Database, 2023, 1–16
DOI: https://doi.org/10.1093/database/baad005
Original article

Chemical identification and indexing in full-text articles: an overview of the NLM-Chem track at BioCreative VII

Robert Leaman[1,‡], Rezarta Islamaj[1,‡], Virginia Adams[2], Mohammed A. Alliheedi[3], João Rafael Almeida[4,5], Rui Antunes[4], Robert Bevan[6], Yung-Chun Chang[7], Arslan Erdengasileng[8], Matthew Hodgskiss[6], Ryuki Ida[9], Hyunjae Kim[10], Keqiao Li[8], Robert E. Mercer[11], Lukrécia Mertová[12], Ghadeer Mobasher[12,13], Hoo-Chang Shin[2], Mujeen Sung[10], Tomoki Tsujimura[9], Wen-Chao Yeh[14] and Zhiyong Lu[1,*]

[1]National Center for Biotechnology Information (NCBI), National Library of Medicine, National Institutes of Health, 8600 Rockville Pike, Bethesda, MD 20894, USA
[2]NVIDIA, 2788 San Tomas Expressway, Santa Clara, CA 95051, USA
[3]Department of Computer Science, Al Baha University, 4781 King Fahd Rd, Al Aqiq 65779, Saudi Arabia
[4]Department of Electronics, Telecommunications and Informatics (DETI), Institute of Electronics and Informatics Engineering of Aveiro (IEETA), University of Aveiro, Campus Universitário de Santiago, Aveiro 3810-193, Portugal
[5]Department of Information and Communications Technologies, University of A Coruña, Camiño do Lagar de Castro, A Coruña 15008, Spain
[6]Informatics Department, Medicines Discovery Catapult, Alderley Park, Block 35, Mereside, Macclesfield SK10 4ZF, UK
[7]Graduate Institute of Data Science, Taipei Medical University, No. 172-1, Section 2, Keelung Rd, Da'an District, Taipei City, Taipei 106, Taiwan
[8]Department of Statistics, Florida State University, 117 N. Woodward Ave, Tallahassee, FL 32306, USA
[9]Computational Intelligence Laboratory, Toyota Technological Institute, 2-12-1 Hisakata, Tempaku-ku, Nagoya, Aichi 468-8511, Japan
[10]Department of Computer Science and Engineering, Korea University, 145 Anam-ro, Seongbuk-gu, Seoul 02841, South Korea
[11]Department of Computer Science, The University of Western Ontario, Room 355, Middlesex College, Ontario, London N6A 5B7, Canada
[12]Scientific Databases and Visualization Group, Heidelberg Institute for Theoretical Studies (HITS gGmbH), Schloss-Wolfsbrunnenweg 35, Heidelberg 69118, Germany
[13]Institute of Computer Science, Heidelberg University, Im Neuenheimer Feld 205, Heidelberg 69120, Germany
[14]Institute of Information Systems and Applications, National Tsing Hua University, No. 101, Section 2, Kuang-Fu Road, Hsinchu 30013, Taiwan

[*]Corresponding author: Tel: +1-301-594-7089; Fax: +1-301-480-2290; Email: zhiyong.lu@nih.gov
[‡]Co-first authors.

Citation details: Leaman, R., Islamaj, R., Adams, V. *et al.* Chemical identification and indexing in full-text articles: an overview of the NLM-Chem track at BioCreative VII. *Database* (2023) Vol. 2023: article ID baad005; DOI: https://doi.org/10.1093/database/baad005

Abstract

The BioCreative National Library of Medicine (NLM)-Chem track calls for a community effort to fine-tune automated recognition of chemical names in the biomedical literature. Chemicals are one of the most searched biomedical entities in PubMed, and—as highlighted during the coronavirus disease 2019 pandemic—their identification may significantly advance research in multiple biomedical subfields. While previous community challenges focused on identifying chemical names mentioned in titles and abstracts, the full text contains valuable additional detail. We, therefore, organized the BioCreative NLM-Chem track as a community effort to address automated chemical entity recognition in full-text articles. The track consisted of two tasks: (i) chemical identification and (ii) chemical indexing. The chemical identification task required predicting all chemicals mentioned in recently published full-text articles, both span [i.e. named entity recognition (NER)] and normalization (i.e. entity linking), using Medical Subject Headings (MeSH). The chemical indexing task required identifying which chemicals reflect topics for each article and should therefore appear in the listing of MeSH terms for the document in the MEDLINE article indexing. This manuscript summarizes the BioCreative NLM-Chem track and post-challenge experiments. We received a total of 85 submissions from 17 teams worldwide. The highest performance achieved for the chemical identification task was 0.8672 *F*-score (0.8759 precision and 0.8587 recall) for strict NER performance and 0.8136 *F*-score (0.8621 precision and 0.7702 recall) for strict normalization performance. The highest performance achieved for the chemical indexing task was 0.6073 *F*-score (0.7417 precision and 0.5141 recall). This community challenge demonstrated that (i) the current substantial achievements in deep learning technologies can be utilized to improve automated prediction accuracy further and (ii) the chemical indexing task is substantially more challenging. We look forward to further developing biomedical text–mining methods to respond to the rapid growth of biomedical literature. The NLM-Chem track dataset and other challenge materials are publicly available at https://ftp.ncbi.nlm.nih.gov/pub/lu/BC7-NLM-Chem-track/.

Database URL: https://ftp.ncbi.nlm.nih.gov/pub/lu/BC7-NLM-Chem-track/

Introduction

Identifying named entities is an essential building block for many complex knowledge extraction tasks. Errors in identifying relevant biomedical entities are a key impediment to accurate article retrieval, classification and further understanding of textual semantics, such as relation

Received 19 August 2022; Revised 6 January 2023; Accepted 15 February 2023

extraction (1). Chemical entities appear throughout the biomedical research literature and are among the most frequently searched entity types in PubMed (2). Accurate automated identification of the chemicals mentioned in journal publications can translate to improvements in many downstream natural language processing tasks and biomedical fields and, in the near term, specifically in retrieving relevant articles, greatly assisting researchers, indexers and curators (3, 4).

Automated methods to recognize and identify chemicals in biomedical text have a long history (5–7). Previous work in biomedical named entity recognition (NER) and normalization [i.e. entity linking (EL)] for chemicals includes several community challenges, including the Chemical Compound and Drug Name Recognition (CHEMDNER) (8) and BioCreative 5 Chemical-Disease Relation (BC5CDR) (9) tasks at previous BioCreative workshops. A 2014 review of entity annotations in biomedical corpora found chemicals to be one of the types most frequently annotated (10), and chemicals are one of many entity types included in the Colorado Richly Annotated Full Text (CRAFT) corpus of full-text articles (11). More recent work has included chemical patents (12, 13)—which are an important source of information for chemical compounds—and chemical reactions (14, 15), with an increased focus on chemical reactions. Automated indexing, on the other hand, has typically addressed all biomedical entity types simultaneously (16, 17). The specific challenges that chemical names pose for automated indexing have long been recognized (18), however, and some recent efforts focus specifically on chemicals (19).

Few of these efforts have addressed full-text articles, however, which are required for the indexing and curation tasks and where information retrieval and extraction differ. For example, the full text frequently contains more detailed information, such as chemical compound properties, biological effects and interactions with diseases, genes and other chemicals. Previous work released the first version of the National Library of Medicine (NLM)-Chem corpus (3), containing 150 full-text articles annotated for chemical spans and identifiers. This work also demonstrated that training systems with full-text articles (rather than only abstracts) results in substantially higher performance for both NER and normalization.

During the coronavirus disease 2019 pandemic, the biomedical literature experienced a rapid influx of articles as researchers searched to understand severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) and provide effective prevention strategies and treatments (20). The resulting information overload underscores the importance of automated methods to assist in searching, categorizing and extracting information from the literature (21). Correctly identifying chemicals is essential to these tasks, especially for the timely identification of potential treatments of both the acute and long-term effects of SARS-CoV-2 infection (22).

To support the efforts of increasing the efficiency and accuracy of the current state-of-the-art algorithms and foster the efforts of researching novel methods and achievements, the NLM-Chem track at BioCreative VII brought together the community to address two tasks:

- Chemical identification in full text: predict all chemicals mentioned in recently published full-text articles, both span (i.e. NER) and normalization (i.e. EL) using Medical Subject Headings (MeSH, https://www.nlm.nih.gov/mesh) (23).
- Chemical indexing task: predict which chemicals mentioned in recently published full-text articles should be indexed, i.e. appear in the list of MeSH terms for the document.

We developed a rich and comprehensive full-text corpus of chemical mentions to support the challenge. Each article in this resource contains manual annotations for chemical entities mentioned in the full text and manual indexing for the chemical substances representing the topic and content of the article with respect to chemicals. This resource is detailed in a study by Islamaj *et al.* (24) and is available from https://ftp.ncbi.nlm.nih.gov/pub/lu/BC7-NLM-Chem-track/.

The NLM-Chem track attracted worldwide participation. Ultimately, 17 teams submitted predictions. We received 53 valid submissions for the chemical identification task, of which 50 were official, and the rest submitted after the deadline. We received 18 valid submissions for the chemical indexing task, of which 5 were official, and the rest were submitted after the deadline. For the chemical identification task, 73% and 29% of the teams, respectively, had higher performance than the benchmark system provided by the organizers for the strict NER and normalization metrics. For the chemical indexing task, 50% of the teams had higher strict performance than the benchmark indexing system provided by the organizers. Participating teams explored different methodologies, with the majority focusing on deep learning architectures. Both data and evaluation scripts are available from the workshop webpage, the same link mentioned earlier. We encourage further participation from interested teams in developing biomedical text–mining methods to predict chemical mentions (identification) and chemical topic terms (indexing) in biomedical full-text articles.

After the challenge concluded, we performed several additional experiments. First, we identified a small set of straightforward improvements most frequently reported by the task participants and used these to update the benchmark systems. For the chemical identification task, these improvements resulted in a statistically significant 0.0331 increase in the strict *F*-score for the NER evaluation. For the chemical indexing task, these improvements resulted in a statistically significant increase of 0.1382 in the strict *F*-score over the original benchmark system.

Second, after the conclusion of the challenge, we created a silver-standard corpus for chemical identification in full-text articles (24). To the best of our knowledge, this corpus is the first of its kind. We created the silver-standard corpus from an ensemble of all 53 valid submissions to the chemical identification task, plus the original benchmark system. This manuscript provides a summary of the methods for the ensemble predictions and for creating the silver-standard corpus; the complete details are given in a study by Islamaj *et al.* (24). In this manuscript, we show that adding the silver-standard corpus to the training data for the NER component of the updated chemical identification benchmark system results in an additional statistically significant increase of 0.0167 in the strict *F*-score measure for NER.

