Supplementary Information - nach0: Multimodal Natural and Chemical Languages Foundation Model

Micha Livne[1†], Zulfat Miftahutdinov[2†], Elena Tutubalina[2†], Maksim Kuznetsov[3†], Daniil Polykovskiy[3], Annika Brundyn[1], Aastha Jhunjhunwala[1], Anthony Costa[1], Alex Aliper[4], Alán Aspuru-Guzik[5*], Alex Zhavoronkov[2*]

[1]NVIDIA, 2788 San Tomas Expressway, Santa Clara, 95051, CA, US.
[2]Insilico Medicine Hong Kong Ltd., Unit 310, 3/F, Building 8W, Phase 2, Hong Kong Science Park, Pak Shek Kok, New Territories, Hong Kong.
[3]Insilico Medicine Canada Inc., 3710-1250 René-Lévesque west, Montreal, Quebec, Canada.
[4]Insilico Medicine AI Ltd., Level 6, Unit 08, Block A, IRENA HQ Building, Masdar City, Abu Dhabi, United Arab Emirates.
[5]University of Toronto, Lash Miller Building 80 St. George Street, Toronto, Ontario, Canada.

*Corresponding author(s). E-mail(s): alan@aspuru.com; alex@insilicomedicine.com;
†These authors contributed equally to this work.

1 NLP Ablation

To examine the impact of cross-domain data on multi-task fine-tuning, we conducted training on mono-domain data. The results of four pre-trained checkpoints fine-tuned exclusively on NLP data are presented in Supplementary Information, Tab. 1. Several noteworthy observations can be made based on these findings.

Firstly, when considering average performance, nach0, SciFive, and FLAN exhibit similar results. However, each model demonstrates superior performance on different tasks. FLAN, being a general-domain model, outperforms others in textual entailment, binary QA, and sentence similarity. On the other hand, the domain-specific SciFive

shows best results in NER, while nach0 – in relation extraction, classification, and multi-choice QA.

Secondly, MolT5 achieves lower scores compared to the other models. This can be related to the pre-training strategy, where molecules and natural language texts share the same tokens in the semantic space. In contrast, nach0 utilizes specialized tokenization for molecular data, which does not significantly impact overall performance on NLP tasks compared to SciFive and FLAN.

Table 1 Performance of nach0 on NLP tasks in comparison with FLAN, SciFive, MolT5. We list the scores for each task (see Sec. 2 about datasets and metrics). All models are base models.

	nach0	FLAN-T5	SciFive	MolT5
Named Entity Recognition	80.63%	75.01%	**81.14%**	56.48%
BC5-chem	91.14%	87.56%	91.81%	64.28%
BC5-disease	81.72%	76.61%	82.33%	61.56%
NCBI-disease	84.43%	79.46%	85.33%	54.74%
BC2GM	72.44%	61.75%	72.76%	45.87%
JNLPBA	73.42%	69.68%	73.45%	55.93%
PICO extraction	67.10%	**68.94%**	67.62%	66.39%
EBM PICO	67.10%	68.94%	67.62%	66.39%
Textual Entailment	86.03%	**87.53%**	86.96%	55.63%
MedNLI	81.28%	81.75%	82.90%	55.67%
SciTail	90.77%	93.31%	91.01%	55.58%
Relation Extraction	**84.06%**	73.84%	73.22%	63.38%
ChemProt	89.40%	84.48%	82.77%	75.98%
DDI	89.67%	72.85%	66.08%	63.23%
GAD	73.11%	64.19%	70.82%	50.93%
Sentence similarity	27.45%	**32.78%**	1.17%	14.95%
BIOSSES	27.45%	32.78%	1.17%	14.95%
Document Classification	**83.83%**	75.48%	82.49%	70.99%
HoC	83.83%	75.48%	82.49%	70.99%
Question answering (Yes/No)	63.87%	**65.04%**	63.66%	51.6%
PubMedQA	51.32%	50.36%	52.04%	47.20%
BioASQ	76.43%	79.71%	75.29%	56.00%
Question answering (Multi Choice)	**27.71%**	25.61%	26.29%	25.54%
MedMCQA and MMLU	27.71%	25.61%	26.29%	25.54%
Question answering (Open)	**2.43%**	2.34%	2.25%	1.83%
MedMCQA-Open	2.43%	2.34%	2.25%	1.83%

2 Datasets

2.1 Chemistry

We've integrated several chemical domain tasks from widely-used benchmarks and datasets. It covers distribution match, molecular property prediction, reaction prediction and related problems. Where it's possible, we use the provided standard train/validation/test split procedures, otherwise, we employ the random data split. We choose this data preparation strategy to enable comparison with baseline models, however, we don't guarantee that one can't find chemical objects with similar structures in the different subsets.

