# A comprehensive large scale biomedical knowledge graph for AI powered data driven biomedical research

Yuan Zhang[1,2,┼], Xin Sui[1,┼], 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[2,6], 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 Zhang
**Email:** jinfeng@insilicom.com

[┼] These authors contributed equally.

**Abstract**

To address the rapid growth of scientific publications and data in biomedical research, knowledge graphs (KGs) have become a critical tool for integrating large volumes of heterogeneous data to enable efficient information retrieval and automated knowledge discovery (AKD). However, transforming unstructured scientific literature into KGs remains 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 (2022) to construct a large-scale KG named iKraph using all PubMed abstracts. The extracted information matches human expert annotations and significantly exceeds the content of 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. This KG facilitates 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 applied it to real-time COVID-19 drug repurposing from March 2020 to May 2023. Our method identified 600-1400 candidate drugs per month, with one-third of those discovered in the first two months later supported by clinical trials or PubMed publications. These outcomes are very challenging to attain through alternative approaches that lack a thorough understanding of the existing literature. A cloud-based platform (https://biokde.insilicom.com) was developed for academic users to 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 innovative 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 innovative 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 a structured form, facilitating efficient and accurate information retrieval. Graph algorithms can be employed on KGs to infer potential 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[21,24]. 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[43–54]. 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 perform comparably or better to humans in multiple tasks, marking a significant breakthrough in the field.

Recently, LLMs like GPT-4 have been explored for their integration with KGs, aiming to enhance tasks such as named entity recognition (NER), relation extraction, and event detection through techniques like zero-shot prompting, in-context learning, and multi-turn question answering[58–60]. While these models excel in generalization and large-scale data processing, they still struggle with domain-specific challenges, including handling long-tail entities[61], directional entailments[62], and inconsistencies in retrieving knowledge from paraphrased or low-frequency phenomena[63,64]. Experiments[60] on datasets like DuIE2.0[65], Re-TACRED[66], and SciERC[67] highlight that fine-tuned small models continue to outperform GPT-like LLMs in KG-related tasks. Despite these limitations, LLMs have shown significant adaptability in augmenting KGs, particularly when structured data is limited, positioning them as a complementary tool for AKD[58].

To facilitate the methodology development and identification of the most effective methods for KG construction, the National Center for Advancing Translational Sciences (NCATS) of the National Institutes of Health (NIH) 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 (https://ncats.nih.gov/funding/challenges/winners/litcoin-nlp).

In this study, we applied our LitCoin NLP Challenge-winning information retrieval pipeline to all PubMed abstracts (cutoff date: May 2023) to construct a large-scale Biomedical Knowledge Graph (iKraph, the abbreviation for Insilicom's Knowledge Graph). Manual verification confirmed the pipeline's accuracy at a human annotator level. By annotating the directions for the relations in the LitCoin dataset and training a model to predict the direction of relations, we constructed a causal knowledge graph (CKG) capable of making indirect causal inferences. To further enhance the coverage of iKraph, 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[68]. The ascent in costs largely stems from prior research exhausting more straightforward drug targets, necessitating a shift towards more complex ones[69]. In this setting, knowledge graphs play a pivotal role in automated knowledge discovery (AKD)[2,70–72], particularly in the domain of drug target identification and drug repurposing[73–77]. 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 a thorough search of the 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. This enabled us to measure recall and observed positive rate (OPR), which is infeasible in prior drug repurposing research.

We demonstrate the power of our approach by conducting several drug repurposing studies: drug repurposing for COVID-19, cystic fibrosis, ten diseases without satisfactory treatment, and ten commonly prescribed drugs. For COVID-19 and cystic fibrosis, we performed retrospective, real-time drug repurposing exercises. Our method identified numerous viable candidates, supported by substantial literature evidence connecting the drug and disease entities. This level of interpretability is invaluable when determining the necessity of subsequent research endeavors.

## Results

### Building a Large-Scale Biomedical Knowledge Graph (iKraph)

To facilitate the methodology development and identification of the most effective methods for KG construction, NIH organized the LitCoin natural language processing (NLP) challenge between Nov 2021 and Feb 2022 (https://ncats.nih.gov/funding/challenges/litcoin). Our team, JZhangLab@FSU, participated in the challenge and won first place (Table 1). In the summer of 2023, we also participated in the BioRED track of the BioCreative Challenge VIII. In the end-to-end KG construction task, our team also achieved the highest score (Table 1)[78]. The LitCoin NLP Challenge dataset comprises 500 PubMed abstracts, each annotated with six distinct entity types and eight types of relations at the abstract level. We used the pipeline initially developed for the LitCoin challenge to process all PubMed abstracts available before May 2023, creating a large-scale Knowledge Graph, iKraph. In constructing iKraph, we processed over 34 million PubMed abstracts, resulting in 10,686,927 unique entities and 30,758,640 unique relations. We incorporated entity normalization into our pipeline, as this was not a component of the LitCoin NLP challenge (see Supplementary Materials Section 1.2, 1.3 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, including 1583 entity pairs. The results shown in Supplementary Table 3 indicate that our information extraction performance rivals that of human annotations. A more in-depth analysis is available in the Supplementary Materials Section 2.

Fig. 1A shows the number 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 articles focus exclusively on diseases. In contrast, gene mentions often coexist with other entities, such as chemicals and diseases. Fig. 1B depicts the number of PubMed abstracts containing one or more of the five major types of relations, offering insight into the distribution of topics in biomedical research.

Fig. 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 relations in each abstract, especially those involving sequence variant entities[79]. This explains the performance difference of our pipeline on these two datasets. Relations from PubMed and public databases are also quite complimentary.

Fig. 1D shows the number of novel discoveries for different entity pairs over the year. We have observed 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 iKraph, and $p(k)$ is the probability of an entity having degree $k$ (Fig. 1E). We found that iKraph exhibits a scale-free topology with an alpha parameter value of around 3.0 (more details in Supplementary Materials Section 3).

Supplementary Table 4 compares the numbers of relations for five types of entity pairs from all the public databases integrated into iKraph, 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 the one hand, iKraph has significantly more numbers of relations than those from public databases. 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.

**Constructing a causal knowledge graph**

We developed a model to predict the direction of correlation relations in the LitCoin dataset, identifying whether the relation flows from entity1 to entity2 or entity2 to entity1. Adding this directional information transformed correlations into potential causal relationships, allowing us to construct a directed knowledge graph for knowledge discovery applications.

**PSR for inferring indirect relationships**

With directional information, we can infer relations between indirectly connected entities using straightforward reasoning. To this end, we designed the probabilistic semantic reasoning (PSR) algorithm, which is both efficient and interpretable. PSR enables all-against-all drug repurposing for all drugs and diseases with limited computational resources and allows efficient updates of newly inferred relations. For instance, freshly published PubMed articles can be processed daily to extract discoveries and generate hypotheses for timely dissemination. In contrast, most machine learning methods struggle to achieve this level of efficiency and interpretability.

**Drug repurposing for Covid-19 using iKraph.**

Using the PSR algorithm, we conducted a retrospective, real-time drug repurposing study for COVID-19 spanning from March 2020 to May 2023 (Fig. 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 has at least one directed path to COVID-19 through an intermediate gene. We checked whether any repurposed drugs were later validated by either PubMed literature or clinical trials through a monthly assessment. The validation 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. Notably, 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 iKraph 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.

Fig. 2B shows a timeline of repurposed drug validation. Notably, there is a surge in validated drugs during the first year, which subsequently shows a month-to-month decline. This trend suggests most repurposed drugs align with practitioners' early assessments. Some drugs were validated only in the second or third year, indicating they were less immediately evident. The number of drugs validated through publications matches those validated via clinical trials. While numerous drug repurposing studies for COVID-19 exist[80–83], 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 results highlight iKraph's ability to identify promising drug candidates for specific diseases in real-time.

We then conducted drug repurposing for COVID-19 in the current timeframe (Fig. 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. Fig. 2C displays the top 50 repurposed drugs. Notably, most of these drugs (36 out of 50) have published studies mentioning either their potential therapeutic efficacy or demonstrated therapeutic efficacy for phenotypes associated with COVID-19. Among the remaining 14, 11 have been proposed as potential treatments for COVID-19 (citations provided in Supplementary Table 3). For each drug, numerous genes that link COVID-19 with the drug were identified (y-axis of Fig. 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 cystic fibrosis using iKraph.**

We applied PSR to uncover indirect relationships between drugs and cystic fibrosis (CF) from 1985 to 2022 (Fig. 3). Since the early 1990s, at least 50 potential repurposed drugs have been identified annually. A drug was considered validated if later reported as directly therapeutic for CF. Historically, estimating these metrics was challenging due to reliance on manual literature searches. We calculated recall (percentage of known direct relations successfully repurposed) and observed positive rate (OPR, percentage of repurposed cases with reported direct relations). Unlike precision, OPR accounts for potential candidates awaiting validation. From 1990 to 2022, the average recall is 0.635 (Fig. 3B), and the average OPR from 1985 to 2011 is 0.159. Different time intervals were used because OPR requires earlier predictions, while recent predictions need time for validation.