Third, we updated the indexing for the 1333 articles in the NLM-Chem-BC7 Chemical Indexing corpus (24). We identified frequent differences between the indexing terms predicted by the challenge participants and the publicly available MeSH indexing. These were then presented to 11 expert indexers at the NLM to be accepted or rejected without knowing the source. This manuscript describes the methods used to create the indexing submissions; a description of the manual annotation is provided in a study by Islamaj *et al.* (24). In this manuscript, we provide an evaluation of both the chemical indexing submissions and the updated benchmark system with respect to the updated annotations.

This article provides an extended description of the BioCreative NLM-Chem track. First, it summarizes the NLM-Chem-BC7 corpus and the evaluation methods used. It further describes the methods used by the participating teams and the benchmark system. Finally, the results of the task are presented in detail along with several post-challenge experiments to (i) improve the benchmark systems using the insights from the participants and (ii) create a silver-standard corpus. We show that using the silver-standard corpus as training data improves performance on the benchmark systems.

Methods

We announced the BioCreative VII NLM-Chem chemical recognition challenge in full-text articles in Spring 2021. The NLM-Chem corpus was made available as the training dataset in May 2021, with the BC5CDR and CHEMDNER corpora as additional data. A webinar was held in May 2021 for interested teams to introduce them to the motivation for the challenge and associated data. The testing dataset for the chemical identification task, which complements the NLM-Chem-BC7 Chemical Identification task corpus, was manually annotated in April–June 2021. The testing dataset for the chemical indexing task was manually indexed via the regular indexing pipeline at the NLM in September 2021. All materials provided to the participants, including a recording of the webinar, are publicly available at https://ftp.ncbi.nlm.nih.gov/pub/lu/BC7-NLM-Chem-track/.

The NLM-Chem-BC7 corpus

The BioCreative community challenges aim to evaluate text-mining and information extraction systems applied to the biological domain. The main emphasis is on comparing methods for scientific progress rather than on the purely competitive aspects. The most accurate measure of progress for identifying and indexing chemicals from full-text articles is recently published and previously unseen. Therefore, an appropriate training dataset consists of articles that span a variety of journals, are rich in chemical mentions and cover a plethora of chemical-related topics to be representative of biomedical literature publications that contain chemical mentions (3). Furthermore, given the goal of providing training data for chemical identification and indexing in full-text articles, the dataset needs to contain expert-annotated full-text articles. We describe the NLM-Chem-BC7 dataset in detail in a study by Islamaj *et al.* (24) and give an overview in the following.

We specifically selected the NLM-Chem-BC7 corpus articles to (i) have no restrictions on sharing and distribution, (ii) be useful for other downstream biomedical text–mining tasks and (iii) be suitable for testing real-world tasks. The article selection was therefore focused on recently published articles.

The organizers provided three collections of articles to the participants.

The NLM-Chem-BC7 Chemical Identification task training corpus consists of the 150 full-text articles described in a study by Islamaj *et al.* (3), doubly annotated by 12 expert NLM indexers for all chemical mentions and their corresponding MeSH identifiers. Typically, an article is indexed with all medical subject terms that describe and identify its content, but for our study, we focus on chemical mentions. For this challenge, we augmented the NLM-Chem dataset with 54 additional full-text articles, recently published in Spring 2021, to serve as the testing dataset for the chemical identification task. These articles were doubly annotated by the same group of experts, following the same annotation guidelines. These articles were enriched with the indexing terms corresponding to their chemical substances, as assigned during the regular manual indexing process.

We re-purposed the CHEMDNER (8) and the BC5CDR (9) corpora for the NLM-Chem track challenge. The CHEMDNER documents contain title/abstract annotations for chemical NER and do not include chemical normalization. However, as this could still be useful for training deep learning strategies, we converted all articles and their annotations into the BioC format (25), the same format as the NLM-Chem-BC7 corpus. The BC5CDR corpus contains title/abstract chemical annotations and their MeSH identifiers; they were also converted to the BioC format. All articles were enriched with their chemical substance indexing terms assigned by the NLM indexers during the normal indexing process. The indexing data were filtered to select only the indexing terms representing chemical substances and provided in the same format.

Finally, the last collection, the NLM-Chem-BC7 Chemical Indexing corpus, consisted of 1333 recently published articles, which served as the test dataset of the chemical indexing task. These articles were published in Spring 2021 and underwent the normal manual indexing process at the NLM during September 2021 after completing the NLM-Chem track challenge. These indexed labels were used as gold-standard data for task evaluation.

Benchmark methods for chemical identification and indexing

In previous work (3), the NLM-Chem track organizers (authors Re.I., R.L. and Z.L.) described a benchmark tool for chemical entity recognition and normalization to illustrate the value of the NLM-Chem full-text corpus. This tool was based on BlueBERT (https://github.com/ncbi-nlp/bluebert), a variant of the Bidirectional Encoder Representations from Transformers (BERT) transformer language model (26) trained on PubMed abstracts and clinical notes, which was introduced with the Biomedical Language Understanding Evaluation (BLUE) benchmark (27). We fine-tuned the BlueBERT model (BlueBERT-Base, Uncased, PubMed + MIMIC-III) to perform chemical NER using the combined BC5CDR (9) and NLM-Chem training sets. Chemical mentions are assigned MeSH identifiers by our sieve-based normalization system

Multiple Terminology Candidate Resolution (MTCR), which is optimized for chemical mention normalization. MTCR first resolves all abbreviations that appear in the mention text using abbreviation definitions identified by Ab3P in full-text articles (28). Then, each mention text is mapped to a set of candidate MeSH concepts using multiple string-matching methods, applied in sequence, with the first method that returns a non-zero number of MeSH concepts used as the overall result. The earlier methods in the sequence provide higher precision, while the later methods provide higher recall. These methods can be briefly summarized as follows: 'exact match' to terminology vocabulary (MeSH); 'relaxed match' allows for certain lexical variations in the sequence; 'relaxed plural match' processes tokens using a conservative plural stemmer and 'relaxed match' to multiple chemical terminologies.

We adapted this chemical tagger to provide comparison methods for the chemical identification and chemical indexing tasks. For the chemical identification task, we updated the transformer NER model to BioBERT and the normalization component to use the 2021 version of MeSH. Thus, the comparison method for chemical identification sets a very high benchmark. For the chemical indexing task, we added a component to return the MeSH identifiers from annotations found in the title and abstract as the set of indexed chemicals. The indexing component thus represents a straightforward baseline approach with relatively low precision but higher recall.

After the conclusion of the challenge, we made two improvements to the benchmark methods, which are hereafter called the improved benchmark systems. Both improvements are based on reports from the task participants and are discussed further in the Discussion section. First, while BERT-based transformer models consistently provide strong performance for chemical NER, many participants in the chemical identification task reported noticeable performance increases when using a variant of BERT trained with a biomedical vocabulary. We, therefore, replaced the BERT variant used in the NER component with PubMedBERT (29). Second, for the chemical indexing task, we noted that the information most frequently used to determine if a chemical should be indexed is the document structure and how frequently the chemical is mentioned, echoing previous work on identifying focus entities in scientific documents (30). Therefore, we created an improved indexing benchmark system that uses the number of times a chemical is mentioned in each document section (e.g. title, abstract, introduction and methods) as features for a logistic regression model to predict whether that chemical should be indexed. This model must be trained using the output of the chemical identification system as input; however, the performance of the chemical identification system would be artificially high using the training data. We, therefore, created a separate training set by selecting ∼1200 open-access full-text articles published within the last 20 years with publicly available indexing. Half were chosen randomly from journals whose name includes either 'chem' or 'molec', and half were chosen completely at random. We then trained the model using the output of the improved benchmark system for chemical identification as features and the publicly available MeSH indexing as the gold-standard labels. We applied a threshold to the probabilistic predictions to trade off between precision and recall.

Ensemble methods

We explored ensemble predictions for both chemical identification and chemical indexing tasks. Ensemble methods combine predictions from multiple sources, frequently providing higher performance than any single prediction alone. Ensemble predictions over a sufficiently large corpus can therefore be used as a silver-standard corpus for training later automated systems (8). Ensemble methods have been applied previously in many biomedical natural language processing contexts, notably in the CHEMDNER task for chemical NER (8).

To create the ensemble predictions for the chemical identification task, we performed separate procedures for the NER and normalization subtasks. For NER, we gathered the set of spans from the chemical mentions in all submissions, including the original benchmark results. We then added a score to each mention span representing the proportion of submissions that included the span. Next, we applied a threshold, dropping any mentions below the threshold. Any mentions that overlapped after applying the threshold were combined into a single mention using the lowest start index of the overlapping mentions as the new start index and, similarly, the highest end index of the overlapping mentions as the new end index. For normalization, we gather all (document identifier, mention text, MeSH identifier) tuples from the submissions and perform normalization by choosing the MeSH identifier most frequently associated with each (document identifier, mention text) pair.

We combined the results from all submissions to the chemical indexing task into an ensemble prediction, again including the original benchmark results. This ensemble gathers the (document identifier, MeSH identifier) tuples from the chemical indexing task submissions and adds a score representing the proportion of the submissions that contain the indexing value.

Finally, while the participants returned chemical identification annotations for 1387 articles, only 54 were used for evaluating the chemical identification task. Therefore, we performed an experiment to determine if the ensemble prediction could be used to annotate the remainder as a silver-standard training set, focusing primarily on NER. We created the NLM-Chem-BC7 Chemical Indexing silver-standard corpus by applying the same procedure as the ensemble prediction to the 1333 articles not used to evaluate the chemical identification task. The resulting dataset contains 392 838 chemical mention spans, corresponding to 33 209 unique mention texts and 12 301 MeSH identifiers. The full description of the NLM-Chem-BC7 Chemical Indexing silver-standard corpus can be found in a study by Islamaj *et al.* (24). We use this corpus as a silver-standard training set for NER by choosing and applying a threshold and then combining overlapping mentions. While silver-standard training sets may be used in several ways, in this work we simply add it to the training data for the NER model, so that the full training data consist of the NLM-Chem training set, the BC5CDR training set and the NLM-Chem-BC7 Chemical Indexing silver-standard corpus.

