The Largest Knowledge Graph in Materials Science - Entities, Relations, and Link Prediction through Graph Representation Learning

Anonymous Author(s)
Affiliation
Address
email

Abstract

This paper introduces MatKG, a novel graph database of key concepts in material science spanning the traditional material-structure-property-processing paradigm. MatKG is autonomously generated through transformer-based, large language models and generates pseudo ontological schema through statistical co-occurrence mapping. At present, MatKG contains over 2 million unique relationship triples derived from 80,000 entities. This allows the curated analysis, querying, and visualization of materials knowledge at unique resolution and scale. Further, Knowledge Graph Embedding models are used to learn embedding representations of nodes in the graph which are used for downstream tasks such as link prediction and entity disambiguation. MatKG allows the rapid dissemination and assimilation of data when used as a knowledge base, while enabling the discovery of new relations when trained as an embedding model.

1 Introduction

Comprehensive knowledge of a given material requires the integration of disparate streams of information that include compositional data, thermodynamic parameters, applications, phase/symmetry labels, synthesis and processing routines, as well as physical, chemical, thermal, optical, and functional properties[1]. In general, it is difficult to find all this information in one place, with the result that comprehensive knowledge of a given material is often missing, even when the data are available. Given the rate at which new data are being accumulated, the amount of available data is far greater than what can be accessed or assimilated. The standard paradigm of data sharing and storage - through peer reviewed scientific publications and relational databases - remains inadequate for the Materials Genome Age where artificial intelligence is increasingly employed to accelerate materials discovery and design [2, 3, 4, 5]. The task of data organization has been approached through custom ontologies that build relations between data points through manual expert input. While several domain specific ontologies such as Nanomine[6], Chemos[7], etc. have been written over the years, no field-wide ontology exists focused on materials science. Given the onerous task of assigning a relation among individual pairs of data, even highly generalizable ontologies such as SKOS[8] have not been applied to materials at scale.

In this paper, we introduce MatKG, a novel graph database that links major conceptual entities in the discipline using transformer-based large language models. The database is autonomously extracted from over 4 million papers on the topic of materials and includes chemistry, structure, property, application, synthesis, and characterization data that are aggregated in the form of relational triples <subject, predicate, object>. MatKG has over 2 million unique relationships among over 80,000 unique entities.

2 Methods

Entity Generation: A Named Entity Recognition (NER) [9] model was used to extract 80,000 unique entities from the abstracts and figure captions of over 4 million scientific publications [10] in the field of material science. Being information dense, these contain low 'noise' and are hence particularly suitable for large scale autonomous data mining[11, 12, 13]. The NER model follows the scheme developed in MatScholar [14] and is built on MatBERT [14], a Large Language Model (LLM) trained on a material science text corpus that classifies text tokens into one of the following seven categories: Material (CHM), Property (PRO), Application (APL), Synthesis Method (SYN), Characterization Method (CMT), Descriptor (DSC), and Symmetry/Phase Label (SPL). Derived from the traditional structure-property-processing-application paradigm in material science[1], these entities encapsulate the sum total of the knowledge of any given concept, be it a particular chemistry, process, property, or application. Where possible, each entity is linked to an identifier in Wikipedia using procedure developed in [15] or the corresponding descriptor page in the Materials Project[16]. This allows the mapping of entities to broader knowledge bases such as DBpedia[17] and YAGO[18], thereby allowing holistic integration of MatKG with the larger knowledge graph community.

Link Generation : If entities e_1 and e_2 have the NER tags $T[e_1]$ and $T[e_2]$, they are assigned the relationship $T[e_1]\_T[e_2]$ and the weight $v(e_1, e_2)$ according to the method detailed in Appendix 5.1. Subsequently, they are either filtered based on a predefined threshold to form knowledge triples of the form <e_1, $T[e_1]\_T[e_2]$, e_2> (1) or as a quartet of the form <e_1, $T[e_1]\_T[e_2]$, e_2, $v(e_1, e_2)$> (2) (See Appendix 5.1). (1) allows the extraction of 160,000 high fidelity links between about 12,000 unique entities, while (2) results in 2 million relations from up to 80,000 unique entities, thereby demonstrating that a weighted link extraction approach captures far more data - increasing the scope of the knowledge base.

