scientific data

OPEN

DATA DESCRIPTOR

EnzChemRED, a rich enzyme chemistry relation extraction dataset

Po-Ting Lai[1,4], Elisabeth Coudert[2,4], Lucila Aimo[2], Kristian Axelsen[2], Lionel Breuza[2], Edouard de Castro[2], Marc Feuermann[2], Anne Morgat[2], Lucille Pourcel[2], Ivo Pedruzzi[2], Sylvain Poux[2], Nicole Redaschi[2], Catherine Rivoire[2], Anastasia Sveshnikova[2], Chih-Hsuan Wei[1], Robert Leaman[1], Ling Luo[3], Zhiyong Lu[1]✉ & Alan Bridge[2]✉

Expert curation is essential to capture knowledge of enzyme functions from the scientific literature in FAIR open knowledgebases but cannot keep pace with the rate of new discoveries and new publications. In this work we present EnzChemRED, for Enzyme Chemistry Relation Extraction Dataset, a new training and benchmarking dataset to support the development of Natural Language Processing (NLP) methods such as (large) language models that can assist enzyme curation. EnzChemRED consists of 1,210 expert curated PubMed abstracts where enzymes and the chemical reactions they catalyze are annotated using identifiers from the protein knowledgebase UniProtKB and the chemical ontology ChEBI. We show that fine-tuning language models with EnzChemRED significantly boosts their ability to identify proteins and chemicals in text (86.30% F_1 score) and to extract the chemical conversions (86.66% F_1 score) and the enzymes that catalyze those conversions (83.79% F_1 score). We apply our methods to abstracts at PubMed scale to create a draft map of enzyme functions in literature to guide curation efforts in UniProtKB and the reaction knowledgebase Rhea.

Background & Summary

Knowledge of enzyme functions is critical to our understanding of how biological systems function and interact in complex communities[1–3], how genetic variation and disease affect those systems[4–6], and for efforts to engineer those systems to synthesize beneficial compounds such as drugs and biofuels or break down harmful environmental pollutants[7–13]. Expert curated knowledgebases such as the UniProt Knowledgebase (UniProtKB)[14,15], Rhea[16], MetaCyc[17], KEGG[18], BRENDA[19], SABIO-RK[20], Reactome[21] and the Gene Ontology (GO)[22] capture knowledge of enzymes and the reactions they catalyze from peer reviewed publications using human- and machine-readable chemical ontologies and cheminformatics descriptors in forms that are Findable, Accessible, Interoperable, and Reusable (FAIR)[23]. These databases play a critical role in biological and biomedical research but face significant challenges in keeping pace with the discovery and publication of new enzymes and reactions, with the result that much of our knowledge of how enzymes function remains "locked" in peer-reviewed publications in forms that are difficult for both humans and machines to access.

Natural Language Processing (NLP) methods offer a potential means to address this issue by accelerating the expert curation of enzyme functions, with large language models based on the transformer architecture such as BERT[24] among the most promising methods developed to date. The models are pre-trained using self-supervised approaches on large corpora of unannotated text using a language modeling objective and can be fine-tuned to perform specific tasks using curated domain-specific corpora in a process of transfer learning[25–28]. At the current time there is no freely available curated domain-specific dataset for fine-tuning language models to extract enzyme functions from text, and rule-based[29] and weakly supervised machine learning approaches[30] have been used instead.

[1]National Center for Biotechnology Information (NCBI), National Library of Medicine (NLM), National Institutes of Health (NIH), Bethesda, MD, 20894, USA. [2]Swiss-Prot group, SIB Swiss Institute of Bioinformatics, Centre Medical Universitaire, CH-1211, Geneva, 4, Switzerland. [3]School of Computer Science and Technology, Dalian University of Technology, 116024, Dalian, China. [4]These authors contributed equally: Po-Ting Lai, Elisabeth Coudert. ✉e-mail: zhiyong.lu@nih.gov; alan.bridge@sib.swiss

Year	Dataset	Documents	Chemical, protein, and gene mentions	Unique IDs	Relations	Application
2008	Corbett[32]	500 abstracts, 42 papers	11,571	—	—	NER
2008	SCAI[33]	100 abstracts	1,206	—	—	NER
2012	ADE[39]	300 case reports	5,063 drugs	—	6,821 drug adverse effects 279 drug dosage	RE
2013	DDI[43]	1,025, including texts from DrugBank and abstracts	18,502 drugs	—	5,028 drug-drug interactions	RE
2015	CHEMDNER[34]	10,000 abstracts	84,355 chemicals	—	—	NER
2016	BC5CDR[35]	1,500 articles	15,935 chemicals 12,850 diseases	4,409 MeSH	chemically induced diseases	NER, NEN, RE
2017	N-ary drug-gene-mutation[42]	—	—	—	137,469 drug–gene 3,192 drug–mutation	RE
2017	ChemProt[40]	2,482 abstracts	32,514 chemicals 30,922 genes	—	chemical-protein	RE
2019	DrugProt[41]	5,000 abstracts	65,561 chemicals 61,775 proteins	—	24,526 chemical-protein	RE
2020	EBED[38]	4,200, including abstracts, paragraphs, figure legends, and patents	16,715 chemicals 56,059 genes	5,161 ChEBI 12,563 Entrez	chemically induced diseases	NER, NEN, RE
2021	ChEMU 2020[44]	1,500 patent extracts	17,834 chemicals	—	chemical reaction steps	NER, RE
2022	BioRED[37,75]	1,000 abstracts	7,021 chemicals 12,412 genes	1,096 MeSH 2,605 Entrez	chemical-(chemical/disease/gene/variant)	NER, NEN, RE
2024	EnzChemRED (this work)	1,210 abstracts	18,887 chemicals 13,028 proteins	3,155 ChEBI 2,569 UniProtKB	chemical-chemical and (chemical-chemical)-protein	NER, NEN, RE

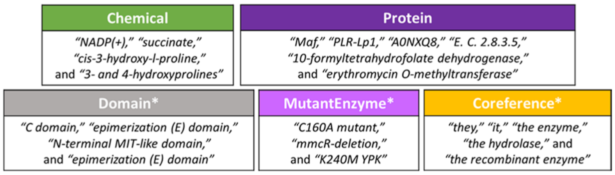
Table 1. Overview of datasets for chemical NLP.

In this work, we set out to develop a dataset that could be used to fine-tune language models and other NLP methods to assist curators in extracting knowledge of enzymes and their reactions from text. This dataset, EnzChemRED, for Enzyme Chemistry Relation Extraction Dataset, consists of 1,210 abstracts from PubMed in which the chemical conversions and the enzymes that catalyze them are curated using stable unique identifiers from UniProtKB and the chemical ontology ChEBI[31]. We propose a methodology to extract knowledge of enzyme functions from the literature by framing the problem as a series of NLP tasks, beginning with named entity recognition (NER), to identify text spans that denote enzymes and the chemicals they act on, named entity normalization (NEN), to link text spans to database identifiers, and relation extraction (RE), to link mentions of pairs of chemicals that define conversions (binary relations) and to link those conversions to mentions of enzymes that catalyze them (ternary relations). To establish a baseline for future research we present the results from fine-tuning pre-trained language models using EnzChemRED, achieving an F_1 score (harmonic mean of precision and recall) of 86.30% for NER, and 86.66% for binary RE and 83.79% for ternary RE. We combined these methods in a prototype end-to-end pipeline that performs literature triage, NER, NEN, and RE, and applied this pipeline to PubMed abstracts to create a draft map of enzyme functions in literature to guide and assist curation efforts in UniProtKB/Swiss-Prot and Rhea. We hope that the EnzChemRED benchmark dataset will be of broad utility for NLP researchers and the wider community of knowledgebase developers and biocurators of other resources, and welcome feedback and suggestions for further improvements to both the dataset and the methods described here.

Related works. Several prior works have described curated datasets to train and benchmark methods for the extraction of information about chemicals and their relations from scientific literature. Table 1 provides a chronology and summary of their main characteristics. They include datasets that address the problem of chemical NER, such as that of Corbett *et al.*[32], the SCAI dataset of Fluck and colleagues[33], and the CHEMDNER dataset[34], datasets that address chemical NEN, such as the BC5CDR dataset[35,36] and the Biomedical Relation Extraction Dataset (BioRED)[37], which both map chemical mentions to identifiers from the Medical Subject Headings (MeSH), and the EBED dataset[38], which maps chemical mentions to ChEBI, and datasets that address chemical RE, including the BC5CDR and EBED datasets that link chemicals to chemically induced diseases, the ADE dataset[39], that links drugs and adverse drug effects, the ChemProt[40] and DrugProt[41] datasets, that link chemicals to proteins, and the *n*-ary dataset of Peng and colleagues that links drugs, genes, and mutations[42]. Of the very few datasets that include chemical-chemical associations, most focus on interactions between drugs, such as the drug-drug interaction (DDI) dataset[43]. Only the BioRED and ChEMU lab 2020 datasets[44] include chemical conversions, but the BioRED dataset includes only seventeen chemical conversion pairs, which is too few for any meaningful assessment, while the ChEMU lab 2020 dataset focuses on organic chemistry from patents relevant to drug synthesis, and not on descriptions of enzyme functions. EnzChemRED differs from these prior works in that it focuses specifically on the problem of extracting chemical conversions catalyzed by enzymes, considers not only chemical conversions (binary relations that link pairs of chemicals) but also the enzymes that catalyze them (ternary relations that link conversions to enzymes), and addresses all the steps needed to extract those relations, namely NER, NEN, and RE.

```
PREFIX rh: <http://rdf.rhea-db.org/>
   PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
   PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
   PREFIX ECO: <http://purl.obolibrary.org/obo/ECO_>
   PREFIX up: <http://purl.uniprot.org/core/>
   SELECT ?rhea ?catalyzedReaction ?source
   WHERE {
      {
         SERVICE <https://sparql.rhea-db.org/sparql> {
            SELECT DISTINCT ?rhea
            WHERE {
               ?rhea rdfs:subClassOf rh:Reaction .
               ?rhea rh:side/rh:contains/rh:compound ?compound .
               ?compound rh:chebi ?chebi ;
               rdfs:subClassOf rh:SmallMolecule .
            }
         }
      }
      ?catalyzedReaction up:catalyzedReaction ?rhea .
      ?reif rdf:object ?catalyzedReaction ;
      up:attribution ?attr .
      ?attr up:evidence ECO:0000269 ;
      up:source ?source .
      ?source a up:Citation .
   }
```

Fig. 1 SPARQL query used to identify papers for abstract curation in EnzChemRED.



Chemical
"NADP(+)," "succinate," "cis-3-hydroxy-l-proline," and *"3- and 4-hydroxyprolines"*

Protein
"Maf," "PLR-Lp1," "A0NXQ8," "E. C. 2.8.3.5," "10-formyltetrahydrofolate dehydrogenase," and *"erythromycin O-methyltransferase"*

Domain*
"C domain," "epimerization (E) domain," "N-terminal MIT-like domain," and *"epimerization (E) domain"*

MutantEnzyme*
"C160A mutant," "mmcR-deletion," and *"K240M YPK"*

Coreference*
"they," "it," "the enzyme," "the hydrolase," and *"the recombinant enzyme"*

Fig. 2 Entity curation in EnzChemRED. We curated all Chemical and Protein mentions, but not of Domain, MutantEnzyme and Coreference (denoted by '*'), for which curation focused on those mentions that participate in conversions. The latter are not included in our evaluations.

Methods

The first subsection describes the development of the EnzChemRED, which is the main focus of this paper. The subsections that follow describe the development of a prototype end-to-end NLP pipeline for enzyme functions that makes use of EnzChemRED for fine-tuning and benchmarking. The last two subsections describe the combination of methods to create the end-to-end pipeline, as well as methods to process and visualize the output.

EnzChemRED development.

