BioKG: a comprehensive, large-scale biomedical knowledge graph for AI-powered, data-driven biomedical research

Yuan Zhang[1,+], Xin Sui[2,+], Feng Pan[2,+], Kaixian Yu[2,+], Keqiao Li[1], Shubo Tian[1], Arslan Erdengasileng[1], Qing Han[1], Wanjing Wang[1], Jianan Wang[2], Jian Wang[3], Donghu Sun[2], Henry Chung[2], Jun Zhou[2], Eric Zhou[2], Ben Lee[2], Peili Zhang[4], Xing Qiu[5], Tingting Zhao[6], Jinfeng Zhang[1,2,*]

[1] Department of Statistics, Florida State University, Tallahassee, FL 32306
[2] Insilicom LLC, Tallahassee, FL 32303
[3] 977 Wisteria Ter., Sunnyvale, CA 94086
[4] Forward Informatics, Winchester, Massachusetts, 01890
[5] Department of Biostatistics and Computational Biology, University of Rochester, Rochester, NY 14642
[6] Department of Geography, Florida State University, Tallahassee, FL 32306

[*] Correspondence: jinfeng@insilicom.com
[+] These authors contributed equally.

Abstract:

To cope with the rapid growth of scientific publications and data in biomedical research, knowledge graphs (KGs) have emerged as a powerful data structure for integrating large volumes of heterogeneous data to facilitate accurate and efficient information retrieval and automated knowledge discovery (AKD). However, transforming unstructured content from scientific literature into KGs has remained a significant challenge, with previous methods unable to achieve human-level accuracy. In this study, we utilized an information extraction pipeline that won first place in the LitCoin NLP Challenge to construct a large-scale KG using all PubMed abstracts. The quality of the large-scale information extraction rivals that of human expert annotations, signaling a new era of automatic, high-quality database construction from literature. Our extracted information markedly surpasses the amount of content in manually curated public databases. To enhance the KG's comprehensiveness, we integrated relation data from 40 public databases and relation information inferred from high-throughput genomics data. The comprehensive KG enabled rigorous performance evaluation of AKD, which was infeasible in previous studies. We designed an interpretable, probabilistic-based inference method to identify indirect causal relations and achieved unprecedented results for drug target identification and drug repurposing. Taking lung cancer as an example, we found that 40% of drug targets reported in literature could have been predicted by our algorithm about 15 years ago in a retrospective study, demonstrating that substantial acceleration in scientific discovery could be achieved through automated hypotheses generation and timely dissemination. A cloud-based platform (https://www.biokde.com) was developed for academic users to freely access this rich structured data and associated tools.

Introduction

The sheer volume of information produced daily in scientific literature, expressed in natural languages, makes it impractical to manually read all publications, even within relatively narrow research areas. Additionally, advances in high-throughput technologies have led to the creation of enormous quantities of research data, much of which remains underutilized in various databases. This information explosion poses a major challenge for researchers to identify and develop new ideas using all the available data. Automated knowledge discovery (AKD, a.k.a. automated hypothesis generation) can help mitigate this problem by automating the process of data analysis, identifying patterns, and generating new insights and hypotheses(*1*). In recent years, knowledge graphs (KGs) have been proposed as a powerful data structure for integrating heterogeneous data and for AKD(*2–8*). KGs, with entities as nodes and their relationships as edges, represent human knowledge in structured form, facilitating efficient and accurate information retrieval. Graph algorithms can be employed on KGs to infer potential new relationships as plausible hypotheses between known entities.

Computational construction of KGs from unstructured text entails two steps: named entity recognition (NER) to identify key biological entities and relation extraction (RE) to extract relationships among entities. Historically, NER and RE have been collectively referred to as information retrieval tasks. Early automated methods mainly fell into two categories: rule-based and machine learning-based. The rule-based approach systematically extracted specific data based on predefined rules (*9–14*), while the machine learning-based approaches inferred rules from annotated data usually with increased recall and overall performance (*15–30*). The advent of machine learning led to more sophisticated methods that leveraged semantic information and sentence structure, resulting in significant improvements in information extraction effectiveness(*20, 23*). However, a gap remained compared to human proficiency.

The emergence of deep learning models has allowed for a more nuanced utilization of information, such as semantic content and grammatical structures. By expanding the use of features and enhancing expressive capabilities, deep models have significantly improved the overall effectiveness of information

extraction(*31–43*). Recently, the technique of pretraining and large language models (LLMs) have garnered considerable attention, expanding both model complexity and the amount of training data and achieving remarkable progress in information retrieval tasks(*42*, *44–54*). This was evidenced by the significant results in the BioCreative VII Challenge in 2021, where finetuning BERT-based models was widely used, and the top performance in some tasks closely matched human annotator performance. Subsequently, a highly advanced series of pre-trained models, like GPT-4, emerged(*55–57*). These models have been proven to outperform humans in multiple more general tasks beyond information extraction, marking a significant breakthrough in the field.

To facilitate the methodology development and identification of the most effective methods for KG construction, the NCATS (National Center for Advancing Translational Sciences) of NIH (National Institutes of Health) organized the LitCoin natural language processing (NLP) challenge between Nov 2021 and Feb 2022. In the LitCoin NLP Challenge dataset, six common biological entity types were annotated: diseases, genes/proteins, chemical compounds, species, genetic variants, and cell lines. Eight relation types were also annotated for the entities: association, binding, comparison, conversion, cotreatment, drug interaction, positive correlation, and negative correlation. These entity types and relations are highly relevant in translational research and drug discoveries. Our team, JZhangLab@FSU, participated in the challenge and won first-place(*58*).

In this study, we applied the information retrieval pipeline we developed for the LitCoin NLP Challenge to all PubMed abstracts (with a cutoff date in May 2022) to construct a large-scale Biomedical Knowledge Graph (BioKG). Manual verification showed that this pipeline has achieved accuracy at the level of a human annotator. By annotating the directions for the relations in the LitCoin dataset and training a model to predict the direction of relations, we were able to construct a causal knowledge graph (CKG) capable of making indirect causal inferences. To further enhance the coverage of BioKG, we integrated relation data from public databases and high-throughput genomics datasets, making it the most comprehensive, high quality biomedical knowledge graph to date. To make causal inferences among the

entities that are not directly connected in the KG, we designed a probabilistic based approach, probabilistic semantic reasoning (PSR). PSR is highly interpretable as it directly infers indirect relations using direct relations through straightforward reasoning principles.

Navigating the modern drug development terrain is intricate and resource-intensive(*59*). The ascent in costs largely stems from prior research exhausting more straightforward drug targets, necessitating a shift towards more complex ones(*60*). In this setting, knowledge graphs play a pivotal role in automated knowledge discovery (AKD)(*8*, *61–63*), particularly in the domain of drug target identification and drug repurposing (*64–68*). A significant challenge in developing methods for such applications has been to comprehensively assess the effectiveness of these studies. For example, in the case of drug repurposing, collecting all the known therapeutic associations of a particular disease or drug requires thorough search of literature. Without such knowledge, it is impossible to rigorously evaluate drug repurposing methods. In our investigation, for each repurposing objective, we extracted all therapeutic associations documented in PubMed abstracts, which enables us to measure recall and observed positive rate (OPR), infeasible in prior drug repurposing research.