2.1.1 MOSES

MOSES dataset [1] is a benchmarking platform that provides a large dataset and set of metrics to compare generative models on an unconditional molecular generation task. The dataset provided by MOSES contains almost 2 million samples filtered by MCF, PAINS, and additional rules. The metrics set estimates the quality of the generative model from several points of view: validity of generated structures, molecular distribution matching quality, and the ability of the model to produce novel, diverse molecules.

Technical Details

Evaluation metric: The MOSES benchmark provides established set metrics for assessing the ability of models to produce unique, diverse, valid molecules similar to ground-truth distribution. In our work, we adopt several metrics: uniqueness, validity, novelty, internal diversity, similarity to a nearestneighbor (SNN), fragment similarity, scaffold similarity and FCD [2]. We've generated 30000 new molecules to compute these metrics.

Example on molecular distribution matching:

input text with prompt: Generate random molecule from MOSES dataset.

output text: `CC1C2CCC(C2)C1CN(CCO)C(=O)c1ccc(Cl)cc1`.

2.1.2 Mol-Instructions

The recently published Mol-Instructions dataset [3] covers three significant modalities: molecule-oriented instructions, protein-oriented instructions, and biomolecular text instructions. In our study, we specifically focus on the first subset, which is the most relevant and contains chemical tasks.

Example on descriptor-guided molecule generation:

input text with prompt: Synthesize a molecule that matches the given characteristics: The molecule is the (R)-enantiomer of aceprometazine. It is an enantiomer of a (S)-aceprometazine.

output text: `CC(=O)c1cc2c(cc1)Sc1ccccc1N2C[C@@H](C)N(C)C`.

Example on molecular description generation:

input text with prompt: What can you tell me about this molecule?: `C(=C(Cl)Cl)(Cl)Cl`.

output text: The molecule is a manufactured chemical that is widely used for dry cleaning of fabrics and for metal-degreasing. It is also used to make other chemicals and is used in some consumer products.

Example on forward reaction prediction:

input text with prompt: With the provided reactants and reagents, propose a potential product.: `CO.C[Si](C)(C)C#Cc1ccc(C=O)cc1.ClCCl.O=C([O-])[O-].[K+].[K+]`.

output text: `C#Cc1ccc(C=O)cc1`.

Example on reagent prediction:

input text with prompt: Please propose potential reagents that might have been utilized in the provided chemical reaction:

COC(=O)c1c(F)cc(NC(=O)c2cc(C(C)C)c(C(C)C)s2)cc1F ≫
CC(C)c1c(C(C)C)sc(C(=O)Nc2cc(F)c(C(=O)O)c(F)c2)c1.
output text: [Na+].[OH-]

Example on retrosynthesis:

input text with prompt: Provide a list of potential reactants that may have produced the given product.: Cc1ccc(-c2ccccc2N)cc1

output text: Cc1ccc(B(O)O)cc1.Nc1ccccc1I

2.1.3 Property Prediction

We adopt several binary classification and regression tasks from the MoleculeNet benchmark to assess the model's ability to predict molecular properties.

Technical Details

Evaluation metric: Binary classification tasks include BBBP, HIV, and BACE datasets from MoleculeNet [4] and use balanced accuracy as the main metric. Regression tasks involve ESOL, FreeSOLV and Lipo datasets from MoleculeNet [4], QM9 dataset from MolInstructions [3] and rely on the R2 metric. In our work, we utilized the code provided by the MoleculeNet benchmark to prepare data splits.

