1521-0081/75/4/714–738$35.00
PHARMACOLOGICAL REVIEWS

dx.doi.org/10.1124/pharmrev.122.000715
Pharmacol Rev 75:714–738, July 2023

ASSOCIATE EDITOR: HABIBEH KHOSHBOUEI

Review of Natural Language Processing in Pharmacology

Dimitar Trajanov, Vangel Trajkovski, Makedonka Dimitrieva, Jovana Dobreva, Milos Jovanovik, Matej Klemen, Aleš Žagar, and Marko Robnik-Šikonja

Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University in Skopje, North Macedonia (D.T., V.T., M.D., J.D., M.J.); Computer Science Department, Metropolitan College, Boston University, Boston, Massachusetts (D.T.); and Faculty of Computer and Information Science, University of Ljubljana, Slovenia (M.K., A.Ž., M.R.- Š.)

Address correspondence to: Dimitar Trajanov, ul. Rudzer Boshkovikj 16, P.O. 393, 1000 Skopje, Republic of North Macedonia. E-mail: dimitar.trajanov@finki.ukim.mk

This work is partially based on COST Action CA18209–NexusLinguarum "European Network for Web-Centred Linguistic Data Science," supported by COST (European Cooperation in Science and Technology). The work in this paper was partially financed by the Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University in Skopje. The work was partially supported by the Slovenian Research Agency (ARRS) core research programme P6-0411 and the young researchers grant.

No author has an actual or perceived conflict of interest with the contents of this article.

A preprint of this article was deposited in arXiv [https://doi.org/10.48550/arXiv.2208.10228].

dx.doi.org/10.1124/pharmrev.122.000715.

***Abstract*——Natural language processing (NLP) is an area of artificial intelligence that applies information technologies to process the human language, understand it to a certain degree, and use it in various applications. This area has rapidly developed in the past few years and now employs modern variants of deep neural networks to extract relevant patterns from large text corpora. The main objective of this work is to survey the recent use of NLP in the field of pharmacology. As our work shows, NLP is a highly relevant information extraction and processing approach for pharmacology. It has been used extensively, from intelligent searches through thousands of medical documents to finding traces of adversarial drug interactions in social media. We split our coverage into five categories to survey modern NLP: methodology, commonly addressed tasks, relevant textual data, knowledge bases, and useful programming libraries. We split each of the five categories into appropriate subcategories, describe their main properties and ideas, and summarize them in a tabular form. The resulting survey presents a comprehensive overview of the area, useful to practitioners and interested observers.**

***Significance Statement*——The main objective of this work is to survey the recent use of NLP in the field of pharmacology in order to provide a comprehensive overview of the current state in the area after the rapid developments that occurred in the past few years. The resulting survey will be useful to practitioners and interested observers in the domain.**

I. Introduction

Information processing is indispensable to modern drug design, production, and application. A significant amount of information is stored in textual format and located in scientific papers, clinical notes, ontologies, knowledge bases, social media posts, and newspaper articles. Extraction and retrieval of this information rely on natural language processing (NLP). NLP is a broad scientific area based on computer science, linguistics, and artificial intelligence (Jurafsky and Martin, 2008, 2022). As the whole area of artificial intelligence, it has been completely transformed in recent years by deep learning (Goodfellow et al., 2016). It has witnessed numerous new techniques and successful applications, such as intelligent search, machine translation, and speech recognition.

Many general NLP techniques and approaches can be applied to the pharmacological area. However, often NLP techniques have to be adapted to the specifics of the field in terms of available knowledge sources, text representation, specific methods, terminology, and so on. In this work, we survey modern NLP methodology, tasks, resources, knowledge bases, and tools used and adapted to the area of pharmacology. The review aims to inform practitioners working in the area of the pharmacology of exciting recent development and to give a solid starting reference material to new entrants.

Several surveys summarize NLP in pharmacology but only cover specific areas of NLP methods. One of the first reviews of NLP for clinical decision support (CDS) (Demner-Fushman et al., 2009) was published in 2009. The authors observed that many CDS data are textual and reviewed existing NLP developments for CDS. Luo et al. (2017) present a structured review of NLP for narratives in electronic health records (EHR) for pharmacovigilance. Dreisbach et al. (2019) review NLP of symptoms from electronic patient-authored text data. A review of NLP in languages other than English for clinic-related texts is presented by Névéol et al. (2018). Chen et al. (2021b) survey NLP addressing challenges related to the COVID-19 pandemic. They present details related to several NLP tasks like information retrieval, named entity recognition, literature-based discovery (LBD), question answering (QA), topic modeling, entiment and emotion analysis, caseload forecasting, and misinformation detection. In contrast to the listed surveys, we aim for a comprehensive overview of NLP in pharmacology.

NLP is a subfield of much broader areas of machine learning (ML) and artificial intelligence (AI). Many ML algorithms not related to text are applicable to pharmacological tasks. While the focus of this paper is the application of NLP in pharmacology, we here refer to several recent survey articles that cover ML application in specific pharmacological tasks, like drug discovery (Stephenson et al., 2019; Carracedo-Reboredo et al., 2021; Dara et al., 2022), drug-target interaction prediction (Le and Le, 2016; Chen et al., 2018), drug repurposing (Yang et al., 2022), drug-drug interactions (Han et al., 2022), ML applications for COVID-19 (Kamalov et al., 2022),

ABBREVIATIONS: ADE, adverse drug event; ADR, adverse drug reaction; AI, artificial intelligence; BERT, Bidirectional Encoder Representations from Transformers; Bi-LSTM, bidirectional long-short term memory; BLUE, Biomedical Language Understanding Evaluation; CDS, clinical decision support; CORD-19, COVID-19 Open Research Database; DDA, drug-disease association; DDI, drug-drug interaction; DNN, deep neural network; EHR, electronic health records; ELMo, Embeddings from Language Model; EMBASE, Excerpta Medica Database; KG, knowledge graph; LBD, literature-based discovery; LOD, Linked Open Data; LLM, large language model; LM, language model; MIMIC, Medical Information Mart for Intensive Care; ML, machine learning; MLM, masked language modeling; NCBI, National Center for Biotechnology Information; NER, named entity recognition; NLI, natural language inference; NLP, natural language processing; NLTK, Natural Language Toolkit; POS, part of speech; RDF, Resource Description Framework; PharmaCoNER, Pharmacological Substances, Compounds and Proteins, Named Entity Recognition; RoBERTa, A Robustly Optimized BERT Pretraining Approach; QA, question answering; SHAP, SHapley Additive exPlanations; UMLS, Unified Medical Language System.

pharmacometrics (Janssen et al., 2022; McComb et al., 2022), cancer management (Kumar and Saha, 2022), microbiome therapeutics (McCoubrey et al., 2021), exploratory pharmacovigilance (Kaas-Hansen et al., 2023), biomaterials (Kerner et al., 2021), and many more.

Recently, the primary methodological approach to NLP has been deep learning. Deep neural networks (DNNs) require that text is transformed (embedded) into numeric vectors in a process called representation learning. We present general text embeddings as well as specific variants relevant to the area of life sciences and pharmacology. As pharmacology is a knowledge-intensive area where relevant information is not stored only in text documents but also in databases, ontologies, and linked data, we survey recent attempts to inject knowledge into DNNs. Due to the need to understand the decisions and biases of DNNs, we discuss techniques that make their output more transparent.

Some NLP tasks are particularly important to the area of pharmacology. While they are often based on general approaches, they are strongly adapted and use specific pharmacological resources. We discuss general tasks such as named entity recognition, relation extraction, LBD, QA, and field-specific tasks such as detection of adverse drug reactions (ADR) and drug discovery.

The basic precondition for applying NLP is the availability of language resources. In multiple studies, EHRs are the main source of information (Wang et al., 2009; Li et al., 2018; Jagannatha et al., 2019; Liu et al., 2019a; Wunnava et al., 2019). EHRs contain patient data such as diagnoses, hospital admissions, prescriptions, and adversary drug effects. The data in EHRs are well structured and can be readily processed; however, different EHR components are difficult to integrate. Many authors use molecular data (Suthram et al., 2010; Park et al., 2011), which can be integrated with diseases (Goh et al., 2007). Other important sources of information are clinical data (Jung and Lee, 2013) [used in, e.g., drug repurposing (Yang et al., 2017; Deftereos et al., 2011)], linked data, and the pharmacology-related semantic web.

Linked data and knowledge graphs have recently emerged as general formalisms to represent knowledge in artificial intelligence and the semantic web. Linked (open) data movement introduced new standards for representing, storing, and retrieving data over the web (Bizer et al., 2008, 2009; Heath and Bizer, 2011; Wood et al., 2014; Hogan et al., 2021), which enabled new distributed data sources and new applications. Knowledge graphs allow generating, consolidating, and contextually linking structured data. We present several knowledge graphs from the biomedical domain and outline several COVID-19-related knowledge graphs.

Software tools and libraries are essential for using NLP in pharmacological research and practice. Mostly, these support the Python language. We present many general NLP tools and libraries as well as life science and pharmacology-specific variants.

We organize the survey along with five main areas: methodology, common tasks, datasets, knowledge graphs, and software libraries. In Section II, we structure NLP methodologies into three groups: representation learning (i.e., different embeddings), approaches to inject domain-specific knowledge into deep neural networks, and explainable AI techniques used in pharmacology. The most frequently used NLP tasks in pharmacology are presented in Section III. We cover the named entity recognition, relation extraction, ADRs, LBD, and QA. In Section IV, we first outline the approaches to finding data resources, followed by a survey of existing data. We organize the overview into five categories: patient data, drug usage data, drug structure data, QA datasets, and general text processing datasets. Knowledge graphs used in the biomedical domain and a specific example of COVID-19 disease knowledge graphs are covered in Section V. We give an overview of useful NLP software libraries and tools for the pharmacological domain as well as useful general NLP libraries in Section VI. We conclude the survey in Section VII.

II. Natural Language Processing Methodology in Pharmacology

Recently, NLP has switched entirely to deep neural networks, mostly large language models (LLMs) that are pretrained on huge quantities of text to capture various linguistic, general, and domain-specific knowledge. LLMs embed the text data into a numeric representation preserving semantic relations between words. To be used for specific tasks, LLMs are fine-tuned with problem-specific data.

In Section II.A, we give an overview of modern text representations. We present static and contextual embeddings (i.e., LLMs) and specific variants relevant to the area of life sciences and pharmacology. While most of the work is focused on English, we present some notable exceptions in other languages. As pharmacology is a knowledge-intensive area where relevant information is not stored only in text documents but also in databases, ontologies, and linked-data, we survey recent attempts to inject knowledge into deep neural networks in Section II.B. Unfortunately, deep neural networks often appear as black-box models, lacking transparency on how the decisions are taken. In Section II.C, we present general explanation techniques applicable to text prediction and focus on successful applications related to pharmacology.

A. Representation Learning

In NLP, text representation is a crucial issue and research direction. Various text embeddings emerged that capture both syntax and semantics of a given text. While traditional approaches were based on sparse

representations such as bag-of-words, dense representations such as word2vec (Mikolov et al., 2013), Embeddings from Language Model (ELMo) (Peters et al., 2018), and Bidirectional Encoder Representations from Transformers (BERT) (Devlin et al., 2019) are based on neural networks and offer much more semantically valid and computationally efficient representations. A common trait of these embeddings is to train a neural network on self-supervised text classification tasks and use the weights of the trained neural network or the whole trained network to represent different text units (words, sentences, or documents). The labels required for training these classifiers originate from large corpora of general texts, e.g., web crawl, news, and Wikipedia. The usual classification tasks used in training these representation models are predicting the next and previous word in a sequence or filling in missing words [also called masked language modeling (MLM)]. Representation learning can be extended with other related tasks, such as prediction if two sentences are sequential. The positive instances for learning are obtained from the text in the given corpus, while the negative instances are mostly sampled from instances that are unlikely to be related.

We first briefly describe the principle of the most frequently used static embeddings, called word2vec, followed by large language models such as contextual BERT. Next, we cover the adaptations of these representation techniques for life sciences and pharmacology domains. We provide a summary of the presented embeddings in Table 1.

1. Static Embeddings. The word2vec word embedding method (Mikolov et al., 2013) trains a shallow (one hidden layer) neural network predicting the neighboring words of a given input word. The trained weights of the hidden layer produce a static embedding in the sense that we get a single vector for each word. For example, the term *bank* may denote a financial institution or land

TABLE 1
Representation models (i.e., embeddings for text and biologic sequences) useful for pharmacology

Name	Description	Trained on	Usage
Static embeddings			
word2vec (Mikolov et al., 2013)	General static word embeddings	Any collection of text, e.g., Wikipedia dump	Any general noncontextual text processing
PubMed-PMC, WikiPubMed-PMC (Pyysalo et al., 2013)	Word2vec adapted to life sciences	PubMed abstracts and articles; in combination with Wikipedia	Any noncontextual life science text processing, e.g., biomedical NER for genes, chemicals, and diseases (Habibi et al., 2017)
BioVec, ProtVec, GeneVec (Asgari and Mofrad, 2015) dna2vec (P. Ng et al., preprint, DOI: https://doi.org/10.48550/arXiv.1701.06279)	Word2vec style embeddings for biologic sequences, genes, and proteins	Different biologic sequences, e.g., Swiss-Prot	Proteomics and genomics, e.g., structure prediction for proteins
Contextual embeddings			
BERT (Devlin et al., 2019), RoBERTa (Y. Liu et al., preprint, DOI: https://doi.org/10.48550/arXiv.1907.11692)	General contextual text embeddings	Large general text corpora such as Wikipedia and Common Crawl	Any general text processing
SciBERT (Beltagy et al., 2019)	Contextual embeddings for scientific texts	Scientific papers from Semantic Scholar	NER for clinical use, text classification, relation classification
Character BERT (El Boukkouri et al., 2020)	Character-level input allows for easy adaptation to different areas	Clinical texts and PubMed abstracts	Medical NER, NLI, RE, and clinical sentence similarity
BioMed-RoBERTa (Gururangan et al., 2020)	RoBERTa adaptation for life sciences	Standard texts, abstracts, and full papers from PubMed	Chemical-protein-disease annotations, sequential sentence classification task
BioBERT (Lee et al., 2019)	BERT adapted to life sciences	BERT further trained on PubMed abstracts and papers	Biomedical NER, RE, and QA
Bio+Clinical BERT (Alsentzer et al., 2019)	BioBERT adapted to clinical texts	BioBERT further pretrained with clinical notes and discharge summaries	Clinical NER and medical NLI.
Clinical BERT (K. Huang et al., preprint, DOI: https://doi.org/10.48550/arXiv.1904.05342)	Suitable for clinical texts	Clinical notes from EHR for patients in intensive care units	Clinical readmission prediction
BLUE BERT (Peng et al., 2019)	Suitable for clinical texts	PubMed abstracts and clinical notes	Good performance on BLUE benchmark, including NER and RE
CovBERT (Khadhraoui et al., 2022)	BERT adapted to COVID-19	BERT further pretrained on PubMed abstracts with COVID-19 relevant contents	Tasks related to COVID-19
RuDR-BERT (Tutubalina et al., 2021)	Multilingual BERT adapted to pharmacology in Russian	mBERT further pretrained on consumer reviews about pharmaceutical products	NER and multiclass classification

RE, relation extraction.

alongside a river, but it is represented with a single vector.

The word2vec method pretrains a feed-forward neural network on a huge corpus, and the weights of the hidden layer in this network are used as word embeddings. Pretrained word vectors for many languages are publicly available. The published vectors are typically 100- or 300-dimensional; e.g., Google published vectors for 3 million English words and phrases (https://code.google.com/archive/p/word2vec/). While the word2vec algorithm consists of two related methods, we describe only the skip-gram method, which mostly produces more favorable results. The method constructs a neural network to classify cooccurring words by taking a word and predicting its d preceding and succeeding words, e.g., ± 5 words. In the actual neural network, one word is on the input (the central word) and one word is on the output, where both are represented with one-hot encoding. The words and their contexts appearing in the training corpus constitute the training instances of the classification problem. The first word of the training pair is presented at the network's input in the one-hot-encoding representation, and the network is trained to predict the second word. The difference in prediction is evaluated using a loss function. For a sequence of T training words $w_1, w_2, w_3, \ldots, w_T$, the skip-gram model maximizes the average log probability:

$$\frac{1}{T}\sum_{t=1}^{T}\sum_{-d\le j\le d, j\neq 0}\log p(w_{t+j}|w_t). \quad (1)$$

Once the network is trained with word2vec, vectors for each word in the vocabulary can be generated. As one-hot encoding of the input word only activates one input connection for each hidden layer neuron, the weights on these connections constitute the embedding vector for the given input word.

The resulting word embeddings' properties depend on the context's size. For a small number of neighboring words (e.g., ± 5 words), we get embeddings that perform better on syntactic tasks. For larger neighborhoods (e.g., ± 10 words), the embeddings better express semantic properties.

Word2vec has attracted the immense attention of NLP researchers and practitioners. The word2vec precomputed embeddings soon became a default choice for the first layer of many classification deep neural networks. Several domain-specific variants have also been created and made publicly available. For life sciences, a well-known example is the work of Pyysalo et al. (2013), who released two sets of word2vec vectors. The first, denoted PubMed-PubMed Central, was trained on 23 million PubMed abstracts and 0.7 million PubMed Central articles. The second model, Wiki-PubMed-PubMed Central, was prepared using the same two corpora combined with 4M English Wikipedia articles. These static embeddings were successfully used in many life science applications (the paper had received 492 citations by March 13, 2022). For example, Habibi et al. (2017) have successfully applied the two embeddings to the biomedical named entity recognition problem to detect genes, chemicals, and diseases.

Note that the same technology to represent text can be applied to represent biologic sequences, such as DNA, RNA, and proteins (Asgari and Mofrad, 2015). The created biovectors (BioVec) refer to biologic sequences in general, protein vectors are called ProtVec, and gene vectors are named GeneVec. A similar attempt to represent biologic sequences is dna2vec vectors (P. Ng et al., preprint, DOI: https://doi.org/10.48550/arXiv.1701.06279).

