Supplementary data for

BERN2: an advanced neural biomedical named entity recognition and normalization tool

Mujeen Sung [1,†], Minbyul Jeong [1,†], Yonghwa Choi [1], Donghyeon Kim [2], Jinhyuk Lee [1,*] and Jaewoo Kang [1,3,*]

[1] Department of Computer Science and Engineering, Korea University, Seoul, 02841, Republic of Korea
[2] AIRS Company, Hyundai Motor Group, Seoul, 06620, Republic of Korea
[3] AIGEN Sciences, Seoul, 04778, Republic of Korea

[†] These authors wish it to be known that they should be regarded as joint first authors.
[*] To whom correspondence should be addressed. These authors wish it to be known that they should be regarded as joint last authors.

Application website: http://bern2.korea.ac.kr
Contact: jinhyuk_lee@korea.ac.kr, kangj@korea.ac.kr

A Usage of BERN2

BERN2 is available as a web service (http://bern2.korea.ac.kr/) or as a local installation (source code and installation manual in https://github.com/dmis-lab/BERN2) depending on the users' preferences. BERN2 web service is useful for users to get the annotations of biomedical texts without installing or to get pre-computed annotations in an instant. Using BERN2 locally, on the other hand, provides users with a more stable and customizable way, allowing them to add external modules on their own.

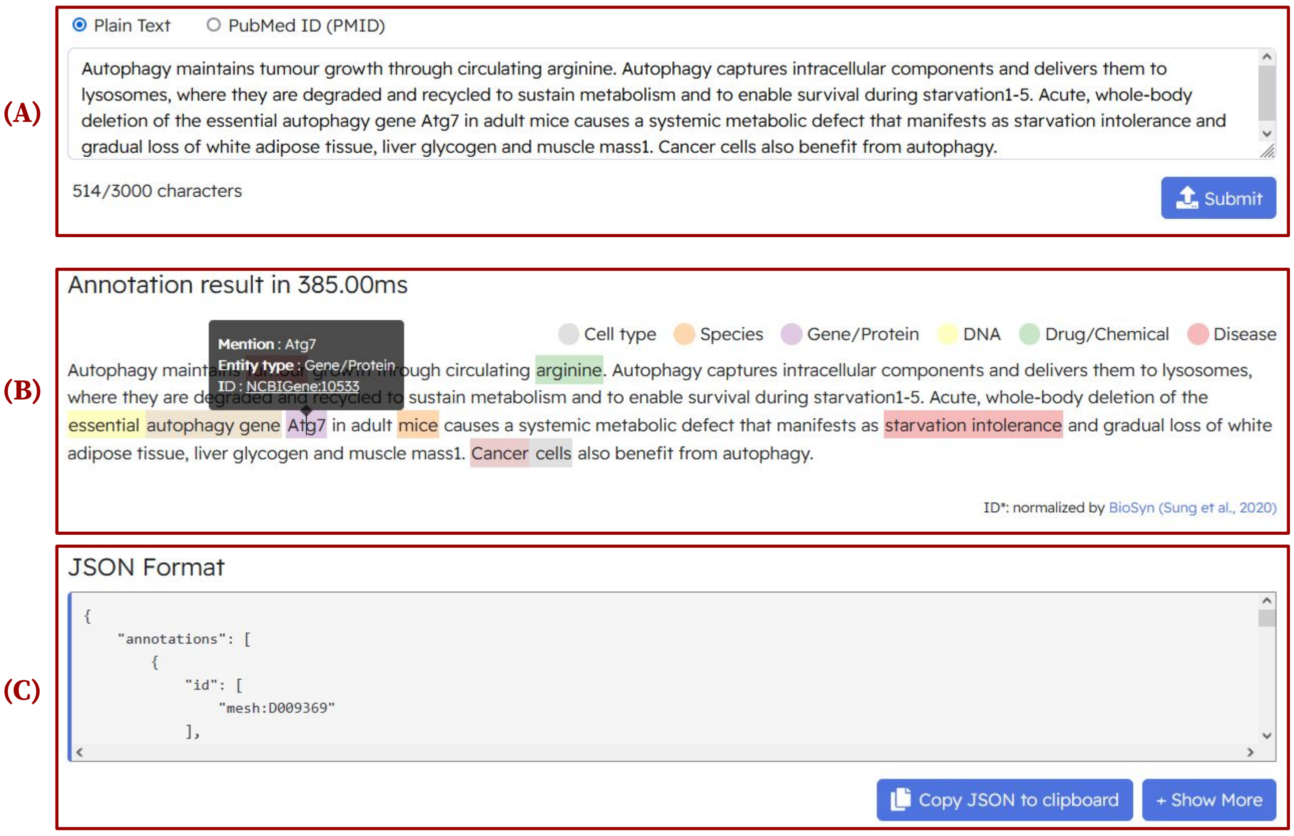
In this section, we describe the web service in terms of the interactive web demo and RESTful APIs, but please note that the usages described here can be applied to local BERN2 as well.

