Graph-Enriched Biomedical Language Models: A Research Proposal

Andrey Sakhovskiy[1,2,3], Alexander Panchenko[3,4], and Elena Tutubalina[1,2]
[1]Sber AI, [2]Kazan Federal University, [3]Skolkovo Institute of Science and Technology,
[4]Artificial Intelligence Research Institute
{andrey.sakhovskiy, panchenko.alexander, tutubalinaev}@gmail.com

Abstract

Recent advancements in biomedical NLP have been driven by domain-specific pre-trained language models (LMs), yet the challenge of effectively storing extensive biomedical factual knowledge remains. Despite the superior performance of fine-tuned LMs in downstream NLP tasks, these models exhibit limitations in ontology memorization, reasoning abilities, and capturing complex specialized domain terminology. To address these issues, we present four research questions that explore the integration of LMs with large knowledge graphs (KGs) like the Unified Medical Language System (UMLS). Our proposal introduces novel alignment methods to bridge LMs with the UMLS KG, with the aim of leveraging structured background knowledge to enhance the reasoning and generalization capabilities of biomedical LMs. The research proposal discusses multilingual specifics of KBs and evaluation metrics across various datasets.

1 Introduction

Recent years have witnessed significant progress in various biomedical Natural Language Processing (NLP) caused by domain-specific pre-trained Language Models (LMs) (Lee et al., 2020; Peng et al., 2019; Alsentzer et al., 2019; Beltagy et al., 2019; Michalopoulos et al., 2021; Gu et al., 2022; Yasunaga et al., 2022b). Although, these models demonstrate superior performance on Biomedical Language Understanding and Reasoning Benchmark (BLURB) (Gu et al., 2022) and BigBio benchmark (Fries et al., 2022), their ability to store extensive biomedical factual knowledge remains an open question. In the general domain, Large LMs (LLMs) were shown to have limited ontology memorization and reasoning abilities (Wu et al., 2023). Existing research on biomedical knowledge probing task indicate that the biomedical LMs struggle to capture complex specialized domain terminology (Meng et al., 2022), are highly biased towards certain prompts, and are unaware of synonyms (Sung et al., 2021). Making LM well-informed about in-domain facts could assist various NLP applications including drug discovery (Wu et al., 2018; Khrabrov et al., 2022; Zitnik et al., 2018), clinical decision making (Sutton et al., 2020; Peiffer-Smadja et al., 2020), and biomedical research (Lee et al., 2016; Fiorini et al., 2018; Soni and Roberts, 2021).

In the biomedical domain, vast multilingual Knowledge Bases (KBs) such as the Unified Medical Language System (UMLS) (Bodenreider, 2004) are available, making the infusion of factual knowledge into LMs possible. Over 166 lexicons/thesauri with over 4M concepts and 15M concept names from 27 languages are present in the UMLS. However, as seen from Tab. 1, severe language imbalance is a great challenge for processing texts in low-resource languages.

In KBs, factual information is usually stored in the form of knowledge triples (h, r, t). Each triple reflects the fact that concept h is in relation to type r with concept t. The combination of concept set V and relation triples $E \in \{V \times R \times V\}$ can be seen as a knowledge graph (KG) $G = G(V, E, R)$ where R is a set of possible relation types. Although plenty of research focused on developing effective knowledge-augmented general-purpose pre-training methods for LMs, this topic remains challenging. One approach is to apply an LM on textual sequences augmented by KB triples (Wang et al., 2019a; Mannion et al., 2023; Xu et al., 2023; Liu et al., 2020). These approaches share two major limitations (Ke et al., 2021). First, the fully connected nature of the attention mechanism present in modern LMs contradicts the sparse structure of the existing KB graphs. Second, the linearization of a KB graph prevents a direct alignment between the textual and the KB modalities. Wang et al. (2021) obtained representations for Wikipedia en-

Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics: Student Research Workshop, pages 82–92