Despite the successful use of static embeddings such as word2vec, contextual embedding models such as BERT have become even more successful. Therefore, we skip the detailed review of static embedding models and focus on contextual models.

2. Contextual Word Embeddings. The problem with word2vec embeddings is their failure to express polysemous words. During its training, all senses of a given word (e.g., *paper* as a material, as a newspaper, as a scientific work, and as an exam) contribute relevant neighboring words in proportion to their frequency in the training corpus. This causes the final vector to be placed somewhere in the weighted middle of all words' meanings. Consequently, rare meanings of words are poorly expressed with word2vec, and the resulting vectors do not offer good semantic representations. For example, none of the 50 closest vectors of the word *paper* is related to science.

The idea of contextual word embeddings is to generate a different vector for each word's context. The context is typically defined sentence-wise. This solves the problems with word polysemy. The context of a sentence is mostly enough to disambiguate different meanings of a word for humans and learning algorithms. Several contextual embeddings have been developed, e.g., ELMo, Universal Language Model Fine-tuning for Text Classification, and BERT. As the latter achieves the best results in most NLP tasks, we describe it next.

Contextual embeddings are based on the idea of language models, which predict either the next, previous, or missing word in a sequence. Training often combines several of these and other related tasks. Due to the network's depth, extracting vector representations from the network is no longer trivial, i.e., the trained deep networks store their knowledge in weights spread over several layers. A frequently used approach concatenates weights from several layers into a vector. Still, often it is more convenient to use the whole pretrained neural language model as a starting point and fine-tune its weights further during the training on a specific task.

BERT embeddings (Devlin et al., 2019) generalize the idea of language models (LMs) to masked language models, inspired by the gap-filling tests. The masked

language model randomly masks some of the tokens from the input. The task of an LM is then to predict each missing token based on its neighborhood. BERT uses the transformer architecture of neural networks (Vaswani et al., 2017) in a bidirectional sense (forward and backward). It introduces another task of predicting whether two sentences appear in a sequence. The input representation of BERT is sequences of tokens representing subword units. The input is constructed by summing the corresponding token, segment, and position embeddings.

Using BERT for classification requires adding connections between its last hidden layer and new neurons corresponding to the number of classes in the intended task. The fine-tuning process is typically applied to the whole network. All the BERT parameters and new class-specific weights are fine-tuned jointly to maximize the log-probability of the correct labels.

BERT has shown excellent performance on many NLP tasks and is now a de facto standard in NLP. In the initial evaluation (Devlin et al., 2019), BERT showed improved performance on all eight tasks from the general language understanding evaluation benchmark suite (Wang et al., 2018), consisting of QA, named entity recognition, and common-sense inference. A variant of BERT, called A Robustly Optimized BERT Pretraining Approach (RoBERTa) (Y. Liu et al., preprint, DOI: https://doi.org/10.48550/arXiv.1907.11692), which only uses masked language model training but on a larger dataset and for a longer time, has become a popular practical choice due to its improved robustness and better parallel training capability.

Due to its success, BERT has spurred an immense tide of research analyzing its capabilities and using and adapting it for different purposes. An overview of research on BERT capabilities and inner workings is presented by Rogers et al. (2020). Here we overview the adaptations and applications relevant to pharmacology.

3. BERT Variants Relevant to Pharmacology. BERT has many extensions in architecture, training, and fine-tuning. A general improvement for science-related text processing is SciBERT (Beltagy et al., 2019) that was trained on 1.14 million scientific papers (3.17 billion tokens) from Semantic Scholar instead of general text. The training data consisted of 18 percent computer science papers and 82 percent papers from the biomedical domain. Upon its introduction, the SciBERT was compared with BERT and achieved improved performance in a study involving four classification tasks based on scientific publications: named entity recognition (NER), extraction of participants, interventions, comparisons, and outcomes in clinical trial papers, text classification, relation classification, and dependency parsing. The SciBERT has attracted considerable attention of the scientific community with more than 1000 citations recorded by Google Scholar at the time of this writing.

In life sciences, there are several popular domain adaptations of BERT. BioMed-RoBERTa-base (Gururangan et al., 2020) (almost 600 Google Scholar citations at the time of this writing) is an adaptation of RoBERTa (Y. Liu et al., preprint, DOI: https://doi.org/10.48550/arXiv.1907.11692), using long pretraining on 160GB of standard texts and an additional 47GB (7.55 billion tokens from 2.68 million papers) of abstracts and full papers randomly sampled from PubMed repository. Using this domain-adapted pretrained model, the authors improved classification for two domain-specific tasks. First, they improved the classification compared with the baseline RoBERTa model for 2.3 micro F_1 percent on the Chem-Prot database (Kringelum et al., 2016) that contains chemical-protein-disease annotations enabling the study of systems pharmacology for a small molecule across multiple layers of complexity from molecular to clinical levels. Second, they tested the BioMed-RoBERTa on the PubMed sequential sentence classification task (Dernoncourt and Lee, 2017) and achieved 0.4 micro F_1 percent advantage over RoBERTa.

The BioBERT (Lee et al., 2019) representation model (almost 2000 Google Scholar citations at the time of this writing) was initialized with BERT weights and then pretrained using domain-specific literature, namely PubMed abstracts (4.5 billion words) and PubMed Central full-text articles (13.5 billion words). The resulting model was successfully fine-tuned for three biomedical text mining tasks: biomedical named entity recognition, biomedical relation extraction, and biomedical QA. The BioBERT model was further pretrained for clinical texts using 2 million generic clinical notes and discharge summaries (Alsentzer et al., 2019). The resulting Bio+Clinical BERT showed superior results on clinical NER tasks and medical natural language inference (NLI) task.

Clinical BERT (K. Huang et al., preprint, DOI: https://doi.org/10.48550/arXiv.1904.05342) is similar to the Bio+Clinical BERT model, but it is trained on 2,083,180 anonymized clinical notes from the Medical Information Mart for Intensive Care (MIMIC-III) database (Johnson et al., 2016) that consists of the electronic health records of 58,976 unique hospital admissions from 38,597 patients in the intensive care unit between 2001 and 2012. The model performed better than BERT on the clinical readmission prediction problem. A similar model is BLUE BERT (Peng et al., 2019), trained on more than 4 billion PubMed abstracts and 500 million MIMIC-III clinical notes. The model showed good performance on BLUE (Biomedical Language Understanding Evaluation) benchmark that includes several tasks relevant to pharmacology, like named entity recognition (see Section III.A) and relation extraction (see Section III.B).

In the light of COVID-19 epidemics, Khadhraoui et al. (2022) have prepared a specialized BERT model, called CovBERT, intended to improve the COVID-19 literature review. The model, based on BERT, was pretrained on

4304 PubMed abstracts on several topics such as COVID-19 treatment, COVID-19 symptoms, virology, public health, and mental health. CovBERT showed better classification accuracy on this dataset compared with baseline RoBERTa, AlBERT, SciBERT, BioBERT, and Bio+Clinical BERT.

Another popular adaptation to specific terminological areas is named CharacterBERT (El Boukkouri et al., 2020). Instead of using subword tokenization, this approach starts with characters and first constructs words with a convolutional neural network. The pretraining used around 1 billion tokens from the MIMIC-III clinical dataset and PubMed abstracts. The effectiveness of this approach was originally demonstrated in the biomedical domain using four tasks: medical entity recognition, medical NLI, relation extraction (Chem-Prot database and drug-drug interactions), and clinical sentence similarity. The resulting CharacterBERT models performed on par or better than BERT.

As evident from many citations, the BERT enhancements received, these models were successfully applied to many relevant pharmacological problems. We list a sample of works addressing a few relevant problems and approaches in Section III.

4. Languages Other than English. While the majority of NLP in pharmacology is focused on English, there are also some exceptions. Akhtyamova (2020) trains a domain-specific BERT model for Spanish on a relatively small dataset (87 million tokens) and successfully applies it to the problem of NER in Spanish. In the context of the annual workshop on BioNLP Open Shared Tasks, in 2019 (https://2019.bionlp-ost.org/) one of the tasks, Pharmacological Substances, Compounds and Proteins and Named Entity Recognition (PharmaCoNER) track, addressed the mentioning of chemicals and drugs in Spanish medical texts. The task included two tracks: one for the NER offset and entity classification and the other for the concept indexing. In their entry, Xiong et al. (2019) devised a system based on BERT for the NER offset and entity classification and bidirectional long-short term memory (Bi-LSTM) with max/mean pooling for concept indexing. On the same tasks, Sun et al. (2021) compared several BERT variants (see Section II.A.3): BLUE BERT (Peng et al., 2019), multilingual BERT (Devlin et al., 2019), SciBERT (Beltagy et al., 2019), BioBERT (Lee et al., 2019), and Spanish BERT (Canete et al., 2020). The results show that domain-specific pretraining is successful and better than the language-specific BERT variant.

For the ADR relation extraction in Russian, Sboev et al. (2022) have preliminary trained multilingual XLM-RoBERTa (Conneau et al., 2020), and Russian RuBERT (K. Kuratov and M. Arkhipov, preprint, DOI: https://doi.org/10.48550/arXiv.1905.07213) models on Russian drug review texts, followed by fine-tuning on the created training dataset. The results showed that the former multilingual model is advantageous. Tutubalina et al. (2021) have created a consumer reviews corpus in Russian about pharmaceutical products for the detection of health-related named entities and the assessment of pharmaceutical product effectiveness. Using this corpus and the multilingual BERT, they created domain-specific RuDR-BERT, which showed favorable performance on medical named entity recognition and multilabel sentence classification.

B. Injecting Pharmacological Knowledge into Deep Neural Networks

While large pretrained language models have significantly increased the performance of ML approaches for most NLP tasks, many shortcomings still make the approaches less robust as desired. Examples of weaknesses are processing of negation, uncertainty about factual knowledge, and lack of problem-specific knowledge (Rogers et al., 2020).

The knowledge injection approaches attempt to address the shortcomings of large pretrained models by utilizing external knowledge resources in various forms, such as knowledge graphs (KGs; see Section V) and other types of knowledge bases. This can reduce the need for ever-larger language models while improving their interpretability. In general, knowledge injection approaches differ in time of injection (during a pretraining phase, as an intermediate task, or in a downstream task), type of injected knowledge (facts, linguistic knowledge, commonsense reasoning), and type of evaluation (general language, domain-specific language, or probing).

To improve pretrained language models for the biomedical domain, the existing approaches usually use the Unified Medical Language System (UMLS) knowledge base. UMLS is a medical terminology database with hundreds of biomedical vocabulary entries, including definitions of terms and relationships between them. The basic BERT model (Devlin et al., 2019) or any of the specific biomedical BERT models mentioned in Section II.A.3, are used as a baseline where the knowledge is injected.

Next we present several approaches to knowledge injection in pharmacology. We divide them into the ones that modify existing pretraining tasks with general improvements in mind and those that focus on better concept representation for a specific task. We summarize the presented models in Table 2.

1. Modification of Existing Pretraining Tasks for General Improvement. This group of knowledge injection approaches focuses on developing new pretraining tasks or adding new modules to existing pretrained LMs.

Hao et al. (2020) improve biomedical LMs for medical downstream tasks by infusing the knowledge base information into the pretraining phase of the Clinical BERT. The authors used the MIMIC-III dataset and continued pretraining on the MLM task and next sentence prediction. They also introduced the task of predicting whether a relationship exists between two

TABLE 2
Summary of knowledge enhanced models
All methods are evaluated on more than one pretrained model. Here we report the one that achieved the best results. For CODER, we report the best monolingual in multilingual versions of the models.

Name	External knowledge	Pretrained model	Evaluation tasks
Hao et al. (2020)	MIMIC-III, UMLS	ALBERT	NER, NLI
UmlsBERT (Michalopoulos et al., 2021)	MIMIC-III, UMLS	Bio_ClinicalBERT	NER, NLI
Meng et al. (2021)	UMLS	PubMedBERT	Document classification, NLI, QA
CODER (Yuan et al., 2022)	UMLS	PubMedBERT, mBERT	Term normalization
SAPBERT (Liu et al., 2021)	UMLS	PubMedBERT	Entity linking
Mao and Fung (2020)	UMLS	BioWordVec	Semantic relatedness, WSD

WSD, word sense disambiguation.

concepts in the UMLS knowledge base. Positive instances for this task are taken from the existing relations in ULMS, while negative ones are created through negative sampling as relations in ULMS are very sparse. The final loss function used in training is a combination of all three tasks. The resulting knowledge-enhanced Clinical BERT was evaluated on two named entity recognition datasets and one NLI dataset, and the results showed an improvement over the baseline biomedical models BioBERT and Clinical BERT.

UmlsBERT (Michalopoulos et al., 2021) also integrates external knowledge resources to improve biomedical language models. The authors updated the MLM in the pretraining step with the associations between the words specified in the UMLS. First, at the input level, medical terms are enhanced by their semantic types (UMLS contains 44 unique semantic types). For example, the model receives information that "lungs" are "body part," "organ," and so on. This represents an additional input layer that must be trained. Words without semantic type are represented by a zero-filled vector. Second, the MLM task is modified: instead of predicting one missing token, the model predicts all words associated with the same concept unique identifier. For instance, where the standard MLM task predicts only "lung," the modified one predicts "lungs" and "pulmonary" as well. UmlsBERT achieves the best results in four out of the five tasks (one NLI and four NER tasks). The ablation study checking if semantic type information improves the performance shows that the model performs significantly worse on all tasks without it.

Meng et al. (2021) improve biomedical BERTs by partitioning a very large KG into smaller subgraphs and infusing this knowledge into various BERT models using adapters. Adapters (Houlsby et al., 2019) are BERT additions that add only a few new trainable parameters while the original weights remain fixed. This reduces the inefficiency of fine-tuning large models for each task and allows a high degree of parameter sharing. Meng et al. (2021) construct two KGs from the UMLS knowledge graph. The METIS algorithm (Karypis and Kumar, 1998) partition the knowledge graph into n subgraphs. Following that, they train an adapter module for each subgraph to predict the tail entity of a triplet from the subgraph. Finally, they use AdapterFusion mixture layers (Pfeiffer et al., 2021) to combine the knowledge from adapter modules. They experimentally determined that 20 subgraphs and PubMedBERT yielded the best results. Their approach improves performance on QA, NLI, and document classification tasks in the biomedical domain.

2. Improved Concept Representation for Specific Tasks. This group of knowledge injection approaches focuses on improving concept representations for specific tasks.

The same medical concepts can be represented by a variety of nonstandard names, misspellings, and abbreviations. Term normalization is a task that addresses this problem. CODER (Yuan et al., 2022) proposes dual contrastive learning simultaneously on both terms and relation triplets from the UMLS KG. The approach is motivated by examples such as that it is better to have "rheumatoid arthritis" closer to "osteoarthritis" than "rheumatoid pleuritis" because both are subtypes of arthritis. Relations between terms express that and thus provide useful information during the training. CODER maximizes similarities between positive term-term pairs and term-relation-term pairs from the KG. They evaluate their approach on datasets in different languages consisting of term normalization, relation classification, and conceptual similarity tasks. Their approach significantly outperforms existing medical embeddings in zero-shot term normalization.

Liu et al. (2021) address the problem of entity linking, specifically, the heterogeneous naming of medical concepts. The authors pretrain a transformer-based language model on the UMLS biomedical KG. They propose a metric learning framework that learns to cluster synonyms of the same concept. The goal of a self-alignment pretraining step is to learn such concept embeddings that maximize the similarity between two concepts based on the cosine similarity measure. The learning setup consists of triplets in the form (x_a, x_p, x_n), where x_p is a positive match for x_a and x_n is its negative match. This approach first samples hard triplets (triplets that contain negative pairs closer in space than positive pairs with basic BERT embedding by some margin). It learns to push negative pairs away from each other and positive pairs together by considering the multisimilarity loss function. The resulting SAPBERT improves

the accuracy across six medical entity linking tasks (up to 20 percent) compared with the domain-specific BERT models and achieves state-of-the-art results.

Mao and Fung (2020) tackle the problem of measuring semantic relatedness between biomedical concepts (UMLS concepts). Semantic similarity expresses the relatedness of two concepts in their meaning and is an important tool for automatic spelling correction, information retrieval, and word sense disambiguation. Authors use pretrained word embedding models (e.g., BioWordVec (Zhang et al., 2019), variations of BERT to generate concept sentence embeddings from UMLS, and various graph embedding models [e.g., graph convolutional networks (Welling and Kipf, 2016), TransE (Bordes et al., 2013) and its variants]. In addition, they combined both concept sentence embeddings and graph embeddings by concatenation. The similarity score between two embeddings was computed using the cosine similarity measure. The combined word and graph embeddings produced the best results on three semantic relatedness datasets and a one-word sense disambiguation dataset.

C. Explainable Natural Language Processing in Pharmacology

Deep learning models commonly surpass standard ML models in terms of predictive performance. However, their decision-making process is typically opaque, meaning that it is difficult to explain why the model made a certain prediction. Understanding models' inner workings are helpful for debugging errors, possibly improving their performance, and gaining scientific insights into the modeled process, e.g., why two drugs interact in the drug-drug interaction identification. Additionally, as pharmacology is concerned with drugs affecting humans, it is essential that predictions are safe and verifiable.

Depending on the time when an explanation is created, there are two types of explanation methods: intrinsic and post hoc (Madsen et al., 2022). Intrinsic methods use a model's architecture or its components to construct an explanation. A simple example is a linear regression model using binary bag-of-words features. The learned weights associated with the input words represent an explanation of the prediction for the given input (positive weights indicate the positive impact of words on the decision, and negative weights indicate negative impact). Another commonly used intrinsic method used for large pretrained transformer models is the inspection of attention weights, which intuitively represent the parts of the input the model focuses on. Attention, being the key component of the currently dominant transformer-based models, is easy to compute. However, multiple attention heads may be difficult to comprehend, and the alignment between attention explanations and the underlying model behavior (i.e., actual explanations) is questionable (Jain and Wallace, 2019; Wiegreffe and Pinter, 2019).