A.1 Interactive Web Demo

We provide the interactive BERN2 web demo so that users can easily access our tool. Fig. S1 shows an example web page from our demo. When a user types in plain text or a PubMed ID (PMID) in the input box and presses the submit button (Fig. S1-A), the annotated results (Fig. S1-B) and their JSON results (Fig. S 1-C) are displayed. For the annotation results, recognized entity spans are colored with their entity types, and their CUIs appear by clicking on them.

A.2 RESTful APIs

As shown in Table S1, BERN2 offers RESTful APIs to allow users to get annotation results for plain texts or PMIDs in a programmable way. The URL format for a single or multiple PubMed abstracts is http://bern2.korea.ac.kr/pubmed/<PMID> using the HTTP GET method. For plain text, the format of the URL is http://bern2.korea.ac.kr/plain using the HTTP POST method. Listing 1 shows an example of the JSON results from BERN2.



Autophagy maintains tumour growth through circulating arginine. Autophagy captures intracellular components and delivers them to lysosomes, where they are degraded and recycled to sustain metabolism and to enable survival during starvation1-5. Acute, whole-body deletion of the essential autophagy gene Atg7 in adult mice causes a systemic metabolic defect that manifests as starvation intolerance and gradual loss of white adipose tissue, liver glycogen and muscle mass1. Cancer cells also benefit from autophagy.

Mention : Atg7
Entity type : Gene/Protein
ID : NCBIGene:10533

Autophagy maintains tumour growth through circulating arginine. Autophagy captures intracellular components and delivers them to lysosomes, where they are degraded and recycled to sustain metabolism and to enable survival during starvation1-5. Acute, whole-body deletion of the essential autophagy gene Atg7 in adult mice causes a systemic metabolic defect that manifests as starvation intolerance and gradual loss of white adipose tissue, liver glycogen and muscle mass1. Cancer cells also benefit from autophagy.

```
{
    "annotations": [
        {
            "id": [
                "mesh:D009369"
            ],
```

Fig. S1: Interactive web demo of BERN2 (http://bern2.korea.ac.kr). (A) Input box and submit button. (B) Annotation results. (C) Annotation results in the JSON format.

Table S1. BERN2 APIs and URL examples

API	HTTP Method	URL example	Data
Single PMID	GET	http://bern2.korea.ac.kr/pubmed/29446767	-
PMIDs	GET	http://bern2.korea.ac.kr/pubmed/29446767,2568119	-
Plain text	POST	http://bern2.korea.ac.kr/plain	{"text":"tumour growth through arginine"}

Listing 1: An example of BERN2 annotations in the JSON format

```
{
    "annotations": [
        {
            "id": [
                "mesh:D009369"
            ],
            "is_neural_normalized": false,
            "prob": 0.9999922513961792,
            "mention": "tumour",
            "obj": "disease",
            "span": {
                "begin": 20,
                "end": 26
            }
        },
        {
            "id": [
                "mesh:D001120"
            ],
            "is_neural_normalized": false,
            "prob": 0.9819278717041016,
            "mention": "arginine",
            "obj": "drug",
            "span": {
                "begin": 54,
                "end": 62
            }
        }
    ],
    "text": "Autophagy maintains tumour growth through circulating arginine.",
    "timestamp": "Thu Dec 23 04:12:28 +0000 2021"
}
```

B Named Entity Recognition for BERN2

The goal of named entity recognition (NER) is to detect biomedical entity spans given a biomedical text. While BERN uses the biomedical language model, BioBERT (Lee *et al.*, 2020), for high-performance NER, it has some limitations from using separate single-task NER models for each entity type (i.e., five separate NER models required for annotating five entity types). This approach requires larger GPU memory for parallelization but is very slow at inference when used sequentially (Fig. S2a). To overcome these limitations, we adopt a biomedical multi-task NER model motivated by Wang *et al.* (2019). In this section, we describe the design of our multi-task NER model, the decision rules for resolving overlapping entities, and its performance on out-of-distribution datasets.

