Assessing Scientific Research Papers with Knowledge Graphs

Kexuan Sun
University of Southern California
Information Sciences Institute
kexuansu@usc.edu

Zhiqiang Qiu
University of Southern California
Information Sciences Institute
qiuz@usc.edu

Abel Salinas
University of Southern California
Information Sciences Institute
abelsali@usc.edu

Yuzhong Huang
University of Southern California
Information Sciences Institute
yuzhongh@usc.edu

Dong-Ho Lee
University of Southern California
Information Sciences Institute
dongho.lee@usc.edu

Daniel Benjamin
Nova Southeastern University
dbenjam1@nova.edu

Fred Morstatter
University of Southern California
Information Sciences Institute
fredmors@isi.edu

Xiang Ren
University of Southern California
Information Sciences Institute
xiangren@usc.edu

Kristina Lerman
University of Southern California
Information Sciences Institute
lerman@isi.edu

Jay Pujara
University of Southern California
Information Sciences Institute
jpujara@isi.edu

ABSTRACT

In recent decades, the growing scale of scientific research has led to numerous novel findings. Reproducing these findings is the foundation of future research. However, due to the complexity of experiments, manually assessing scientific research is laborious and time-intensive, especially in social and behavioral sciences. Although increasing reproducibility studies have garnered increased attention in the research community, there is still a lack of systematic ways for evaluating scientific research at scale. In this paper, we propose a novel approach towards automatically assessing scientific publications by constructing a *knowledge graph* (KG) that captures a holistic view of the research contributions. Specifically, during the KG construction, we combine information from two different perspectives: *micro*-level features that capture knowledge from published articles such as sample sizes, effect sizes, and experimental models, and *macro*-level features that comprise relationships between entities such as authorship and reference information. We then learn low-dimensional representations using language models and knowledge graph embeddings for entities (nodes in KGs), which are further used for the assessments. A comprehensive set of experiments on two benchmark datasets shows the usefulness of leveraging KGs for scoring scientific research.

CCS CONCEPTS

• **Computing methodologies → Machine learning**; • **Applied computing → Law, social and behavioral sciences**.

KEYWORDS

Knowledge graph; Social and behavioral sciences; Reproducibility

Conference'17, July 2017, Washington, DC, USA
© 2022 Association for Computing Machinery.
ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

ACM Reference Format:
Kexuan Sun, Zhiqiang Qiu, Abel Salinas, Yuzhong Huang, Dong-Ho Lee, Daniel Benjamin, Fred Morstatter, Xiang Ren, Kristina Lerman, and Jay Pujara. 2022. Assessing Scientific Research Papers with Knowledge Graphs. In *Proceedings of ACM Conference (Conference'17).* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

In recent years, there has been an explosion in the number of scientific articles published in journals and conferences and posted on pre-print servers. In order for the conclusions of scientific publications to be trusted and accepted by the research community, the underlying methods and techniques must be reproducible [18, 23]. Since new research findings build upon prior results, reproducibility is an essential component of scientific research. Unfortunately, a growing body of research suggests that results in scientific literature are not as reproducible as expected [5, 6, 14, 25]. Some other researchers, from a range of disciplines such as psychology [24], biomedicine [13], economics [11] and social sciences [3], revisited a variety of published scientific papers, manually assessed the credibility of them by conducting direct replication studies, and further confirmed the reproducibility issue. The underlying difficulties of reproducible research coupled with the growing rate of new publications motivates the urgent need for large-scale models to curate information about research methods and assess the reproducibility of scientific results.

Although the replication studies have provided ways to identify credible publications, they also showed the difficulty of assessing publications at scale. This is because, in many research fields such as social and behavioral sciences, the process of reproducing experimental results is resource-intensive, and researchers are deincentivized from running replication studies since novel results advance careers. For example, replicating a social-psychology study requires domain experts to understand the experimental design and different groups of participants for comparison. Since researchers usually have limited resources, it is impractical to evaluate all related work manually. Therefore, it becomes more and more important to have automated systems to perform the assessments and provide insights for fellow researchers.