Example on the BBBP classification task:

input text with prompt: Can CN(C)[C@H]1[C@@H]2C[C@H]3C(=C(O)c4c(O)
cccc4[C@@]3(C)O)C(=O)[C@]2(O)C(=O)
C(=C(/O)NCN5CCCC5)C1=O penetrate the BBB?

output text: 1

Example on HIV classification task:

input text: Is CCC1=[O+][Cu-3]2([O+]=C(CC)C1)[O+]=C(CC)CC(CC)=[O+]2 an HIV inhibitor?

output text: 0

Example on BACE classification task:

input text with prompt: Please evaluate the ability of S(=O)(=O)(CCCCC)C[C@@H](NC(=O)c1cccnc1)C(=O)N[C@H]([C@H](O)C[NH2+]
Cc1cc(ccc1)CC)Cc1cc(F)cc(F)c1 to inhibit human beta-secretase

output text: 1

Example on logS prediction task:

input text with prompt: Given molecule with SMILES OCC2OC(Oc1ccccc1CO)C(O)C(O)C2O, predict its logS

output text: 1.083897

Example on HFE prediction task:

input text with prompt: What hydration free energy does COc1cc(c(c(c1O)OC)Cl)C=O have?

output text: -1.013714

Example on logD prediction task:

input text with prompt: What is the lowest unoccupied molecular orbital (LUMO) energy of this molecule? : O=C1OC2C3CC1OC32

output text: 0.0035

Example on HOMO-LUMO prediction task:

input text with prompt: lipophilic is `COc1cc(OC)c(cc1NC(=O)CCC(=O)O)S(=O)(=O)NCc2ccccc2N3CCCCC3`?

output text: -0.720000

2.2 NLP

2.2.1 Named entity recognition

Named entity recognition (NER) is a fundamental aspect of natural language processing, involving the identification and classification of entities in a given text into predefined categories. In biomedical NER, the focus lies in extracting mentions of diseases, genes, chemicals, and other biologically relevant entity types. To conduct this study, we carefully selected five datasets:

- BC2GM [5];
- BC5CDR-Disease [6];
- BC5CDR-Chemical [6];
- JNLPBA [7];
- NCBI-Disease [8].

BC2GM

The BC2GM dataset encompasses an extensive collection of over 20,000 sentences extracted from the MEDLINE database, spanning the years 1991 to 2003. Each document in this dataset is annotated with gene mention spans, amounting to a total of 24,583 mentions.

BC5CDR

The BioCreative V CDR dataset was specifically designed for named entity recognition tasks involving disease and chemical entity types. It contains 12,850 disease and 15,935 chemical mentions, drawn from 1,500 PubMed articles.

JNLPBA

The JNLPBA involves gene mention annotations across more than 2,000 PubMed abstracts. The creation of this dataset entailed a meticulous search on the MEDLINE database, using specific MeSH terms such as 'human', 'blood cells', and 'transcription factors'. In total, JNLPBA comprises 59,963 gene mention spans.

NCBI-Disease

The NCBI-disease corpus, developed by the National Center for Biotechnology Information (NCBI), constitutes a vast collection of 793 PubMed abstracts that have undergone meticulous annotation by domain experts. These annotations include disease names and their corresponding concept IDs, sourced from the Medical Subject Headings (MeSH) vocabulary [9].

Technical Details

In order to train the neural network in a text-to-text format, we designed five prompts. Each prompt asks to highlight the spans corresponding to mentions of specific entity. In order to achieve this, we insert specific tokens before and after the mention of an entity in the text.

Evaluation metric: the evaluation of the NER task's quality is performed using the entity level F-measure.

Example:

input text with prompt: Please find all instances of diseases in the given text. Each mention should be surrounded by "diso*" and "*diso": Identification of APC2, a homologue of the adenomatous polyposis coli tumor suppressor;

output text: Identification of APC2 , a homologue of the diso* adenomatous polyposis coli tumour *diso suppressor.

2.2.2 Question Answering

Question Answering (QA) is an important area of NLP research. The objective of QA is to develop intelligent systems that can understand and accurately answer questions posed in natural language. Within the biomedical domain, QA refers to the specific applications and models designed to address questions related to biomedical and healthcare information. It is required for model to understand and respond to questions pertaining to medical knowledge, clinical data, scientific literature, drug information, and other relevant biomedical topics. In this study, we conducted experiments on four biomedical QA datasets:

- BioASQ [10];
- PubMedQA [11];
- MedMCQA [12];
- MMLU [13].

The first two datasets are employed to evaluate the neural network's ability to answer binary Yes/No questions, while the remaining two datasets are used in scenarios that involve multi-choice and open question answering.

BioASQ and PubMedQA

BioASQ (Biomedical Question Answering) is a widely recognized dataset in the biomedical domain, specifically designed for evaluating question answering systems. Following the [14] we restrict the dataset to yes/no questions. We use the official train/dev/test split where each contains 670/75/140 questions respectively.