Evaluation measures

The evaluation metrics used to assess team predictions were micro-averaged recall, precision and *F*-scores. Three different

result types were scored: false negative (FN) results correspond to incorrect negative predictions, false positive (FP) predictions correspond to incorrect positive predictions and true positives (TP) results correspond to correct predictions. Recall r (also known as coverage, sensitivity, true positive rate or hit rate) is the percentage of correctly labeled positive results over all positive cases, $r = TP/(TP+FN)$. Precision p (positive predictive value) is the percentage of correctly labeled positive results over all positive labeled results, $p = TP/(TP+FP)$. The *F*-measure F_β is the harmonic mean between precision and recall, where β is a parameter for the relative importance of precision over recall. $F_\beta = ((1+\beta^2)\cdot p\cdot r)/(\beta^2 p + r)$. The balanced *F*-measure (F_β), referred to as '*F*-score' in this work, can be simplified to $F_1 = 2\cdot p\cdot r/(p+r)$ and is the primary evaluation metric.

We measure the precision, recall and *F*-score measures in a strict and approximate evaluation setting. Furthermore, the chemical identification task consists of chemical NER and normalization using MeSH identifiers. The strict evaluation for NER requires an exact match between the predicted mention span and the annotated mention span. The strict evaluation for normalization tasks requires an exact match between the set of predicted MeSH identifiers and the set of annotated MeSH identifiers for the full text. The approximate evaluation for NER considers a predicted mention span to match an annotated mention span if they overlap. For chemical entity normalization, which is evaluated in the chemical identification task and the chemical indexing task, the approximate evaluation is the least common ancestor *F*-score (31). This measure identifies an approximately minimal set of ancestor identifiers sufficient to ensure that all predicted and annotated identifiers have at least one common ancestor in the set. Both the set of predicted identifiers and the set of annotated identifiers are then augmented with the ancestor set. The augmented sets are then evaluated as for the strict normalization measure. This procedure allows partial credit for predictions that are closely related to the reference identifier. Note, however, that if many ancestors must be added, the approximate performance measures may be lower than the strict performance measures.

Similar to previous BioCreative challenges, statistical significance was measured using bootstrap resampling (32, 33). While the test set for the chemical identification task contains tens of thousands of chemical mentions, it only contains 54 documents. Each sample for the chemical identification task was created by sampling 54 documents with replacement and then sampling passages with replacement within each document until achieving the same number of passages as the original document. This procedure results in pseudo-documents that vary from the original but remain representative—simply sampling passages, for example, skews the normalization evaluation since it is calculated with respect to the set of identifiers at the document level. We selected 10 000 samples, calculating the precision, recall and *F*-score for each submission. We then considered a submission to have statistically significantly higher performance than another submission if it has a higher *F*-score for at least 95% of the samples. We only calculated statistical significance for the strict measures and followed the same procedure for both the NER and normalization evaluations. We also used bootstrap resampling for the chemical indexing task but instead sampled entire documents. We again selected 10 000 samples and considered a submission to have statistically significantly higher performance than another submission if it has a higher *F*-score for at least 95% of the samples.

Participation and team descriptions

We received 85 submissions from a total of 17 teams. The participating teams represent nine nations from Europe, Asia and North America. Two teams were from industry, with the remainder from universities. The teams reported sizes of 2–7 (average 4), typically with backgrounds in natural language processing, machine learning, information retrieval and computer science. Seventeen teams submitted a total of 59 runs for the chemical identification task, 6 of which failed the evaluation script and were not evaluated further. Three of the remaining 53 runs were considered unofficial because they were submitted after the deadline. For the chemical indexing task, 8 teams submitted a total of 26 runs, 8 of which failed the evaluation script and were not evaluated further. Of the remaining 18 runs, 13 were considered unofficial because they were submitted after the deadline. To illustrate the variety of solutions for these tasks, a subset of the teams submitted system descriptions for this article. These teams are summarized in Table 1, along with their affiliations and the number of submissions to each task. The system descriptions follow in order of increasing team number.

Team 110: DETI/IEETA, University of Aveiro; Rui Antunes and João Rafael Almeida (identification and indexing tasks)

We participated in the NLM-Chem track using a three-stage pipeline composed of deep learning and rule-based strategies that solved chemical recognition, normalization and indexing steps.

Table 1. Summary of the teams contributing system descriptions for this overview of the NLM-Chem track, along with the number of valid submissions for each task

Team number and citation	Team affiliation	Submission count	
		Identification	Indexing
110 (34)	DETI/IEETA, University of Aveiro	5	5, 5[a]
114 (35)	University of Western Ontario and Al Baha University	1	
121 (36)	Graduate Institute of Data Science, Taipei Medical University	5	
128 (37)	Department of Statistics, Florida State University	4	4[a]
130 (38)	Scientific Databases and Visualization Group, Heidelberg Institute for Theoretical Studies	1[a]	
139 (39)	DMIS Lab at Korea University	5	
141 (40)	Medicines Discovery Catapult	1	1[a]
143 (41)	NVIDIA	2	
157 (42)	Toyota Technological Institute	5	3[a]

[a]Unofficial submissions.

In the first step, we tackled the problem of detecting chemical mentions as a traditional NER task using the beginning, inside, outside (BIO) tagging scheme for token-level classification (34). The text was represented using contextualized word embeddings, calculated from PubMedBERT (29), which were forwarded through a conditional random field (35) layer. We experimented with several corpora for training our deep learning model. Besides the original NLM-Chem corpus (3, 36) and the BC5CDR (9) and CHEMDNER (8) datasets prepared by the organizers, we also preprocessed the DrugProt dataset (37) and used some of the datasets—CRAFT, BioNLP11ID, BioNLP13CG and BioNLP13PC—prepared by Crichton *et al.* (38).

In the next step, a sieve-based normalization process linked the chemical mentions recognized with their corresponding MeSH identifiers. This workflow was composed of two sequential components: a rule-based system and a deep learning model. The rule-based component used exact text matching to map chemical mentions to their corresponding MeSH identifiers. Mentions were mapped using a dictionary consisting of entry terms and concept identifiers from MeSH, filtered to only retain concepts belonging to the 'Drugs and Chemical' category; abbreviation mappings obtained through the use of the Ab3P tool (28) and finally, mappings from training gold-standard annotations to improve dictionary coverage. Any remaining unmapped mentions were processed using a deep learning approach: First, SapBERT (39) was used to create dense representations for the remaining entities and all MeSH identifiers. Then, we measured the cosine similarity between each entity representation and all MeSH identifier representations, and for each entity, we selected the MeSH identifier with the highest similarity and above a specific threshold.

For the chemical indexing task, we tested a rule-based strategy and another based on term frequency-inverse document frequency (TF-IDF) scores. The rule-based approach started by extracting the MeSH identifier present in specific parts of the documents, namely the title, abstract and table and figure captions. We hypothesized that these sections of the document would be the most appropriate to find mentions of MeSH terms of interest. For each section, we defined frequency occurrence thresholds to reduce false positives. For example, if the MeSH identifier was recognized in the title, this identifier needed a percentage of occurrence equal or superior to 10% within the document. In the second strategy, we modeled the importance of each MeSH identifier using the TF-IDF weighting scheme.

Team 114: University of Western Ontario and Al Baha University; Robert E. Mercer and Mohammed A. Alliheedi (identification task)

For the NLM-Chem track, our team decided to take an already trained model and modify its performance with a rule-based system rather than design and train a neural model from scratch. This method achieved moderate success.

The first step was to use Stanza to tokenize the text (40), and then some rules are invoked to modify this tokenization. The initial tokenization step uses the package from the genome information acquisition project (41). The output from this step is then analyzed with some rules that dealt with parentheses and other punctuation. Once the text is tokenized, it is processed using the Stanza named entity recognizer bc4chemd. Stanza tags each token as 'Other', or as single-token chemical names, or as multi-word chemical names using 'Begin', 'Internal' and 'End' tags. The output from Stanza is analyzed with 12 rules. One rule dealt with acronyms not found by Stanza, two rules tidied up some minor errors and the remaining rules added modifying words and phrases that were annotated as part of the chemical names.

The next step is the normalization process. We downloaded MeSH 2021 and extracted all text items associated with each MeSH term. These are loaded in our Python program as a dictionary and mapped using case-insensitive exact match to provide the normalized MeSH identifier for any chemical name recognized. The last task is to provide the location and length of each chemical name.

The original Stanza finds 62.9% of the unique chemical names in the training set. After adding the *post hoc* rule-based corrections, Stanza finds 71.6% of the unique chemical names in the training set.

Team 121: Graduate Institute of Data Science, Taipei Medical University; Wen-Chao Yeh and Yung-Chun Chang (identification task)

We developed a BERT-based ensemble learning approach to recognize chemical entities. To determine which pretrained BERT model is best suited for this task, we evaluated the performance of BERT pretrained models with biomedical vocabularies using the NLM-Chem corpus under 10-fold cross-validation. The results demonstrated that the model by Alrowili and Shanker (42) and the PubMedBERT (29) model achieved the best performance out of all models tested. We then used ensemble learning with a majority voting mechanism to integrate them, selected via *k*-fold cross-validation.

To perform chemical normalization, we adopted a dynamic programming-based method to link the MeSH identifier of recognized chemical entities. First, we extracted the MeSH identifiers from all training datasets as a knowledge base. Next, we use this knowledge base to map all chemical mentions predicted to their respective identifiers. We quantify the string similarity between chemical mentions identified in the text and terms in the knowledge base with edit distance, retrieving the most similar MeSH term and identifier if the string similarity is >90%. We used parallel programming to save search time for a high volume of identification terms.

For chemical NER, we achieved *F*-scores of 0.8521 and 0.9183 on the strict and approximate evaluations, respectively. Moreover, our model achieves high performance for chemical normalization, with *F*-scores of 0.8072 and 0.8015 on the strict and approximate evaluations, respectively.

Team 128: Department of Statistics, Florida State University; Arslan Erdengasileng and Keqiao Li (identification and indexing tasks)

We developed a hybrid pipeline for chemical NER, chemical normalization and chemical indexing as follows:

Chemical NER: we first tried several variants of the BERT model, and PubMedBERT (29) achieved the highest *F*-score. We also trained models using different combinations of the current and related datasets (NLM-Chem, BC5CDR and CHEMDNER) to see if adding more data

would increase the performance. Our experiment showed that adding more data did not improve the NER performance. We then used a data augmentation technique, replacing each chemical mention (and a subset of non-chemical mentions) with random text, and found the procedure improved the performance (43).