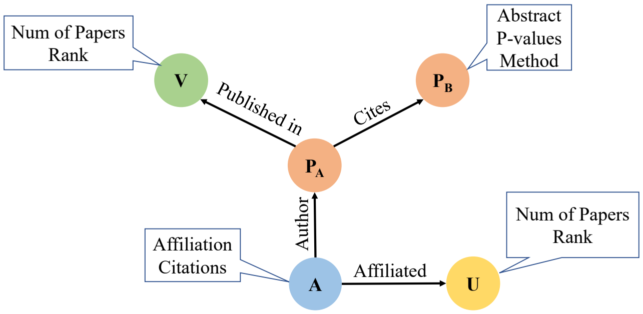

Figure 1: An example KG with additional information associated entities. P_A, P_B are paper entities, A, U and V are author, organization, and venue entities, respectively. Different entities have different features associated with them.

In the spirit of automatically assessing papers, researchers have started applying advanced machine learning techniques. For example, Altmejd et al. [4] train predictive machine learning models to study the effect of different variables in terms of predicting reproducibility. Their experiments identified several basic experimental features, such as the sample and effect sizes of the original papers, that are useful for predicting reproducibility. In addition, Yang et al. [31] collect paper abstracts and train a word embedding model [22] to capture textual information of the papers for making predictions.

Despite the usefulness of these existing approaches, higher-order information is ignored. For example, a finding might be hard to be reproduced if it is purely based on another irreproducible finding. Such kind of information could potentially be captured if relationships between papers are considered. Based on this intuition, in this paper, we propose a novel approach for assessing scientific research using *knowledge graphs (KGs)* which have been successfully used in many applications such as search, question answering and data integration [9, 19, 21]. Our approach incorporates information from two different perspectives: *micro*-level and *macro*-level. Specifically, *micro* features include explicit inter-paper features such as sample sizes and effect sizes, and implicit features that encode paper content with pre-trained language models. *Macro* features include high-level intra-paper relationships between different elements such as author-paper, paper-paper, and paper-venue relationships. We then construct KGs such that entities represent different elements (i.e. papers, authors) and edges represent different relationships between the elements. In addition, each entity may also have additional associated features. We then improve and apply KG embedding methods to learn hidden representations for entities. Finally, a neural network is trained to assess papers using their hidden representations. Figure 1 shows an example KG with task-relevant entities and relations.

In this paper, for the first time, we propose to incorporate features from two perspectives for assessing scientific papers. We construct KGs with both micro- and macro-features to encode rich information for papers. To incorporate different types of features associated to entities, we then adjust the existing KG embedding methods (i.e. LiteralE) for our task to learn hidden representations for papers. We finally experimentally demonstrate the usefulness of our approach on two benchmark datasets.

2 APPROACH

In this section, we introduce our approach in details. We first show two levels of information (i.e. micro and macro) which are leveraged in our approach. We then describe how a knowledge graph (KG) is constructed with additional information associated entities for our task. We finally demonstrate the extended KG embedding method that supports our KGs.

2.1 Micro and Macro Information

We consider micro- and macro-level information in our task. Specifically, micro-level information includes features of entities themselves. For example, models, P-values and sample sizes are features within papers; years and series are features of conferences; counts of citations and counts of papers are features of authors, etc. For assessing papers, these features could be helpful. For example, a large P-value may indicate a low credibility while a high citation count may suggest a high reproducibility.

In addition to micro-level features, we also incorporate macro-level information. We refer to macro-level information as information that capture relationships between base entities, such as the citation relationships between papers, affiliated relationship between authors and organizations, and authorship between papers and authors, etc. We believe such information can potentially be useful due to the intuition that social influence may exist in research studies as well. For example, papers with robust methods are likely to cite other papers with robust methods, papers from a higher-prestige author or institute may be more reproducible, and papers published in top-tier conferences may be more reliable, etc.

2.2 Knowledge Graph Construction

To incorporate both micro and macro information, we propose to construct knowledge graphs such that micro features are used as additional information associated with entities while macro relationships are used to construct the network structure. Formally, a KG can be represented as $G = \{\langle e_i, r_k, e_j \rangle | e_i \in E, e_j \in E, r_k \in R\}$ where E is a set of entities and R is a set of relations. $\langle e_i, r_k, e_j \rangle$ indicates that the relation r_k exists between entities e_i and e_j. In the graph, each node represents an entity and each edge represents a directed relation between two entities. Besides the triples, in our task, each entity e_i has two types of associated information d_i and n_i which represent the encoded description and some other numerical features of e_i, respectively.