Post hoc explanation methods construct an explanation after a model is trained. While intrinsic methods are based on the design of a specific model, post hoc methods are typically model-agnostic. An example of such methods are perturbation-based explanation methods such as Local Interpretable Model-agnostic Explanations (Ribeiro et al., 2016), SHapley Additive exPlanations (SHAP) (Lundberg and Lee, 2017), and Interactions-based Method for Explanation (Štrumbelj and Kononenko, 2013). They work by repeatedly modifying (perturbing) the input, observing the changes in the output, and modeling their associations using a surrogate model. Post hoc methods are convenient due to their flexibility in the choice of used model architectures. However, the faithfulness of the produced explanations may be poor (Slack et al., 2020; Frye et al., 2021) as they explain the model from an external perspective.

Both intrinsic and post hoc methods have been successfully applied to general (Lai et al., 2019) and topic-specific language tasks in biomedicine (Moradi and Samwald, 2021). Next we describe several cases of using explanation methods in pharmacology. We provide an overview of the methods in Table 3.

TABLE 3
An overview of the explanation methods used in NLP for pharmacology

Reference	Explanation type	Short description	Downstream tasks
Jha et al. (2018)	Intrinsic	Interpretable word embedding transformation	Semantic concept categorization
Wawrzinek et al. (2020)	Intrinsic	Embedding arithmetic (analogies)	Drug-disease association prediction
Huang et al. (2020)	Intrinsic	Interpretable subspace	Drug-drug interaction prediction
Yazdani-Jahromi et al. (2022)	Intrinsic	Interpretable component (attention weights)	Drug-target interaction prediction
Bradshaw et al. (2019)	Intrinsic	Interpretable component (reaction predictor)	Molecule generation
Jiang et al. (2019)	Post hoc	Present representative examples	Detection of potential adverse medication effects
Rodríguez-Pérez and Bajorath (2020)	Post hoc	Out-of-the-box method (SHAP)	Structure-activity relationship modeling
Pope et al. (2019)	Intrinsic	Adapt out-of-the-box methods (gradient-based saliency, class activation mapping, excitation backpropagation)	Identification of biologic molecular properties

Jha et al. (2018) make pretrained word embeddings more interpretable by learning a transformation to a more interpretable embedding space with the retained performance. The interpretable word embeddings correspond to categorical embeddings, trained separately using expert-provided definitions and additional knowledge from a biomedical knowledge graph.

Wawrzinek et al. (2020) introduce an entity embedding-based explanation method for drug-disease association (DDA) prediction. They construct explanations following the drug-centric and the disease-centric notion of similarity:

- drug-centric: "if two drugs are chemically similar, they likely have a similar relationship with the target disease";
- disease-centric: "a drug has the same relation for similar diseases."

To obtain the explanation, they embed the drug and the disease from a DDA pair and retrieve k intermediate entities (drugs or diseases) using a cosine similarity-based metric. An explanation instance is created based on the relationship between the drug, disease, and the intermediate entity in existing publications. For example, if the intermediate entity is a drug and the intermediate drug treats the input disease, the input drug is assumed to also treat the input disease with confidence proportional to the embedding similarity. The obtained explanations using k intermediate entities are aggregated into the final DDA prediction, e.g., using a majority vote.

Huang et al. (2020) include an interpretable component in their drug-drug interaction (DDI) prediction system. The component projects the latent embedding of the input drug pair into a more interpretable subspace, whose basis consists of frequently occurring molecular substructures. The substructures are extracted from a database of drug representations by finding substrings with a high enough frequency. The projection into the subspace aims to capture the relevance of the molecular substructures toward the drug interaction prediction.

Yazdani-Jahromi et al. (2022) propose an attention-based drug-target interaction prediction system, using the attention weights as an explanation. They demonstrate the high predictive performance of their system on three benchmark datasets, while they demonstrate the interpretation capability of their model on a drug-target interaction prediction example via visualization.

Bradshaw et al. (2019) present a generator of product molecules from a set of common reactant molecules. It is composed of

- an encoder-decoder model between a latent space and a list of reactant molecules, and
- a reaction prediction model that transforms the reactants into a list of product molecules.

The second component introduces interpretability to the model as it provides some insight on *how* the product molecules are constructed out of the reactants. However, the authors do not put an emphasis on evaluating the interpretability of their approach.

Jiang et al. (2019) present an approach for detecting potential adverse medication effects from social media posts. The detection is posed as a word analogy task: given a known possible side effect of a drug, the task is to find similar pairs of drugs and corresponding side effects with a similar relation. The known possible side effects are taken from the SIDER database (Kuhn et al., 2016), while the static word embeddings are trained on unlabeled tweets. They found potential side effects are subject to human examination along with relevant tweets expressing the effect.

Rodríguez-Pérez and Bajorath (2020) present a usability study of the SHAP explanation method for explaining complex compound activity prediction models. They find that SHAP produces consistent feature attributions across three complex models. Additionally, they demonstrate how the obtained attributions can be used to find potential biases in the models.

Pope et al. (2019) present adaptations of three explanation methods for explaining graph convolutional neural networks: contrastive gradient-based saliency maps, class activation mapping, and excitation backpropagation. They test the methods on molecular graph classification, where the task is to predict whether molecules possess certain properties, such as toxicity. The explanations are salient subgraphs, which can be interpreted as functional groups responsible for the molecular property (according to the model). By analyzing the explanations using automated metrics (fidelity, contrastivity, and sparsity), the authors conclude that the gradient-weighted class activation mapping is the most suitable out of the tested methods, although they emphasize the need for detailed studies of chemical validity of the explanations in future work.

In summary, explanation methods have been adopted across a variety of pharmacology applications. We find that the authors typically use the explanation methods in one of two ways, using the explanations either as a safety mechanism for a semiautomatic use of the model predictions or as a way to obtain plausible hypotheses that are then manually verified, for example using additional experiments. The proposed explanation methods for pharmacology commonly use a connection to an external knowledge source. We believe that the incorporation of external knowledge into explanation methods is a promising direction for further research as the prediction may not be intuitively explainable to humans in terms of only input components. In addition, external

human-curated knowledge may naturally be more intuitive to end-users.

III. Common Natural Language Processing Tasks and Applications

Several NLP tasks are frequently tackled in the pharmacological context. Some of them are adapted from general NLP tasks (e.g., named entity recognition, relation extraction, and QA). In contrast, others are specific to pharmacology (e.g., adverse drug reactions and literature-based drug discovery). We have mentioned some successful uses of contextual BERT models on these tasks in Section II.A.3, but this mainly demonstrated the usability of these models. This section systematically analyzes the most important tasks in life sciences and pharmacology. As hundreds of works tackle these problems exclusively or among other problems, we review a sample of recent works. The overview is presented in Table 4.

A. Named Entity Recognition for Pharmacology

NER—called entity identification, entity chunking, or entity extraction—is one of the most popular NLP techniques that classifies named entities in text into predefined categories such as person, time, location, organization, and so on. In the biomedical context, the entities of interest can be cells, genes, gene sequences, proteins, biologic processes and pathways, diseases, drugs, drug targets, compounds, adverse effects, metabolites, tissues, and organs (Perera et al., 2020; Bonner et al., 2021). NER is often used as the initial stage of analyses to provide semantic interpretations of unstructured text by identifying and categorizing concept references. Various concepts are detected with different degrees of difficulty. The critical issue in recognizing chemicals, for example, is the high variance in concept names and chemical formulas. In contrast, the main challenge in identifying gene functions is the high degree of uncertainty caused by species diversity.

In pharmacology domain, NER is often used as the first step of the relation extraction task (see Section III.B) (Kadir and Bokharaeian, 2013; Gu et al., 2016) or adverse drug reactions task (see Section III.C) (Li et al., 2018). Many authors start with the MADE 1.0 challenge dataset, e.g., Jagannatha et al. (2019) find the medications and their attributes, Chapman et al. (2019) apply the conditional random field method for medication recognition, Yang et al. (2019a) developed the MADEx model based on LSTM networks for the same purpose, and Wunnava et al. (2019) apply the Bi-LSTM model.

B. Relation Extraction for Pharmacology

The relation extraction task is part of information extraction and extracts semantic relationships from texts. The extracted relationships connect two or more entities of the same kind that fit into one of many semantic categories (e.g., people, organizations, or places). Frequently, extracted relations are related to ADR and DDI, relations between medications, between their attributes such as dosage, route, frequency, and duration (Jagannatha et al., 2019). The ability of NLP models to automatically detect adverse drug event (ADE) related terms in textual data helps avoid ADEs. This results in safer and better quality health care services, lower health care expenditures, more educated and engaged customers, and improved health outcomes.

In pharmacology, relation extraction typically processes scientific papers that provide novelties from the pharmacology. Classic approaches extracted semantic relationships with a pattern-based approach to find medical relations in pharmaceutical texts (Rosario and Hearst, 2004; Ben Abacha and Zweigenbaum, 2011). Deep learning approaches brought significant improvements (Li et al., 2018; Yang et al., 2019a). Lately used approaches apply pretrained language models, e.g., SemRep (Kilicoglu et al., 2020). The extracted information is sometimes used to construct graphs encoding

TABLE 4
Overview of tasks related to the pharmacology

Task	Description	Referenced papers
Named Entity Recognition	Identifying pharmaceutical entities in textual data	(Jagannatha et al., 2019) (Chapman et al., 2019) (Wunnava et al., 2019) (Gu et al., 2016) (Kadir and Bokharaeian, 2013) (Li et al., 2018) (Yang et al., 2019a)
Relation Extraction	Finding relation between drugs and diseases from scientific text resources	(Ben Abacha and Zweigenbaum, 2011) (Li et al., 2018) (Chen et al., 2019) (Kilicoglu et al., 2020) (Yang et al., 2019a) (Rosario and Hearst, 2004) (Zhou et al., 2020)
Adverse Drug Reactions	Anticipate danger from future administration and demand avoidance, particular therapy, or dose regimen modification	(A. Breden and L. Moore, preprint, DOI: https://doi.org/10.48550/arXiv.2005.06634) (Li et al., 2020b) (Hussain et al., 2021) (Li et al., 2018) (Wunnava et al., 2019) (Chapman et al., 2019)
Literature Based Drug Discovery	Discovering new pharmacological information from existing literature	(Zhou et al., 2017) (Biswas et al., 2021) (Wei et al., 2019) (X. Wang et al., preprint, DOI: https://doi.org/10.48550/arXiv.2003.1221) (Pinto et al., 2020) (Martinc et al., 2020) (Sang et al., 2018) (Jofche et al., 2023) (Preiss et al., 2015) (Wang et al., 2017) (Dobreva et al., 2020) (Xue et al., 2018) (Gottlieb et al., 2011)
Question Answering	Answers given question with the most relevant response	(Su et al., 2020) (Lee et al., 2020) (Farrar, 2002) (Veisi and Shandi, 2020) (Marginean, 2014)

drug-drug and disease-drug relationships, representing the similarity between them (Zhou et al., 2020). Although most approaches are based on textual data, relations are also discovered through the analysis of EHR data (Chen et al., 2019).

C. Adverse Drug Reactions

ADR is defined as a considerably damaging or unpleasant reaction occurring from an intervention associated with the use of a pharmaceutical product. Adverse reactions frequently anticipate danger from future administration and demand avoidance, particular therapy, or dose regimen modification (Pirmohamed et al., 1998). ADRs have traditionally been divided into two categories. Type A responses are dose-dependent and predicted based on the drug's pharmacology (also known as enhanced reactions). In contrast, Type B responses, often known as weird reactions, are distinctive and unpredictable from the pharmacological point of view.

Implementation-wise, ADR extraction is similar to relation extraction, where ADRs connected to various diseases and drugs are detected. Lately, large pretrained language models, such as BERT, are used in ADR extraction (A. Breden and L. Moore, preprint, DOI: https://doi.org/10.48550/arXiv.2005.06634) (Li et al., 2020b; Hussain et al., 2021). Again, texts are not the sole source of information, and EHRs are often used as additional information in ADR extraction (Li et al., 2018; Chapman et al., 2019; Wunnava et al., 2019).

D. Literature-Based Drug Discovery

LBD is an automatic or semiautomatic method for discovering new information from the literature. The amount of scientific literature is steadily growing, driving researchers to become more specialized and making it challenging to track developments even in restricted fields (Henry and McInnes, 2017). If text is identified that overtly asserts the knowledge that "A is associated with B" and "B is associated with C" in the Swanson ABC co-occurrence model (Swanson and Smalheiser, 1997), then the implicit knowledge of "A may be associated with C" is obtained. LBD is essential for biomedical NLP since it allows finding implicit information that can help to enhance biomedical research. A recent study presents the computational strategies used for LBD in the biomedical area (Gopalakrishnan et al., 2019)

LBD applies several NLP tasks to process the pharmacological and medical literature, with the purpose to detect new medical entities (Sang et al., 2018; Dobreva et al., 2020; X. Wang et al., preprint, DOI: https://doi.org/10.48550/arXiv.2003.1221), extract relations (Preiss et al., 2015; Wang et al., 2017), or reactions (Zhou et al., 2017). Some approaches use scientific texts for protein engineering and visualization (Biswas et al., 2021). Frequent information source is the PubMed engine together with the PubTator model (Wei et al., 2019) for automated annotation. The PharmKE tool (Jofche et al., 2023) labels pharmaceutical entities and the relationships between them. In new diseases, such as COVID-19, the LBD technique has proven useful to extract relevant information (Martinc et al., 2020; Pinto et al., 2020). Another frequent task is drug repositioning, which helps to find another purpose for existing drugs, i.e., to use them in treating similar diseases (Xue et al., 2018). Alternatively, novel drug indications can be discovered by analyzing the medical history, as exemplified in the PREDICT model (Gottlieb et al., 2011).

E. Question Answering

QA is an NLP task that takes a question as input and returns an answer in the form of a ranked list of relevant replies or a summary answer snippet (Coleman and Coleman, 2005). In a classic (preneural) approach, QA incorporates three tasks: information retrieval, retrieving relevant documents or passages for a particular query, and text summarization that summarizes the reply from relevant passages. A related information retrieval task is called "Learning by Doing" and searches the knowledge base for entities most related to the ones mentioned in the question. This task is divided into ranking the texts found in the database and finding the correct answer among the recovered paragraphs.

QA can summarize the pharmacological literature, e.g., for new diseases like COVID-19 (Su et al., 2020). The data are mainly from PubMed articles and, in the case of COVID-19, also news about this disease (Lee et al., 2020). To answer pharmaceutical questions, the QA task can be applied in many languages, even in low-resource languages such as Persian (Veisi and Shandi, 2020). Another source of information can be linked data as used in the GFMed model (Marginean, 2014).

IV. Data Resources

As the application of open science and open data principles is rising (Burgelman et al., 2019), the number of publicly available datasets is steadily growing. This makes finding and discovering appropriate datasets increasingly challenging. There are two strategies to find a dataset suitable for a given task. First, a bottom-up approach starts by searching available datasets and evaluating their utility for the given problem. Second, a top-down approach first finds relevant papers for the tackled topic and then explores the available datasets used in the papers.

We first present an overview of specialized search engines for discovering and finding datasets in Section IV.A. Then we give an overview of the most important datasets used in published papers related to NLP in pharmacology. The covered datasets are organized into five groups: patient data, drug usage data, drug structure data, QA datasets, and general pharmacological data. In Section IV.B, we present datasets containing patients' history and medical notes. The datasets in

Section IV.C contain drug characteristics according to the prescriptions to patients, while in Section IV.D, we cover datasets with information about drugs' chemical composition. Datasets supporting QA systems in pharmacology are described in IV.E. Section IV.F describes general resources useful for successful NLP in pharmacology.

We include public and closed (private/commercial) data in the survey. The summary of datasets is contained in Table 5, where for each dataset, we include a list of references where the dataset was used, a short description, the size of the dataset, and its typical usage.

A. Finding and Discovering Datasets

As the number of datasets rapidly grows, it becomes essential to have effective tools for finding them. As a solution, there are several specialized search engines for discovering and finding datasets.

Google's Dataset Search (https://datasetsearch.research.google.com/) currently indexes more than 30 million publicly available datasets. Filters can limit the results based on licensing (free or premium), format (comma-separated values, images, etc.), and update time. Alternatively, a specialized cloud platform data.world (https://data.world/) hosts an enterprise data catalog with more than 130,000 datasets and knowledge graphs. Another platform hosting public datasets is Kaggle (https://www.kaggle.com/datasets), which is primarily a ML competition platform, but it also includes a dataset search engine.

The NLP community usually publishes the source code and datasets in the Github (https://github.com/) repository so that this source control platform can be used for dataset discovery. A specialized platform indexing the code and data related to research papers is Papers with Code (https://paperswithcode.com/datasets). This platform offers research area-based organization of papers allowing for a convenient discovery and browsing of papers and datasets. One of the most popular development platforms for NLP, the Huggingface, offers a good dataset search engine organized by NLP task, category, language, size, and license (https://huggingface.co/datasets).

A specialized search engine for linked data is the Linked Open Data (LOD) Cloud (https://lod.openlinksw.com/) that allows for text-based search and entity lookup. LOD Cloud is a distributed web of interconnected datasets (more than 1500 datasets) containing open data in a structured and semantically annotated format from multiple domains—life sciences, publications, government, media, and so on. The background on the LOD Cloud is described in Section V.

B. Patient Data

Datasets with the information about patients typically contain patients' medical history or medical notes about them. The main application of these datasets is to find novel relations between drugs and diseases. Next we briefly describe the most commonly used patient datasets.

MIMIC-III (Johnson et al., 2016) (https://mimic.mit.edu/) is a dataset that contains data on patients hospitalized in large tertiary care hospitals in critical care units. It contains information on vital signs, medicines, laboratory measurements, care providers' observations and notes, fluid balance, procedure codes, diagnostic codes, imaging reports, hospital length of stay, survival statistics, and more. This dataset contains data on more than 40 000 patients.