B.1 Multi-task Named Entity Recognition

Following Wang *et al.* (2019), our multi-task NER model consists of a shared backbone language model and separate task-specific layers per entity type as shown in Fig. S2b. We choose the state-of-the-art biomedical language model, Bio-LM (Lewis *et al.*, 2020)[1], as our backbone model and a two-layer MLP network with ReLU activation as a task-specific layer. The outputs of each task-specific layer are the probabilities of three classes (i.e., Begin, Inside, and Outside). At training time, we merge all train sets across all entity types and use cross-entropy objectives to optimize our model. We define a loss function $\mathcal{L}$ for our model as follows,

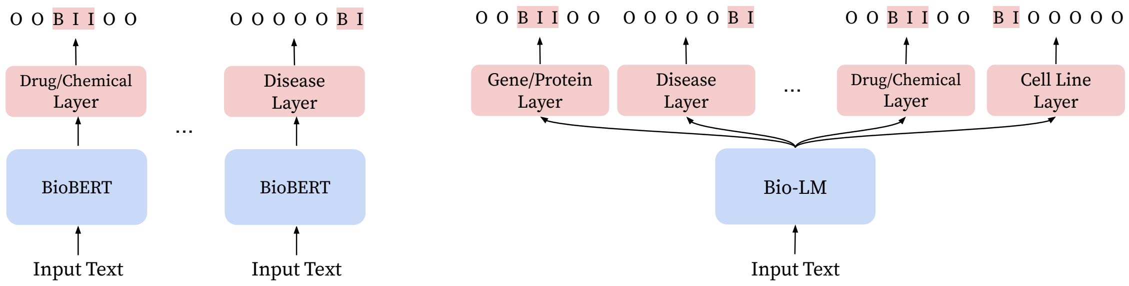
$$\mathcal{L} = -\frac{1}{NM}\sum_{i=1}^{N}\sum_{j=1}^{M} y_{ij} \cdot log(p_{ij}), \tag{1}$$

where N is the number of task-specific layers, M is the max sequence length of input texts, y_{ij} denotes the ground-truth label, and p_{ij} denotes the probability that each task-specific layer produces.

B.2 Overlapping Entity Resolution

Recognized entities in biomedical texts can overlap between entity types. For example, in ‘the androgen is synthesized from ...’, a gene/protein type layer and a drug/chemical type layer can annotate ‘androgen’ as both types because androgen is a natural or synthetic steroid hormone. Therefore, it is important to resolve these ambiguous overlapping entities and provide more appropriate entities based on the context. Following the decision rules proposed by BERN, we resolve overlapping entities based on the highest probability of each type. We also include another heuristic where normalized entities (see

[1] We use the *RoBERTa-large-PM-M3-Voc* checkpoint released in https://github.com/facebookresearch/bio-lm.



(a) BERN (single-task NER)   (b) BERN2 (multi-task NER)

Fig. S2: Comparison of the NER model architectures between BERN and BERN2. BERN2 uses a shared backbone language model and task-specific layers for each entity type, allowing fast parallel inference and efficient memory usage.

Section C) are preferred over entities that failed to be normalized. Lastly, we include mutation entities when recognized by tmVar2.0 (Wei *et al.*, 2018) because it achieves high precision (over 97%) while producing no probability to compare with other entity types. Fig. S3 illustrates the decision rules of our overlapping entity resolution. Table S2 shows statistics of the overlapping entities between pairs of types in randomly sampled 308K abstracts (1% of PubMed). For example, 9,364 entity overlaps are resolved between the gene/protein type and the drug/chemical type in 308K abstracts.