Our KG schema includes the following 6 main types of entities:

(1) *Affiliation*: Entities of universities, companies and other organizations that authors affiliated to. Their names are provided as descriptions and there are numeric features rank, paper count and citation count.
(2) *Author*: Entities of authors of the publications. Similarly, their names are descriptions and there are rank, paper count and citation count as numeric features.
(3) *Field of Study*: Entities of research fields that publications belong to. They have rank as numeric features and their names as descriptions.

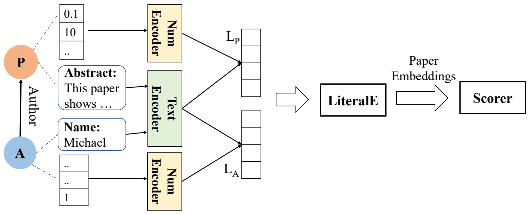

Figure 2: The workflow of our approach. Given a triple $\langle A, Author, P\rangle$, the numeric features of entities are passed to a Number Encoder, and the descriptions are passed to a text encoder. The static number and description vectors are concatenated and passed to LiteralE as literal vectors. Paper vector representations are finally obtained and used as input for predicting reproducibility scores.

(4) *Publication*: Entities of papers. Abstracts (or titles if abstracts are not available) are used as descriptions. There are 24 numeric features including experiment features (e.g. P-values, sample sizes, and number of studies), transparency-related features (e.g. if the data and code open and if there are preregistrations), and network measures (e.g. the paper rank among all papers, the network authority and clustering coefficient of the paper in the citation network.

(5) *Venue*: Both journal and conference entities are included. They have similar numeric features as *Affiliation*.

(6) *Constant*: All other entities are constants. For example, *year* of the publication, *type* and *sub-type* of the venue. No numeric features are considered for these entities.

To connect entities, we consider 10 different relations. For example, an author entity is *affiliated to* an affiliation entity; a publication entity is *published in* a venue entity; a publication *cites* another publication entity; an author entity is *author of* a publication entity. We use the Microsoft Academic Graph (MAG) [1] to collect such relationship information. Given a set of origin publications, we traverse the MAG and keep the publications within two-hop away from the origin publications as well as their author and venue entities.

2.3 Scoring Publications with KG Embeddings

We use the constructed KG to learn representations for publications and score them using their representations (also called embeddings). Most KG embedding methods focus on capturing KG structures and relation patterns. For example, TransX [10, 15, 20] considers relations as translations such that the embedding of source entities are translated into target entities using the relation embeddings. DistMult [30] applies a three-way interactions between entities and relations using matrices. ComplEx [26] proposed to use complex embeddings (real and imaginary) for both entities and relations in order to handle antisymmetric relations. More recently, researchers have been considering involving additional information of entities to assist representation learning. For example, Xie et al. [28], Xu et al. [29] involve structure and textual information to learn representations. LiteralE [17] combines literals and structural embeddings with a learnable function to form the final embeddings.

In this paper, since LiteralE is a general framework which accepts any classic KG embedding methods (i.e. TransE, ComplEx, DistMult, etc), we adopt it as our base model. Figure 2 demonstrates the workflow of our approach. Let $\langle e_s, r, e_o\rangle$ be a triple, d_s and d_o are the descriptions associated to e_s and e_o, n_s and n_o are the lists of numerical features associated to e_s and e_o, respectively. We apply SciBERT [7], a BERT-based [12] model pre-trained on scientific text, to encode the descriptions such that

$$\mathbf{v_d^s} = SciBERT(d_s) \;\; \mathbf{v_d^o} = SciBERT(d_o).$$

$\mathbf{v_d^s}, \mathbf{v_d^o} \in \mathbf{R}^N$ where N is the dimension of the hidden states. The numerical feature lists are converted by an element-wise Exponent Number Converter which collapses numbers into bins [8].

$$\mathbf{v_n^s} = Exp(n_s) \;\; \mathbf{v_n^o} = Exp(n_o).$$

$\mathbf{v_n^s}, \mathbf{v_n^o} \in \mathbf{R}^M$ where M is the number of numerical features. For example, if the initial feature list is $[1.1, 10.5]$, the number converter may take the logarithm of individual numbers and convert the list into an integer list $[0, 2]$. LiteralE learns a function g to convert an original entity embedding into a new embedding with both textual and numeric features involved.