Selection of abstracts for curation. To build EnzChemRED we selected papers curated in UniProtKB/Swiss-Prot that describe enzyme functions. We queried the UniProt SPARQL endpoint (https://sparql.uniprot.org/) to identify papers that provided experimental evidence used to link protein sequence records from UniProtKB/Swiss-Prot to reactions from Rhea (specifically those reactions that involve only small molecules, excluding papers linked to Rhea reactions that involve proteins and other macromolecules) (Fig. 1). UniProtKB/Swiss-Prot uses evidence tags and the Evidence and Conclusions Ontology (ECO)[45] to denote provenance and evidence for functional annotations; our SPARQL query selected papers linked to Rhea annotations in UniProtKB/Swiss-Prot with evidence tags with experimental evidence codes from ECO, such as "ECO:0000269", which denotes "experimental evidence used in manual assertion". We also narrowed the selection of abstracts to those including mentions of at least one pair of reactants found in Rhea, and to those having a score of at least 0.9 according to our LitSuggest[46] model for abstracts relevant to enzyme function (see Section Literature triage). This LitSuggest score threshold is exceeded by 99% of abstracts of papers curated in Rhea. We selected 1,210 abstracts and divided these into 11 groups of 110 abstracts for curation by our team of expert curators.

Curation of chemical and protein mentions. Curation of abstracts in EnzChemRED was performed using the collaborative curation tool TeamTat (www.teamtat.org)[47] following a protocol based on that developed for the curation of the BioRED dataset (available at https://ftp.ncbi.nlm.nih.gov/pub/lu/BioRED/), with modifications described below. We curated five types of entity in EnzChemRED: Chemical, Protein, Domain, MutantEnzyme, and Coreference, which are described below, with examples shown in Fig. 2.

- **Chemical**: a mention of a chemical entity – including chemical structures, chemical classes, and chemical groups. Where possible we normalized chemical mentions to identifiers from ChEBI, which provides chemical structure information, and which is used to describe chemical entities in both Rhea and UniProtKB. Where no ChEBI identifier was available we used MeSH. A small number of chemical mentions have no mapping to either resource.

Relation type	Example	Explanation
Conversion	"**D-Dopachrome tautomerase** $_{P1}$ converts **2-carboxy-2,3-dihydroindole-5,6-quinone** $_{C1}$ (**D-dopachrome**) $_{C2}$ into **5,6-dihydroxyindole** $_{C3}$." – PMID: 9480844	P1 can convert C1 to C3, and C2 is a synonym for C1. We therefore curate two "Conversion" relations of (C1, C3) by P1 as "Converter", and of (C2, C3) by P1 as "Converter".
Indirect_conversion	"Cell extracts of Brucella abortus (British 19) catabolized **erythritol** $_{C1}$ through a series of phosphorylated intermediates to **dihydroxyacetonephosphate** $_{C2}$ and **CO-2** $_{C3}$. " – PMID:163226	C1 can be converted to C2 and C3, but indirectly, via a series of intermediates; no enzyme is mentioned. We therefore curate both (C1, C2) and (C1, C3) as "Indirect_conversion", with no "Converter".
Non_conversion	*"In the amination direction, they catalyze the conversion of* **mesaconate** $_{C1}$ *to yield only* **(2S,3S)-3-methylaspartic acid** $_{C2}$, *with no detectable formation of* **(2S,3 R)-3-methylaspartic acid** $_{C3}$." – PMID:19670200	C1 can be converted to C2, but not C3. We therefore curate (C1, C2) as a "Conversion", and (C1, C3) as a "Non_conversion".

Table 2. Examples of curated relations in EnzChemRED. Chemical mentions and protein mentions are denoted by the numbered subscripts "c" and "p" respectively.

- **Protein**: a mention of a protein, or family of proteins, normalized to UniProtKB accession numbers (UniProtKB ACs). As in BioRED, we included gene names in our annotation, which we also normalized to UniProtKB ACs.
- **Domain**: a mention of a protein domain, normalized to the UniProtKB ACs of the protein in which the domain occurs.
- **MutantEnzyme:** a mention of a mutant protein, normalized to the UniProtKB accession number of the wild-type protein.
- **Coreference:** not a mention *per se*, but rather a reference to a protein or chemical mention that appears elsewhere in the abstract. In the following example, "It" is a coreference to a specific protein mention found in the preceding sentence: "*ABC1 is a hydrolase. It catalyses the hydrolysis of phospholipids*.". Coreferences were normalized to the chemical or protein identifier for the mention being referenced.

We curated all Chemical and Protein mentions found in abstracts, irrespective of whether those mentions were part of descriptions of enzymatic reactions or not. We did not systematically curate Domain, MutantEnzyme and Co-reference mentions, but focused on those that participate in enzymatic reactions. For this reason, we did not consider these three types of mentions in our evaluations of NER, NEN, and RE performance, but these could serve as valuable annotations for the development of a more extensive dataset in the future.

Curation of chemical conversions. We based our schema for the curation of relations relevant for enzyme functions in EnzChemRED on that developed in BioRED, but with three major alterations.

First, we defined two additional relation types in EnzChemRED. BioRED captures chemical reactions by using the relation "Conversion" to link pairs of reactants. In EnzChemRED, we also added "Indirect_conversion" and "Non_conversion", giving three possible relations (Table 2) that are defined as follows:

- **Conversion**: links two chemicals that, according to the text, may participate on opposite sides of a reaction equation – such as one substrate and one product.
- **Indirect_conversion**: links two chemicals that, according to the text, can interconvert, but not directly – such as conversions involving the first and last chemical in a pathway. While these kinds of relations will not give rise to enzyme function annotations, they constitute a significant but minor fraction of relations in the EnzChemRED dataset (see Section Dataset statistics and inter-annotator agreement).
- **Non_conversion**: links two chemicals that, according to the text, were experimentally tested but did not interconvert at all (at least under the experimental conditions used). These are the rarest type of relation in EnzChemRED.

Second, while BioRED features only binary pairs, in EnzChemRED we also introduced ternary tuples, which allow us to link mentions of enzymes to the "Conversions" they catalyze. We assign each enzyme the role of "Converter".

Third, we modified the granularity of annotations in BioRED. While BioRED provides document-level relation pairs, EnzChemRED provides relations annotated at the level of individual mentions and sentences.

Curation workflow. Fig. 3 outlines the curation workflow for EnzChemRED; we describe the main steps below.

Pre-Annotation of chemicals and proteins: We used PubTator[48,49] to pre-tag chemical and gene/protein mentions in the 1,210 abstracts of EnzChemRED prior to their curation. PubTator assigns MeSH IDs for chemical mentions and Entrez IDs for gene and protein mentions, which we converted to ChEBI IDs and UniProtKB ACs using MeSH-to-ChEBI and Entrez-to-UniProtKB mapping tables. ChEBI provides multiple distinct identifiers for different protonation states of a given chemical compound, so we mapped all ChEBI IDs to those of the major protonation state at pH 7.3 – the form used in UniProtKB and Rhea – using a mapping file created for this purpose by Rhea (the file "chebi_pH7_3_mapping.tsv", which is available at https://www.rhea-db.org/help/download).

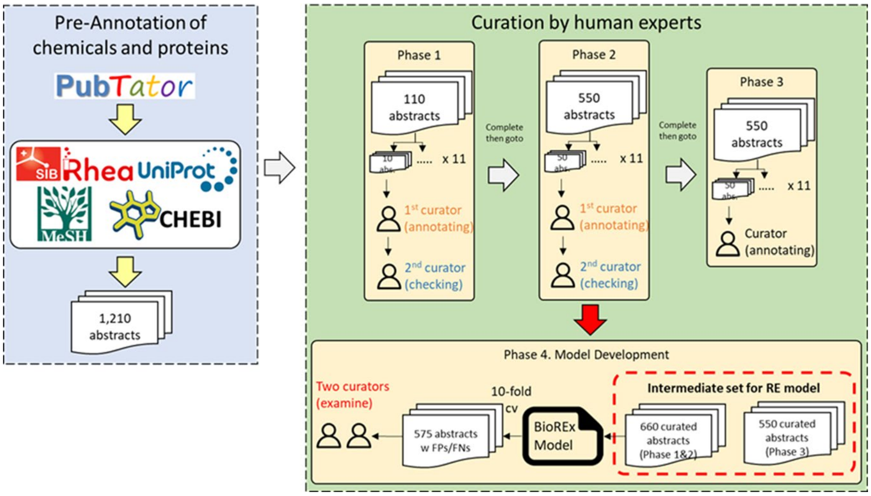
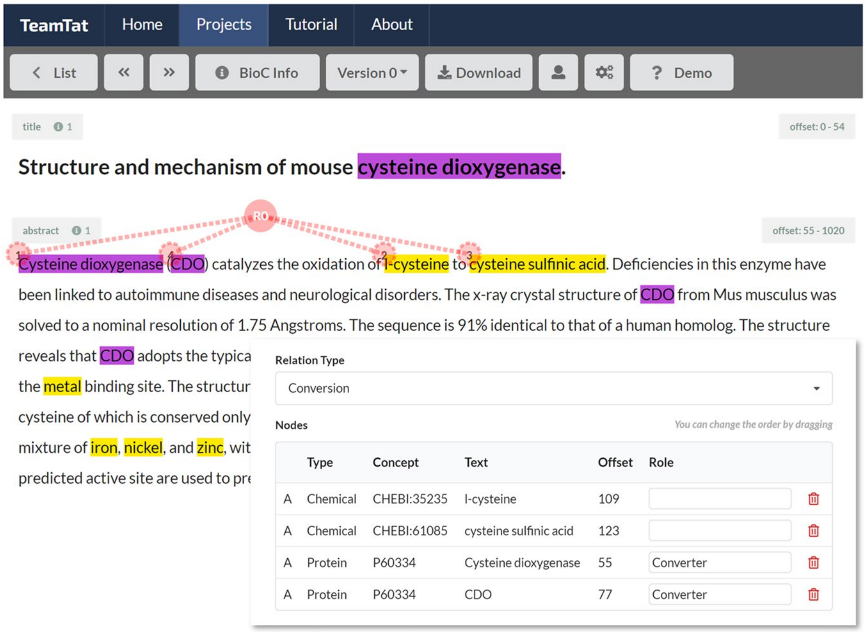


Fig. 3 Curation workflow for the EnzChemRED corpus. A total of 1,210 abstracts were curated by 11 experts in three phases. These 1,210 abstracts were then used to train an interim BioREx model, which was then run on all EnzChemRED abstracts; abstracts with putative FP and FN predictions by the model were then analyzed again in phase 4, and, where necessary, re-curated.



Structure and mechanism of mouse cysteine dioxygenase.

Cysteine dioxygenase (CDO) catalyzes the oxidation of l-cysteine to cysteine sulfinic acid. Deficiencies in this enzyme have been linked to autoimmune diseases and neurological disorders. The x-ray crystal structure of CDO from Mus musculus was solved to a nominal resolution of 1.75 Angstroms. The sequence is 91% identical to that of a human homolog. The structure

reveals that CDO adopts the typica

the metal binding site. The structur

cysteine of which is conserved only

mixture of iron, nickel, and zinc, wit

predicted active site are used to pre

Relation Type

Conversion

Nodes

You can change the order by dragging

	Type	Concept	Text	Offset	Role
A	Chemical	CHEBI:35235	l-cysteine	109	
A	Chemical	CHEBI:61085	cysteine sulfinic acid	123	
A	Protein	P60334	Cysteine dioxygenase	55	Converter
A	Protein	P60334	CDO	77	Converter

Fig. 4 Annotation of EnzChemRED abstracts using TeamTat. In the abstract shown (from ref. [76]) the mentions of "Cysteine dioxygenase" and "CDO" refer to the enzyme (UniProt:P60334) that is responsible for the conversion of L-cysteine (CHEBI:35235) and cysteine sulfinic acid (CHEBI:61085). The inset shows details of the curated relation within the TeamTat tool, including the type of relation ("Conversion"), the text spans that define the participants in that relation, and their offsets, the unique identifiers from UniProtKB and ChEBI that were used to tag those text spans, and the assignment of the role "Converter" to the text spans of the protein mentions.

