# BioCAT5: Concept Mention-Aware Text-to-Text Transfer Transformer for Biomedical Literature

**Mohammad Golam Sohrab**
Artificial Intelligence Research Center (AIRC),
National Institute of Advanced Industrial Science and Technology (AIST),
Tokyo 135-0064, Japan
sohrab.mohammad@aist.go.jp

## Abstract

Integrating biomedical knowledge information like concept name a.k.a concept mention during transfer learning is not a common fashion as it needs an alignment of large-scale concepts with unstructured biomedical text during pretraining. Existing learning objectives on biomedical text mainly rely on masked language modeling or random span denoising where grafting knowledge like integrating medical concepts has never been explored nor been discussed to facilitate the model pretraining. A prevailing assumption is that such concept-aware pretraining on biomedical text can be effective in biomedical task-specific language models. Based on this assumption, we introduce BioCAT5, a biomedical domain-specific text-to-text transfer transformer that integrates concept related knowledge from the unified medical language system (UMLS) during pretraining. BioCAT5 leverages concept-aware denoising – an out-of-the-box language model learning objective that is trained to reconstruct text by masking out the concept spans of the input sequence. To facilitate the BioCAT5 pretraining, large-scale concept mentions from UMLS are considered to align with PubMed abstract text to create the pretraining dataset. A publicly available comprehensive biomedical natural language processing (NLP) benchmark datasets including named entity recognition (NER), relation extraction, question answering are evaluated. Besides, the general language understanding evaluation (GLUE) benchmark datasets are considered to show the adaptability of BioCAT5 in general domain tasks. We conduct a series of experiments with ablation studies to validate our proposed approach. The experiment results show that the alignment of concepts during pretraining is potentially significant for biomedical NLP tasks, especially on the NER datasets. Our code is publicly released on GitHub (Available at: https://github.com/aistairc/biocat5) under the Apache 2.0 License.

## 1 Introduction

Pretrained Language Models (PLMs) such as BERT (Devlin et al., 2019) and its different variant including BioBERT (Lee et al., 2019), SciBERT (Beltagy et al., 2019), ClinicalBERT (Alsentzer et al., 2019), BlueBERT (Peng et al., 2019), PubMedBERT (Gu et al., 2021) in biomedical literature have shown remarkable performance in many natural language processing (NLP) tasks. These are encoder only models and are not an unified transfer learning method that limits the model not to design for text generation tasks such as summarization, translation, or question answering. However, following the success of encoder-only PLMs in several biomedical domain-specific tasks – encoder-decoder-based sequence-to-sequence models like BioBART (Yuan et al., 2022), SciFive (Phan et al., 2021a) draw a significant impact in biomedical corpora including PubMed [1], PubMED Central (PMC [2]) etc.

To overcome the limitation of encoder-only PLMs, text-to-text transfer transformer (T5) (Raffel et al., 2020) overcomes the BERT-style limitation by outputting a sequence of text for each input. In an unified transfer learning like T5, a model is first pretrained on an unsupervised data task with a language model objective before fine-tuning it on a task-specific datasets. During the pretraining with PLMs where different objective functions including masked language model (MLM), random span masking, and de-shuffling are common practice for text denoising technique. In MLM, predicting entity or concept mention-aware span related information during transfer learning is not a common fashion as it needs to align by detecting a large number of knowledge base concept mentions a.k.a entity mention with a large-scale unstructured dataset. Such large-scale annotated dataset is not available

[1] https://pubmed.ncbi.nlm.nih.gov/
[2] https://pmc.ncbi.nlm.nih.gov/



*Findings of the Association for Computational Linguistics: IJCNLP-AACL 2026*, pages 1425–1439
November 6-10, 2026 ©2026 Association for Computational Linguistics

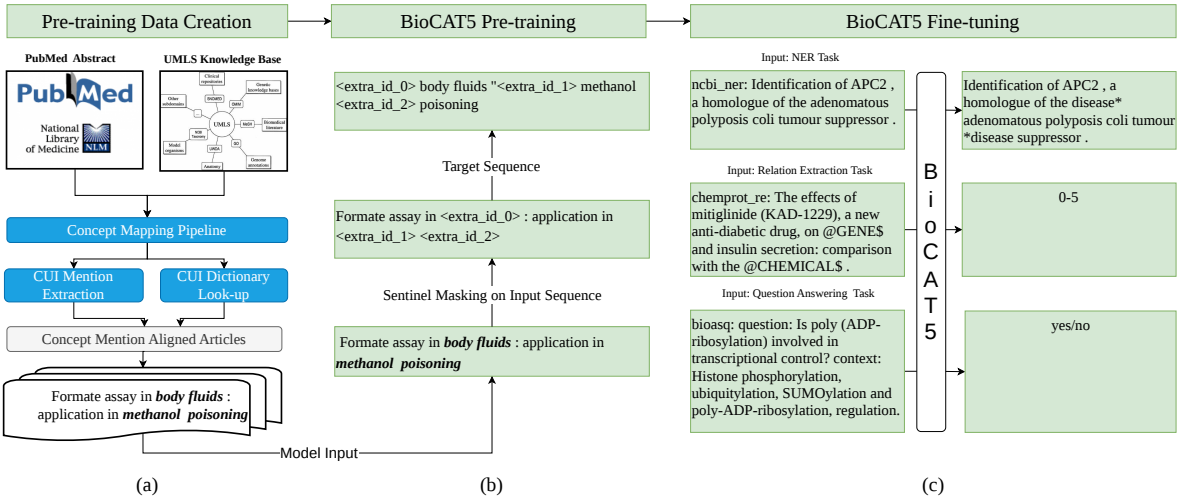



Figure 1: BioCAT5 Transfer Learning Architecture: (a) Pretraining Data Creation (b) BioCAT5 pretraining with concept-mention aligned dataset (c) Fine-tuning with BioCAT5 over the downstream tasks

for transfer learning, therefore, it is common to pretraining the entire model using data-rich unsupervised learning on unlabeled data. To give shed on this gap, we propose a novel domain-specific **Bio**medical **C**oncept mention-**A**ware **T**ext-**t**o-**T**ext **T**ransfer **T**ransformer, that packs more information into the T5 model (Raffel et al., 2020) in biomedical domain, which we call BioCAT5.

BioCAT5 leverages concept mention-aware span masking pretraining by leveraging the concept mentions from the unified medical language system (UMLS) that facilitates over a million of biomedical concept unique identifiers (CUI). Among all CUI in the unique concept name or string in each language in the UMLS, Metathesaurus has a unique identifier known as CUI. Concept spans are sometimes referred to as entities mentioned in the biomedical domain.

During transfer learning a model is first pretrained on a large-scale unsupervised data task and the most successful approaches have been variants of masked language models (MLMs), which are denoising autoencoders that are trained to reconstruct text by masking out a random subset of the input sequence. Our baseline model, T5 investigates different objective tasks, including masked language model (MLM), random span, and deshuffling, where the model is limited to exploring integrating knowledge like concept-aware denoising during pretraining. Here, we push the limits of this model by grafting UMLS's concept name information by mapping PubMed and UMLS's metathesaurus data during pretraining.

Our implementation of the BioCAT5 model is based on Huggingface transformers. To the best of our knowledge, this is the first model to handle such a large-scale CUI's name alignment during pretraining. We aligned the PubMed abstract and UMLS source vocabularies concept name to pretrain the BioCAT5 model to shed light on this challenging task. We are interested in measuring whether this concept-aware dataset can enrich the downstream tasks by leveraging the knowledge during pretraining.

Our contributions are summarized as follows:

- For BioCAT5 pretraining data creation, we perform a large-scale concept's or CUI's name mapping that grafts medical knowledge during model pretraining.
- We introduce encoder-decoder based BioCAT5-base and BioCAT5-3B models that leverage concept-aware language model text denoising objectives by replacing the traditional random span masking.
- Publicly available datasets are used and compared with the state-of-the-art (SOTA) biomedical models where the BioCAT5 outperformed four out of five NER datasets.