$$\mathbf{v_e^s} = g(\mathbf{v^e}, \mathbf{v_d^s} \| \mathbf{v_n^s}), \mathbf{v_e^o} = g(\mathbf{v^o}, \mathbf{v_d^o} \| \mathbf{v_n^o})$$

where $\mathbf{v^e}$ and $\mathbf{v^o}$ are original entity embeddings. With the literal-enriched vectors, g is learned according to the scoring function $f(\mathbf{v_e^s}, \mathbf{v_r}, \mathbf{v_e^o})$ where f is determined by the base model used in LiteralE. For TransE,

$$f = |\mathbf{v_e^s} + \mathbf{v_r} - \mathbf{v_e^o}|,$$

For DistMult,

$$f = \mathbf{v_e^s}^T M_r \mathbf{v_e^o}$$

where M_r is a diagnoal matrix for r. For ComplEx,

$$f = Re(\mathbf{v_e^s}^T M_r \overline{\mathbf{v_e^o}})$$

where Re means the real part of the vector.

After training the KG embedding model, we take static learned publication representations and apply a Multi-Layer Perceptron (MLP) on them to predict continuous scores. For the MLP, we use Mean Squared Error as training loss.

3 EXPERIMENTS

In this section, we first provide statistics of the constructed KGs of two different datasets. We then compare our method with several baseline methods and analyze the results.

3.1 Datasets

We use two datasets in our experiments. The first dataset is from the Reproducibility Project: Psychology [24]. To determine a successful replication, we use the significance of the meta-analytic combination between the original and replication study (coded as a binary). It contains 70 paper in total from the psychology domain. The second dataset SCORE is from the SCORE project [3]. The dataset contains 2362 papers from several different social and behavioral domains. In this dataset, the scores indicate the credibility of the claims in the papers. For both datasets, we consider continuous scores between 0 and 1 such that a larger value indicates a higher credibility/reproducibility.

[1] https://www.microsoft.com/en-us/research/project/microsoft-academic-graph

Table 1: Statistics of the KGs

	# Nodes	# Edges
RPP	148,983	1,769,883
SCORE	2,287,066	36,144,015

For both datasets, we create the KG by starting from the papers within the dataset and traversing over the paper citation and author academic graphs. The papers within two hops from the root papers are kept. All information (authors, affiliations, venues, etc) directly related to the selected papers are also kept in the KG. Table 1 shows more details of the constructed KGs.

3.2 Experimental Settings

All our experiments are run on a single Nvidia Quadro RTX 8000 GPU. We use Python to implement the models [2]. We use PyKEEN [3] [2] library as the backbone of the KG embedding methods. For both KGs, we set the embedding dimension to 100, batch_size to be 2048, negative sample size to be 16, and learning rate to be 0.0001. We run the KG embedding models 10 epochs for RPP and 3 epochs for SCORE. For our methods, we apply a simple linear layer in LiteralE to fuse description and numeric features. During running the downstream task, since different methods may have different input dimensions, we freeze the embedding models and use a MLP which first projects the input into a vector with 50 dimensions and then predicts a continuous score. For all methods, we run 5-fold cross validations. For each experiment, we set batch_size to 4, learning rate to 1e-5 and run 100 epochs. We set the random seed to be 42. We report the average performance.

3.3 Results

Compared Methods. We evaluate the following models on two benchmark datasets:

- **Random**: Each score is a continuous value randomly sampled between 0 and 1.
- **SciBERT**: Entity descriptions are encoded by SciBERT [7]. The hidden states are used as inputs for scoring papers. They are also a part of our method.
- **Numeric**: Only numeric features are used as inputs.
- **Yang**: The approach proposed in [31] that leverages word embeddings trained on paper abstracts collected from MAG.
- **TransE** [10]: A classic translation-based KG embedding method. Only graph structures are leveraged in the method.
- **DistMult** [30]: A KG embedding method based on bilinear interactions between entity and relation representations.
- **ComplEx** [26]: A KG embedding method that represents an entity from both the real and imaginary perspectives.
- **Ours (TransE)**: Our LiteralE-based [17] method that involves both description features, numeric features and KG structures. TransE is used as the base model in LiteralE.
- **Ours (DistMult)**: DistMult is the base model in LiteralE.
- **Ours (ComplEx)**: ComplEx is the base model in LiteralE.