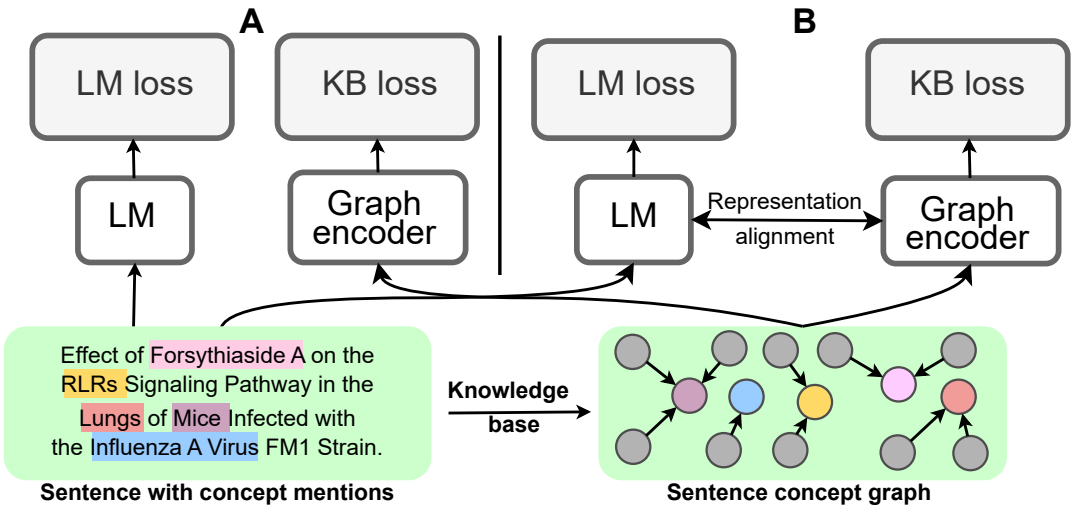

Figure 1: Visualization of different approaches towards knowledge-enhanced LM training. **A**: During KB-enhanced LM pre-training or fine-tuning, a text encoder and a graph encoder independently minimize a textual loss (i.e., masked language modeling) and a graph-related task (i.e., link prediction) with an implicit interaction between the two encoders. An implicit interaction may be in the form of LM embeddings being initial node representations for the graph model. **B**: A less common approach is to add an explicit alignment loss to stimulate an information exchange between two modalities. Named entities can serve as anchor points for this kind of intermodal interaction.

tities by encoding short textual entity and relation descriptions with an LM, which is not feasible in the biomedical domain since most biomedical concepts lack a textual description.

As LM pre-training from scratch requires extensive computational resources, a cheaper alternative is a task-specific KB-aware fine-tuning. Recently, a series of studies focused on the utilization of the UMLS concept names and inter-concept relations for improved Biomedical Concept Normalization (BCN) (Liu et al., 2021a,b; Yuan et al., 2022b; Sakhovskiy et al., 2023). While GEBERT proposed by Sakhovskiy et al. (2023) explicitly learns the identity between synonymous concept names and concept node representations, the model is extremely tied to BCN and leaves no room for its generalization to other biomedical tasks. Recently proposed Question Answering (QA) (Yasunaga et al., 2022a, 2021a; Zhang et al., 2022b) systems adopt Message Passing (MP) (Gilmer et al., 2017) graph neural networks to perform well-grounded reasoning over KB which results in an improved quality in both general and biomedical domains. These models rely on implicit interaction between an LM and a graph encoder and do not explicitly learn an alignment between two modalities, thus limiting LM's ability to memorize KB facts.

2 Related work

An extensive comparison of various biomedical knowledge representation learning approaches was conducted by Chang et al. (2020). They compared semantic matching methods, such as TransE (Bordes et al., 2013), DistMult (Yang et al., 2015), ComplEx (Trouillon et al., 2016), SimplE (Kazemi and Poole, 2018), and RotatE (Sun et al., 2019), for link prediction quality on SNOMED-CT dataset. Although these methods outperform simpler Snomed2Vec (Agarwal et al., 2019) and Cui2Vec (Beam et al., 2020) baselines, they fall short of LM-based approaches (Wang et al., 2019a).

Several attempts to integrate a pre-trained biomedical LM with an external KB have increased performance in various downstream tasks. Sakhovskiy et al. (2021); Sakhovskiy and Tutubalina (2022) employed DrugBank (Wishart et al., 2008, 2017), a drug-oriented chemical database, to combine LM embeddings with drug chemical features in a classification layer to detect texts that mention an adverse drug reaction. SapBERT (Liu et al., 2021a,b) achieved state-of-the-art Medical Concept Normalization (MCN) performance by applying a contrastive objective to learn from synonymous biomedical concept names from the Unified Medical Language System

(UMLS) ontology. CODER (Yuan et al., 2022b) and GEBERT (Sakhovskiy et al., 2023) extended the idea by introducing additional graph-based contrastive objectives to capture inter-concept relations from the UMLS graph. CODER (Yuan et al., 2022b) and multilingual SapBERT (Liu et al., 2021b) achieve a normalization improvement in both monolingual English and multilingual setups.

In both general and biomedical domains, numerous state-of-the-art QA solutions retrieve a relevant subgraph from a KB (Lin et al., 2019; Feng et al., 2020; Yasunaga et al., 2021a; Zhang et al., 2022b,a; Yasunaga et al., 2022a) to perform a knowledge-aware reasoning. Yasunaga et al. (2022a) proposed a language-knowledge DRAGON model that benefits from joint language modeling and graph completion objectives and bidirectional interaction between text and graph encoders in both general and biomedical domains.

Thus, the existing knowledge-enhanced text processing models possess at least one of the following key limitations. First, they are too tied to a specific downstream task, such as MCN or QA. Second, they provide no explicit alignment between a biomedical concept and its mention in a text but instead rely on implicit interaction between textual and graph encoders. Third, except for multilingual BCN methods, they mostly focus on English, which has the most extensive KBs, ignoring a low-resource case.

Language	# concept names	percentage
English	11,280,428	70.78%
Spanish	1,589,581	9.97%
French	431,527	2.71%
Portuguese	423,826	2.66%
Japanese	332,099	2.08%
Dutch	293,817	1.84%
Russian	293,031	1.84%
Italian	251,912	1.58%
German	235,736	1.48%
Czech	198,115	1.24%
Korean	147,217	0.92%
Hungarian	109,271	0.69%
Chinese	81,916	0.51%
Norwegian	63,797	0.4%
Polish	51,778	0.32%
Turkish	51,597	0.32%
Estonian	31,183	0.2%
Swedish	30,439	0.19%
Finnish	25,489	0.16%
Croatian	10,035	0.06%
Greek	2,286	0.01%
Latvian	1405	0.01%
Danish	723	0.1%
Basque	695	<0.1%
Hebrew	485	<0.1%

Table 1: UMLS statistics on the number of concept names.

