 768-dimensional vectors capture semantic meaning. The example queries listed in Supplementary Table 2 show the striking similarity between the queried concept and its nearest neighbours.

Link prediction

To statistically assess the performance of our different link prediction models (see Methods for a detailed description), we evaluate their performance on a held-out test set for edge formation in the period between 2020 and 2022. The test set consisted of 2,000,000 node pairs, including 307 (0.015%) positive pairs, that is, emerging edges. We complement this with a qualitative analysis of the real-world applicability of the models based on human expert knowledge.

Figure 3a shows the receiver operator characteristic (ROC) curves for predicting link formation during the test period, as they illustrate the capacities of the model to distinguish between classes across all possible classification thresholds. ROC curves are particularly useful for imbalanced datasets because they evaluate performance independently of class distribution[27]. More information about test set creation, along with detailed results (Precision/Recall@k), can be found in Supplementary Note 3. Although the 'Baseline' model (a modified version of Krenn et al.[14]) performs slightly better (area under the curve (AUC) 0.9109) than the Concept Embeddings (MatSciBERT) model (AUC 0.8855), the performance of the latter already shows that our model architecture can use the semantic information contained in the concept embeddings.

https://doi.org/10.1038/s42256-026-01206-y

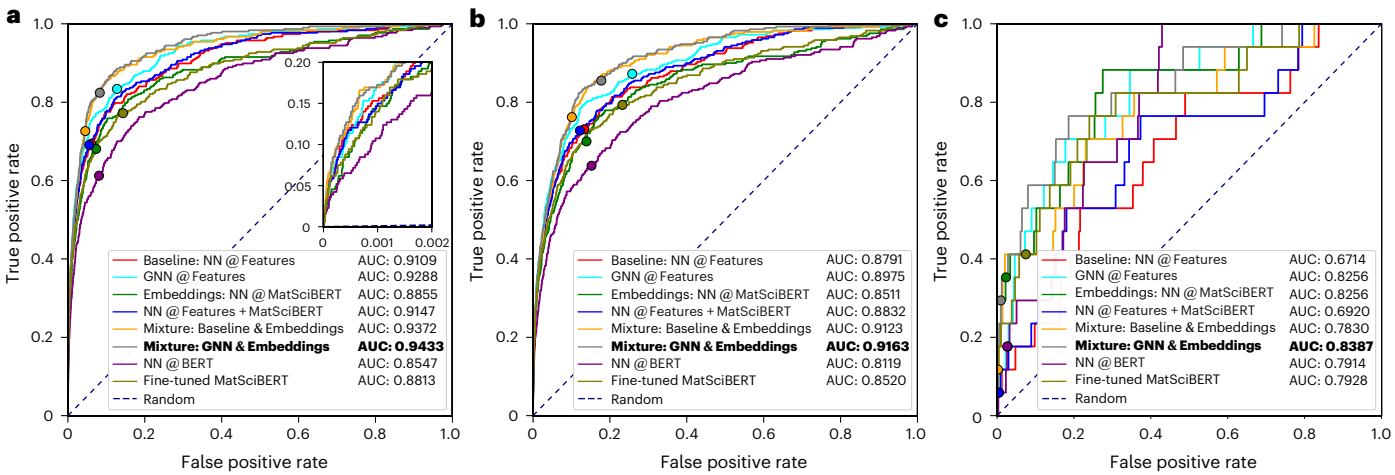


Fig. 3 | Performance metrics (ROC and the respective AUC) for our link prediction models on the test set (T_{test} = [2020, 2022]). Markers highlight the performances at a threshold of 0.5. **a**, ROC curves on all data points with a zoomed-in view of the low-false-positive-rate region in the inset. **b**,**c**, The respective performance metrics for $d_{prev} = 2$ (**b**) and $d_{prev} = 3$ (**c**). Best result is in bold.

A GNN model based on the GraphSAGE architecture surpasses the Baseline model with an AUC of 0.9288, suggesting that while both models access the same input features, the GNN effectively leverages additional structural signals to improve performance. The Pure Text Baseline (implemented via a fine-tuned MatSciBERT[17]) also exploits this semantic information and performs similarly overall, but at a 5× higher inference cost; while overall performance was similar, worse results were achieved for emerging links between nodes which were previously connected through more than one intermediate node. Furthermore, the performance of the three hybrid models demonstrates that the link prediction task benefits from incorporating semantic knowledge on top of local graph features: While the Combination of features model already shows a slightly improved AUC of 0.9147, the Mixture of Baseline and Embeddings and Mixture of GNN and Embeddings models approach exhibits a substantial performance leap in the AUC metric, which reaches a maximum of 0.9433 when scaling the GNN and Concept embeddings model predictions by 0.5 and 0.5, respectively. We speculate that gradient descent optimization on a unified feature vector (concatenating the features of the Baseline and Concept embeddings model)—as it is done in the Combination of features model—might not be as effective as optimizing them individually. The distinct nature of baseline features versus high-dimensional concept embeddings could lead to the gradient for each batch becoming a suboptimal compromise between the gradients suited for each feature type in isolation. While MatSciBERT may under-represent emerging or interdisciplinary concepts, it still benefits from the base BERT knowledge, and tokenization ensures meaningful embeddings (Fig. 4; Methods). In our experiments, MatSciBERT (AUC 0.8855) outperformed BERT (AUC 0.8547), indicating an advantage of domain-specific embeddings, although BERT still offers a reasonable baseline.

We further investigated the predicted new links with regard to the previous node distance d_{prev}, that is the shortest path distance between two nodes in the time range of the test set T_{test} = [2020, 2022] before they become directly connected. An analysis of the graph shows its dense interconnectedness as certain prevalent concepts in materials science, such as 'mechanical property' and 'X ray diffraction' (Supplementary Table 1), have edges with the majority of nodes, which leads to a large number of short distances between many concept pairs. Although the graph consists of 137,000 nodes, nearly all of the unlinked concept pairs in the test set were already connected through one ($d_{prev} = 2$, 43.3%) or two ($d_{prev} = 3$, 56.5%) concepts. The distribution of d_{prev} is even more biased towards short paths for the positive samples, meaning the connections that actually formed during the test period. Of the 307 emerging edges (94.5%) in the test set, 290 were found to have $d_{prev} = 2$, while only 17 (5.5%) had a previous distance of 3, which shows that the proximity of two nodes in the concept graph increases the probability for a new edge to form between them. Samples at $d_{prev} = 4$ were scarce (0.2% of the total samples, all negatives) and therefore excluded from further analysis.