Table 2: Main results. Each number represents a RMSE/KT score. The best performing scores are highlighted and the second best scores are underlined.

	RPP		SCORE	
	RMSE↓	KT↑	RMSE↓	KT↑
Random	0.6080	-0.0642	0.3233	0.0040
SciBERT	0.4765	0.0694	0.1344	0.1675
Numeric	0.4731	0.0697	0.1379	0.1032
Yang	0.4404	0.1191	0.1326	**0.1926**
TransE	0.4931	-0.0856	0.1619	-0.0210
DistMult	0.5010	-0.0508	0.1674	-0.0127
ComplEx	0.5131	0.0573	0.1526	-0.0283
Ours (TransE)	**0.4041**	**0.2454**	0.1323	0.1809
Ours (DistMult)	0.4215	0.2437	**0.1314**	0.1894
Ours (ComplEx)	0.4420	0.1846	0.1316	0.1852

Table 3: Ablation study on encoders. The best performing scores under individual categories are highlighted.

	RPP		SCORE	
	RMSE↓	KT↑	RMSE↓	KT↑
Full (TransE)	0.4041	0.2454	**0.1323**	**0.1809**
- Remove Num Encoder	0.4665	0.0536	0.1396	0.0651
- Use BERT	**0.4015**	**0.3299**	0.1330	0.1664
- Use LongFormer	0.4528	0.1320	0.1341	0.1588
Full (DistMult)	**0.4215**	**0.2437**	**0.1314**	0.1894
- Remove Num Encoder	0.4661	-0.0342	0.1387	0.0674
- Use BERT	0.4258	0.1912	0.1316	**0.1907**
- Use LongFormer	0.4519	0.1553	0.1329	0.1614
Full (ComplEx)	**0.4420**	**0.1846**	0.1316	0.1852
- Remove Num Encoder	0.4691	-0.1669	0.1351	0.1427
- Use BERT	**0.4420**	0.1054	**0.1309**	**0.1957**
- Use LongFormer	0.4606	0.0788	0.1340	0.1557

Metrics. We consider continuous scores and leverage two metrics: Root Mean Squared Error (RMSE) and Kendall's Tau (KT) [16]. Given a list of ground-truth scores and predicted scores, KT measures the correlation between the two lists and RMSE shows the difference between them. KT scores are within the range $[-1, 1]$ such that the larger the better. RMSE scores are non-negative values such that the smaller the better.

Main Results. We report the main results in Table 2. We find: 1) Both micro- and macro- features perform better than random guess. 2) Although SciBERT and Numeric perform better than pure KG embedding methods, literal-involved KG embedding methods achieve better performance by jointly learning useful information from both micro and macro perspectives and also generally outperform the previous published method. 3) Comparing different KG embedding methods, DistMult performs better than TransE and ComplEx by a small margin. 4) The RMSEs of SCORE are generally

[2] https://github.com/kianasun/kg4rr
[3] https://pykeen.readthedocs.io

smaller than those of RPP because ground-truth scores in RPP are all binary while the scores in SCORE are all continuous values.

Ablation Study. We show ablation studies in Table 3. We first show the effectiveness of the Exponent number encoder. We remove the number encoder and use the original numeric features to learn the embeddings and compare their results. In all cases, RMSE increases and KT decreases after removing the encoder. This shows that dividing numbers into bins could be easier for models to learn representations. In the second experiment, we replace the SciBERT by another two language models BERT and LongFormer. After using BERT, in 3 out of 5 cases (for ComplEx, the two variants achieve the same performance), RMSE increases, and in 3 out of 6 cases, KT decreases. This experiment shows that SciBERT achieves slightly better performance than the vanilla BERT. After using LongFormer, the performance decreases by a small margin in all cases. Comparing different KG embedding methods, ComplEx performs the best on SCORE and TransE performs the best on RPP.

4 CONCLUSIONS AND FUTURE WORK

In this paper, we proposed to apply knowledge graphs to assess scientific papers at scale. Our proposed approach leverages explicit features extracted from papers, hidden features encoded by pre-trained language models, and higher-order relationships between papers, authors and venues. We also applied a literal-involved knowledge graph embedding method to learn representations for paper entities and use the learned hidden vectors to provide scores for papers. We hope the scores generated by our system could provide more insights for researchers and help resources in replication studies to be used for evaluating most important papers.