From the PubMedBERT output, we added some post-processing steps. First, Ab3P, an abbreviation definition detector trained on PubMed abstracts, was used to recognize abbreviations in the text (28). The full names and their abbreviations are linked within the same articles, and all occurrences received the same NER label. Second, chemical names can be part of other entity names, such as protein entities. We trained another BioBERT-based protein NER model to detect protein entities. The goal was to remove wrongly labeled chemical names that are part of protein names. Our best *F*-score on the test set is 0.8600.

Chemical normalization: we built a sieve-based pipeline using multiple dictionaries: MeSH (https://www.nlm.nih.gov/mesh), Unified Medical Language System (UMLS) (44), PubTator annotations (45) and the NLM-Chem corpus. First, we scanned through dictionaries in a specific order. Second, we used Ab3P to process those unmapped chemicals to detect abbreviations and their long forms. Last, we scanned the long forms of abbreviations detected through the dictionaries in the same order in the first step. Our best *F*-score on the test set is 0.8101.

Chemical indexing: we built a binary MeSH indexing classification system using a PubMedBERT model with engineered features. In this strategy, we dealt with one MeSH term at a time by predicting whether it should be indexed. To remove the noise from the long text, we broke up full texts into sentences and selected the sentences with chemical mentions of the corresponding MeSH terms as input to the model. The labels are simply 'True' or 'False' based on whether or not the MeSH terms were used for indexing the articles. We added engineered features before the sentences, such as the section where the sentences were taken from and the chemical names whose MeSH terms were to be predicted. Our best *F*-score on the test set is 0.5681.

Team 130: Scientific Databases and Visualization Group, Heidelberg Institute for Theoretical Studies; Ghadeer Mobasher and Lukrécia Mertová (identification task)

We based our chemical entity identification model on the biomedical pretrained language model BioBERT (46), fine-tuned using the BC5CDR-chemicals (9) and CHEMDNER corpora (8), which is publicly available (https://huggingface.co/alvaroalon2/biobert_chemical_ner). We conducted our work on a professional-grade NVIDIA Tesla cluster, using one of its graphics processing unit nodes to run the model on the NLM-Chem corpora and evaluating fine-tuned BioBERT using the evaluation script provided by the organizers.

For chemical EL, we adopt a rule-based approach (47). The key concept is parsing the chemical entity into meaningful sub-word components and then creating a comprehensive comparison. Each chemical is analyzed and parsed via the semantic analysis module and then processed via a similarity search module that queries internally stored reference databases. For this purpose, we enriched the MeSH database of descriptors (23) with synonyms from the PubChem database (48). The semantic analysis module uses rules according to International Union of Pure and Applied Chemistry nomenclature (49) to separate components and assign different priorities according to the position and type. The similarity estimation provided by the similarity search module deals with chemical-based errors and typos by considering the similarity of each chemical feature separately.

We have evaluated the performance of the fine-tuned BioBERT, trained on BC5CDR chemicals and CHEMDNER corpora, using the evaluation script provided by the organizers. The parameter set used remains unmodified from the default setting. For our future work, we aim to resolve the issue arising from BERT tokenizers, develop joint learning of chemical NER and EL using pretrained transformer-based models and compare their performance with our preliminary approach (50).

Team 139: DMIS Lab at Korea University; Hyunjae Kim and Mujeen Sung (identification task)

Our overall system consists of two separate components, each responsible for NER and normalization. NER is accomplished through three steps: transfer learning, model ensemble and majority voting. In transfer learning, NER models, which are Bio-LM-large models with a linear layer (51), are first trained on a single source data and then fine-tuned on the target data (i.e. NLM-Chem). For source data, existing datasets such as BC5CDR and CHEMDNER and the new synthetic data NLM-Chem(syn) are used, which are created by replacing entity mentions in the original NLM-Chem data with their synonyms using the Comparative Toxicogenomics Database (CTD) version released on 1 April 2021 (http://ctdbase.org/). Twenty differently trained NER models are combined to form an ensemble model, where the model aggregates predictions of every single model and outputs the majority as the final prediction. The output predictions of the ensemble model are modified by majority voting, a rule-based post-processing method. Majority voting collects all predictions within an article, identifies the most common prediction for each input and changes all other predictions for the same input to match the most common.

In normalization, predicted entities in the NER stage are mapped into MeSH identifiers by a hybrid model consisting of a dictionary model and a neural model. The dictionary model first performs normalization based on exact matching between predicted entities and the dictionary, which is the same CTD version as in synonym replacement. The neural model further performs the process on entities that are not matched by the dictionary model. For the neural model, we use BioSyn (52) with the SapBERT encoder (39).

Team 141: Medicines Discovery Catapult; Robert Bevan and Matthew Hodgskiss (identification task)

We experimented with two approaches to chemical entity recognition while developing our system. First, we fine-tuned a PubMedBERT transformer model (29) using the NLM-Chem dataset. Next, we fine-tuned two additional transformers: one using the CHEMDNER dataset (8) and one using the BC5CDR dataset (9). We then trained a stacking model using the outputs of the three models in the hope it would offer an improvement over the single transformer model. We observed no such improvement and used the single transformer model in our final system. An additional challenge to consider when

working with the NLM-Chem dataset is that it comprises full articles, which contain passages with lengths greater than the 512 tokens PubMedBERT can process. We excluded these passages during training. For the evaluation, we split long passages into overlapping sequences of 512 tokens, processed these sequences individually and averaged the outputs.

Our normalization system is built on the normalization sieve approaches presented in previous work (3, 53). We began by extending the MeSH database by adding synonyms from the following sources: Chemical Entities of Biological Interest (ChEBI) (54), UMLS (44), NCI Thesaurus (55) and Unique Ingredient Identifier (UNII) (https://fdasis.nlm.nih.gov/) and PubChem (https://pubchem.ncbi.nlm.nih.gov/). Next, we normalized each of the terms in the database. More specifically, we lowercased and removed whitespace from each term and substituted Greek characters with their English equivalents. We implemented a set of string manipulation and substitution methods to try to improve recall. The string manipulation methods included stemming and non-alphanumeric character removal. We adopted two string substitution methods: abbreviation expansion using Ab3P (28), where abbreviations are substituted for their typically less ambiguous long form and similar entity substitution using BioWordVec embeddings (56). Similar entity substitution works as follows: an entity is converted to a BioWordVec embedding, and the 10 most similar embeddings—measured by cosine similarity—are identified. Embeddings with cosine similarity below some empirically determined threshold are discarded. The strings corresponding to the remaining embeddings are then resolved to MeSH identifiers where possible, and the final MeSH identifier is determined by majority vote. We evaluated the precision of each data source and string manipulation/substitution method combinations and used this information to construct the optimal normalization sieve.

For the task of predicting whether an identified MeSH identifier belongs to an article's index, we trained a logistic regression model using the following features: mention count, normalized mention count, the number of different passages the identifier is mentioned in, passage-type mention count (e.g. identifier mentioned three times in the abstract) and the identifier identity.

Team 143: NVIDIA; Virginia Adams and Hoo-Chang Shin (identification task)

For the NLM-Chem track, we perform NER using prompting-based approaches with generative pre-training (GPT) (57) and T5-(58) style generative models. We use BERT-based NER models as our baseline. Our experiments show that the BioMegatron BERT–based model (59) performed better than BioBERT (46) baseline, possibly due to larger model and pre-training text corpus sizes. We use model ensemble to boost the final performance by an additional ∼5%.

For our prompting experiments, our GPT (57) and T5 (58) models have a similar number of parameters as our BioMegatron model (345 M parameters). Text-to-text, or prompting-based, methods are attractive, as many different tasks can be expressed as natural language questions and answers and the fine-tuning objective for prompt-style tasks aligns with the pretraining objective of these generative models. Additionally, because all tasks share the same fine-tuning objective, a model can be tuned for multiple tasks at once, potentially benefiting from shared knowledge between tasks. We adopt this approach to convert NER into a natural language, text-to-text problem. When preprocessing the training data and converting them into a format conducive to prompt-style question answering, long paragraphs are split into two to four sentences and a manually formatted prompt is added to the sentence. For example, we found prefacing sentences with the phrase 'find entities:' to be successful. In addition to the NLM-Chem track NER task, we also used a similar prompting-based set-up for the topic indexing and EL tasks with varying levels of success.

For EL, we also adopt the self-alignment pretraining approach described by Liu *et al.* (39). The idea behind this approach is to use a metric-learning loss to reshape the initial BioBERT embedding space such that synonyms of the same concept are pulled closer together and unrelated concepts are pushed further apart. The concept embeddings from this reshaped space can then be used to build a knowledge base embedding index. This index stores MeSH identifiers mapped to their respective concept embeddings in a format conducive to efficient nearest neighbor search. We link query concepts to their canonical forms in the knowledge base by performing a nearest neighbor search—matching concept query embeddings to the most similar concept embedding in the knowledge base index. We use the UMLS dataset to pretrain BioBERT for EL.

Team 157: Toyota Technological Institute; Tomoki Tsujimura and Ryuki Ida (identification and indexing tasks)

We built fully neural NER, linking and indexing models along with a TF-IDF-based indexing model to tackle the NLM-Chem track. We treated the NER task as a sequential tagging problem using the beginning, inside, last, outside, unit scheme, which identifies the beginning, inside and last tokens of multi-token mentions, as well as tokens outside of an entity mention or mentions comprising a single token unit. We decoded the entity labels via the Viterbi algorithm. Our NER model was a fine-tuned model of SciBERT (60) with a classification layer for the task. Our NER model achieved a strict *F*-score of 0.8284.

We employed a model built at n2c2 2019 shared-task Track 3 (61) for the EL task. We regarded the EL task as a classification problem over all entities in the MeSH thesaurus. Our model took the surface form of an entity as the input and encoded it using the SciBERT encoder. The encoded representation was then fed to the cosine similarity-based output layer, producing the probability distribution over all entities. We automatically built additional training instances from synonyms recorded in the MeSH thesaurus to cover unseen entities during training. We also tried to augment the training instance using the BC5CDR dataset (9), but the model without augmentation produced our best result. To mitigate the underfitting problem, we overwrote the weight matrix of the output layer with the mean of the representations from the input once during training, which we also performed when we participated in n2c2 2019. Finally, we employed Ab3P (28) at preprocessing to expand the abbreviations in the target documents. Our best test *F*-score was 0.7954.