## 2 Related Work

T5 (Raffel et al., 2020) - the basic idea underlying this work is to treat every text processing problem as a "text-to-text" problem, i.e. taking text as input and producing new text as output. The

model achieves SOTA results on many benchmarks covering summarization, question answering, text classification, and more.

Pretrained language models such as BERT (Devlin et al., 2019), BioBERT (Lee et al., 2019), T5 (Raffel et al., 2020), ClinicalBERT (Alsentzer et al., 2019), SciBERT (Beltagy et al., 2019), PubMedBERT (Gu et al., 2021) etc. are neural networks trained on large-scale datasets that can be fine-tuned on task-specific datasets that shows a remarkable success in a wide range of NLP tasks.

SciFive (Phan et al., 2021a), a biomedical domain-specific text-to-text model that achieves SOTA results in many downstream tasks including NER, relation extraction (RE), question answering (QA) etc. The findings based on SciFive support that text-to-text (text generation) models are highly versatile and broadly applicable within domain-specific contexts. These models can be used for common tasks which require a longer sequence of text as an output. BioBERT (Lee et al., 2019) is a domain-specific language representation model pretrained on large-scale biomedical corpora. BioBERT has outperformed previous models on biomedical NER, RE, and QA tasks.

PubMedBERT (Gu et al., 2021) This paper challenge a prevailing assumption in pretraining and show that domain-specific pretraining from scratch can significantly outperform mixed-domain pretraining such as continual pretraining from a general-domain language model, leading to new SOTA results for a wide range of biomedical NLP applications. To further facilitate this study, BLURB creates a diverse set of biomedical tasks such as NER, RE, document classification, and QA.

BART-NAR-BERT (BnB) (Sohrab et al., 2024) – a pretrained non-autoregressive S2S model, which employs BERT as the backbone for the encoder and decoder for natural language understanding and generation tasks. The model outperformed several SOTA models in non-autoregressive benchmark and has shown comparable performance in autoregressive models. Since the model follows a sequence-to-sequence manner, we also compare our model over the language understanding tasks.

KAT5 (Sohrab and Miwa, 2024) – Knowledge-aware Text-to-Text Transfer Transformer that integrates knowledge like entity and coreference information during transfer learning by aligning Wikipedia abstract and Wikidata triples to facilitate the KAT5 pretraining. The experiment results show that knowledge-aware pretraining is effective for various downstream tasks. This work motivates us to integrate medical concept-aware knowledge by aligning biomedical unstructured text and UMLS during model pretraining.

## 3 BioCAT5

We consider T5 (Raffel et al., 2020) as a baseline encoder-decoder architecture close to the original architecture of transformer (Vaswani et al., 2017) to facilitate our BioCAT5 model. The entire transfer learning of BioCAT5 can be decomposed into three components: 1) Concept-aware Pretraining Dataset, 2) Model Pretraining, and 3) Model Fine-tuning. Figure 1 shows the overall transfer learning architecture of BioCAT5.

### 3.1 Concept-aware Pretraining Dataset

In the transfer learning architecture, the model first needs to be pretrained on a large amount of unsupervised data that teaches the model generalized knowledge. With a language model objective, pretraining dataset plays a significant role for model generalizable knowledge. The T5 model used the Colossal Clean Crawled Corpus (C4) dataset for pretraining by downloading about 750 GB of text extracted from the Web. In contrast, our BioCAT5 model is based on integrating knowledge like concept name information during pretraining; therefore, it is a challenging task to integrate such knowledge in a vast amount of unstructured text like the PubMed abstract dump. One possible way is to create such a concept-aware pretraining dataset by aligning PubMed abstract and UMLS source vocabularies. To create the concept-aware pretraining dataset, we align PubMed abstract with the unified medical language system (UMLS) Metathesaurus data. The UMLS Metathesaurus is a large biomedical thesaurus that is organized by concept, or meaning. It links synonymous names from over 200 different source vocabularies [3]. Among these source vocabularies we use the 21 most popular source vocabularies [4]. Based on 21 most popular knowledge-bases, we extract all candidates CUI and its term name along with its all associated aliases or synonyms (i.e., atom) to create a mapping dictionary based on concepts. We then detect

[3] https://www.nlm.nih.gov/research/umls/sourcereleasedocs/

[4] MSH, CSP, NCI, MTH, FMA, UWDA, NIH, HPO, HGNC, OMIM, RXNORM, NDFRT, NCBI, MDR, ICD9CM, ICD10CM, ICD10, CPT, GO, MEDLINEPLUS, SNOMEDCT_US

the mention based on conducting simple dictionary look-up matching for each mention in the given PubMed abstract text.

**UMLS Data Preparation** We download the full release of UMLS2020AA[5] version and extract all English entries based on the 21 most popular source vocabularies from `MRCONSO.RFF` raw file. After preprocessing, a list of 407,917 unique concepts (CUI) are extracted to create UMLS synonym dictionary where the dictionary size is 1,193,241. With this dictionary string matching approach we align a total of 362,023,532 entities. We refer to Appendix A for more details.

### 3.2 Model Pretraining

As an unsupervised objective during pretraining, a model needs an objective function that does not require labels but teaches the model generalizable knowledge and will be useful to transfer that knowledge into downstream tasks. Apart from casual language modeling objectives for pretraining, recently denoising a.k.a. MLM (Devlin et al., 2019) shows better performance and becomes a standard unsupervised learning objective in many natural language processing (NLP) tasks. In the MLM objective, the model is trained to predict missing or corrupted tokens by adding `<MASK>` in the input sequence. Inspired by BERT's MLM objective, T5 follows a random span masking objective to corrupt 15% of tokens in the input sequence where all consecutive spans of dropped-out tokens are replaced by a single sentinel token, a.k.a. unique mask tokens. We adapt the T5 masking strategies and design an objective that samples concept related spans in the input sequence using 100 sentinel tokens. For a given input sequence, each sentinel token represents a unique mask token starting as `<extra_id_0>, ..., <extra_id_99>` unlike BERT's MLM (Devlin et al., 2019) objective. In concept masking objective in BioCAT5, we ignore 15% text denoising but corrupt the text whenever concept mentions appear in the text since concepts are rare in the text.

Figure 1(b) shows a pretraining architecture where the left-side of the figure 1(a) represents the pretraining data creation by mapping PubMed abstract and UMLS source vocabularies where the italic bold texts `body fluids`, `methanol`, and `poisoning` are aligned as concept mentions to further facilitate the concept-aware masking for BioCAT5. During self-supervised learning in Figure 1(b), a text input sequence is given where the model tries to learn a target sequence using concept-aware sentinel MLM. During sentinel masking in BioCAT5, unique mask tokens are used to corrupt the input text by replacing the concept related spans. Finally, the output sequence consists of the dropped-out concept spans, delimited by the sentinel tokens used to replace them in the input.

### 3.3 Model Fine-tuning

During task specific supervised training, BioCAT5 follows the T5 fine-tuning approach where the proposed BioCAT5 treats every text processing problem as a test-to-text problem, i.e. giving text as input to the model and producing new text as output. Figure 1(c) shows a fine-tuning approach of BioCAT5. A full fine-tuning where all model parameters are updated during task-specific learning. During fine-tuning on the base and 3B parameters models, we initialize each task-specific dataset with BioCAT5-base and BioCAT5-3B models.

## 4 Experiments

Our proposed BioCAT5 is an encoder-decode-based generative model on the basis of domain-specific biomedical literature. To judge the effectiveness of our models both in generative and language understanding tasks, we employ popular benchmark biomedical datasets grouped into three different domain specific tasks including NER, QA, and RE for generative tasks. In contrast to language understanding task, we consider the General Language Understanding Evaluation (GLUE) benchmark (Wang et al., 2018), consists of nine general language understanding tasks since language understanding in biomedical literature is somewhat missing.