While the Baseline model tends to correctly predict emerging edges primarily at a distance of 2 (212 of 213 true positives have $d_{prev} = 2$) with a recall of 73.1%, it performs much worse for $d_{prev} = 3$ (recall 5.9%). By contrast, the Concept embeddings model achieves a significantly better recall of 35.3% ($P < 0.05$, DeLong test[28] for $d_{prev} = 3$ while only slightly compromising on the recall at $d_{prev} = 2$ (70.0%). Notably, the GNN model matches this performance at $d_{prev} = 3$, demonstrating that improved structural processing can rival the benefits of semantic embeddings for distant connections. The results are summarized in Extended Data Table 1. The high number of false positives, especially at $d_{prev} = 3$ is not a problem in itself because those combinations may remain scientifically plausible and will subsequently be evaluated by human scientists. Hence, we prioritize recall over precision in order not to miss valuable ideas.

Note that optimizing the classification metrics by changing the classification threshold of 0.5 for a positive prediction is outside the scope of this work, which mainly aims at rating non-existing links with respect to their future emergence rather than accurately predicting whether a new link will form or not.

In addition, we also separately calculated the ROC curves and the corresponding AUCs for both $d_{prev} = 2$ and $d_{prev} = 3$, and the results (Fig. 3b,c) emphasize the Baseline model's failure to correctly categorize most positives with $d_{prev} = 3$. This not only highlights the inherent challenge of predicting positives at greater distances but also indicates that integrating semantic information enhances the model's ability to forecast connections between concept pairs that are further apart in the graph. However, these emerging connections with larger previous node distances are particularly interesting and hold great potential to broaden the scientific scope beyond the more obvious new research directions. Ultimately, the Mixture of GNN and Embeddings yields the highest AUC for these distant connections, outperforming the individual models by effectively combining structural and semantic signals.

We also evaluated our Baseline model on the Science4Cast benchmark, where it achieved an area under the receiver operating characteristic of 0.9088, ranking second among all reported approaches by Krenn et al.[14]. This finding demonstrates that a deep NN trained on a large set of semantically meaningful features can outperform most competing

https://doi.org/10.1038/s42256-026-01206-y

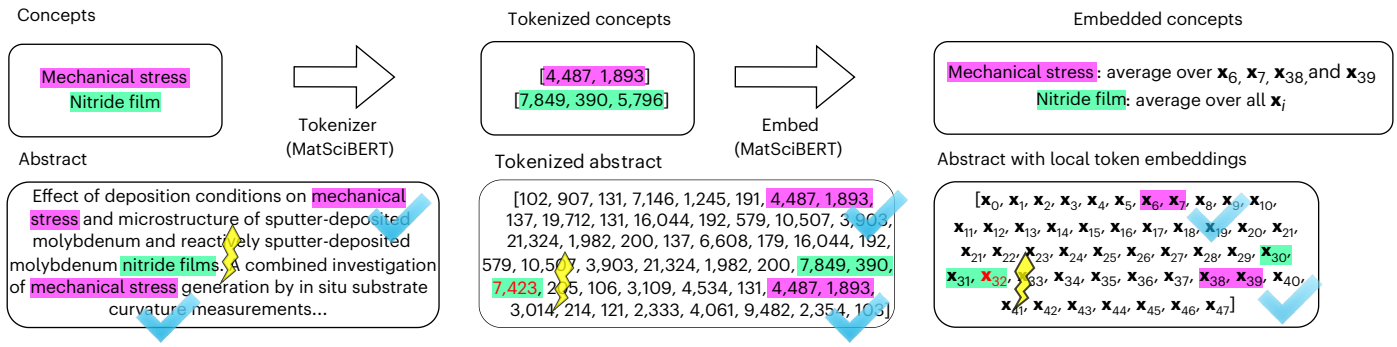


Fig. 4 | Example of calculating concept embeddings from an abstract. Embeddings of verbatim concepts ('mechanical stress') are calculated by averaging all local MatSciBert embeddings of the corresponding tokens (4,487 and 1,893). Embeddings of non-verbatim concepts ('nitride film' is only present in the abstract in its unnormalized form 'nitride films') are calculated as the average of all token embeddings. **x** represents the embedding vectors of the tokens.

methods, including those based on common neighbours and node-2vec embeddings combined with a Transformer architecture[6,29,30]. As Science4Cast does not contain any semantic or text information about the meaning of nodes, we could not apply and benchmark our embedding-based models on the Science4Cast challenge.

Analysis of human expert evaluation

As the second part of the model performance analysis, we conducted interviews with ten materials scientists (human experts). Each expert received an individualized report containing recommended concept combinations suggested by our prediction model (Mixture of Baseline and Embeddings). The personalized recommendations were generated using the Mixture of Baseline and Embeddings model, which is marginally less performant (<0.01 AUC) than the Mixture of GNN and Embeddings model, because the GNNs were only tested at a later stage of our study. The suggestions were subsequently discussed in the interviews to assess and clarify the proposed concept combinations. The small sample size of the interviewees and a potential selection bias limit the robustness of our conclusions and allow only a qualitative analysis and anecdotal findings. Nonetheless, the expert feedback sheds light on the usefulness of the suggestions provided by our model.

Report generation. An overview of the report generation is shown in Extended Data Fig. 1. First, a set of individualized concepts C_{own} is generated as the intersection of (1) all concepts extracted from the abstracts of the recent publications of the respective researcher and (2) all known concepts C_{known} in the concept graph. Based on these two sets of concepts, we generated researcher-specific suggestions, that is combinations of concepts: The first two report sections, $S_{\text{own}\times\text{own}}$ and $S_{\text{own}\times\text{other}}$, contain the top 25 combinations of the own concepts with themselves and with all other concepts, respectively. We applied two heuristics (avoid generic concepts and avoid combinations that are too similar and too unrelated based on their semantic embeddings) to filter the suggestions in the second category, resulting in a section $S^{\text{filtered}}_{\text{own}\times\text{other}}$. To take the researcher's full profile into account, the next section $S_{(\text{many own})\times\text{other}}$ contains the top 20 concepts with highly scored connections to many own concepts. For the final section of the report ('LLM curation'), LLMs were queried to select interesting combinations from the previous sets of combinations and to write a short paragraph with more information on how the concepts can be combined and why these specific combinations are promising new research directions. A technical definition of each section is given in Supplementary Note 4.

Classification of the suggestions. Based on the individual reports described above, a 30-minute interview was conducted with each researcher, in which the suggested combinations of concepts were classified as already known (A), nonsensical or not understandable (B) and novel, interesting or inspiring (C). In section 4 ($S_{(\text{many own})\times\text{other}}$), suggestions were generously already counted as overall interesting (category C) if one of the many concepts in C_{own} was inspiring in conjunction with the proposed other concept. To account for cases in which the participants were unsure whether to label a suggestion as B or C, an additional category D was introduced in the analysis. For further analysis, the first class was divided into already published combinations (A1), which were likely missed during dataset generation (for example, very recent publications or publications outside of the analysed literature corpus) and obvious, trivial or very general combinations (A2), which are not necessarily mentioned together in an abstract.