We calculated the typical duration for these repurposed drugs to be validated. Remarkably, our proposed drugs typically appeared in literature 2 to 33 years later, with a median validation time of 9.4 years (Fig. 3A). Assuming experimental validation takes 2 years on average, iKraph could hypothetically reduce this time from 9 to 2 years if predictions were acted on immediately. With over 63% recall and a 9-year median lag, our findings highlight iKraph's potential to accelerate drug repurposing and validation for cystic fibrosis treatment.

**Drug repurposing for 10 diseases and 10 drugs**

To evaluate our method's versatility, we extended drug repurposing to ten diseases lacking satisfactory treatments and ten commonly prescribed drugs (Supplementary 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). Impressively, our findings revealed average recall values of 0.76 for disease repurposing and 0.86 for drug repurposing. This exceptional recall rate emphasizes the potency of iKraph 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.

We extended our analysis by using relations from a database, alongside those extracted from PubMed abstracts, to make drug repurposing predictions. Fig. 4 illustrates the F1 scores for these comparisons based on the top 50 predicted repurposed drugs and the top 250 predicted indications. In each panel, blue bars represent PubMed-based predictions, while orange bars represent database-based predictions. Most repurposed drugs or diseases showed higher F1 scores using PubMed-derived predictions, likely due to the greater amount of information available in PubMed, which databases cannot match.

## Discussion

Converting unstructured scientific literature into structured data has been a long-standing challenge in natural language processing (NLP). Successfully addressing this issue can potentially revolutionize the pace of scientific discoveries. Although numerous studies have been conducted over the years, computational methods have yet to achieve the precision of manual annotation in relation to 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 iKraph, stands out as perhaps the most all-encompassing biomedical knowledge graph constructed so far. The coverage of iKraph is much larger than public databases for the relations we have extracted. The construction of a causal knowledge graph and the design of an interpretable PSR algorithm allows us to perform automated knowledge discovery very effectively. The exhaustive nature of iKraph 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 a structured form, one must perform a 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 iKraph-enabled capabilities in Supplementary Material Box S1, 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 iKraph for knowledge discovery tasks, such as drug repurposing, 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 iKraph's proficiency in identifying treatments for pandemics, marking it as an indispensable asset for potential future outbreaks.

Many users might inquire about how iKraph handles noisy information from low-quality papers. Our approach involves aggregating the probabilities of relations (e.g., between A and B) across multiple papers. Each paper assigns a probability to a specific relation, and these probabilities are combined to form an overall score. The more papers that mention the relation, the higher the final probability, making it less susceptible to noise. Relations with low-quality evidence tend to appear in fewer papers, resulting in lower scores. However, while this method provides a strong foundation for handling noisy data, future improvements could involve weighting papers based on factors like journal impact factor, citation count, and publication date. Integrating such metrics aligns with approaches demonstrated in prior work, where features like author diversity, institutional independence, and publication density were found to predict the robustness and reproducibility of scientific claims[84]. Integrating these metrics would allow us to refine the score further by giving more weight to higher-quality sources. For example, papers published in high-

impact journals or those widely cited in the scientific community may provide stronger evidence for a relation than those from less reputable sources. Additionally, the publication date can be factored in to balance the relevance of older versus newer findings, ensuring that the most current and impactful research plays a more prominent role in shaping the final probability. This reflects insights from robust scientific literature, where combining high-throughput experimental data with features predictive of reliability has shown promise in assessing the reproducibility of claims[84]. This holistic approach would help iKraph remain robust against misinformation while continuously improving the accuracy of its predictions through adaptive weighting.

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 iKraph with LLMs can effectively mitigate these limitations. To this end, we are actively developing a comprehensive question-answering system, combining iKraph with an open-source LLM.

In the Supplementary Materials Section 6, we delve into future research avenues and the challenges we've faced. In summary, iKraph serves as a powerful enabler for more effective and efficient information retrieval and automated knowledge discovery.

## Methods

### 1. Information extraction pipeline

We utilized the pipeline crafted during the LitCoin NLP Challenge (https://ncats.nih.gov/funding/challenges/litcoin) to process all PubMed abstracts available until 2022, along with data from several renowned biomedical databases, leading to the creation of the Knowledge Graph, iKraph. The construction of iKraph involves three primary stages: named entity recognition (NER), relation extraction (RE), and novelty classification. The details of the methods can be found in the Method section in the Support Information. When developing the pipeline for LitCoin Challenge, we tested a large set of pre-trained language models including BERT[46], BioBERT[48], PubMedBERT[85] abstract only, PubMedBERT fulltext, sentence BERT[86], RoBERTa[87], T5[88], BlueBERT[89], SciBERT[54], and ClinicalBERT[90]. We tested many ideas, such as different loss functions, data augmentations, different settings of label smoothing, different ways of ensemble learning, etc. Our final pipelines contain the following components: (1) Improved in-house script for data processing, including sentence split, tokenization, and entity tagging; (2) RoBERTa large and PubMedBERT models as baseline models for NER task; (3) Ensemble modeling strategy that combines models trained with different parameter settings, different random seeds and at different epochs for both NER and RE; (4) Label smoothing for both NER and RE; (5) Using Ab3P[91] for handling abbreviations for NER; (6) Modified classification rule tailored for LitCoin scoring method; and (7) Training a multi-sentence model for predicting relations at document level, which gave a very competitive baseline for relation extraction.

We used the pipeline developed in the LitCoin challenge to process all the abstracts in the PubMed database, which contains over 34 million abstracts, resulting in 10,686,927 unique entities and 30,758,640 unique relations.

## 2. Constructing a causal knowledge graph

To infer causal relations, we first annotated causal direction for 4,572 relations in the LitCoin dataset. Among them, 2,009 cases have direction from the first entity to the second; 1,611 cases have direction from the second entity to the first; and 952 cases have no direction. This annotation allowed us to train a model for predicting the directions for relations, which achieved an F1 score of 0.924 in a 5-fold cross-validation test on the LitCoin dataset. Using a causal knowledge graph, we can infer indirect causal relations more effectively for entities not directly connected in the knowledge graph – an essential task in automated knowledge discovery.

To make inferences using the causal knowledge graph, we designed a probabilistic semantic reasoning (PSR) algorithm, which calculates the probability of a true relation between two entities connected directly or indirectly. For two entities with a direct edge (a relation mentioned in the literature), there can be multiple mentions in different articles. It is necessary to estimate an overall probability for this pair, which will be used for estimating probabilities for indirectly connected entity pairs. PSR is highly interpretable, which is key for the validation of predictions.

The overall drug repurposing strategy and validation approach are depicted in Fig. 5, with some details provided in the figure legend.

### 2.1 Probabilistic Semantic Reasoning (PSR)

To make inferences using the causal knowledge graph, we designed a probabilistic framework, probabilistic semantic reasoning (PSR), for inferring indirect causal relations. PSR is highly interpretable, which is critical for the validation of predictions. There can be multiple mentions in different articles of two entities with a direct edge (a relation mentioned in the literature). It is necessary to estimate an overall probability for this pair, which will be used for estimating probabilities for indirectly connected entity pairs.

To simplify the discussion, let's assume we want to infer the indirect relation from A to C using direct relations from A to B and the relation from B to C. To infer the indirect relation, we first extract the two direct relations. As mentioned earlier, relation A to B and B to C will likely occur many times in different PubMed abstracts. We calculate the overall probability of whether two entities have a particular relation using the formula:

$$P_{A\to B} = 1 - \prod_{i=1}^{N}(1 - p_{A\to B}^{i}) \quad (1)$$

In Equation (1), $P_{A\to B}$ is the overall probability of A-B entity pair having a particular relation, $p_{A\to B}^{i}$ is the probability of being true for the i-th occurrence of these two entities in a PubMed abstract, $1 - p_{A\to B}^{i}$ is the probability of this occurrence being false, and $\prod_{i=1}^{N}(1 - p_{A\to B}^{i})$is the probability that all the occurrences being false (assuming the predictions for these occurrences are independent). The probability of all occurrences being false, when subtracted by 1, gives the probability that at least one of them is true, which is the desired probability. It is also possible that several different relation types will be inferred for a single pair of entities. Often, only one relation type is true, and others may be wrong predictions. To simplify the inference, we selected the relation type with the highest probability as the true relation type for any pair of entities. In reality, there can be multiple entities linking A to C. We denote one of them as $B_j$. Then, the probability of A to C through $B_j$ can be calculated as:

$$P_{A,B_j,C} = P_{A,B_j} \cdot P_{B_j,C} \quad (2)$$

Equation (2) is straightforward since for the indirect relation between A and C (direction from A to C) to be true, both the direct relations must be true. Again, we assume the predictions for the two direct relations are independent. The probability between A to C through m intermediate nodes can then be calculated as

$$P_{A,\cdot,C} = 1 - \prod_{i=1}^{m}(1 - P_{A,B_i,C}) \quad (3)$$

In Equation (3), we use $P_{A,\cdot,C}$ to denote the probability that the indirect relation between A and C through any intermediate entity and there is m such intermediate entities that link A and C. The argument for this formula is similar to equation (1). Putting equations 1-3 together, we get

$$P_{A,:,C} = 1 - \prod_{i=1}^{m}[1 - [1 - \prod_{j=1}^{n}(1 - p_{A,B_i}^{j})] \cdot [1 - \prod_{k=1}^{l}(1 - p_{B_i,C}^{k})]] \quad (4)$$

In Equation (4), *m* entities are l A and C, *n* instances of A-$B_i$ relations in literature, and l instances of $B_i$-C relations in literature. Extending this to multiple intermediate nodes between A and C is relatively straightforward. The above probabilistic framework will allow us to rank all the indirect relations that can be inferred. To infer the relation type (positive correlated or negative correlated) between two entities, which multiple intermediate entities could link, we use 1 to represent positive correlations, -1 to represent negative correlations, and 0 to represent unknown correlation type between any two entities connected by a direct edge and multiply all the correlations together. The resulting value, 1, -1, or 0 will give us the correlation type between the two entities. If there is at least one unknown correlation type (0) between the two entities, the overall correlation type is unknown. If there is no 0 and an even number of negative correlations, then the overall correlation type will be a positive correlation; otherwise, it is a negative one. For A-C entity pair to have a non-zero probability, there must be a path from A to C with all the directions going from A to C, such as A->B->D->C, while A->B<-D->C is not a valid path from A to C.