### 4.1 Generative Datasets

To evaluate the BioCAT5 on generative datasets, we conduct experiments on NER, QA, and RE tasks.

#### 4.1.1 Named Entity Recognition

For named entity recognition (NER), we evaluate our models on five commonly used datasets for biomedical NER including NCBI disease (Doğan et al., 2014), BC5CDR disease (Li et al., 2016), BC5CDR chemical (Li et al., 2016), BC4CHEMD (Krallinger et al., 2015)

[5] https://download.nlm.nih.gov/kss/2020AA/umls-2020AA-full.zip

| Model | Named Entity Recognition | | | | |
|---|---|---|---|---|---|
| | NCBI-D | BC5CDR-D | BC5CDR-C | BC2GM | JNLPBA |
| - Encoder Model - | | | | | |
| $BERT_{base}$ (Phan et al., 2021b) | 85.63 | 82.41 | 91.16 | 81.79 | 74.94 |
| $BioBERT_{base}$ (Phan et al., 2021b) | 89.71 | 87.15 | 93.47 | 84.72 | 77.49 |
| SciBERT (Gu et al., 2021) | 88.25 | 84.70 | 92.51 | 83.36 | 78.51 |
| ClinicalBERT (Gu et al., 2021) | 86.32 | 83.04 | 90.80 | 81.71 | 78.07 |
| BlueBERT (Gu et al., 2021) | 88.04 | 83.69 | 91.19 | 81.87 | 77.71 |
| PubMedBERT (Gu et al., 2021) | 87.82 | 85.62 | 93.33 | 84.52 | 79.10 |
| - Encoder-Decoder Model - | | | | | |
| T5 (Phan et al., 2021b) | 88.54 | 86.83 | 93.61 | 82.29 | 74.56 |
| SciFive (Phan et al., 2021b) | | | | | |
| + base + PMC + PubMed | 88.79 | 87.33 | 94.34 | 84.16 | 75.29 |
| + base + PubMed | 87.96 | 87.44 | 94.35 | 83.92 | 75.60 |
| + base + PMC | 89.39 | 87.23 | 94.18 | 84.29 | 75.89 |
| + large + PMC + PubMed | 89.11 | 87.59 | **94.76** | 83.84 | 75.99 |
| + large + PubMed | 89.17 | 86.98 | 94.66 | 83.60 | 76.08 |
| + large + PMC | 88.46 | 87.62 | 94.61 | 83.57 | 77.55 |
| - Encoder-Decoder Model - | | | | | |
| BioCAT5 + base | 87.39 | 84.71 | 89.32 | 82.37 | 74.93 |
| BioCAT5 + 3B | **90.07** | **88.52** | 93.78 | **88.95** | **88.43** |

Table 1: Performance comparison over the NER related datasets. NCBI-D/BC5CDR-D/BC5CDR-C represents NCBI-Disease/BC5CDR-Disease/BC5CDR-Chemical datasets respectively. Bold and underlined denotes the best and second-best scores within BioCAT5 and Baseline Models.

| Model | Question Answering | |
|---|---|---|
| | BioAsq | PubMedQA |
| - Encoder - | | |
| $BERT_{base}$ (Gu et al., 2021) | 74.44 | 49.96 |
| $BioBERT_{base}$ (Gu et al., 2021) | 81.14 | **60.24** |
| SciBERT (Gu et al., 2021) | 74.22 | 51.50 |
| ClinicalBERT (Gu et al., 2021) | 68.50 | 49.08 |
| BlueBERT (Gu et al., 2021) | 68.71 | 48.44 |
| PubMedBERT (Gu et al., 2021) | **87.56** | 55.84 |
| - Encoder Decoder - | | |
| BioCAT5 + base | 87.14 | 60.0 |
| BioCAT5 + 3B | 77.14 | 52.8 |

Table 2: Performance comparison over the QA datasets.

| Model | Relation Extraction | | |
|---|---|---|---|
| | CP | DDI | GAD |
| - Encoder - | | | |
| BERT-base (Gu et al., 2021) | 71.5 | 79.3 | 79.6 |
| BioBERT (Gu et al., 2021) | 76.1 | 80.9 | 82.4 |
| SciBERT (Gu et al., 2021) | 75.0 | 81.2 | 81.4 |
| ClinicalBERT (Gu et al., 2021) | 72.0 | 78.2 | 80.5 |
| BlueBERT (Gu et al., 2021) | 71.5 | 77.8 | 79.2 |
| PubMedBERT (Gu et al., 2021) | 77.2 | 82.4 | 83.9 |
| - Encoder Decoder - | | | |
| T5-base (Phan et al., 2021b) | 84.8 | 82.0 | - |
| T5-large (Phan et al., 2021b) | 85.4 | 83.0 | - |
| SciFive-base (Phan et al., 2021b) | **88.8** | 83.2 | - |
| SciFive-large (Phan et al., 2021b) | 87.9 | 83.7 | - |
| - Encoder Decoder - | | | |
| BioCAT5 + base | 74.7 | 80.2 | **85.4** |
| BioCAT5 + 3B | 77.8 | **84.4** | 82.4 |

Table 3: Performance comparison over the relation extraction datasets. '-' indicates a certain model with corresponding dataset is not evaluated.

BC2GM (Smith et al., 2008a), and JNLPBA (Collier et al., 2004). We follow the processing pipeline and the train/validation/test split similar to (Lee et al., 2019). For all NER tasks, we evaluate the performance based on F1 score. Details of NER datasets are stated in Appendix B.

#### 4.1.2 Question Answering

Question answering (QA) involves generating an answer if given a question and a sequence of text containing the answer to that question. We evaluate our models on QA tasks based on two datasets: PubMedQA (Jin et al., 2019) and BioAsq (Nentidis et al., 2020). The PubMedQA train/validation/test set consists of 450/50/500 samples respectively. In BioAsq, we use the official train/dev/test split of 670/75/140 questions by focusing on the yes/no answers based on BioAsq task 7b.

#### 4.1.3 Relation Extraction

Relation Extraction (RE) involves identifying relationships within text (i.e. gene-disease). We evaluate our models on three RE tasks: (1) the Chemical Protein interaction corpus (ChemProt) (Islamaj Doğan et al., 2019), (2) Drug-Drug Interaction corpus (DDI) (Herrero-Zazo et al., 2013), and (3) the Genetic Association Database corpus (GAD) (Bravo et al., 2015). We follow the same pre-processing technique as (Peng et al., 2019) and each class in the RE datasets are evaluated on the basis of F1-score.

| Model | MNLI 392k ACC | QQP 363k F1/ACC | QNLI 108k ACC | SST-2 67k ACC | COLA 8.5k MCC | STS-B 5.7k P.C. | MRPC 3.5k F1 | RTE 2.5k ACC | Avg. |
|---|---|---|---|---|---|---|---|---|---|
| OpenAI GPT (Devlin et al., 2019) | 82.1 | 70.3 | 87.4 | 91.3 | 45.4 | 80.0 | 82.3 | 56.0 | 75.1 |
| BERT (Devlin et al., 2019) | 84.6 | 71.2 | 90.5 | 93.5 | 52.1 | 85.8 | 88.9 | 66.4 | 79.6 |
| Optimus (Li et al., 2020) | 83.4 | 90.9 | 90.8 | 92.4 | 57.3 | 88.8 | 87.3 | 69.7 | 82.5 |
| BnB (Sohrab et al., 2024) | 82.2 | 90.3 | 89.8 | 90.7 | 54.3 | 87.9 | 88.0 | 66.8 | 81.3 |
| RoBERTa (Warner et al., 2025) | 90.2 | 92.2 | 94.7 | 96.4 | 68.0 | 92.4 | 90.9 | 86.6 | 88.9 |
| DeBERTav3 (Warner et al., 2025) | **91.8** | **93.3** | **96.0** | 96.9 | **75.3** | **93.0** | 92.2 | **92.7** | **91.4** |
| ModernBERT (Warner et al., 2025) | 90.8 | 92.7 | 95.2 | **97.1** | 71.4 | 92.8 | 91.7 | 92.1 | 90.5 |
| T5-Base (Raffel et al., 2020) | 87.1 | 89.4 | 93.7 | 95.2 | 51.1 | 89.4 | 90.7 | 80.1 | 84.6 |
| T5-Large (Raffel et al., 2020) | 89.9 | 89.9 | 94.8 | 96.3 | 61.2 | 89.9 | **92.4** | 87.2 | 86.4 |
| BioCAT5 | | | | | | | | | |
| + Run1 | 89.5 | 91.7 | 93.8 | 96.1 | 61.2 | 89.2 | 88.5 | 64.6 | 84.3 |
| + Run2 | 89.5 | 91.7 | 93.8 | 96.1 | 61.2 | 90.2 | 90.9 | 80.2 | 86.7 |