Of 292 categorized suggestions, 71 suggestions were classified by the interviewees as already known (class A1), 36 as trivial (class A2), 99 as nonsense (class B), 77 as interesting (class C) and 9 as uncertain (class D); an overview of the categorized suggestion is given in Table 2. Thus, overall, the interviewees considered 26% of all suggested concept

Table 2 | Amount of suggestions categorized by researchers across all interviews, broken down by section of the report

Section	Occurance		
	Category	Number	Fraction
$S_{\text{own}\times\text{own}}$	A1	31	30.7%
	A2	21	20.8%
	B	25	24.8%
	C	21	20.8%
	D	3	3.0%
$S_{\text{own}\times\text{other}}$	A1	17	22.1%
	A2	8	10.4%
	B	35	45.4%
	C	14	18.2%
	D	3	3.9%
$S_{(\text{many own})\times\text{other}}$	A1	6	23.1%
	A2	1	03.9%
	B	3	11.5%
	C	16	61.5%
$S^{\text{filtered}}_{\text{own}\times\text{other}}$	A1	17	19.3%
	A2	6	06.8%
	B	36	40.9%
	C	26	29.5%
	D	3	3.4%

combinations interesting. An excerpt of combinations per category can be found in Supplementary Tables 7 and 8. As mentioned above, the number of interview partners is insufficient for a reliable statistical analysis; for example, the total number of classifications per researcher ranged between 18 and 48, with per-participant variance ranging from 5.04 to 51.36. An overview of the classified combinations per researcher and per-participant variance can be found in Supplementary Fig. 20 and Supplementary Table 6.

To evaluate the usefulness of the 'LLM Curation' approach, we analysed how many of the combinations suggested by an LLM are later labelled as interesting. Supplementary Table 5 presents the confusion matrix of the variables 'Suggested by an LLM' and 'Is Interesting' (rated as C by a human expert). We observed a rounded precision of 47% regarding the LLM selecting interesting concepts; that is, of the 53 concepts suggested by the LLM as interesting, 24 were also labelled as interesting by the scientists. This is a substantial improvement in precision compared to 61 of 266 (23%) concept combinations analysed in total. In this context, the recall is not of primary interest as the LLM was only allowed to select a limited number of combinations as described above. An analysis of the previous node distances of the suggested combinations across all reports showed that 5 of 9 concept pairs at $d_{prev} = 3$ were rated as category C. This high ratio of interesting combinations underpins our previous assumption that including semantic information in the prediction model increases its capability to foster out-of-the-box thinking.

In retrospect, groundbreaking ideas sometimes seemed absurd at first. Interviewees repeatedly categorized a combination as B ('nonsense') only to change their minds after some reconsideration or when seeing that suggestion again in the 'LLM Curation' section with an elaboration on how the combination could be realized. The exemplary paragraph often helped researchers to judge the usefulness of concept combinations. We speculate that the task of generating one's own hypotheses on how concepts could be connected is inherently more difficult than judging an existing proposal.

Furthermore, many combinations could not be classified, especially in the sections 'Other Concept Combinations' and 'Filtered Other Concept Combinations'. This is attributable to the vast number of concepts in C_{known}, many of which the interviewees had never heard of. To allow researchers to navigate unknown suggestions, additional context, such as the original abstract, might prove helpful.

Examples of human expert evaluation

To illustrate the value of the suggested combinations of concepts in more detail, we discuss five selected examples of category C suggestions by putting them into context and describing why these combinations are interesting. A more detailed discussion of all five concept combinations can be found in Supplementary Information.

Suggestion 1 ('conventional ceramic' + 'graphene oxide'). This suggestion associates 'conventional ceramics' with 'graphene oxide', two domains seldomly combined. Conventional oxide ceramics provide chemical, thermal and structural stability. Graphene oxide offers a high-surface-area, electronically conductive carbon framework. Their union could yield composites marrying ceramic robustness with rapid charge and heat transport, relevant to batteries, catalysis and thermal barriers. Existing studies mix pre-synthesized oxides with graphene derivatives, giving limited interfacial contact. We show preliminary data of a ~200-nm iron oxide shell on multilayer graphene in Extended Data Fig. 2 (unpublished). The in situ process creates intimate oxide–graphene interfaces and a continuous conductive network. Electrochemical tests show a high reversible capacity and enhanced redox kinetics in Li-ion conversion cells. These findings demonstrate the model's capacity to provide inspiration for overlooked but feasible synthesis strategies. Systematic exploration of AI-highlighted ceramic/graphene hybrids may accelerate multifunctional material discovery.

Suggestion 2 ('tensile strain' + 'molecular architecture'). Thin-film organic and perovskite solar cells comprise multi-layers with mismatched thermal expansion coefficients. Temperature excursions during coating, annealing and operation impose tensile strain at these organic–inorganic interfaces. Such strain drives delamination and point defect formation, accelerating performance loss. While strain engineering is routine in inorganic semiconductors, it is rarely applied in soft matter photovoltaics. Molecular architecture provides a complementary lever: Greater torsional flexibility lowers the film modulus and dissipates stress. Thus, the AI-proposed link 'tensile strain + molecular architecture' highlights an under-exploited stability pathway. In 2024, Brabec and Friederich showed that hole transport layers containing triphenylamine derivatives show an enhanced power conversion efficiency by accommodating strain[31]. These studies corroborate strain-aware molecular design as a broadly applicable interface strategy. Systematic exploration could extend device lifetimes without compromising performance.

Suggestion 3 ('multiphase structure' + 'selective laser melting'). Microstructure denotes all internal structural features—from lattice arrangement to point, line, planar and volumetric defects—across relevant length scales. These features dictate the mechanical and functional response and thus allow a rational selection of materials. A central attribute is the spatial distribution of phases, each possessing uniform crystal structure and composition. Technical alloys and ceramics are typically multiphase, generated through controlled thermo-mechanical processing. Phase topology manipulation enables simultaneous optimization of strength, toughness, corrosion resistance and functional properties. Selective laser melting (SLM) fabricates components by layer-wise laser melting of metal powders directly from digital models. The extreme heating-cooling rates inherent in SLM impose strong non-equilibrium solidification conditions. The resulting parts frequently exhibit metastable, compositionally heterogeneous multiphase microstructures. These structures can elevate hardness and corrosion resistance, yet they may also induce residual stresses. Therefore, elucidating phase-formation pathways during SLM represents a critical avenue for advanced materials design.