Lastly, we demonstrate the power of BioKG using two important problems in drug discovery: drug target identification and drug repurposing. Our method identified numerous viable candidates, supported by substantial literature evidence connecting the drug (or drug target) and disease entities. This level of interpretability is invaluable when determining the necessity of subsequent research endeavors.

Results

Constructing a large-scale and high-quality biomedical knowledge graph (BioKG)

In constructing BioKG, we processed over 34 million PubMed abstracts, resulting in 10,686,927 unique entities and 30,758,640 unique relations. The performance of the method in the LitCoin Challenge can be found in supplementary materials (Table S2). We incorporated entity normalization into our

pipeline, as this was not a component of the LitCoin challenge (see Supplementary Materials for more details).

We evaluated the accuracy of our large-scale relation extraction (RE) and our novelty prediction results using a sample of 50 randomly selected PubMed abstracts (refer to Table S3). The results indicate that the performance of our information extraction rivals that of human annotations, with a more in-depth analysis available in the Supplementary Materials.

We further integrated relation data from 40 public databases and from analysis results of some commonly used public genomics datasets (Supplementary Materials). After data integration, BioKG has 11,479,285 unique entities across 12 types (e.g., genes, diseases, chemical compounds) and 42,504,077 relations from 52 types (e.g., associations, positive correlation, cotreatment, etc.).

Figure 1A shows the numbers of PubMed abstracts containing one or more of the four major types of entities: diseases, genes, chemicals, and sequence variants. It is evident that diseases are the most common topic with over 20 million articles referencing at least one disease entity, and nearly half of these focusing exclusively on diseases. In contrast, gene mentions often coexist with other entities, such as chemicals and diseases. Figure 1B depicts the numbers of PubMed abstracts containing one or more of the five major types of relations, offering insight into the distribution of topics in biomedical research.

Figure 1C compares the relations extracted from PubMed with those from databases and the LitCoin dataset. There is a clear difference between the LitCoin dataset and general PubMed abstracts, as the former contains more relation in each abstract, especially those involving sequence variant entities(*69*), which explains the performance difference of our pipeline on these two datasets. Relations from PubMed and public databases are also quite complimentary to each other.

Figure 1D shows the number of novel discoveries over the year for different entity pairs. We observe a remarkable upswing in disease-gene relations since 2005, which underscores the tangible outcomes of translational initiatives promoted by federal agencies. Furthermore, the increasing number of disease-gene

relations signifies an improved understanding of disease mechanisms at the molecular level, thereby bolstering efforts in drug discovery. Of particular note is the rapid escalation of chemical-disease relations in recent years, particularly around 2020, which is anticipated to continue in the foreseeable future.

We plotted $p(k)$ vs k, where k is the degree of an entity in BioKG and $p(k)$ is the probability of an entity having degree k (Figure 1E). We found that BioKG exhibits a scale-free topology with an alpha parameter value around 3.0 (more details in Supplementary Materials).

Figure 1F compares the numbers of relations for five types of entity pairs from all the public databases integrated in BioKG, those extracted from PubMed, and the numbers extracted if we use a simple co-occurrence rule, which considers two entities having a relation if they co-occur in an abstract. On one hand, BioKG has significantly more numbers of relations than those from public databases (the y-axis shows Log(count)). On the other hand, the numbers of co-occurrences are much larger than relations extracted from PubMed, indicating a substantial noise reduction by explicitly extracting relations from literature compared to retrieval using keywords.

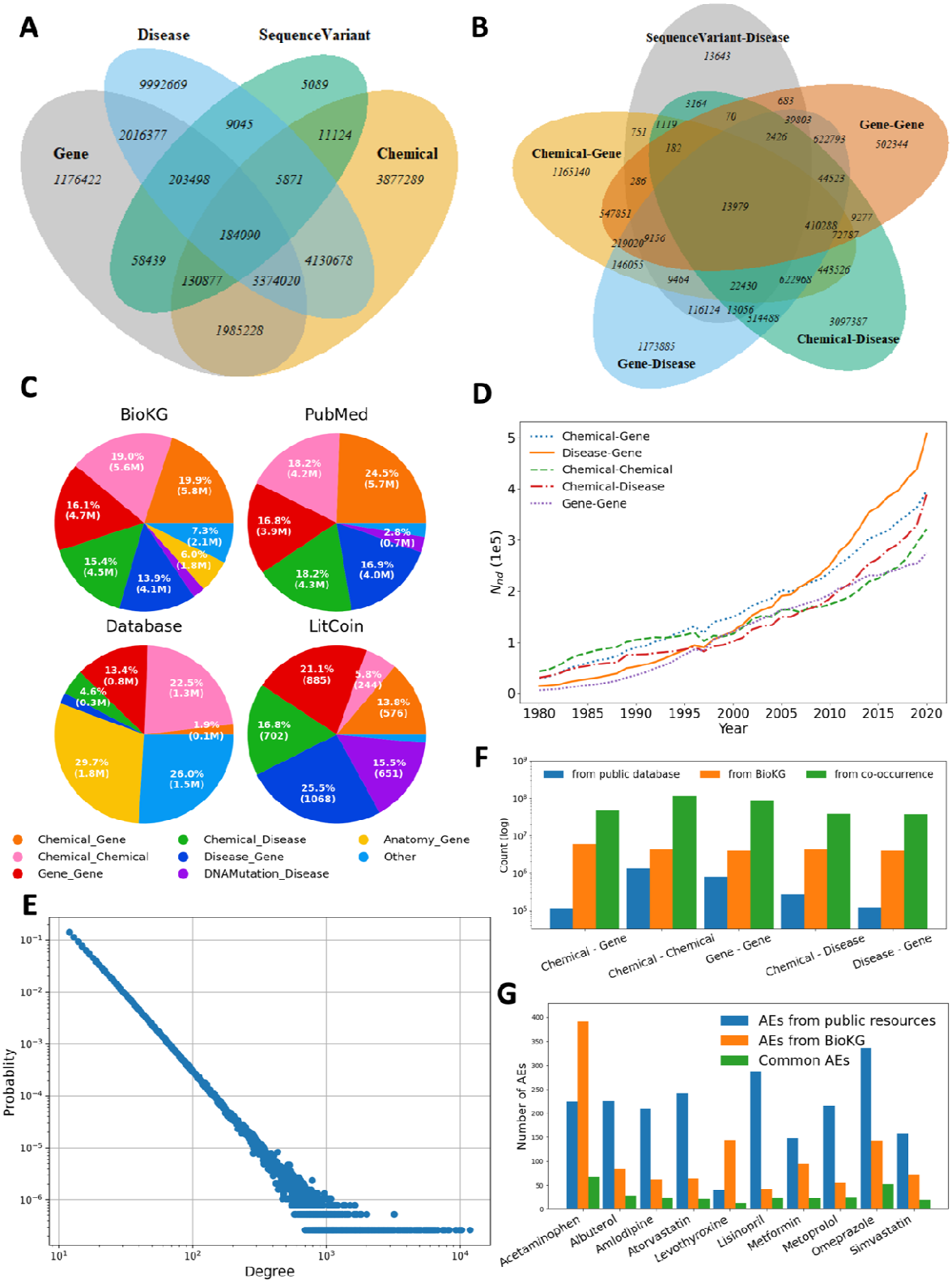
A
Disease
SequenceVariant
Gene
Chemical
9992669
5089
2016377
9045
11124
1176422
203498
5871
3877289
184090
58439
4130678
130877
3374020
1985228
B
SequenceVariant-Disease
13643
Gene-Gene
502344
Chemical-Gene
1165140
3164
683
70
39803
751
1119
2426
622793
182
286
44525
13979
547851
410288
9277
72787
219020
9156
146055
445526
9404
22430
622968
116124
13056
514488
3097387
Chemical-Disease
1173885
Gene-Disease
C
BioKG
19.0% (5.6M)
19.9% (5.8M)
16.1% (4.7M)
7.3% (2.1M)
6.0% (1.8M)
15.4% (4.5M)
13.9% (4.1M)
PubMed
18.2% (4.2M)
24.5% (5.7M)
16.8% (3.9M)
2.8% (0.7M)
18.2% (4.3M)
16.9% (4.0M)
Database
13.4% (0.8M)
22.5% (1.3M)
4.6% (0.3M)
1.9% (0.1M)
29.7% (1.8M)
26.0% (1.5M)
LitCoin
21.1% (885)
5.8% (244)
13.8% (576)
16.8% (702)
15.5% (651)
25.5% (1068)
Chemical_Gene
Chemical_Chemical
Gene_Gene
Chemical_Disease
Disease_Gene
DNAMutation_Disease
Anatomy_Gene
Other
D
Chemical-Gene
Disease-Gene
Chemical-Chemical
Chemical-Disease
Gene-Gene
N_{nd} (1e5)
Year
E
Probablity
Degree
F
from public database
from BioKG
from co-occurrence
Count (log)
Chemical - Gene
Chemical - Chemical
Gene - Gene
Chemical - Disease
Disease - Gene
G
AEs from public resources
AEs from BioKG
Common AEs
Number of AEs
Acetaminophen
Albuterol
Amlodipine
Atorvastatin
Levothyroxine
Lisinopril
Metformin
Metoprolol
Omeprazole
Simvastatin