In this manuscript, we show one application of PSR using our iKraph that calculates the indirect relationship between two entities: discover the repurposed drug. We present two study cases for identifying repurposed drugs for COVID-19 and cystic fibrosis, along with an additional study that involves predicting both repurposed drug candidates for 10 common diseases and potential additional uses for 10 common drugs. The details can be found in the Section 1.7 in SI. Fig. 5E illustrates Genistein's repurposing for COVID-19 treatment, interacting with 25 human genes. It negatively affects 3 genes that have a positive impact on COVID-19 while positively influencing the remaining 22 genes, which are negatively associated with the disease. This evidence supports Genistein's potential as a COVID-19 treatment candidate.

### 3. Integrating relations from public databases

To integrate the relations in the public databases, we downloaded the relations from two databases that have integrated data from a large number of databases recently, Hetionet[73] and primeKG[92], where Hetionet has integrated data from 29 databases and primeKG has integrated data from 20 databases. The total number of unique databases from both sources is 38. In addition, we extracted drug-target relations from the Therapeutic Target Database (TTD)[93] and GO annotation[94,95]. In total, we integrated relation data from 40 public databases. The KG covers twelve common entity types: diseases, chemical compounds, species, genes/proteins, mutations, cell lines, anatomy, biological processes, cellular components, molecular function, pathway, and pharmacologic class. It covers fifty-three different relation types. Among them, eight were annotated in the LitCoin dataset: association, positive correlation, negative correlation, bind, cotreatment, comparison, drug interaction, and conversion. Other relation types came from public databases. When incorporating relations from public databases to maintain the quality of the resulting KG, we excluded relations generated by high-throughput experiments, which are well-known to have a high proportion of false positives, and those predicted by previous machine learning models.

### 4. Incorporating relations from analyzing RNASeq data

We downloaded more than 300,000 human RNASeq profiles from the recount3 database[96]. We performed two types of analysis: differential gene expression analysis (DGEA) and gene regulatory network inference (GRNI). DGEA gave 92,628 differentially expressed genes for 36 different diseases, which correspond to 92,628 disease-gene relations, either positive or negative correlation, depending on whether the genes are up or downregulated in the corresponding diseases. GRNI gave 101,392 gene regulatory relations overall. We added close to 200,000 additional relations by analyzing this RNASeq dataset.

## Data Availability

The datasets used in this study are available on the GitHub repository at https://github.com/myinsilicom/iKraph[97]. Due to size limitations, additional large datasets can be accessed via Zenodo at https://zenodo.org/records/14851275[98]. We used BioRED dataset to train our NER and RE models, and the BioRED dataset can be accessed through https://ftp.ncbi.nlm.nih.gov/pub/lu/BioRED/. The complete knowledge graph is hosted on the cloud-based platform: https://www.biokde.com. The downloadable version of the complete iKraph can be accessed via Zenodo at https://zenodo.org/records/14851275[98].

## Code Availability

The code and datasets generated during this study can be found via the GitHub repository at: https://github.com/myinsilicom/iKraph[97].


## Acknowledgments:

We thank the LitCoin NLP Challenge organizers for generating the valuable challenge dataset, which made the work possible.

This research was partially supported by the National Institutes of Health (NIH) under grant R21LM014277 (JZ), contract 75N91024C00007 (JZ), and contract 75N93024C00034 (JZ); by the National Science Foundation (NSF) under grants 2335357 (JZ) and 2403911 (JZ); and by the National Cancer Institute, National Institutes of Health, under Prime Contract No. 75N91019D00024, Task Order No. 75N91024F00030 (JZ). The content of this publication does not necessarily reflect the views or policies of the Department of Health and Human Services, nor does mention of trade names, commercial products, or organizations imply endorsement by the U.S. Government. The funders had no role in the study design, data collection and analysis, decision to publish, or preparation of the manuscript.


### Author Contributions Statement

Y.Z., X.S., F.P., and K.Y. contributed equally to this work. Y.Z., X.S., F.P., K.L., S.T., A.E., Q.H., W.W., Jianan W., and Jian W. collected data and developed models and pipelines. Y.Z., F.P., and J.Z. analyzed the data and developed methods. D.S., H.C., J.Zh., E.Z., B.L., T.Z., and J.Z. developed the iExplore platform interface. K.Y. and J.Z. conceptualized and designed the study. Y.Z., F.P., K.Y., and J.Z. wrote the manuscript. X.Q., T.Z., and P.Z. provided consultation and manuscript revision. J.Z. supervised the study and is the corresponding author.

### Competing Interests Statement

Jinfeng Zhang and Tingting Zhao are owners of Insilicom LLC. The remaining authors declare no competing financial or non-financial interests.

**Table 1**: The Performance of the top teams in the LitCoin NLP challenge and BioCreative challenge VIII (BC8) BioRED track subtask 2(End-to-End KG construction). The LitCoin NLP Challenge results include NER and RE

scores, reported as Jaccard scores, and the overall score. Our team, JZhangLab@FSU, achieved the highest overall score. The BioCreative VIII BioRED track Subtask 2 results present F1 scores for various evaluation metrics. Our team, Insilicom, ranked first. '+' indicates additional task-specific scores. Bold values indicate the highest-performing scores within each column in each competition.

| Competition | Team | NER | RE | Overall | +ID | +Entity pair | +Relation type | +Novelty |
|---|---|---|---|---|---|---|---|---|
| LitCoin | JZhangLab@FSU | 0.9177 | **0.6332** | **0.7186** | - | - | - | - |
| | UTHealth SBMI | **0.9309** | 0.5941 | 0.6951 | - | - | - | - |
| | UIUCBioNLP | 0.9068 | 0.5681 | 0.5681 | - | - | - | - |
| BC8 | 156 (Insilicom) | **0.8926** | - | - | 0.8407 | **0.5584** | **0.4303** | **0.3275** |
| | 129 | 0.7858 | - | - | 0.7635 | 0.4127 | 0.3103 | 0.2334 |
| | 127 | 0.7830 | - | - | 0.7598 | 0.3945 | 0.2976 | 0.2280 |
| | 118 | 0.7831 | - | - | 0.7561 | 0.3903 | 0.2859 | 0.2248 |
| | GPT-3.5 + PubTator3 | 0.8652 | - | - | **0.8565** | 0.3021 | 0.1081 | 0.0714 |
| | GPT-4+PubTator3 | 0.8652 | - | - | **0.8565** | 0.3449 | 0.1704 | 0.1277 |

## Figure Legends

**Fig. 1. A**. Venn diagram for the number of PubMed articles containing certain types of entities; B. Venn diagram for the number of PubMed articles containing certain types of relations; **C**. The composition of relations in iKraph, PubMed abstracts, public databases, and LitCoin dataset; **D**. The numbers of novel discoveries by entity pair type from 1980 to 2020. **E**. Degree distribution of iKraph, 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.

**Fig. 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 (36 out of 50). Among those that were not reported as treatments for COVID-19, 11 out of 14 were hypothesized as potential treatments.

**Fig. 3. Drug repurposing for cystic fibrosis**. **A.** The number of repurposed drugs from 1985 to 2022. The dark yellow bar shows the number of validated drugs. The blue line shows the time it takes to validate the targets for the corresponding years. **B.** The number of drugs reported in PubMed from 1985 to 2022. The dark yellow bar shows the drugs that have been predicted previously.

**Fig. 4. F1 scores for drug repurposing prediction for 10 diseases and 10 common drugs**. The calculation is based on the top 50 repurposed drugs and the top 250 repurposed indications. The blue bars represent predictions made using relations extracted from PubMed abstracts, and the orange bars represent predictions made using relations from the database. Panels A and C are validated using therapeutic relations reported in the database, while panels B and D are validated using therapeutic relations extracted from PubMed abstracts.

**Fig. 5. The overview of our drug repurposing strategy and validation approach. A.** Our method infers drug-disease therapeutic relations through identifying an intermediate entity, the drug target of the disease, with causal relations from the drug to a drug target, and from the drug target to the disease. **B.** Two scenarios correspond to a drug-disease therapeutic relation: drug activates a target, and the target represses the disease, or drug inhibits the target, and the target promotes the disease. **C.** To infer an indirect relation from A (i.e. a drug) to C (i.e. a disease), there can be many intermediate potential targets between them. For each relation formed by two entities, there could be many scientific articles mentioning it. The probabilistic semantic reasoning (PSR) algorithm aggregates all the information to make the inference, while identifying polypharmacological candidates. **D.** To validate a drug repurposing study, we use a time-sensitive approach. We select many cutoff time points (lightbulb with a question mark) and use the knowledge published before the cutoff time (indicated by an orange arrow) to generate predictions and use the knowledge published after the cutoff time (lightbulbs on the green arrow) to validate our predictions. The figure shows three sets of discoveries (blue, red, and yellow bulbs) using three different cutoff times. **E**. Genistein's repurposing for COVID-19 treatment. Connected via 25 human genes represented by yellow ovals. The diagram shows drug-gene and gene-disease correlations, with red lines indicating negative correlations and green lines positive ones. Each connecting line is supported by multiple articles from which the correlation probabilities are derived.