MADE1.0 Database (Jagannatha et al., 2019) (https://bio-nlp.org/index.php/announcements) is an EHR database that is a part of the MADE1.0 competition. The structured dataset contains information on taken drugs, experienced ADEs, and indications and symptoms of patients. The competition addressed three tasks: NER, relation identification, and a joint NER-RI task. The dataset contains 1089 patient notes with detailed named entity and relation annotations.

n2c2 NLP Research Database (Henry et al., 2020) (https://portal.dbmi.hms.harvard.edu/projects/n2c2-nlp/) is database used for Track 2 of the 2018 National NLP Clinical Challenges shared task. The data is extracted from the MIMIC-III clinical care database. The records were chosen using a query that looked for ADEs in the description of records' International Classification of Diseases code. The retrieved records were manually inspected to ensure that at least one ADE was present and adequately annotated. The dataset contains 505 discharge summaries in textual format.

MarketScan (Adamson et al., 2008) (https://www.ibm.com/products/marketscan-research-databases) dataset is a collection of administrative claims databases that includes information on in-patient and out-patient claims, out-patient prescription claims, clinical usage records, and healthcare costs in United States. The three main databases each contain a convenience sample for one of the following patient populations: (1) employees with contributing employers' health insurance, (2) Medicare beneficiaries with employer-paid supplemental insurance, and (3) Medicaid recipients in 1 of 11 participating states. The data are not in textual format but can be used with NLP applications. The database contains data on approximately 43.6 million persons.

C. Drug Usage Data

Datasets described in this section provide information on drugs' usage, usage instructions, effects, pharmaceutical properties, and composition.

DailyMed Database (National Institutes of Health., 2014) (https://dailymed.nlm.nih.gov/dailymed/index.cfm) is a web database provided by the National Library of Medicine in the United States. The US Food and Drug Administration updates the material daily. The DailyMed contains prescription and nonprescription medications for human and animal usage, medical gases,

TABLE 5
Different types of pharmacology-relevant datasets

Name	Description	Entries	Usage
Patient Data			
MarketScan (Adamson et al., 2008) https://www.ibm.com/products/marketscan-research-databases	Collection of administrative claims	43,600,000	NER, ADE, DDI
MIMIC-III (Johnson et al., 2016) https://mimic.mit.edu/	Data on patients hospitalized	40,000	Drug discovery, ADE, DDI
MADE 1.0 (Jagannatha et al., 2019) https://bio-nlp.org/index.php/announcements	A challenge dataset with 21 EHRs of cancer patients	1089	NER, ADE
n2c2 (Henry et al., 2020) https://portal.dbmi.hms.harvard.edu/projects/n2c2-nlp/	Unstructured notes from the Research Patient Data	505	ADE
Drug Usage Data			
DailyMed (National Institutes of Health , 2014) https://dailymed.nlm.nih.gov/dailymed/index.cfm	Drug label database	142,981	NER, DDI, ADE
DrugBank (Wishart et al., 2018) https://go.drugbank.com/	Database of drugs and drug products	14,665	ADE, pharmacovigilance, standardization, interactions
Drug Structure Data			
ChEMBL (Gaulton et al., 2012) https://www.ebi.ac.uk/chembl/	Binding, functional, and ADMET data	2.4 million	ADE, pharmacovigilance, standardization, interaction
UMLS (Bodenreider, 2004) http://umlsks.nlm.nih.gov	Biomedical vocabularies	2 million	ADE
PDB (Protein Data Bank Contributors, 1971) http://www.rcsb.org/pdb/	biologic macromolecules	133,920	ADE, pharmacovigilance, standardization, interaction
ChemProt (Taboureau et al., 2011) https://biocreative.bioinformatics.udel.edu/news/corpora/chemprot-corpus-biocreative-vi/	Biologic annotations	1820	ADE
Question Answering Data			
MQP (McCreery et al., 2020) https://github.com/curai/medical-question-pair-dataset	Collection of medical related pairs of questions and answers	3048	QA
COVID-Q (Wei et al., 2020) https://paperswithcode.com/dataset/covid-q	Collection of COVID-19-related questions divided into 15 general categories and 207 specific question classes	1690	QA
CovidQA (Zhao et al., 2020) https://aclanthology.org/2020.nlpcovid19-acl.18/	Collection of question–article–answer triplets taken from 85 different articles in CORD-19	124	QA
General Pharmacological Data			
Wikipedia (Wikipedia, 2004) https://en.wikipedia.org/	Online free encyclopedia	15 billion	ADE, DDI, drug discovery, NER
PubMed (Canese and Weis, 2013) https://pubmed.ncbi.nlm.nih.gov/	Web engine for searching health articles	30 million	ADE, DDI, drug discovery, NER
LitCovid (Chen et al., 2021a) https://www.ncbi.nlm.nih.gov/research/coronavirus/	Scientific PubMed articles related with COVID-19	255,935	ADE, DDI, drug discovery, NER
CORD-19 (L.L. Wang et al., preprint, DOI: https://doi.org/10.48550/arXiv.2004.10706) https://www.kaggle.com/datasets/allen-institute-for-ai/CORD-19-research-challenge	Scientific papers relevant to COVID-19 research	52,000	ADE, DDI, drug discovery, NER
DBpedia (Auer et al., 2007) https://www.dbpedia.org/	Articles and structured data on e.g., drugs and diseases	10,000	ADE, pharmacovigilance, standardization, interactions
EMBASE http://www.embase.com	Biologic and pharmacological bibliographic database	32 million	ADE, pharmacovigilance, standardization, interactions
ClinicalTrials.gov (Zarin et al., 2011) https://www.clinicaltrials.gov/	Clinical trials database	329,000	ADE, pharmacovigilance, standardization, interactions

ADMET, absorption, distribution, metabolism, excretion, and toxicity; PBD, protein data bank.

gadgets, cosmetics, nutritional supplements, and medical foods. The labeled drugs describe the composition, form, packaging, and other properties of drug products according to the HL7 Reference Information Model. These details are given in the descriptive text format. The database contains 142,981 labels.

DrugBank Database (Wishart et al., 2018) (https://go.drugbank.com/) is one of the largest drug databases. Besides drugs, it contains drug paths that show how the drug travels in the human body and allows search for indications and drug targets. For an individual drug, the database contains all the brand names, background information in the text form, its type, structure, weight, formula, other names it is called by, what it is used for, what therapies it is used in, indications, doses, interactions, and more. All the details for each drug are available online and given in the descriptive text format. The database contains descriptions of 14,665 drug entries.

D. Drug Structure Data

Datasets covered in this section contain drug characteristics regarding their chemical composition. Mainly, they are used for discovering new drugs or finding protein-protein interactions between drugs.

ChEMBL Database (Gaulton et al., 2012) (https://www.ebi.ac.uk/chembl/) is an open-source database that contains binding, functional, and chemical absorption, distribution, metabolism, excretion, and toxicity data for a wide range of drug-like bioactive chemicals. These data are regularly manually extracted from the published literature, then selected and standardized to enhance their quality and usability across a variety of chemical biology and drug-discovery research uses. The database includes 2.4 million bioassay measurements spanning 622,824 chemicals, including 24,000 natural products. The contents were produced by sifting through over 34,000 papers published in 12 medicinal chemistry journals. The data from the journals containing details can also be used.

UMLS (Bodenreider, 2004) (http://umlsks.nlm.nih.gov) is a database of biomedical vocabularies. The National Center for Biotechnology Information (NCBI) taxonomy, Gene Ontology, Medical Subject Headings, Online Mendelian Inheritance in Man, and the Digital Anatomist Symbolic Knowledge Base are all included in the UMLS MetaThesaurus. The UMLS is not a textual database but is frequently used in NLP tasks, such as extracting concepts, relationships, or knowledge of pharmacological entities from texts. The UMLS has about 2 million names for more than 900,000 concepts from more than 60 biomedical vocabularies and 12 million relationships between them.

PDB: The Protein Data Bank Database (Protein Data Bank Contributors, 1971) (http://www.rcsb.org/pdb/) is a global repository of structural data for biologic macromolecules. To obtain the data, depositors used X-ray crystal structure determination, nuclear magnetic resonance, cryo-electron microscopy, and theoretical modeling. The search queries also return the literature from which the data are extracted, e.g., the abstracts from medical articles that can be further used for NLP. The number of papers accessible in the textual format is not available, but the database contains 133,920 Biologic Macromolecular Structures, each accompanied by a related abstract.

ChemProt Database (Taboureau et al., 2011) (https://biocreative.bioinformatics.udel.edu/news/corpora/chemprot-corpus-biocreative-vi/) is a biology annotated database based on several chemical-protein annotation resources, together with disease-associated protein-protein interactions. ChemProt was used in the BioCreative VI text mining chemical-protein interactions shared task. The data contains PubMed abstracts in textual format together with annotated entities and interactions. The database has 1820 abstracts.

E. Question-Answering Data

This section covers some datasets that can be used to build pharmacological QA models.

MQP Database (McCreery et al., 2020) (https://github.com/curai/medical-question-pair-dataset) comprises 3048 question-answer pairs that are categorized as similar or distinct by medical experts (i.e., not particular to COVID-19). Two doctors collaborated on the annotation and their agreement on 836 question pairings in the test set was above 85 percent.

COVID-Q Database (Wei et al., 2020) (https://paperswithcode.com/dataset/covid-q) is a collection of 1690 COVID-19-related questions divided into 15 general categories and 207 specific question classes. The dataset was annotated in three stages by many curators. First, two curators discussed and categorized the questions. Second, an external curator reviewed the work and, if necessary, proposed adjustments to the categories. Third, questions from more than four different question classes were sampled and allocated to three different Amazon Mechanical Turk workers. The validation was based on the majority vote.

CovidQA Database (Zhao et al., 2020) (https://aclanthology.org/2020.nlpcovid19-acl.18/) is made up of 124 question–article–answer triplets taken from 85 different articles in COVID-19 Open Research Database (CORD-19) Kaggle challenge and covers 27 different categories. Five curators created annotations by synthesizing questions from the challenge organizers' categories, then manually discovered relevant articles and replies.

F. General Pharmacological Data

In this section, we describe five resources that are general and useful for many tasks.

Wikipedia (Wikipedia, 2004) (https://en.wikipedia.org/) is a well-known encyclopedia and web-based collaborative database consisting of more than 15 billion articles. Wikipedia contains articles from different scientific fields written in many languages.

PubMed (Canese and Weis, 2013) (https://pubmed.ncbi.nlm.nih.gov/) is a free web engine for primarily MEDLINE, bibliographic database encompassing medicine, nursing, dentistry, veterinary medicine, the health care system, and preclinical sciences like molecular biology. More than 4600 biomedical journals are indexed in MEDLINE, together with bibliographic citations and author abstracts. PubMed indexes more than 30 million articles and abstracts.

LitCovid Database (Chen et al., 2021a) (https://www.ncbi.nlm.nih.gov/research/coronavirus/) is a curated literature site for tracking up-to-date scientific knowledge regarding the COVID-19 disease. It is the most comprehensive resource on the topic with central access to more than 255,935 relevant PubMed articles. The articles are updated daily and divided into categories based on research themes and geographical areas.

CORD-19 (L.L. Wang et al., preprint, DOI: https://doi.org/10.48550/arXiv.2004.10706) (https://www.kaggle.com/datasets/allen-institute-for-ai/CORD-19-eresearch-challenge) contains metadata about papers related to COVID-19. The main sources are PubMed, World Health Organization, bioRxiv, and medRxiv. This database contains more than 52,000 papers.

DBpedia (Auer et al., 2007) (https://www.dbpedia.org/) is a structured open-source database with information extracted from Wikipedia articles. For drugs, it contains basic information on uses, contained chemicals, drug type, links to other languages, Wikipedia links, and other links used to extract information. The database contains more than 10,000 drug type entries.

Excerpta Medica database (EMBASE) (http://www.Embase.com) is a biologic and pharmacological bibliographic database of published literature. It was created to assist information managers and pharmacovigilance in adhering to the regulatory requirements of a licensed medicine. The EMBASE database, created in 1947, contains more than 32 million entries from more than 8500 published journals.

ClinicalTrials.gov (Zarin et al., 2011) (https://www.clinicaltrials.gov/) is a clinical trial registry and the biggest clinical trials database. It is managed by the National Institutes of Health and contains registrations for more than 329,000 studies from 209 countries.

V. Knowledge Graphs

The concepts of linked data and knowledge graphs introduced new standards for representing, storing, and retrieving data over the web, both publicly and privately (Bizer et al., 2008, 2009; Heath and Bizer, 2011; Wood et al., 2014; Hogan et al., 2021). As a result of years of adoption of the linked data principles by various data publishers, the LOD Cloud (https://lod-cloud.net) has been created and populated with 1541 interlinked datasets from the domains of geography, government, life sciences, linguistics, media, publications, social networking, user-generated, and cross-domain.

Knowledge graphs, the latest trend in the semantic web and linked data, enable the generation, consolidation, and contextual linking of structured data. The standards and technologies for knowledge graphs solve the problem of having separate "data silos" in traditional relational database systems, which have to be explicitly mapped to other isolated databases to take advantage of interconnected data (Jovanovik and Trajanov, 2017).

The pharmaceutical industry is leading in using knowledge graph-based NLP techniques, especially in patient disease identification, clinical decision support systems, and pharmacovigilance (Dumitriu et al., 2021). The problem of identifying patients with specific diseases can be mitigated by knowledge graphs generated from structured and unstructured data from medical records, which capture explicit disease–symptom relationships (Chen et al., 2019). Recently, knowledge graphs improved the classification of rare-disease patients (Li et al., 2019). In the area of clinical decision support, the combination of NLP and knowledge graphs is employed in inferring drug-related knowledge that is not immediately observed in data, inferring cuisine-drug interactions based on knowledge graphs of drugs and recipes, improving user interaction with relevant medical data, and so on (Jovanovik et al., 2015a; Goodwin and Harabagiu, 2016; Liu et al., 2018; Xia et al., 2018, Ruan et al., 2019; F. Xia et al., preprint, DOI: https://doi.org/10.48550/arXiv.2204.09220). In pharmacovigilance, the struggles of NLP engines to understand complex language components (e.g., negation, doubt, historical medical statements, family medical history) from individual case study reports have been significantly mitigated with the use of knowledge graphs (Perera et al., 2013). Other examples include the use of knowledge graphs to improve NLP pipelines for detecting medication and ADEs from EHRs (Ngo et al., 2018), as well as from Medline abstracts (Yeleswarapu et al., 2014).

Section V.A presents several knowledge graphs from the biomedical domain used in the mentioned application areas. Given the ongoing COVID-19 pandemic, we outline several recent COVID-19-related knowledge graphs in Section V.B.

A. Biomedical Knowledge Graphs

Several projects worked on the transformation of pharmacology-related and health care data into linked data and knowledge graphs. Currently, 341 life science datasets are present in the LOD Cloud. These datasets contain health care data from various subdomains, such as drugs, diseases, genes, interactions, clinical trials, enzymes, and more. The most notable of them are presented next and are outlined in Table 6.

Bioportal (Whetzel et al., 2011) (https://bioportal.bioontology.org/) project hosts ontologies covering drugs,

diseases, genes, clinical procedures, and others. With more than 980 biomedical ontologies, which define a total of more than 13.9 million classes, it represents the largest such repository in the life science domain.

Bio2RDF (Callahan et al., 2013a) (https://bio2rdf.org) is an open-source project that creates Resource Description Framework (RDF) datasets from various life science resources and databases and interconnects them into one network (Belleau et al., 2008; Callahan et al., 2013a,b). The latest release of Bio2RDF contains around 11 billion triples, which are part of 35 datasets. These datasets contain various healthcare data: clinical trials (ClinicalTrials), drugs (DrugBank, LinkedSPL, NDC), diseases (Orphanet), bioactive compounds (ChEMBL), genes (GenAge, GenDR, GOA, HGNC, HomoloGene, MGD, NCBI Gene, Online Mendelian Inheritance in Man, PharmGKB, SGD, WormBase), proteins (InterPro, iProClass, iRefIndex), gene-protein interactions (CTD), biomedical ontologies (BioPortal), side effects (SIDER), terminology (Resource Registry, Medical Subject Headings, NCBI taxonomy), mathematical models of biologic processes (BioModels), publications (PubMed), and more.

Macedonian drug data is drug data from the Health Insurance Fund of North Macedonia that has been transformed into a knowledge graph and linked to other LOD Cloud datasets (Jovanovik et al., 2013). This knowledge graph was further extended with linked data about Macedonian medical institutions and drug availability lists from pharmacies (Jovanovik et al., 2015b).

Cuisine-drug interactions is a project that used two knowledge graphs for analysis of connections between drugs and their interactions with food, and recipes from different national cuisines, resulting in findings that uncovered the ingredients and cuisines most responsible for negative food-drug interactions in different parts of the world (http://viz.linkeddata.finki.ukim.mk) (Jovanovik et al., 2015a).

The global drug data is a research project that is a pipeline-based platform created to collect, clean, align, consolidate, and create a publicly available knowledge graph of drug products registered in various countries (http://drugs.linkeddata.finki.ukim.mk) (Jovanovik and Trajanov, 2017). The source of the data are the official country drug registers. The generated RDF knowledge graph is publicly available through a web-based app (http://godd.finki.ukim.mk).

B. COVID-19 Knowledge Graphs

The COVID-19 pandemic turned the attention of many researchers to life sciences and health care domains. Next we list some recent COVID-19-related knowledge graphs.

TypeDB Bio (Covid) knowledge graph (https://github.com/typedb-osi/typedb-bio) contains data extracted from COVID-19 papers and from datasets on proteins, genes, disease-gene associations, coronavirus proteins, protein expression, biologic pathways, and drugs. For instance, it allows querying for specific viruses giving associated human proteins related to the virus (e.g., a protein that helps in the replication of the virus). From here, it is possible to identify drugs that inhibit the detected proteins, meaning they can be prioritized in research as potential treatments for patients with the virus. To check the plausibility of this association and the implications, the graph can be used to identify relevant papers in the COVID-19 literature where this protein has been studied.

Covid-19-DS (Pestryakova et al., 2022) (https://dice-research.org/COVID19DS) is an RDF knowledge graph of scientific publications. The base of the graph is the CORD-19 dataset (L.L. Wang et al., preprint, DOI: https://doi.org/10.48550/arXiv.2004.10706) that is regularly updated. The graph generation pipeline applies NER, entity linking, and link discovery to the CORD-19 data. The current version of the resulting graph contains more than 69 000 000 RDF triples and is linked to 9 other datasets with more than 1 million links.