We built a neural indexing model that took a bag of sentences as the input for the indexing task. For each MeSH entity in the target document, we gathered sentences containing the

Table 2. Official chemical identification results for strict and approximate NER measures

		Strict				Approximate		
Rank	Team/run	Precision	Recall	*F*-score	Signif.	Precision	Recall	*F*-score
1	139/3	**0.8759**	0.8587	**0.8672**	2–16	**0.9373**	0.9161	**0.9266**
2	128/1	0.8544	**0.8658**	0.8600	4–16	0.9220	0.9304	0.9262
3	143/1	0.8535	0.8608	0.8571	4–16	0.9271	0.9235	0.9253
4	121/2	0.8461	0.8583	0.8521	6–16	0.9152	0.9215	0.9183
5	141/1	0.8338	0.8654	0.8493	9–16	0.8953	**0.9309**	0.9127
6	104/2	0.8687	0.8249	0.8463	9–16	0.9273	0.8791	0.9025
7	148/1	0.8692	0.8239	0.8459	9–16	0.9277	0.8761	0.9011
8	110/4	0.8394	0.8515	0.8454	11–16	0.9040	0.9229	0.9134
9	149/1	0.8226	0.8614	0.8416	11–16	0.8951	0.9204	0.9076
10	146/4	0.8219	0.8622	0.8415	11–16	0.8945	0.9235	0.9088
11	157/1	0.8476	0.8101	0.8284	12–16	0.9128	0.8670	0.8893
12	Benchmark	0.8440	0.7877	0.8149	14–16	0.9156	0.8492	0.8811
13	155/1	0.8312	0.7967	0.8136	14–16	0.9009	0.8596	0.8798
14	114/1	0.7219	0.5897	0.6492	15–16	0.8348	0.6919	0.7567
15	130/1[a]	0.7208	0.5211	0.6049	16	0.8933	0.6331	0.7410
16	116/3	0.8234	0.1916	0.3109		0.9196	0.215	0.3485

[a]Unofficial runs. This table only shows the run for each team with the highest strict *F*-score; complete results are shown in Supplementary Materials. Runs are ordered by strict *F*-score, descending. The strict *F*-score for each row is statistically significantly higher ($P < 0.05$) than the rows marked in the Signif. column. The highest value in each column is marked in bold.

entity and formed a bag of sentences. We inserted special tokens into the sentences to mark the entity spans. Our model encoded each sentence with the special tokens to the sentence representation by the SciBERT encoder. Then, the attention module aggregated the sentence representations to form the bag representation. The bag representation was fed to the output layer to predict the probability of being the index. We also built an indexing model based on TF-IDF (62). The TF-IDF model calculated the TF-IDF value for each entity and made decisions by comparing the value with the threshold. We found that our neural indexing model is more sensitive to noise than the TF-IDF model. As a result, our neural model and TF-IDF model got *F*-scores of 0.2736 and 0.4745, respectively. As for the training, we trained each neural model five times with different random seeds and took the average of the outputs as the ensemble for the test submission. We tuned the hyperparameters using Optuna (63).

Results

Chemical identification results

The official chemical identification results for the NER measures (strict and approximate) are shown in Table 2. Each team was allowed up to five submissions, labeled Run 1 through Run 5; we show only the run with the highest strict *F*-score for each team and provide the complete results in Supplementary Material. The highest strict *F*-score was 0.8672 (Team 139, Run 3), and the highest-performing submission of 11 of the 15 teams had a higher strict *F*-score than the original benchmark. The approximate *F*-scores are highly correlated with the strict *F*-scores, even though ordering by approximate *F*-score produces slightly different rankings for the intermediate results (i.e. Ranks 5–10). While the absolute difference between many of the results is relatively small, the average difference between ranks which is statistically significant is 1.80. We also observe that the strict *F*-score of the Rank 1 submission is statistically significantly higher than the Rank 2 submission.

The official chemical identification results for the normalization measures (strict and approximate) are shown in Table 3. Each team was again allowed up to five submissions, labeled Run 1 through Run 5. We again only show the submission with the highest strict *F*-score for each team, with the complete results provided in Supplementary Material. Note that the team submission reported in Table 3 may differ from the submission reported for the same team in Table 2 since each table only shows the submission with the highest strict *F*-score for each evaluation. The highest strict *F*-score was 0.8136 (Team 110, Run 4), and the highest-scoring submission for 4 of the 14 teams had a higher strict *F*-score than the original benchmark. The approximate *F*-scores are highly correlated with the strict *F*-scores, even though slightly less than for NER. While the same submission had the highest strict *F*-score and approximate *F*-score, ordering by approximate *F*-score produces significantly different rankings overall. For example, the submissions ranked 2, 3 and 4 according to approximate *F*-scores were ranked 3, 7 and 2, respectively, when ranked by strict *F*-score. Considering statistical significance, we see that the Rank 1 submission does not have statistically significantly higher performance than the Rank 2 submission, unlike the NER evaluation. There are also a few submissions starting at Rank 5 (the original benchmark) with performances that are not statistically significantly different. We observe that the Rank 7 submission (Team 139, Run 2) has statistically significantly higher performance than the Rank 8 submission, even though the Rank 5 and 6 submissions do not; we believe this is due to the sampling procedure slightly favoring the high recall of this submission. Finally, while the absolute difference between many of the results is again relatively small, the average difference between ranks which is statistically significant is 1.64.

Chemical indexing results

Table 4 reports the chemical indexing results (strict and approximate) for the run using the NLM-Chem-BC7 Chemical Indexing corpus. Each team was allowed up to five official

Table 3. Official chemical identification results for strict and approximate normalization measures

		Strict				Approximate		
Rank	Team/run	Precision	Recall	*F*-score	Signif.	Precision	Recall	*F*-score
1	110/4	**0.8621**	0.7702	**0.8136**	3–15	**0.8302**	0.7867	**0.8030**
2	128/2	0.7792	0.8434	0.8101	4–15	0.7258	0.8679	0.7864
3	121/1	0.7874	0.8281	0.8072	5–15	0.7530	0.8643	0.8015
4	157/3	0.7338	**0.8683**	0.7954	5–15	0.6954	**0.8976**	0.7760
5	Benchmark	0.8151	0.7644	0.7889	9–15	0.7917	0.7889	0.7857
6	141/1	0.7890	0.7849	0.7870	9–15	0.7192	0.8254	0.7628
7	139/2	0.7256	0.8505	0.7831	8–15	0.7113	0.8966	0.7883
8	155/1	0.7886	0.7644	0.7763	9–15	0.7309	0.7917	0.7551
9	104/1	0.6720	0.7475	0.7078	10–15	0.6319	0.8097	0.7039
10	149/2	0.6645	0.7451	0.7025	12–15	0.6260	0.8174	0.7033
11	148/4	0.6481	0.7629	0.7008	12–15	0.6043	0.8281	0.6923
12	146/1	0.5931	0.7816	0.6744	13–15	0.5418	0.8558	0.6581
13	114/1	0.8334	0.4645	0.5965	14–15	0.8273	0.5279	0.6368
14	143/1	0.4326	0.6541	0.5208	15	0.4418	0.8108	0.5664
15	130/1[a]	0.4575	0.4449	0.4511		0.5461	0.6093	0.5662

[a]Unofficial runs. This table only shows the run for each team with the highest strict *F*-score; complete results are shown in Supplementary Materials. Runs are ordered by strict *F*-score, descending. The strict *F*-score for each row is statistically significantly higher ($P < 0.05$) than the rows marked in the Signif. column. The highest value in each column is marked in bold.

Table 4. Official chemical indexing results with both strict and approximate measures using the updated indexing

		Strict				Approximate		
Rank	Team/Run	Precision	Recall	*F*-score	Signif.	Precision	Recall	*F*-score
1	110/1	**0.7417**	0.5141	**0.6073**	2–5	**0.7526**	0.6779	**0.6743**
2	128/1[a]	0.5506	0.5867	0.5681	3–5	0.6437	0.7545	0.6507
3	Benchmark	0.4057	**0.7005**	0.5138	4–5	0.5233	0.8536	0.6135
4	157/1[a]	0.3553	0.7137	0.4745	5	0.4614	**0.8729**	0.5687
5	141/1[a]	0.4958	0.3061	0.3785		0.5459	0.4626	0.4687

[a]Unofficial runs. Runs are ordered by strict *F*-score, descending. The strict *F*-score for each row is statistically significantly higher ($P < 0.05$) than the rows marked in the Signif. column. The highest value in each column is marked in bold.

submissions and an additional five unofficial submissions, labeled Run 1 through Run 10; we show only the run with the highest strict *F*-score for each team and provide the complete results in Supplementary Material. The NLM-Chem-BC7 Chemical Indexing corpus was updated by expert NLM indexers in a double-reviewed, blind annotation experiment using frequent differences between the indexing terms predicted by the challenge participants and the publicly available MeSH indexing. Comparing the results for all submissions using the NLM-Chem-BC7 Chemical Indexing corpus to the results using the original indexing gold standard, reported in previous work (64), demonstrates both higher precision and higher recall for all submissions, with the *F*-score increasing an average of 0.0982 (values shown in Supplementary Material). We also observe differences with the original submission rankings; this is expected, however. The highest strict *F*-score was 0.6073, and two of the four teams have a strict *F*-score higher than the original benchmark. The approximate *F*-scores are substantially higher than the strict scores (+0.0887 on average). The approximate *F*-scores are highly correlated with the strict *F*-scores, and ranking by approximate *F*-score is very similar to ranking by strict *F*-score. Finally, we observe that all differences in the strict *F*-score are statistically significant.

Ensemble results

We evaluated the ensemble NER results using the strict measure, determining the precision, recall and *F*-score as a function of the proportion of submissions containing the mention. Any overlapping mentions remaining after thresholding were combined into a single mention using the lowest start index of the overlapping mentions as the new start index and, similarly, the highest end index of the overlapping mentions as the new end index. The results are presented in Figure 1. The maximum *F*-score is 0.8732 at a score threshold of 0.3519. When the threshold is high, precision is high, and recall is low, as expected. However, when the threshold is low, precision and recall are both low due to combining overlapping mentions.

We evaluated the normalization ensemble results using the strict normalization measures. The normalization ensemble approach does not allow for tuning, so we determined the performance as a function of the proportion of submissions containing the annotation at the NER level. The results are presented in Figure 2. The maximum *F*-score is 0.8092 at a score threshold of 0.6667. As with the NER, when the threshold is high, precision is high, and recall is low, as expected. However, when the threshold is low, precision and recall are both low due to combining overlapping mentions.