As an initial attempt, we applied a few classic KG embedding methods and simply concatenated two different types of information during training. In the future, more advanced KG embedding methods [1, 27] can be experimented on and more information fusion strategies can potentially be integrated into the model.

5 ACKNOWLEDGEMENTS

This work was funded by the Defense Advanced Research Projects Agency with award W911NF-19-20271. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the Defense Advanced Research Projects Agency. The authors acknowledge the Center for Advanced Research Computing (CARC) at the University of Southern California for providing computing resources that have contributed to the research results reported within this publication. We also thank anonymous reviewers for their valuable suggestions.

REFERENCES

[1] Ralph Abboud, İsmail İlkan Ceylan, Thomas Lukasiewicz, and Tommaso Salvatori. 2020. BoxE: A Box Embedding Model for Knowledge Base Completion. In *Proceedings of the Thirty-Fourth Annual Conference on Advances in Neural Information Processing Systems.*

[2] Mehdi Ali, Max Berrendorf, Charles Tapley Hoyt, Laurent Vermue, Sahand Sharifzadeh, Volker Tresp, and Jens Lehmann. 2021. PyKEEN 1.0: A Python Library for Training and Evaluating Knowledge Graph Embeddings. *Journal of Machine Learning Research* 22, 82 (2021), 1–6.

[3] Nazanin Alipourfard, Beatrix Arendt, Daniel M Benjamin, Noam Benkler, Michael M Bishop, Mark Burstein, Martin Bush, James Caverlee, Yiling Chen, Chae Clark, and et al. 2021. Systematizing Confidence in Open Research and Evidence (SCORE). https://doi.org/10.31235/osf.io/46mnb

[4] Adam Altmejd, Anna Dreber, Eskil Forsell, Juergen Huber, Taisuke Imai, Magnus Johannesson, Michael Kirchler, Gideon Nave, and Colin Camerer. 2019. Predicting the replicability of social science lab experiments. 14, 12 (2019). https://doi.org/10.1371/journal.pone.0225826

[5] Monya Baker. 2016. IS THERE A REPRODUCIBILITY CRISIS? *Nature* 533 (05 2016), 452–454.

[6] C. Glenn Begley and Lee M. Ellis. 2012. Raise standards for preclinical cancer research | Nature. 83, 7391 (2012), 531–533. https://doi.org/doi.org/10.1038/483531a

[7] Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. SciBERT: Pretrained Language Model for Scientific Text. In *EMNLP.*

[8] Taylor Berg-Kirkpatrick and Daniel Spokoyny. 2020. An Empirical Investigation of Contextualized Number Prediction. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing.* 4754–4764. https://doi.org/10.18653/v1/2020.emnlp-main.385

[9] Kurt Bollacker, Colin Evans, Praveen Paritosh, Tim Sturge, and Jamie Taylor. 2008. Freebase: A Collaboratively Created Graph Database for Structuring Human Knowledge. In *Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data.* 1247–1250. https://doi.org/10.1145/1376616.1376746

[10] Antoine Bordes, Nicolas Usunier, Alberto Garcia-Duran, Jason Weston, and Oksana Yakhnenko. 2013. Translating Embeddings for Modeling Multi-relational Data. In *Advances in Neural Information Processing Systems*, Vol. 26.

[11] Colin F. Camerer, Anna Dreber, Eskil Forsell, Teck-Hua Ho, Jürgen Huber, Magnus Johannesson, Michael Kirchler, Johan Almenberg, Adam Altmejd, Taizan Chan, Emma Heikensten, Felix Holzmeister, Taisuke Imai, Siri Isaksson, Gideon Nave, Thomas Pfeiffer, Michael Razen, and Hang Wu. 2016. Evaluating replicability of laboratory experiments in economics. *Science* 351, 6280 (2016), 1433–1436. https://doi.org/10.1126/science.aaf0918

[12] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies.* 4171–4186. https://doi.org/10.18653/v1/N19-1423

[13] Timothy M Errington, Courtney K Soderberg Maya Mathur, Alexandria Denis, Nicole Perfito, Elizabeth Iorns, and Brian A Nosek. 2021. Investigating the replicability of preclinical cancer biology. (2021). https://doi.org/10.7554/eLife.71601

[14] John P. A. Ioannidis. 2005. Why Most Published Research Findings Are False. *PLOS Medicine* 2 (08 2005), null. https://doi.org/10.1371/journal.pmed.0020124

[15] Guoliang Ji, Shizhu He, Liheng Xu, Kang Liu, and Jun Zhao. 2015. Knowledge Graph Embedding via Dynamic Mapping Matrix. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing.* 687–696. https://doi.org/10.3115/v1/P15-1067

[16] M. G. Kendall. 1938. A New Measure of Rank Correlation. *Biometrika* 30, 1/2 (1938), 81–93.