3 Research plan

3.1 Research questions

Although a wide range of knowledge-aware Language Modeling techniques have been proposed, several fundamental research questions remain unanswered. In this proposal, we formulate some important questions as well as possible trajectories for answering them. First of all, we see three major knowledge fusion strategies:

1. Knowledge-enhanced LM pre-training from scratch;

2. KB-augmented task-specific fine-tuning;

3. Alignment between pre-trained LM and informative KB representations.

RQ1. What is an optimal knowledge fusion strategy?

While existing knowledge-enhanced general-domain and biomedical LMs benefit from pre-training with external knowledge, they usually share at least one of the following critical limitations. First, they imply a modification of an LM architecture (Peters et al., 2019; Zhang et al., 2022b; Yasunaga et al., 2022a). Second, they require additional pre-training of all model parameters on textual inputs augmented with external knowledge (Wang et al., 2021; Lauscher et al., 2020; El Boukkouri et al., 2022; Yuan et al., 2022a; Mannion et al., 2023). Both limitations lead to a resource-intensive pre-training of all the LM parameters from scratch which might not be feasible. Recently proposed FROMAGe (Koh et al., 2023b) and GILL (Koh et al., 2023a) in text-and-image domain propose to align image representations with their textual captions via contrastive InfoNCE (Oord et al., 2018) objective in a significantly more light-weighted scenario of frozen textual encoder. With far less trainable parameters, these alignment meth-

ods manage to even outperform fully trainable bi-modal Transformer (Vaswani et al., 2017) models. Inspired by the success of alignment-based strategy in text and image tasks, we strive to explore its applicability and effectiveness in the biomedical domain in comparison with the remaining two strategies.

RQ2. How to align KB and LM in the biomedical domain?

To the best of our knowledge, no LM and biomedical KB representation alignment method is proposed so far. A direct adaptation of GILL and FROMAGe to biomedical texts and KBs is hindered by two critical issues. First, both models rely on Transformer encoder-decoder architecture and adopt text generation tasks, while the majority of the existing state-of-the-art biomedical LMs are encoder-only BERT models (Alsentzer et al., 2019; Peng et al., 2019; Beltagy et al., 2019; Lee et al., 2020; Gu et al., 2022; Liu et al., 2021a; Mannion et al., 2023). Second, while image-to-text and text-to-image tasks are inherently bi-modal, it is not the case for most biomedical NLP tasks (i.e., only textual sequence is provided during fine-tuning and evaluation).

3.1.1 RQ3. How to enrich an LM with biomedical knowledge?

Current biomedical knowledge probing benchmarks (Sung et al., 2021; Meng et al., 2022) indicate that the existing domain-specific LMs lack factual knowledge. This might be caused by either of two reasons: (i) imperfection of prompting approaches or (ii) an actual absence or incompleteness of knowledge in LMs. We believe, the integration of in-domain knowledge from biomedical KBs (e.g., interaction between biomedical concepts from the UMLS) remains an open challenge and requires a thorough exploration.

RQ4. How to exploit rich English KBs for low-resource languages?

Most existing research in biomedical NLP employ extensive English data leaving low-resource languages out-of-scope. While the alignment of multilingual UMLS concept names was shown to significantly improve the BCN quality in uni-modal setting (Liu et al., 2021b; Yuan et al., 2022b), they still struggle to deal with severe language imbalance of the UMLS concept names (see Table 1). Alternatively, the UMLS KB can be approached from a bi-modal text and graph perspective with graph modality capturing language-independent concept node's features.

3.2 Proposed methodology

3.2.1 Representation alignment

Currently, the alignment of textual and KB representations remains under-expored topic. To answer **RQ1** and **RQ2** we plan to develop novel alignment methods. To align textual representations with KB knowledge, we plan to use biomedical concept representations obtained from their contextualized mention embeddings in texts. We foresee two possible alignment approaches: (i) implicit alignment via an auxiliary KB-guided training objective and (ii) via an explicit alignment of textual and graph representations.

Implicit alignment One of the ways to enable information exchange between two or more modalities is to introduce a multi-modal objective. Prior work on general domain QA (Yasunaga et al., 2022a; Ke et al., 2021) introduced multi-task text and graph restoration objectives to learn from aligned textual sequences and KB subgraphs of entities mention in a text. However, this approaches rely on implicit interaction between text and graph modalities and do not explicitly inform the model that the subgraph is induced by text and is in fact its alternative representation obtained from another modality. In our work, we plan to adopt and extend the idea of graph restoration objective the idea and consider two its following cases:

- **Single modality graph restoration**: Following Yasunaga et al. (2022a) and Ke et al. (2021) we will treat text and graph restoration tasks as separate uni-modal tasks with a single graph encoder to encode both head and tail concepts of a triple;

- **Mixed-modality graph restoration**: As LM- and graph-based representations of a concept are complementary, we propose to initialize a head concept with an embedding of the first modality and a tail concept with an embedding of the second one.