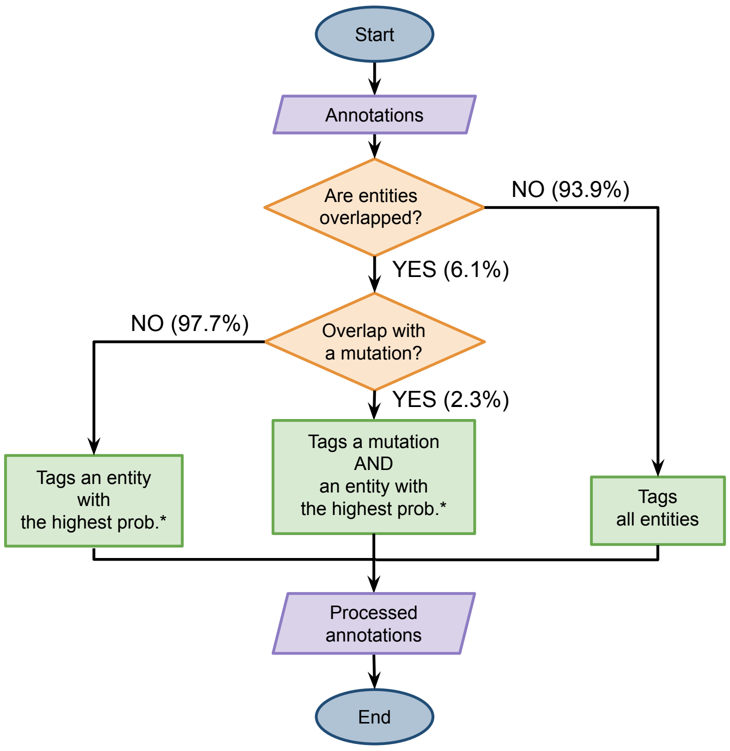


Fig. S3: The decision rules for resolving overlapping entities in BERN2. Between two entities, a normalized entity is preferred over an entity that failed to be normalized (*).

Table S2. Overlap statistics of recognized entities in randomly sampled 308K PubMed abstracts. The numbers in each cell indicate the count of completely overlapping cases of the entity pair. The percentages in parentheses indicate the ratio of overlapping entities of the entity type pair.

	Gene/Protein	Disease	Drug/Chemical	Mutation	Species	Cell Line	Cell Type	DNA	RNA
Gene/Protein	-	2,332 (0.10%)	9,364 (0.41%)	1,582 (0.15%)	1,144 (0.06%)	3,681 (0.29%)	4,260 (0.29%)	108,744 (7.60%)	14,496 (1.30%)
Disease	2,332 (0.10%)	-	862 (0.04%)	2 (0.0002%)	27,919 (1.26%)	8,838 (0.60%)	7,945 (0.48%)	752 (0.05%)	431 (0.03%)
Drug/Chemical	9,364 (0.41%)	862 (0.04%)	-	138 (0.01%)	30 (0.001%)	2,549 (0.17%)	769 (0.05%)	9,297 (0.57%)	869 (0.07%)
Mutation	1,582 (0.15%)	2 (0.0002%)	138 (0.01%)	-	4 (0.0004%)	246 (0.09%)	6 (0.001%)	5,252 (1.29%)	727 (0.79%)
Species	1,144 (0.06%)	27,919 (1.26%)	30 (0.001%)	4 (0.0004%)	-	4,331 (0.35%)	6,011 (0.42%)	1,345 (0.10%)	527 (0.05%)
Cell Line	3,681 (0.29%)	8,838 (0.60%)	2,549 (0.17%)	246 (0.09%)	4,331 (0.35%)	-	109,467 (16.11%)	3,457 (0.54%)	91 (0.03%)
Cell Type	4,260 (0.29%)	7,945 (0.48%)	769 (0.05%)	6 (0.001%)	6,011 (0.42%)	109,467 (16.11%)	-	1,616 (0.20%)	512 (0.10%)
DNA	108,744 (7.60%)	752 (0.05%)	9,297 (0.57%)	5,252 (1.29%)	1,345 (0.10%)	3,457 (0.54%)	1,616 (0.20%)	-	7,777 (1.66%)
RNA	14,496 (1.30%)	431 (0.03%)	869 (0.07%)	727 (0.79%)	527 (0.05%)	91 (0.03%)	512 (0.10%)	7,777 (1.66%)	-