Graph Representation Learning : The vector representations for the entities in the graph are learnt using knowledge graph embedding models (KGE)[19],[20], [21]. The models are evaluated using mean reciprocal rank (MRR) and hits@(1,10,100) metrics on the test set as described in KGE literature [22]. All models are implemented using the publicly available AmpliGraph Library[23]. The model with the highest MRR on the test set was used to perform downstream tasks that are described later.

3 Results

3.1 Knowledge base creation

The autonomously created highly interconnected knowledge graph for materials consists of the seven NER categories and 49 relations (including inverse relations such as APL_PRO and PRO_APL). The KG is thus a bidirectional digraph. The three most common types of entities are PRO, CHM, and CMT, while the most frequent relations are CHM_PRO, PRO_DSC, and CHM_CMT (see Appendix, Table 1, 2). The large number of material-property (108 k) and material-application (89 k) triples could correspond to the type of information usually present in abstracts, while characterization related information originate from figure captions. Many papers in the corpus relate to inorganic synthesis[10] which explains the high number of SMT_CHM (80 k) and SMT_PRO (67 k) relations.

Together, the acquired data allows the extraction of subgraphs corresponding to wildcard triples such as <TiO_2, CHM_PRO, ?>, which correspond to the customized query: "what are the properties of TiO_2?". Further, by accounting for the co-occurrence frequency, a confidence score can be assigned to each triple as is visually represented in Fig 1(a, b) where the applications and phase labels of TiO_2 are separately extracted and presented as individual bipartite graphs such that the size of the node is proportional to $v(TiO_2, e)$. We see that the most common symmetry/phase labels associated with TiO_2 are 'rutile' and 'anastase', while the most frequent applications are as electrodes, catalyts and for coating. These are in agreement with the widely available literature on the material[24]. There is much less information on CdTe by comparison (18153 vs 1500 triples), but Fig 1(c, d) extracted from MatKG still enables a high-level understanding with some specificity, such as the knowledge that CdTe is used in solar cells and electrodes, and is an optical material as deduced from its properties[25].

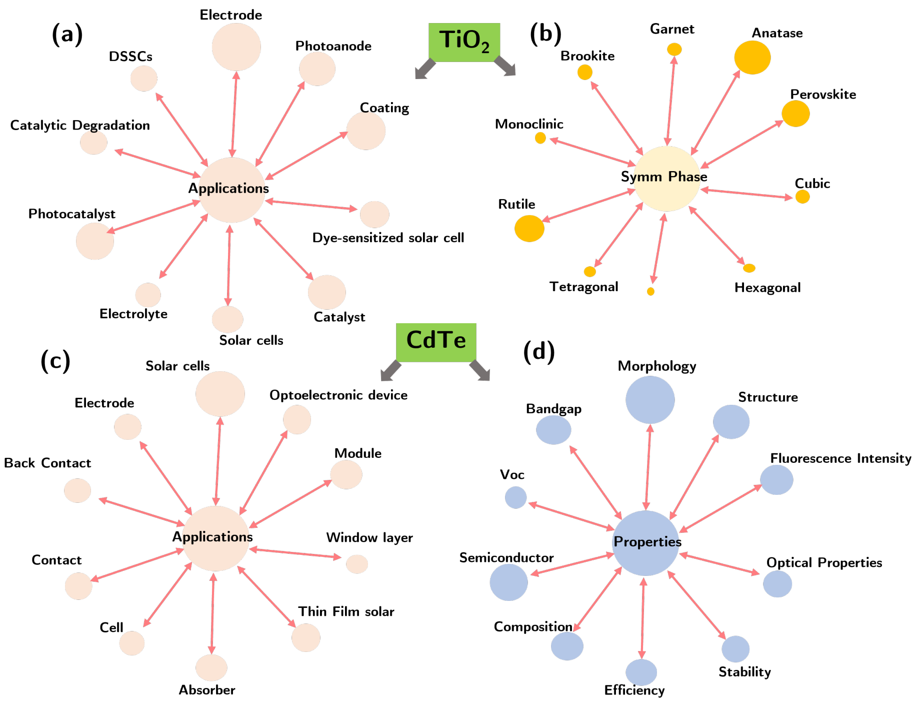

Figure 1: (a) Applications and (b) Symmetry Phase Labels of TiO_2. (c) Applications and (d) Properties of CdTe. The size of the node is proportional to the co-occurence frequency of the link.