Figure 1. A. Venn diagram for the numbers of PubMed articles containing certain types of entities; **B.** Venn diagram for the numbers of PubMed articles containing certain types of relations; **C.** The composition of relations in BioKG, PubMed abstracts, public databases, and LitCoin dataset; **D.** The numbers of novel discoveries by entity pair type from 1980 to 2022. **E.** Degree distribution of BioKG, where the x-axis represents the degree k of an individual entity, and the y-axis *p(k)* denotes the corresponding probability of any entity exhibiting that degree. **F.** The numbers of relations between entity types in BioKG compared with 40 public databases combined and compare with the numbers from co-occurrence. **G.** Adverse events for 10 common drugs in BioKG compared to those in the public domain.

A Case Study in Pharmacovigilance

Drug safety represents a critical component in pharmaceutical development and production processes. We evaluate the adverse event (AE) information in BioKG by comparing it with those in the OpenFDA (*70*) and SIDER (*65*) databases for ten well-known drugs (Supplementary Materials). The AE information in BioKG is the positive correlation relation between chemical entities and diseases (including phenotypical features) with direction from chemicals to diseases. Figure 1G shows that BioKG contains significant numbers of AEs not documented in public databases for all ten drugs (Table S5 in Supplementary Materials). We conducted a manual examination of ten unique AEs for each drug that were in BioKG but absent in the other databases and found that a majority of these AEs were indeed true, with a precision of 97%. To further validate that we were indeed uncovering new, previously unrecorded AEs, as opposed to identifying synonyms of known AEs, we employed ChatGPT to check for synonyms for an additional layer of verification. We found, on average, over 78% of the 100 randomly sampled AEs for the ten drugs were novel findings. This result demonstrated the effectiveness of our literature-based approach for uncovering a significant proportion of AEs overlooked in professional databases. It highlights the challenges of accessing such information without accurate information extraction tools, positioning BioKG as a vital resource in drug safety.

Constructing a causal knowledge graph

We annotated the direction information for the relations in the LitCoin dataset and trained a model for predicting the direction of relations. This allowed us to construct a causal knowledge graph (Supplementary Materials) for knowledge discovery applications.

Probabilistic semantic reasoning (PSR) for inference of relationships between indirectly connected entities

With the direction information available, we can infer relations between entities not directly connected using straight forward reasoning. To the end, we designed the probabilistic semantic reasoning (PSR) algorithm (Supplementary Materials). The algorithm is highly efficient and interpretable. The efficiency of PSR makes it possible to perform all-against-all drug repurposing for all drugs and all diseases using a limited computational resource. It also allows efficient updating of newly inferred relations. For example, we can download newly published articles from PubMed, make inferences based on extracted new discoveries, and publish the automatically generated hypotheses on a daily basis to disseminate them timely. Machine learning based methods would have difficulties achieving either such efficiency or interpretability.

Drug repurposing for Covid-19 using causal knowledge graph

Using the PSR algorithm, we conducted a retrospective, real-time drug repurposing study for COVID-19 (72) spanning from March 2020 to May 2023 (Figure 2). During this period, we consistently discovered repurposed drugs based on the drug targets reported for COVID-19 between March and June 2020. A candidate drug is defined as one that has a directed path through a gene to COVID-19. Our monthly assessments involved scrutinizing whether these repurposed drugs had been subsequently tested in clinical trials documented on ClinicalTrials.gov or had published therapeutic efficacy in COVID-19 patients in PubMed abstracts. It is noteworthy that drugs identified in clinical trials may not always translate into effective treatments for COVID-19. Nevertheless, they serve as valuable hypotheses, aligning with the primary objective of our drug repurposing approach. As shown in **Error! Reference source not found.**A, we were able to identify nearly 600 to 1,400 candidate drugs from our causal KG using PSR. Remarkably, one-third of the repurposed drugs identified during the initial two months were later validated as effective treatments or plausible potential treatments worthy of clinical trials.

Importantly, even drugs that did not achieve validation status remain viable hypotheses, warranting further investigation, particularly when existing treatments prove less than optimal.

Figure 2B presents a timeline showcasing the validation of repurposed drugs. Notably, there is a surge in validated drugs during the first year, which subsequently shows a month-to-month decline. This pattern suggests that a majority of the repurposed drugs align with practitioners' assessments. Interestingly, some drugs only received validation in the second or even the third year, hinting that these might not have been as immediately evident or persuasive as those tested earlier. A deeper analysis is warranted to discern any significant variances between drugs validated at different times. The count of drugs validated through publications is on par with those through clinical trials. While numerous drug repurposing studies for COVID-19 exist (*71–74*), as per our understanding, no prior research has as thoroughly validated such a vast quantity of repurposed drugs as we have in this research. These findings underscore our causal KG's proficiency in pinpointing promising drug candidates for specific diseases in real-time scenarios.