Table 4: Comparison of OpenAI GPT, BERT, Optimus, BnB, RoBERTa, DeBERTav3, ModernBERT, T5, and BioCAT5 over the GLUE Tasks. ACC and P.C stand for accuracy and Pearson correlation coefficient, respectively.

### 4.2 Language Understanding Datasets

To fine-tune the model on GLUE tasks (Wang et al., 2018), we consider applying our pretrained BioCAT5 model over the eight most popular downstream tasks by ignoring the problematic nature of the WNLI dataset in GLUE tasks.

### 4.3 Experimental Setups

**BioCAT5 Pretraining** During BioCAT5 pretraining, we initialize the BioCAT5-base and BioCAT5-3B with the Hugging Face SciFive-base and T5-3B checkpoints respectively for continual pretraining and updates the entire model parameters during concept-aware text denoising objective of our conceptually aligned dataset.

- BioCAT5-Base: We use a learning rate of 0.001, a linear warm-up of 2k steps, a gradient accumulation of 5 steps, and a maximum sequence length of 512 tokens. The BioCAT5 model is trained on 13.6B tokens, where we employ a batch size of 61,440 tokens with a maximum step of 222k steps which is around 3 times smaller than the original T5 model. The BioCAT5 model is optimized end-to-end using an Adafactor optimizer with a concept-aware span denoising learning objective.

- BioCAT5-3B: We use the same learning rate and warmup steps. BioCAT5-3B is a large language model (LLM), because of computational complexity to train such a LLM, we reduce the maximum sequence length of 256 tokens. However, the BioCAT5-3B model is trained on 1.06B tokens, where we employ a batch size of 40,960 tokens with a maximum step of 26000 steps.

We refer to Appendix C for more details about model hyper-parameters.

**BioCAT5 Fine-tuning** During fine-tuning with BioCAT5 on downstream datasets, we mainly consider single-task settings: a single model on a single dataset is learned initializing from BioCAT5 checkpoint. In this setting, we add the dataset name followed by the task separator : (for example, "ncbi_ner :" for NCBI dataset) as a prefix to each input sentence. We fine-tune on top of BioCAT5 for a maximum of 10-30 epochs based on the data size in all our downstream tasks.

### 4.4 Experimental Results

In this section, we show that our BioConcept-aware T5 (BioCAT5) can effectively solve the structure prediction tasks that match or exceed the previous SOTA on multiple datasets. All our experiments initialize from a pretrained BioCAT5 and to keep our model as simple as possible, hyper-parameters are the same across all experiments.

**Performance on NER** Table 1 shows the NER performances where we grouped the results into two folds: (1) Performance comparison of BioCAT5 with the encoder-only models including BERT, BioBERT, SciBERT, ClinicalBERT, BlueBERT, and PubMEdBERT and (2) Performance comparison over the encoder decoder-based models including $T5_{base}$, $T5_{large}$, $SciFIve_{base}$, and $SciFive_{large}$. Our BioCAT5 base model shows comparable performances among all the reported scores. In contrast, the BioCAT5 3B parameter model outperforms all the encoder-only and encoder-decoder approaches over the NCBI-Disease(D), BC5CDR-Disease(D), BC2GM, and

JNLPBA. It shows a little drop in BC5CDR-Chemical(C) in comparison to the SciFive base and large models, though it outperforms the encoder-only models.

**Performance on Question Answering** In Table 2, we compare our model based on encoder-only models as encoder-decoder models in Table 1 are not evaluated over the PubMedQA dataset. For BioAsq, as we focus on the 7b task on this dataset here the encode-decoder models are missing to evaluate this dataset. In this table, PubMedBERT and BioBERT show its superiority over the other models in BioAsq and PubMedQA datasets respectively, where the proposed BioCAT5 achieved second best score. We refer to Appendix D for more details about the input and output format of NER, question answering, relation extraction, and the pretraining data format of our models.

**Performance on Relation Extraction** We evaluate our BioCAT5 over the ChemProt (CP), DDI, and GAD datasets. In Table 3 the SciFive-large model shows its superiority over the other models including BioCAT5 in ChemProt. Our BioCAT5 models achieve mixed performances, where the BioCAT5 with base and 3B models shows its superiority over the GAD and DDI datasets respectively.

#### 4.4.1 Performance on GLUE

We compare our model against SOTA GLUE models. GLUE is not biomedical but rather it shows the adaptability of our model in general domain tasks. Table 4 shows the performance comparison of our BioCAT5 over the eight different GLUE tasks in different existing models. We employ the Optimus evaluation script[6] to evaluate the scores and select the best performance model. In this table, the SOTA models on GLUE tasks such as RoBERTa, DeBERTav3, and ModernBERT show its superiority over all the other models including BioCAT5. These models are especially trained for general-domain dataset. In this table, we report our results on two runs based on two random seeds. BioCAT5 with first run shows little drop from our baseline T5 model which is trained on general-domain dataset. In the second run, our model shows its superiority over the T5. Our GLUE results suggest that it overcomes the catastrophic forgetting that does not severely harm the general capabilities of T5 during model pretraining.

[6] https://github.com/ChunyuanLI/Optimus

Apart from SOTA models in GLUE tasks, the BioCAT5 obtains a significant improvement over the earlier contribution of different models including OpenAI GPT, BERT, Optimus, and BnB.

### 4.5 Ablation Study

We conduct ablation studies to investigate the proposed model performances in different scenarios including 1) Random VS concept-aware masking, 2) single-dataset vs multi-dataset, 3) language model selection, 4) LLM behavior, and 5) the effect of multiple runs for model reliability over the different downstream tasks.

#### 4.5.1 Random VS Concept-Aware Masking

To evaluate the effectiveness of concept-aligned masking over the random span masking we split the pretraining dataset on the basis of 10% of the entire PubMed abstract to conduct a small pretraining ablation study. The continual pretraining over the 10% dataset is based on two aspects: 1) T5-style random-span masking and 2) proposed concept-aligned masking, where we follow the same training and evaluation strategies for both aspects. The models pretraining for random span and concept-aware masking is conducted based on one single epoch.

In random span masking, we follow the same strategy like T5 by corrupting 15% random spans in the input sequence where dropped-out spans are replaced by the sentinel tokens. During continual pretraining with random span masking the model is initialized from the T5-base model. In contrast to concept-aware text denoising during pretraining, the model is initialized from the same T5-base model with same size of pretraining dataset, where masking strategies are based on the concept related spans in the given input sequence.

Table 5 shows the performance of random span and concept-aware masking. In this table, we evaluate the models based on three learning rates such as 2e-4, 3e-4, and 4e-4 over the four NER datasets including BC5CDR-Disease, BC5CDR-Chemical, BC2GM, and JNLPBA. The results show a steady gain of concept-aligned masking over the random-span masking in three different learning rates.