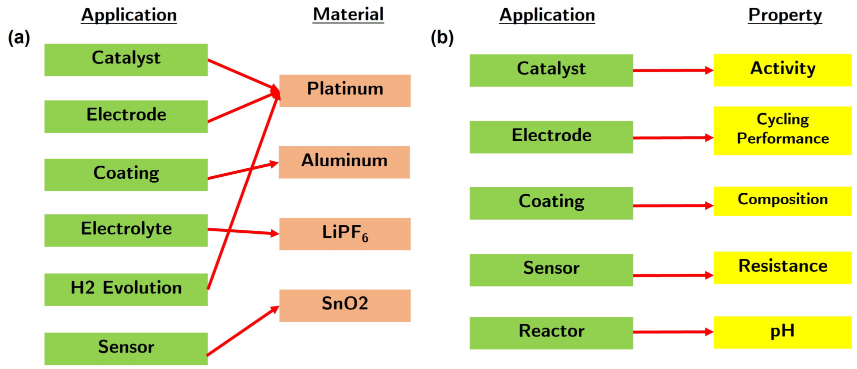

Figure 2: Partitioned (a) application-material (b) application-property subgraphs showing the highest weighted material and property for some select applications

In addition to material specific queries, MatKG can be partitioned into relation specific subgraphs such as the application-material and application – property graphs in Fig 2 (a-b), which shows the highest weighted material and property respectively for some select applications. Platinum, perhaps the most widely used metallic catalyst, appears with both ‘catalyst’ and ’hydrogen evolution’. Aluminum is identified as a coating material while $LiPF_6$ is seen to be an electrolyte, both of which are well known applications of each respectively. In Fig 2 (b) the most common property associated with electrodes is ’cycling performance’, while that of catalyst is ‘activity’. Both are in accordance with our understanding of these concepts. Therefore, MatKG allows the curated visualization and querying of materials specific data directly extracted from literature at unprecedented scale and resolution.

Table 1: Selected entities and their similarities, demonstrating semantic convergence at the embedding level

Entities	Similarity
(qspr, quantitative structure property relationship)	0.90
(qmom, quadrature method of moments)	0.91
(electromagnetic acoustic resonance, emar)	0.89
(ner, net energy ratio)	0.92
(let, linear energy transfer)	0.91

3.2 Embedding representation learning

The TransE[19] model with 150 dimensions is found to have the highest MRR (0.49) on the test set. This model was chosen for discovering new links and for performing entity disambiguation.

Entity Linking: The similarity between embeddings can be used as a measure of the semantic similarity between entities, in turn becoming a useful tool for both co-reference resolution as well as similar – chemical mapping. As shown in Table 1, several pairs of entities such as 'qspr' and 'quantitative structure property relationship', or 'ner' and 'net energy ratio' occupy almost identical positions in MatKG and consequently have very similar graph embeddings. This suggests that they are the same semantic token, even though their lexical distance can be substantial. This form of co-reference resolution is currently not an easy task, especially for the sciences[26].

Link Prediction: Finally, the KGE model was used to make new link predictions between existing entities in the graph. In this way, the model can be used to discover new applications and properties of existing materials, new properties that can be useful to a given application, or a new characterization method for an existing property, etc. This results in a fuller and more integrated knowledge graph, allowing a holistic analysis of structure-property-processing relations, even when such data is absent in the training literature.