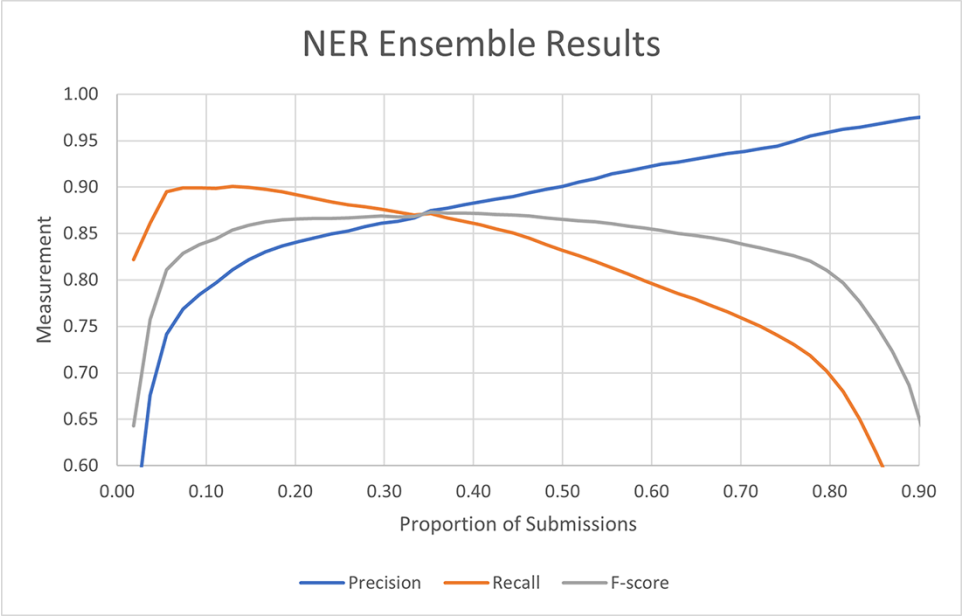

Figure 1. Performance of the ensemble of all submissions using the strict NER measures as a function of the proportion of submissions containing the mention.

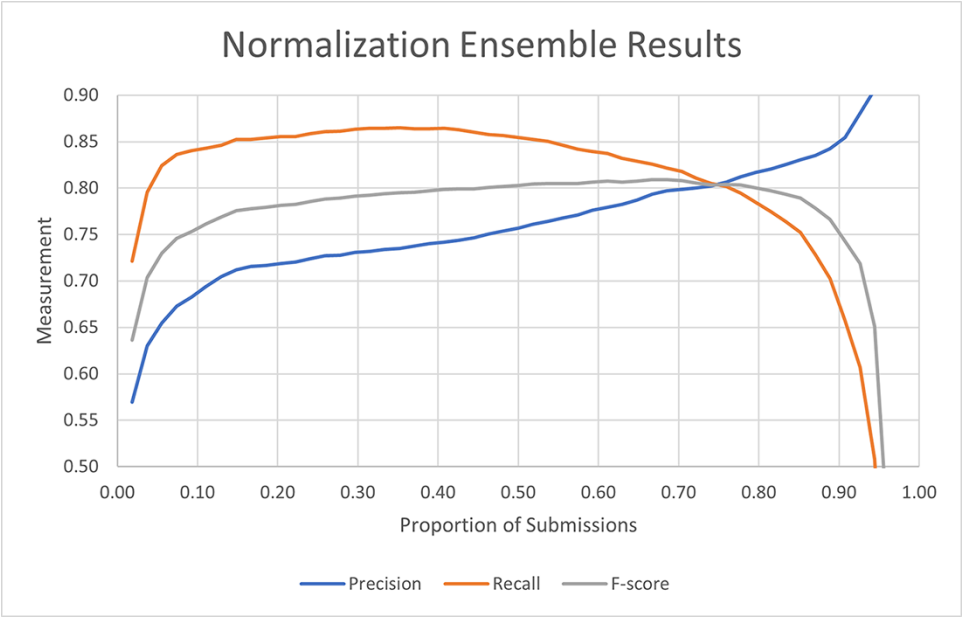

Figure 2. Performance of the ensemble of all submissions using the strict normalization measures as a function of the proportion of submissions containing the mention.

The evaluation of the indexing ensemble results can be seen in Figure 3. The score represents the proportion of submissions to the chemical indexing task that included the identifier. The maximum *F*-score is 0.7474 at a score threshold of 0.4737.

Improved benchmark and silver-standard results

Table 5 summarizes the results of the improved benchmark systems before and after adding the silver-standard corpus to the training data for the NER component. We show only the *F*-score of the strict evaluation for each task to highlight the improvements. All increases in the *F*-score are statistically significant, using the same procedures as for used to determine statistical significance for the official submissions ($P < 0.05$). For the silver-standard experiment, we chose a threshold of 0.35, which is equivalent to the threshold that produced the highest *F*-score for the NER ensemble. The improved indexing benchmark also requires a threshold; we used 0.07, which maximized the performance on the training data.

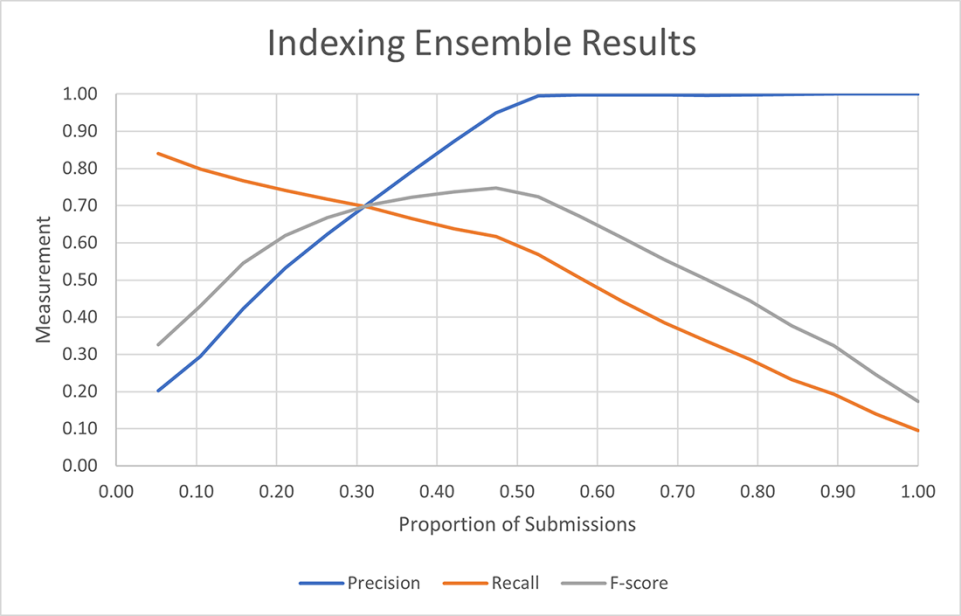

Figure 3. Indexing performance of the ensemble of all submissions as a function of the proportion of submissions including the identifier.

Table 5. Comparison of results of the original benchmark systems, improved benchmark systems and the improved benchmark systems with the NER component trained on the silver-standard corpus

Task and evaluation	Original benchmark system	Improved benchmark system (change)	Improved benchmark + silver standard (change)
NER, strict	0.8149	0.8480 (+0.0331)	0.8647 (+0.0498)
Normalization, strict	0.7889	0.8149 (+0.0260)	0.8213 (+0.0324)
Indexing (updated)	0.5138	0.6520 (+0.1382)	n/a

All values are strict *F*-scores, and all increases in the strict *F*-score are statistically significant ($P < 0.05$). Since the silver-standard corpus and the indexing evaluation set comprise the same documents, we do not report indexing values using the NER component trained on the silver-standard corpus.

Discussion

In this section, we analyze the submissions to the NLM-Chem track and the post-challenge experiments. First, we analyze the submissions to the chemical identification task, followed by the chemical indexing task. We then analyze the results of the ensemble experiments and the silver-standard experiment.

Analysis of chemical identification task team submissions

Most teams separated the chemical identification task into distinct NER and normalization components, combined using a pipeline approach. For NER, systems based on BERT transformer models such as BioBERT (46) were popular and performed well. However, BioBERT uses a general vocabulary, and many teams noted that BERT variants using a vocabulary intended for biomedical text, such as PubMedBERT (29) and BioMegatron (59), have noticeably higher performance on the NLM-Chem corpus (43, 65–67). We further note that the one update to the benchmark for the chemical identification task was changing from BioBERT to PubMedBERT, resulting in a substantial performance improvement (0.0331 *F*-score for the strict NER evaluation). The strong performance improvement provided by using a BERT model with a biomedical vocabulary may reflect the specialized terminology used by chemical names.

In line with previous work on NER in chemicals (68), Teams 121, 139, 141 and 143 reported that ensemble methods are effective (65–67, 69). In addition, several teams reported that fine-tuning a BERT model directly on the additional datasets (BC5CDR and CHEMDNER) resulted in lower performance than models fine-tuned on only the NLM-Chem data. However, Team 139 reported increased performance by pretraining on the additional datasets, followed by fine-tuning on NLM-Chem (67). Finally, Teams 128 and 139—the teams with the highest NER performance—augmented their training sets with synthetic data, either by replacing the annotated chemical mentions with chemical names from a lexicon or random strings (43, 67).

The normalization methods were significantly more varied than those for NER; however, most teams used a hybrid of two or more methods, often in a sieve configuration (53). Almost all teams used a direct string match as their primary approach, employing an approximate match approach only if the direct match was unsuccessful. While abbreviations have many potential senses, these can be handled separately, and polysemy is otherwise not a critical issue for this task. The direct string-matching approach provides high precision and, if used with string transformations, also provides modest recall.

Most teams also used an approximate match for higher recall, but these methods varied. Teams 110, 139, 141 and 157 reported converting chemical mentions and names to dense vectors and then identifying the closest matches using cosine similarity (67, 69–71). The results show that this approach seems to have been particularly effective at achieving high recall; the five runs submitted by Team 157 are notable for

achieving the five highest recall values for the strict normalization evaluation (71). Team 121 found edit distance to be a useful approximate match method, although computationally expensive (66). Usage of additional chemical name resources seems to have been relatively limited, even though teams did report using UMLS (44), PubChem (48), NCI Thesaurus (55), ChEBI (54), UNII (https://fdasis.nlm.nih.gov/srs/) and also the chemical mappings of PubTator (45). The teams did not seem to employ a separate step to filter non-chemicals from the results, instead relying on the NER results to determine whether the span was a chemical or not. The three teams that predicted the greatest number of chemical annotations with composite identifiers (i.e. more than one MeSH identifier per mention) were also the three teams with the highest *F*-scores for the strict normalization evaluation. Overall, the normalization scores were significantly lower than the NER scores and not as tightly clustered toward the high end of performance. We note that strong NER performance is necessary for good normalization performance but is not sufficient. In particular, the submission with the highest normalization score was ranked 15th in the NER scores. Finally, some teams reported cascading errors from the NER system, suggesting that an end-to-end approach might be beneficial.