Curation by human experts: Curation was performed using TeamTat (Fig. 4) by a team of 11 professional curators with expertise in biochemistry and the curation of UniProtKB/Swiss-Prot and Rhea. Curators were required to review all PubTator tagging results for gene/protein and chemical mentions (both text spans and IDs) and correct them and add missing protein and chemical annotations and identifiers as necessary. Curators were allowed to use external information sources, including the full text of the article, as well as knowledge resources such as UniProtKB, Rhea, ChEBI, MeSH, and PubChem[50], when curating chemical and protein mentions. Following curation of all protein and chemical mentions, curators were then required to link chemical mentions that participate in relations of the type "Conversion", "Indirect_conversion", or "Non_conversion",

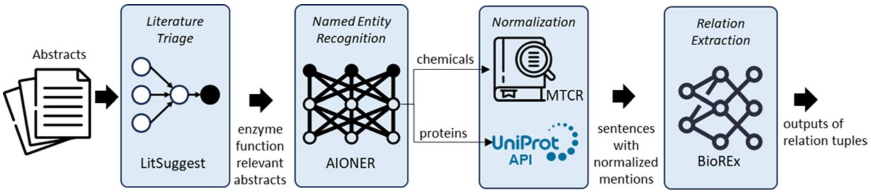


Fig. 5 Overview of the end-to-end pipeline.

thereby creating binary (chemical-chemical) pairs, as well as mentions of enzymes that catalyze those conversions ("Converter") if applicable, creating ternary [enzyme-(chemical-chemical)] tuples. Curators were prohibited from using external information, such as the full text of the publication or prior knowledge of the chemistry or enzymes involved in the reactions, when annotating relations of any type, or linking converters to conversions. Put another way, all the evidence needed to create a conversion, and to link a converter to it, had to be contained in the abstract itself, either within one sentence, or across multiple sentences.

We divided the curation process into four phases.

1. **Phase 1**. We provided each curator with 10 abstracts for familiarization with the curation workflow and guidelines. Following curation, the abstracts were frozen, and a copy was made, which was reviewed and corrected by a second curator. The curation team then met to discuss curation issues, revise guidelines, and finalize the set of abstracts from phase 1. The output from phase 1 consisted of 110 curated abstracts, each reviewed and where necessary revised, and a set of revised guidelines.
2. **Phase 2**. We provided each curator with 50 additional abstracts for curation. Curated abstracts were frozen, and a copy was made, which was reviewed and corrected by a second curator. The curation team then met to discuss curation issues, revise guidelines, and finalize the set of abstracts from phase 2. The output from phase 2 consisted of a further 550 curated abstracts, each reviewed and where necessary revised, and a set of revised guidelines.
3. **Phase 3**. We provided each curator with 50 additional abstracts for curation. These were not reviewed after curation. The output from phase 3 consisted of a further 550 curated abstracts that had not been reviewed.
4. **Phase 4**. We performed a round of "model guided" re-curation of abstracts, using a "preliminary" BioREx model (see Section Relation extraction) to identify potential curation errors, such as missed chemical conversions. We trained this model using the set of 1,210 abstracts curated in phases 1–3 and used it to perform RE on the entire dataset of 1,210 abstracts. We identified potential false positive (FP) or false negative (FN) predictions in 575 of the 1,210 abstracts. Each potential FP or FN prediction identified by the preliminary BioREx model was then examined by two curators, who were free to compare and discuss their interpretations of the models' predictions. In some cases, the potential FP and FN errors from the model were deemed to be correct and were re-curated as TP or TN as appropriate, and the curation guidelines were updated if needed. We used this Phase 4 EnzChemRED dataset to train our final models for NER (see Section Named entity recognition) and RE (see Section Relation extraction).

Overview of the end-to-end pipeline.

Fig. 5 shows the four main steps of our end-to-end NLP pipeline for enzyme function extraction, which are:

1. Literature triage, to identify relevant papers about enzyme functions.
2. Named entity recognition (NER), to tag chemical and protein mentions.
3. Named entity normalization (NEN), to link chemical and protein mentions to stable unique database identifiers.
4. Relation extraction (RE), to extract information about chemical conversions and the enzymes that catalyze them.

The following sections describe the methods used in each of the steps, the combination of methods to create the end-to-end pipeline, and methods to process and visualize the output from it.

Literature triage.

The goal of the literature triage step is to identify relevant abstracts, and reduce the number of irrelevant abstracts for processing during the subsequent steps of NER, NEN, and RE. For literature triage we used LitSuggest (https://www.ncbi.nlm.nih.gov/research/litsuggest/)[46], a web-based machine-learning framework for literature recommendations. LitSuggest frames literature recommendation as a document classification task and addresses that task using a stacking ensemble learning method. LitSuggest uses a variety of fields from each publication, including the journal name, publication type, title, abstract, registry numbers (for substance identifiers and names), and user-submitted keywords. Text of the fields is concatenated and converted into a bag-of-words representation, which serves as inputs/features for a diverse array of classifiers for text-mining available through the scikit-learn library. The outputs from the individual classifiers are fed into a logistic regression model, which ensemble and obtain the optimal classification. In addition to literature triage, we also used LitSuggest to confirm the relevance of abstracts selected for curation in EnzChemRED (see Section Selection of abstracts for curation).

Positive training examples for LitSuggest consisted of abstracts from papers that provided experimental evidence used to annotate enzymes in UniProtKB/Swiss-Prot with Rhea reactions (dataset created November 7th, 2020). As with the EnzChemRED dataset, we defined papers that provided experimental evidence as those linked to an evidence tag with an experimental evidence code from the Evidence and Conclusions Ontology (ECO), such as "ECO:0000269", and excluded abstracts of papers linked to Rhea reactions that involve proteins and other macromolecules, such as DNA. This exclusion criterion strongly reduced the propensity of LitSuggest models to retrieve papers about signalling pathways, where proteins are modified by enzymes.

We trained and tested LitSuggest models using a set of 9,055 positive abstracts split into 5 sets of 7,244 positive abstracts for training and 1,811 positive abstracts for testing, with 14,488 negative abstracts for training selected at random from PubMed using the LitSuggest curation interface. LitSuggest models provide a score of 0-1 for each abstract, with scores above 0.5 denoting that the abstract is relevant (belongs to the same class as the positive training data). The five LitSuggest models had a mean recall of 98% when tasked with classifying the 1,811 abstracts left out. To test precision and recall using more realistic ratios of relevant and irrelevant literature, we also performed "spike-in" tests that mixed 250 relevant papers (from the set of 1,811 abstracts left out) with a set of 100,000 abstracts selected from PubMed using NCBI e-utils (so a ratio of 400:1 irrelevant to relevant abstracts). At a score threshold of 0.8, our best performing LitSuggest model had precision of 90.1%, sensitivity of 94.8%, and F_1 score of 92.4% in these "spike-in" tests. Swiss-Prot curators now use this LitSuggest model as a tool to triage literature for curation of protein sequences in UniProtKB/Swiss-Prot with reactions from Rhea on a weekly basis. It is available at https://www.ncbi.nlm.nih.gov/research/litsuggest/project/5fa57e75bf71b3730469a83b. We also used this model in the first step of our end-to-end pipeline for enzyme function extraction from PubMed abstracts.

Named entity recognition (NER). The goal of Named entity recognition (NER) is to identify chemical and protein mentions in text. The NER task is framed as one of sequence labelling; text is represented as a sequence of tokens $x = (x_1, x_2, \ldots, x_n)$, where n denotes the length of the text, and the goal is to classify the sequence of tokens x into a corresponding label sequence $y = (y_1, y_2, \ldots, y_n)$, where $y_i \in Y$. Y is a label set for the model, and each label represents the entity type and position in the text. A common scheme for label sets in biomedical NER is the IOB2 tagging scheme, where IOB stands for "inside, outside, beginning". In IOB2 the label set Y consists of the "O" label, which denotes a text chunk that is "outside" the tagging schema (so not a chemical or protein entity in the case of EnzChemRED), and entity type labels, with the prefixes "B-", denoting the beginning of the entity token, and "I-", denoting the inside of the entity token. So, under the IOB2 schema, a mention of a "membrane fatty acid" would be labelled as "membrane [O] fatty [B-Chemical] acid [B-Chemical]".

For the EnzChemRED task we used AIONER[51] (https://github.com/ncbi/AIONER), an all-in-one scheme-based biomedical NER tool that integrates multiple datasets into a single task by adding task-oriented tagging labels, allowing the model to learn the synonyms present in texts covering multiple subjects. AIONER performs optimally on 14 BioNER benchmark datasets, such as BioRED, BC5CDR[52], GNormPlus/GNorm2[53], NLM-Gene[54], and NLM-Chem[55]. It has been shown to achieve competitive performance with Multi-Task Learning (MTL) methods for multiple NER tasks, while being more stable and requiring fewer model modifications. AIONER replaces the "O" (outside) label with a specific form of O-label for each NER dataset, such as "O-Gene" for the task of gene finding in the case of GNormPlus/GNorm2 and NLM-Gene. For EnzChemRED we use "O-Reaction" to signify the NER task, giving five labels for our dataset, namely "O-Reaction", "B-Gene", "I-Gene", "B-Chemical", and "I-Chemical". So, under the IOB2 schema used by AIONER to integrate EnzChemRED, our mention of a "membrane fatty acid" would be labelled as "membrane [O-Reaction] fatty [B-Chemical] acid [I-Chemical]".

We illustrate the training process for the AIONER model using EnzChemRED in Fig. 6. We employed spaCy (https://spacy.io/) for sentence detection and tokenization, allowing us to convert our entire biochemical reaction text dataset for AIONER. Once the dataset was converted to the AIONER representations of input and label sequence, we optimized our model using the fine-tuning script provided by AIONER on GitHub. A similar procedure was followed during the testing phase without inputting the label sequence. We tested four different pre-trained language models (PLMs) for NER, namely Bioformer[56], PubMedBERT[28], AIONER-Bioformer[51], and AIONER-PubMedBERT[51]. We performed 10-fold cross validation for each model using EnzChemRED, fine-tuning the PLMs using the training set partition and evaluating them on the test set partition.

Named entity normalization (NEN). Named entity normalization (NEN) takes the entities identified during NER a step further. It aims to determine the exact meaning of each mention in context by mapping it to a unique identifier from a knowledgebase or ontology such as UniProtKB (for proteins) or ChEBI (for chemical entities). This process helps clarify and standardize the entities detected in text and is an essential step in transforming natural language into a structured knowledgebase.

NEN can be formulated as follows: given a named entity e in context and a lexicon L (essentially a list of IDs and their corresponding synonyms, where an ID can have multiple synonyms), the goal is to find the unique ID in L for e. For EnzChemRED we used the Multiple Terminology Candidate Resolution (MTCR) pipeline to map chemical mentions in abstracts to ChEBI and MeSH IDs. MTCR is a structured approach for linking entities in the biomedical domain, including chemical terminologies; a similar process, referred to as sieve-based entity linking, has been described by D'Souza and Ng[57]. There are three main steps in the MTCR pipeline: abbreviation resolution, candidate lookup, and post-processing.

1. During the abbreviation resolution phase, the pipeline identifies pairs of short and long forms in each document using the Ab3P Abbreviation Resolution tool[58]. Short forms are then expanded into long forms before looking up (so "TPP" to "triphenyl phosphate").
2. The candidate lookup step involves finding the candidate IDs of e in lexicon L. The process starts with a precise lookup and proceeds to higher recall queries, stopping once a match is found. The steps are as

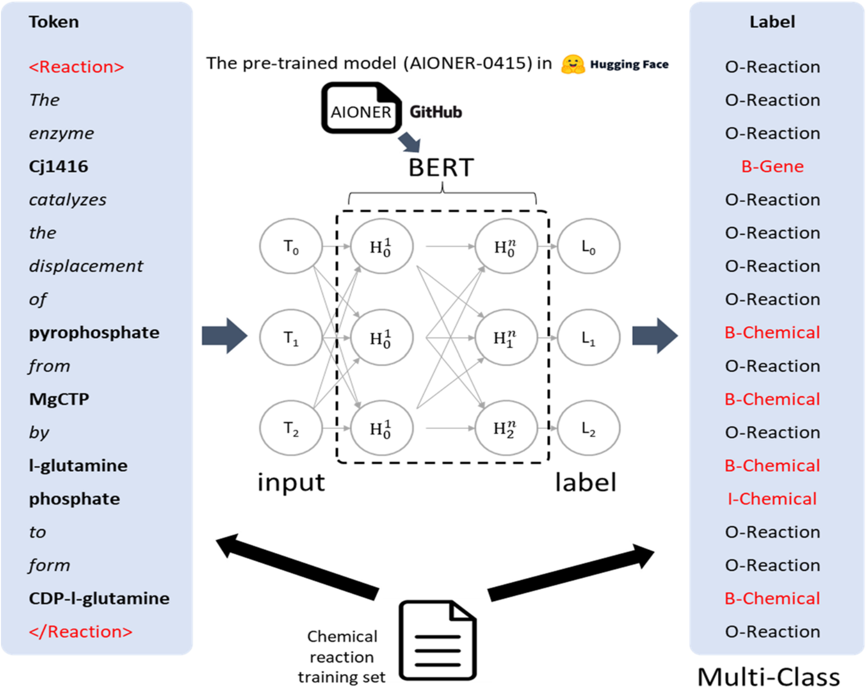


Fig. 6 Overview of the fine-tuning process for AIONER on EnzChemRED.

follows: (1) search for *e* in lexicon *L*; (2) lower the case of e and strip non-alphanumeric characters from it, creating e_p, then search for e_p in lexicon *L*; (3) stem e_p, and search for stemmed forms of e_p in lexicon *L*; (4) search for e_p in L_{all} and map to *L*, where "L_{all}" refers to all lexicons; (5) Search for stemmed *e* in L_{all} and map to *L*. Mappings can take place in two ways: 'single', where terms are mapped directly to the target using cross references in the two lexicons, and 'pivot', where terms are mapped directly to the target through identifiers shared with another lexicon, such as the International Chemical Identifier (InChI) (or it's hashed form, the InChIKey) (https://www.inchi-trust.org)[59], the SMILES (Simplified Molecular-Input Line-Entry System) (http://opensmiles.org), a linear notation for chemical structures, the Chemical Abstracts Service (CAS) number, or others.