While the MRR and hits@(1,10,100) are good measures of link predictiveness of the model[27], it is desirable to quantify this inference ability further. To this end, 150 random link predictions were generated by the model across all relationship categories. The top three entities with the highest score for each prediction is manually ranked according to the following criteria: Rank1 if the relationship can be classified as of type *SKOS: Narrow*, Rank 2 if it is of type *SKOS: Broad*, and 3 otherwise, where 'narrow' and 'broad' are ontological schema specificed in SKOS [8]. An example triple is shown in Table 2, Appendix, which lists the top three model predictions for the applications of Fe_2O_3. Some lithium-ion batteries use lithium-iron-oxide as an electrode, which is usually made by the solid-state reaction of Li_2CO_3 and Fe_2O_3, which could explain the first prediction. Since 'lithium-ion batteries' is not a direct application of Fe_2O_3, this triple is ranked 2. However, 'air batteries' directly use iron/iron oxide as an electrode[28] and hence this triple is assigned rank 1.

Of the 150 x 3 predictions made by the model, 47 % were found to have a rank 1, 29 % had a rank of 2, and the rest had a rank of 3 (See Appendix, Table 3 for examples). The utility of this approach is seen in Fig 3, Appendix where previously empty application and characterization subgraphs of Bismuth Telluride (as extracted from MatKG) are populated with meaningful entities through successful link prediction.

4 Broader Impact

MatKG is the first step towards the complete synthesis of materials knowledge that allows for the richer databases not just for materials but also for applications, properties, and characterization methods. The ability to predict new links between entities in the graph allows the discovery of new materials for existing applications and properties, in finding new applications of existing materials, and novel correlations between synthesis, characterizations and properties. Consequently, MatKG has broad impact for all the three categories of **AI-guided materials design**, **Automated Synthesis** and for **Automated Characterization** .

References

[1] D William and J Callister. The structure of crystalline solids. In *Materials Science and Engineering*, pages 64–65. Wiley, 1989.

[2] Vahe Tshitoyan, John Dagdelen, Leigh Weston, Alexander Dunn, Ziqin Rong, Olga Kononova, Kristin A Persson, Gerbrand Ceder, and Anubhav Jain. Unsupervised word embeddings capture latent knowledge from materials science literature. *Nature*, 571(7763):95–98, 2019.

[3] Alden Dima, Sunil Bhaskarla, Chandler Becker, Mary Brady, Carelyn Campbell, Philippe Dessauw, Robert Hanisch, Ursula Kattner, Kenneth Kroenlein, Marcus Newrock, et al. Informatics infrastructure for the materials genome initiative. *Jom*, 68(8):2053–2064, 2016.

[4] Juan J de Pablo, Nicholas E Jackson, Michael A Webb, Long-Qing Chen, Joel E Moore, Dane Morgan, Ryan Jacobs, Tresa Pollock, Darrell G Schlom, Eric S Toberer, et al. New frontiers for the materials genome initiative. *npj Computational Materials*, 5(1):1–23, 2019.

[5] Juan J De Pablo, Barbara Jones, Cora Lind Kovacs, Vidvuds Ozolins, and Arthur P Ramirez. The materials genome initiative, the interplay of experiment, theory and computation. *Current Opinion in Solid State and Materials Science*, 18(2):99–117, 2014.

[6] James P McCusker, Neha Keshan, Sabbir Rashid, Michael Deagen, Cate Brinson, and Deborah L McGuinness. Nanomine: A knowledge graph for nanocomposite materials science. In *International Semantic Web Conference*, pages 144–159. Springer, 2020.

[7] Loïc M Roch, Florian Häse, Christoph Kreisbeck, Teresa Tamayo-Mendoza, Lars PE Yunker, Jason E Hein, and Alán Aspuru-Guzik. Chemos: orchestrating autonomous experimentation. *Science Robotics*, 3(19):eaat5559, 2018.

[8] Alistair Miles and Sean Bechhofer. Skos simple knowledge organization system reference. *W3C recommendation*, 2009.

[9] David Nadeau and Satoshi Sekine. A survey of named entity recognition and classification. *Lingvisticae Investigationes*, 30(1):3–26, 2007.

[10] Edward Kim, Kevin Huang, Stefanie Jegelka, and Elsa Olivetti. Virtual screening of inorganic materials synthesis parameters with deep learning. *npj Computational Materials*, 3(1):1–9, 2017.

[11] Vineeth Venugopal, Sourav Sahoo, Mohd Zaki, Manish Agarwal, Nitya Nand Gosvami, and NM Anoop Krishnan. Looking through glass: Knowledge discovery from materials science literature using natural language processing. *Patterns*, 2(7):100290, 2021.