#### 4.5.2 Effect of Single- VS Multi-dataset

During the fine-tuning on each task-specific dataset, we add the dataset name as a prefix to each input sentence (as mentioned in Figure 1(c)), therefore, apart from single-task setting, the model can be

| Dataset | Random Span | | | Concept Span | | |
|---|---|---|---|---|---|---|
| | 2e-4 | 2e-4 | 3e-4 | 2e-4 | 3e-4 | 4e-4 |
| BC5-D | 82.8 | 83.0 | 83.0 | **83.6** | **83.2** | **83.2** |
| BC5-C | 90.1 | 89.0 | 88.7 | **90.3** | **90.1** | **90.4** |
| BC2GM | 79.4 | 78.6 | 78.5 | **79.6** | **79.2** | **79.1** |
| JNLPBA | 74.2 | 74.2 | 74.5 | **74.6** | **75.1** | **75.1** |

Table 5: Random-span VS Concept-aligned masking. BC5-D, and BC5-C denote, BC5CDR-D, and BC5CDR-C datasets respectively. Bold text indicates learning rate-wise best performance.

| Data | Named Entity Recognition | |
|---|---|---|
| | Single-Dataset | Multi-Dataset |
| - BioCAT5-base - | | |
| NCBI-Disease | 87.39 | **88.14**(+0.75) |
| BC5CDR-Disease | 84.71 | **85.23**(+0.52) |
| BC5CDR-Chemical | 89.32 | **93.71**(+4.39) |
| BC2GM | 82.37 | **83.14**(+0.77) |
| JNLPBA | 74.93 | **75.62**(+0.69) |

Table 6: Performance comparison on the single- vs multi-dataset settings over the NER datasets.

further facilitated on multi-task setting by merging several datasets into one to train a single model that can cover many task-specific datasets. To judge the effectiveness of multi-task learning in contrast to single-task setting, first, we merge all the NER datasets and fine-tune on top of the BioCAT5 base model for 30 epochs. Table 6 shows a comparison of single- vs multi-task fine-tuning where the multi-dataset results show +0.75, +0.52, +4.39, +0.77, and +0.69 improvements in the NCBI-Disease, BC5CDR-Disease, BC5CDR-Chemical, BC2GM, and JNLPBA datasets respectively in the multi-dataset setting.

#### 4.5.3 Large Language Model Selection

Our proposed BioCAT5 can be initialized from the different variations of available google 3B parameters model to facilitate the continual pretraining over the large language model (LLM). The BioCAT5 + 3B model is initialized from the publicly available `umt5-xl` [7] model which is a multilingual T5-based LLM. To judge the effectiveness of `umt5-xl` over the `t5-v1_1-xl` [8] model, we split the GLUE tasks into two scenarios: 1) MNLI, QQP, QNLI, and SST2 as a datasets that contains more than 10k training samples, and 2) COLA, STS-B, MRPC, and RTE as smaller datasets where each contain less than 10k training samples.

[7] https://huggingface.co/google/umt5-xl

[8] https://huggingface.co/google/t5-v1_1-xl

| Model | MNLI 392k ACC | QQP 363k F1/ACC | QNLI 108k ACC | SST-2 67k ACC |
|---|---|---|---|---|
| t5-v1_1-xl (3B) | 74.5 | 65.3 | 56.1 | 72.8 |
| umt5-xl (3B) | **90.2** | **89.4** | **95.3** | **95.0** |

Table 7: Performance comparison over the larger datasets from GLUE tasks.

| Model | CoLA 8.5k M.C | STS-B 5.7k PCorr | MRPC 3.5k F1 | RTE 2.5k ACC |
|---|---|---|---|---|
| t5-v1_1-xl (3B) | 0.0 | 10.4 | 81.1 | 45.8 |
| umt5-xl (3B) | **25.4** | **85.7** | **81.2** | **80.9** |

Table 8: Performance comparison over the smaller datasets from GLUE tasks.

| Dataset | UMT5-3B | | | BioCAT5-3B | | |
|---|---|---|---|---|---|---|
| | E:1 | E:2 | E:3 | E:1 | E:2 | E:3 |
| NCBI-D | 66.2 | 72.7 | 78.7 | **71.1** | **78.2** | **79.6** |
| BC5-D | 17.5 | 74.8 | 77.1 | **61.8** | **75.0** | **79.1** |
| BC5-C | 2.5 | **86.7** | 45.8 | **74.2** | 86.0 | **86.1** |
| BC2GM | 39.8 | 52.7 | 73.8 | **66.5** | **74.4** | **79.6** |
| JNLPBA | 70.1 | 75.6 | **78.1** | **75.2** | **77.3** | 76.9 |

Table 9: UMT5-3B VS BioCAT5. Here E, BC5-D, and BC5-C denote Epoch, BC5CDR-D, and BC5CDR-C respectively. Bold text indicates epoch-wise best performance.

**Effect of Larger Datasets** Table 7 shows LLMs performances over the larger datasets from GLUE tasks. The `umt5-xl` shows a significant improvement over the `t5-v1_1-xl` model.

**Effect of Smaller Datasets** Table 8 shows LLMs performances over the smaller datasets from GLUE tasks. The `umt5-xl` shows a significant improvement and even shows its performance consistency over the `t5-v1_1-xl` model in four datasets.

#### 4.5.4 Effect of LLM Behavior

To evaluate the LLM stability during the earlier stage of fine-tuning, we consider the first three epochs over the five NER datasets. In table 9, the BioCAT5-3B model shows stable and steady performances over the five datasets even at the very early learning stage of the model in comparison to the UMT5-3B parameter model.

#### 4.5.5 Effect of Multiple Runs

Mostly reported results are based on single run per downstream tasks since multiple runs based on random seeds, different learning rates, or other hyper-parameters etc. are computationally very time consuming and expensive to judge the reliability of the model. In Tables 10 and 11, we consider

| | BioCAT5-3B | | |
|---|---|---|---|
| Dataset | Run1 | Run2 | Run3 |
| BC5CDR-D | 88.5 | 88.6 | 88.9 |
| BC5CDR-C | **93.8** | **90.2** | **90.0** |
| BC2GM | 89.0 | 88.9 | 89.7 |
| JNLPBA | 88.4 | 88.7 | 87.9 |

Table 10: Performance comparison over the multiple runs of named entity recognition tasks.

| | Relation Extraction | | | |
|---|---|---|---|---|
| Model | Run | ChemProt | DDI | GAD |
| BioCAT5-base | 1 | **74.7** | 78.1 | 82.5 |
| | 2 | 73.7 | **80.2** | **85.4** |
| | 3 | 74.3 | 79.5 | 85.4 |
| BioCAT5-3B | 1 | 76.7 | **84.4** | 82.1 |
| | 2 | **77.8** | 84.0 | **82.4** |
| | 3 | 76.9 | 82.5 | 80.5 |

Table 11: Performance comparison over the multiple runs of relation extraction tasks.

three different runs based on three random seeds in NER and relation extraction tasks respectively. The steady and comparable results on both tasks based on different random seeds indicates the reliability of the BioCAT5 performances.

### 4.6 Discussion

We present BioCAT5, a novel, simple, and easy-to-implement concept-aware span masking language model S2S pretraining model that leverages millions of biomedical CUI aligning with PubMed literature to better capture concept mention related knowledge during pretraining. We demonstrate strong performance of NER tasks and stable performances in QA datasets. We evaluate BioCAT5 with the most popular models including BERT, BioBERT, SciBERT, ClinicalBERT, BlueBERT, PubmedBERT T5, and SciFive over the ten popular biomedical datasets. Experiment results show that BioCAT5 outperformed in four datasets, where the other SOTA models like SciFive, PubMedBERT, and BioBERT outperformed in three, two, and one datasets respectively.

During fine-tuning, SOTA models mostly train the models for a long period of time by setting early stopping criteria that leads to long and expensive training time. In this study, we mainly set 10-30 epochs based on data size to address all the downstream tasks to mainly avoid longer and expensive training time.