We then conducted drug repurposing for COVID-19 in the current timeframe (Figure 2C). We did not exclude drugs already reported as treatments for COVID-19 (direct relations). This was to observe if our repurposing efforts agree with existing treatment choices for COVID-19. Figure 2C displays the top 50 repurposed drugs. Notably, a majority of them (38 out of 50) were known treatments for COVID-19. Among the remaining 12, 11 have been proposed as potential treatments for COVID-19 (citations provided in Supplementary Materials Table S4). For each of the drugs, numerous genes that link COVID-19 with the drug were identified (y-axis of Figure 2C). Additionally, each of these relations, whether drug-gene or gene-COVID-19, is supported by one or multiple articles. To our knowledge, none of the previous literature-based COVID-19 repurposing studies has yielded such comprehensive findings.

Drug repurposing for 10 diseases and 10 drugs

To further assess our method's applicability across varied conditions, we extended the drug repurposing to ten diseases that lack satisfactory treatments and ten commonly prescribed drugs (as illustrated in Figure 3). Our PSR algorithm identified a vast array of candidates for these drugs and

indications. For each drug (or disease) assessed, we calculated both the recall and the observed positive rate (OPR). Historically, these metrics posed challenges in estimation due to the reliance on manual literature searches. The recall is defined as the percentage of known direct relations successfully repurposed, while OPR quantifies the percentage of repurposed cases that have reported direct relations. Instead of precision, we opted for OPR, considering that numerous repurposed candidates could be valid prospects awaiting further validation. Impressively, our findings revealed average recall values of 0.91 for disease repurposing and 0.904 for drug repurposing. This exceptional recall rate emphasizes the potency of BioKG coupled with our PSR algorithm in spotlighting viable drug repurposing candidates. Notably, these elevated recall rates were achieved without an excessive number of predictions. The observed OPRs remained commendable at 0.197 for diseases and 0.07147 for drugs. Importantly, a significant proportion of indications repurposed for these drugs are not associated with any treatments in PubMed abstracts. This suggests that certain ailments might still be without treatments, and these widely used drugs could potentially fill those therapeutic gaps.

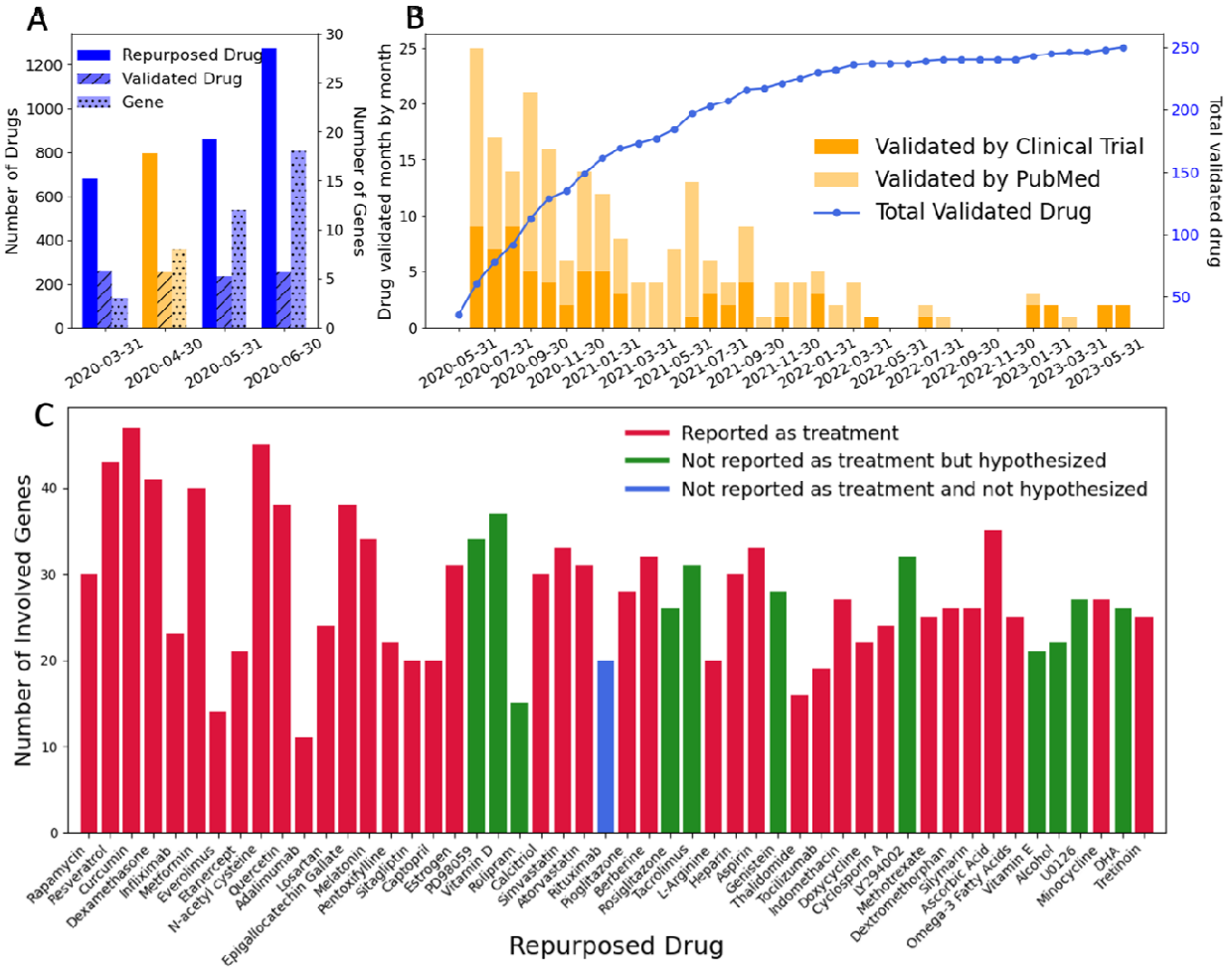

Figure 2. Drug repurposing for COVID-19. **A.** The number of repurposed drugs, number of verified drugs, and number of COVID-19 related genes for the first four months of the COVID-19 pandemic (March to June 2020); **B.** The number of verified drugs each month for those repurposed for Apr 2020; **C.** The number of genes involved in the drugs repurposed at present time (March 2023). The figure shows the top 50 repurposed drugs sorted from left to right, with those on the left having higher scores. Almost all the repurposed drugs interact with many genes (height of the bar) related to COVID-19. The majority of the drugs were reported as treatment for COVID-19 (38 out of 50). Among those that were not reported as treatments for COVID-19, 11 out of 12 were hypothesized as potential treatments.