[17] Agustinus Kristiadi, Mohammad Asif Khan, Denis Lukovnikov, Jens Lehmann, and Asja Fischer. [n.d.]. Incorporating Literals into Knowledge Graph Embeddings. In *The Semantic Web.* 347–363. https://doi.org/10.1007/978-3-030-30793-6_20

[18] Imre Lakatos. 1970. Criticism and the Growth of Knowledge (Proceedings of the International Colloquium in the Philosophy of Science, London 1965, Volume 4). (1970).

[19] Jens Lehmann, Robert Isele, Max Jakob, Anja Jentzsch, Dimitris Kontokostas, Pablo Mendes, Sebastian Hellmann, Mohamed Morsey, Patrick Van Kleef, Sören Auer, and Christian Bizer. 2014. DBpedia - A Large-scale, Multilingual Knowledge Base Extracted from Wikipedia. *Semantic Web Journal* 6 (01 2014). https://doi.org/10.3233/SW-140134

[20] Yankai Lin, Zhiyuan Liu, Maosong Sun, Yang Liu, and Xuan Zhu. 2015. Learning Entity and Relation Embeddings for Knowledge Graph Completion. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence.* 2181–2187.

[21] Farzaneh Mahdisoltani, Joanna Asia Biega, and Fabian M. Suchanek. 2015. YAGO3: A Knowledge Base from Multilingual Wikipedias. In *CIDR.*

[22] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient Estimation of Word Representations in Vector Space. (2013). arXiv:1301.3781

[23] Brian A. Nosek and Timothy M. Errington. 2020. What is replication? *PLOS Biology* 18 (03 2020), 1–8. https://doi.org/10.1371/journal.pbio.3000691

[24] Open Science Collaboration. 2015. Estimating the reproducibility of psychological science | Science. 349, 6251 (2015). https://doi.org/10.1126/science.aac4716

[25] Florian Prinz, Thomas Schlange, and Khusru Asadullah. 2011. Believe it or not: how much can we rely on published data on potential drug targets? 10, 9 (2011), 712–712. https://doi.org/10.1038/nrd3439-c1

[26] Théo Trouillon, Johannes Welbl, Sebastian Riedel, Eric Gaussier, and Guillaume Bouchard. 2016. Complex Embeddings for Simple Link Prediction. In *Proceedings of The 33rd International Conference on Machine Learning (Proceedings of Machine Learning Research)*, Vol. 48. 2071–2080.

[27] Shikhar Vashishth, Soumya Sanyal, Vikram Nitin, and Partha Talukdar. 2020. Composition-based Multi-Relational Graph Convolutional Networks. In *International Conference on Learning Representations*.

[28] Ruobing Xie, Zhiyuan Liu, Jia Jia, Huanbo Luan, and Maosong Sun. 2016. Representation Learning of Knowledge Graphs with Entity Descriptions. *Proceedings of the AAAI Conference on Artificial Intelligence* 30, 1 (2016).

[29] Jiacheng Xu, Xipeng Qiu, Kan Chen, and Xuanjing Huang. 2017. Knowledge Graph Representation with Jointly Structural and Textual Encoding. In *IJCAI*.

[30] Bishan Yang, Wen-tau Yih, Xiaodong He, Jianfeng Gao, and Li Deng. 2015. Embedding Entities and Relations for Learning and Inference in Knowledge Bases. In *3rd International Conference on Learning Representations, ICLR 2015*.

[31] Yang Yang, Wu Youyou, and Brian Uzzi. 2020. Estimating the deep replicability of scientific findings using human and artificial intelligence. 117, 20 (2020), 10762–10768. https://doi.org/10.1073/pnas.1909046117