Similar to BioASQ, the PubMedQA dataset as well presents questions with limited number of answers. In contrast to the previous dataset, the answers to the questions in PubMedQA are selected from yes, no, or maybe. We use the original train/dev/test split with 450, 50, and 500 questions, respectively.

MedMCQA and MMLU

For multiple choice question answering, we employ the concatenation of the MedMCQA and MMLU datasets from [3], resulting in a total of 12,398 multiple-choice

questions. As [3] does not provide train/dev/test partitions, we randomly split the dataset into a ratio of 75:25.

To perform open question answering, we adopted a dataset introduced in [3], which comprises 27,574 question-answer pairs. This dataset was curated from the MedMCQA dataset.

Technical Details

Evaluation metric: to evaluate the performance of yes/no and multiple-choice question-answering tasks, we utilized the accuracy metric. For open-ended question-answering tasks, we adopted the BLEU-2 metric as our evaluation criterion.

Yes/No QA example:

input text with prompt: Given a passage: De novo DNA methylation in Arabidopsis thaliana is catalyzed by the methyltransferase DRM2, a homolog of the mammalian de novo methyltransferase DNMT3. Here we describe DNA methyltransferase genes from both Arabidopsis and maize that show a high level of sequence similarity to Dnmt3, suggesting that they encode plant de novo methyltransferases. Relative to all known eukaryotic methyltransferases, these plant proteins contain a novel arrangement of the motifs required for DNA methyltransferase catalytic activity. The N termini of these methyltransferases contain a series of ubiquitin-associated (UBA) domains. BLASTX searches and phylogenetic analysis suggested that five cDNAs belonged to four classes (Dnmt1, Dnmt2, CMT and Dnmt3) of DNA methyltransferase genes,

answer the question: Are there any DNMT3 proteins present in plants?;

output text: Yes.

Multi-choice QA example:

input text with prompt: Which of the following is antifibrinolytic drug: What of the following is the right choice?

(A) Tenecteplase
(B) Heparin
(C) Urokinase
(D) Tranexaemic acid

output text: The final answer is (D).

Open ended QA example:

input text with prompt: 1,25 dihydrocholecalciferol acts on?

output text: Intranuclear receptors

2.2.3 Relation Extraction

Relation extraction (RE) is a NLP task that involves identifying and classifying the relationships between entities mentioned in a text. In the biomedical domain, RE refers to the specific application of RE techniques and models to extract and classify relationships between biomedical entities mentioned in text. Biomedical RE focuses on identifying and categorizing the associations between various biomedical entities, including genes, proteins, diseases, drugs, and other molecular entities. For experiments, we use three corpora:

- ChemProt [15];

- DDI [16];
- GAD [17].

ChemProt

The ChemProt dataset is a widely used benchmark for the task of chemical-protein RE. The dataset comprises PubMed abstracts that are annotated with chemical-protein interactions, where the chemicals typically represent drug compounds or small molecules, and the proteins denote specific biological targets or enzymes. Each annotated interaction is labeled with the corresponding chemical and protein mentions, along with the following types of relationship: upregulator, downregulator, antagonist, agonist, and substrate. The training set of the dataset contains 9,995 relation pairs, and the test set contains 5,744 relation pairs.

DDI

The DDI (Drug-Drug Interaction) corpus is a dataset designed for the purpose of identifying drug-drug interactions mentioned in biomedical texts. The corpus consists of annotated sentences or text passages that describe interactions between pairs of drugs. Each annotated interaction is labeled with the names of the drugs involved and the specific type of interaction. We employ the train/test split produced in [14], where the training set contains 4,021 relation pairs and the test set contains 979 relation pairs.

GAD

The GAD dataset is a comprehensive collection of genetic association information that was semi-automatically compiled using the Genetic Association Archive. In our study, we utilize an existing preprocessed version of GAD and its corresponding train/test split, which was created by Lee et al. [18]. The training set of the dataset consists of 4,796 relation pairs, while the testing set includes 534 relation pairs.

Technical details

In our experimental framework, we adopt a binary classification approach for relation extraction. Here, the positive class indicates the presence of the specified type of relationship between two entities.

Evaluation metric: to evaluate the quality of RE tasks we utilize the F-1 measure of positive class.