3. Post-processing removes annotations for mentions identified as non-chemical, which is particularly relevant when the target terminology is broad, such as MeSH. Ambiguous mentions (those with two or more potential target identifiers) are resolved using unambiguous mentions.

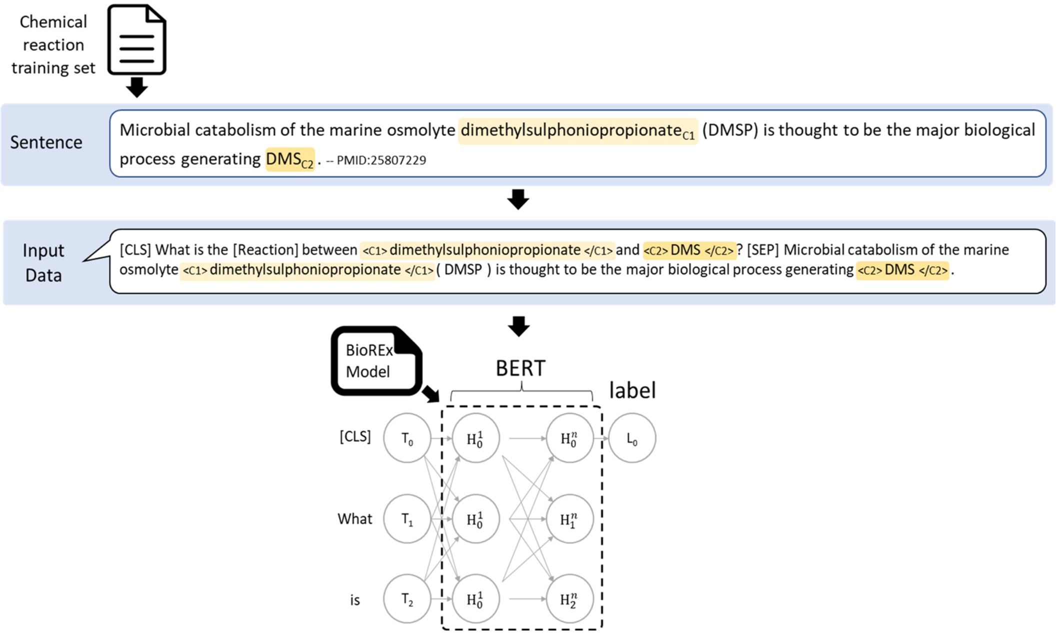
The MTCR pipeline has been benchmarked for the BioCreative VII NLM Chem task[60] with MeSH as the target terminology. BlueBERT[27] showed higher NER performance, but MTCR demonstrated outstanding NEN performance, with a precision of 81.5% and a recall of 76.4%, with only 29% of teams outperforming MTCR in NEN.

Relation extraction (RE). We frame the problem of extracting pairs of chemical reactants as one of relation classification. There are two tasks, binary pair classification, for (chemical-chemical) relations that link reaction participants, and ternary tuple classification, for (protein-(chemical-chemical)) relations, that link chemical reactants, and the enzymes that catalyze their conversion.

- For the task of binary pair classification for (chemical-chemical) relations, given a chemical mention pair (c_1, c_2) and the corresponding sentence *s*, the objective is to classify the relation type *r* of the chemical pair (c_1, c_2).
- For the task of ternary tuple classification for (protein-(chemical-chemical)) relations, given a protein mention *p*, a chemical mention pair (c_1, c_2), and the corresponding sentence *s*, the objective is to classify the relation type *r* of the ternary tuple (*p* (c_1, c_2)).

Valid relation types for binary pairs and ternary tuples are "Conversion", "Indirect_conversion", and "Non_conversion", which are curated (see Section Curation of chemical conversions), and "None", which is assigned automatically during evaluation. For binary pairs, "None" is assigned to all pairs of chemical mentions that are not curated using one of the three valid relation types. For ternary tuples, "None" is assigned to all ternary tuples that include pairs of chemical mentions not curated using one of the three valid relation types, and to all ternary tuples that include pairs of chemicals curated with a valid relation but that also include a protein mention that was not linked to them by a curator (i.e. it is not the enzyme responsible).

We performed relation classification using PubMedBERT and BioREx[61], which is a PubMedBERT model trained on the BioRED dataset and eight other common biomedical RE benchmark datasets (PubMedBERT is essentially the same model but without this additional training step). BioREx offers a reliable and effective



Microbial catabolism of the marine osmolyte dimethylsulphoniopropionate$_{C1}$ (DMSP) is thought to be the major biological process generating DMS$_{C2}$. -- PMID:25807229

[CLS] What is the [Reaction] between <C1> dimethylsulphoniopropionate </C1> and <C2> DMS </C2> ? [SEP] Microbial catabolism of the marine osmolyte <C1> dimethylsulphoniopropionate </C1> (DMSP) is thought to be the major biological process generating <C2> DMS </C2> .

Fig. 7 An illustration of the specific input representation for the EnzChemRED dataset and the fine-tuning process of the BioREx model.

approach to chemical reaction extraction and has shown consistently high performance for relation classification across seven different entity pairs. In PubMedBERT and BioREx, each input sequence $x = (x_1, x_2, \ldots, x_n)$ is prefixed with a special [CLS] token x_{CLS}. This token is processed through the neural network layers of the model along with the sequence, and the output corresponding to the x_{CLS} token is a high-dimensional vector that aggregates, or summarizes, the contextual information from the entire sequence. We denote the output embedding of the [CLS] token as h_{CLS}.

To adapt the [CLS] vector h_{CLS} for relation classification, it is passed through a linear neural network layer, which aims to map the high-dimensional h_{CLS} vector into a lower-dimensional vector space suitable for the classification task. In our case, this is a four-dimensional vector, with each dimension corresponding to one relation type label - Conversion, Indirect_conversion, Non_conversion, or None. Mathematically, we can express the operation of the linear layer as:

$$r = W \times h_{CLS} + b$$

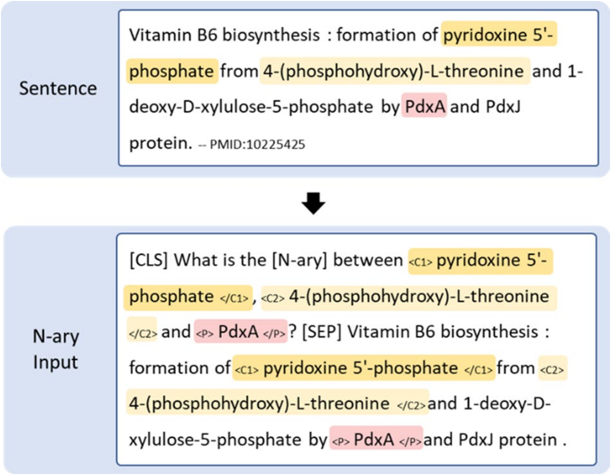
where W is the weight matrix, b is the bias vector, and r is the output vector. The length of r matches the number of classes in our task, which is four.

Fig. 7 illustrates the process of fine-tuning on EnzChemRED using BioREx. EnzChemRED is annotated at both mention and sentence levels, with locations specified, unlike BioRED, which uses document-level annotation, and where relations are given in the format of ID pairs without specifying the exact locations of the entity mentions involved. We therefore adjust the fine-tuning procedure used in BioRED, replacing the [Corpus] tag with a [Reaction] tag, and using individual sentences as input rather than full documents.

Fig. 8 illustrates an example input representation of a ternary tuple. The classification of ternary tuples follows similar rules to that for binary pairs. We insert additional boundary tags, "<P>" and "</P>", to denote the enzyme in the input instance, but otherwise follow the same procedure as for binary pair RE. As with binary pairs, ternary tuples are annotated at both sentence and mention levels, such that if the same enzyme appears more than once in a sentence, each occurrence would be treated as a different instance.

End-to-end pipeline.

We combined the best performing methods for NER (AIONER-PubMedBERT, fine-tuned using EnzChemRED) and RE (BioREx, fine-tuned using EnzChemRED) with MTCR for chemical NEN to create a prototype end-to-end pipeline for enzyme function extraction from text. We applied this pipeline to EnzChemRED abstracts for cross validation purposes, and to relevant PubMed abstracts (up to December 2023) identified using the LitSuggest model described in the Section **Literature triage** to map enzyme functions in literature. The latter necessitated comparison of chemical pairs extracted from PubMed abstracts to pairs of chemical reactants from Rhea, which was accomplished as follows. To create the set of Rhea pairs for comparison, we extracted pairs of chemical reactants from Rhea using a heuristic procedure in which we removed the top 100 most frequently occurring compounds in Rhea reactions such as water, oxygen, and protons, and then enumerated all possible pairs of the remaining compounds within each Rhea reaction. We also removed pairs of identical ChEBI IDs from the Rhea set, which in Rhea can occur as part of transport reactions. To prepare the chemical pairs extracted from PubMed abstracts for comparison to Rhea, we first normalized their ChEBI IDs to those representing the major microspecies at pH7.3, the form used in Rhea, removed pairs that include any of the top



Vitamin B6 biosynthesis : formation of pyridoxine 5'-phosphate from 4-(phosphohydroxy)-L-threonine and 1-deoxy-D-xylulose-5-phosphate by PdxA and PdxJ protein. -- PMID:10225425

[CLS] What is the [N-ary] between <C1> pyridoxine 5'-phosphate </C1>, <C2> 4-(phosphohydroxy)-L-threonine </C2> and <P> PdxA </P>? [SEP] Vitamin B6 biosynthesis : formation of <C1> pyridoxine 5'-phosphate </C1> from <C2> 4-(phosphohydroxy)-L-threonine </C2> and 1-deoxy-D-xylulose-5-phosphate by <P> PdxA </P> and PdxJ protein .

Fig. 8 An example of the input representation for a ternary tuple in EnzChemRED.

100 most frequently occurring compounds in Rhea reactions, and removed pairs where both members had the same ChEBI ID. This can occur due to errors in NER and NEN, where erroneous text spans can cause distinct but related chemical names to be mapped to the same identifier. After processing we compared the degree of overlap in the two sets (chemical pairs from Rhea reactions and PubMed abstracts) using their ChEBI IDs.

Visualization of chemical pairs from PubMed abstracts and Rhea. To visualize chemical pairs from PubMed abstracts and Rhea we used the Tree Map (TMAP) algorithm to create TMAP trees using code from http://tmap.gdb.tools as described[62], clustering chemical pairs in TMAP trees according to their Differential Reaction Fingerprint (DRFP), calculated according to the method of Probst[63]. We used the degree of atom conservation between the members of each chemical pair to filter the output of our end-to-end NLP pipeline. To calculate atom conservation, we first converted molecular structures into graphs by replacing all bond types with single bonds. This ensures a standardized representation of molecular structures, simplifying subsequent analyses. We then computed the Maximum Common Substructure (MCS) using the rdkit.Chem.rdFMCS.FindMCS function (from the open-source cheminformatics toolkit RDKit, at www.rdkit.org) with a permissive ring fusion parameter. The MCS represents the largest common atomic framework shared by the two molecules (after conversion into a graph of atoms linked by single bonds). The atom conservation is the average of the percentage of common atoms, as given by:

$$\%\ \text{atom conservation} = 1/2 \times (n_{MCS}/n_L + n_{MCS}/n_R) \times 100$$

where n_{MCS} is the number of atoms in the maximum common substructure, n_L is the number of atoms in the molecule on the left side of the pair, and n_R is the number of atoms in the molecule on the right side of the pair. This metric provides a standardized measure of structural similarity, facilitating the comparison of chemical compounds in each pair.