# References


1. Kitano, H. Nobel Turing Challenge: creating the engine for scientific discovery. *npj Systems Biology and Applications* **7**, 29 (2021).
2. Li, L. *et al.* Real-world data medical knowledge graph: construction and applications. *Artificial Intelligence in Medicine* **103**, 101817 (2020).
3. Yu, S. *et al.* BIOS: An Algorithmically Generated Biomedical Knowledge Graph. (2022).
4. Nicholson, D. N. & Greene, C. S. Constructing knowledge graphs and their biomedical applications. *Computational and Structural Biotechnology Journal* **18**, 1414–1428 (2020).
5. Gao, Z., Ding, P. & Xu, R. KG-Predict: A knowledge graph computational framework for drug repurposing. *Journal of Biomedical Informatics* **132**, 104133 (2022).
6. Li, N. *et al.* KGHC: a knowledge graph for hepatocellular carcinoma. *BMC Medical Informatics and Decision Making* **20**, 135 (2020).
7. Ernst, P., Siu, A. & Weikum, G. KnowLife: a versatile approach for constructing a large knowledge graph for biomedical sciences. *BMC Bioinformatics* **16**, 157 (2015).
8. Zheng, S. *et al.* PharmKG: a dedicated knowledge graph benchmark for bomedical data mining. *Briefings in Bioinformatics* **22**, bbaa344 (2021).
9. Petasis, G. *et al.* 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* 426–433 (Association for Computational Linguistics, 2001). doi:10.3115/1073012.1073067.
10. Kim, J.-H. & Woodland, P. C. A rule-based named entity recognition system for speech input. in *6th International Conference on Spoken Language Processing (ICSLP 2000)* vols 1, 528-531–0 (ISCA, 2000). doi:10.21437/ICSLP.2000-131.
11. Miyao, Y., Sagae, K., Sætre, R., Matsuzaki, T. & Tsujii, J. Evaluating contributions of natural language parsers to protein–protein interaction extraction. *Bioinformatics* **25**, 394–400 (2009).

12. Lee, J., Kim, S., Lee, S., Lee, K. & Kang, J. On the efficacy of per-relation basis performance evaluation for PPI extraction and a high-precision rule-based approach. *BMC Medical Informatics and Decision Making* **13**, S7 (2013).
13. Raja, K., Subramani, S. & Natarajan, J. PPInterFinder—a mining tool for extracting causal relations on human proteins from literature. *Database* **2013**, bas052 (2013).
14. Kim, J.-H., Kang, I.-H. & Choi, K.-S. Unsupervised named entity classification models and their ensembles. in *Proceedings of the 19th international conference on Computational linguistics* - 1–7 (Association for Computational Linguistics, 2002). doi:10.3115/1072228.1072316.
15. Li, L., Zhou, R. & Huang, D. Two-phase biomedical named entity recognition using CRFs. *Computational Biology and Chemistry* **33**, 334–338 (2009).
16. Tikk, D., Thomas, P., Palaga, P., Hakenberg, J. & Leser, U. A Comprehensive Benchmark of Kernel Methods to Extract Protein–Protein Interactions from Literature. *PLoS Computational Biology* **6**, e1000837 (2010).
17. Bui, Q.-C., Katrenko, S. & Sloot, P. M. A. A hybrid approach to extract protein–protein interactions. *Bioinformatics* **27**, 259–265 (2011).
18. Patra, R. & Saha, S. K. A Kernel-Based Approach for Biomedical Named Entity Recognition. *The Scientific World Journal* **2013**, 1–7 (2013).
19. Hong, L. *et al.* A novel machine learning framework for automated biomedical relation extraction from large-scale literature repositories. *Nature Machine Intelligence* **2**, 347–355 (2020).
20. Zhang, H.-T., Huang, M.-L. & Zhu, X.-Y. A Unified Active Learning Framework for Biomedical Relation Extraction. *Journal of Computer Science and Technology* **27**, 1302–1313 (2012).
21. Yu, K. *et al.* Automatic extraction of protein-protein interactions using grammatical relationship graph. *BMC Medical Informatics and Decision Making* **18**, 42 (2018).
22. Chowdhary, R., Zhang, J. & Liu, J. S. Bayesian inference of protein–protein interactions from biological literature. *Bioinformatics* **25**, 1536–1542 (2009).

23. Corbett, P. & Copestake, A. Cascaded classifiers for confidence-based chemical named entity recognition. *BMC Bioinformatics* **9**, S4 (2008).
24. Lung, P.-Y., He, Z., Zhao, T., Yu, D. & Zhang, J. Extracting chemical–protein interactions from literature using sentence structure analysis and feature engineering. *Database* **2019**, (2019).
25. Bell, L., Chowdhary, R., Liu, J. S., Niu, X. & Zhang, J. Integrated Bio-Entity Network: A System for Biological Knowledge Discovery. *PLoS ONE* **6**, e21474 (2011).
26. Kim, S., Yoon, J. & Yang, J. Kernel approaches for genic interaction extraction. *Bioinformatics* **24**, 118–126 (2008).
27. Bell, L., Zhang, J. & Niu, X. 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* 371–375 (ACM, 2011). doi:10.1145/2147805.2147853.
28. Florian, R., Ittycheriah, A., Jing, H. & Zhang, T. Named entity recognition through classifier combination. in *Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003* - 168–171 (Association for Computational Linguistics, 2003). doi:10.3115/1119176.1119201.
29. Leaman, R., Wei, C.-H. & Lu, Z. tmChem: a high performance approach for chemical named entity recognition and normalization. *J Cheminform* **7**, S3 (2015).
30. Qu, J. *et al*. Triage of documents containing protein interactions affected by mutations using an NLP based machine learning approach. *BMC Genomics* **21**, 773 (2020).
31. Nguyen, T. H. & Grishman, R. Relation Extraction: Perspective from Convolutional Neural Networks. in *Proceedings of the 1st Workshop on Vector Space Modeling for Natural Language Processing* 39–48 (Association for Computational Linguistics, 2015). doi:10.3115/v1/W15-1506.
32. He, D., Zhang, H., Hao, W., Zhang, R. & Cheng, K. A Customized Attention-Based Long Short-Term Memory Network for Distant Supervised Relation Extraction. *Neural Computation* **29**, 1964–1985 (2017).
33. Li, F., Zhang, M., Fu, G. & Ji, D. A neural joint model for entity and relation extraction from biomedical text. *BMC Bioinformatics* **18**, 198 (2017).

34. Crichton, G., Pyysalo, S., Chiu, B. & Korhonen, A. A neural network multi-task learning approach to biomedical named entity recognition. *BMC Bioinformatics* **18**, 368 (2017).
35. Luo, L. *et al.* An attention-based BiLSTM-CRF approach to document-level chemical named entity recognition. *Bioinformatics* **34**, 1381–1388 (2018).
36. Guo, Z., Zhang, Y. & Lu, W. Attention Guided Graph Convolutional Networks for Relation Extraction. in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* 241–251 (Association for Computational Linguistics, 2019). doi:10.18653/v1/P19-1024.
37. Gridach, M. Character-level neural network for biomedical named entity recognition. *Journal of Biomedical Informatics* **70**, 85–91 (2017).
38. Lim, S. & Kang, J. Chemical–gene relation extraction using recursive neural network. *Database* **2018**, (2018).
39. Gu, J., Sun, F., Qian, L. & Zhou, G. Chemical-induced disease relation extraction via convolutional neural network. *Database* **2017**, (2017).
40. Habibi, M., Weber, L., Neves, M., Wiegandt, D. L. & Leser, U. Deep learning with word embeddings improves biomedical named entity recognition. *Bioinformatics* **33**, i37–i48 (2017).
41. Liu, S. *et al.* Extracting chemical–protein relations using attention-based neural networks. *Database* **2018**, (2018).
42. Wu, H. & Huang, J. Joint Entity and Relation Extraction Network with Enhanced Explicit and Implicit Semantic Information. *Applied Sciences* **12**, 6231 (2022).
43. Akbik, A., Bergmann, T. & Vollgraf, R. Pooled Contextualized Embeddings for Named Entity Recognition. in *Proceedings of the 2019 Conference of the North* 724–728 (Association for Computational Linguistics, 2019). doi:10.18653/v1/N19-1078.
44. Eberts, M. & Ulges, A. Span-based Joint Entity and Relation Extraction with Transformer Pre-training. (2019) doi:10.3233/FAIA200321.