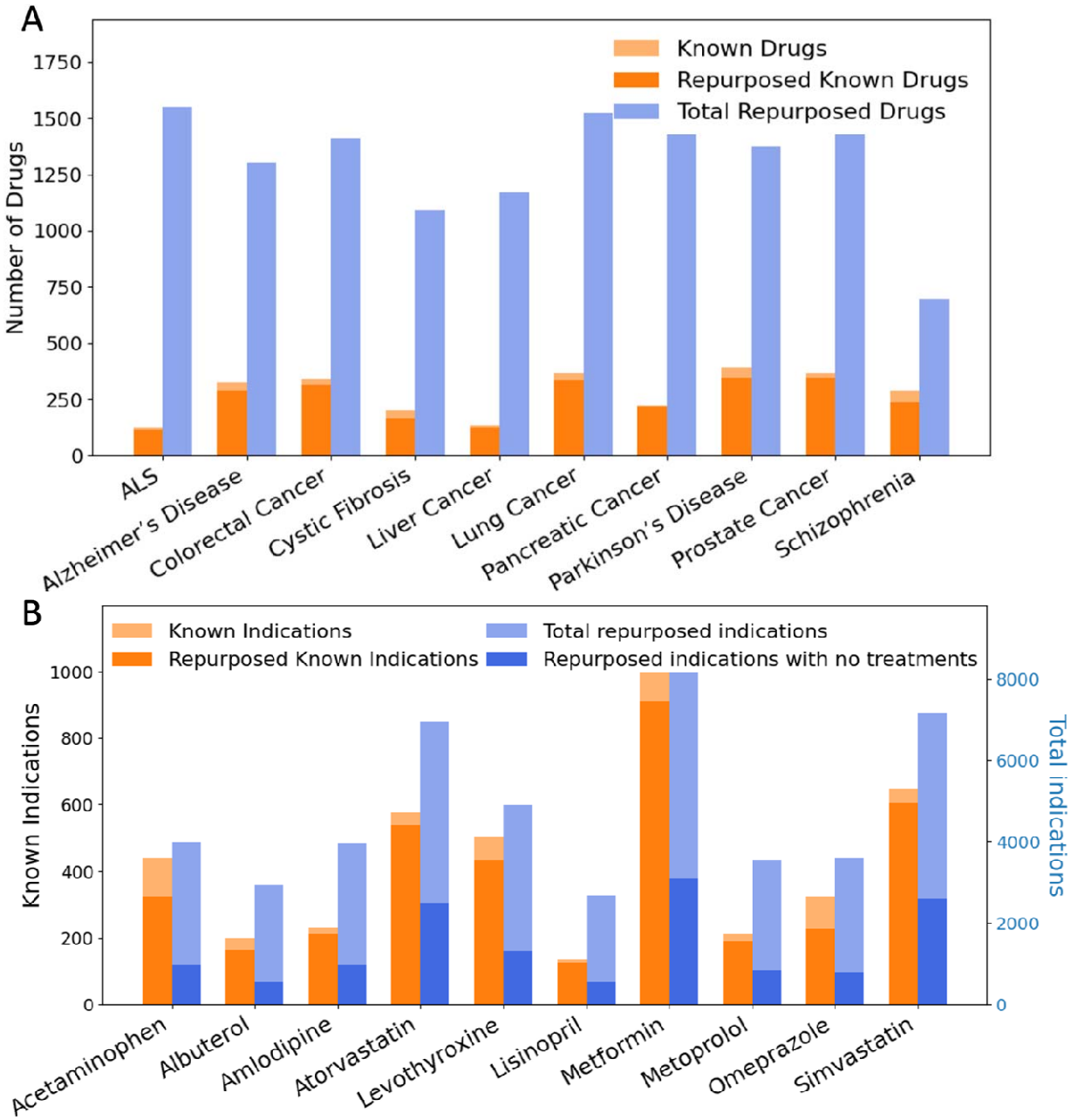

Figure 3. Drug repurposing for 10 diseases and 10 common drugs. Our method identified large numbers of candidates. The known drugs or indications are those extracted from PubMed abstracts, not FDA approved drugs or indications. On average, our method has repurposed more than 90% of the known drugs or indications. **A.** Drug repurposing for 10 diseases without satisfactory treatments; **B.** Drug repurposing for 10 common drugs. Among the repurposed indications, some (dark blue bar) do not have reported treatments in PubMed abstracts, which suggests unmet medical needs.

Drug target identification for lung cancer

In addition to drug repurposing, a causal knowledge graph serves as a powerful tool for pinpointing potential drug targets (genes) for diseases. These targets are entities not directly linked with the disease in

existing literature but possess an indirect, more subtle, connection through other genes or entities. As a case study to demonstrate the power of BioKG for target identification (TI), we applied PSR to infer indirect relation from genes/proteins to lung cancer. The TI procedure was conducted annually from 1986 to 2009 (Figure 4A and 4B). Notably, since the start of this century, we have consistently proposed at least 100 potential drug targets each year. If a proposed target was later reported as a direct relation with direction from the gene to the disease, we consider the proposed target was validated later. The average recall from 1989 to 2022 is 0.4071 (Figure 4B) and the average OPR from 1988 to 2009 is 0.3237 (Figure 4A). These two metrics were calculated for different time intervals because to calculate recall we need to first predict targets for some years and targets predicted for recent years still need time to be validated. We also calculated the average duration it typically takes for these targets to be validated. Intriguingly, our proposals for potential drug targets precede their appearance in literature from 4.5 to 24 years with a median of 15 years (Figure 4A). If we assume that validating a drug target experimentally takes on average 5 years, then in a hypothetical scenario, BioKG may cut the time from 15 years to 5 if the researchers had known the potential drug targets right after they were predicted and started to confirm them experimentally. With a recall rate exceeding 40% and a lag of 15 years, it is evident that many of our present-day discoveries could have been accelerated, demonstrating the potential of the causal KG for significantly expediting the discovery of new drug targets and new drug development.