Data Record

The EnzChemRED dataset is available for download in BioC format[64] from the Rhea ftp site at https://ftp.expasy.org/databases/rhea/nlp/ and at https://zenodo.org/records/11067998[65]. The EnzChemRED.tar.gz contains 1,210 PubMed abstracts with annotations of gene/protein and chemical mentions using UniProtKB and ChEBI respectively, chemical conversions - relations that link pairs of chemical mentions - and the enzymes that catalyze those conversions, when available. The BioREx_EnzChemRED_PubMed.tsv file provides chemical conversion pairs (binary pairs) identified in PubMed using our prototype end-to-end pipeline, including the PMID and sentence from which they derive, and their BioREx score. The BioREx_EnzChemRED_PubMed_normalized.tsv file provides unique chemical conversion (binary pairs) identified in PubMed using our prototype end-to-end pipeline, with key characteristics such as PubMed count, maximum BioREx score, atom conservation, and the minimum DRFP distance to a ChEBI pair in Rhea.

Technical Validation

Dataset statistics and inter-annotator agreement (IAA). Table 3 provides an overview of the EnzChemRED dataset, highlighting key statistics including counts of documents, entity mentions, and binary and ternary relations. The inter-annotator agreement (IAA) of our dataset stands at 92.82% for the curation of entity mentions - where we define agreement as the selection of matching text spans and matching identifiers (covering NER+NEN) - and at 87.03% for the curation of binary pair relations.

Evaluation of NER methods using EnzChemRED. We evaluated the effects of fine-tuning using EnzChemRED on NER using four pre-trained language models: Bioformer, PubMedBERT, AIONER-Bioformer, and AIONER-PubMedBERT. Bioformer and PubMedBERT are the base models used to test the effect of fine-tuning using EnzChemRED alone; AIONER-Bioformer and AIONER-PubMedBERT

Type		Entity mentions, pairs, tuples	Unique IDs, pairs, or tuples
Entity	All	31,915	5,724
	Chemical	18,887	3,155
	Protein	13,028	2,569
Binary pair	All	4,817	2,679
	Conversion	4,386	2,434
	Indirect_conversion	411	292
	Non_conversion	20	17
Ternary tuple	All	4,195	2,120
	Conversion	3,966	1,997
	Indirect_conversion	229	133
	Non_conversion	0	0

Table 3. EnzChemRED summary statistics.

	Chemical			Protein			Overall		
PLM	P	R	F	P	R	F	P	R	F
AIONER-Bioformer	86.13	85.80	85.97	76.37	78.79	77.56	82.08	82.95	82.51
AIONER-PubMedBERT	87.10	85.52	86.30	79.53	74.99	77.19	84.10	81.23	82.64
Bioformer + EnzChemRED	85.83	86.87	86.35	83.13	84.75	83.93	84.73	86.01	85.36
PubMedBERT + EnzChemRED	86.92	87.21	87.07	83.19	85.07	84.11	85.38	86.33	85.86
AIONER-Bioformer + EnzChemRED	86.20	86.95	86.58	81.90	85.35	83.59	84.41	86.30	85.35
AIONER-PubMedBERT + EnzChemRED	87.16	87.37	**87.26**	83.80	86.09	**84.93**	85.77	86.85	**86.30**

Table 4. NER performance of pre-trained language models on EnzChemRED. P, precision; R, recall; F, F_1 score.

were pretrained on a wide range of chemical and gene NER datasets. For evaluation, we used the F_1 score and considered the PMID, entity type (chemical or protein), and the start and end characters of named entities. An entity detected by the NER method is considered a true positive only if both the entity type and the start and end character positions match.

Table 4 shows NER performance for each of the four models. All models performed well in chemical NER, with similar performance for models that were either trained with AIONER, or fine-tuned using EnzChemRED, suggesting that the chemical datasets included in AIONER and the annotations in EnzChemRED provide similar benefits to the models for NER. All models performed less well in protein NER, which may be due to the greater variation in text spans for genes and proteins, but fine-tuning with EnzChemRED significantly improved the performance of all four models for proteins too. The best performing model overall is AIONER-PubMedBERT fine-tuned using EnzChemRED, with F_1 scores of 87.26% for chemical NER and 84.93% for gene/protein NER, which is comparable to SOAT performance for chemical NER (F_1 score 84.79%)[66] on the NLM-Chem dataset[60] and for gene/protein NER (F_1 score 86.70%)[48] on the GNormPlus dataset[67].

Evaluation of RE methods using EnzChemRED.

We evaluated the effects of fine-tuning using EnzChemRED on RE using two pre-trained language models: PubMedBERT and BioREx. To evaluate binary pair RE using EnzChemRED we consider the PMID, the start and end character positions of the chemical pair within the sentence, their ChEBI or MeSH IDs, and the relation type. For ternary tuple RE using EnzChemRED we also consider the start and end character positions of the protein mentions within the sentence and the UniProtKB AC. For both binary pair and ternary tuple evaluation, we used two types of classification: binary and multi-class classification. Binary classification considers "Conversion" and "Indirect_conversion" as equivalent, while multi-class classification considers them as distinct. For training purposes, a "None" relation type is utilized, which is assigned to all chemical pairs that are not curated, and which are presumed true negatives. Chemical and protein mentions that lack identifiers are also considered in the evaluations, with their IDs being treated as empty strings.

Table 5 shows RE performance (both binary pairs and ternary tuples). Both baseline models PubMedBERT and BioREx were poor predictors of binary and multi-class relations for binary and ternary tuples, but the performance of both models increased significantly after fine-tuning with EnzChemRED, with BioREx achieving an F_1 score of 86.66% for binary relation classification for chemical pairs, and consistently outperforming PubMedBERT for all RE tasks. We therefore chose to employ BioREx in our end-to-end pipeline. Performance of both models decreased slightly as the complexity of the classification task increased, with reduced F_1 score for multi-class relation classification, which requires identification of conversions that require multiple steps in pathways, and for ternary tuples relative to binary pairs, which requires that the enzyme be correctly identified.

Analysis of BioREx RE errors in EnzChemRED.

While BioREx performs well in binary pair and ternary tuple relation classification on EnzChemRED, analysis of error cases may identify areas for further improvement.

		Evaluation					
		Binary pair			Ternary tuple		
Model	Classification	P	R	F	P	R	F
PubMedBERT	Binary class	21.83	79.16	34.23	12.92	94.21	22.72
	Multi-class	19.64	71.21	30.79	12.19	88.89	21.44
BioREx	Binary class	20.59	98.34	34.05	13.29	99.74	23.46
	Multi-class	19.33	89.31	31.78	12.57	94.28	22.18
PubMedBERT + EnzChemRED	Binary class	83.18	87.79	85.43	80.16	83.03	81.57
	Multi-class	75.73	79.93	77.77	76.55	79.28	77.89
BioREx + EnzChemRED	Binary class	85.20	88.17	**86.66**	83.13	84.46	**83.79**
	Multi-class	77.49	80.20	**78.82**	79.40	80.67	**80.23**

Table 5. RE performance of pre-trained language models on EnzChemRED. P, precision; R, recall; F, F_1 score.

We randomly selected and analyzed a set of around 50 false positive (FP) cases (binary and ternary), where BioREx predicts a relation that was not curated in EnzChemRED, and 50 false negative (FN) cases (binary and ternary), where BioREx failed to predict a relation that was curated in EnzChemRED. Broadly speaking there appear to be two main categories of FP predictions and two main categories of FN predictions, with examples provided below. In each example chemical mentions and protein mentions are denoted by the numbered subscripts "*c*" and "*p*" respectively.

FP predictions for binary pairs. The first common category of FP predictions (~20% of all FPs) occurred in sentences with at least two chemical pairs, with all substrates listed first in order, followed by all products in order, and the two lists linked using the term "respectively", as in this example:

"*Wild-type strain L108 and* ***mdpJ*** $_{P1}$ *knockout mutants formed* ***isoamylene*** $_{C1}$ *and* ***isoprene*** $_{C2}$ *from* ***TAA*** $_{C3}$ *and* ***2-methyl-3-buten-2-ol*** $_{C4}$*, respectively.*" (PMID: 22194447).

This sentence provides evidence for two curated instances of "Conversion", isoamylene (C1) from TAA (C3), and isoprene (C2) from 2-methyl-3-buten-2-ol (C4). Our BioREx model correctly identified both, but also predicted a third conversion, isoamylene (C1) from 2-methyl-3-buten-2-ol (C4), which constitutes an FP prediction.

The second common category of FP predictions (~23%) occurred in sentences with multiple chemical mentions, and where the chemical mentions involved in the FP predictions also participated in other curated TP "Conversion", as in this example:

"*Most methanogenic Archaea contain an unusual cytoplasmic fumarate reductase which catalyzes the reduction of* ***fumarate*** $_{C1}$ *with* ***coenzyme M*** $_{C2}$ *(**CoM-S-H*** $_{C3}$*) and* ***coenzyme B*** $_{C4}$ *(**CoB-S-H*** $_{C5}$*) as electron donors forming* ***succinate*** $_{C6}$ *and* ***CoM-S-S-CoB*** $_{C7}$ *as products.*" (PMID: 9578488).

This sentence provides evidence for the conversion of fumarate (C1), coenzyme M (C2) (CoM-S-H (C3)) and coenzyme B (C4) (CoB-S-H (C5)) to succinate (C6) and CoM-S-S-CoB (C7). Our BioREx model identified all these relations but also erroneously identified one "Conversion" involving fumarate (C1) and CoM-S-H (C3). These are in fact a pair of substrates and do not convert one to the other.

The remainder of FP cases did not fall into any single clearly definable category.

FN predictions for binary pairs. The most common category of FN cases (~43%) represent instances of missed "Indirect_conversion", as in this example:

"*We provide NMR and crystallographic evidence that the PucG* $_{P1}$ *protein from Bacillus subtilis catalyzes the transamination between an unstable intermediate ((S)-ureidoglycine* $_{C1}$*) and the end product of* ***purine*** $_{C2}$ *catabolism (**glyoxylate*** $_{C3}$*) to yield oxalurate* $_{C4}$ *and glycine* $_{C5}$*.*" (PMID: 20852637).

This sentence provides evidence for a curated "Indirect_conversion" between purine (C2), which is the input for the catabolic pathway that yields glyoxylate (C3) as a product, which BioREx failed to identify. BioREx did correctly identify the "Conversion" relations that link participants in the reaction in which (*S*)-ureidoglycine (C1) and glyoxylate (C3) are converted to oxalurate (C4) and glycine (C5). Note that "Indirect_conversion" relations do not provide direct knowledge of reactions and enzyme functions, and so FN cases will not have a major impact on our goal of identifying enzyme functions in literature.

The second major category of FN cases (~20%) are those where the text describes multiple substrates and products of a reaction, but where our BioREx model fails to classify all the possible pairwise relations between them, as in this example:

"*The purified enzyme catalyzed transacetylation of the acetyl group not only from* ***PAF*** $_{C1}$ *to* ***lysoplasmalogen*** $_{C2}$ *forming* ***plasmalogen analogs of PAF*** $_{C3}$*, but also to* ***sphingosine*** $_{C4}$ *producing* ***N-acetylsphingosine*** $_{C5}$ *(**C2-ceramide*** $_{C6}$*).*" (PMID:10085103).

Task	ChEBI matching required	End-to-end, NER – NEN – RE		
		P	R	F
Binary pair, binary classification	ChEBI exact	61.96	40.91	49.28
	ChEBI relaxed	62.61	41.17	49.55
Ternary tuple, binary classification	ChEBI exact	25.39	17.45	20.69
	ChEBI relaxed	25.46	17.50	20.74

Table 6. Performance of end-to-end pipeline combining NER, NEN, and RE on EnzChemRED. P, precision; R, recall; F, F_1 score.

This sentence describes the conversion of PAF (C1) and lysoplasmalogen (C2) to plasmalogen analogs of PAF (C3), and of PAF (C1) and sphingosine (C4) to acetylsphingosine (C5) (C2-ceramide (C6)). Our BioREx model correctly identified the binary conversions of lysoplasmalogen (C2) to plasmalogen analogs of PAF (C3) but failed to identify the curated binary "conversions" involving PAF (C1).