[12] Vineeth Venugopal, Scott R Broderick, and Krishna Rajan. A picture is worth a thousand words: applying natural language processing tools for creating a quantum materials database map. *MRS Communications*, 9(4):1134–1141, 2019.

[13] Vineeth Venugopal, Suresh Bishnoi, Sourabh Singh, Mohd Zaki, Hargun Singh Grover, Mathieu Bauchy, Manish Agarwal, and NM Anoop Krishnan. Artificial intelligence and machine learning in glass science and technology: 21 challenges for the 21st century. *International journal of applied glass science*, 12(3):277–292, 2021.

[14] Leigh Weston, Vahe Tshitoyan, John Dagdelen, Olga Kononova, Amalie Trewartha, Kristin A Persson, Gerbrand Ceder, and Anubhav Jain. Named entity recognition and normalization applied to large-scale information extraction from the materials science literature. *Journal of chemical information and modeling*, 59(9):3692–3702, 2019.

[15] Valentin I Spitkovsky and Angel X Chang. A cross-lingual dictionary for english wikipedia concepts. 2012.

[16] Anubhav Jain, Shyue Ping Ong, Geoffroy Hautier, Wei Chen, William Davidson Richards, Stephen Dacek, Shreyas Cholia, Dan Gunter, David Skinner, Gerbrand Ceder, et al. Commentary: The materials project: A materials genome approach to accelerating materials innovation. *APL materials*, 1(1):011002, 2013.

[17] Sören Auer, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak, and Zachary Ives. Dbpedia: A nucleus for a web of open data. In *The semantic web*, pages 722–735. Springer, 2007.

[18] Fabian M Suchanek, Gjergji Kasneci, and Gerhard Weikum. Yago: a core of semantic knowledge. In *Proceedings of the 16th international conference on World Wide Web*, pages 697–706, 2007.

[19] Antoine Bordes, Nicolas Usunier, Alberto Garcia-Duran, Jason Weston, and Oksana Yakhnenko. Translating embeddings for modeling multi-relational data. In *NIPS*, pages 2787–2795, 2013.

[20] Bishan Yang, Scott Wen-tau Yih, Xiaodong He, Jianfeng Gao, and Li Deng. Embedding entities and relations for learning and inference in knowledge bases. In *ICLR*, 2015.

[21] Théo Trouillon, Johannes Welbl, Sebastian Riedel, Éric Gaussier, and Guillaume Bouchard. Complex embeddings for simple link prediction. In *ICML*, pages 2071–2080, 2016.

[22] Hongyun Cai, Vincent W Zheng, and Kevin Chen-Chuan Chang. A comprehensive survey of graph embedding: Problems, techniques, and applications. *IEEE Transactions on Knowledge and Data Engineering*, 30(9):1616–1637, 2018.

[23] Luca Costabello, Sumit Pai, Chan Le Van, Rory McGrath, Nicholas McCarthy, and Pedro Tabacof. AmpliGraph: a Library for Representation Learning on Knowledge Graphs, 2019. URL `https://doi.org/10.5281/zenodo.2595043`.

[24] Qing Guo, Chuanyao Zhou, Zhibo Ma, and Xueming Yang. Fundamentals of tio2 photocatalysis: concepts, mechanisms, and challenges. *Advanced Materials*, 31(50):1901997, 2019.

[25] SH Shin, J Bajaj, LA Moudy, and DT Cheung. Characterization of te precipitates in cdte crystals. *Applied Physics Letters*, 43(1):68–70, 1983.

[26] Ozlem Uzuner, Andreea Bodnari, Shuying Shen, Tyler Forbush, John Pestian, and Brett R South. Evaluating the state of the art in coreference resolution for electronic medical records. *Journal of the American Medical Informatics Association*, 19(5):786–791, 2012.

[27] Vivek Khetan, Erin Wetherley, Elena Eneva, Shubhashis Sengupta, Andrew E Fano, et al. Knowledge graph anchored information-extraction for domain-specific insights. *arXiv preprint arXiv:2104.08936*, 2021.