While the first case is conventional, the mixed-modality problem statement is, to the best of our knowledge, under-explored. For both cases, we will employ TransE or ComplEx which model a tail concept as a relation-based transformation of a head concept.

Explicit alignment Another way to combine multiple modalities is to explicitly inform the model that text and graph embeddings are two complementary representations of a single concept.

Early attempts of alignments of graphs with linguistic models were presented by Biemann et al. (2018): sparse representations of graphs were linked with sparse distributional representations of word senses. Nikishina et al. (2022) attempts to align standard text BERT model with graph-based BERT by learning projections of their internal representations. Similarly, projections between static graph and text embeddings can be used for computing similarity search in graphs given text e.g. for question answering (Huang et al., 2019).

In prior work (Sakhovskiy et al., 2023), a contrastive objective was applied to learn from bi-modal positive pairs consisting of a concept name and a concept node. GILL and FROMAGe benefited from aligning in-context LM tokens and images via a contrastive objective and a small alignment model. In our research, we plan to combine these two approaches and perform an in-context alignment of contextualized concept mentions and their graph representations obtained from the UMLS via a graph encoder. We expect to introduce either the Multi-Similarity (Wang et al., 2019b) or the InfoNCE (Oord et al., 2018) loss function, to directly minimize the distance between textual and graph representations of the same biomedical concept.

3.2.2 Knowledge probing

Two possible ways to improve LM capabilities as KB and answer **RQ3** are (i) an improvement of prompting strategies and (ii) a modification of LM and its training pipeline. Although Meng et al. (2022) and Sung et al. (2021) have observed a probing quality improvement after a proper prompt tuning, the task is still far from being solved with about only 10% in terms of accuracy. We will stick to the second option and attempt to improve the knowledge awareness of biomedical LMs through alignment with KB modality: both implicit and explicit. As current biomedical knowledge probing benchmarks require filling masked concepts in a prompt inferred from a knowledge triple, we will investigate the knowledge infusion as a bi-modal problem and focus on the following knowledge probing problem statements:

Uni-modal textual approach involves filling masked concept slot using solely an LM;

Bi-modal text and KB approach reformulates triplet completion baseline as a bi-modal text-to-graph task: given a textual prompt, the goal is to predict the best matching KG node.

While text-only approaches commonly struggle with multi-word concept names, we aim at exploring whether the reformulation of the task will help overcome the issue. Moreover, the second approach enables the incorporation of aforementioned modality alignment strategies: both explicit and implicit.

3.2.3 Cross-lingual alignment

We expect to address the **RQ4** with the cross-lingual cross-modal representation alignment. While a fixed concept name is monolingual, a concept itself is multilingual from the language perspective and is independent of language from the graph perspective. While cross-lingual concept name alignment improved BCN quality (Liu et al., 2021b; Yuan et al., 2022b), our goal is to investigate whether cross-modal alignment could further boost the performance. Unfortunately, application on other biomedical tasks is hindered by the lack of non-English data but the experiments on BCN could serve as a good starting point.

3.3 Experimental setting

Training data As training data for various alignment methods, we will utilize PubMed abstracts. To recognize and align textual concept mentions with UMLS concepts, we will adopt BERN2 (Sung et al., 2022), a recently proposed biomedical entity recognition and normalization tool.

Text and graph encoders To obtain language representations, we will adopt PubMedBERT (Gu et al., 2022), a state-of-the-art biomedical LM pretrained on PubMed abstracts. To produce graph representations, we will adopt the Message Passing framework (Gilmer et al., 2017) and obtain concept node embeddings with either GraphSAGE (Hamilton et al., 2017) or GAT (Veličković et al., 2018) encoder. Each node will be initialized with a PubMedBERT embedding of its concept name at random.

Computational efficiency Since for alignment strategy, we assume both textual and graph encoder are already well-trained, we strive to explore if

we can reduce the computational burden of the alignment procedure. For each encoder, we will consider three cases: (i) fully frozen encoder with a small external alignment model, (ii) partially frozen encoder, (iii) fully trainable encoder.

Concept masking To enforce a KB-aligned LM learns from full context rather than concept mentions only, we will mask concept mentions with a fixed probability. Similarly, to stimulate graph encoder pass more informative messages from concept neighboring concepts in a KG, we will mask a concept name of an anchor. Masking is expected to improve model's compatibility with knowledge probing benchmarks.

3.4 Evaluation

Fries et al. (2022) released BigBio, a large data-centric benchmark that includes 126 biomedical NLP datasets, covering 13 tasks, including QA and BCN in more than 10 languages. To answer the **RQ1** and **RQ2**, we will primarily focus on QA and BCN since these tasks already have knowledge-enhanced task-specific solutions to compare with. To explore the **RQ4**, we will compare against current state-of-the-art cross-lingual models for BCN (Liu et al., 2021b; Yuan et al., 2022b; Sakhovskiy et al., 2023) and additionally adopt two cross-lingual BCN benchmarks for zero-shot ranking-based avaluation: (i) the one (Alekseev et al., 2022) based on Mantra corpus (Kors et al., 2015) and (ii) XL-BEL (Liu et al., 2021b). We will adopt KG-enhanced state-of-the-art QA models: QA-GNN (Yasunaga et al., 2021b), GreaseLM (Zhang et al., 2022b), JointGT (Ke et al., 2021), and DRAGON (Yasunaga et al., 2022a) as knowledge-enhanced QA baselines. For both BCN and QA as well as other tasks, we will adopt strong domain-specific biomedical LMs, e.g., BioBERT (Lee et al., 2020).