FP and FN predictions for ternary tuples. The classification of ternary tuples by BioREx suffered from many of the types of errors as the classification of binary pairs, as in this example of a FP ternary tuple, where multiple conversions and enzymes are found in the same sentence:

"*The proteins encoded by **YhfQ** $_{P1}$ and YhfN $_{P2}$ were overexpressed in E. coli, purified, and shown to catalyze the ATP $_{C1}$ -dependent phosphorylation of fructoselysine $_{C2}$ to a product identified as fructoselysine 6-phosphate $_{C3}$ by 31 P NMR (YhfQ $_{P3}$), and the reversible conversion of **fructoselysine 6-phosphate** $_{C4}$ and water to lysine and **glucose 6-phosphate** $_{C5}$ (**YhfN** $_{P4}$).*" (PMID:12147680).

The sentence provides evidence for the conversion of fructoselysine 6-phosphate (C4) to glucose 6-phosphate (C5), by YhfN (P4). BioREx erroneously classified the relation linking this pair of chemical mentions to the protein mention YhfQ (P1) as a "Conversion" when the correct enzyme is YhfN (P4).

FN ternary tuples were also observed when a single protein catalyzed multiple conversions:

"*We conclude that **acs1** $_{P1}$ encodes a bifunctional enzyme that converts **ribulose 5-phosphate** $_{C1}$ into **ribitol 5-phosphate** $_{C2}$ and further into **CDP-ribitol** $_{C3}$, which is the activated precursor form for incorporation of ribitol 5-phosphate into the H. influenzae type a capsular polysaccharide.*" (PMID: 10094675).

Our BioREx model correctly identified the conversion of ribulose 5-phosphate (C1) to ribitol 5-phosphate (C2), and of ribitol 5-phosphate (C2) to CDP-ribitol (C3) but failed to link the second of these conversions to acs1 (P1).

Evaluation of end-to-end pipeline on EnzChemRED.

We combined the best performing methods for NER (AIONER-PubMedBERT + EnzChemRED) and RE (BioREx + EnzChemRED) with MTCR for NEN to create a prototype end-to-end pipeline for enzyme function extraction from text. This achieved precision of 61.54%, recall of 37.93%, and F_1 score of 49.39% for the task of extraction and binary classification of binary chemical pairs from EnzChemRED abstracts (Table 6).

Significant losses in both precision and recall occurred during the normalization of chemical mentions to identifiers from the ChEBI ontology, with precision of 78.46%, recall of 67.29%, and F_1 score of 72.45% using AIONER and MTRC for this step. Relaxing the evaluation criterion – to allow matching to either the annotated ChEBI or to a direct parent or child node in the ChEBI ontology – only slightly increased the overall performance of the end-to-end pipeline.

Normalization of protein mentions to UniProt ACs was also problematic. In addition to MTCR we attempted UniProtKB AC normalization using the recently developed gene normalization tool GNorm2 (https://github.com/ncbi/GNorm2)[53], which employs a straightforward Entrez ID to UniProtKB AC mapping table to link gene mentions to unique UniProtKB ACs. Performance of GNorm2 on EnzChemRED was relatively modest, with precision, recall, and F_1 score of only 39.66%/11.43%/17.74% respectively. GNorm2 utilizes Entrez IDs to map to UniProt ACs, and many enzymes in EnzChemRED have no Entrez ID. We also attempted mapping of gene and protein mentions using the UniProt API (https://www.uniprot.org/help/query-fields), retaining the first match as the UniProt AC for each protein name. NER+NEN performance for proteins remained marginal, with precision, recall, and F_1 score of 30.41%, 56.43%, and 39.52%, respectively. The F_1 score for ternary tuples was therefore significantly lower than that for binary pairs, at 20.74%.

Application of end-to-end pipeline to PubMed abstracts.

We used the prototype end-to-end pipeline for enzyme function extraction described in the preceding section to extract candidate reaction pairs from 32 million PubMed abstracts (up to December 2023) (Table 7). We identified 680,426 relevant abstracts using our LitSuggest model, which together contained 158,837 distinct mentions of chemical pairs corresponding to one of the three valid relation types – "Conversion", "Indirect_conversion", or "Non_conversion" (BioREx_EnzChemRED_PubMed.tsv in Data Record). We then extracted all unique pairs of ChEBI IDs from these valid relation types (see Data Record BioREx_EnzChemRED_PubMed_normalized.tsv) and compared them to unique

		Number	% of total
Abstracts	Total processed	32,000,000	100
	Relevant according to LitSuggest	680,426	2.13
	Containing predicted binary chemical- chemical relations (any type)	64,077	0.20
Binary chemical- chemical relations	All relation types	176,084	100
	Conversion	158,837	90.21
	Indirect_conversion	16,889	9.59
	Non_conversion	358	0.20

Table 7. Results of applying the end-to-end pipeline on PubMed abstracts.

		In PubMed	In PubMed and in Rhea	% in Rhea
Unique pairs of ChEBI IDs identified by our pipeline	All relation types	37,715	3,152	8.36
	Conversion	30,661	2,721	8.87
	Indirect_conversion	6,986	428	6.13
	Non_conversion	68	3	4.41

Table 8. Comparison of chemical conversion pairs from PubMed abstracts to Rhea.

pairs of ChEBI IDs extracted from Rhea reactions (Rhea release 130 of 8th November 2023) as described in the Section **End-to-end pipeline of Methods**. The results of this comparison are shown in Table 8.

Most of the chemical conversion pairs identified in PubMed abstracts – over 91% – are not found in Rhea and may correspond to novel chemical reactions that are potential candidates for curation in both Rhea and UniProtKB/Swiss-Prot. Fig. 9 shows one approach to selecting potential high priority candidates among them using the Tree MAP (TMAP) algorithm to visualize the results from the end-to-end pipeline in chemical space. TMAP is an algorithm for the generation of intuitive visualizations of large data sets in the form of trees, which facilitates visual inspection of the data, including the detailed structure of clusters within the data and the relationships between clusters. TMAP is a popular method for visualizing chemical datasets, both molecules and reactions, and can accurately reflect similarities and differences in high-dimensional chemical space. The TMAP tree in Fig. 9 clusters chemical conversion pairs from PubMed abstracts and from Rhea reactions (those pairs for which both members have a defined chemical structure, or InChIKey) according to the similarity of chemical conversions they undergo, encoded using the differential reaction fingerprint (DRFP), which considers the differences in the circular substructures in the SMILES of each chemical pair. Clusters in the TMAP tree group chemical conversion pairs that have similar differences – chemical conversion pairs that, assuming that they are indeed pairs of reactants, would undergo similar changes during a reaction. To reduce the noise from the end-to-end pipeline we can apply additional filters to the chemical conversion pairs from PubMed abstracts and Rhea reactions in the TMAP tree. For instance, we would expect the main substrate-product pairs of curated reactions to share a high proportion of atoms on average, which would not be the case for erroneously identified chemical conversion pairs. For Rhea reaction pairs with InChIKeys the mean atom conservation is around 79.76%, and we have removed pairs from PubMed abstracts and Rhea reactions that fall below this threshold during the construction of the TMAP tree shown in Fig. 9. These types of filters can help curators to identify, and focus on, true novel chemical conversion pairs from the end-to-end pipeline and reduce noise.

Most of the chemical conversion pairs that are unique to PubMed abstracts form clusters in the TMAP tree with pairs from Rhea reactions, including several large clusters. This suggests that while the end-to-end pipeline may have identified many new chemical conversion pairs in PubMed, these generally undergo types of chemical conversion that have been previously curated in Rhea. The relatively low number of novel clusters from PubMed (all-blue clusters) that include no pairs from Rhea (red) suggests that coverage of published reaction chemistries in Rhea is already high, at least for those reactions that are described in abstracts and that involve chemical entities that can be mapped to ChEBI. Those all-blue clusters that are observed may represent novel reaction chemistries that are high priority targets for curation in UniProtKB/Swiss-Prot and Rhea, and curators are addressing these as a matter of priority. The 3,152 pairs of ChEBI IDs identified by our pipeline that do map to Rhea covered only 3,800 reactions of the total of 16,112 Rhea reactions at the time of analysis (Rhea release 130). The Rhea reactions that were not identified by our pipeline may have been missed due to low recall of our methods, or may only be described in full text and/or figures, which provide more information for NLP pipelines[68].

Usage Notes

This work describes EnzChemRED, a high-quality expert curated dataset of PubMed abstracts designed to support the development and benchmarking of NLP methods to extract knowledge of enzyme functions from scientific literature. We demonstrate the benefits of using EnzChemRED by fine-tuning pre-trained language models for NER and RE and combining these models in a prototype end-to-end pipeline that we used to build a draft map of enzyme functions from PubMed abstracts. While EnzChemRED was originally conceived to support the curation of enzyme functions in UniProtKB/Swiss-Prot using the reaction knowledgebase Rhea and the chemical ontology ChEBI, we hope that EnzChemRED can serve as a useful training and benchmarking set for NLP method developers and for other knowledgebases and resources.

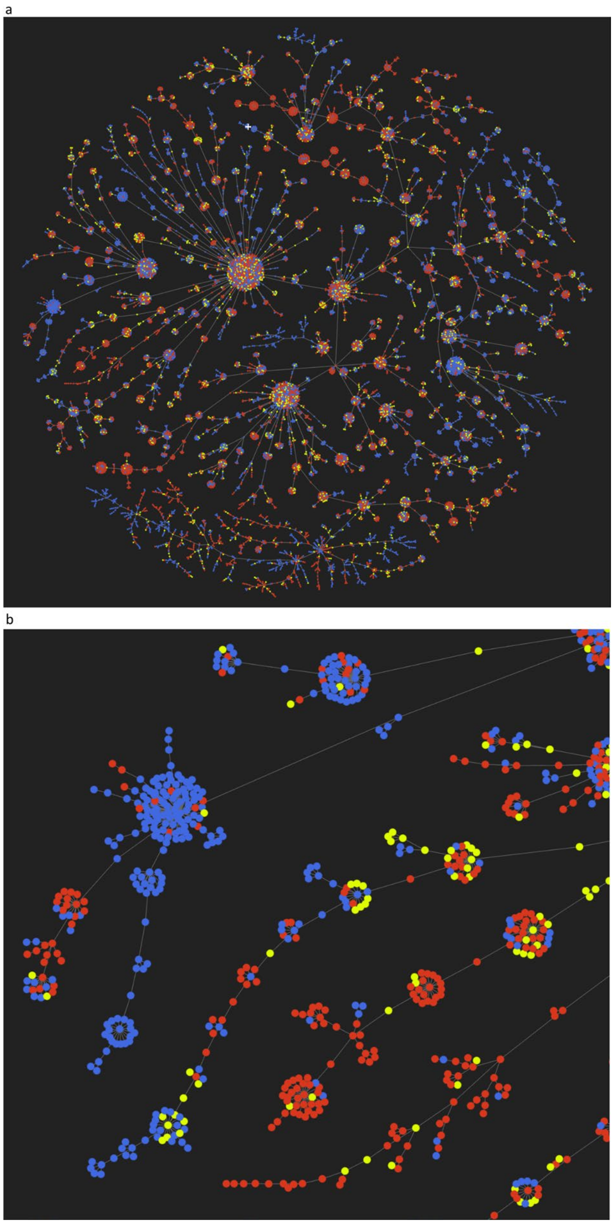


Fig. 9 TMAP tree of chemical conversion pairs extracted from PubMed abstracts and from Rhea. Each point in the TMAP tree corresponds to a chemical conversion pair, which are clustered according to the similarity of their Differential Reaction Fingerprint (DRFP). Blue, chemical conversion pairs from PubMed abstracts only; red, chemical conversion pairs from Rhea reactions only; yellow, chemical conversion pairs common to both PubMed abstracts and Rhea (as determined by ChEBI ID matching). (**a**) View of complete TMAP tree. (**b**) Zoom on the TMAP tree reveals further details of clustering of chemical pairs. Novel chemical pairs from PubMed abstracts, with chemical differences that differ from those curated in Rhea, appear as blue nodes that do not cluster with red, and may be one priority for curation at Rhea and UniProt.