45. Zhuang, L., Wayne, L., Ya, S. & Jun, Z. A Robustly Optimized BERT Pre-training Approach with Post-training. in *Proceedings of the 20th Chinese National Conference on Computational Linguistics* 1218–1227 (Chinese Information Processing Society of China, 2021).
46. Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv:1810.04805* (2018).
47. Nguyen, D. Q., Vu, T. & Nguyen, A. T. BERTweet: A pre-trained language model for English Tweets. (2020).
48. Lee, J. *et al.* BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics* (2019) doi:10.1093/bioinformatics/btz682.
49. Liang, C. *et al.* 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* 1054–1064 (ACM, 2020). doi:10.1145/3394486.3403149.
50. Wadden, D., Wennberg, U., Luan, Y. & Hajishirzi, H. Entity, Relation, and Event Extraction with Contextualized Span Representations. Preprint at https://doi.org/10.48550/arXiv.1909.03546 (2019).
51. Zhang, Z. *et al.* ERNIE: Enhanced Language Representation with Informative Entities. in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* 1441–1451 (Association for Computational Linguistics, 2019). doi:10.18653/v1/P19-1139.
52. Chang, H., Xu, H., Genabith, J. van, Xiong, D. & Zan, H. JoinER-BART: Joint Entity and Relation Extraction with Constrained Decoding, Representation Reuse and Fusion. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 1–14 (2023) doi:10.1109/TASLP.2023.3310879.
53. Yamada, I., Asai, A., Shindo, H., Takeda, H. & Matsumoto, Y. LUKE: Deep Contextualized Entity Representations with Entity-aware Self-attention. in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)* 6442–6454 (Association for Computational Linguistics, 2020). doi:10.18653/v1/2020.emnlp-main.523.
54. Beltagy, I., Lo, K. & Cohan, A. 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)* 3613–3618 (Association for Computational Linguistics, 2019). doi:10.18653/v1/D19-1371.

55. Radford, A. *et al.* Language Models are Unsupervised Multitask Learners. in (2019).
56. Radford, A. & Narasimhan, K. Improving Language Understanding by Generative Pre-Training. (2018).
57. Brown, T. B. *et al.* Language Models Are Few-Shot Learners. in *Proceedings of the 34th International Conference on Neural Information Processing Systems* (Curran Associates Inc., 2020).
58. Wei, X. *et al.* ChatIE: Zero-Shot Information Extraction via Chatting with ChatGPT. Preprint at https://doi.org/10.48550/arXiv.2302.10205 (2024).
59. Pan, J. Z. *et al.* Large Language Models and Knowledge Graphs: Opportunities and Challenges. Preprint at https://doi.org/10.48550/arXiv.2308.06374 (2023).
60. Zhu, Y. *et al.* LLMs for Knowledge Graph Construction and Reasoning: Recent Capabilities and Future Opportunities. Preprint at https://doi.org/10.48550/arXiv.2305.13168 (2024).
61. Kandpal, N., Deng, H., Roberts, A., Wallace, E. & Raffel, C. Large Language Models Struggle to Learn Long-Tail Knowledge. Preprint at https://doi.org/10.48550/arXiv.2211.08411 (2023).
62. Li, T., Hosseini, M. J., Weber, S. & Steedman, M. Language Models Are Poor Learners of Directional Inference. in *Findings of the Association for Computational Linguistics: EMNLP 2022* 903–921 (Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, 2022). doi:10.18653/v1/2022.findings-emnlp.64.
63. Elazar, Y. *et al.* Measuring and Improving Consistency in Pretrained Language Models. *Transactions of the Association for Computational Linguistics* **9**, 1012–1031 (2021).
64. Heinzerling, B. & Inui, K. Language Models as Knowledge Bases: On Entity Representations, Storage Capacity, and Paraphrased Queries. in *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume* 1772–1791 (Association for Computational Linguistics, Online, 2021). doi:10.18653/v1/2021.eacl-main.153.

65. Zheng, Q., Guo, K. & Xu, L. A large-scale Chinese patent dataset for information extraction. *Systems Science & Control Engineering* **12**, 2365328 (2024).

66. Stoica, G., Platanios, E. A. & Póczos, B. Re-TACRED: Addressing Shortcomings of the TACRED Dataset. Preprint at https://doi.org/10.48550/arXiv.2104.08398 (2021).

67. Luan, Y., He, L., Ostendorf, M. & Hajishirzi, H. Multi-Task Identification of Entities, Relations, and Coreference for Scientific Knowledge Graph Construction. in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing* 3219–3232 (Association for Computational Linguistics, Brussels, Belgium, 2018). doi:10.18653/v1/D18-1360.

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

69. Lovering, F., Bikker, J. & Humblet, C. Escape from Flatland: Increasing Saturation as an Approach to Improving Clinical Success. *Journal of Medicinal Chemistry* **52**, 6752–6756 (2009).

70. Cui, L. *et al.* DETERRENT: Knowledge Guided Graph Attention Network for Detecting Healthcare Misinformation. in *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* 492–502 (ACM, 2020). doi:10.1145/3394486.3403092.

71. Mohamed, S. K., Nounu, A. & Nováček, V. Biological applications of knowledge graph embedding models. *Briefings in Bioinformatics* **22**, 1679–1693 (2021).

72. Wang, C., Yu, H. & Wan, F. 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, 2018). doi:10.2991/icammce-18.2018.65.

73. Himmelstein, D. S. *et al.* Systematic integration of biomedical knowledge prioritizes drugs for repurposing. *eLife* **6**, (2017).

74. Azuaje, F. Drug interaction networks: an introduction to translational and clinical applications. *Cardiovascular research* **97**, 631–41 (2013).

75. Ye, H., Liu, Q. & Wei, J. Construction of drug network based on side effects and its application for drug repositioning. *PloS one* **9**, e87864 (2014).

76. Chen, H., Zhang, H., Zhang, Z., Cao, Y. & Tang, W. Network-Based Inference Methods for Drug Repositioning. *Computational and Mathematical Methods in Medicine* **2015**, 1–7 (2015).

77. Luo, Y. *et al.* A network integration approach for drug-target interaction prediction and computational drug repositioning from heterogeneous information. *Nature Communications* **8**, 573 (2017).

78. Islamaj, R., Lai, P.-T., Wei, C.-H., Luo, L. & Lu, Z. The overview of the BioRED (Biomedical Relation Extraction Dataset) track at BioCreative VIII. (2023) doi:10.5281/ZENODO.10351131.

79. Luo, L., Lai, P.-T., Wei, C.-H., Arighi, C. N. & Lu, Z. BioRED: a rich biomedical relation extraction dataset. *Briefings in Bioinformatics* **23**, (2022).

80. Ahmed, F. *et al.* SperoPredictor: An Integrated Machine Learning and Molecular Docking-Based Drug Repurposing Framework With Use Case of COVID-19. *Frontiers in Public Health* **10**, (2022).

81. Ahmed, F. *et al.* A comprehensive review of artificial intelligence and network based approaches to drug repurposing in Covid-19. *Biomedicine & Pharmacotherapy* **153**, 113350 (2022).

82. Zhou, Y. *et al.* Network-based drug repurposing for novel coronavirus 2019-nCoV/SARS-CoV-2. *Cell Discovery* **6**, 14 (2020).

83. Aghdam, R., Habibi, M. & Taheri, G. Using informative features in machine learning based method for COVID-19 drug repurposing. *Journal of Cheminformatics* **13**, 70 (2021).

84. Belikov, A. V., Rzhetsky, A. & Evans, J. Prediction of robust scientific facts from literature. *Nat Mach Intell* **4**, 445–454 (2022).