For biomedical knowledge probing task and **RQ3**, we will adopt the aforementioned MedLAMA and BioLAMA benchmarks. We will evaluate against the existing biomedical LMs, such as the BioBERT (Lee et al., 2020), Bio-LM (Lewis et al., 2020), and PubMedBERT (Gu et al., 2022).

4 Conclusion

In this paper, we identify critical limitations of the existing domain-specific pre-trained biomedical LMs and current state-of-the-art domain-specific solutions for solving downstream NLP tasks. We raise four important research questions and present a plan for exploring them. Modern LMs are unable to reveal the potential of factual knowledge fully and lack an explicit text-KB alignment procedure in current pre-training pipelines. While the usage of KB has already advanced the quality of biomedical concept normalization and question answering, a method for the fusion of domain knowledge into a general-purpose biomedical LM awaits to be explored. To overcome the existing LM limitations, we propose ideas for explicit alignment of KB concepts and their representatives in texts. The completion of our research plan is expected to deepen the understanding of text-KB interaction and give a better understanding of an optimal strategy for KB utilization in biomedical NLP.

Acknowledgements The work has been supported by the Russian Science Foundation grant # 23-11-00358.

5 Ethics, limitations, and risks

Large domain-specific graphs. We plan to employ a large biomedical knowledge graph, the Unified Medical Language System (UMLS), which contains over 4 million concepts and 15 million concept names. It is important to note that using knowledge graphs for different domains with a smaller number of nodes and edges may affect the performance. The knowledge graph's size and complexity can significantly impact the model's ability to learn and make accurate predictions.

Biases. Consequently, it is important to acknowledge that trained models can inherit biases and toxic behaviors present in the language models and knowledge graphs used for their initialization. Language models, for instance, have been demonstrated to incorporate biases about race, gender, and other demographic attributes. Biomedical research and clinical trials may not adequately represent certain populations. Likewise, a knowledge graph may incorporate stereotypes instead of providing unbiased, commonsense knowledge.

Diversity of biomedical concepts. It is important to highlight that the datasets and knowledge graphs primarily focus on well-documented medical concepts found in the literature. This limits the exposure of models to infrequent or uncommon occurrences. Consequently, adapting trained models to handle rare biomedical events may require additional effort and attention.

References

Khushbu Agarwal, Tome Eftimov, Raghavendra Addanki, Sutanay Choudhury, Suzanne Tamang, and Robert Rallo. 2019. Snomed2vec: Random walk and poincaré embeddings of a clinical knowledge base for healthcare analytics. *CoRR*, abs/1907.08650.

Anton Alekseev, Zulfat Miftahutdinov, Elena Tutubalina, Artem Shelmanov, Vladimir Ivanov, Vladimir Kokh, Alexander Nesterov, Manvel Avetisian, Andrei Chertok, and Sergey Nikolenko. 2022. Medical crossing: a cross-lingual evaluation of clinical entity linking. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4212–4220, Marseille, France. European Language Resources Association.

Emily Alsentzer, John Murphy, William Boag, Wei-Hung Weng, Di Jindi, Tristan Naumann, and Matthew McDermott. 2019. Publicly available clinical BERT embeddings. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 72–78, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Andrew L. Beam, Benjamin Kompa, Allen Schmaltz, Inbar Fried, Griffin M. Weber, Nathan P. Palmer, Xu Shi, Tianxi Cai, and Isaac S. Kohane. 2020. Clinical concept embeddings learned from massive sources of multimodal medical data. In *Pacific Symposium on Biocomputing 2020, Fairmont Orchid, Hawaii, USA, January 3-7, 2020*, pages 295–306.

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. SciBERT: A pretrained language model for scientific text. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3615–3620, Hong Kong, China. Association for Computational Linguistics.

Chris Biemann, Stefano Faralli, Alexander Panchenko, and Simone Paolo Ponzetto. 2018. A framework for enriching lexical semantic resources with distributional semantics. *Natural Language Engineering*, 24(2):265–312.

Olivier Bodenreider. 2004. The unified medical language system (UMLS): integrating biomedical terminology. *Nucleic Acids Res.*, 32(Database-Issue):267–270.

Antoine Bordes, Nicolas Usunier, Alberto García-Durán, Jason Weston, and Oksana Yakhnenko. 2013. Translating embeddings for modeling multi-relational data. In *Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States*, pages 2787–2795.

David Chang, Ivana Balažević, Carl Allen, Daniel Chawla, Cynthia Brandt, and Andrew Taylor. 2020. Benchmark and best practices for biomedical knowledge graph embeddings. In *Proceedings of the 19th SIGBioMed Workshop on Biomedical Language Processing*, pages 167–176, Online. Association for Computational Linguistics.

Hicham El Boukkouri, Olivier Ferret, Thomas Lavergne, and Pierre Zweigenbaum. 2022. Specializing static and contextual embeddings in the medical domain using knowledge graphs: Let's keep it simple. In *Proceedings of the 13th International Workshop on Health Text Mining and Information Analysis (LOUHI)*, pages 69–80, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Yanlin Feng, Xinyue Chen, Bill Yuchen Lin, Peifeng Wang, Jun Yan, and Xiang Ren. 2020. Scalable multi-hop relational reasoning for knowledge-aware question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1295–1309, Online. Association for Computational Linguistics.