[28] J Requies, MB Güemez, S Perez Gil, VL Barrio, JF Cambra, U Izquierdo, and PL Arias. Natural and synthetic iron oxides for hydrogen storage and purification. *Journal of Materials Science*, 48(14):4813–4822, 2013.

Checklist

For all authors...

1. Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
2. Did you describe the limitations of your work? [Yes]
3. Did you discuss any potential negative societal impacts of your work? [Yes]
4. Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]

If you are including theoretical results...

1. Did you state the full set of assumptions of all theoretical results? [N/A]
2. Did you include complete proofs of all theoretical results? [N/A]

If you ran experiments...

1. Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [No] The data is open-source and freely available. The code and project are still a work-in-progress and code will be released upon full publication of this work at a later date.
2. Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [No] This will done upon the full publication of this work.

3. Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [N/A]
4. Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [N/A]

If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...

1. If your work uses existing assets, did you cite the creators? [Yes]
2. Did you mention the license of the assets? [Yes]
3. Did you include any new assets either in the supplemental material or as a URL? [N/A]
4. Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
5. Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]

If you used crowdsourcing or conducted research with human subjects...

1. Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
2. Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
3. Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]

5 Appendix

5.1 KG construction

For every entity e, the lexical frequency $L(e)$ is defined as the fraction of documents where e is present at least once, where a document could either be one of N_a abstracts or N_c figure captions. For every pair of entities (e_1, e_2) in a given document, a co-occurrence function $CO(e_1, e_2)$ is defined such that:

$$CO(e_1, e_2) = \begin{cases} 1 & \text{if both } e_1 \text{ and } e_2 \text{ present in the document} \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

The co-occurrence frequency $v(e_1, e_2)$ is then defined as :

$$v(e_1, e_2) = \frac{\sum^{N_a+N_c} CO(e_1, e_2)}{N_a + N_c} \tag{2}$$

$v(e_1, e_2)$ therefore is a measure of how many times the given pair of entities (e_1, e_2) co-occur in the document corpus. Subsequently, two approaches are employed to assign a link to (e_1, e_2). Approach (1) is based on the premise that if $v(e_1, e_2) > \frac{L(e_1)*L(e_2)}{(N_a+N_c)^2}$, then the entities e_1 and e_2 are strongly correlated as they occur far more often than their conditional probabilities allow. Approach (2) however, retains all entity pairs but appends their co-occurrence frequency as a weight in the knowledge representation model

$$Relation(e_1, e_2) = \begin{cases} < e_1, T[\mathrm{e}_1]\_\mathrm{T}[e_2]\ , \mathrm{e}_2 > & v(\mathrm{e}_1, e_2) > \beta \text{ , where beta is a threshold} \\ < e_1, T[e_1]\_T[e_2], e_2,\mathrm{v}(e_1, e_2) > & v(\mathrm{e}_1, e_2) > \epsilon \text{ where } \epsilon \approx 10 \end{cases}$$

Table 2: NER Categories in MatKG and the number of unique entities in each category.

NER Category	Number of Entities
Property (PRO)	27048
Chemical (CHM)	23438
Characterization Method	10908
Synthesis Method	8547
Application	7009

Table 3: Selected relationships and their instance count in MatKG.

Relationship	Number of Triple
CHM_CHM	499994
PRO_PRO	368381
CHM_PRO	252714
PRO_DSC	146929
CMT_CHM	141955
CHM_DSC	139740
CMT_PRO	108233
CMT_CMT	100675
APL_PRO	91466
CHM_APL	89117
CHM_SMT	80349

Table 4: Model predictions for the triple <Fe_2O_3, CHM_PRO, X> where X is a property. The triples are ranked according to the scheme described in Results

Subject	relationship	Object	Rank
Fe_2O_3	CHM_APL	$lithium\ ion\ batteries$	2
Fe_2O_3	CHM_APL	$electrocatalyts$	1
Fe_2O_3	CHM_APL	$air\ batteries$	1

Table 5: Top three model predicted links for selected examples with model score, custom rank, and cited doi