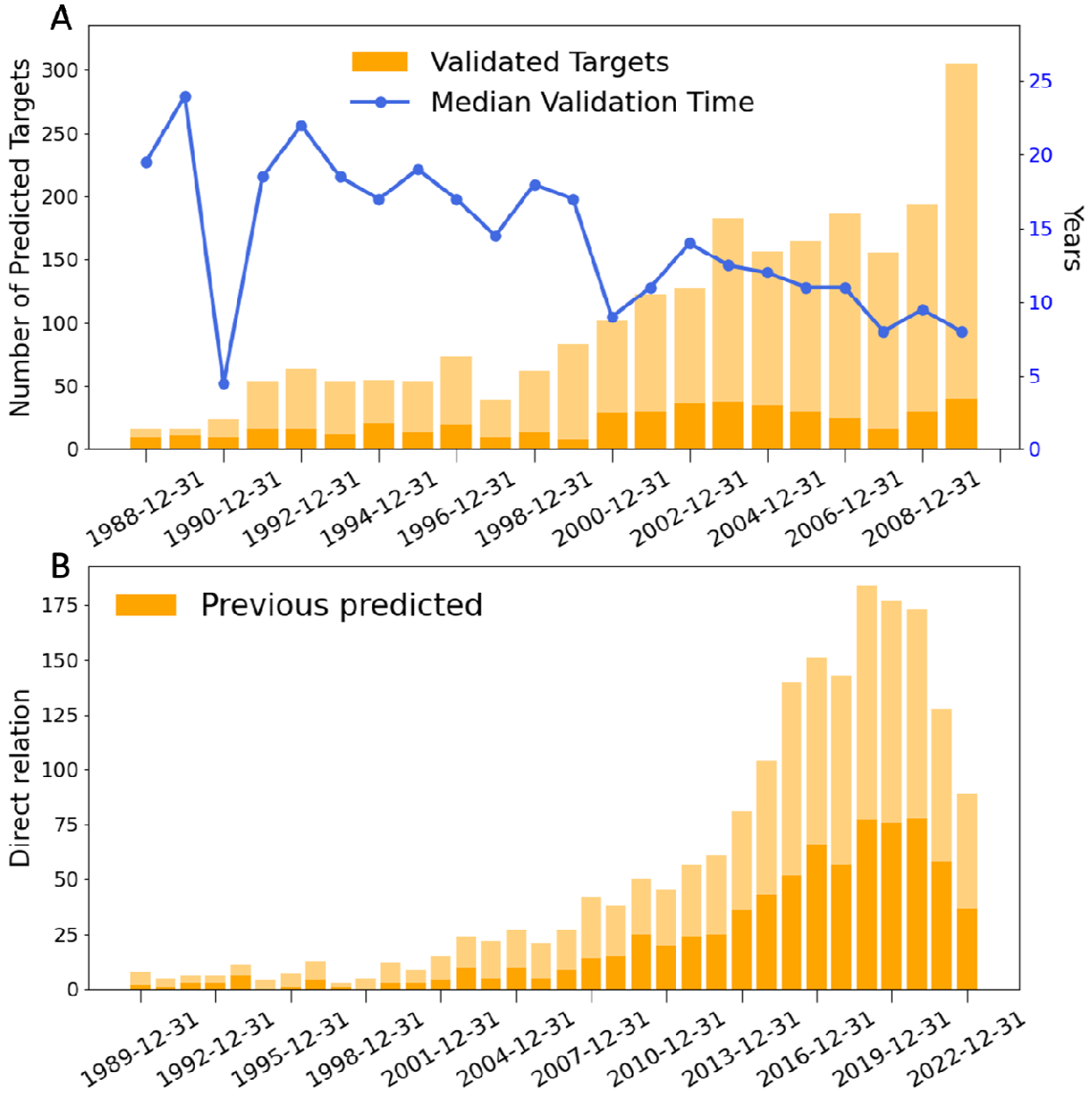

Figure 4. Drug target identification for lung cancer. A. The number of identified targets from the year 1988 to 2009. The dark yellow bar shows the number of validated targets. The blue line shows the time it takes to validate the targets for the corresponding years. **B.** The number of drug targets reported in PubMed from the year 1989 to year 2022. The dark yellow bar shows the targets that have been predicted previously.

Discussion

Converting unstructured scientific literature into structured data has been a long-standing challenge in natural language processing (NLP). Successfully addressing this issue has the potential to revolutionize the pace of scientific discoveries. Although there have been numerous studies over the years, computational methods have yet to achieve the precision of manual annotation in relation extraction, posing a significant hurdle. The emergence of LLMs in recent years has ushered in noteworthy advancements in information extraction through LLM fine-tuning. In this paper, we report the first utilization of a human-level information extraction pipeline to construct a large-scale biomedical knowledge graph by processing all the abstracts in PubMed. By further integrating relation data from 40 public databases and those analyzed from publicly available genomics data, the resulting knowledge graph, dubbed BioKG, stands out as perhaps the most all-encompassing biomedical knowledge graph constructed so far. The coverage of BioKG is much larger than public databases for the relations we have extracted. Construction of a causal knowledge graph and design of an interpretable PSR algorithm allow us to perform automated knowledge discovery very effectively. The exhaustive nature of BioKG allows us to perform research that was infeasible previously. For the first time, we were able to evaluate the performance of automated knowledge discoveries systematically and rigorously by calculating recalls and observed positive rates (OPRs). Without the knowledge of all PubMed abstracts in structured form, one must perform manual search of the literature, which would not be feasible for a relatively large number of predictions. We summarize the notable advances in this study including some unique BioKG-enabled capabilities in Box S1 (Supplementary Material) and discuss some of them below.

The biomedical research community has traditionally invested significant resources and human effort in knowledge curation through manual annotations. Our research suggests a paradigm shift, leveraging the capabilities of modern LLMs. By initially producing a limited set of high-caliber labeled data, it is feasible to train an information extraction model that operates at human-level precision on much larger

text datasets. This methodology could notably expand the reach of public databases without compromising data quality.

Utilizing BioKG for knowledge discovery tasks, such as drug repurposing and drug target identification, has yielded a vast array of credible candidates, supported by an unparalleled volume of literature evidence. This underscores the potential of structured knowledge in hastening scientific breakthroughs. In our drug repurposing endeavors for COVID-19, we highlighted BioKG's proficiency in identifying treatments for pandemics, marking it as an indispensable asset for potential future outbreaks. Moreover, with the drug target analysis for lung cancer, we demonstrated that BioKG can notably expedite the drug target identification process for researchers.

Finally, we would like to put our study in the context of the LLMs popular in the current NLP research community. While LLMs have showcased exceptional capabilities in understanding and generating natural language, they aren't without shortcomings. A notable limitation is their fixed knowledge cut-off date, which restricts their awareness of the very latest developments. Furthermore, in biomedical research, where precision is crucial, relying solely on LLMs to answer specific questions risks inaccuracies due to their limited knowledge base. Additionally, LLMs possess a propensity to generate text that, while convincingly articulated, may lack factual accuracy. This propensity raises concerns regarding the veracity of answers generated by LLMs, necessitating mechanisms for verification and the production of more substantiated results, possibly with appropriate citations. We believe that integrating knowledge graphs like BioKG with LLMs can effectively mitigate these limitations. To this end, we are actively developing a comprehensive question-answering system, combining BioKG with an open-source LLM.

In the Supplementary Materials, we delve into future research avenues and the challenges we've faced. In summary, BioKG serves as a powerful enabler for more effective and efficient information retrieval and automated knowledge discovery.

Acknowledgements

This work was supported in part by the National Institute of General Medical Sciences of the National Institute of Health under award number R01GM126558 for J.Z. The funders had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript.

Bibliography

1. H. Kitano, Nobel Turing Challenge: creating the engine for scientific discovery. *NPJ Syst Biol Appl* **7**, 29 (2021).

2. S. Yu, Z. Yuan, J. Xia, S. Luo, H. Ying, S. Zeng, J. Ren, H. Yuan, Z. Zhao, Y. Lin, K. Lu, J. Wang, Y. Xie, H.-Y. Shum, BIOS: An Algorithmically Generated Biomedical Knowledge Graph. (2022).

3. D. N. Nicholson, C. S. Greene, Constructing knowledge graphs and their biomedical applications. *Comput Struct Biotechnol J* **18**, 1414–1428 (2020).

4. Z. Gao, P. Ding, R. Xu, KG-Predict: A knowledge graph computational framework for drug repurposing. *J Biomed Inform* **132**, 104133 (2022).

5. N. Li, Z. Yang, L. Luo, L. Wang, Y. Zhang, H. Lin, J. Wang, KGHC: a knowledge graph for hepatocellular carcinoma. *BMC Med Inform Decis Mak* **20**, 135 (2020).

6. P. Ernst, A. Siu, G. Weikum, KnowLife: a versatile approach for constructing a large knowledge graph for biomedical sciences. *BMC Bioinformatics* **16**, 157 (2015).

7. S. Zheng, J. Rao, Y. Song, J. Zhang, X. Xiao, E. F. Fang, Y. Yang, Z. Niu, PharmKG: a dedicated knowledge graph benchmark for bomedical data mining. *Brief Bioinform* **22** (2021).