B.3 Evaluation on Out-of-Distribution Datasets

Apart from the evaluation on in-domain datasets described in the main paper, we also evaluate BERN2 on out-of-distribution datasets (Bada *et al.*, 2012; Pyysalo *et al.*, 2013; Kim *et al.*, 2019a) to measure the generalization ability of our NER model. Table S3 shows that BERN2 outperforms BERN by 2.0% macro-averaged F1 score. This reveals that the Bio-LM NER model (Lewis *et al.*, 2020) trained in multi-task learning detects biomedical entities from out-of-distribution contexts more robustly than the BioBERT NER model (Lee *et al.*, 2020) trained in single-task learning. However, BERN2 underperforms HunFlair (Weber *et al.*, 2021), which uses the Flair model (Akbik *et al.*, 2019) with LSTM-CRF-based prediction layers trained on 23 NER corpora. We leave an investigation on making BERN2 more generalizable in out-of-distribution circumstances as future work.

C Named Entity Normalization for BERN2

A named entity normalization (NEN) model aims to find the concept unique IDs (CUI) of entities recognized by the NER model. In addition to the rule-based NEN models used in Kim *et al.* (2019b)[2], BERN2 applies neural network-based NEN models (Sung *et al.*, 2020) to increase the number of correctly normalized entities. Specifically, BERN2 use the neural network-based NEN models to normalize entities that failed to be normalized by rule-based NEN models. We call this hybrid NEN. We employ the hybrid NEN model for gene/protein, disease, and drug/chemical types where the public

[2] GNormPlus (Wei *et al.*, 2015) for gene/protein, Sieve-based entity linking (D’Souza and Ng, 2015) for disease, tmChem (Leaman *et al.*, 2015) without Ab3P for drug/chemical, tmVar2.0 (Wei *et al.*, 2018) for mutation, and dictionary lookup for species.

Table S3. Results on out-of-distribution biomedical NER benchmarks. F1 score is reported. Following the evaluation in Weber et al. (2021), Misc refers to tmChem (Leaman et al., 2015) for Chemical, GNormPlus (Wei et al., 2015) for Gene and Species, and DNorm (Leaman et al., 2013) for Disease.

Dataset	Type	Misc	HunFlair	BERN	BERN2
CRAFT	Gene/Protein	64.9	**72.2**	56.9	57.7
(Bada *et al.*, 2012)	Drug/Chemical	42.9	**59.7**	54.9	55.4
	Species	81.2	85.1	82.4	**86.5**
BioNLP CG	Gene/Protein	69.0	**87.7**	76.8	78.5
(Pyysalo *et al.*, 2013)	Disease	55.6	**65.1**	63.2	63.7
	Drug/Chemical	72.2	**81.8**	73.0	76.1
	Species	80.5	76.5	81.3	**85.6**
PDR (Kim *et al.*, 2019a)	Disease	80.6	**83.4**	78.8	79.3
Average		68.4	**76.4**	70.9	72.9

datasets (Morgan *et al.*, 2008; Li *et al.*, 2016) are available for training neural network-based NEN models. For four entity types, mutation, species, cell line, and cell type, we only use the rule-based NEN models due to the lack of public normalization datasets. In this section, we describe the dictionaries used for NEN and the statistics of recognized and normalized entities in randomly sampled 308K abstracts.

C.1 Dictionaries and Statistics

BERN2 can normalize entities of seven different types.[3] Table S4 shows the dictionaries used for each type and their statistics. For instance, the gene/protein dictionary, NCBI Gene (Brown *et al.*, 2015), has 67,370 CUIs and the corresponding 277,944 entity names. We also report the number of normalized entities on randomly sampled 308K abstracts (1% of PubMed) in Table S5. For example, for the species type, 974,935 entities are recognized and 864,904 (88.7%) are normalized among them. Since entities of three types (gene/protein, disease, and drug/chemical) are normalized into the CUIs by the hybrid NEN models, there could be incorrect CUIs unlike when solely using the dictionaries or rule-based approaches. Taking this into account, we report the estimated number of correctly normalized entities considering the accuracies of the neural NEN models. For example, for the gene/protein type, when 436,851 entities are normalized by the rule-based model and 585,583 entities are normalized by the neural network-based NEN model, which has 91.3% accuracy on the gene/protein normalization benchmark, we estimate the ratio of the correctly normalized gene/protein type entities as follows: (436,851 + 585,583 * 0.913) / (436,851 + 585,583) = 95.0%. In the case of two entity types, cell line and cell type, the number of normalized entities is low due to the use of the simple dictionary lookup method.