Apart from domain-specific biomedical domain, we also compare our models in the general domain datasets, especially in the most popular general language understanding task to judge the effectiveness of how our domain-specific model can further contribute in the general downstream tasks like GLUE. The results of BioCAT5 show the adaptability of our model in general-domain language understanding tasks. It shows the comparable or improved performances over the base T5 model which indicate the model still remembers the previously learned task after being trained on a new biomedical dataset. In our BioCAT5 pretraining for both base and 3B parameter models, we aligned UMLS concept with PubMed abstract only since aligning full-text articles from PubMed Central (PMC) will substantially increase billions of words that will cost a huge computational cost. The BioCAT5 shows its superiority over the NER datasets which indicates the concept mention-aware masking is aligned during the model pretraining that leads to generalization of a knowledge enriched concept-aware pretraining model. In contrast to relation extraction and QA datasets, the BioCAT5 shows comparable performance. It is expected since the target output format of QA and relation extraction datasets are based on yes/no and 0-5 numerical values respectively. Here our generative model fails to show its superiority in comparison to the different variation of BERT-based discriminative models.

## 5 Conclusion

This paper introduces an efficient BioConcept-aware T5 (BioCAT5) S2S method with encoders and decoders that integrates concepts mentioned as knowledge during pretraining. To introduce such a concept-aware approach, we perform large-scale alignments between PubMed abstract and UMLS concepts to facilitate our pretraining BioCAT5-base and BioCAT5-3B models. The BioCAT5-base and BioCAT5-3B models are trained based on our concept-aware pretraining dataset. Experiment results show that the proposed model outperforms almost all the NER related datasets where it shows comparable performances over the question answering and relation extraction datasets.

In the future, we plan to extend our BioCAT5 model into a larger parameter model with more training steps and knowledge-aware data by leveraging biomedical dataset along with general domain dataset like Wikipedia. Besides, a mixture masking objective function that includes random span, entity-aware Wikipedia dataset, and concept-aware biomedical dataset will be introduced.

## Limitations

The BioCAT5-3B model is trained only on 1.06B tokens which is comparatively very small in consideration to continual LLM pretraining. Therefore, one may further train the models for a longer period with more steps to achieve better concept-aware representations.

Hyper-parameter tuning on large language models is computationally very costly, we choose 10-30 epochs depending on the training data size for the fine-tuning tasks without conducting a wide-level of empirical studies. Besides, the model performances on the basis of different learning rates are under explored.

BioCAT5 leverages concept-aware span masking pretraining by leveraging the concept mention from the UMLS that facilitates over a million of biomedical concept unique identifiers that replaces the original T5 random span masking. A mixture of concept mention-aware span masking and random span masking would be interesting, therefore, one may further train the model by integrating random span along with concept mention-aware span masking to observe the performance.

The pretraining data is based on the UMLS2020AA version, since UMLS is updated regularly, therefore, creating pretraining data with new releases is under explored. This work is focused on transfer learning from model pretraining to fine-tuning tasks, which is a continual pretraining adaptation where scopes are always there to improve the learning model with more training data, new releases and concepts alignment with newer versions of PubMed articles. In addition, increasing source vocabularies where we restrict our search space based on 21 most popular biomedical source vocabularies among more than 200 source vocabularies. So there is always continual scope for the research community to further improve the model performances integrating with more resources and data.

## Ethics Statement

We use only publicly available datasets and relatively low compute amounts while conducting our experiments to enable reproducibility. This work is purely methodological, and all models and results are intended for research purposes only. They are not designed or intended to cause harm, enable surveillance, or mislead any individual or group. We do not perform any studies on other humans or animals in this research.

## Acknowledgments

This work has been partially supported by JSPS KAKENHI Grant Number JP24K15097. We thank all the anonymous reviewers for their thoughtful comments and constructive suggestions.

## References

Emily Alsentzer, John Murphy, William Boag, Wei-Hung Weng, Di Jindi, Tristan Naumann, and Matthew McDermott. 2019. Publicly available clinical BERT embeddings. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 72–78, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. SciBERT: A pretrained language model for scientific text. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3615–3620, Hong Kong, China. Association for Computational Linguistics.

Àlex Bravo, Janet Piñero, Núria Queralt-Rosinach, Michael Rautschka, and Laura I Furlong. 2015. Extraction of relations between genes and diseases from text and large-scale data analysis: implications for translational research. *BMC bioinformatics*, 16:55.

Nigel Collier, Tomoko Ohta, Yoshimasa Tsuruoka, Yuka Tateisi, and Jin-Dong Kim. 2004. Introduction to the bio-entity recognition task at JNLPBA. In *Proceedings of the International Joint Workshop on Natural Language Processing in Biomedicine and its Applications (NLPBA/BioNLP)*, pages 73–78, Geneva, Switzerland. COLING.

Gamal Crichton, Sampo Pyysalo, Billy Chiu, and Anna Korhonen. 2017. A neural network multi-task learning approach to biomedical named entity recognition. *BMC Bioinformatics*, 18.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Rezarta Islamaj Doğan, Robert Leaman, and Zhiyong Lu. 2014. Ncbi disease corpus: A resource for disease name recognition and concept normalization. *Journal of Biomedical Informatics*, 47:1–10.

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2021. Domain-specific language model pretraining for biomedical natural language processing. *ACM Trans. Comput. Healthcare*, 3(1).

María Herrero-Zazo, Isabel Segura-Bedmar, Paloma Martínez, and Thierry Declerck. 2013. The ddi corpus: An annotated corpus with pharmacological substances and drug–drug interactions. *Journal of Biomedical Informatics*, 46(5):914–920.

Rezarta Islamaj Doğan, Sun Kim, Andrew Chatr-aryamontri, Chih-Hsuan Wei, Donald C Comeau, Rui Antunes, Sérgio Matos, Qingyu Chen, Aparna Elangovan, Nagesh C Panyam, Karin Verspoor, Hongfang Liu, Yanshan Wang, Zhuang Liu, Berna Altınel, Zehra Melce Hüsünbeyi, Arzucan Özgür, Aris Fergadis, Chen-Kai Wang, and 8 others. 2019. Overview of the biocreative vi precision medicine track: mining protein interactions and mutations for precision medicine. *Database*, 2019:bay147.

Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William Cohen, and Xinghua Lu. 2019. PubMedQA: A dataset for biomedical research question answering. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2567–2577, Hong Kong, China. Association for Computational Linguistics.

Martin Krallinger, Obdulia Rabal, Florian Leitner, Miguel Vazquez, David Salgado, Zhiyong lu, Robert Leaman, Yanan Lu, Donghong Ji, Daniel Lowe, Roger Sayle, Riza Batista-Navarro, Rafal Rak, Torsten Huber, Tim Rocktäschel, Sérgio Matos, David Campos, Buzhou Tang, Wang Qi, and Alfonso Valencia. 2015. The chemdner corpus of chemicals and drugs and its annotation principles. *Journal of Cheminformatics*, 7:S2.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2019. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Chunyuan Li, Xiang Gao, Yuan Li, Baolin Peng, Xiujun Li, Yizhe Zhang, and Jianfeng Gao. 2020. Optimus: Organizing sentences via pre-trained modeling of a latent space. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4678–4699, Online. Association for Computational Linguistics.

Jiao Li, Yueping Sun, Robin J. Johnson, Daniela Sciaky, Chih-Hsuan Wei, Robert Leaman, Allan Peter Davis, Carolyn J. Mattingly, Thomas C. Wiegers, and Zhiyong Lu. 2016. Biocreative v cdr task corpus: a resource for chemical disease relation extraction. *Database*, 2016:baw068.