8. L. Li, P. Wang, J. Yan, Y. Wang, S. Li, J. Jiang, Z. Sun, B. Tang, T.-H. Chang, S. Wang, Y. Liu, Real-world data medical knowledge graph: construction and applications. *Artif Intell Med* **103**, 101817 (2020).

9. J.-H. Kim, P. C. Woodland, “A rule-based named entity recognition system for speech input” in *6th International Conference on Spoken Language Processing (ICSLP 2000)* (ISCA, ISCA, 2000), p. vol. 1, 528-531-0.

10. Y. Miyao, K. Sagae, R. Sætre, T. Matsuzaki, J. Tsujii, Evaluating contributions of natural language parsers to protein–protein interaction extraction. *Bioinformatics* **25**, 394–400 (2009).

11. J. Lee, S. Kim, S. Lee, K. Lee, J. Kang, On the efficacy of per-relation basis performance evaluation for PPI extraction and a high-precision rule-based approach. *BMC Med Inform Decis Mak* **13**, S7 (2013).

12. K. Raja, S. Subramani, J. Natarajan, PPInterFinder—a mining tool for extracting causal relations on human proteins from literature. *Database* **2013** (2013).

13. J.-H. Kim, I.-H. Kang, K.-S. Choi, "Unsupervised named entity classification models and their ensembles" in *Proceedings of the 19th International Conference on Computational Linguistics - -* (Association for Computational Linguistics, Morristown, NJ, USA, 2002), pp. 1–7.

14. G. Petasis, F. Vichot, F. Wolinski, G. Paliouras, V. Karkaletsis, C. D. Spyropoulos, "Using machine learning to maintain rule-based named-entity recognition and classification systems" in *Proceedings of the 39th Annual Meeting on Association for Computational Linguistics - ACL '01* (Association for Computational Linguistics, Morristown, NJ, USA, 2001), pp. 426–433.

15. D. Tikk, P. Thomas, P. Palaga, J. Hakenberg, U. Leser, A Comprehensive Benchmark of Kernel Methods to Extract Protein–Protein Interactions from Literature. *PLoS Comput Biol* **6**, e1000837 (2010).

16. Q.-C. Bui, S. Katrenko, P. M. A. Sloot, A hybrid approach to extract protein–protein interactions. *Bioinformatics* **27**, 259–265 (2011).

17. R. Patra, S. K. Saha, A Kernel-Based Approach for Biomedical Named Entity Recognition. *The Scientific World Journal* **2013**, 1–7 (2013).

18. L. Hong, J. Lin, S. Li, F. Wan, H. Yang, T. Jiang, D. Zhao, J. Zeng, A novel machine learning framework for automated biomedical relation extraction from large-scale literature repositories. *Nat Mach Intell* **2**, 347–355 (2020).

19. H.-T. Zhang, M.-L. Huang, X.-Y. Zhu, A Unified Active Learning Framework for Biomedical Relation Extraction. *J Comput Sci Technol* **27**, 1302–1313 (2012).

20. K. Yu, P.-Y. Lung, T. Zhao, P. Zhao, Y.-Y. Tseng, J. Zhang, Automatic extraction of protein-protein interactions using grammatical relationship graph. *BMC Med Inform Decis Mak* **18**, 42 (2018).

21. R. Chowdhary, J. Zhang, J. S. Liu, Bayesian inference of protein–protein interactions from biological literature. *Bioinformatics* **25**, 1536–1542 (2009).

22. P. Corbett, A. Copestake, Cascaded classifiers for confidence-based chemical named entity recognition. *BMC Bioinformatics* **9**, S4 (2008).

23. P.-Y. Lung, Z. He, T. Zhao, D. Yu, J. Zhang, Extracting chemical–protein interactions from literature using sentence structure analysis and feature engineering. *Database* **2019** (2019).

24. L. Bell, R. Chowdhary, J. S. Liu, X. Niu, J. Zhang, Integrated Bio-Entity Network: A System for Biological Knowledge Discovery. *PLoS One* **6**, e21474 (2011).

25. S. Kim, J. Yoon, J. Yang, Kernel approaches for genic interaction extraction. *Bioinformatics* **24**, 118–126 (2008).

26. L. Bell, J. Zhang, X. Niu, "Mixture of logistic models and an ensemble approach for protein-protein interaction extraction" in *Proceedings of the 2nd ACM Conference on Bioinformatics, Computational Biology and Biomedicine* (ACM, New York, NY, USA, 2011), pp. 371–375.

27. R. Florian, A. Ittycheriah, H. Jing, T. Zhang, "Named entity recognition through classifier combination" in *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003* - (Association for Computational Linguistics, Morristown, NJ, USA, 2003), pp. 168–171.

28. R. Leaman, C.-H. Wei, Z. Lu, tmChem: a high performance approach for chemical named entity recognition and normalization. *J Cheminform* **7**, S3 (2015).

29. J. Qu, A. Steppi, D. Zhong, J. Hao, J. Wang, P.-Y. Lung, T. Zhao, Z. He, J. Zhang, Triage of documents containing protein interactions affected by mutations using an NLP based machine learning approach. *BMC Genomics* **21**, 773 (2020).

30. L. Li, R. Zhou, D. Huang, Two-phase biomedical named entity recognition using CRFs. *Comput Biol Chem* **33**, 334–338 (2009).

31. D. He, H. Zhang, W. Hao, R. Zhang, K. Cheng, A Customized Attention-Based Long Short-Term Memory Network for Distant Supervised Relation Extraction. *Neural Comput* **29**, 1964–1985 (2017).

32. F. Li, M. Zhang, G. Fu, D. Ji, A neural joint model for entity and relation extraction from biomedical text. *BMC Bioinformatics* **18**, 198 (2017).

33. G. Crichton, S. Pyysalo, B. Chiu, A. Korhonen, A neural network multi-task learning approach to biomedical named entity recognition. *BMC Bioinformatics* **18**, 368 (2017).

34. L. Luo, Z. Yang, P. Yang, Y. Zhang, L. Wang, H. Lin, J. Wang, An attention-based BiLSTM-CRF approach to document-level chemical named entity recognition. *Bioinformatics* **34**, 1381–1388 (2018).

35. Z. Guo, Y. Zhang, W. Lu, "Attention Guided Graph Convolutional Networks for Relation Extraction" in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (Association for Computational Linguistics, Stroudsburg, PA, USA, 2019), pp. 241–251.

36. M. Gridach, Character-level neural network for biomedical named entity recognition. *J Biomed Inform* **70**, 85–91 (2017).

37. S. Lim, J. Kang, Chemical–gene relation extraction using recursive neural network. *Database* **2018** (2018).

38. J. Gu, F. Sun, L. Qian, G. Zhou, Chemical-induced disease relation extraction via convolutional neural network. *Database* **2017** (2017).

39. M. Habibi, L. Weber, M. Neves, D. L. Wiegandt, U. Leser, Deep learning with word embeddings improves biomedical named entity recognition. *Bioinformatics* **33**, i37–i48 (2017).

40. S. Liu, F. Shen, R. Komandur Elayavilli, Y. Wang, M. Rastegar-Mojarad, V. Chaudhary, H. Liu, Extracting chemical–protein relations using attention-based neural networks. *Database* **2018** (2018).