Subject	relationship	Object	Score	Rank	Citation url
optical material	APL_CHM	In_2O_3	5.5	1	https://en.wikipedia.org/wiki/Indium(III)_oxide
optical material	APL_CHM	CdO	5.27	1	https://en.wikipedia.org/wiki/Cadmium_oxide
optical material	APL_CHM	Zinc Oxide	5.26	1	https://en.wikipedia.org/wiki/Zinc_oxide
anodic electrode	APL_CHM	Graphite	3.00	1	10.1016/j.ensm.2020.12.027
anodic electrode	APL_CHM	Carbon-fiber	3.00	1	10.1016/C2015-0-00574-3
anodic electrode	APL_CHM	$LiClO_4$	2.90	2	https://en.wikipedia.org/wiki/Lithium_perchlorate
nuclear reactor	APL_CHM	Beryllium	7.02	1	https://www.energy.gov/ehss/about-beryllium
nuclear reactor	APL_CHM	Carbide	6.41	2	https://en.wikipedia.org/wiki/Uranium_carbide
nuclear reactor	APL_CHM	Tungsten	6.38	1	10.1016/j.ijhydene.2016.02.019
smes	APL_PRO	dmain	0.34	3	N/A
smes	APL_PRO	transmitted current	0.28	1	https://en.wikipedia.org/wiki/Superconducting_magnetic_energy_storage
smes	APL_PRO	u11	0.20	3	N/A
reverse water gas shift reaction	APL_CHM	C_6H_5OH	5.22	3	N/A
reverse water gas shift reaction	APL_CHM	Naphtha	5.22	3	N/A
reverse water gas shift reaction	APL_CHM	diethylether	4.79	3	N/A

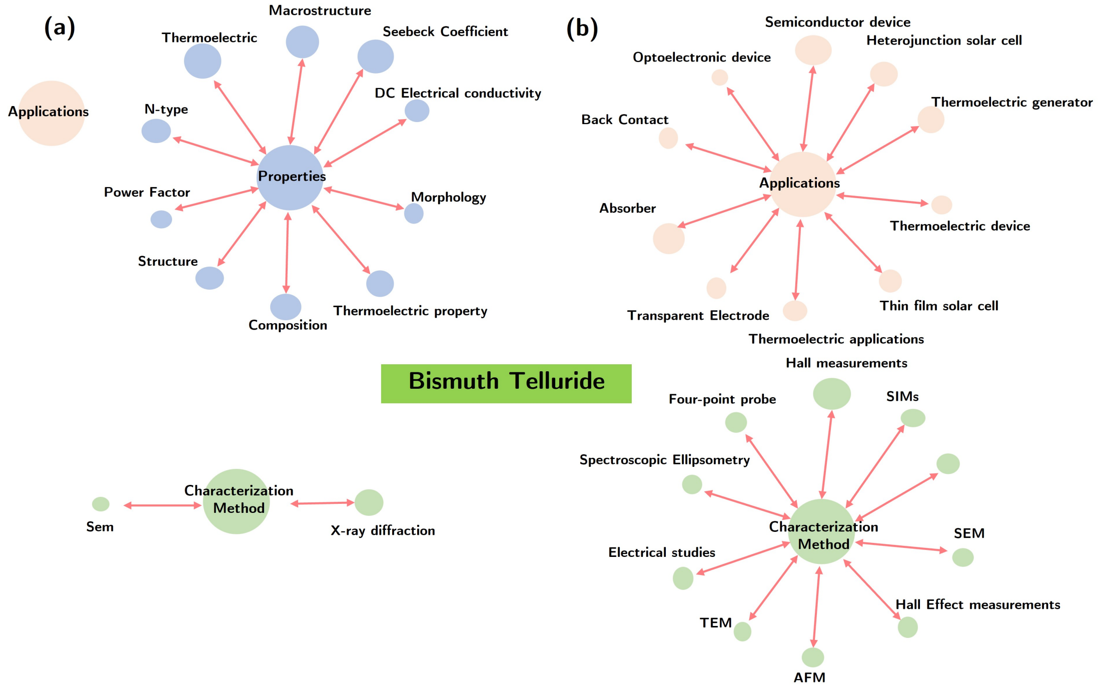

Figure 3: (a) Original Triples extracted from MatKG and (b) model predicted triples for Bismuth Telluride demonstrating the utility of KGE in complementing material knowledge bases