Limitations. EnzChemRED has two main limitations. First, EnzChemRED was developed using PubMed abstracts rather than full text. We chose to focus on abstracts as the full text of many articles is not freely available, although full text may contain more information than abstracts[68], including complete descriptions of reactions that are necessary for the curation of Rhea. Processing open access full text therefore remains a desirable end goal, and we plan to address this in future work. We also plan to test computational methods to convert chemical conversion pairs from abstracts into complete candidate reactions, including by adding known co-substrates from existing Rhea reactions (such as NAD+/NADH). Combinations of chemical pairs from PubMed and co-substrates from Rhea that sum to balanced reactions where all atoms can be mapped between substrates and products, and whose reaction fingerprints resemble those of known reactions, will be prioritised for validation by curators. Second, EnzChemRED is also fairly small compared to some RE datasets (Table 1), consisting of only 1,210 PubMed abstracts, but our experiments indicate that EnzChemRED is already comprehensive enough to support the fine-tuning of language models that perform well at both NER and RE.

Tests of our end-to-end pipeline on EnzChemRED revealed limitations of the methods used, with NEN of gene/protein mentions to UniProtKB ACs using MTCR and GNorm2 proving to be a major challenge. Many of the proteins in EnzChemRED lack the Entrez IDs used in GNorm2, and we plan to expand the gene vocabulary of this tool using UniProt and other data sources. New corpora for species recognition may further assist gene and protein name disambiguation[69]. Normalization of chemical mentions to ChEBI IDs was also challenging due to the high complexity of the ChEBI ontology and the propensity of authors to use ambiguous chemical names that may be applicable to several isomers, including stereoisomers, as well as to related compounds or classes of compound, within the ChEBI ontology. For example, authors generally refer in abstracts to L-amino acids without specifying stereochemistry, so "L-serine" would generally be referred to as "serine", which our MTCR approach mapped to CHEBI:35243, a serine zwitterion of undefined stereochemistry. In contrast in the EnzChemRED dataset, mentions of "serine" were curated to CHEBI:33384, the "L-serine zwitterion", when information in the full text and other sources allowed. We found that allowing fuzzy matching to parents or children in the ChEBI ontology during NEN improved performance by a small margin (Table 7). We also expect that performance of NEN methods for chemicals may be lower for PubMed abstracts generally than for those abstracts that form part of EnzChemRED. The latter were drawn from publications used to curate UniProtKB/Swiss-Prot entries with Rhea reactions, meaning that the chemical mentions in EnzChemRED abstracts would have been targeted for curation in ChEBI; we would therefore expect a lower rate of false negatives for NEN on EnzChemRED abstracts than for NEN on other PubMed abstracts. One route to reducing FNs in chemical NEN may be to use larger chemical dictionaries such as PubChem, which we intend to explore. The best solution would be for authors to use unambiguous machine readable chemical identifiers such as SMILES or InChIs/InChIKeys, and some journals now recommend their use in publications[70]. Methods to extract reaction data from figures[71,72] may also provide a useful complement to the text-based methods tested here.

Feedback and large language models (LLMs). We will continue to develop the EnzChemRED dataset in response to feedback from users and to work to improve the methods for protein and chemical NER and NEN used in our prototype end-to-end pipeline. One area we are exploring is to extend the curation of EnzChemRED to include mappings of sentences and larger passages to other vocabularies, including Gene Ontology terms and the hierarchical enzyme classification of the IUBMB (EC numbers), to test embeddings and other approaches for semantic annotation of text. Curators of UniProtKB and Rhea are now using the output from the end-to-end pipeline to guide the curation of new enzymes and reactions in UniProtKB/Swiss-Prot and Rhea, including to link disconnected reactions in Rhea and UniProtKB/Swiss-Prot, analogous to gap-filling for genome scale metabolic models[73]. We are also testing the ability of large language models (LLMs) such as GPT-3.5 and GPT-4 to perform the tasks described here, as well as to extract knowledge of enzymes and their reactions directly using other kinds of simple prompt, with zero- and few-shot learning and fine-tuning. While smaller domain-specific models such as PubMedBERT and BioREx have to date proven to be at least as effective as much larger commercially available LLMs for most common NLP tasks tested[74,75], these much larger LLMs may offer a route to creating high performing RE systems using smaller training datasets for fine-tuning, or for augmentation of training datasets by approaches such as rephrasing.

Code availability

The source code for our data pipeline is publicly available from our GitLab repositories. AIONER-Bioformer and AIONER-PubMedBERT source codes are available at https://github.com/ncbi/AIONER. PubMedBERT and BioREx source codes are available at https://github.com/ncbi/biored and https://github.com/ncbi/biored, respectively. GNorm2 source code is available at https://github.com/ncbi/GNorm2. The PubMedBERT and BioREx models are available at https://github.com/ncbi/biored and https://github.com/ncbi/BioREx. The LitSuggest model described in this work is available at https://www.ncbi.nlm.nih.gov/research/litsuggest/project/5fa57e75bf71b3730469a83b.

Received: 1 May 2024; Accepted: 23 August 2024;
Published online: 09 September 2024

References

1. Ankrah, N. Y. D. *et al*. Enhancing Microbiome Research through Genome-Scale Metabolic Modeling. *mSystems* **6**, e0059921, https://doi.org/10.1128/mSystems.00599-21 (2021).
2. Thiele, I. *et al*. Personalized whole-body models integrate metabolism, physiology, and the gut microbiome. *Mol Syst Biol* **16**, e8982, https://doi.org/10.15252/msb.20198982 (2020).
3. Robinson, J. L. *et al*. An atlas of human metabolism. *Sci Signal* **13** https://doi.org/10.1126/scisignal.aaz1482 (2020).
4. Paneghetti, L., Bellettato, C. M., Sechi, A., Stepien, K. M. & Scarpa, M. One year of COVID-19: infection rates and symptoms in patients with inherited metabolic diseases followed by MetabERN. *Orphanet J Rare Dis* **17**, 109, https://doi.org/10.1186/s13023-022-02247-3 (2022).
5. Ambikan, A. T. *et al*. Multi-omics personalized network analyses highlight progressive disruption of central metabolism associated with COVID-19 severity. *Cell Syst* **13**, 665–681 e664, https://doi.org/10.1016/j.cels.2022.06.006 (2022).
6. Foguet, C. *et al*. Genetically personalised organ-specific metabolic models in health and disease. *Nat Commun* **13**, 7356, https://doi.org/10.1038/s41467-022-35017-7 (2022).
7. Probst, D. *et al*. Biocatalysed synthesis planning using data-driven learning. *Nat Commun* **13**, 964, https://doi.org/10.1038/s41467-022-28536-w (2022).
8. Sveshnikova, A., MohammadiPeyhani, H. & Hatzimanikatis, V. ARBRE: Computational resource to predict pathways towards industrially important aromatic compounds. *Metab Eng* **72**, 259–274, https://doi.org/10.1016/j.ymben.2022.03.013 (2022).
9. MohammadiPeyhani, H. *et al*. NICEdrug.ch, a workflow for rational drug design and systems-level analysis of drug metabolism. *Elife* **10**, e65543, https://doi.org/10.7554/eLife.65543 (2021).

10. Herisson, J. *et al*. The automated Galaxy-SynBioCAD pipeline for synthetic biology design and engineering. *Nat Commun* **13**, 5082, https://doi.org/10.1038/s41467-022-32661-x (2022).
11. Sankaranarayanan, K. *et al*. Similarity based enzymatic retrosynthesis. *Chem Sci* **13**, 6039–6053, https://doi.org/10.1039/d2sc01588a (2022).
12. Zheng, S. *et al*. Deep learning driven biosynthetic pathways navigation for natural products with BioNavi-NP. *Nat Commun* **13**, 3342, https://doi.org/10.1038/s41467-022-30970-9 (2022).
13. Levin, I., Liu, M., Voigt, C. A. & Coley, C. W. Merging enzymatic and synthetic chemistry with computational synthesis planning. *Nat Commun* **13**, 7747, https://doi.org/10.1038/s41467-022-35422-y (2022).
14. The UniProt Consortium. UniProt: the universal protein knowledgebase in 2021. *Nucleic Acids Res* **49**, D480–D489, https://doi.org/10.1093/nar/gkaa1100 (2021).
15. Morgat, A. *et al*. Enzyme annotation in UniProtKB using Rhea. *Bioinformatics* **36**, 1896–1901, https://doi.org/10.1093/bioinformatics/btz817 (2020).
16. Bansal, P. *et al*. Rhea, the reaction knowledgebase in 2022. *Nucleic Acids Res* **50**, D693–D700, https://doi.org/10.1093/nar/gkab1016 (2022).
17. Caspi, R. *et al*. The MetaCyc database of metabolic pathways and enzymes - a 2019 update. *Nucleic Acids Res* **48**, D445–D453, https://doi.org/10.1093/nar/gkz862 (2020).
18. Kanehisa, M., Furumichi, M., Sato, Y., Ishiguro-Watanabe, M. & Tanabe, M. KEGG: integrating viruses and cellular organisms. *Nucleic Acids Res* **49**, D545–D551, https://doi.org/10.1093/nar/gkaa970 (2021).
19. Chang, A. *et al*. BRENDA, the ELIXIR core data resource in 2021: new developments and updates. *Nucleic Acids Res* **49**, D498–D508, https://doi.org/10.1093/nar/gkaa1025 (2021).
20. Wittig, U., Rey, M., Weidemann, A., Kania, R. & Muller, W. SABIO-RK: an updated resource for manually curated biochemical reaction kinetics. *Nucleic Acids Res* **46**, D656–D660, https://doi.org/10.1093/nar/gkx1065 (2018).
21. Jassal, B. *et al*. The reactome pathway knowledgebase. *Nucleic Acids Res* **48**, D498–D503, https://doi.org/10.1093/nar/gkz1031 (2020).
22. The Gene Ontology Consortium. The Gene Ontology Knowledgebase in 2023. *Genetics* https://doi.org/10.1093/genetics/iyad031 (2023).
23. Wilkinson, M. D. *et al*. The FAIR Guiding Principles for scientific data management and stewardship. *Sci Data* **3**, 160018, https://doi.org/10.1038/sdata.2016.18 (2016).
24. Vaswani, A. *et al*. Attention is all you need. *Proceedings of the 31st International Conference on Neural Information Systems, NIPS' 17*., 6000–6010 https://doi.org/10.5555/3295222.3295349 (2017).
25. Lee, J. *et al*. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics* **36**, 1234–1240, https://doi.org/10.1093/bioinformatics/btz682 (2020).
26. Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *Proceedings of NAACL-HLT* **1**, 4171–4186, https://doi.org/10.18653/v1/N19-1423 (2019).
27. Peng, Y., Yan, S. & Lu, Z. Transfer learning in biomedical natural language processing: an evaluation of BERT and ELMo on ten benchmarking datasets. *Proceedings of the 2019 Workshop on Biomedical Natural Language Processing (BioNLP 2019)*, 58–65 (2019).
28. Gu, Y. *et al*. Domain-Specific Language Model Pretraining for Biomedical Natural Language Processing. *ACM Trans. Comput. Healthcare* **3**, Article 2 https://doi.org/10.1145/3458754 (2021).
29. Czarnecki, J., Nobeli, I., Smith, A. M. & Shepherd, A. J. A text-mining system for extracting metabolic reactions from full-text articles. *BMC Bioinformatics* **13**, 172, https://doi.org/10.1186/1471-2105-13-172 (2012).
30. Mallory, E. K. *et al*. Extracting chemical reactions from text using Snorkel. *BMC Bioinformatics* **21**, 217, https://doi.org/10.1186/s12859-020-03542-1 (2020).
31. Hastings, J. *et al*. ChEBI in 2016: Improved services and an expanding collection of metabolites. *Nucleic Acids Res* **44**, D1214–1219, https://doi.org/10.1093/nar/gkv1031 (2016).
32. Corbett, P. & Copestake, A. Cascaded classifiers for confidence-based chemical named entity recognition. *BMC Bioinformatics* **9**, S4, https://doi.org/10.1186/1471-2105-9-S11-S4 (2008).
33. Kolárik, C., Klinger, R., Friedrich, C. M., Hofmann-Apitius, M. & Fluck, J. Chemical Names: Terminological Resources and Corpora Annotation. *Workshop on Building and evaluating resources for biomedical text mining (6th edition of the Language Resources and Evaluation Conference)* **36**, 51–58 (2008).
34. Krallinger, M. *et al*. The CHEMDNER corpus of chemicals and drugs and its annotation principles. *J Cheminform* **7**, S2, https://doi.org/10.1186/1758-2946-7-S1-S2 (2015).
35. Wei, C. H. *et al*. Assessing the state of the art in biomedical relation extraction: overview of the BioCreative V chemical-disease relation (CDR) task. *Database (Oxford)* **2016**, baw032, https://doi.org/10.1093/database/baw032 (2016).
36. Islamaj, R. *et al*. The corpus of the BioRED Track at BioCreative VIII. *Proceedings of the BioCreative VIII Challenge and Workshop: Curation and Evaluation in the era of Generative Models*. (2023).
37. Luo, L., Lai, P. T., Wei, C. H., Arighi, C. N. & Lu, Z. BioRED: a rich biomedical relation extraction dataset. *Brief Bioinform* **23**, bbac282, https://doi.org/10.1093/bib/bbac282 (2022).
38. Huang, M. S. *et al*. Biomedical named entity recognition and linking datasets: survey and our recent development. *Brief Bioinform* **21**, 2219–2238, https://doi.org/10.1093/bib/bbaa054 (2020).
39. Gurulingappa, H. *et al*. Development of a benchmark corpus to support the automatic extraction of drug-related adverse effects from medical case reports. *J Biomed Inform* **45**, 885–892, https://doi.org/10.1016/j.jbi.2012.04.008 (2012).
40. Krallinger, M. *et al*. Overview of the BioCreative VI chemical-protein interaction Track. *Proceedings of the sixth BioCreative challenge evaluation workshop* **1**, 141–146 (2017).
41. Miranda-Escalada, A. *et al*. Overview of DrugProt task at BioCreative VII: data and methods for large-scale text mining and knowledge graph generation of heterogenous chemical-protein relations. *Database (Oxford)* **2023**, baad080, https://doi.org/10.1093/database/baad080 (2023).
42. Peng, N., Poon, H., Quirk, C., Toutanova, K. & Yih, W. Cross-Sentence N-ary Relation Extraction with Graph LSTMs. *Transactions of the Association for Computational Linguistics* **5**, 101–115, https://doi.org/10.1162/tacl_a_00049 (2017).
43. Herrero-Zazo, M., Segura-Bedmar, I., Martinez, P. & Declerck, T. The DDI corpus: an annotated corpus with pharmacological substances and drug-drug interactions. *J Biomed Inform* **46**, 914–920, https://doi.org/10.1016/j.jbi.2013.07.011 (2013).
44. He, J. *et al*. ChEMU 2020: Natural Language Processing Methods Are Effective for Information Extraction From Chemical Patents. *Front Res Metr Anal* **6**, 654438, https://doi.org/10.3389/frma.2021.654438 (2021).
45. Nadendla, S. *et al*. ECO: the Evidence and Conclusion Ontology, an update for 2022. *Nucleic Acids Res* **50**, D1515–D1521, https://doi.org/10.1093/nar/gkab1025 (2022).
46. Allot, A., Lee, K., Chen, Q., Luo, L. & Lu, Z. LitSuggest: a web-based system for literature recommendation and curation using machine learning. *Nucleic Acids Res* **49**, W352–W358, https://doi.org/10.1093/nar/gkab326 (2021).
47. Islamaj, R., Kwon, D., Kim, S. & Lu, Z. TeamTat: a collaborative text annotation tool. *Nucleic Acids Res* **48**, W5–W11, https://doi.org/10.1093/nar/gkaa333 (2020).
48. Wei, C. H., Allot, A., Leaman, R. & Lu, Z. PubTator central: automated concept annotation for biomedical full text articles. *Nucleic Acids Res* **47**, W587–W593, https://doi.org/10.1093/nar/gkz389 (2019).