Suggestion 4 ('stress-induced phase transformation' + 'hexagonal boron nitride'). Stress-induced phase transformation toughening, exemplified by the tetragonal to monoclinic switch in zirconia, suppresses crack advance through local volume expansion. An alternative route uses elastic anisotropy; in pearlitic wires, load-parallel micro-cracks raise both strength and toughness. Applying these principles to boron nitride asks whether hexagonal BN (h-BN) can function as a transformation- or anisotropy-assisted toughener. Cubic BN (c-BN) is a dense, superhard phase, and pressure-driven c-BN to h-BN transitions could release crack-tip stresses. h-BN displays strong in-plane versus out-of-plane stiffness contrast, enabling guided micro-crack arrays akin to the pearlite mechanism. A coupled h-BN anisotropy and c-BN/h-BN transformation would thus offer simultaneous crack deflection and compressive shielding. Recent c-BN/h-BN composites show higher hardness and fracture energy, indicating technological promise. However, the role of a reversible transformation in these gains remains experimentally unverified. Targeted high-pressure mechanical tests with in situ diffraction are required to resolve transformation kinetics and toughening contributions. Establishing these links could generalize anisotropy-assisted transformation toughening to lightweight nitride coatings.

Suggestion 5 ('in-plane polarization' + 'organic solar cell'). Ferroelectric in-plane polarization, recently demonstrated in $MAPbI_3$ perovskites, spatially separates photocarriers and channels them towards electrodes. The model's suggestion indicates that similar lateral dipole fields could be engineered in organic absorbers. Asymmetric polar moieties, oriented during self-assembly or within covalent

organic frameworks, may supply the required non-centrosymmetry. The resulting internal field should enhance carrier separation and transport, while a higher dielectric constant lowers exciton binding and monomolecular recombination. Ferroelectricity has so far been verified only for halide perovskites[32–34], and is absent in silicon or conventional organic cells[35,36]. Piezoelectric polymers such as polyvinylidene fluoride already exploit oriented dipoles in sensing devices, suggesting viable processing routes. Earlier attempts to raise organic permittivity show limited success, and deliberate in-plane polarization remains unexplored[37]. Hence, the predicted concept defines a tractable, new direction for photovoltaic materials research.

Discussion

In the first part of this study, we showed that the power of LLMs, especially LLaMa-2-13B, can be harnessed to extract scientific concepts—vaguely defined as key phrases—from scientific texts. We established a methodology for fine-tuning open-source LLMs based on small manually labelled abstracts, which guides the LLM to extract only relevant concepts. The initial training data can be iteratively extended by humanly corrected LLM annotations to further improve the extraction process, but no human verification is required to check the final 221,000 labelled data points. Follow-up studies may investigate whether prioritizing quality over quantity[38] in the annotated training examples, by using fewer but carefully selected data points, could yield more accurate and representative extracted concepts, and whether including synthetic data can help to accelerate the annotation process and further enhance model performance.

In addition, we created a concept graph, derived from the previously extracted materials science concepts and the dates of the corresponding publications. This graph was successfully used to predict emerging links between previously unconnected concepts, underscoring that a simple graph representation suffices for this task. Finally, we demonstrated that integrating semantic knowledge in the form of concept embeddings boosts the predictive performance of our model. Combining the GNN approach with semantic features is possible and will be explored in future work. We investigated the usefulness of our model in a real-world scenario through qualitative interviews with domain experts, who rated 77 out of 292 (26%) generated recommendations as interesting. While this rate may sound modest, each 30-minute session still yielded several promising ideas, making the outcome practical for guiding research.

In summary, we demonstrated that ML tools can be used to automatically process the vast amount of scientific literature and to predict future research directions that have not previously been explored to foster innovation and advancements. While this work focused on the material sciences as a use case, the developed approach can easily be extended to other research areas. By suggesting potential new research directions, we hope to drive innovation and collaboration in the field.

Methods

The key steps of our approach are depicted in Extended Data Fig. 3. After gathering the abstracts of a large number of research publications in the domain of materials sciences, we extracted the main concepts, that is short key phrases consisting of few words from these abstracts, and used them as nodes in a concept graph that mirrors the (time-dependent) connectivity of the materials science concepts in literature. In the final step of our workflow, we performed link prediction on this graph based on both network properties, for example, connectivity information and semantic knowledge about the concepts captured in aggregated word embeddings.

Dataset

We prepared a dataset of published papers related to materials science. Data were obtained from OpenAlex by querying all publications listed at materials science-related journals, conferences and other venues[39]. The retrieved papers were filtered based on language, length and whether they had an abstract. For each publication, the title and abstract were cleaned and concatenated. Chemical formulae were extracted, stored separately and later merged with the extracted concepts. The resulting dataset comprised approximately 221,000 articles published between 1955 and 2022, with relevant attributes being 'title', 'abstract' and 'publication date'. A more detailed description of the dataset generation is given in Supplementary Note 7.

Concept extraction

Previous work used RAKE for concept extraction in conjunction with manual filtering to remove errors. These errors included phrases that did not represent semantic information and that were introduced by imperfections in RAKE's statistical analysis[12–14]. Instead, we opted to extract concepts using fine-tuned LLMs (Fig. 1). To create a dataset for fine-tuning, 100 randomly chosen abstracts were first manually annotated by extracting and partially adjusting or even paraphrasing relevant and meaningful concepts as a preliminary step. Manual annotation is particularly sensitive to the labeler because there is no unique way of extracting and defining concepts. Subsequently, we fine-tuned the LLaMa-2-13B base model[40,41] on our manually annotated abstracts for 4 epochs, using a learning rate of 5×10^{-4} and a batch size of 1 (Supplementary Note 8). Llama-2 models were state of the art when the work was performed in 2023, but future iterations of this work will use newer models. The size of the model is a trade-off between accuracy and cost, as it is the largest model that can process 20 abstracts at once on an A100 GPU with 80 GB of video random access memory. To accelerate training and especially inference, we incorporated 8-bit quantization[42] and low-rank adaptation techniques[43,44] using Hugging Face's parameter-efficient fine-tuning module[45].

Similar to Dunn et al.'s assisted annotation process[4], the fine-tuned model's outputs were compared in the third step to the concepts extracted using GPT-3.5[7] to efficiently identify and correct common mistakes made by the fine-tuned model. To do so, we labelled 100 additional abstracts, and the base model was again fine-tuned on a larger dataset of 200 abstracts. The process of iteratively adding more automatically extracted and manually corrected concepts to fine-tune the model with more data points could, in principle, be repeated more often, but 200 labelled abstracts were enough for our use case. Finally, the resulting model was employed to extract concepts from approximately 221,000 abstracts in our dataset, requiring approximately 160 GPU hours. Future updates of our concept graph would be substantially less demanding, as only incremental (delta) extraction is required. Future developments in LLM research might enable a complete re-evaluation with higher quality and reliability. After extraction, we conducted minor post-processing of the extracted concepts by removing the remaining plural forms. We note that some bias was introduced by the selection of 200 abstracts with 3,102 distinct concepts, as they were all from materials science, and observe that our method did not extract, for example, core-biology concepts.