Anastasios Nentidis, Konstantinos Bougiatiotis, Anastasia Krithara, and Georgios Paliouras. 2020. Results of the seventh edition of the bioasq challenge. In *Machine Learning and Knowledge Discovery in Databases*, pages 553–568, Cham. Springer International Publishing.

Yifan Peng, Shankai Yan, and Zhiyong Lu. 2019. Transfer learning in biomedical natural language processing: An evaluation of BERT and ELMo on ten benchmarking datasets. In *Proceedings of the 18th BioNLP Workshop and Shared Task*, pages 58–65, Florence, Italy. Association for Computational Linguistics.

Long Phan, James T. Anibal, Hieu Trung Tran, Shaurya Chanana, Erol Bahadroglu, Alec Peltekian, and Grégoire Altan-Bonnet. 2021a. Scifive: a text-to-text transformer model for biomedical literature. *ArXiv*, abs/2106.03598.

Long N. Phan, James T. Anibal, Hieu Tran, Shaurya Chanana, Erol Bahadroglu, Alec Peltekian, and Grégoire Altan-Bonnet. 2021b. Scifive: a text-to-text transformer model for biomedical literature. *Preprint*, arXiv:2106.03598.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

L. Smith, L. Tanabe, R. Ando, C. Kuo, I-Fang Chung, C. Hsu, Y. Lin, Roman Klinger, Christoph Friedrich, K. Ganchev, M. Torii, Hongfang Liu, Barry Haddow, Craig Struble, Richard Povinelli, Andreas Vlachos, William Baumgartner Jr, Lawrence Hunter, B. Carpenter, and W. Wilbur. 2008a. Overview of biocreative ii gene mention recognition. *Genome Biology*, 9.

Larry Smith, Lorraine K. Tanabe, Rie Ando, Cheng Ju Kuo, I. Fang Chung, Chun Nan Hsu, Yu Shi Lin, Roman Klinger, Christoph M. Friedrich, Kuzman Ganchev, Manabu Torii, Hongfang Liu, Barry Haddow, Craig A. Struble, Richard J. Povinelli, Andreas Vlachos, William A. Baumgartner, Lawrence Hunter, Bob Carpenter, and 15 others. 2008b. Overview of biocreative ii gene mention recognition. *Genome biology*, 9(SUPPL. 2).

Mohammad Golam Sohrab, Masaki Asada, Matīss Rikters, and Makoto Miwa. 2024. Bert-nar-bert: A non-autoregressive pre-trained sequence-to-sequence model leveraging bert checkpoints. *IEEE Access*, 12:23–33.

Mohammad Golam Sohrab and Makoto Miwa. 2024. Kat5: Knowledge-aware transfer learning with a text-to-text transfer transformer. In *Machine Learning and Knowledge Discovery in Databases. Applied Data Science Track*, pages 157–173, Cham. Springer Nature Switzerland.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium. Association for Computational Linguistics.

Benjamin Warner, Antoine Chaffin, Benjamin Clavié, Orion Weller, Oskar Hallström, Said Taghadouini, Alexis Gallagher, Raja Biswas, Faisal Ladhak, Tom Aarsen, Griffin Thomas Adams, Jeremy Howard, and Iacopo Poli. 2025. Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2526–2547, Vienna, Austria. Association for Computational Linguistics.

Hongyi Yuan, Zheng Yuan, Ruyi Gan, Jiaxing Zhang, Yutao Xie, and Sheng Yu. 2022. BioBART: Pretraining and evaluation of a biomedical generative language model. In *Proceedings of the 21st Workshop on Biomedical Language Processing*, pages 97–109, Dublin, Ireland. Association for Computational Linguistics.

## A Pretraining Data Preparation

We downloaded the full release of UMLS2020AA [9] version. The unified medical language system (UMLS) Metathesaurus includes names and codes from standard biomedical vocabularies including SNOMED CT, RxNorm, LOINC, MeSH, CPT, ICD-10-CM, MedDRA, Human Phenotype Ontology, etc. and extract all English entries based on the 21 most popular source vocabularies stated in Table 12 to restrict our search space of a fill list of UMLS vocabulary [10]. After preprocessing, a list of 407,917 unique concepts (CUI) are extracted to create UMLS synonym dictionary where the dictionary size is 1,193,241. From the UMLS2020AA release, we processed 1219 `xml.gz` pubmed files that aligned a total of 362,023,532 concept mentions based on string matching and generates 1219 JSON files equivalent to 29GB of text as an input of the BioCAT5-base and BioCAT5-3B models. Table 13 shows a concept-aware sequence masking and one sample pretraining JSON input out of 1219 JSON files is included in the GitHub data directory.

[9] https://download.nlm.nih.gov/kss/2020AA/umls-2020AA-full.zip

[10] https://www.nlm.nih.gov/research/umls/sourcereleasedocs/

## B Details on NER Datasets

- NCBI-Disease: The Natural Center for Biotechnology Information Disease corpus (Doğan et al., 2014) contains 793 PubMed abstracts with 6892 annotated disease mentions linked to 790 distinct disease entities. It consists of 5,424/923/940 samples in the training/validation/test set.

- BC5CDR-Chemical and Disease: The BioCreative V Chemical-Disease Relation corpus (Smith et al., 2008b) is frequently used as a NER corpus for detecting chemical (drug) and disease entities which was created for evaluating relation extraction of drug-disease interactions based on 1500 PubMed abstracts. We use the same pre-processed version of BC5CDR chemical and BC5CDR disease that are evaluated by BioBERT and SciFive. The BC5CDR chemical and BC5CDR disease datasets consist of 4560/4581/4797 samples in the training/validation/test set.

- BC2GM: The Biocreative II Gene Mention corpus [53] consists of sentences from PubMed abstracts with manually labeled gene and alternative gene entities. Following prior work, we focus on the gene entity annotation. In its original form, BC2GM contains 15000 train and 5000 test sentences. We use a pre-processed version of the dataset by following the processing pipeline of (Lee et al., 2019) training/validation/test set that consist of 12574/2519/5038 samples respectively.

- JNLPBA: The Joint Workshop on Natural Language Processing in Biomedicine and its Applications (Collier et al., 2004) shared task is a NER corpus on PubMed abstracts where the entity types are protein, DNA, RNA, cell line, and cell type. We follow the same train/validation/test set (Crichton et al., 2017) of 14690/3856/3856 samples respectively.

## C Details on Model Hyper-parameters

Our implementation is based on the Hugging Face encoder-decoder transformers. We used two different computing infrastructure for the BioCAT5-base and BioCA5-large models.

- BioCAT5-Base: For the base model, NVIDIA RTX 6000 ada with 48GB GDDR6 memory is used. During pretraining four GPUs are assigned where the batch size for each GPU is set to 30 with a maximum sequence length of 512 tokens. We initialized the BioCAT5 base model from the publicly available checkpoint [11]. We use a learning rate of 0.001, a linear warm-up of 2k steps, a gradient accumulation of 5 steps. The BioCAT5 model is trained on 13.6B tokens, where we employ a batch size of 61,440 tokens with a maximum step of 221,600 steps which is around 3 times smaller than the original T5 model.

- BioCAT5-3B: BioCAT5-3B is a large language model (LLM) and it is not compatible to train the BioCAT5-3B model with NVIDIA RTX 6000 ada (48GB GPU memory) computing resources. Instead, we used the NVIDIA H200 SXM 141GB GPU for the continual pretraining of BioCAT5-3B. During continual pretraining of BioCAT5, 16 GPUs are assigned where the batch for each GPU is set to 10 with a maximum sequence length of 256 to reduce the computational complexity. The model is initialized from the publicly available checkpoint [12]. However, the BioCAT5-3B model is trained on 1.06B tokens, where we employ a batch size of 40,960 tokens with a maximum step of 26,000 steps. We used the same learning rate and warmup steps as like the BioCAT5-base model.

The BioCAT5-base and BioCAT5-3B models are optimized end-to-end using an Adafactor optimizer and updates the entire model parameters with a concept-aware span denoising learning objective on the training split of our conceptually aligned dataset.