KG-Covid-19 (Reese et al., 2021) (https://github.com/Knowledge-Graph-Hub/kg-covid-19/wiki) is a framework that allows users to download and transform COVID-19 related datasets and generate a knowledge graph that can be used in ML. The project also provides access to prebuilt knowledge graphs along with public querying.

VI. Tools and Libraries

This section focuses on the technical part of NLP applications in pharmacology. In Section VI.A, we cover software libraries and tools that help to build machine learning models for the tasks mentioned in Sections II and III. For each library, we also mention its recorded use in pharmacology. In Section VI.B, we present general text processing libraries. Most covered libraries and tools are accessible as Python packages. Table 7 gives an overview.

TABLE 6
Covered knowledge graphs from the biomedical domain and their characteristics

Name	Unique Entities	RDF Statements
Bio2RDF (Callahan et al., 2013a)	1,107,871,027	11,895,348,562
HIFM (Jovanovik et al., 2013, 2015b)	3,000	21,233
LinkedDrugs (Jovanovik and Trajanov, 2017)	248,746	99,235,032
Covid-19-DS	262,954	69,434,763
KG-Covid-19 (Reese et al., 2021)	574,778	24,145,556

A. Machine Learning Libraries

Natural Language Toolkit (NLTK) (Bird et al., 2009) (https://www.nltk.org/) is one of the most powerful and popular NLP libraries. NLTK is a suite of open-source Python modules, data sets, and tutorials on language processing. The toolkit consists of baseline text processing such as sentence splitting, tokenization, and part of speech (POS) tagging. These tools may help in NER, to identify known medications, to detect ADEs (Chapman et al., 2019), or in evaluation of entity indicators for relation extraction (Qin et al., 2021).

MetaMap Transfer (Aronson, 2001) (https://github.com/theislab/MetaMap) is an extensively used, Java-based, NER tool that maps biomedical free-form text to UMLS Metathesaurus concepts. In the process of creating the first DDI corpus that, besides drugs, contains pharmacokinetic DDIs and pharmacodynamic DDIs, the UMLS MetaMap Transfer tool preannotates the documents with pharmacological substance entities; i.e., it is used to parse the documents to automatically recognize drug types (Herrero-Zazo et al., 2013). MetaMap's intrinsic function—identification of medical concepts—was used for extracting drug indication information from structured product labels (Fung et al., 2013).

CRFsuite (Okazaki, 2007) (http://www.chokkan.org/software/crfsuite/) implements the Conditional Random Fields ML algorithm for labeling sequential data. It is used for NER in the MADEx system for detecting medications and ADEs and their relations from clinical notes (Yang et al., 2019a).

Library for Support Vector Machines (Chang and Lin, 2011) (https://www.csie.ntu.edu.tw/∼cjlin/libsvm/) is an open-source package that implements the sequential minimal optimization algorithm for kernelized support vector machines, supporting both classification and regression. The library was used to classify relation types in the MADEx system (Yang et al., 2019a).

Stanford CoreNLP toolkit (Manning et al., 2014) (https://stanfordnlp.github.io/CoreNLP/) was initially developed for English but now supports German, French, Arabic, Chinese, and Spanish. The Stanford CoreNLP toolkit is a pipeline of NLP Java tools for linguistic annotations, such as tokenization, sentence splitting, POS tagging, morphologic analysis, NER, syntactic parsing, and coreference resolution. In pharmacology, CoreNLP was applied in a joint model for entity and relation extraction from biomedical text, providing POS tagging and dependency parsing (Li et al., 2017).

BRAT annotation tool (Stenetorp et al., 2012) (https://brat.nlplab.org/introduction.html) is an online environment for annotating structured text, i.e., notes in a predefined form. The tool was used to create a corpus from Twitter messages and PubMed sentences to understand drug reports better (Alvaro et al., 2017).

SpaCy library (Honnibal et al., 2020) (https://spacy.io/) is a free, open-source library for NLP. It contains ML models for NER, POS tagging, dependency parsing, sentence segmentation, text classification, entity linking, morphologic analysis, and so on. The library is employed for entity recognition for Pharmaceutical Organizations and Drugs in PharmKE—a text analysis platform focused on the pharmaceutical domain (Jofche et al., 2023).

Document Metadata Exchange Organizer (DOMEO) Annotation Toolkit (Ciccarese et al., 2011) (https://github.com/domeo/domeo) (also called SWAN Annotation Tool) is a web application enabling users to manually, semiautomatically, or automatically create ontology-based annotation metadata. DOMEO can be customized with additional plugins, e.g., for annotation of PDDI mentions in structured product labels (Hochheiser et al., 2016) (https://github.com/rkboyce/DomeoClient).

Transformers—Hugging Face (Wolf et al., 2020) (https://huggingface.co/) package contains many state-of-the-art NLP models, such as BioBERT (Lee et al., 2019), RoBERTa (Y. Liu et al., preprint, DOI: https://doi.org/10.48550/arXiv.1907.11692), CharacterBERT (El Boukkouri et al., 2020), and more. The package offers also tokenizers for several languages and tasks, as well as some popular datasets for NLP tasks such as NER, NLI, QA, and so on.

Medical Concept Annotation Tool (MedCat Tool) (Kraljevic et al., 2021) (https://github.com/CogStack/MedCAT) is an open-source tool that uses unsupervised methods for NER and NEL in the biomedical field. The tools were validated with the MIMIC-III program and MedMentions (biomedical papers annotated with mentions from critical care databases). Dobreva et al. (2022) highlighted drug entities with the help of this tool in the process of extracting drug-disease relations and drug effectiveness.

AllenNLP (Gardner et al., 2018) (https://allenai.org/allennlp) is an open-source research library, built on PyTorch, for developing deep learning models for a wide variety of linguistic tasks. The PharmKe (Jofche et al., 2023) model uses AllenNLP for NER of drugs and pharmaceutical organizations that appear in texts.

Flair (Akbik et al., 2019) (https://github.com/flairNLP/flair) is a simple yet powerful framework for NLP, such as NER, POS tagging, and text classification. The framework supports training new models and is used in many research projects and industrial applications; e.g., Sun et al. (2021) use FLAIR to find subword embeddings.

Gensim (Řehůřek and Sojka, 2010) (https://radimrehurek.com/gensim/) is a Python library for topic modeling—extraction of unknown topics from a large volume of text (feeds from social media, customer reviews, user feedback, e-mails of complaints, and so on), document indexing, and similarity retrieval from large corpora. The library can handle large text files without having to

TABLE 7
Commonly used machine learning and NLP software libraries and tools

Name	Usage	Referenced Papers
Natural Language Toolkit (NLTK) (Bird et al., 2009) https://www.nltk.org/	tokenization, lemmatization, POS tagging, NER, word similarity	(Segura-Bedmar and Martínez, 2017) (Khadhraoui et al., 2022) (Jagannatha et al., 2019) (Liu et al., 2019b) (Aldahdooh et al., 2021) (Chen et al., 2020) (Li et al., 2020a) (Chapman et al., 2019) (Bird et al., 2009) (Turina et al., 2021) (Sivasankari et al., 2017) (Prabadevi et al., 2019) (Mahatpure et al., 2019) (Rabhi et al., 2019) (Ren, 2021) (Romasanta et al., 2020) (Sjögren et al., 2020) (Raghupathi et al., 2018)
MetaMap Transfer tool (MMTx) (Aronson, 2001) https://github.com/theislab/MetaMap	NER, DDI	(Schriml et al., 2012) (Aronson, 2001) (Ben Abacha and Zweigenbaum, 2011) (Fung et al., 2013) (Gottlieb et al., 2011) (Sang et al., 2018) (Preiss et al., 2015) (Kilicoglu et al., 2020) (Yang et al., 2011) (Jagannatha et al., 2019) (Yang et al., 2019a) (Perera et al., 2020) (Kamp et al., 2013) (Mattes et al., 2013) (Chiaramello et al., 2016) (Jiang and Zheng, 2013)
CRFsuite library (Okazaki, 2007) http://www.chokkan.org/software/crfsuite/	NER, drug discovery, ADE	(Pyysalo et al., 2013) (Chapman et al., 2019) (Yang et al., 2019a) (Habibi et al., 2017) (Bamburová and Neverilová, 2019) (Hakala and Pyysalo, 2019) (Soysal et al., 2018) (Ngo et al., 2018) (Liu et al., 2015)
LibSVM (Chang and Lin, 2011) https://www.csie.ntu.edu.tw/~cjlin/libsvm/	Classification, regression.	(Yang et al., 2019a) (Shan and Song, 2019) (Kumari et al., 2010) (Yesmin, 2016) (Huang and Li, 2004)
Stanford CoreNLP toolkit (Manning et al., 2014) https://stanfordnlp.github.io/CoreNLP/	Tokenization, lemmatization, POS tagging, NER, word similarity	(Yang et al., 2019a) (Wang et al., 2018) (Dernoncourt and Lee, 2017) (Li et al., 2020a) (Li et al., 2017) (Tang et al., 2019) (Filannino and Uzuner, 2018) (Gu et al., 2016) (Kilicoglu et al., 2020) (Dobreva et al., 2020) (Perera et al., 2020) (Jofche et al., 2023) (Cunha et al., 2019) (Žunić et al., 2020)
BRAT (Stenetorp et al., 2012) https://brat.nlplab.org/introduction.html	Annotating structured text	(Yang et al., 2021) (Levitan et al., 2011)
SpaCy library (Honnibal et al., 2020) https://spacy.io/	Tokenization, lemmatization, POS tagging, NER, word similarity, SRL	(Peng et al., 2019) (Li et al., 2020a) (Jofche et al., 2023) (Mao and Fung, 2020) (Dobreva et al., 2020) (Y. Liu et al., preprint, DOI: https://doi.org/10.48550/arXiv.1907.11692) (Lai et al., 2019) (Gururangan et al., 2020) (Chen et al., 2020) (K. Huang et al., preprint, DOI: https://doi.org/10.48550/arXiv.1904.05342) (Rivera and Martínez, 2019) (D'souza et al., 2021) (A.K. Tarcar et al., preprint, DOI: https://doi.org/10.48550/arXiv.1910.11241) (Oyewusi et al., 2021) (Zeng et al., 2022) (Jang et al., 2020) (Ramachandran and Arutchelvan, 2021)
DOMEO (Ciccarese et al., 2011) https://github.com/domeo/domeo	Annotating structured text	(Hochheiser et al., 2016) (Boyce et al., 2012)
Transformers (Wolf et al., 2020) https://huggingface.co/	NER, NLI, QA, SRL, classification, embeddings	(K. Kuratov and M. Arkhipov, preprint, DOI: https://doi.org/10.48550/arXiv.1905.07213) (Hussain et al., 2021) (Xiong et al., 2019) (Beltagy et al., 2019) (K. Huang et al., preprint, DOI: https://doi.org/10.48550/arXiv.1904.05342) (El Boukkouri et al., 2020) (Lee et al., 2019) (Aldahdooh et al., 2022) (Canete et al., 2020) (Michalopoulos et al., 2021) (Li et al., 2020b) (Sun et al., 2021) (Akhtyamova, 2020) (Peng et al., 2019) (Dobreva et al., 2020) (Rogers et al., 2020) (Gururangan et al., 2020) (Jofche et al., 2023) (Pfeiffer et al., 2021) (A. Breden and L. Moore, preprint, DOI: https://doi.org/10.48550/arXiv.2005.06634) (Houlsby et al., 2019) (Khadhraoui et al., 2022) (Alsentzer et al., 2019) (Li et al., 2020a) (Sboev et al., 2022) (Mao and Fung, 2020) (Lai et al., 2019) (Moradi and Samwald, 2021) (Liu et al., 2021) (Perera et al., 2020) (Y. Liu et al., preprint, DOI: https://doi.org/10.48550/arXiv.1907.11692) (Conneau et al., 2020) (Aldahdooh et al., 2021) (Yuan et al., 2022) (Qin et al., 2021)

(*continued*)

TABLE 7—*Continued*

Name	Usage	Referenced Papers
MedCat Tool (Kraljevic et al., 2021) https://github.com/CogStack/MedCAT	NER+L	(Dobreva et al., 2022) (Alicante et al., 2016)
AllenNLP (Gardner et al., 2018) https://allenai.org/allennlp	NER, NLI, QA, SRL, classification, embeddings	(Jofche et al., 2023) (Beltagy et al., 2019) (Wang et al., 2018) (Dobreva et al., 2020) (Li et al., 2020a) (Peng et al., 2019) (Gururangan et al., 2020) (Yang et al., 2021) (Li et al., 2020b)
Flair (Akbik et al., 2019) https://github.com/flairNLP/flair	NER, POS tagging, classification	(Sun et al., 2021) (Akhtyamova, 2020) (Conneau et al., 2020)
Gensim (Řehůřek and Sojka, 2010) https://radimrehurek.com/gensim/	Text summarization, embeddings	(Dobreva et al., 2020) (Habibi et al., 2017) (Joshi et al., 2022) (Dhrangadhariya et al., 2020) (Zhu et al., 2020)
JIEBA tool (Sun, 2012) https://github.com/fxsjy/jieba	Chinese words: POS tagging, TF-IDF, text-rank	(Zhong, 2021) (Yang et al., 2019b) (Li et al., 2021a) (Yang et al., 2020) (Lan and Zhang, 2020)
TextBlob (Loria et al., 2018) https://textblob.readthedocs.io/en/dev/	NER, NLI, QA, SRL, classification, embeddings	(Sivasankari et al., 2017) (Saad et al., 2021) (Ribeiro et al., 2021)
Polyglot (Nystrom et al., 2003) https://github.com/aboSamoor/polyglot	NER, POS tagging, sentiment analysis, embedding	(Li et al., 2020a) (Prasad and Sha, 2013) (Ceusters and Bouquet, 2000)
Quepy (Andrawos et al., 2012) https://github.com/machinalis/quepy	NLP, question transformation to queries	(Marginean and Marc, 2013)

load the entire file into memory, has efficient multicore implementations of popular algorithms, is platform-independent, and supports distributed computing. Dobreva et al. (2020) apply Gensim to NER.

B. General Natural Language Processing Libraries

JIEBA tool (Sun, 2012) (https://github.com/fxsjy/jieba) supports Chinese word segmentation based on word frequency statistics with several functions such as POS tagging, TF-IDF weighting, and TextRank keyword extraction. It was used to generate POS tags of words (Qin et al., 2021).

TextBlob (Loria et al., 2018) (https://textblob.readthedocs.io/en/dev/) is a simple Python library, built on top of NLTK and Pattern, that supports complex analysis and operations on text data. The library supports noun phrase extraction, POS tagging, sentiment analysis, classification (Naive Bayes, Decision Tree), tokenization, word and phrase frequencies, parsing, n-grams, word inflection (pluralization and singularization) and lemmatization, spelling correction, and more.

Polyglot (Nystrom et al., 2003) (https://github.com/aboSamoor/polyglot) is an NLP pipeline that supports multilingual applications and offers a wide range of analyses. It features tokenization (165 languages), language detection (196 languages), NER (40 languages), POS tagging (16 languages), sentiment analysis (136 languages), word embeddings (137 languages), morphologic analysis (135 languages), and transliteration (69 languages).

Quepy (Andrawos et al., 2012) (https://github.com/machinalis/quepy) is a Python framework to transform natural language questions to queries in a database query language.

In Table 7, we overview the mentioned libraries, together with references from the papers where they are used.

VII. Conclusion

Text is an important source of information in pharmacology. To extract that information from increasingly large collections of structured and unstructured documents, NLP is an essential approach. We present a survey of recent NLP developments relevant to the pharmacological domain.

Our survey comprises five main pillars, each presented in its section: a modern methodology based on pretrained large language models, frequently used tasks, useful datasets, knowledge bases, and software libraries. Each main topic is further split into several components, giving our review a comprehensible hierarchical structure. We compress the main contributions of each section into overview tables at the end of each section. In summary, our survey testifies to swift developments in NLP and a surprising breadth of its use in pharmacology.

While we reviewed more than 250 works in our survey, the coverage is by no means exhaustive. In a few years, when next such a survey will be needed, we

expect the most exciting developments in the use and integration of multimodal resources, such as text, images, and 3D structural databases. In AI, there is a tendency for large language models, called foundation models (R. Bommasani et al., preprint, DOI: https://doi.org/10.48550/arXiv.2108.07258), to capture as much human knowledge as possible, coupled with the ability for logical and commonsense reasoning. We expect that life sciences and pharmacology will be one of the first areas where domain-specific knowledge will be integrated into such models.

Finally, NLP is a subfield of ML and AI, which have many uses in pharmacology beyond NLP. We are not aware of any review comprehensively covering their applications in pharmacology, but such a work would complement ours. Due to broadness and rapid progress in ML and AI, such a review would require several research groups and a monograph format.

Authorship Contributions

Participated in research design: Trajanov, Robnik-Šikonja.

Conducted research: Trajanov, Trajkovski, Dimitrieva, Dobreva, Jovanovik, Klemen, Žagar, Robnik-Šikonja.

Wrote or contributed to the writing of the manuscript: Trajanov, Trajkovski, Dimitrieva, Dobreva, Jovanovik, Klemen, Žagar, Robnik-Šikonja.

References

Adamson DM, Chang S, and Hansen LG (2008) *Health Research Data for the Real World: The Marketscan Databases*. New York: Thompson Healthcare.

Akbik A, Bergmann T, Blythe D, Rasul K, Schweter S, and Vollgraf R (2019) FLAIR: An easy-to-use framework for state-of-the-art NLP, in *NAACL 2019, 2019 Annual Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*; 2019 June 2–7; Minneapolis, MN, pp 54–59.

Akhtyamova L (2020) Named entity recognition in Spanish biomedical literature: Short review and BERT model, in *2020 26th Conference of Open Innovations Association (FRUCT)*; 2020 April 20–24; Yaroslavl, Russia, pp 1–7. DOI: 10.23919/FRUCT48808.2020.9087359.