Analysis of chemical indexing task team submissions

The task participants reported finding the chemical indexing task significantly more challenging than the chemical indexing task. This subjective evaluation is supported by the number and types of submissions—a total of 18 valid submissions, with 13 unofficial—and by the significantly lower performance relative to the chemical identification task. However, the evaluated performance of all submissions to the chemical indexing task is higher using the NLM-Chem-BC7 chemical indexing corpus than the original MeSH indexing (24), suggesting that the real utility of automated algorithms for MeSH term indexing is higher than can be measured using the publicly available MeSH indexing terms.

The teams reported that the most readily available information for determining whether a specific chemical identifier should be indexed is the document structure—where the chemical is mentioned in the document—followed by frequency. This observation is in line with previous work on identifying key entities in scientific documents (30) and was the basis for the improved benchmark method for the chemical indexing task, which used the number of times the chemical appeared in each section as features, resulting in a substantial performance improvement. Team 128 found a binary classifier with engineered features to be effective (43). Teams 110 and 157 both reported using hybrid methods, including a TF-IDF variant, to prioritize the chemical identifiers found during the identification task (70, 71).

Analysis of ensemble and silver-standard results

The ensemble results for NER show higher performance than any single participant in the task, underscoring the effectiveness of ensemble methods for NER tasks. Moreover, the *F*-score is not overly sensitive to the threshold chosen for the proportion of submissions. For normalization, however, the maximum performance is somewhat less than the performance achieved by the submission with highest performance. We note that the maximum performance for normalization is at a threshold of 0.67, much higher than the threshold for the maximum NER performance (0.35). This higher threshold implies a need for higher precision and may be related to the normalization evaluation ignoring the number of times the chemical was identified in the document.

Error analysis of the NER ensemble results shows that the mention texts most often missed are considerably shorter than mentions on average. These mentions primarily comprise abbreviations (e.g. 'PET'), with some chemical formulas (e.g. 'H_2O_2') and common names (e.g. 'water'). The mention texts most often predicted erroneously are similar. However, we also see some partial mentions such as 'sulfur' instead of 'sulfur isotopes' or chemicals that are a part of a name for a non-chemical, such as 'glutathione peroxidase'.

Error analysis of the normalization ensemble results, on the other hand, is somewhat more diverse. For missed identifiers, the most apparent pattern is that many are compound, such as 'MeSH:D012965,MeSH:D014325', referring to sodium chloride and tromethamine, from the mention 'Tris-buffered saline'. While compound identifiers are uncommon overall, many participants made no attempt to identify them. Other identifiers frequently missed include common chemicals (e.g. 'water'), especially high-level concepts (e.g. 'lipids'). Identifiers often predicted incorrectly include technetium (MeSH:D013667) for the mention 'TC', which typically refers to total cholesterol.

The silver-standard experiment demonstrates that an ensemble of systems can be used to produce training data that increase the accuracy of the NER model. Moreover, the improved results for NER also translate to improved performance for normalization. Unfortunately, we could not evaluate the improvements for the indexing since the silver-standard documents are the same as those for the indexing evaluation. However, it seems reasonable to expect some indexing improvements.

The ensemble results for indexing show substantially higher performance than the submission with the highest performance (+0.1401). Moreover, above a score threshold of ~0.57, the precision of the ensemble is >0.99. Error analysis of the indexing ensemble shows that missed identifiers form two types. The first type is very common chemicals, such as glucose, water and reactive oxygen species. The second type is not mentioned directly in the article where they are indexed; for example, glycosides (MeSH:D005960) is indexed in 12 articles but is not directly mentioned in any, including in the full text.

Conclusion

We presented the NLM-Chem track at BioCreative VII, consisting of two tasks: chemical identification—corresponding to chemical NER and normalization (EL)—and chemical indexing. Of the teams that submitted to the chemical identification task, 73% achieved a higher strict *F*-score than the benchmark system for NER, while 29% achieved the same for normalization. Participants described several improvements that reliably improved the quality of the results for the chemical identification task. We demonstrated improvements to the benchmark identification system using only one of these: updating the deep learning transformer NER system to use a model trained with a biomedical vocabulary. Of the teams

that submitted to the chemical indexing task, 50% achieved a higher strict *F*-score than the benchmark system. Our results suggest that the utility of automated indexing predictions may be higher than that can be demonstrated using the publicly available MeSH indexing. Finally, we create a straightforward but substantially improved benchmark for chemical indexing.

Supplementary material

Supplementary material is available at *Database* online.

Data availability

The NLM-Chem track dataset and other challenge materials are publicly available at https://ftp.ncbi.nlm.nih.gov/pub/lu/BC7-NLM-Chem-track/.

Funding

This research was supported in part by the NIH Intramural Research Program, National Library of Medicine. R.A. and J.R.A. are supported by Portuguese national funds through the Foundation for Science and Technology (FCT) under the grants SFRH/BD/137000/2018 and SFRH/BD/147837/2019, respectively. R.E.M. is supported by the Natural Sciences and Engineering Research Council of Canada. M.A.A. is supported by Scientific Research Deanship, Al Baha University, Saudi Arabia (grant number 1441/2). G.M. is part of the PoLiMeR-ITN (http://polimer-itn.eu/) and is supported by European Union's Horizon 2020 research and innovation program under the Marie Skłodowska–Curie grant agreement PoLiMeR, No. 81261. V.A. and H.C.S. are employees of NVIDIA Corporation.

Conflict of interest

None declared.

References

1. Leaman,R., Wei,C.H., Allot,A. *et al.* (2020) Ten tips for a text-mining-ready article: how to improve automated discoverability and interpretability. *PLoS Biol.*, **18**, e3000716.
2. Islamaj Dogan,R., Murray,G.C., Neveol,A. *et al.* (2009) Understanding PubMed user search behavior through log analysis. *Database (Oxford)*, **2009**, bap018.
3. Islamaj,R., Leaman,R., Kim,S. *et al.* (2021) NLM-Chem, a new resource for chemical entity recognition in PubMed full text literature. *Sci. Data*, **8**, 91.
4. Kim,S., Thiessen,P.A., Cheng,T. *et al.* (2016) Literature information in PubChem: associations between PubChem records and scientific articles. *J. Cheminform.*, **8**, 32.
5. Johnson,H.L., Cohen,K.B., Baumgartner,W.A., Jr. *et al.* (2006) Evaluation of lexical methods for detecting relationships between concepts from multiple ontologies. *Pac. Symp. Biocomput.*, **2006**, 28–39.
6. Corbett,P., Batchelor,C. and Teufel,S. (2007) Annotation of chemical named entities. In: *Biological, translational, and clinical language processing*. Prague, Czech Republic, June 29, 2007, pp. 57–64.
7. Klinger,R., Kolarik,C., Fluck,J. *et al.* (2008) Detection of IUPAC and IUPAC-like chemical names. *Bioinformatics*, **24**, i268–i276.
8. Krallinger,M., Rabal,O., Leitner,F. *et al.* (2015) The CHEMDNER corpus of chemicals and drugs and its annotation principles. *J. Cheminform.*, **7**, 1–17.
9. Li,J., Sun,Y., Johnson,R.J. *et al.* (2016) BioCreative V CDR task corpus: a resource for chemical disease relation extraction. *Database (Oxford)*, **2016**.
10. Neves,M. (2014) An analysis on the entity annotations in biological corpora. *F1000Res*, **3**, 96.
11. Bada,M., Eckert,M., Evans,D. *et al.* (2012) Concept annotation in the CRAFT corpus. *BMC Bioinform.*, **13**, 161.
12. Krallinger,M., Rabal,O., Lourenco,A. *et al.* (2017) Information retrieval and text mining technologies for chemistry. *Chem. Rev.*, **117**, 7673–7761.
13. He,J., Nguyen,D.Q., Akhondi,S.A. *et al.* (2021) ChEMU 2020: natural language processing methods are effective for information extraction from chemical patents. *Front. Res. Metr. Anal.*, **6**, 654438.
14. Guo,J., Ibanez-Lopez,A.S., Gao,H. *et al.* (2022) Automated chemical reaction extraction from scientific literature. *J. Chem. Inf. Model*, **62**, 2035–2045.
15. Yoshikawa,H., Nguyen,D.Q., Zhai,Z. *et al.* (2019) Detecting chemical reactions in patents. In: *The 17th Annual Workshop of the Australasian Language Technology Association*. Sydney, Australia, pp. 100–110.
16. Aronson,A.R. (2001) Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. *AMIA Symp.* 2001, 17–21.
17. Mork,J.G., Jimeno-Yepes,A. and Aronson,A.R. (2013) The NLM Medical Text Indexer System for indexing biomedical literature. In: *The First Workshop on Bio-Medical Semantic Indexing and Question Answering, a Post-Conference Workshop of Conference and Labs of the Evaluation Forum 2013 (CLEF 2013)*. CEUR Workshop Proceedings, Valencia, Spain.
18. Wilbur,W.J., Hazard,G.F., Jr., Divita,G. *et al.* (1999) Analysis of biomedical text for chemical names: a comparison of three methods. *Proc. AMIA Symp.*, 176–180.
19. Savery,M.E., Rogers,W.J., Pillai,M. *et al.* (2020) Chemical entity recognition for MEDLINE indexing. *AMIA Jt. Summits Transl. Sci. Proc.*, **2020**, 561–568.
20. Chen,Q., Allot,A. and Lu,Z. (2021) LitCovid: an open database of COVID-19 literature. *Nucleic Acids Res.*, **49**, D1534–D1540.
21. Chen,Q., Leaman,R., Allot,A. *et al.* (2021) Artificial intelligence in action: addressing the COVID-19 pandemic with natural language processing. *Annu. Rev. Biomed. Data Sci.*, **4**, 313–339.
22. Leaman,R., Islamaj,R., Allot,A. *et al.* (2022) Comprehensively identifying long Covid articles with human-in-the-loop machine learning. *Patterns (N Y)*, **4**, 100659.
23. Lipscomb,C.E. (2000) Medical Subject Headings (MeSH). *Bull. Med. Libr. Assoc.*, **88**, 265–266.
24. Islamaj,R., Leaman,R., Cissel,D. *et al.* (2022) NLM-Chem-BC7: manually annotated full-text resources for chemical entity annotation and indexing in biomedical articles. *Database (Oxford)*, **2022**.
25. Comeau,D.C., Islamaj Dogan,R., Ciccarese,P. *et al.* (2013) BioC: a minimalist approach to interoperability for biomedical text processing. *Database (Oxford)*, **2013**, bat064.
26. Devlin,J., Chang,M.-W., Lee,K. *et al.* (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, Volume 1 (Long and Short Papers)*. Association for Computational Linguistics, Minnesota, Minneapolis, pp. 4171–4186.
27. 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: *18th BioNLP Workshop and Shared Task*. Association for Computational Linguistics, Florence, Italy. pp. 58–65.