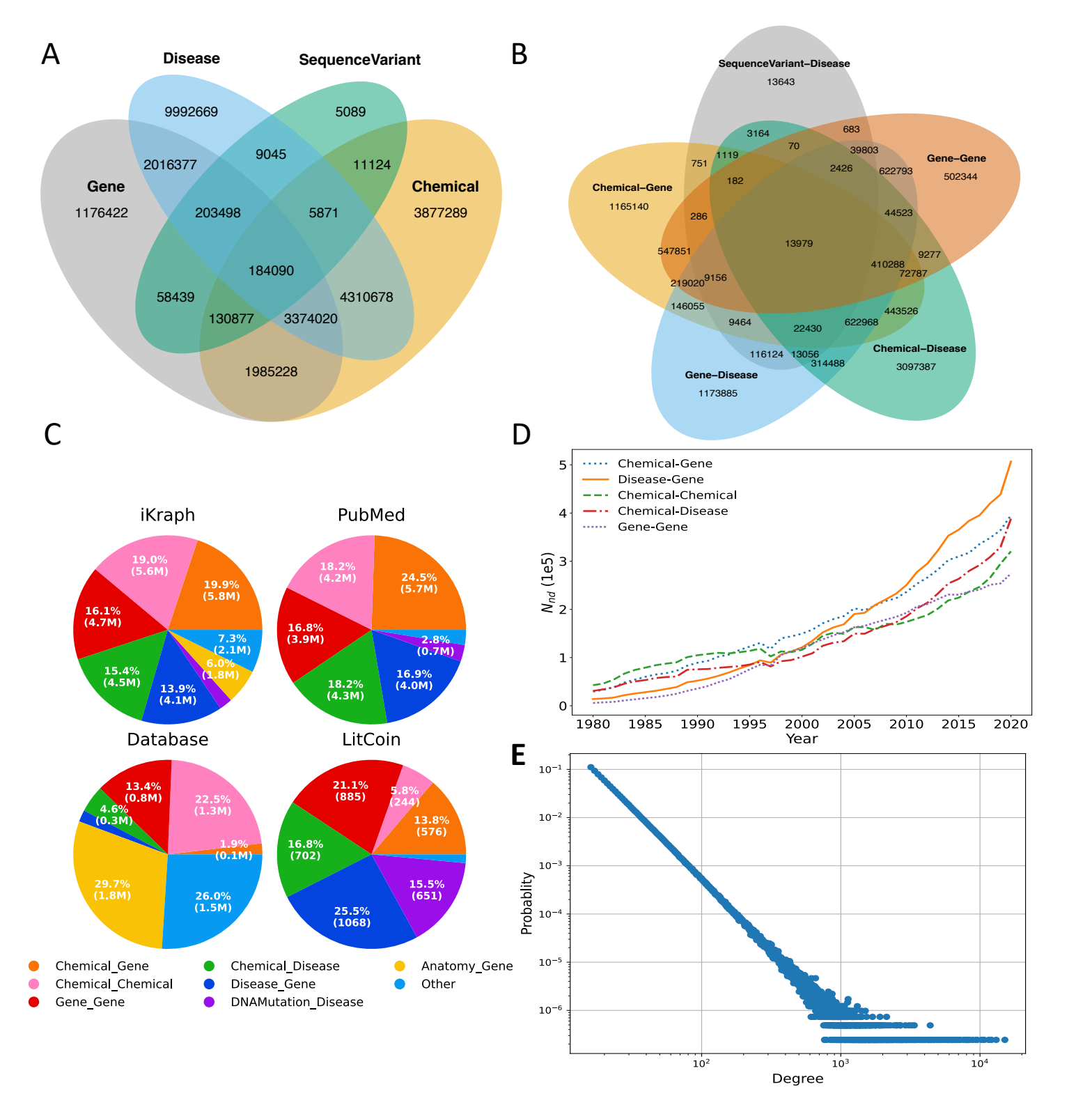
85. Gu, Y. *et al.* Domain-Specific Language Model Pretraining for Biomedical Natural Language Processing. *ACM Transactions on Computing for Healthcare* **3**, 1–23 (2022).

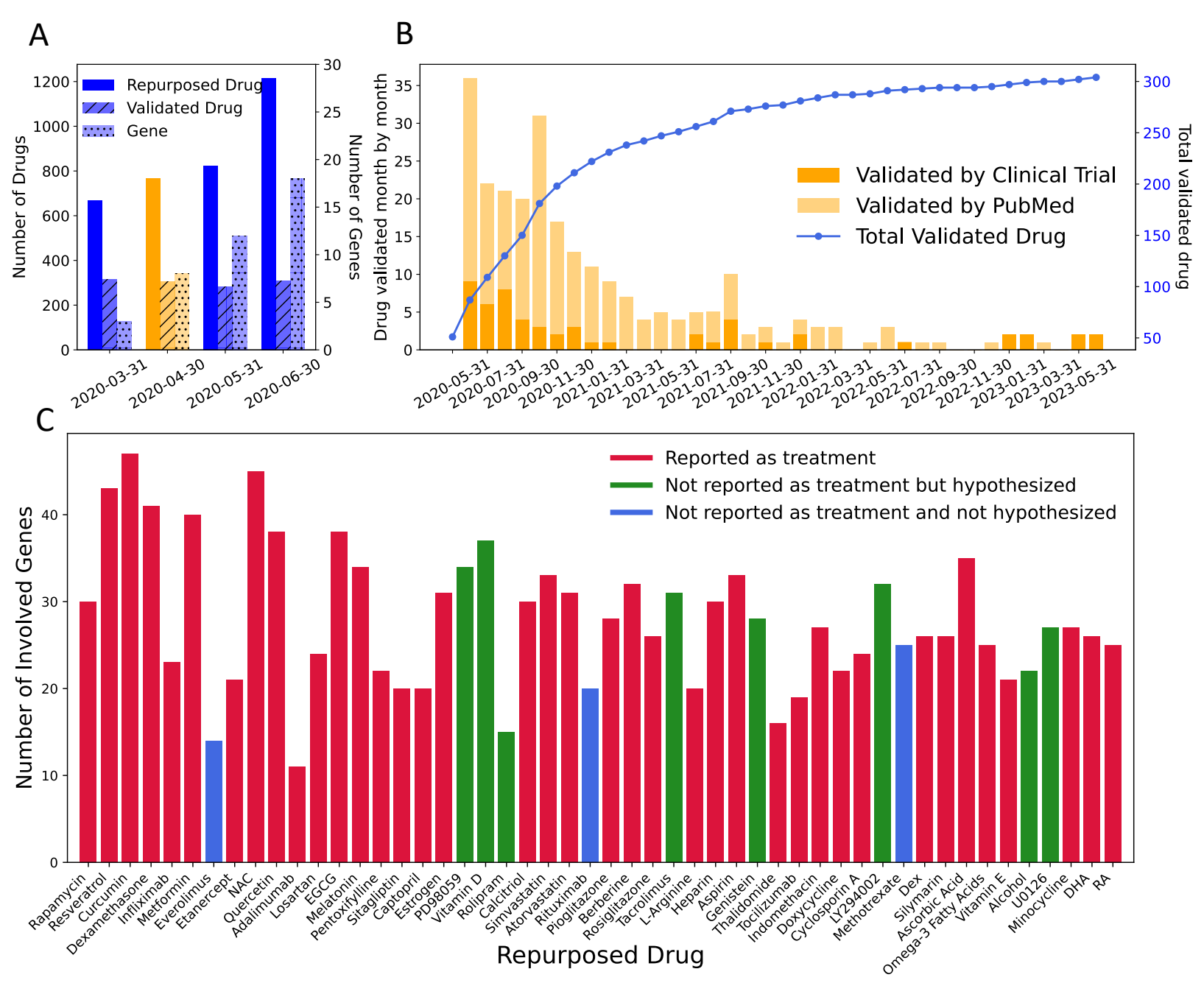
86. Reimers, N. & Gurevych, I. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. (2019).

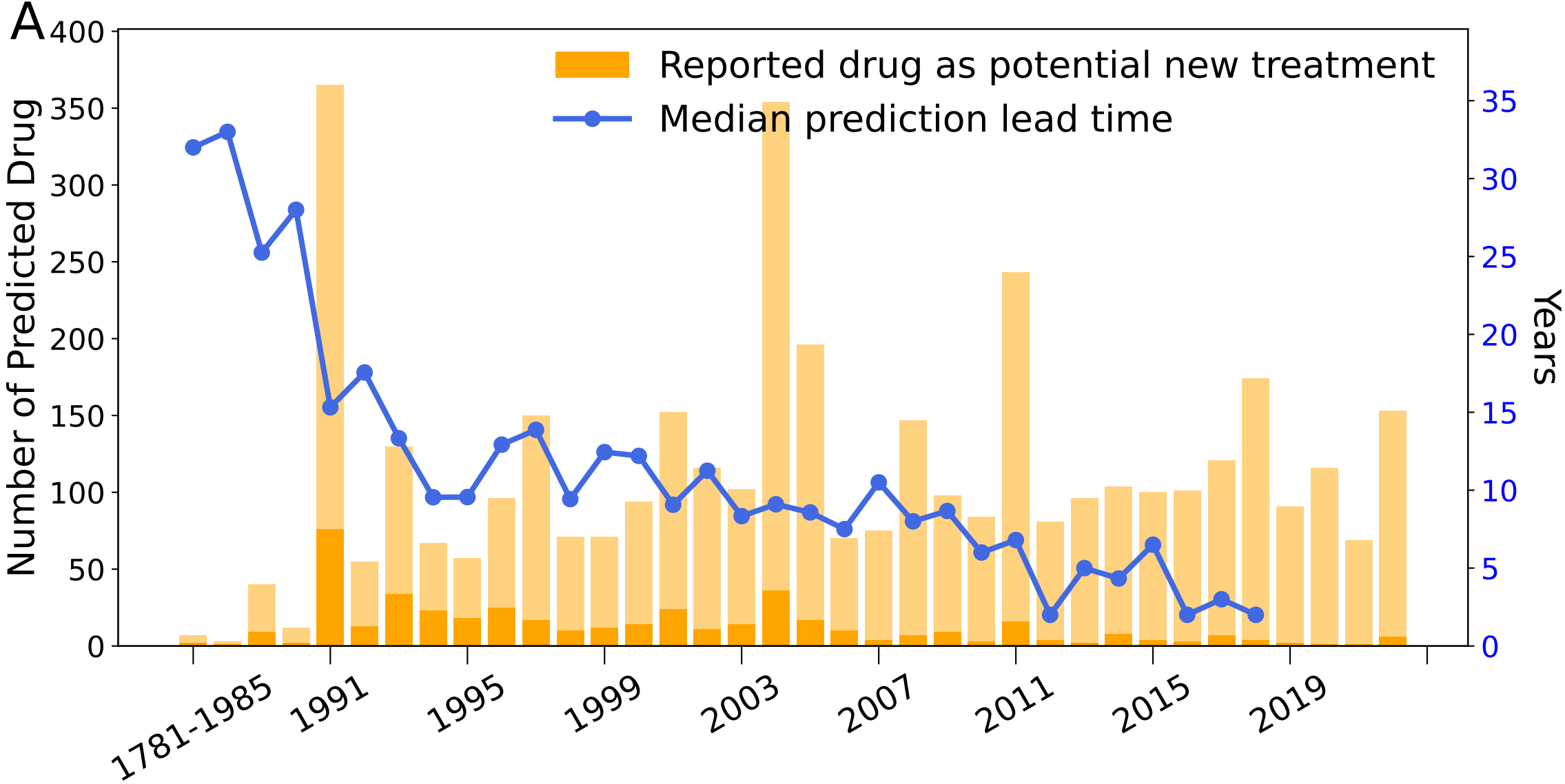
87. Liu, Y. *et al.* RoBERTa: A Robustly Optimized BERT Pretraining Approach. (2019).