Aldahdooh JM, Tanoli Z, and Tang J (2021) R-BERT-CNN: Drug-target interactions extraction from biomedical literature, in *Proceedings of the BioCreative VII Challenge Evaluation Workshop*; 2021 November 8–10, pp 102–106.

Aldahdooh J, Vähä-Koskela M, Tang J, and Tanoli Z (2022) Using BERT to identify drug-target interactions from whole PubMed. *BMC Bioinformatics* **23:**345.

Alicante A, Corazza A, Isgrò F, and Silvestri S (2016) Unsupervised entity and relation extraction from clinical records in Italian. *Comput Biol Med* **72:**263–275.

Alsentzer E, Murphy J, Boag W, Weng WH, Jindi D, Naumann T, and McDermott M (2019) Publicly available clinical BERT embeddings, in *Proceedings of the 2nd Clinical Natural Language Processing Workshop*; 2019 June 7; Minneapolis, MN, pp 72–78. DOI: 10.18653/v1/W19-1909.

Alvaro N, Miyao Y, and Collier N (2017) TwiMed: Twitter and PubMed comparable corpus of drugs, diseases, symptoms, and their relations. *JMIR Public Health Surveill* **3:**e24.

Andrawos E, García Berrotarán G, Carrascosa R, Alonso i Alemany L, Durán H (2012) Quepy-transform natural language to database queries. Available from: https://github.com/machinalis/quepy

Aronson AR (2001) Effective mapping of biomedical text to the UMLS Metathesaurus: The MetaMap program, in *Proceedings of the AMIA Symposium;* 2001 November 3–7; Washington, DC, p. 17. American Medical Informatics Association.

Asgari E and Mofrad MR (2015) Continuous distributed representation of biological sequences for deep proteomics and genomics. *PLoS One* **10:**e0141287.

Auer S, Bizer C, Kobilarov G, Lehmann J, Cyganiak R, and Ives Z (2007) DBpedia: A nucleus for a web of open data, in *The Semantic Web: 14th International Conference, ESWC 2017;* 2017 May 28–June 1; Portorož, Slovenia, pp 722–735.

Bamburová M and Neverilová Z (2019) Structured information extraction from pharmaceutical records, in *RASLAN 2019;* 2019 December 6–8; Karlova Studánka, Czech Republic, pp 55–62.

Belleau F, Nolin MA, Tourigny N, Rigault P, and Morissette J (2008) Bio2RDF: Towards a mashup to build bioinformatics knowledge systems. *J Biomed Inform* **41:**706–716.

Beltagy I, Lo K, and Cohan A (2019) SciBERT: A pretrained language model for scientific text, in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*; 2019 November; Hong Kong, China, pp 3615–3620. DOI: 10.18653/v1/D19-1371.

Ben Abacha A and Zweigenbaum P (2011) Automatic extraction of semantic relations between medical entities: a rule based approach. *J Biomed Semantics* **2(Suppl 5):**S4.

Bird S, Klein E, and Loper E (2009) *Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit*. Sabastopol, CA: O'Reilly Media, Inc.

Biswas S, Khimulya G, Alley EC, Esvelt KM, and Church GM (2021) Low-N protein engineering with data-efficient deep learning. *Nat Methods* **18:**389–396.

Bizer C, Heath T, and Berners-Lee T (2009) Linked data—the story so far. *Int J Semantic Web Inf Syst* **5:**1–22.

Bizer C, Heath T, Idehen K, and Berners-Lee T (2008) Linked data on the web (LDOW2008), in *Proceedings of the 17th International Conference on World Wide Web;* 2008 April 21–25; Beijing, China, pp 1265–1266. ACM.

Bodenreider O (2004) The Unified Medical Language System (UMLS): integrating biomedical terminology. *Nucleic Acids Res* **32:**D267–D270.

Bonner S, Barrett IP, Ye C, Swiers R, Engkvist O, and Hamilton W (2021) A review of biomedical datasets relating to drug discovery: A knowledge graph perspective. *Brief Bioinform* **23:**bbac404.

Bordes A, Usunier N, Garcia-Duran A, Weston J, and Yakhnenko O (2013) Translating embeddings for modeling multi-relational data, in *Neural Information Processing Systems*; 2013 December 8–10; Vancouver, BC, pp 1–9. NIPS.

Boyce R, Gardner G, and Harkema H (2012) Using natural language processing to identify pharmacokinetic drug-drug interactions described in drug package inserts, in *Proceedings of the 2012 Workshop on Biomedical Natural Language Processing*; 2012 June; Montreal, Canada, pp 206–213.

Bradshaw J, Paige B, Kusner MJ, Segler M, and Hernández-Lobato JM (2019) A model to search for synthesizable molecules, in *Advances in Neural Information Processing Systems*; 2019 December 8–14; Vancouver, BC. Vol. 32, pp 7937–7949.

Burgelman JC, Pascu C, Szkuta K, Von Schomberg R, Karalopoulos A, Repanas K, and Schouppe M (2019) Open science, open data, and open scholarship: European policies to make science fit for the twenty-first century. *Front Big Data* **2:**43.

Callahan A, Cruz-Toledo J, Ansell P, and Dumontier M (2013a) Bio2RDF release 2: Improved coverage, interoperability and provenance of life science linked data, in *The Semantic Web: Semantics and Big Data: 10th International Conference, ESWC 2013*; 2013 May 26–30; Montpellier, France, pp 200–212, Springer, New York.

Callahan A, Cruz-Toledo J, and Dumontier M (2013b) Ontology-based querying with Bio2RDF's linked open data. *J Biomed Semantics* **4(Suppl 1):**S1.

Canese K, Weis S (2013) *PubMed: The Bibliographic Database. The NCBI Handbook*, 2nd ed, National Center for Biotechnology Information, Bethesda.

Canete J, Chaperon G, Fuentes R, Ho JH, Kang H, and Pérez J (2020) Spanish pre-trained BERT model and evaluation data., in *Proceedings of Practical ML for Developing Countries (PML4DC) at ICLR*; 2020 April 26; Addis Ababa, Ethiopia.

Carracedo-Reboredo P, Liñares-Blanco J, Rodríguez-Fernández N, Cedrón F, Novoa FJ, Carballal A, Maojo V, Pazos A, and Fernandez-Lozano C (2021) A review on machine learning approaches and trends in drug discovery. *Comput Struct Biotechnol J* **19:**4538–4558.

Ceusters W and Bouquet L (2000) Language engineering and information mapping in pharmaceutical medicine: dealing successfully with information overload. *J Belg Med Inform Assoc.* **7:**26–34.

Chang CC and Lin CJ (2011) LIBSVM: A library for support vector machines. ACM Transactions on Intelligent Systems and Technology. 2:27:1–27:27. Software available at http://www.csie.ntu.edu.tw/∼cjlin/libsvm.

Chapman AB, Peterson KS, Alba PR, DuVall SL, and Patterson OV (2019) Detecting adverse drug events with rapidly trained classification models. *Drug Saf* **42:**147–156.

Chen IY, Agrawal M, Horng S, and Sontag D (2019) Robustly extracting medical knowledge from EHRs: A case study of learning a health knowledge graph, in *Pacific Symposium on Biocomputing 2020*; 2020 January 3–7; Fairmont Orchid, Hawaii, pp 19–30, World Scientific, Singapore.

Chen L, Gu Y, Ji X, Sun Z, Li H, Gao Y, and Huang Y (2020) Extracting medications and associated adverse drug events using a natural language processing system combining knowledge base and deep learning. *J Am Med Inform Assoc* **27:**56–64.

Chen Q, Allot A, and Lu Z (2021a) LitCovid: An open database of COVID-19 literature. *Nucleic Acids Res* **49(D1):**D1534–D1540.

Chen Q, Leaman R, Allot A, Luo L, Wei CH, Yan S, and Lu Z (2021b) Artificial intelligence in action: addressing the COVID-19 pandemic with natural language processing. *Annu Rev Biomed Data Sci* **4:**313–339.

Chen R, Liu X, Jin S, Lin J, and Liu J (2018) Machine learning for drug-target interaction prediction. *Molecules* **23:**2208.

Chiaramello E, Pinciroli F, Bonalumi A, Caroli A, and Tognola G (2016) Use of "off-the-shelf" information extraction algorithms in clinical informatics: A feasibility study of MetaMap annotation of Italian medical notes. *J Biomed Inform* **63:**22–32.

Ciccarese P, Ocana M, Clark T (2011) DOMEO: A web-based tool for semantic annotation of online documents, in *Bio-Ontologies 2011*; 2011 July 19–21; Vienna, Austria.

Coleman J and Coleman JS (2005) *Introducing Speech and Language Processing*. Cambridge, UK: Cambridge University Press.

Conneau A, Khandelwal K, Goyal N, Chaudhary V, Wenzek G, Guzmán F, Grave É, Ott M, Zettlemoyer L, and Stoyanov V (2020) Unsupervised cross-lingual representation learning at scale, in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*; 2020 July, pp 8440–8451.

Cunha AM, Belloze KT, and Guedes GP (2019) Recognizing pharmacovigilance named entities in Brazilian Portuguese with CoreNLP, in *Anais do XIII Brazilian e-Science Workshop;* 2019 July 17–18; Lisbon, Portugal, pp 76–79.

Dara S, Dhamercherla S, Jadav SS, Babu CM, and Ahsan MJ (2022) Machine learning in drug discovery: A review. *Artif Intell Rev* **55:**1947–1999.

Deftereos SN, Andronis C, Friedla EJ, Persidis A, and Persidis A (2011) Drug repurposing and adverse event prediction using high-throughput literature analysis. *Wiley Interdiscip Rev Syst Biol Med* **3:**323–334.

Demner-Fushman D, Chapman WW, and McDonald CJ (2009) What can natural language processing do for clinical decision support? *J Biomed Inform* **42:**760–772.
Dernoncourt F and Lee JY (2017) PubMed 200k RCT: A dataset for sequential sentence classification in medical abstracts, in *Proceedings of the Eighth International Joint Conference on Natural Language Processing*; 2017 November; Taipei, Taiwan. **Vol. 2**: Short Papers, pp 308–313.
Devlin J, Chang MW, Lee K, and Toutanova K (2019) BERT: Pre-training of deep bidirectional transformers for language understanding., in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*; 2019 June; Minneapolis, MN. **Vol. 1**: Long and Short Papers, pp 4171–4186.
Dhrangadhariya A, Hilfiker R, Schaer R, and Müller H (2020) Machine learning assisted citation screening for systematic reviews. *Stud Health Technol Inform* **270**:302–306.
Dobreva J, Jofche N, Jovanovik M, and Trajanov D (2020) Improving NER performance by applying text summarization on pharmaceutical articles, in *International Conference on ICT Innovations*; 2020 September 24–26; Skopje, North Macedonia, pp 87–97, Springer, New York.
Dobreva J, Jovanovik M, and Trajanov D (2022) DD-RDL: Drug-disease relation discovery and labeling, in *International Conference on ICT Innovations;* 2022 September 29–October 1; Skopje, North Macedonia, pp 98–112, Springer, New York.
Dreisbach C, Koleck TA, Bourne PE, and Bakken S (2019) A systematic review of natural language processing and text mining of symptoms from electronic patient-authored text data. *Int J Med Inform* **125:**37–46.
D'souza S, Nazareth D, Vaz C, and Shetty M (2021) Blockchain and AI in pharmaceutical supply chain. Available at *SSRN* 3852034.
Dumitriu A, Molony C, and Daluwatte C (2021) Graph-based natural language processing for the pharmaceutical industry, in *Provenance in Data Science: From Data Models to Context-Aware Knowledge Graphs* (Sikos LF, Seneviratne OW, and McGuinness DL, eds) pp 75–110, Springer International Publishing, Cham, Switzerland. DOI: 10.1007/978-3-030-67681-06.
El Boukkouri H, Ferret O, Lavergne T, Noji H, Zweigenbaum P, and Tsujii J (2020) CharacterBERT: Reconciling ELMo and BERT for word-level open-vocabulary representations from characters, in *International Conference on Computational Linguistics*; 2020 December 8–13, pp 6903–6915.
Farrar S (2002) The Arizona virtual patient: Using question-answering technology to enhance dialogue processing, in *Proceedings of the Second International Conference on Human Language Technology Research*; 2002 March 24–27; San Diego, CA, pp 222–225.
Filannino M and Uzuner Ö (2018) Advancing the state of the art in clinical natural language processing through shared tasks. *Yearb Med Inform* **27:**184–192.
Frye C, de Mijolla D, Begley T, Cowton L, Stanley M, and Feige I (2021) Shapley explainability on the data manifold, in *International Conference on Learning Representations*; 2021 May 3–7.
Fung KW, Jao CS, and Demner-Fushman D (2013) Extracting drug indication information from structured product labels using natural language processing. *J Am Med Inform Assoc* **20:**482–488.
Gardner M, Grus J, Neumann M, Tafjord O, Dasigi P, Liu NF, Peters M, Schmitz M, and Zettlemoyer L (2018) AllenNLP: A deep semantic natural language processing platform, in *Proceedings of Workshop for NLP Open Source Software (NLP-OSS)*; 2018 July; Melbourne, Australia, pp 1–6. Association for Computational Linguistics. DOI: 10.18653/v1/W18-2501.
Gaulton A, Bellis LJ, Bento AP, Chambers J, Davies M, Hersey A, Light Y, McGlinchey S, Michalovich D, Al-Lazikani B et al. (2012) ChEMBL: A large-scale bioactivity database for drug discovery. *Nucleic Acids Res* **40:**D1100–D1107.
Goh KI, Cusick ME, Valle D, Childs B, Vidal M, and Barabási AL (2007) The human disease network. *Proc Natl Acad Sci USA* **104:**8685–8690.
Goodfellow I, Bengio Y, Courville A, and Bengio Y (2016) *Deep Learning*. MIT Press, Cambridge, MA.
Goodwin TR and Harabagiu SM (2016) Medical question answering for clinical decision support, in *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management. CIKM '16*; 2016 October 24–28, Indianapolis, IN, pp 297–306, Association for Computing Machinery, New York. DOI: 10.1145/2983323.2983819.
Gopalakrishnan V, Jha K, Jin W, and Zhang A (2019) A survey on literature based discovery approaches in biomedical domain. *J Biomed Inform* **93:**103141.
Gottlieb A, Stein GY, Ruppin E, and Sharan R (2011) PREDICT: A method for inferring novel drug indications with application to personalized medicine. *Mol Syst Biol* **7:**496.
Gu J, Qian L, and Zhou G (2016) Chemical-induced disease relation extraction with various linguistic features. *Database (Oxford)* **2016:**baw042.
Gururangan S, Marasović A, Swayamdipta S, Lo K, Beltagy I, Downey D, and Smith NA (2020) Don't stop pretraining: Adapt language models to domains and tasks, in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics;* 2020 July, pp 8342–8360.
Habibi M, Weber L, Neves M, Wiegandt DL, and Leser U (2017) Deep learning with word embeddings improves biomedical named entity recognition. *Bioinformatics* **33:**i37–i48.
Hakala K and Pyysalo S (2019) Biomedical named entity recognition with multilingual BERT, in *Proceedings of The 5th Workshop on BioNLP Open Shared Tasks;* 2019 November; Hong Kong, China, pp 56–61.
Han K, Cao P, Wang Y, Xie F, Ma J, Yu M, Wang J, Xu Y, Zhang Y, and Wan J (2022) A review of approaches for predicting drug–drug interactions based on machine learning. *Front Pharmacol* **12:**814858.
Hao B, Zhu H, and Paschalidis I (2020) Enhancing clinical BERT embedding using a biomedical knowledge base, in *Proceedings of the 28th International Conference on Computational Linguistics*; 2020 December; Barcelona, Spain, pp 657–661.
Heath T and Bizer C (2011) *Linked Data: Evolving the Web into a Global Data Space. Synthesis Lectures on the Semantic Web: Theory and Technology.1*. Morgan and Claypool, San Rafael, CA.
Henry S, Buchan K, Filannino M, Stubbs A, and Uzuner O (2020) 2018 n2c2 shared task on adverse drug events and medication extraction in electronic health records. *J Am Med Inform Assoc* **27:**3–12.
Henry S and McInnes BT (2017) Literature based discovery: Models, methods, and trends. *J Biomed Inform* **74:**20–32.
Herrero-Zazo M, Segura-Bedmar I, Martínez P, and Declerck T (2013) The DDI corpus: An annotated corpus with pharmacological substances and drug-drug interactions. *J Biomed Inform* **46:**914–920.
Hochheiser H, Ning Y, Hernandez A, Horn JR, Jacobson R, and Boyce RD (2016) Using nonexperts for annotating pharmacokinetic drug-drug interaction mentions in product labeling: A feasibility study. *JMIR Res Protoc* **5:**e40.
Hogan A, Blomqvist E, Cochez M, D'amato C, Melo GD, Gutierrez C, Kirrane S, Gayo JEL, Navigli R, Neumaier S et al. (2021) Knowledge graphs. *ACM Comput Surv* **54:**14–37.
Honnibal M, Montani I, Van Landeghem S, and Boyd A (2020) spaCy: Industrial-strength natural language processing in Python. DOI: 10.5281/zenodo.1212303.
Houlsby N, Giurgiu A, Jastrzebski S, Morrone B, De Laroussilhe Q, Gesmundo A, Attariyan M, and Gelly S (2019) Parameter-efficient transfer learning for NLP, in *International Conference on Machine Learning*; 2019 June 9–13; Long Beach, CA, pp 2790–2799. PMLR.
Huang K, Xiao C, Hoang T, Glass L, and Sun J (2020) CASTER: Predicting drug interactions with chemical substructure representation, in *Proceedings of the 34th AAAI Conference on Artificial Intelligence*; 2020 February 7–12; New York, NY, pp 702–709.
Huang Y and Li Y (2004) Classifying g-protein coupled receptors with support vector machine; in *International Symposium on Neural Networks*; 2004 August 19–21; Dalian, China, pp 448–452, Springer, New York.
Hussain S, Afzal H, Saeed R, Iltaf N, and Umair MY (2021) Pharmacovigilance with transformers: A framework to detect adverse drug reactions using BERT fine-tuned with FARM. *Comput Math Methods Med* **2021:**5589829.
Jagannatha A, Liu F, Liu W, and Yu H (2019) Overview of the first natural language processing challenge for extracting medication, indication, and adverse drug events from electronic health record notes (MADE 1.0). *Drug Saf* **42:**99–111.
Jain S and Wallace BC (2019) Attention is not explanation, in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2019 June; Minneapolis, MN, **Vol. 1** (Long and Short Papers), pp 3543–3556. DOI: 10.18653/v1/N19-1357.
Jang H, Rempel E, Carenini G, and Janjua N (2020) Exploratory analysis of COVID-19 related tweets in North America to inform public health institutes; in *Proceedings of the 1st Workshop on NLP for COVID-19 (Part 2) at EMNLP;* 2020 December. Association for Computational Linguistics. DOI: 10.18653/v1/2020.nlpcovid19-2.18.
Janssen A, Bennis FC, and Mathôt RAA (2022) Adoption of machine learning in pharmacometrics: An overview of recent implementations and their considerations. *Pharmaceutics* **14:**1814.
Jha K, Wang Y, Xun G, and Zhang A (2018) Interpretable word embeddings for medical domain; in *2018 IEEE International Conference on Data Mining (ICDM)*; 2018 November 17–20; Singapore, pp 1061–1066. DOI: 10.1109/ICDM.2018.00135.
Jiang K, Chen T, Huang L, Gupta R, Calix RA, and Bernard GR (2019) An explainable approach of inferring potential medication effects from social media data, in *Artificial Intelligence in Medicine: Knowledge Representation and Transparent and Explainable Systems* (Marcos M, Juarez JM, Lenz R, Nalepa GJ, Nowaczyk S, Peleg M, Stefanowski J, and Stiglic G eds) pp 82–92. Springer, New York.
Jiang K and Zheng Y (2013) Mining twitter data for potential drug effects, in *International Conference on Advanced Data Mining and Applications*; 2013 December 14–16; Hangzho, China, pp 434–443, Springer, New York.
Jofche N, Mishev K, Stojanov R, Jovanovik M, Zdravevski E, Trajanov D (2023) Pharmke: Knowledge extraction platform for pharmaceutical texts using transfer learning. *Computers* **12**:17.
Johnson AE, Pollard TJ, Shen L, Lehman LW, Feng M, Ghassemi M, Moody B, Szolovits P, Celi LA, and Mark RG (2016) MIMIC-III, a freely accessible critical care database. *Sci Data* **3:**160035.
Joshi C, Attar VZ, and Kalamkar SP (2022) An unsupervised topic modeling approach for adverse drug reaction extraction and identification from natural language text, in *Advances in Data and Information Sciences*, pp 505–514, Springer.
Jovanovik M, Bogojeska A, Trajanov D, and Kocarev L (2015a) Inferring cuisine–drug interactions using the linked data approach. *Sci Rep* **5**:9346.
Jovanovik M, Najdenov B, Strezoski G, and Trajanov D (2015b) Linked Open Data for Medical Institutions and Drug Availability Lists in Macedonia, in *New Trends in Database and Information Systems II* (Tiwari S, Trivedi MC, Kolhe ML, Singh BK eds) pp 245–256, Springer, New York.
Jovanovik M, Najdenov B, and Trajanov D (2013) Linked open drug data from the Health Insurance Fund of Macedonia, in *10th International Conference for Informatics and Information Technology*; 2013 April, pp 56–61, Faculty of Computer Science & Engineering, Skopje, North Macedonia.
Jovanovik M and Trajanov D (2017) Consolidating drug data on a global scale using linked data. *J Biomed Semantics* **8:**3.
Jung J and Lee D (2013) Inferring disease association using clinical factors in a combinatorial manner and their use in drug repositioning. *Bioinformatics* **29:**2017–2023.
Jurafsky D and Martin JH (2008) *Speech and Language Processing: An Introduction to Speech Recognition, Computational Linguistics and Natural Language Processing*, 2nd ed. Prentice Hall, Hoboken, NJ.
Jurafsky D and Martin JH (2022) Speech and language processing, 3rd edition draft. Available from: https://web.stanford.edu/~jurafsky/slp3