D Use Cases of BERN

We conclude our supplementary data by introducing two use cases of BERN (Kim *et al.*, 2019b) in biomedical downstream tasks. Despite its applicability for a variety of tasks, its bottleneck is the

[3] NEN for DNA and RNA is not currently supported due to the lack of available dictionaries

Table S4. The dictionaries used for normalization and their statistics.

Entity type	**Dictionaries**	**# of CUIs**	**# of names**	**Avg. # of names per ID**
Gene/Protein	NCBI Gene (Brown *et al.*, 2015)	67,370	277,944	4.1
Disease	MeSH (Lipscomb, 2000), OMIM (Hamosh *et al.*, 2005)	12,174	141,497	11.6
Drug/Chemical	MeSH (Lipscomb, 2000), CHEBI (Degtyarenko *et al.*, 2007)	212,317	1,147,293	5.4
Mutation	dbSNP (Sherry *et al.*, 2001), ClinVar (Landrum *et al.*, 2016)	208,474	302,498	1.5
Species	NCBI Taxonomy (Federhen, 2012)	398,037	3,119,005	7.8
Cell Line	Cellosaurus (Bairoch, 2018)	128,806	220,824	1.7
Cell Type	Cell Ontology (Diehl *et al.*, 2016)	2,525	4,970	2.0

Table S5. Normalization models and the statistics of recognized and normalized entities in randomly sampled 308K PubMed abstracts. For the types that use hybrid NEN, we report the estimated number of correctly normalized entities based on the performance of the neural network-based NEN models.

Entity type	**Normalization model**	**# of recognized entities**	**# of normalized entities**	**# of abstracts with each entity types**	**Avg. # of recognized entities per abstract**	**Avg. # of normalized entities per abstract**
Gene/Protein	Hybrid	1,022,434	971,312 (95.0%)	117,367	8.7	8.3
Disease	Hybrid	1,210,336	1,178,867 (97.4%)	178,233	6.8	6.6
Drug/Chemical	Hybrid	1,233,942	1,226,538 (99.4%)	152,692	8.1	8.1
Mutation	tmVar2.0	14,570	14,570 (100%)	3,827	3.8	3.8
Species	Dictionary lookup	974,935	864,904 (88.7%)	227,345	4.3	3.8
Cell Line	Dictionary lookup	180,355	13,480 (7.5%)	66,653	2.7	0.2
Cell Type	Dictionary lookup	356,065	36,788 (10.3%)	107,460	3.3	0.3

relatively slow processing time required to annotate large-scale biomedical texts. Therefore, we believe that our BERN2 with much faster inference will be widely used to solve lots of tasks that require biomedical entities.

Biomedical knowledge graph construction Biomedical texts contain a considerable amount of expert knowledge. Constructing biomedical knowledge graphs from biomedical texts can help researchers navigate this information more easily. Biomedical entities extracted by NER tools can be used as essential components (e.g., nodes in the graphs) when building biomedical knowledge graphs. For example, Xu *et al.* (2020) use BERN to automatically extract biomedical entities from 29M PubMed abstracts and build a PubMed knowledge graph, utilizing the extracted entities as nodes.

Biomedical entity-based search engine As biomedical documents are rich in domain-specific terminology, extracting named entities can enhance search engines to find documents related to the entities in queries. Vapur (Köksal *et al.*, 2020), a search engine for finding protein compounds in COVID-19 literature, pre-processes a document into a set of triples (*Entity1*, *Relation*, *Entity2*). They adopt BERN to recognize and normalize the biomedical named entities in a document so that relevant documents that contain the same biomedical concepts can be retrieved. For example, if "IL1B" is given as a query, Vapur retrieves documents that not only contain "IL1B", but also "Interleukin1b", since they are normalized to the same concept (EntrezGene:3553) by BERN.