Example:

input text with prompt: does the Chlorprothixene and lithium are said to have mechanism type of interaction in the following passage:

Chlorprothixene may increase the plasma-level of concomitantly given lithium. In order to avoid lithium intoxication, lithium plasma levels should be monitored closely. If chlorprothixene is given concomitantly with opioids, the opioid dose should be reduced (by approx. 50%), because chlorprothixene amplifies the therapeutic actions and side-effects of opioids massively. Avoid the concomitant use of chlorprothixene and tramadol (Ultram). Massive seizures may be encountered with this combination.

Consider additive sedative effects and confusional states to emerge, if chlorprothixene is given with benzodiazepines or barbituates. Choose particular low doses of these drugs. Exert particular caution in combining chlorprothixene with other anticholinergic drugs (tricyclic antidepressants and antiparkinsonian agents): Particularly the elderly may develop delirium, high fever, severe obstipation, even ileus and glaucoma.

output text: Yes

2.2.4 Textual Entailment

Textual entailment (TE) is a natural language processing task that involves determining the logical relationship between two pieces of text: a text fragment known as the "premise" and another text fragment known as the "hypothesis." The task is to decide whether the meaning of the hypothesis can be logically inferred or entailed from the meaning of the premise. For conducting our experiments, we utilize the following corpora:

- MedNLI [19];
- SciTail [20];

MedNLI

MedNLI (Medical Natural Language Inference) is a specialized dataset designed to facilitate research in natural language inference within the medical and healthcare domain. It consists of pairs of sentences, where each pair comprises a premise and a hypothesis. The premise represents a clinical or biomedical context, while the hypothesis is a medical statement or claim that may or may not logically follow from the premise. Each sentence pair is annotated with one of three labels: "entailment," indicating that the hypothesis can be logically inferred from the premise; "contradiction," suggesting that the hypothesis contradicts the information in the premise; and "neutral," signifying that there is no logical relationship between the two sentences. The dataset comprises a total of 12,627 sentence pairs in the training set and 1,422 sentence pairs in the testing set.

SciTail

The SciTail dataset is similar to the MedNLI dataset was designed for the task of natural language inference. Except that it covers a broader scientific domain. The train part of the corpora contains 24900 sentence pairs and the test part of the corpora contains 2126.

Technical Details

Evaluation metric: to evaluate the quality of TE tasks we utilize the Accuracy score.
Example:

input text with prompt: Given that "At [**Hospital 1456**] Hospital the patient was experiencing 10 out of 10 chest pain and received nitropaste two inches, three sublingual nitroglycerins, morphine 4 mg intravenously, Lopressor 5 mg intravenously." Does it follow that " The patient is asymptomatic."

yes or no?

output text: No

2.2.5 Sentence similarity

Textual similarity tasks in the biomedical domain involve assessing the degree of semantic similarity or relatedness between pairs of biomedical texts. The goal of these tasks is to determine how closely two pieces of text, such as sentences or documents, are semantically or conceptually aligned. To conduct our experiments, we employ the BIOSSES dataset [21].

BIOSSES

The BIOSSES (Biomedical Sentence Similarity Benchmark) dataset is a specialized dataset designed to evaluate sentence similarity models in the biomedical domain. It contains pairs of biomedical sentences that are carefully selected to represent different levels of semantic similarity. Each sentence pair is annotated with a similarity score that represents the degree of semantic relatedness between the two sentences. The scores are typically on a continuous scale, indicating how similar or dissimilar the sentences are in meaning. The dataset comprises a total of 80 sentence pairs in the training set and 20 sentence pairs in the testing set.

Technical Details

Evaluation metric: to evaluate the quality of Textual Similarity tasks we utilize the Pearson corellation score.
Example:

input text with prompt: Please assess the similarity between these two sentences on a scale of 0.0 (lowest) to 4.0 (highest). First sentence: "It has recently been shown that Craf is essential for Kras G12D-induced NSCLC." Second sentence:"It has recently become evident that Craf is essential for the onset of Kras-driven non-small cell lung cancer. "

output text: 4.0

2.2.6 Document Classification

In the biomedical domain, the document classification task involves categorizing entire documents, such as scientific articles, research papers, or clinical reports, into predefined categories or classes. The goal is to automatically assign each document to the most relevant category based on its content and subject matter. For our experimental purposes, we utilize the Hallmarks of Cancer dataset.