Kaas-Hansen BS, Gentile S, Caioli A, and Andersen SE (2023) Exploratory pharmacovigilance with machine learning in big patient data: A focused scoping review. *Basic Clin Pharmacol Toxicol* **132:**233–241.
Kadir RA and Bokharaeian B (2013) Overview of biomedical relations extraction using hybrid rulebased approaches. *J Ind and Intell Inf* **1:**169–173.
Kamalov F, Cherukuri A, Sulieman H, Thabtah F, and Hossain A (2022) Machine learning applications for COVID-19: A state-of-the-art review, in *2022 Advances in Science and Engineering Technology International Conference*; 2022 February 21–24, pp 56–61.
Kamp HG, Walk T, Ishikawa G, Moeller N, and van Ravenzwaay B (2013) The application of metabolomics in vivo for early detection of systemic toxicity in drug safety testing, in *Annual Meeting of the Japanese Society of Toxicology The 40th Annual Meeting of the Japanese Society of Toxicology*; 2013 January 31–February 1; Tsukuba, Japan, pp 150418. Japanese Society of Toxicology.
Karypis G and Kumar V (1998) A fast and high quality multilevel scheme for partitioning irregular graphs. *SIAM J Sci Comput* **20:**359–392.
Kerner J, Dogan A, and von Recum H (2021) Machine learning and big data provide crucial insight for future biomaterials discovery and research. *Acta Biomater* **130:**54–65.
Khadhraoui M, Bellaaj H, Ammar MB, Hamam H, and Jmaiel M (2022) Survey of BERT-base models for scientific text classification: COVID-19 case study. *Appl Sci (Basel)* **12:**2891.
Kilicoglu H, Rosemblat G, Fiszman M, and Shin D (2020) Broad-coverage biomedical relation extraction with SemRep. *BMC Bioinformatics* **21:**188.
Kraljevic Z, Searle T, Shek A, Roguski L, Noor K, Bean D, Mascio A, Zhu L, Folarin AA, Roberts A et al. (2021) Multi-domain clinical natural language processing with MedCAT: The Medical Concept Annotation Toolkit. *Artif Intell Med* **117:**102083.
Kringelum J, Kjaerulff SK, Brunak S, Lund O, Oprea TI, and Taboureau O (2016) ChemProt-3.0: A global chemical biology diseases mapping. *Database (Oxford)* **2016:**bav123.
Kuhn M, Letunic I, Jensen LJ, and Bork P (2016) The SIDER database of drugs and side effects. *Nucleic Acids Res* **44(D1)**:D1075–D1079.
Kumar R and Saha P (2022) A review on artificial intelligence and machine learning to improve cancer management and drug discovery. *Int J Res Appl Sci Biotech.* **9:**149–156.
Kumari T, Pant B, and Pardasani K (2010) A SVM model for AAC based classification of class B GPCRs, in 6th World Congress of Biomechanics (WCB 2010); 2010 August 1–6; Singapore, pp 1607–1610. Springer, New York.
Lai V, Cai Z, and Tan C (2019) Many faces of feature importance: Comparing built-in and post-hoc feature importance in text classification, 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)*; 2019 November; Hong Kong, China, pp 486–495. DOI: 10.18653/v1/D19-1046.
Lan W and Zhang P (2020) Research on adaptive learning methods of Chinese medicine based on big data; in *2020 International Conference on Public Health and Data Science (ICPHDS)*. 2020 November 20–22; Guangzho, China, pp 90–93. IEEE.
Le DH and Le L (2016) Systems pharmacology: A unified framework for prediction of drug-target interactions. *Curr Pharm Des* **22**:3569–3575.
Lee J, Yi SS, Jeong M, Sung M, Yoon W, Choi Y, Ko M, and Kang J (2020) Answering questions on COVID-19 in real-time, in *Proceedings of the 1st Workshop on NLP for COVID-19 (Part 2) at EMNLP* December 2020. Association for Computational Linguistics. DOI: 10.18653/v1/2020.nlpcovid19-2.1.
Lee J, Yoon W, Kim S, Kim D, Kim S, So CH, and Kang J (2019) BioBERT: A pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics* **36:**1234–1240.
Levitan BS, Andrews EB, Gilsenan A, Ferguson J, Noel RA, Coplan PM, and Mussen F (2011) Application of the BRAT framework to case studies: Observations and insights. *Clin Pharmacol Ther* **89:**217–224.
Li F, Liu W, and Yu H (2018) Extraction of information related to adverse drug events from electronic health record notes: Design of an end-to-end model based on deep learning. *JMIR Med Inform* **6:**e12159.
Li F, Zhang M, Fu G, and Ji D (2017) A neural joint model for entity and relation extraction from biomedical text. *BMC Bioinformatics* **18:**198.
Li J, Sun A, Han J, and Li C (2020a) A survey on deep learning for named entity recognition. *IEEE Trans Knowl Data Eng* **34:**50–70.
Li M, Du L, Xu J, and Guo C (2021a) A hypergraph-based method for pharmaceutical data similarity retrieval, in *2021 4th International Conference on Big Data Technologies*; 2021 September 24–26; Zibo, China, pp 134–140.
Li X, Wang Y, Wang D, Yuan W, Peng D, and Mei Q (2019) Improving rare disease classification using imperfect knowledge graph; in *2019 IEEE International Conference on Healthcare Informatics (ICHI);* 2019 June 10–13; Xi'an, China, pp 1–2. DOI: 10.1109/ICHI.2019.8904588.
Li Z, Lin H, and Zheng W (2020b) An effective emotional expression and knowledge-enhanced method for detecting adverse drug reactions. *IEEE Access* **8**:87083–87093.
Li Z, Yang Z, Wang L, Zhang Y, Lin H, and Wang J (2021b) Lexicon knowledge boosted interaction graph network for adverse drug reaction recognition from social media. *IEEE J Biomed Health Inform* **25:**2777–2786.
Liu F, Jagannatha A, and Yu H (2019a) Towards drug safety surveillance and pharmacovigilance: Current progress in detecting medication and adverse drug events from electronic health records. *Drug Saf* **42:**95–97.
Liu F, Shareghi E, Meng Z, Basaldella M, and Collier N (2021) Self-alignment pretraining for biomedical entity representations, in *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*; 2021 June, pp 4228–4238.
Liu J, Abeysinghe R, Zheng F, and Cui L (2019b) Pattern-based extraction of disease drug combination knowledge from biomedical literature, in *2019 IEEE International Conference on Healthcare Informatics (ICHI)*; 2019 June 10–13; Xi'an, China, pp. 1–7, IEEE.
Liu S, Tang B, Chen Q, and Wang X (2015) Effects of semantic features on machine learning-based drug name recognition systems: Word embeddings vs. manually constructed dictionaries. *Information (Basel)* **6:**848–865.
Liu Z, Peng E, Yan S, Li G, and Hao T (2018) T-Know: A knowledge graph-based question answering and information retrieval system for traditional Chinese medicine, in *Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations*; 2018 August; Santa Fe, NM, pp 15–19.
Loria S (2018) textblob Documentation. *Release 0.15*, *2*(8). Available from: https://textblob.readthedocs.io/en/dev/
Lundberg SM and Lee SI (2017) A unified approach to interpreting model predictions. In: *Advances in Neural Information Processing Systems NIPS 2017*; 2017 December 4–9; Long Beach, CA, Vol. 30.
Luo Y, Thompson WK, Herr TM, Zeng Z, Berendsen MA, Jonnalagadda SR, Carson MB, and Starren J (2017) Natural language processing for EHR-based pharmacovigilance: A structured review. *Drug Saf* **40:**1075–1089.
Madsen A, Reddy S, and Chandar S (2022) Post-hoc interpretability for neural NLP: A survey. *ACM Comput Surv* **55:**155.
Mahatpure J, Motwani M, and Shukla PK (2019) An electronic prescription system powered by speech recognition, natural language processing and blockchain technology. *Int J Sci Technol Res* **8:**1454–1462.
Manning CD, Surdeanu M, Bauer J, Finkel JR, Bethard S, and McClosky D (2014) The Stanford CoreNLP natural language processing toolkit, in *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations;* 2014 June; Baltimore, MD, pp 55–60.
Mao Y and Fung KW (2020) Use of word and graph embedding to measure semantic relatedness between Unified Medical Language System concepts. *J Am Med Inform Assoc* **27:**1538–1546.
Marginean A (2014) GFMed: Question answering over biomedical linked data with grammatical framework, in *CLEF (Working Notes)*; 2014 September 15–18; Sheffield, UK, pp 1224–1235.
Marginean A and Marc O (2013) Towards querying bioinformatic linked data in natural language, in *2013 IEEE 9th International Conference on Intelligent Computer Communication and Processing (ICCP)*; 2013 September 5–7; Cluj-Nopoca, Romania, pp 23–26, IEEE.
Martinc M, Škrlj B, Pirkmajer S, Lavrač N, Cestnik B, Marzidovšek M, and Pollak S (2020) COVID-19 therapy target discovery with context-aware literature mining, in *International Conference on Discovery Science;* 2020 October 19–20, pp 109–123, Springer.
Mattes WB, Kamp HG, Fabian E, Herold M, Krennrich G, Looser R, Mellert W, Prokoudine A, Strauss V, van Ravenzwaay B et al. (2013) Prediction of clinically relevant safety signals of nephrotoxicity through plasma metabolite profiling. *BioMed Res Int* **2013:**202497.
McComb M, Bies R, and Ramanathan M (2022) Machine learning in pharmacometrics: Opportunities and challenges. *Br J Clin Pharmacol* **88:**1482–1499.
McCoubrey LE, Elbadawi M, Orlu M, Gaisford S, and Basit AW (2021) Harnessing machine learning for development of microbiome therapeutics. *Gut Microbes* **13:**1–20.
McCreery CH, Katariya N, Kannan A, Chablani M, and Amatriain X (2020) Effective transfer learning for identifying similar questions: Matching user questions to COVID-19 FAQs, in *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining;* 2020 July 6–10, pp 3458–3465.
Meng Z, Liu F, Clark T, Shareghi E, and Collier N (2021) Mixture-of-partitions: Infusing large biomedical knowledge graphs into BERT, in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing;* 2021 November; Punta Cana, Dominican Republic, pp 4672–4681.
Michalopoulos G, Wang Y, Kaka H, Chen H, and Wong A (2021) UmlsBERT: Clinical domain knowledge augmentation of contextual embeddings using the unified medical language system metathesaurus, in: *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies;* 2021 June, pp 1744–1753.
Mikolov T, Sutskever I, Chen K, Corrado GS, and Dean J (2013) Distributed representations of words and phrases and their compositionality, in *Advances in Neural Information Processing Systems* (Jordan MI, LeCun Y, Solla SA eds) pp 3111–3119, MIT Press, Cambridge, MA.
Moradi M and Samwald M (2021) Explaining black-box models for biomedical text classification. *IEEE J Biomed Health Inform* **25:**3112–3120.
Névéol A, Dalianis H, Velupillai S, Savova G, and Zweigenbaum P (2018) Clinical natural language processing in languages other than English: Opportunities and challenges. *J Biomed Semantics* **9:**12.
Ngo DH, Metke-Jimenez A, and Nguyen A (2018) Knowledge-based feature engineering for detecting medication and adverse drug events from electronic health records, in *International Workshop on Medication and Adverse Drug Event Detection;* 2018 May 4, pp 31–38. PMLR.
Nystrom N, Clarkson MR, and Myers AC (2003) Polyglot: An extensible compiler framework for Java, in *International Conference on Compiler Construction*; 2003 April 7–11; Warsaw, Poland, pp 138–152. Springer.
National Institutes of Health (2014) DailyMed database. https://dailymed.nlm.nih.gov/dailymed/
Okazaki N (2007) CRFsuite: A fast implementation of conditional random fields (CRFs). https://www.chokkan.org/software/crfsuite/
Oyewusi WF, Adekanmbi O, Okoh I, Salami MI, Osakuade O, Ibejih S, and Onuigwe V (2021) Artificial intelligence for pharmacovigilance in Nigerian social media text, in *AI for Public Health Workshop at ICLR'21;* 2021 May 7.
Park S, Yang JS, Shin YE, Park J, Jang SK, and Kim S (2011) Protein localization as a principal feature of the etiology and comorbidity of genetic diseases. *Mol Syst Biol* **7:**494.
Peng Y, Yan S, and Lu Z (2019) Transfer learning in biomedical natural language processing: An evaluation of BERT and ELMo on ten benchmarking datasets, in