Concept graph

The concepts extracted from the materials science literature in our dataset are represented in a multi-graph $G = (V, E)$, where V and E are the respective sets of nodes and edges. In this concept graph, each node $v \in V$ represents a distinct concept and each edge $e \in E$ represents the co-occurrence of two concepts in a single abstract. Each edge is labelled with a timestamp t, indicating the publication date of the abstract containing both concepts, where G_t denotes a subgraph of G defined by the edge list that only includes all edges with time stamps $\leq t$. Therefore, every abstract generates a fully connected clique of its concepts in the concept graph. Multiple edges can exist between each pair of nodes if the concepts co-occurred in more than one abstract.

To enrich the nodes of our concept graph with semantic information, we calculated concept embeddings and used them as node

features. Figure 4 summarizes the procedure of calculating concept embeddings using MatSciBERT[17]: First, both the entire abstract and the previously extracted concepts are tokenized and the embeddings are then calculated for the tokenized abstract. The next step consists of locating all instances of a concept in the abstract and averaging the embeddings of the tokens corresponding to the concept. For example, the concept 'mechanical stress' is tokenized as [4,487, 1,893], and its embedding is calculated as the average of the corresponding representations in the embedded abstract at the positions of the sequence [4,487,1,893] in the tokenized abstract. To derive a singular representation for each concept per abstract, we average the embeddings of all its occurrences. In cases where a concept does not appear verbatim in an abstract—for example, due to the normalization processes during the initial concept extraction—we take the mean embedding of all tokens (while excluding the start and end token in the abstract as its representation). As the final step, we calculated the average embedding for identical concepts across different abstracts to obtain a single embedding for each concept and thus for each node. To prevent information leakage, embeddings used for training and testing were computed only from text available up to the corresponding cutoff year.

Link prediction

The previous method used by Krenn et al. to predict new links in their concept graph exclusively relied on abstract local graph properties, either through a purely graph-theoretical approach using hand-crafted features in conjunction with ML or through employing end-to-end ML methods[12,14]. Here we investigate whether integrating semantic knowledge about the concepts can improve link prediction. In particular, we use concept embeddings—that is, high-dimensional vectors that capture semantic information—to make the semantic information integrable into the link prediction task.

Given the concept graph G, we treat link prediction as a binary classification task. Thus, the objective of the ML model is to predict whether a new edge is formed between an arbitrary pair of previously unconnected vertices (u, v) in the time range $T = [T_{\text{start}}, T_{\text{end}}]$. We chose $T_{\text{start,train}} = 2017$ and $T_{\text{end,train}} = 2019$ for training, which means that our model had access to the entire data up to and including 2016 while its predictions were made for the years 2017, 2018 and 2019. We illustrate link prediction on a rudimentary concept graph in Extended Data Fig. 4.

The prediction task has an inherent strong label imbalance as the likelihood of a randomly selected vertex pair (u, v) forming a link throughout 3 years is extremely low. While there are, for example, 18.7 billion possible new edges that could form between 2017 and 2019, only 1.3 million new edges (0.007%) were observed during this period. To address this imbalance, we oversampled positive labels by using a fixed percentage (30%) of positive examples per batch in the training process. This oversampling during training shifts the trade-off between precision and recall in imbalanced tasks towards higher recall at the cost of losing precision, thus favouring the generation of larger sets of suggestions that may contain inspiring concept combinations over smaller sets in which some valuable ideas might not be included.

A modified version of Krenn et al.'s densely connected NN[14], which relies purely on graph properties at different points in time, was used as the Baseline model. Specifically, the degree of a node u ($\sum_{i=1}^{n} A_{u,i}$) and the sum of all 2-length paths from u ($\sum_{i=1}^{n} A_{u,i}^2$) were calculated for different years in the range $t = [T_{\text{start,train}} - 5, T_{\text{start,train}} - 1]$, where A_t denotes the binary adjacency matrix of G_t. These features were then concatenated for a given pair of nodes (u, v) to result in a 20-dimensional baseline feature vector. In the second Concept embeddings (MatSciBERT) model, the concatenated concept embeddings of u and v were used instead as the (1,536-dimensional) feature vector for the NN classifier, to test their information content and relevance for the link prediction task. We repeat the embedding generation process with BERT, yielding a modified Concept embeddings (BERT) model. To explore another way of utilizing semantic information, we fine-tuned the MatSciBERT model directly to predict the likelihood of two concepts becoming connected in our concept graph, thus yielding the Pure Text Baseline model.

The Baseline model was then combined with the Concept embeddings model in two hybrid models, the first of which (Combination of features) used a concatenation of the feature vectors from the two previous models as the input. The hyperparameters of all NNs were optimized using a comprehensive grid search varying the number of neurons per layer, the percentage of positive samples in each batch, the learning rate and the dropout probability (see Supplementary Table 9 for a list of optimized hyperparameters).

The second hybrid model (Mixture of Baseline and Embeddings) uses a weighted output of the optimized Baseline and Concept embeddings (MatSciBERT) models, where an optimal weighting of 3:2 was determined using hyperparameter optimization. We acknowledge many other possible hybrid models exist aside from concatenating the two input vectors and calculating the weighted average of the output probabilities, as the two parts of the input could be passed through a first set of layers separately before the two outputs are concatenated and passed through a second set of layers. However, optimizing the architecture of the NN was outside the scope of this study, as our goal mainly consisted of showing that including the concept embeddings improves link prediction.

To explicitly capture local neighbourhood structures through message passing, we implemented a GNN model. Given the large-scale and hub-heavy nature of the network, we employed neighbour sampling to enable efficient training, initializing the node representations with the topological vectors from the Baseline model. This architecture utilizes a 2-layer GraphSAGE encoder[46] with neighbour sampling to compute node embeddings based on the graph topology at $T_{\text{start,train}}$. A multilayer perceptron decoder was employed to classify the concatenated node embeddings.

Finally, we constructed a Mixture of GNN and Embeddings model as a third hybrid model, analogous to the Mixture of Baseline and Embeddings approach described above. This ensemble calculates a weighted average (1:1) of the output probabilities from the GNN and the Concept embeddings (MatSciBERT) models.

To avoid overfitting, we monitored the AUC—a summary metric we derive from the ROC—on a potentially out-of-distribution validation set with $T_{\text{start,validation}} = 2020$ and $T_{\text{end,validation}} = 2022$.

Human domain experts

The human domain experts who participated in the interviews were affiliated with different institutes at different institutions and were recruited to cover a wide range of topics within materials science. Of the 13 researchers who were invited, 10 agreed to participate in the study and the interviews. All participants were professors or independent group leaders. No interviews were excluded from the study.