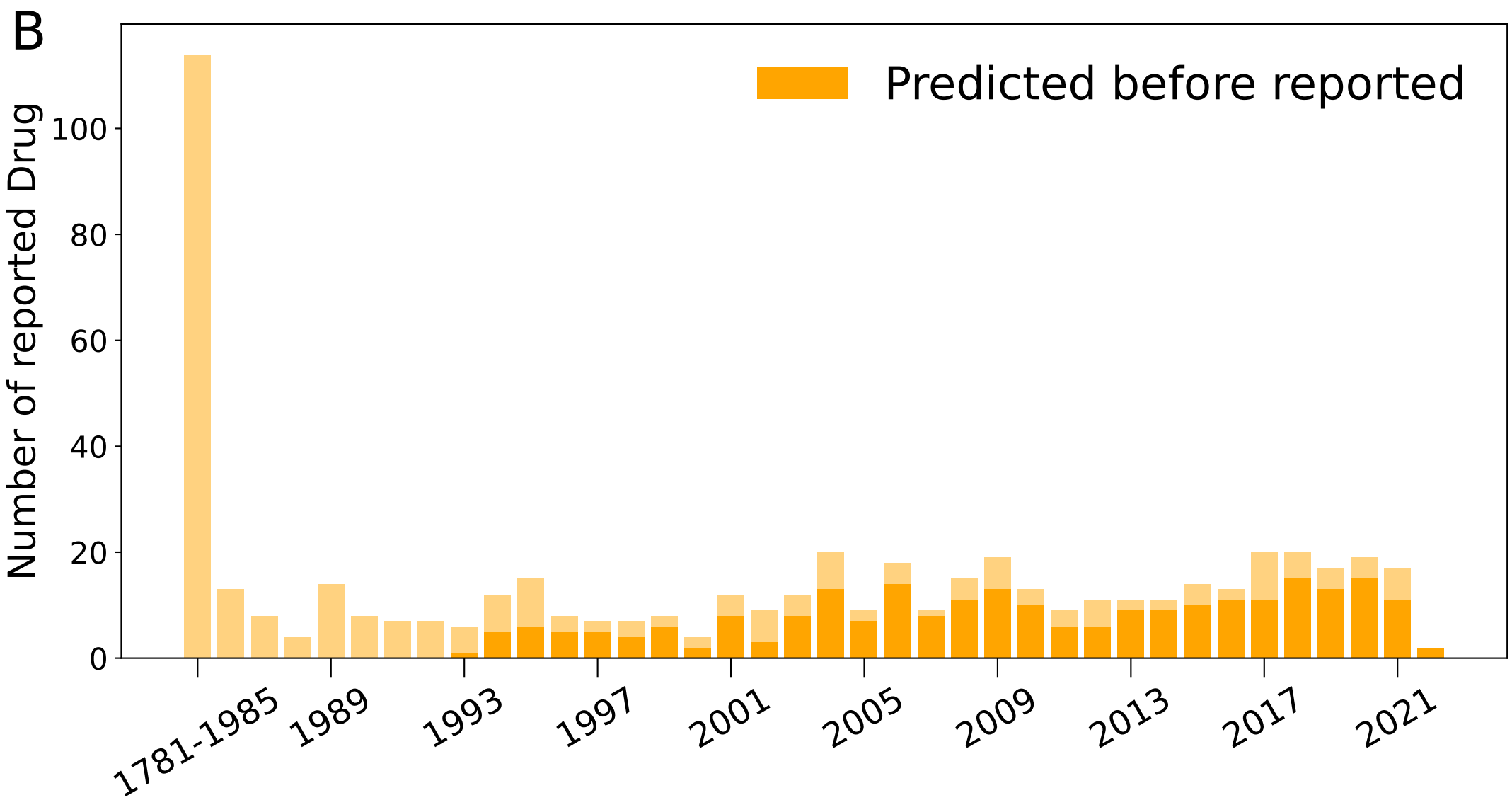
88. Raffel, C. *et al.* Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. (2019).

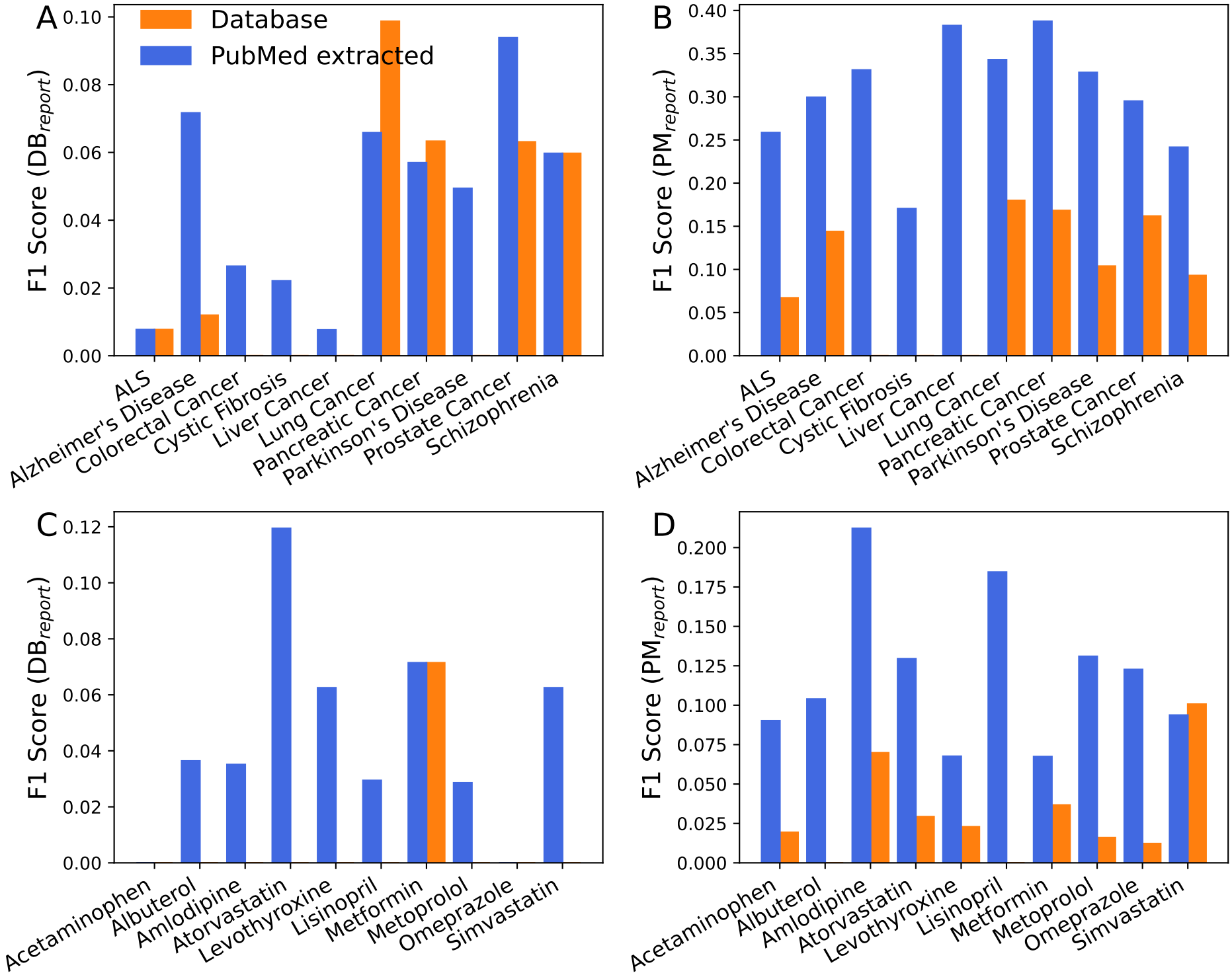
89. Peng, Y., Yan, S. & Lu, Z. Transfer Learning in Biomedical Natural Language Processing: An Evaluation of BERT and ELMo on Ten Benchmarking Datasets. in *Proceedings of the 18th BioNLP Workshop and Shared Task* 58–65 (Association for Computational Linguistics, 2019). doi:10.18653/v1/W19-5006.
90. Alsentzer, E. *et al.* Publicly Available Clinical BERT Embeddings. in *Proceedings of the 2nd Clinical Natural Language Processing Workshop* 72–78 (Association for Computational Linguistics, 2019). doi:10.18653/v1/W19-1909.
91. Sohn, S., Comeau, D. C., Kim, W. & Wilbur, W. J. Abbreviation definition identification based on automatic precision estimates. *BMC Bioinformatics* **9**, 402 (2008).
92. Chandak, P., Huang, K. & Zitnik, M. Building a knowledge graph to enable precision medicine. *Scientific Data* **10**, (2023).
93. Zhou, Y. *et al.* TTD: Therapeutic Target Database describing target druggability information. *Nucleic Acids Research* (2023) doi:10.1093/nar/gkad751.
94. Ashburner, M. *et al.* Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet* **25**, 25–29 (2000).
95. Gene Ontology Consortium *et al.* The Gene Ontology knowledgebase in 2023. *Genetics* **224**, iyad031 (2023).
96. Wilks, C. *et al.* recount3: summaries and queries for large-scale RNA-seq expression and splicing. *Genome Biology* **22**, 323 (2021).
97. Zhang, Y. *et al.* myinsilicom/iKraph: 1.0.0. Zenodo https://doi.org/10.5281/ZENODO.14577964 (2024).
98. Zhagn, Y. *et al.* iKraph: a comprehensive, large-scale biomedical knowledge graph for AI-powered, data-driven biomedical research. Zenodo https://doi.org/10.5281/ZENODO.14846820 (2025).