Nicolas Fiorini, Kathi Canese, Grisha Starchenko, Evgeny Kireev, Won Kim, Vadim Miller, Maxim Osipov, Michael Kholodov, Rafis Ismagilov, Sunil Mohan, et al. 2018. Best match: new relevance search for pubmed. *PLoS biology*, 16(8):e2005343.

Jason A. Fries, Leon Weber, Natasha Seelam, Gabriel Altay, Debajyoti Datta, Samuele Garda, Sunny Kang, Rosaline Su, Wojciech Kusa, Samuel Cahyawijaya, Fabio Barth, Simon Ott, Matthias Samwald, Stephen H. Bach, Stella Biderman, Mario Sänger, Bo Wang, Alison Callahan, Daniel León Periñán, Théo Gigant, Patrick Haller, Jenny Chim, José D. Posada, John M. Giorgi, Karthik Rangasai Sivaraman, Marc Pàmies, Marianna Nezhurina, Robert Martin, Michael Cullan, Moritz Freidank, Nathan Dahlberg, Shubhanshu Mishra, Shamik Bose, Nicholas Broad, Yanis Labrak, Shlok Deshmukh, Sid Kiblawi, Ayush Singh, Minh Chien Vu, Trishala Neeraj, Jonas Golde, Albert Villanova del Moral, and Benjamin Beilharz. 2022. Bigbio: A framework for data-centric biomedical natural language processing. In *NeurIPS*.

Justin Gilmer, Samuel S. Schoenholz, Patrick F. Riley, Oriol Vinyals, and George E. Dahl. 2017. Neural message passing for quantum chemistry. In *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, volume 70 of *Proceedings of Machine Learning Research*, pages 1263–1272. PMLR.

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2022. Domain-specific language model pretraining for biomedical natural language processing. *ACM Trans. Comput. Heal.*, 3(1):2:1–2:23.

William L. Hamilton, Zhitao Ying, and Jure Leskovec. 2017. Inductive representation learning on large

graphs. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 1024–1034.

Xiao Huang, Jingyuan Zhang, Dingcheng Li, and Ping Li. 2019. Knowledge graph embedding based question answering. In *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining, WSDM 2019, Melbourne, VIC, Australia, February 11-15, 2019*, pages 105–113. ACM.

Seyed Mehran Kazemi and David Poole. 2018. Simple embedding for link prediction in knowledge graphs. In *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, pages 4289–4300.

Pei Ke, Haozhe Ji, Yu Ran, Xin Cui, Liwei Wang, Linfeng Song, Xiaoyan Zhu, and Minlie Huang. 2021. JointGT: Graph-text joint representation learning for text generation from knowledge graphs. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2526–2538, Online. Association for Computational Linguistics.

Kuzma Khrabrov, Ilya Shenbin, Alexander Ryabov, Artem Tsypin, Alexander Telepov, Anton Alekseev, Alexander Grishin, Pavel Strashnov, Petr Zhilyaev, Sergey Nikolenko, and Artur Kadurin. 2022. nablaDFT: Large-Scale conformational energy and hamiltonian prediction benchmark and dataset. *Phys. Chem. Chem. Phys.*, 24(42):25853–25863.

Jing Yu Koh, Daniel Fried, and Ruslan Salakhutdinov. 2023a. Generating images with multimodal language models. *NeurIPS*.

Jing Yu Koh, Ruslan Salakhutdinov, and Daniel Fried. 2023b. Grounding language models to images for multimodal inputs and outputs. In *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pages 17283–17300. PMLR.

Jan A. Kors, Simon Clematide, Saber A. Akhondi, Erik M. van Mulligen, and Dietrich Rebholz-Schuhmann. 2015. A multilingual gold-standard corpus for biomedical concept recognition: the mantra GSC. *J. Am. Medical Informatics Assoc.*, 22(5):948–956.

Anne Lauscher, Ivan Vulić, Edoardo Maria Ponti, Anna Korhonen, and Goran Glavaš. 2020. Specializing unsupervised pretraining models for word-level semantic similarity. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 1371–1383, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinform.*, 36(4):1234–1240.

Sunwon Lee, Donghyeon Kim, Kyubum Lee, Jaehoon Choi, Seongsoon Kim, Minji Jeon, Sangrak Lim, Donghee Choi, Sunkyu Kim, Aik-Choon Tan, et al. 2016. Best: next-generation biomedical entity search tool for knowledge discovery from biomedical literature. *PloS one*, 11(10):e0164680.

Patrick Lewis, Myle Ott, Jingfei Du, and Veselin Stoyanov. 2020. Pretrained language models for biomedical and clinical tasks: Understanding and extending the state-of-the-art. In *Proceedings of the 3rd Clinical Natural Language Processing Workshop*, pages 146–157, Online. Association for Computational Linguistics.

Bill Yuchen Lin, Xinyue Chen, Jamin Chen, and Xiang Ren. 2019. KagNet: Knowledge-aware graph networks for commonsense reasoning. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2829–2839, Hong Kong, China. Association for Computational Linguistics.