Data availability

The minimum dataset required to interpret, verify and extend the results of this study (including the processed graph, feature vectors, models and their evaluation metrics) is available via figshare at https://doi.org/10.6084/m9.figshare.29315819 (refs. 47).

Code availability

The code used to produce the results of this study is publicly available via Zenodo at https://doi.org/10.5281/zenodo.18466587 (ref. 48). The source code is licenced under the GPL 3.0+ Licence and the active development repository can be accessed at https://github.com/aimat-lab/materials_concepts.

References

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

2. Varshney, L. R. et al. A big data approach to computational creativity: the curious case of Chef Watson. *J. Res. Dev.* **63**, 7:1–7:18 (2019).
3. Pinel, F., Varshney, L. & Bhattacharjya, D. in *Computational Creativity Research: Towards Creative Machines* (eds Besold, T. et al.) Ch. 16 (Atlantis, 2014).
4. Dagdelen, J. et al. Structured information extraction from scientific text with large language models. *Nat. Commun.* **15**, 1418 (2024).
5. Evans, J. A. & Rzhetsky, A. Advancing science through mining libraries, ontologies, and communities. *J. Biol. Chem.* **286**, 23659–23666 (2011).
6. Vaswani, A. et al. Attention is all you need. *Adv. Neural Inf. Process. Syst.* **30**, 5998–6008 (2017).
7. Brown, T. B. et al. Language models are few-shot learners. *Adv. Neural Inf. Process. Syst.* **33**, 1877–1901 (2020).
8. Torfi, A., Shirvani, R. A., Keneshloo, Y., Tavaf, N. & Fox, E. A. Natural language processing advancements by deep learning: a survey. Preprint at https://arxiv.org/abs/2003.01200 (2021).
9. Kojima, T., Gu, S. S., Reid, M., Matsuo, Y. & Iwasawa, Y. Large language models are zero-shot reasoners. *Adv. Neural Inf. Process. Syst.* **35**, 22199–22213 (2022).
10. Bubeck, S. et al. Sparks of artificial general intelligence: early experiments with GPT-4. Preprint at https://arxiv.org/abs/2303.12712 (2023).
11. Norouzi, E., Hertling, S. and Sack, H. ConExion: concept extraction with large language models. Preprint at https://arxiv.org/abs/2504.12915 (2025).
12. Krenn, M. & Zeilinger, A. Predicting research trends with semantic and neural networks with an application in quantum physics. *Proc. Natl Acad. Sci. USA* **117**, 1910–1916 (2020).
13. Rose, S., Engel, D., Cramer, N. and Cowley, W. in *Text Mining: Applications and Theory* (eds Berry, M. W. & Kogan J.) Ch. 1 (Wiley, 2010).
14. Krenn, M. et al. Forecasting the future of artificial intelligence with machine learning-based link prediction in an exponentially growing knowledge network. *Nat. Mach. Intell.* **5**, 1326–1335 (2023).
15. Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proc. 2019 Conference of the North American Chapter of the ACL: Human Language Technologies* Vol. 1, 4171–4186 (ACL, 2019).
16. Tenney, I., Das, D. & Pavlick, E. BERT rediscovers the classical NLP pipeline. In *Proc. 57th Annual Meeting of the ACL* 4593–4601 (ACL, 2019).
17. Gupta, T., Zaki, M. & Krishnan, N. M. A. & Mausam, MatSciBERT: a materials domain language model for text mining and information extraction. *npj Comput. Mater.* **8**, 102 (2022).
18. Ghafarollahi, A. & Buehler, M. J. SciAgents: automating scientific discovery through bioinspired multi-agent intelligent graph reasoning. *Adv. Mater.* **37**, 2413523 (2025).
19. Buehler, M. J. In situ graph reasoning and knowledge expansion using Graph-PRefLexOR. *Adv. Intell. Discov.* **1**, e202500006 (2025).
20. Gu, X. & Krenn, M. Interesting scientific idea generation using knowledge graphs and LLMs: evaluations with 100 research group leaders. Preprint at https://arxiv.org/abs/2405.17044 (2025).
21. Wang, Q., Downey, D., Ji, H. & Hope, T. SciMON: scientific inspiration machines optimized for novelty. In *Proc. 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (eds Ku L.-W. et al.) 279–299 (ACL, 2024).
22. Baek, J., Jauhar, S. K., Cucerzan, S. & Hwang, S. J. ResearchAgent: iterative research idea generation over scientific literature with large language models. In *Proc. 2025 Conference of the North American Chapter of the ACL: Human Language Technologies* (ACL, 2025).
23. Keya, F. et al. SCI-IDEA: context-aware scientific ideation using token and sentence embeddings. Preprint at https://arxiv.org/abs/2503.19257 (2025).
24. Schmidhuber, J. Artificial scientists & artists based on the formal theory of creativity. In *Proc. 3rd Conference on Artificial General Intelligence* (eds Goertzel, B. et al.) 148–153 (Atlantis, 2010).
25. McInnes, L., Healy, J., Saul, N. & Großberger, L. UMAP: uniform manifold approximation and projection. *J. Open Source Softw.* **3**, 861 (2018).
26. Pedregosa, F. et al. Scikit-learn: machine learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).
27. Fawcett, T. ROC Graphs: notes and practical considerations for researchers. *Mach. Learn.* **31**, 1–38 (2004).
28. DeLong, E. R., DeLong, D. M. & Clarke-Pearson, D. L. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics* **44**, 837–845 (1988).
29. Liben-Nowell, D. & Kleinberg, J. The link-prediction problem for social networks. *J. Ame. Soc. Inf. Sci.* **58**, 1019–1031 (2007).
30. Grover, A. & Leskovec, J. node2vec: scalable feature learning for networks. In *Proc. 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (eds Krishnapuram, B. et al.) 855–864 (ACM, 2016).
31. Wu, J. et al. Inverse design workflow discovers hole-transport materials tailored for perovskite solar cells. *Science* **386**, 1256–1264 (2024).
32. Rossi, D. et al. On the importance of ferroelectric domains for the performance of perovskite solar cells. *Nano Energy* **48**, 20–26 (2018).
33. Röhm, H., Leonhard, T., Hoffmann, M. & Colsmann, A. Ferroelectric domains in methylammonium lead iodide perovskite thin-films. *Energy Environ. Sci.* **10**, 950–955 (2017).
34. Schulz, A., Braun, M., Colsmann, A., Hinterstein, M. & Röhm, H. Ferroelectricity and crystal phases in mixed-cation lead iodide perovskite solar cells. *Solar RRL* **6**, 2200808 (2022).
35. Röhm, H. et al. Ferroelectric properties of perovskite thin-films and their implications for solar energy conversion. *Adv. Mater.* **31**, 1806661 (2019).
36. Breternitz, J., Lehmann, F., Barnett, S. A., Nowell, H. & Schorr, S. Role of the iodide-methylammonium interaction in the ferroelectricity of $CH_3NH_3PbI_3$. *Angew. Chem. Int. Ed.* **59**, 424–428 (2020).
37. Röhm, H., Leonhard, T., Hoffmann, M. J. & Colsmann, A. Ferroelectric poling of methylammonium lead iodide thin films. *Adv. Funct. Mater.* **30**, 1908657 (2020).
38. Zhou, C. et al. LIMA: less is more for alignment. In *Proc. Advances in Neural Information Processing Systems 36* (eds Oh, A. et al.) 55006–55021 (Curran Associates, 2023).
39. Priem, J., Piwowar, H. & Orr, R. OpenAlex: a fully-open index of scholarly works, authors, venues, institutions, and concepts. Preprint at https://arxiv.org/abs/2205.01833 (2022).
40. Touvron, H. et al. LLaMA: open and efficient foundation language models. Preprint at https://arxiv.org/abs/2302.13971 (2023).
41. Touvron, H. et al. Llama 2: open foundation and fine-tuned chat models. Preprint at https://arxiv.org/abs/2307.09288 (2023).
42. Dettmers, T., Lewis, M., Belkada, Y. & Zettlemoyer, L. LLM.int8(): 8-bit matrix multiplication for transformers at scale. *Adv. Neural Inf. Process. Syst.* **35**, 30318–30332 (2022).
43. Hu, E. J. et al. LoRA: low-rank adaptation of large language models. In *Proc. 10th International Conference on Learning Representations* (ICLR, 2022).
44. Dettmers, T., Pagnoni, A., Holtzman, A. & Zettlemoyer, L. QLoRA: efficient finetuning of quantized LLMs. *In Proc. Advances in Neural Information Processing Systems 36* (eds Oh, A. et al.) 10088–10115 (Curran Associates, 2023).