Hallmarks of Cancer

The Hallmarks of Cancer (HoC) dataset serves as a document classification task, centered around the concept of cancer hallmarks as established in the referenced work [22]. This corpus comprises PubMed abstracts, each labeled with binary annotations, denoting the presence of specific discussions related to individual cancer hallmarks. We utilize the train/test split from [14] which comprises 13917 sentences in train part and 3547 sentences in test part.

Technical Details

Evaluation metric: to evaluate the quality of Document Classification tasks we utilize the F-1 score.

Example:

input text with prompt: Pick one category for the following text. The options are - activating invasion and metastasis, avoiding immune destruction, cellular energetics, enabling replicative immortality, evading growth suppressors, genomic instability and mutation, inducing angiogenesis, resisting cell death, none, sustaining proliferative signaling, tumor promoting inflammation.

Biopsy of a skin lesion showed lymphoproliferative infiltration of the dermis with a follicular and angiocentric growth pattern and regional epidermal necrosis.

output text: resisting cell death

2.2.7 PICO extraction

PICO extraction is an essential NLP task that aims to automatically identify and extract specific fragments of text pertaining to the Patient (P), Intervention (I), Comparator (C), and Outcome (O) elements from unstructured biomedical texts, such as research articles and clinical trial reports. Typically, Comparator labels are omitted from the annotations, as they conform to established clinical trial norms, with "placebo" as the passive control and "standard of care" as the active control. To conduct our study, we leveraged the EBM PICO [23] dataset for this purpose.

EBM PICO

The EBM PICO dataset was specifically created to facilitate PICO extraction tasks. It employs token-level labeling, where each token is categorized into one of the PIO classes (Patient, Intervention, Outcome). The dataset comprises a total of 4,800 labeled abstracts for training purposes and 200 labeled abstracts for testing purposes.

Technical Details

To conduct the PICO extraction task in a text-to-text format, we adopted the same prompt style as used for the Named Entity Recognition (NER) dataset.

Evaluation metric: to evaluate the quality of PICO extraction tasks we utilize the word-level F-1 score.

Example:

input text with prompt: Please find all instances of Interventions in the given text. Each mention should be surrounded by "Intervention*" and "*Intervention": Study protocol : Rehabilitation including Social and Physical activity and Education in Children and Teenagers with Cancer (RESPECT)

output text: Study protocol : Intervention* Rehabilitation including Social and Physical activity and Education *Intervention in Children and Teenagers with Cancer (RESPECT) .

References

[1] Polykovskiy, D., Zhebrak, A., Sanchez-Lengeling, B., Golovanov, S., Tatanov,

O., Belyaev, S., Kurbanov, R., Artamonov, A., Aladinskiy, V., Veselov, M., *et al.*: Molecular sets (moses): a benchmarking platform for molecular generation models. Frontiers in pharmacology **11**, 565644 (2020)

[2] Preuer, K., Renz, P., Unterthiner, T., Hochreiter, S., Klambauer, G.: Fréchet chemnet distance: A metric for generative models for molecules in drug discovery. Journal of Chemical Information and Modeling **58**(9), 1736–1741 (2018) https://doi.org/10.1021/acs.jcim.8b00234 https://doi.org/10.1021/acs.jcim.8b00234. PMID: 30118593

[3] Fang, Y., Liang, X., Zhang, N., Liu, K., Huang, R., Chen, Z., Fan, X., Chen, H.: Mol-instructions: A large-scale biomolecular instruction dataset for large language models. arXiv preprint arXiv:2306.08018 (2023)

[4] Wu, Z., Ramsundar, B., Feinberg, E.N., Gomes, J., Geniesse, C., Pappu, A.S., Leswing, K., Pande, V.: Moleculenet: a benchmark for molecular machine learning. Chemical science **9**(2), 513–530 (2018)

[5] Smith, L., Tanabe, L., Ando, R., et al.: The biocreative ii-critical assessment for information extraction in biology challenge. DOI: https://doi. org/10.1186/gb-2008-9-s2-s2 (2008)

[6] Li, J., Sun, Y., Johnson, R.J., Sciaky, D., Wei, C.-H., Leaman, R., Davis, A.P., Mattingly, C.J., Wiegers, T.C., Lu, Z.: BioCreative V CDR task corpus: a resource for chemical disease relation extraction. Database **2016**, 068 (2016) https://doi.org/10.1093/database/baw068 https://academic.oup.com/database/article-pdf/doi/10.1093/database/baw068/8224483/baw068.pdf