Fangyu Liu, Ehsan Shareghi, Zaiqiao Meng, Marco Basaldella, and Nigel Collier. 2021a. Self-alignment pretraining for biomedical entity representations. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4228–4238, Online. Association for Computational Linguistics.

Fangyu Liu, Ivan Vulić, Anna Korhonen, and Nigel Collier. 2021b. Learning domain-specialised representations for cross-lingual biomedical entity linking. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 565–574, Online. Association for Computational Linguistics.

Weijie Liu, Peng Zhou, Zhe Zhao, Zhiruo Wang, Qi Ju, Haotang Deng, and Ping Wang. 2020. K-BERT: enabling language representation with knowledge graph. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 2901–2908. AAAI Press.

Aidan Mannion, Didier Schwab, and Lorraine Goeuriot. 2023. UMLS-KGI-BERT: Data-centric knowledge integration in transformers for biomedical entity recognition. In *Proceedings of the 5th Clinical Natural Language Processing Workshop*, pages 312–322, Toronto, Canada. Association for Computational Linguistics.

Zaiqiao Meng, Fangyu Liu, Ehsan Shareghi, Yixuan Su, Charlotte Collins, and Nigel Collier. 2022. Rewire-then-probe: A contrastive recipe for probing biomedical knowledge of pre-trained language models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4798–4810, Dublin, Ireland. Association for Computational Linguistics.

George Michalopoulos, Yuanxin Wang, Hussam Kaka, Helen Chen, and Alexander Wong. 2021. Umls-BERT: Clinical domain knowledge augmentation of contextual embeddings using the Unified Medical Language System Metathesaurus. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1744–1753, Online. Association for Computational Linguistics.

Irina Nikishina, Alsu Vakhitova, Elena Tutubalina, and Alexander Panchenko. 2022. Cross-modal contextualized hidden state projection method for expanding of taxonomic graphs. In *Proceedings of TextGraphs-16: Graph-based Methods for Natural Language Processing*, pages 11–24, Gyeongju, Republic of Korea. Association for Computational Linguistics.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*.

Nathan Peiffer-Smadja, Timothy Miles Rawson, Raheelah Ahmad, Albert Buchard, P Georgiou, F-X Lescure, Gabriel Birgand, and Alison Helen Holmes. 2020. Machine learning for clinical decision support in infectious diseases: a narrative review of current applications. *Clinical Microbiology and Infection*, 26(5):584–595.

Yifan Peng, Shankai Yan, and Zhiyong Lu. 2019. Transfer learning in biomedical natural language processing: An evaluation of BERT and ELMo on ten benchmarking datasets. In *Proceedings of the 18th BioNLP Workshop and Shared Task*, pages 58–65, Florence, Italy. Association for Computational Linguistics.

Matthew E. Peters, Mark Neumann, Robert Logan, Roy Schwartz, Vidur Joshi, Sameer Singh, and Noah A. Smith. 2019. Knowledge enhanced contextual word representations. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 43–54, Hong Kong, China. Association for Computational Linguistics.

Andrey Sakhovskiy, Zulfat Miftahutdinov, and Elena Tutubalina. 2021. KFU NLP team at SMM4H 2021 tasks: Cross-lingual and cross-modal BERT-based models for adverse drug effects. In *Proceedings of the Sixth Social Media Mining for Health (#SMM4H) Workshop and Shared Task*, pages 39–43, Mexico City, Mexico. Association for Computational Linguistics.

Andrey Sakhovskiy, Natalia Semenova, Artur Kadurin, and Elena Tutubalina. 2023. Graph-enriched biomedical entity representation transformer. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction - 14th International Conference of the CLEF Association, CLEF 2023, Thessaloniki, Greece, September 18-21, 2023, Proceedings*, volume 14163 of *Lecture Notes in Computer Science*, pages 109–120. Springer.

Andrey Sakhovskiy and Elena Tutubalina. 2022. Multimodal model with text and drug embeddings for adverse drug reaction classification. *Journal of Biomedical Informatics*, 135:104182.

Sarvesh Soni and Kirk Roberts. 2021. An evaluation of two commercial deep learning-based information retrieval systems for COVID-19 literature. *Journal of the American Medical Informatics Association*, 28(1):132–137.

Zhiqing Sun, Zhi-Hong Deng, Jian-Yun Nie, and Jian Tang. 2019. Rotate: Knowledge graph embedding by relational rotation in complex space. In *International Conference on Learning Representations*.

Mujeen Sung, Minbyul Jeong, Yonghwa Choi, Donghyeon Kim, Jinhyuk Lee, and Jaewoo Kang. 2022. BERN2: an advanced neural biomedical named entity recognition and normalization tool. *Bioinformatics*, 38(20):4837–4839.

Mujeen Sung, Jinhyuk Lee, Sean Yi, Minji Jeon, Sungdong Kim, and Jaewoo Kang. 2021. Can language models be biomedical knowledge bases? In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4723–4734, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Reed T. Sutton, David Pincock, Daniel C. Baumgart, Daniel C. Sadowski, Richard N. Fedorak, and Karen I. Kroeker. 2020. An overview of clinical decision support systems: benefits, risks, and strategies for success. *NPJ digital medicine*, 3.