45. Mangrulkar, S. et al. Peft: state-of-the-art parameter-efficient fine-tuning methods. *GitHub* https://github.com/huggingface/peft (2022).
46. Hamilton, W. L., Ying, Z. & Leskovec, J. Inductive representation learning on large graphs. In *Proc. Advances in Neural Information Processing Systems 30* (eds Guyon, I. et al.) 1024–1034 (Curran Associates, 2017).
47. Marwitz, T. Data for 'predicting new research directions in materials science using llms and concept graphs'. *figshare* https://doi.org/10.6084/m9.figshare.29315819 (2026).
48. Marwitz, T. aimat-lab/materials_concepts: code for paper. *Zenodo* https://doi.org/10.5281/zenodo.18466587 (2026).

Acknowledgements

Creating a chemical element parser from scratch is hard. We thank Chris Konop for providing an outline of how a chemical formula can be recursively defined. This made our implementation much more straightforward. We acknowledge support from the Federal Ministry of Education and Research (BMBF) under Grant No. 01DM21001B (German-Canadian Materials Acceleration Center) (P.F.). We acknowledge funding by the German Research Foundation (DFG) under Germany's Excellence Strategy via the Excellence Cluster '3D Matter Made to Order' (3DMM2O, EXC-2082/1-390761711) (P.F.). Funding by the German Research Foundation (DFG) through the collaborative research center CRC 1249 N-Heteropolycycles as Functional Materials (SFB 1249, Project C13) and through DFG Projekt 436506789 is gratefully acknowledged (P.F.). This work was partly carried out with the support of the Karlsruhe Nano Micro Facility (KNMFi, www.knmf.kit.edu), a Helmholtz Research Infrastructure at Karlsruhe Institute of Technology (KIT, www.kit.edu) (B.B.). We acknowledge funding from the Helmholtz Metadata Collaboration (HMC) through the project AIMWORKS (P.F.). This work was performed on the HoreKa supercomputer funded by the Ministry of Science, Research and the Arts Baden-Württemberg and by the Federal Ministry of Education and Research. We acknowledge support from the state of Baden-Württemberg through bwHPC (P.F.).

Author contributions

T.M., T.S. and P.F. wrote the first draft of the paper. All authors commented on previous versions of the paper, and all authors read and approved the final paper. P.F. worked on conceptualization and supervision. T.M. and P.F. worked on the methodology and formal analysis. All authors contributed to the investigation.

Funding

Open access funding provided by Karlsruher Institut für Technologie (KIT).

Competing interests

The authors declare no competing interests.

Additional information

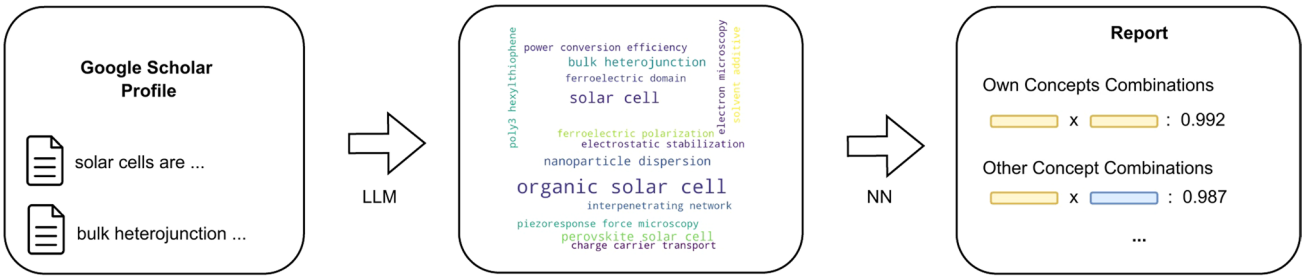
Extended data is available for this paper at https://doi.org/10.1038/s42256-026-01206-y.

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s42256-026-01206-y.

Correspondence and requests for materials should be addressed to Pascal Friederich.