49. Wei, C.-H. *et al*. PubTator 3.0: an AI-powered Literature Resource for Unlocking Biomedical Knowledge. *Nucleic Acids Research* (2024).
50. Kim, S. *et al*. PubChem 2023 update. *Nucleic Acids Res* **51**, D1373–D1380, https://doi.org/10.1093/nar/gkac956 (2023).
51. Luo, L. *et al*. AIONER: all-in-one scheme-based biomedical named entity recognition using deep learning. *Bioinformatics* **39** https://doi.org/10.1093/bioinformatics/btad310 (2023).
52. Li, J. *et al*. BioCreative V CDR task corpus: a resource for chemical disease relation extraction. *Database (Oxford)* **2016**, baw068, https://doi.org/10.1093/database/baw068 (2016).
53. Wei, C. H., Luo, L., Islamaj, R., Lai, P. T. & Lu, Z. GNorm2: an improved gene name recognition and normalization system. *Bioinformatics* **39**, btad599, https://doi.org/10.1093/bioinformatics/btad599 (2023).
54. Islamaj, R. *et al*. NLM-Gene, a richly annotated gold standard dataset for gene entities that addresses ambiguity and multi-species gene recognition. *J Biomed Inform* **118**, 103779, https://doi.org/10.1016/j.jbi.2021.103779 (2021).
55. Islamaj, R. *et al*. NLM-Chem, a new resource for chemical entity recognition in PubMed full text literature. *Sci Data* **8**, 91, https://doi.org/10.1038/s41597-021-00875-1 (2021).
56. Fang, L., Chen, Q., Wei, C.-H., Lu, Z. & Wang, K. Bioformer: an efficient transformer language model for biomedical text mining. *arXiv* https://doi.org/10.48550/arXiv.2302.01588 (2023).
57. D'Souza, J. & Ng, V. Sieve-based entity linking for the biomedical domain. *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)*, 297–302 (2015).
58. Sohn, S., Comeau, D. C., Kim, W. & Wilbur, W. J. Abbreviation definition identification based on automatic precision estimates. *BMC Bioinformatics* **9**, 402, https://doi.org/10.1186/1471-2105-9-402 (2008).
59. Heller, S. R., McNaught, A., Pletnev, I., Stein, S. & Tchekhovskoi, D. InChI, the IUPAC International Chemical Identifier. *J Cheminform* **7**, 23, https://doi.org/10.1186/s13321-015-0068-4 (2015).
60. Leaman, R. *et al*. Chemical identification and indexing in full-text articles: an overview of the NLM-Chem track at BioCreative VII. *Database (Oxford)* **2023** https://doi.org/10.1093/database/baad005 (2023)
61. Lai, P.-T., Wei, C.-H., Luo, L., Chen, Q. & Lu, Z. BioREx: Improving Biomedical Relation Extraction by Leveraging Heterogeneous Datasets. *Journal of Biomedical Informatics* **146** (2023).
62. Probst, D. & Reymond, J. L. Visualization of very large high-dimensional data sets as minimum spanning trees. *J Cheminform* **12**, 12, https://doi.org/10.1186/s13321-020-0416-x (2020).
63. Probst, D., Schwaller, P. & Reymond, J. L. Reaction classification and yield prediction using the differential reaction fingerprint DRFP. *Digit Discov* **1**, 91–97, https://doi.org/10.1039/d1dd00006c (2022).
64. Comeau, D. C. *et al*. BioC: a minimalist approach to interoperability for biomedical text processing. *Database (Oxford)* **2013**, bat064, https://doi.org/10.1093/database/bat064 (2013).
65. Lai, P. T. *et al*. EnzChemRED, a rich enzyme chemistry relation extraction dataset. *Zenodo*. https://doi.org/10.5281/zenodo.11067997 (2024).
66. Tong, Y. *et al*. Improving biomedical named entity recognition by dynamic caching inter-sentence information. *Bioinformatics* **38**, 3976–3983, https://doi.org/10.1093/bioinformatics/btac422 (2022).
67. Wei, C. H., Kao, H. Y. & Lu, Z. GNormPlus: An Integrative Approach for Tagging Genes, Gene Families, and Protein Domains. *Biomed Res Int* **2015**, 918710, https://doi.org/10.1155/2015/918710 (2015).
68. Westergaard, D., Staerfeldt, H. H., Tonsberg, C., Jensen, L. J. & Brunak, S. A comprehensive and quantitative comparison of text-mining in 15 million full-text articles versus their corresponding abstracts. *PLoS Comput Biol* **14**, e1005962, https://doi.org/10.1371/journal.pcbi.1005962 (2018).
69. Luoma, J. *et al*. S1000: a better taxonomic name corpus for biomedical information extraction. *Bioinformatics* **39**, btad369, https://doi.org/10.1093/bioinformatics/btad369 (2023).
70. Schymanski, E. L. & Bolton, E. E. FAIR chemical structures in the Journal of Cheminformatics. *J Cheminform* **13**, 50, https://doi.org/10.1186/s13321-021-00520-4 (2021).
71. Wilary, D. M. & Cole, J. M. ReactionDataExtractor 2.0: A Deep Learning Approach for Data Extraction from Chemical Reaction Schemes. *J Chem Inf Model* **63**, 6053–6067, https://doi.org/10.1021/acs.jcim.3c00422 (2023).
72. Qian, Y., Guo, J., Tu, Z., Coley, C. W. & Barzilay, R. RxnScribe: A Sequence Generation Model for Reaction Diagram Parsing. *J Chem Inf Model* **63**, 4030–4041, https://doi.org/10.1021/acs.jcim.3c00439 (2023).
73. Pan, S. & Reed, J. L. Advances in gap-filling genome-scale metabolic models and model-driven experiments lead to novel metabolic discoveries. *Curr Opin Biotechnol* **51**, 103–108, https://doi.org/10.1016/j.copbio.2017.12.012 (2018).
74. Chen, Q. *et al*. Large language models in biomedical natural language processing: benchmarks, baselines, and recommendations. *arXiv preprint arXiv:2305.16326* (2023).
75. Tian, S. *et al*. Opportunities and Challenges for ChatGPT and Large Language Models in Biomedicine and Health. *Briefings in Bioinformatics* **25**, bbad493 (2024).
76. McCoy, J. G. *et al*. Structure and mechanism of mouse cysteine dioxygenase. *Proc Natl Acad Sci USA* **103**, 3084–3089, https://doi.org/10.1073/pnas.0509262103 (2006).

Acknowledgements

We would like to thank the Cheminformatics and Metabolism Team of EMBL-EBI for their work in maintaining and developing ChEBI, without which Rhea would not be possible, particularly Adnan Malik for expert curation. We are indebted to Blanca Cabrera Gil at Swiss-Prot for stimulating discussions and advice on machine learning in chemistry. We gratefully acknowledge the software contributions of Chemaxon (https://www.chemaxon.com/products/marvin/). Expert curation and evaluation of EnzChemRED at Swiss-Prot were supported by the Swiss Federal Government through the State Secretariat for Education, Research and Innovation (SERI) and the National Human Genome Research Institute (NHGRI), Office of Director [OD/DPCPSI/ODSS], National Institute of Allergy and Infectious Diseases (NIAID), National Institute on Aging (NIA), National Institute of General Medical Sciences (NIGMS), National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK), National Eye Institute (NEI), National Cancer Institute (NCI), National Heart, Lung, and Blood Institute (NHLBI) of the National Institutes of Health [U24HG007822], and by the European Union's Horizon Europe Framework Programme project Recon4IMD (grant number 101080997), supported in Switzerland through the State Secretariat for Education, Research and Innovation (SERI). This research was also supported by the NIH Intramural Research Program, National Library of Medicine and by the Fundamental Research Funds for the Central Universities [DUT23RC(3)014 to L.L.]. Funding for open access charge: NIH.

Author contributions

P.-T.L. – data processing, model development, analysis, writing. E.C. – data processing, data annotation, analysis, writing. L.A. – data annotation, analysis, writing. K.A. –data annotation, analysis, writing. L.B. – data annotation, analysis, writing. E. de C. – data processing, analysis, writing. M.F. – data annotation, analysis, writing. A.M. – data annotation, analysis, writing. L.P. – data annotation, analysis, writing. I.P. – data annotation, analysis, writing. S.P. – data annotation, analysis, writing. N.R. – data processing, analysis, writing. C.R. – data annotation, analysis, writing. A.S. – data processing, analysis, writing. C.-H.W. – data processing, model development, analysis, writing. R.L. – model development, analysis, writing. L.L. – data processing, model development, analysis, writing. Z.L. – project conception, data acquisition, writing. A.B. – project conception, data acquisition, data annotation, analysis, writing.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to Z.L. or A.B.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.