[7] Collier, N., Ohta, T., Tsuruoka, Y., Tateisi, Y., Kim, J.-D.: Introduction to the bio-entity recognition task at JNLPBA. In: Collier, N., Ruch, P., Nazarenko, A. (eds.) Proceedings of the International Joint Workshop on Natural Language Processing in Biomedicine and Its Applications (NLPBA/BioNLP), pp. 73–78. COLING, Geneva, Switzerland (2004). https://aclanthology.org/W04-1213

[8] Doğan, R.I., Leaman, R., Lu, Z.: Ncbi disease corpus: a resource for disease name recognition and concept normalization. Journal of biomedical informatics **47**, 1–10 (2014)

[9] Lipscomb, C.E.: Medical subject headings (mesh). Bulletin of the Medical Library Association **88**(3), 265 (2000)

[10] Nentidis, A., Bougiatiotis, K., Krithara, A., Paliouras, G.: Results of the seventh edition of the bioasq challenge. In: Machine Learning and Knowledge Discovery in Databases: International Workshops of ECML PKDD 2019, Würzburg, Germany, September 16–20, 2019, Proceedings, Part II, pp. 553–568 (2020). Springer

[11] Jin, Q., Dhingra, B., Liu, Z., Cohen, W.W., Lu, X.: Pubmedqa: A dataset for

biomedical research question answering. arXiv preprint arXiv:1909.06146 (2019)

[12] Pal, A., Umapathi, L.K., Sankarasubbu, M.: Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering. In: Conference on Health, Inference, and Learning, pp. 248–260 (2022). PMLR

[13] Hendrycks, D., Burns, C., Basart, S., Zou, A., Mazeika, M., Song, D., Steinhardt, J.: Measuring massive multitask language understanding. arXiv e-prints, 2009 (2020)

[14] Gu, Y., Tinn, R., Cheng, H., Lucas, M., Usuyama, N., Liu, X., Naumann, T., Gao, J., Poon, H.: Domain-specific language model pretraining for biomedical natural language processing. ACM Transactions on Computing for Healthcare (HEALTH) **3**(1), 1–23 (2021)

[15] Krallinger, M., Rabal, O., Akhondi, S.A., Pérez, M.P., Santamaría, J., Rodríguez, G.P., Tsatsaronis, G., Intxaurrondo, A., López, J.A., Nandal, U., *et al.*: Overview of the biocreative vi chemical-protein interaction track. In: Proceedings of the Sixth BioCreative Challenge Evaluation Workshop, vol. 1, pp. 141–146 (2017)

[16] Herrero-Zazo, M., Segura-Bedmar, I., Martínez, P., Declerck, T.: The ddi corpus: An annotated corpus with pharmacological substances and drug–drug interactions. Journal of biomedical informatics **46**(5), 914–920 (2013)

[17] Bravo, À., Piñero, J., Queralt-Rosinach, N., Rautschka, M., Furlong, L.I.: Extraction of relations between genes and diseases from text and large-scale data analysis: implications for translational research. BMC bioinformatics **16**, 1–17 (2015)

[18] Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C.H., Kang, J.: Biobert: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics **36**(4), 1234–1240 (2020)

[19] Shivade, C., *et al.*: Mednli-a natural language inference dataset for the clinical domain. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium. Association for Computational Linguistics, pp. 1586–1596 (2019)

[20] Khot, T., Sabharwal, A., Clark, P.: Scitail: A textual entailment dataset from science question answering. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 32 (2018)

[21] Soğancıoğlu, G., Öztürk, H., Özgür, A.: Biosses: a semantic sentence similarity estimation system for the biomedical domain. Bioinformatics **33**(14), 49–58 (2017)

[22] Hanahan, D., Weinberg, R.A.: The hallmarks of cancer. cell **100**(1), 57–70 (2000)

[23] Nye, B., Li, J.J., Patel, R., Yang, Y., Marshall, I.J., Nenkova, A., Wallace, B.C.: A corpus with multi-level annotations of patients, interventions and outcomes to support language processing for medical literature. In: Proceedings of the Conference. Association for Computational Linguistics. Meeting, vol. 2018, p. 197 (2018). NIH Public Access