Proceedings of the 18th BioNLP Workshop and Shared Task; 2019 August; Florence, Italy, pp 58–65.
Perera N, Dehmer M, and Emmert-Streib F (2020) Named entity recognition and relation detection for biomedical information extraction. *Front Cell Dev Bio* **8:**673.
Perera S, Sheth A, Thirunarayan K, Nair S, and Shah N (2013) Challenges in understanding clinical notes: Why NLP engines fall short and where background knowledge can help, in *Proceedings of the 2013 International Workshop on Data Management & Analytics for Healthcare*; 2013 November 1; San Francisco, CA, pp 21–26.
Pestryakova S, Vollmers D, Sherif MA, Heindorf S, Saleem M, Moussallem D, and Ngomo AN (2022) CovidPubGraph: A FAIR knowledge graph of COVID-19 publications. *Sci Data* **9:**389.
Peters M, Neumann M, Iyyer M, Gardner M, Clark C, Lee K, and Zettlemoyer L (2018) Deep contextualized word representations, in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*; 2018 June 1–6; New Orleans, LA, **Vol. 1** (Long Papers), pp 2227–2237.
Pfeiffer J, Kamath A, Rücklé A, Cho K, and Gurevych I (2021) AdapterFusion: Non-destructive task composition for transfer learning, in *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume;* 2021 April, pp 487–503.
Pinto BGG, Oliveira AER, Singh Y, Jimenez L, Gonçalves ANA, Ogava RLT, Creighton R, Schatzmann Peron JP, and Nakaya HI (2020) ACE2 expression is increased in the lungs of patients with comorbidities associated with severe COVID-19. *J Infect Dis* **222:**556–563.
Pirmohamed M, Breckenridge AM, Kitteringham NR, and Park BK (1998) Adverse drug reactions. *BMJ* **316:**1295–1298.
Pope PE, Kolouri S, Rostami M, Martin CE, and Hoffmann H (2019) Explainability methods for graph convolutional neural networks, in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*; 2019 June 15–20;Long Beach, CA, pp 10764–10773.
Prabadevi B, Reddy NS, and Deepa B (2019) Heart rate encapsulation and response tool using sentiment analysis. *Iran J Electr Comput Eng* **9:**2585.
Prasad S and Sha MN (2013) NextGen data persistence pattern in healthcare: polyglot persistence, in *2013 Fourth International Conference on Computing, Communications and Networking Technologies (ICCCNT)*; 2013 July 4–6, pp 1–8, IEEE.
Preiss J, Stevenson M, and Gaizauskas R (2015) Exploring relation types for literature-based discovery. *J Am Med Inform Assoc* **22:**987–992.
Protein Data Bank Contributors (1971) Protein data bank. *Nature New Biol* **233:**223.
Pyysalo S, Ginter F, Moen H, Salakoski T, and Ananiadou S (2013) Distributional semantics resources for biomedical text processing, in *Proceedings of LBM 2013;* 2013 December 12–13; Tokyo, Japan, pp 39–44.
Qin Y, Yang W, Wang K, Huang R, Tian F, Ao S, and Chen Y (2021) Entity relation extraction based on entity indicators. *Symmetry (Basel)* **13:**539.
Rabhi S, Jakubowicz J, Metzger MH (2019) Deep learning versus conventional machine learning for detection of healthcare-associated infections in French clinical narratives. *Methods Inf Med* **58:**31–41.
Raghupathi V, Zhou Y, and Raghupathi W (2018) Legal decision support: exploring big data analytics approach to modeling pharma patent validity cases. *IEEE Access* **6:**41518–41528.
Ramachandran R and Arutchelvan K (2021) Named entity recognition on bio-medical literature documents using hybrid based approach. *J Ambient Intell Humaniz Comput* DOI: 10.1007/s12652-021-03078-z [published ahead of print].
Reese JT, Unni D, Callahan TJ, Cappelletti L, Ravanmehr V, Carbon S, Shefchek KA, Good BM, Balhoff JP, Fontana T et al. (2021) KG-COVID-19: A framework to produce customized knowledge graphs for COVID-19 response. *Patterns (N Y)* **2:**100155.
Řehůřek R and Sojka P (2010) Software framework for topic modelling with large corpora, in *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, 2010 May 22; Valletta, Malta, pp 45–50, ELRA. http://is.muni.cz/publication/884893/
Ren J (2021) Variability and functions of lexical bundles in research articles of applied linguistics and pharmaceutical sciences. *J Engl Acad Purposes* **50:**100968.
Ribeiro LA, Cinalli D, and Garcia ACB (2021) Discovering adverse drug reactions from Twitter: A sentiment analysis perspective, in *2021 IEEE 24th International Conference on Computer Supported Cooperative Work in Design (CSCWD)*; 2021 May 5–7; Dalian, China, pp 1172–1177. IEEE.
Ribeiro MT, Singh S, and Guestrin C (2016) "Why should I trust you?" Explaining the predictions of any classifier, in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining;* 2016 August 13–17; San Francisco, CA, pp 1135–1144. DOI: 10.1145/2939672.2939778.
Rivera R and Martínez P (2019) Deep neural model with enhanced embeddings for pharmaceutical and chemical entities recognition in Spanish clinical text, in *Proceedings of the 5th Workshop on BioNLP Open Shared Tasks;* 2019 November; Hong Kong, China, pp 38–46.
Rodríguez-Pérez R and Bajorath J (2020) Interpretation of compound activity predictions from complex machine learning models using local approximations and Shapley values. *J Med Chem* **63:**8761–8777.
Rogers A, Kovaleva O, and Rumshisky A (2020) A primer in BERTology: What we know about how BERT works. *Trans Assoc Comput Linguist* **8:**842–866.
Romasanta AKS, van der Sijde P, and van Muijlwijk-Koezen J (2020) Innovation in pharmaceutical R&D: mapping the research landscape. *Scientometrics* **125:**1801–1832.
Rosario B and Hearst MA (2004) Classifying semantic relations in bioscience texts, in *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL-04)*; 2004 June; Barcelona, Spain, pp 430–437.
Ruan T, Huang Y, Liu X, Xia Y, and Gao J (2019) QAnalysis: A question-answer driven analytic tool on knowledge graphs for leveraging electronic medical records for clinical research. *BMC Med Inform Decis Mak* **19:**82.
Saad E, Din S, Jamil R, Rustam F, Mehmood A, Ashraf I, and Choi GS (2021) Determining the efficiency of drugs under special conditions from users' reviews on healthcare web forums. *IEEE Access* **9:**85721–85737.
Sang S, Yang Z, Wang L, Liu X, Lin H, and Wang J (2018) SemaTyP: A knowledge graph based literature mining method for drug discovery. *BMC Bioinformatics* **19:**193.
Sboev A, Selivanov A, Moloshnikov I, Rybka R, Gryaznov A, Sboeva S, Rylkov G (2022) Extraction of the relations among significant pharmacological entities in Russian-language reviews of internet users on medications. *Big Data Cogn Comput* **6:**10.
Schriml LM, Arze C, Nadendla S, Chang YWW, Mazaitis M, Felix V, Feng G, and Kibbe WA (2012) Disease ontology: A backbone for disease semantic integration. *Nucleic Acids Res* **40:**D940–D946.
Segura-Bedmar I and Martínez P (2017) Simplifying drug package leaflets written in Spanish by using word embedding. *J Biomed Semantics* **8:**45.
Shan Z and Song H (2019) Research on management decision based on machine learning: Taking the decision of location selection of a pharmaceutical retail enterprise as an example, in *Fuzzy Systems and Data Mining V*, pp 564–574, IOS Press, Amsterdam.
Sivasankari S, Kavitha M, and Saranya G (2017) Medical analysis and visualisation of diseases using tweet data. *Res J Pharm Techn.* **10:**4306–4312.
Sjögren R, Stridh K, Skotare T, and Trygg J (2020) Multivariate patent analysis—using chemometrics to analyze collections of chemical and pharmaceutical patents. *J Chemometr* **34:**e3041.
Slack D, Hilgard S, Jia E, Singh S, and Lakkaraju H (2020) Fooling LIME and SHAP: Adversarial attacks on post hoc explanation methods, in *AAAI/ACM Conference on AI, Ethics, and Society;* 2020 February 7–8; New York, NY.
Soysal E, Wang J, Jiang M, Wu Y, Pakhomov S, Liu H, and Xu H (2018) CLAMP—a toolkit for efficiently building customized clinical natural language processing pipelines. *J Am Med Inform Assoc* **25:**331–336.
Stenetorp P, Pyysalo S, Topić G, Ohta T, Ananiadou S, and Tsujii J (2012) BRAT: A web-based tool for NLP-assisted text annotation, in *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics*; 2012 April; Avignon, France, pp 102–107.
Stephenson N, Shane E, Chase J, Rowland J, Ries D, Justice N, Zhang J, Chan L, and Cao R (2019) Survey of machine learning techniques in drug discovery. *Curr Drug Metab* **20:**185–193.
Štrumbelj E and Kononenko I (2013) Explaining prediction models and individual predictions with feature contributions. *Knowl Inf Syst* **41:**647–665.
Su D, Xu Y, Yu T, Siddique FB, Barezi E, and Fung P (2020) CAiRE-COVID: A question answering and query-focused multi-document summarization system for COVID-19 scholarly information management, in *Proceedings of the 1st Workshop on NLP for COVID-19 (Part 2) at EMNLP*; 2020 December. Association for Computational Linguistics. DOI: 10.18653/v1/2020.nlpcovid19-2.14.
Sun C, Yang Z, Wang L, Zhang Y, Lin H, and Wang J (2021) Deep learning with language models improves named entity recognition for PharmaCoNER. *BMC Bioinformatics* **22(Suppl 1):**602.
Sun J (2012) Jieba: Chinese Word Segmentation Tool. Available online at: https://github.com/fxsjy/jieba
Suthram S, Dudley JT, Chiang AP, Chen R, Hastie TJ, and Butte AJ (2010) Network-based elucidation of human disease similarities reveals common functional modules enriched for pluripotent drug targets. *PLOS Comput Biol* **6:**e1000662.
Swanson DR and Smalheiser NR (1997) An interactive system for finding complementary literatures: a stimulus to scientific discovery. *Artif Intell* **91:**183–203.
Taboureau O, Nielsen SK, Audouze K, Weinhold N, Edsgärd D, Roque FS, Kouskoumvekaki I, Bora A, Curpan R, Jensen TS et al. (2011) ChemProt: A disease chemical biology database. *Nucleic Acids Res* **39:**D367–D372.
Tang Y, Yang J, Ang PS, Dorajoo SR, Foo B, Soh S, Tan SH, Tham MY, Ye Q, Shek L et al. (2019) Detecting adverse drug reactions in discharge summaries of electronic medical records using Readpeer. *Int J Med Inform* **128:**62–70.
Turina P, Fariselli P, and Capriotti E (2021) ThermoScan: Semi-automatic identification of protein stability data from PubMed. *Front Mol Biosci* **8:**620475.
Tutubalina E, Alimova I, Miftahutdinov Z, Sakhovskiy A, Malykh V, and Nikolenko S (2021) The Russian Drug Reaction Corpus and neural models for drug reactions and effectiveness detection in user reviews. *Bioinformatics* **37:**243–249.
Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser Ł, and Polosukhin I (2017) Attention is all you need, in *Advances in Neural Information Processing Systems;* 2017 December 4–9; Long Beach, CA, pp 5998–6008.
Veisi H and Shandi HF (2020) A Persian medical question answering system. *Int J Artif Intell Tools* **29:**2050019.
Wang A, Singh A, Michael J, Hill F, Levy O, and Bowman S (2018) GLUE: A multi-task benchmark and analysis platform for natural language understanding, in *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP;* 2018 November; Brussels, Belgium, pp 353–355.
Wang P, Hao T, Yan J, and Jin L (2017) Large-scale extraction of drug–disease pairs from the medical literature. *J Assoc Inf Sci Technol* **68:**2649–2661.
Wang X, Hripcsak G, Markatou M, and Friedman C (2009) Active computerized pharmacovigilance using natural language processing, statistics, and electronic health records: A feasibility study. *J Am Med Inform Assoc* **16:**328–337.
Wawrzinek J, Hussaini SAR, Wiehr O, Pinto JMG, and Balke WT (2020) Explainable word-embeddings for medical digital libraries: A context-aware approach, in *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries in 2020;* 2020 August 1–5; Wuhan, China, pp 299–308. DOI: 10.1145/3383583.3398522.
Wei CH, Allot A, Leaman R, and Lu Z (2019) PubTator central: automated concept annotation for biomedical full text articles. *Nucleic Acids Res* **47(W1):**W587–W593.
Wei J, Huang C, Vosoughi S, and Wei J (2020). What are people asking about COVID-19? A question classification dataset, in *Proceedings of the 1st Workshop on NLP for COVID-19 at ACL*; 2020 July. Association for Computational Linguistics.

Welling M and Kipf TN (2016) Semi-supervised classification with graph convolutional networks, in *International Conference on Learning Representations (ICLR 2017)*; 2016 April 24–26;Toulon, France.

Whetzel PL, Noy NF, Shah NH, Alexander PR, Nyulas C, Tudorache T, and Musen MA (2011) BioPortal: Enhanced functionality via new web services from the National Center for Biomedical Ontology to access and use ontologies in software applications. *Nucleic Acids Res* **39:**W541-5.

Wiegreffe S and Pinter Y (2019) Attention is not not explanation, 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);* 2019 November; Hong Kong, China, pp 11–20. DOI: 10.18653/v1/D19-1002.

Wikipedia (2004) *Wikipedia*, PediaPress, Mainz, Germany.

Wishart DS, Feunang YD, Guo AC, Lo EJ, Marcu A, Grant JR, Sajed T, Johnson D, Li C, Sayeeda Z et al. (2018) DrugBank 5.0: A major update to the DrugBank database for 2018. *Nucleic Acids Res* **46(D1):**D1074–D1082.

Wolf T, Debut L, Sanh V, Chaumond J, Delangue C, Moi A, Cistac P, Rault T, Louf R, Funtowicz M et al. (2020) Transformers: State-of-the-art natural language processing, in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*; 2020 October, pp 38–45.

Wood D, Zaidman M, Ruth L, and Hausenblas M (2014) *Linked Data*. Manning Publications, Shelter Island, NY.

Wunnava S, Qin X, Kakar T, Sen C, Rundensteiner EA, and Kong X (2019) Adverse drug event detection from electronic health records using hierarchical recurrent neural networks with dual-level embedding. *Drug Saf* **42:**113–122.

Xia E, Sun W, Mei J, Xu E, Wang K, and Qin Y (2018) Mining disease-symptom relation from massive biomedical literature and its application in severe disease diagnosis, in *AMIA Annual Symposium Proceedings;* 2018 November 3–7; San Fransico, CA, p 1118. American Medical Informatics Association.

Xiong Y, Shen Y, Huang Y, Chen S, Tang B, Wang X, Chen Q, Yan J, and Zhou Y (2019) A deep learning-based system for PharmaCoNER, in *Proceedings of the 5th Workshop on BioNLP Open Shared Tasks;* 2019 November; Hong Kong, China, pp 33–37. DOI: 10.18653/v1/D19-5706.

Xue H, Li J, Xie H, and Wang Y (2018) Review of drug repositioning approaches and resources. *Int J Biol Sci* **14:**1232–1244.

Yang F, Zhang Q, Ji X, Zhang Y, Li W, Peng S, and Xue F (2022) Machine learning applications in drug repurposing. *Interdiscip Sci* **14:**15–21.

Yang H, Swaminathan R, Sharma A, Ketkar V, and D'Silva J (2011) Mining biomedical text towards building a quantitative food-disease-gene network, in *Learning Structure and Schemas from Documents*, pp 205–225, Springer, New York.

Yang HT, Ju JH, Wong YT, Shmulevich I, and Chiang JH (2017) Literature-based discovery of new candidates for drug repurposing. *Brief Bioinform* **18:**488–497.

Yang W, Zhang Z, and Gao (2020) Extracting online recruitment information based on BiLSTM-Dropout-CRF model, in *2020 IEEE 5th Information Technology and Mechatronics Engineering Conference (ITOEC)*; 2020 June 12–14, pp 1661–1665, IEEE.

Yang X, Bian J, Gong Y, Hogan WR, and Wu Y (2019a) MADEx: A system for detecting medications, adverse drug events, and their relations from clinical notes. *Drug Saf* **42:**123–133.

Yang Y, Cao Z, Zhao P, Zeng DD, Zhang Q, and Luo Y (2021) Extracting impacts of non-pharmacological interventions for COVID-19 from modelling study, in *IEEE International Conference on Intelligence and Security Informatics (ISI);* 2021 November 2–3; San Antonio, TX, pp 1–6.

Yang Y, Li Q, Liu Z, Ye F, and Deng K (2019b) Understanding traditional Chinese medicine via statistical learning of expert-specific electronic medical Records. *Quant Biol* **7:**210–232.

Yazdani-Jahromi M, Yousefi N, Tayebi A, Kolanthai E, Neal CJ, Seal S, Garibay OO (2022). Attentionsitedti: an interpretable graph-based model for drug-target interaction prediction using nlp sentence-level relation classification. *Brief Bioinform* **23:**bbac272.

Yeleswarapu S, Rao A, Joseph T, Saipradeep VG, and Srinivasan R (2014) A pipeline to extract drug-adverse event pairs from multiple data sources. *BMC Med Inform Decis Mak* **14:**13.

Yesmin F (2016) *Identification of pharmaceutical substances with raman spectroscopy.* Ph.D. thesis, Ruhr Universität, Bochum, Germany.

Yuan Z, Zhao Z, Sun H, Li J, Wang F, and Yu S (2022) CODER: Knowledge-infused cross-lingual medical term embedding for term normalization. *J Biomed Inform* **126:**103983.

Zarin DA, Tse T, Williams RJ, Califf RM, and Ide NC (2011) The ClinicalTrials.gov results database—update and key issues. *N Engl J Med* **364:**852–860.

Zeng J, Cruz-Pico CX, Saridogan T, Shufean MA, Kahle M, Yang D, Shaw K, and Meric-Bernstam F (2022) Natural language processing–assisted literature retrieval and analysis for combination therapy in cancer. *JCO Clin Cancer Inform* **6:**e2100109.

Zhang Y, Chen Q, Yang Z, Lin H, and Lu Z (2019) BioWordVec, improving biomedical word embeddings with subword information and MeSH. *Sci Data* **6:**52.

Zhao WM, Song SH, Chen ML, Zou D, Ma LN, Ma YK, Li RJ, Hao LL, Li CP, Tian DM, et al. (2020) The 2019 novel coronavirus resource. *Yi chuan= Hereditas* **42:**212–221.

Zhong Z (2021) Internet public opinion evolution in the COVID-19 event and coping strategies. *Disaster Med Public Health Prep* **15:**e27–e33.

Zhou R, Lu Z, Luo H, Xiang J, Zeng M, and Li M (2020) NEDD: A network embedding based method for predicting drug-disease associations. *BMC Bioinformatics* **21(Suppl 13):**387.

Zhou Z, Li X, and Zare RN (2017) Optimizing chemical reactions with deep reinforcement learning. *ACS Cent Sci* **3:**1337–1344.

Zhu H, Pothukuchi A, and Guo J (2020) *Doc2Vec on Similar Document Suggestion for Pharmaceutical Collections. Technical report*. College of Engineering, University of Michigan, Ann Arbor, MI.

Žunić A, Corcoran P, and Spasić I (2020) Improving the performance of sentiment analysis in health and wellbeing using domain knowledge, in *Healthcare Text Analytics Conference—HealTAC 2020*; 2010 April 23–24; London, UK.