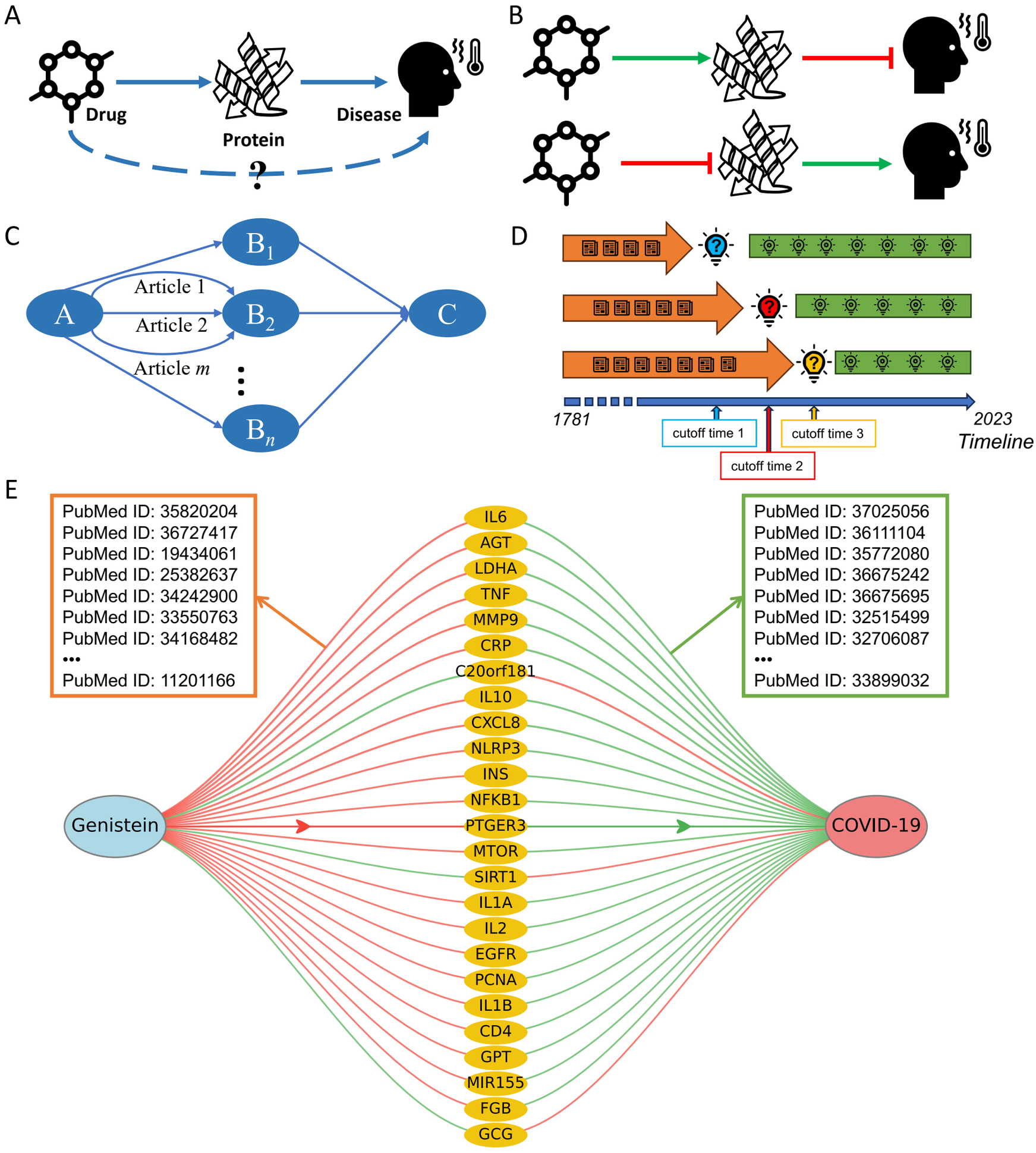

A
Disease
SequenceVariant
Gene
Chemical
9992669
5089
2016377
9045
11124
1176422
203498
5871
3877289
184090
58439
4310678
130877
3374020
1985228
B
SequenceVariant–Disease
13643
Gene–Gene
502344
Chemical–Gene
1165140
Chemical–Disease
3097387
Gene–Disease
1173885
3164
683
70
39803
1119
751
2426
622793
182
286
44523
13979
547851
9277
410288
72787
219020
9156
146055
443526
9464
22430
622968
116124
13056
314488
C
iKraph
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


A
Repurposed Drug
Validated Drug
Gene
Number of Drugs
Number of Genes
2020-03-31
2020-04-30
2020-05-31
2020-06-30
B
Validated by Clinical Trial
Validated by PubMed
Total Validated Drug
Drug validated month by month
Total validated drug
C
Reported as treatment
Not reported as treatment but hypothesized
Not reported as treatment and not hypothesized
Number of Involved Genes
Repurposed Drug
Rapamycin
Resveratrol
Curcumin
Dexamethasone
Infliximab
Metformin
Everolimus
Etanercept
NAC
Quercetin
Adalimumab
Losartan
EGCG
Melatonin
Pentoxifylline
Sitagliptin
Captopril
Estrogen
PD98059
Vitamin D
Rolipram
Calcitriol
Simvastatin
Atorvastatin
Rituximab
Pioglitazone
Berberine
Rosiglitazone
Tacrolimus
L-Arginine
Heparin
Aspirin
Genistein
Thalidomide
Tocilizumab
Indomethacin
Doxycycline
Cyclosporin A
LY294002
Methotrexate
Dex
Silymarin
Ascorbic Acid
Omega-3 Fatty Acids
Vitamin E
Alcohol
U0126
Minocycline
DHA
RA


A
Reported drug as potential new treatment
Median prediction lead time
Number of Predicted Drug
Years
400
350
300
250
200
150
100
50
0
35
30
25
20
15
10
5
0
1781-1985
1991
1995
1999
2003
2007
2011
2015
2019


B
Predicted before reported
Number of reported Drug
100
80
60
40
20
0
1781-1985
1989
1993
1997
2001
2005
2009
2013
2017
2021


A
Database
PubMed extracted
F1 Score (DB$_{report}$)
0.00
0.02
0.04
0.06
0.08
0.10
ALS
Alzheimer's Disease
Colorectal Cancer
Cystic Fibrosis
Liver Cancer
Lung Cancer
Pancreatic Cancer
Parkinson's Disease
Prostate Cancer
Schizophrenia
B
F1 Score (PM$_{report}$)
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0.35
0.40
ALS
Alzheimer's Disease
Colorectal Cancer
Cystic Fibrosis
Liver Cancer
Lung Cancer
Pancreatic Cancer
Parkinson's Disease
Prostate Cancer
Schizophrenia
C
F1 Score (DB$_{report}$)
0.00
0.02
0.04
0.06
0.08
0.10
0.12
Acetaminophen
Albuterol
Amlodipine
Atorvastatin
Levothyroxine
Lisinopril
Metformin
Metoprolol
Omeprazole
Simvastatin
D
F1 Score (PM$_{report}$)
0.000
0.025
0.050
0.075
0.100
0.125
0.150
0.175
0.200
Acetaminophen
Albuterol
Amlodipine
Atorvastatin
Levothyroxine
Lisinopril
Metformin
Metoprolol
Omeprazole
Simvastatin


A
Drug
Protein
Disease
?
B
C
A
B1
B2
Bn
C
Article 1
Article 2
Article m
D
1781
2023
Timeline
cutoff time 1
cutoff time 2
cutoff time 3
E
PubMed ID: 35820204
PubMed ID: 36727417
PubMed ID: 19434061
PubMed ID: 25382637
PubMed ID: 34242900
PubMed ID: 33550763
PubMed ID: 34168482
...
PubMed ID: 11201166
PubMed ID: 37025056
PubMed ID: 36111104
PubMed ID: 35772080
PubMed ID: 36675242
PubMed ID: 36675695
PubMed ID: 32515499
PubMed ID: 32706087
...
PubMed ID: 33899032
Genistein
COVID-19
IL6
AGT
LDHA
TNF
MMP9
CRP
C20orf181
IL10
CXCL8
NLRP3
INS
NFKB1
PTGER3
MTOR
SIRT1
IL1A
IL2
EGFR
PCNA
IL1B
CD4
GPT
MIR155
FGB
GCG