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*, volume 48 of *Proceedings of Machine Learning Research*, pages 2071–2080, New York, New York, USA. PMLR.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008.

Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio.

2018. Graph Attention Networks. *International Conference on Learning Representations*. Accepted as poster.

Quan Wang, Pingping Huang, Haifeng Wang, Songtai Dai, Wenbin Jiang, Jing Liu, Yajuan Lyu, Yong Zhu, and Hua Wu. 2019a. Coke: Contextualized knowledge graph embedding. *arXiv preprint arXiv:1911.02168*.

Xiaozhi Wang, Tianyu Gao, Zhaocheng Zhu, Zhengyan Zhang, Zhiyuan Liu, Juanzi Li, and Jian Tang. 2021. KEPLER: A unified model for knowledge embedding and pre-trained language representation. *Transactions of the Association for Computational Linguistics*, 9:176–194.

Xun Wang, Xintong Han, Weilin Huang, Dengke Dong, and Matthew R. Scott. 2019b. Multi-similarity loss with general pair weighting for deep metric learning. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 5022–5030. Computer Vision Foundation / IEEE.

David S Wishart, Yannick D Feunang, An C Guo, Elvis J Lo, Ana Marcu, Jason R Grant, Tanvir Sajed, Daniel Johnson, Carin Li, Zinat Sayeeda, Nazanin Assempour, Ithayavani Iynkkaran, Yifeng Liu, Adam Maciejewski, Nicola Gale, Alex Wilson, Lucy Chin, Ryan Cummings, Diana Le, Allison Pon, Craig Knox, and Michael Wilson. 2017. DrugBank 5.0: a major update to the DrugBank database for 2018. *Nucleic Acids Research*, 46(D1):D1074–D1082.

David S. Wishart, Craig Knox, Anchi Guo, Dean Cheng, Savita Shrivastava, Dan Tzur, Bijaya Gautam, and Murtaza Hassanali. 2008. Drugbank: a knowledgebase for drugs, drug actions and drug targets. *Nucleic Acids Res.*, 36(Database-Issue):901–906.

Weiqi Wu, Chengyue Jiang, Yong Jiang, Pengjun Xie, and Kewei Tu. 2023. Do PLMs know and understand ontological knowledge? In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3080–3101, Toronto, Canada. Association for Computational Linguistics.

Zhenqin Wu, Bharath Ramsundar, Evan N. Feinberg, Joseph Gomes, Caleb Geniesse, Aneesh S. Pappu, Karl Leswing, and Vijay Pande. 2018. Moleculenet: a benchmark for molecular machine learning. *Chem. Sci.*, 9:513–530.

Hanwen Xu, Jiayou Zhang, Zhirui Wang, Shizhuo Zhang, Megh Bhalerao, Yucong Liu, Dawei Zhu, and Sheng Wang. 2023. Graphprompt: Graph-based prompt templates for biomedical synonym prediction. In *Thirty-Seventh AAAI Conference on Artificial Intelligence, AAAI 2023, Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence, IAAI 2023, Thirteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2023, Washington, DC, USA, February 7-14, 2023*, pages 10576–10584. AAAI Press.

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, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Michihiro Yasunaga, Antoine Bosselut, Hongyu Ren, Xikun Zhang, Christopher D. Manning, Percy Liang, and Jure Leskovec. 2022a. Deep bidirectional language-knowledge graph pretraining. In *NeurIPS*.

Michihiro Yasunaga, Jure Leskovec, and Percy Liang. 2022b. LinkBERT: Pretraining language models with document links. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8003–8016, Dublin, Ireland. Association for Computational Linguistics.

Michihiro Yasunaga, Hongyu Ren, Antoine Bosselut, Percy Liang, and Jure Leskovec. 2021a. QA-GNN: Reasoning with language models and knowledge graphs for question answering. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 535–546, Online. Association for Computational Linguistics.

Michihiro Yasunaga, Hongyu Ren, Antoine Bosselut, Percy Liang, and Jure Leskovec. 2021b. QA-GNN: Reasoning with language models and knowledge graphs for question answering. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 535–546, Online. Association for Computational Linguistics.

Hongyi Yuan, Zheng Yuan, and Sheng Yu. 2022a. Generative biomedical entity linking via knowledge base-guided pre-training and synonyms-aware fine-tuning. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4038–4048, Seattle, United States. Association for Computational Linguistics.

Zheng Yuan, Zhengyun Zhao, Haixia Sun, Jiao Li, Fei Wang, and Sheng Yu. 2022b. Coder: Knowledge-infused cross-lingual medical term embedding for term normalization. *Journal of Biomedical Informatics*, 126:103983.

Miao Zhang, Rufeng Dai, Ming Dong, and Tingting He. 2022a. DRLK: Dynamic hierarchical reasoning with language model and knowledge graph for question answering. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5123–5133, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Xikun Zhang, Antoine Bosselut, Michihiro Yasunaga, Hongyu Ren, Percy Liang, Christopher D. Manning,

and Jure Leskovec. 2022b. Greaselm: Graph reasoning enhanced language models. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*.

Marinka Zitnik, Monica Agrawal, and Jure Leskovec. 2018. Modeling polypharmacy side effects with graph convolutional networks. *Bioinformatics*, 34(13):i457–i466.