## D BioCAT5 In/Output Representation

Tables 14, 15, and 16 show the input and output format during fine-tuning with BioCAT5-base and BioCAT5-3B on the downstream datasets. We mainly consider single-task settings: a single model on a single dataset is learned initializing from the BioCAT5-base and BioCAT5-3B models. In this setting, we add the dataset name followed by the task separator : (for example, "ncbi_ner :" for NCBI dataset) as a prefix to each input sentence. We fine-tune on top of BioCAT5 for a maximum of 10-30 epochs based on the data size in all our downstream tasks.

| **UMLS Metathesaurus Vocabulary** | |
|---|---|
| Abbreviation | Short Name |
| MSH | Medical Subject Headings |
| CSP | CRISP Thesaurus |
| NCI | NCI Thesaurus |
| MTH | Metathesaurus Names |
| FMA | Foundational Model of Anatomy |
| UWDA | Digital Anatomist |
| NIH | NIH |
| HPO | Human Phenotype Ontology |
| HGNC | HUGO Gene Nomenclature Committee |
| OMIM | Online Mendelian Inheritance in Man |
| RXNORM | RXNORM |
| NDFRT | NDFRT |
| NCBI | NCBI Taxonomy |
| MDR | MedDRA |
| ICD9CM | International Classification of Diseases, Ninth Revision, Clinical Modification |
| ICD10CM | Clinical Classifications Software Refined for ICD-10-CM |
| ICD10 | Int. Classification of Diseases and Related Health Problems, Tenth Revision |
| CPT | Current Procedural Terminology |
| GO | Gene Ontology |
| MEDLINEPLUS | MedlinePlus Health Topics |
| SNOMEDCT_US | SNOMED CT, US Edition |

Table 12: UMLS Metathesaurus vocabulary documentation that used during pretraining data preparation.

[11] https://huggingface.co/razent/SciFive-base-Pubmed

[12] https://huggingface.co/google/umt5-xl

| | |
|---|---|
| **Input Sequence:** | Formate assay in body fluids : application in methanol poisoning . |
| **Concept Masking:** | Formate assay in $< extra_id_1 >$ : application in $< extra_id_0 > < extra_id_2 >$ . |
| **Target Sequence:** | $< extra_id_0 >$ methanol $< extra_id_1 >$ body fluids $< extra_id_2 >$ poisoning |

Table 13: An example of concept-aware sequence masking.

| | |
|---|---|
| **NCBI-Disease** | |
| Input Sequence: | ncbi_ner: Identification of APC2 , a homologue of the adenomatous polyposis coli tumour suppressor . |
| Target Sequence: | Identification of APC2 , a homologue of the disease* adenomatous polyposis coli tumour *disease suppressor . |
| **BC5CDR-Disease** | |
| Input Sequence: | bc5cdr_disease_ner: Hemolysis of human erythrocytes induced by tamoxifen is related to disruption of membrane structure . |
| Target Sequence: | disease* Hemolysis *disease of human erythrocytes induced by tamoxifen is related to disruption of membrane structure . |
| **BC5CDR-Chemical** | |
| Input Sequence: | bc5cdr_chem_ner: Effect of intravenous nimodipine on blood pressure and outcome after acute stroke . |
| Target Sequence: | Effect of intravenous chem* nimodipine *chem on blood pressure and outcome after acute stroke . |
| **BC2GM** | |
| Input Sequence: | bc2gm_ner: Accordingly , we designated this gene CTL1 ( capping enzyme RNAtriphosphatase - like 1 ) . |
| Target Sequence: | Accordingly , we designated this gene* gene CTL1 *gene ( gene* capping enzyme RNAtriphosphatase - like 1 *gene ) . |
| **JNLPBA** | |
| Input Sequence: | jnlpba_ner: Concomitant with this loss of a , a new Ets - 1 - containing complex ‘ ‘ c ‘ ‘ appeared . |
| Target Sequence: | Concomitant with this loss of a , a gene* new Ets - 1 - containing complex *gene ‘ ‘ gene* c *gene ‘ ‘ appeared . |

Table 14: An example of input and output sequence of named entity recognition task.

| | |
|---|---|
| **BioAsq** | |
| Input Sequence: | bioasq: question: Does strenuous physical activity affect thyroid hormone metabolism? context: The results of the present study in a unique experimental human model of maximal exposure to altitude and physical exercise demonstrate that extreme HA and strenuous physical exercise are coupled with specific endocrine adaptations. These include increased activity of the GH/IGF-I axis and a low T(3) syndrome 3,5,3'-triiodothyronine (T3) and T4 levels increase during strenuous exercise, and, at the end of the exercise bout, a decrease of T3 and T4 levels, with an increase in TSH during the following 4-5 days, is seen. the obtained results indicate that in intense exercise, causing the rapid development of fatigue, rapid increases in serum levels of hormones of the pituitary-adrenocortical, pituitary-gonadal and pituitary-thyroid systems occur. Mean levels of fasting plasma estradiol, luteinizing hormone, follicle-stimulating hormone, free thyroxine and triiodothyronine were significantly lower in AR compared to ER and SE. Reductions in plasma T4, T3 and T3/T4 ratio are probably due to inhibition of T4 secretion and 5'-monodeiodination with possible conversion of T4 to reverse T3 (rT3). These processes may represent a mechanism for regulation of thyroid hormone metabolism during strenuous and extended flight. Strenuous endurance training seems to have minor changes on the function of the thyroid gland. Depressed T4 levels in runners may rather be due to lowered TBG levels than due to direct effect of training. brief strenuous swimming or moderate bicycle exercise had minor or no effect on thyroid hormone concentrations when consideration was given to the attendant hemoconcentration. thyroxine were determined in 26 men participating in a 90-km cross-country ski race, before, immediately after, and on the following days Total thyroxine and free thyroxine in serum were significantly increased at the end of the race, but had returned to the pre-raced levels during the rest of the observation period. There are controversial results concerning thyroid hormone metabolism during strenuous exercise in adult athletes and only scant data concerning the impact of strenuous exercise on thyroid hormone metabolism in children and adolescents. |
| Target Sequence: | yes |
| **PubMedQA** | |
| Input Sequence: | pubmedqa: question: Juvenile ankylosing spondylitis–is it the same disease as adult ankylosing spondylitis? context: Juvenile and adult forms of ankylosing spondylitis (AS) have been shown to have different clinical presentation and outcome in Caucasians. We did this retrospective analysis to see if similar differences exist in the Indian population. Case records of 210 Indian patients diagnosed with AS according to modified New York criteria were reviewed. Data were collected regarding age of onset, clinical features, drug treatment, and outcome at last follow-up. Patients with onset before 17 years of age were classified as having juvenile AS (JAS) and the rest with adult AS (AAS). There were 150 patients with AAS and 60 with JAS. The latter had higher male preponderance, more frequent onset with peripheral arthritis, and greater involvement of hip and knee joints. Valvular dysfunction was seen only in patients with JAS. |
| Target Sequence: | no |

Table 15: An example of input and output sequence of question answering task.

| | |
|---|---|
| **ChemProt** | |
| Input Sequence: | chemprot_re: The effects of mitiglinide (KAD-1229), a new anti-diabetic drug, on @GENE$ and insulin secretion: comparison with the @CHEMICAL$ . |
| Target Sequence: | 0 |
| **DDI** | |
| Input Sequence: | ddi_re: Prior administration of @DRUG$ has no clinically important effect on the neuromuscular blocking action of @DRUG$. |
| Target Sequence: | 0 |
| **GAD** | |
| Input Sequence: | gad_re: this study proposes that A/A genotype at position -607 in @GENE$ gene can be used as a new genetic maker in Thai population for predicting @DISEASE$ development. |
| Target Sequence: | 0 |

Table 16: An example of input and output sequence of relation extraction task.