Funding

This work was supported in part by NRF-2020R1A2C3010638, NRF-2014M3C9A3063541, HR20C0021, IITP-2021-0-01819.

References

Akbik, A. *et al.* (2019). FLAIR: An easy-to-use framework for state-of-the-art NLP. In *NAACL 2019, 2019 Annual Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 54–59.

Bada, M. *et al.* (2012). Concept annotation in the craft corpus. *BMC bioinformatics*, **13**(1), 1–20.

Bairoch, A. (2018). The cellosaurus, a cell-line knowledge resource. *Journal of biomolecular techniques: JBT*.

Brown, G. R. *et al.* (2015). Gene: a gene-centered information resource at ncbi. *Nucleic acids research*.

Degtyarenko, K. *et al.* (2007). Chebi: a database and ontology for chemical entities of biological interest. *Nucleic acids research*.

Diehl, A. D. *et al.* (2016). The cell ontology 2016: enhanced content, modularization, and ontology interoperability. *Journal of biomedical semantics*.

D'Souza, J. and Ng, V. (2015). Sieve-based entity linking for the biomedical domain. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*.

Federhen, S. (2012). The ncbi taxonomy database. *Nucleic acids research*.

Hamosh, A. *et al.* (2005). Online mendelian inheritance in man (omim), a knowledgebase of human genes and genetic disorders. *Nucleic acids research*.

Kim, B. *et al.* (2019a). A corpus of plant–disease relations in the biomedical domain. *Plos one*, **14**(8), e0221582.

Kim, D. *et al.* (2019b). A neural named entity recognition and multi-type normalization tool for biomedical text mining. *IEEE Access*.

Köksal, A. *et al.* (2020). Vapur: A search engine to find related protein-compound pairs in covid-19 literature. In *Workshop on NLP for COVID-19 at EMNLP*.

Landrum, M. J. *et al.* (2016). Clinvar: public archive of interpretations of clinically relevant variants. *Nucleic acids research*.

Leaman, R. *et al.* (2013). Dnorm: disease name normalization with pairwise learning to rank. *Bioinformatics*, **29**(22), 2909–2917.

Leaman, R. *et al.* (2015). tmchem: a high performance approach for chemical named entity recognition and normalization. *Journal of cheminformatics*, **7**(1), 1–10.

Lee, J. *et al.* (2020). Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*.

Lewis, P. *et al.* (2020). Pretrained language models for biomedical and clinical tasks: Understanding and extending the state-of-the-art. In *Proceedings of the 3rd Clinical Natural Language Processing Workshop*.

Li, J. *et al.* (2016). Biocreative v cdr task corpus: a resource for chemical disease relation extraction. *Database*.

Lipscomb, C. E. (2000). Medical subject headings (mesh). *Bulletin of the Medical Library Association*.

Morgan, A. A. *et al.* (2008). Overview of biocreative ii gene normalization. *Genome biology*.

Pyysalo, S. *et al.* (2013). Overview of the cancer genetics (cg) task of bionlp shared task 2013. In *Proceedings of the BioNLP Shared Task 2013 Workshop*, pages 58–66.

Sherry, S. T. *et al.* (2001). dbsnp: the ncbi database of genetic variation. *Nucleic acids research*.

Sung, M. *et al.* (2020). Biomedical entity representations with synonym marginalization. *ACL*.

Wang, X. *et al.* (2019). Cross-type biomedical named entity recognition with deep multi-task learning. *Bioinformatics*.

Weber, L. *et al.* (2021). Hunflair: an easy-to-use tool for state-of-the-art biomedical named entity recognition. *Bioinformatics*.

Wei, C.-H. *et al.* (2015). Gnormplus: an integrative approach for tagging genes, gene families, and protein domains. *BioMed research international*.

Wei, C.-H. *et al.* (2018). tmvar 2.0: integrating genomic variant information from literature with dbsnp and clinvar for precision medicine. *Bioinformatics*.

Xu, J. *et al.* (2020). Building a pubmed knowledge graph. *Scientific data*.