41. H. Wu, J. Huang, Joint Entity and Relation Extraction Network with Enhanced Explicit and Implicit Semantic Information. *Applied Sciences* **12**, 6231 (2022).

42. A. Akbik, T. Bergmann, R. Vollgraf, "Pooled Contextualized Embeddings for Named Entity Recognition" in *Proceedings of the 2019 Conference of the North* (Association for Computational Linguistics, Stroudsburg, PA, USA, 2019), pp. 724–728.

43. T. H. Nguyen, R. Grishman, "Relation Extraction: Perspective from Convolutional Neural Networks" in *Proceedings of the 1st Workshop on Vector Space Modeling for Natural Language Processing* (Association for Computational Linguistics, Stroudsburg, PA, USA, 2015), pp. 39–48.

44. L. Zhuang, L. Wayne, S. Ya, Z. Jun, "A Robustly Optimized BERT Pre-training Approach with Post-training" in *Proceedings of the 20th Chinese National Conference on Computational Linguistics* (Chinese Information Processing Society of China, Huhhot, China, 2021; https://aclanthology.org/2021.ccl-1.108), pp. 1218–1227.

45. J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv:1810.04805* (2018).

46. D. Q. Nguyen, T. Vu, A. T. Nguyen, BERTweet: A pre-trained language model for English Tweets. (2020).

47. J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So, J. Kang, BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics* **36**, 1234–1240 (2020).

48. C. Liang, Y. Yu, H. Jiang, S. Er, R. Wang, T. Zhao, C. Zhang, "BOND: BERT-Assisted Open-Domain Named Entity Recognition with Distant Supervision" in *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* (ACM, New York, NY, USA, 2020), pp. 1054–1064.

49. D. Wadden, U. Wennberg, Y. Luan, H. Hajishirzi, Entity, Relation, and Event Extraction with Contextualized Span Representations. (2019).

50. Z. Zhang, X. Han, Z. Liu, X. Jiang, M. Sun, Q. Liu, "ERNIE: Enhanced Language Representation with Informative Entities" in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (Association for Computational Linguistics, Stroudsburg, PA, USA, 2019), pp. 1441–1451.

51. H. Chang, H. Xu, J. van Genabith, D. Xiong, H. Zan, JoinER-BART: Joint Entity and Relation Extraction with Constrained Decoding, Representation Reuse and Fusion. *IEEE/ACM Trans Audio Speech Lang Process*, 1–14 (2023).

52. I. Yamada, A. Asai, H. Shindo, H. Takeda, Y. Matsumoto, "LUKE: Deep Contextualized Entity Representations with Entity-aware Self-attention" in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (Association for Computational Linguistics, Stroudsburg, PA, USA, 2020), pp. 6442–6454.

53. I. Beltagy, K. Lo, A. Cohan, "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)* (Association for Computational Linguistics, Stroudsburg, PA, USA, 2019), pp. 3613–3618.

54. M. Eberts, A. Ulges, Span-based Joint Entity and Relation Extraction with Transformer Pre-training. doi: 10.3233/FAIA200321 (2019).

55. A. Radford, K. Narasimhan, Improving Language Understanding by Generative Pre-Training. (2018).

56. T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, D. Amodei, "Language Models Are Few-Shot Learners" in *Proceedings of the 34th International Conference on Neural Information Processing Systems* (Curran Associates Inc., Red Hook, NY, USA, 2020)*NIPS'20*.

57. A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, "Language Models are Unsupervised Multitask Learners" (2019; https://api.semanticscholar.org/CorpusID:160025533).

58. LitCoin Natural Language Processing (NLP) Challenge, *National Center for Advancing Translational Sciences* (2022). https://ncats.nih.gov/funding/challenges/litcoin.

59. O. J. Wouters, M. McKee, J. Luyten, Estimated Research and Development Investment Needed to Bring a New Medicine to Market, 2009-2018. *JAMA* **323**, 844 (2020).

60. F. Lovering, J. Bikker, C. Humblet, Escape from Flatland: Increasing Saturation as an Approach to Improving Clinical Success. *J Med Chem* **52**, 6752–6756 (2009).

61. S. K. Mohamed, A. Nounu, V. Nováček, Biological applications of knowledge graph embedding models. *Brief Bioinform* **22**, 1679–1693 (2021).

62. C. Wang, H. Yu, F. Wan, "Information Retrieval Technology Based on Knowledge Graph" in *Proceedings of the 2018 3rd International Conference on Advances in Materials, Mechatronics and Civil Engineering (ICAMMCE 2018)* (Atlantis Press, Paris, France, 2018).

63. L. Cui, H. Seo, M. Tabar, F. Ma, S. Wang, D. Lee, "DETERRENT: Knowledge Guided Graph Attention Network for Detecting Healthcare Misinformation" in *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* (ACM, New York, NY, USA, 2020), pp. 492–502.

64. F. Azuaje, Drug interaction networks: an introduction to translational and clinical applications. *Cardiovasc Res* **97**, 631–41 (2013).

65. H. Ye, Q. Liu, J. Wei, Construction of drug network based on side effects and its application for drug repositioning. *PLoS One* **9**, e87864 (2014).

66. H. Chen, H. Zhang, Z. Zhang, Y. Cao, W. Tang, Network-Based Inference Methods for Drug Repositioning. *Comput Math Methods Med* **2015**, 1–7 (2015).

67. Y. Luo, X. Zhao, J. Zhou, J. Yang, Y. Zhang, W. Kuang, J. Peng, L. Chen, J. Zeng, A network integration approach for drug-target interaction prediction and computational drug repositioning from heterogeneous information. *Nat Commun* **8**, 573 (2017).

68. D. S. Himmelstein, A. Lizee, C. Hessler, L. Brueggeman, S. L. Chen, D. Hadley, A. Green, P. Khankhanian, S. E. Baranzini, Systematic integration of biomedical knowledge prioritizes drugs for repurposing. *Elife* **6** (2017).

69. L. Luo, P.-T. Lai, C.-H. Wei, C. N. Arighi, Z. Lu, BioRED: a rich biomedical relation extraction dataset. *Brief Bioinform* **23** (2022).

70. openFDA. [Preprint] (2023).

71. F. Ahmed, A. M. Soomro, A. R. Chethikkattuveli Salih, A. Samantasinghar, A. Asif, I. S. Kang, K. H. Choi, A comprehensive review of artificial intelligence and network based approaches to drug repurposing in Covid-19. *Biomedicine & Pharmacotherapy* **153**, 113350 (2022).

72. Y. Zhou, Y. Hou, J. Shen, Y. Huang, W. Martin, F. Cheng, Network-based drug repurposing for novel coronavirus 2019-nCoV/SARS-CoV-2. *Cell Discov* **6**, 14 (2020).

73. R. Aghdam, M. Habibi, G. Taheri, Using informative features in machine learning based method for COVID-19 drug repurposing. *J Cheminform* **13**, 70 (2021).

74. F. Ahmed, J. W. Lee, A. Samantasinghar, Y. S. Kim, K. H. Kim, I. S. Kang, F. H. Memon, J. H. Lim, K. H. Choi, SperoPredictor: An Integrated Machine Learning and Molecular Docking-Based Drug Repurposing Framework With Use Case of COVID-19. *Front Public Health* **10** (2022).