28. Sohn,S., Comeau,D.C., Kim,W. *et al.* (2008) Abbreviation definition identification based on automatic precision estimates. *BMC Bioinform.*, **9**, 402.
29. Gu,Y., Tinn,R., Cheng,H. *et al.* (2021) Domain-specific language model pretraining for biomedical natural language processing. *ACM Trans. Comput. Healthc.*, **3**, 1–23.
30. Yepes,A.J., Albahem,A. and Verspoor,K. (2021) Using discourse structure to differentiate focus entities from background entities in scientific literature. In: *Proceedings of the 19th Annual Workshop of the Australasian Language Technology Association: Australasian Language Technology Association*. Sydney, Australia. pp. 174–178.
31. Tsatsaronis,G., Balikas,G., Malakasiotis,P. *et al.* (2015) An overview of the BIOASQ large-scale biomedical semantic indexing and question answering competition. *BMC Bioinform.*, **16**, 138.
32. Smith,L., Tanabe,L.K., Ando,R.J. *et al.* (2008) Overview of BioCreative II gene mention recognition. *Genome Biol.*, **9**, 1–19.
33. Krallinger,M., Leitner,F., Rabal,O. *et al.* (2015) CHEMDNER: the drugs and chemical names extraction challenge. *J. Cheminform.*, **7**, 1–11.
34. Ramshaw,L. and Marcus,M. (1995) Text chunking using transformation-based learning. In: *Third Workshop on Very Large Corpora*. Cambridge, Massachusetts, June 1995, pp. 82–94.
35. Lafferty,J.D., McCallum,A. and Pereira,F.C.N. (2001) Conditional random fields: probabilistic models for segmenting and labeling sequence data. In: *Eighteenth International Conference on Machine Learning*. June 2001, pp. 282–289.
36. Islamaj,R., Leaman,R., Cissel,D. *et al.* (2021) The chemical corpus of the NLM-Chem BioCreative VII track: full-text chemical identification and indexing in PubMed articles In: *Proceedings of the Seventh BioCreative Challenge Evaluation Workshop*. Virtual conference, November 2021.
37. Miranda,A., Mehryary,F., Luoma,J. *et al.* (2021) Overview of DrugProt BioCreative VII track: quality evaluation and large scale text mining of drug-gene/protein relations. In: *Proceedings of the Seventh BioCreative Challenge Evaluation Workshop*. Virtual Conference, November 2021.
38. Crichton,G., Pyysalo,S., Chiu,B. *et al.* (2017) A neural network multi-task learning approach to biomedical named entity recognition. *BMC Bioinform.*, **18**, 368.
39. Liu,F., Shareghi,E., Meng,Z. *et al.* (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*. Virtual Conference, June 2021, pp. 4228–4238.
40. Zhang,Y., Zhang,Y., Qi,P. *et al.* (2021) Biomedical and clinical English model packages for the Stanza Python NLP library. *J. Am. Med. Inform. Assoc.*, **28**, 1892–1899.
41. Tsuruoka,Y. and Tsujii,J. (2005) Bidirectional inference with the easiest-first strategy for tagging sequence data. In: *Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*. Vancouver, British Columbia, Canada, 6–8 October 2005, pp. 467–474.
42. Alrowili,S. and Shanker,V.. (2021) BioM-transformers: building large biomedical language models with BERT, ALBERT and ELECTRA. In: *20th Workshop on Biomedical Language Processing*. Online, June 11, 2021, pp. 221–227.
43. Erdengasileng,A., Li,K., Han,Q. *et al.* (2021) A BERT-based hybrid system for chemical identification and indexing in full-text articles. In: *Proceedings of the Seventh BioCreative Challenge Evaluation Workshop*. Virtual Conference, November 2021.
44. Bodenreider,O. (2004) The Unified Medical Language System (UMLS): integrating biomedical terminology. *Nucleic Acids Res.*, **32**, D267–D270.
45. Wei,C.H., Allot,A., Leaman,R. *et al.* (2019) PubTator Central: automated concept annotation for biomedical full text articles. *Nucleic Acids Res.*, **47**, W587–W593.
46. Lee,J., Yoon,W., Kim,S. *et al.* (2020) BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, **36**, 1234–1240.
47. Mertová,L. (2020) Framework for automatised annotation of biochemical entities. Masaryk University.
48. Kim,S., Chen,J., Cheng,T. *et al.* (2021) PubChem in 2021: new data content and improved web interfaces. *Nucleic Acids Res.*, **49**, D1388–D1395.
49. McNaught,A.D. and Wilkinson,A. (1997) Compendium of Chemical Terminology - IUPAC Recommendations: Blackwell Science.
50. Mobasher,G., Mertová,L., Ghosh,S. *et al.* (2021) Combining dictionary and rule-based approximate entity linking with tuned BioBERT. In: *Proceedings of the Seventh BioCreative Challenge Evaluation Workshop*. Virtual Conference, November 2021.
51. Lewis,P., Ott,M., and Du,J. (2020) Pretrained language models for biomedical and clinical tasks: understanding and extending the state-of-the-art. In: *3rd Clinical Natural Language Processing Workshop*. Online, November 19, 2020, pp. 146–157.
52. Sung,M., Jeon,H., Lee,J. *et al* (2020) Biomedical entity representations with synonym marginalization. In: *58th Annual Meeting of the Association for Computational Linguistics*. Online, July 2020, pp. 3641–3650.
53. D'Souza,J. and Ng,V. (2015) Sieve-based entity linking for the biomedical domain. In: *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*. Beijing, China, July 2015. pp. 297–302.
54. Hastings,J., de Matos,P., Dekker,A. *et al.* (2013) The ChEBI reference database and ontology for biologically relevant chemistry: enhancements for 2013. *Nucleic Acids Res.*, **41**, D456–D463.
55. Sioutos,N., de Coronado,S., Haber,M.W. *et al.* (2007) NCI Thesaurus: a semantic model integrating cancer-related clinical and molecular information. *J. Biomed. Inform.*, **40**, 30–43.
56. Zhang,Y., Chen,Q., Yang,Z. *et al.* (2019) BioWordVec, improving biomedical word embeddings with subword information and MeSH. *Sci. Data*, **6**, 52.
57. Brown,T., Mann,B., Ryder,N. *et al.* (2020) Language models are few-shot learners. *Adv. Neural Inf. Process Syst.*, **33**, 1877–1901.
58. Raffel,C., Shazeer,N., Roberts,A. *et al.* (2020) Exploring the limits of transfer learning with a unified text-to-text transformers. *J. Mach. Learn. Res.*, **21**, 5485–5551.
59. Shin,H.-C., Zhang,Y., Bakhturina,E. *et al.* (2020) BioMegatron: larger biomedical domain language model. In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Virtual Conference, November 2020, pp. 4700–4706.
60. 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)*. Hong Kong, China, November 2019, pp. 3615–3620.
61. Henry,S., Wang,Y., Shen,F. *et al.* (2020) The 2019 National Natural Language Processing (NLP) Clinical Challenges (n2c2)/Open Health NLP (OHNLP) shared task on clinical concept normalization for clinical records. *J. Am. Med. Inform. Assoc.*, **27**, 1529–1537.
62. Manning,C.D., Raghavan,P. and Schütze,H. (2008) *Introduction to Information Retrieval*. Cambridge University Press, Cambridge, p. 506.
63. Akiba,T., Sano,S., Yanase,T. *et al.* (2019) Optuna: a next-generation hyperparameter optimization framework. In: *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD '19): Association for Computing Machinery*, Anchorage, AK, USA, 2019, pp. 2623–2631.

64. Leaman,R., Islamaj,R. and Lu,Z. (2021) Overview of the NLM-Chem BioCreative VII track: full-text chemical identification and indexing in PubMed articles. In: *Proceedings of the Seventh BioCreative Challenge Evaluation Workshop*. Virtual Conference, November 2021.
65. Adams,V., Shin,H.-C., Anderson,C. *et al.* (2021) Chemical identification and indexing in PubMed articles via BERT and text-to-text approaches. In: *Proceedings of the Seventh BioCreative Challenge Evaluation Workshop*. Virtual Conference, November 2021.
66. Chiu,Y.-W., Yeh,W.-C., Lin,S.-J. *et al.* (2021) Recognizing chemical entity in biomedical literature using a BERT-based ensemble learning methods for the BioCreative 2021 NLM-Chem track. In: *Proceedings of the Seventh BioCreative Challenge Evaluation Workshop*. Virtual Conference, November 2021.
67. Kim,H., Sung,M., Yoon,W. *et al.* (2021) Improving tagging consistency and entity coverage for chemical identification in full-text articles In: *Proceedings of the Seventh BioCreative Challenge Evaluation Workshop*. Virtual Conference, November 2021.
68. Leaman,R., Wei,C.H. and Lu,Z. (2015) tmChem: a high performance approach for chemical named entity recognition and normalization. *J. Cheminform.*, **7**, 1–10.
69. Bevan,R. and Hodgskiss,M. (2021) Fine-tuning transformers for automatic chemical entity identification in PubMed articles. In: *Proceedings of the Seventh BioCreative Challenge Evaluation Workshop*. Virtual Conference, November 2021.
70. Almeida,T., Antunes,R., Silva,J.F. *et al.* (2021) Chemical detection and indexing in PubMed full text articles using deep learning and rule-based methods. In: *Proceedings of the Seventh BioCreative Challenge Evaluation Workshop*. Virtual Conference, November 2021, pp. 119–123.
71. Tsujimura,T., Ida,R., Oiwa,I. *et al.* (2021) TTI-COIN at BioCreative VII Track 2: fully neural NER, linking, and indexing models. In: *Proceedings of the Seventh BioCreative Challenge Evaluation Workshop*. Virtual Conference, November 2021.