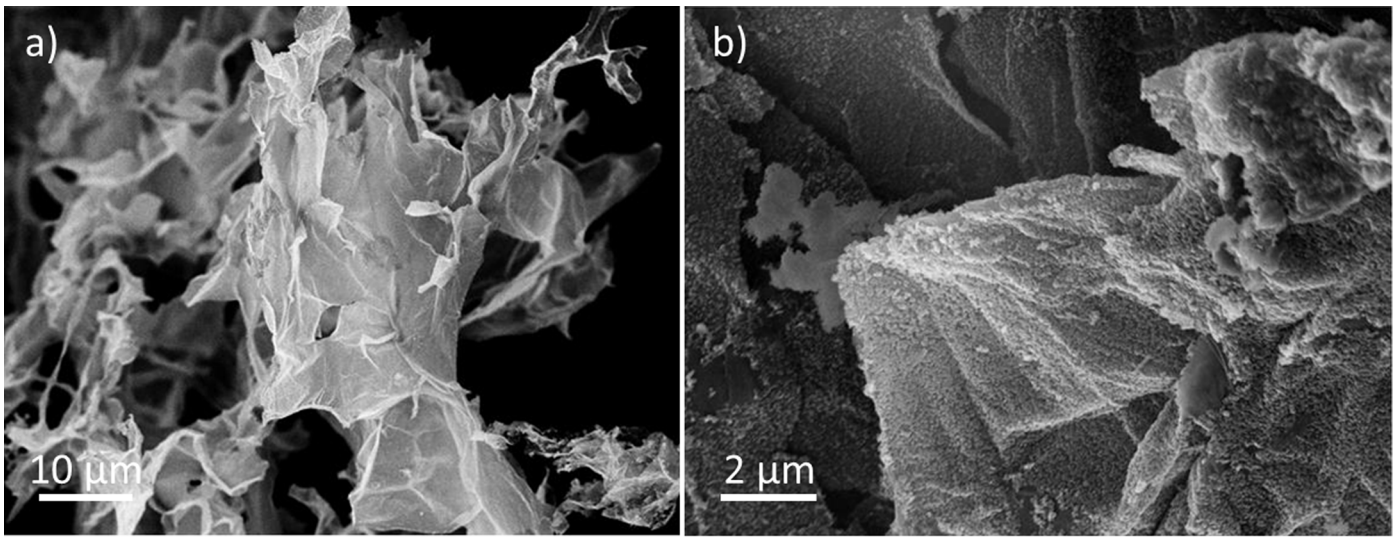
Peer review information *Nature Machine Intelligence* thanks Anurag Bajpai and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Peer reviewer reports are available.

https://doi.org/10.1038/s42256-026-01206-y



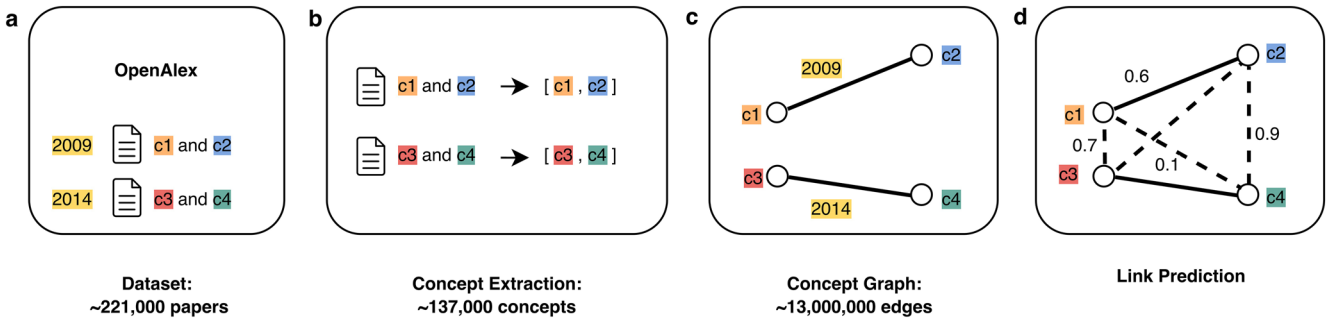
Extended Data Fig. 1 | Overview of the report generation. (**1**) Selection of abstracts of recent publications, (**2**) Extraction of individual concepts using our fine-tuned LLM, and (**3**) Suggestion of combinations in a standardized report based on our best ML model for link prediction.

https://doi.org/10.1038/s42256-026-01206-y



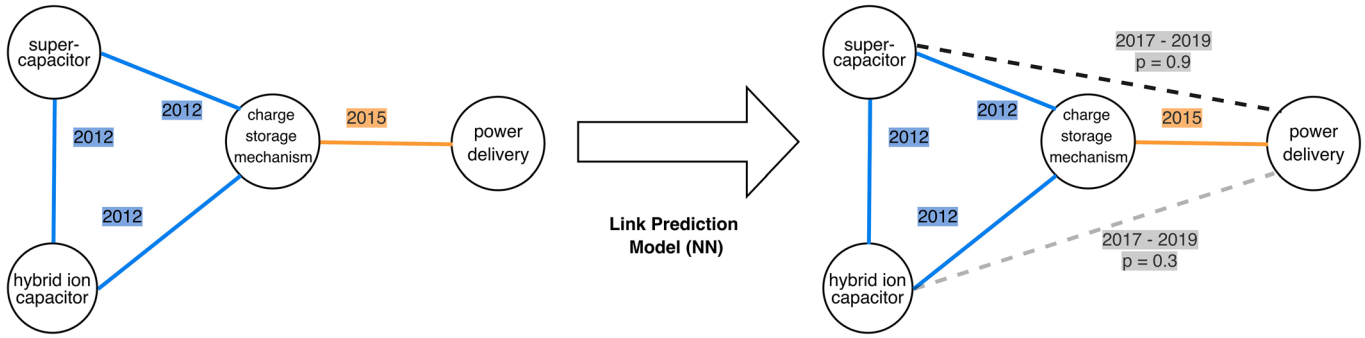
Extended Data Fig. 2 | Exfoliated graphene oxide (resulting in multilayer graphene) covered with 200 nm thick iron oxide layers. (**a**) flake morphology of the exfoliated layers, (**b**) iron oxide layer covering the whole surface.

https://doi.org/10.1038/s42256-026-01206-y



Extended Data Fig. 3 | Overview of the link prediction workflow. (**a**) Gathering of materials science abstracts with OpenAlex[39], (**b**) Extraction of concepts using LLMs, (**c**) creation of the semantics-aware concept graph, and (**d**) prediction of new research directions.

https://doi.org/10.1038/s42256-026-01206-y



Extended Data Fig. 4 | Example of a concept graph. The concept graph is derived from two articles published 2012 (blue) in and 2015 (orange) with one overlapping concept. Possible new edges are marked as dotted grey lines together with their predicted probability of formation in the years 2017 to 2019.

https://doi.org/10.1038/s42256-026-01206-y

Extended Data Table 1 | Confusion Matrix Classes

Model	d_{prev}	Positives			Negatives	
		TP	FN	Recall	FP	TN
Baseline	2	212	78	73.1%	115,807	751,021
	3	1	16	5.9%	5,342	1,124,056
	4	0	0	–	0	3,467
Concept embeddings	2	203	87	70.0%	120,575	746,253
	3	6	11	35.3%	26,209	1,103,189
	4	0	0	–	27	3,440
GNN	2	253	37	87.2%	223,637	643,191
	3	3	14	17.6%	31,512	1,097,886
	4	0	0	–	9	3,458
Mixture of GNN and Embeddings	2	248	42	85.5%	153,455	713,373
	3	5	12	29.4%	11,753	1,117,645
	4	0	0	–	6	3,461

Confusion matrix classes for the link predictions of the baseline, Concept Embeddings, GNN, and GNN-Mixture models split by previous node distances.