RELATION EXTRACTION FOR CHEMICAL AND PROTEIN INTERACTIONS FROM BIOMEDICAL DOCUMENTS

by

Hilal Dönmez

B.S., Computer Engineering, Boğaziçi University, 2018

Submitted to the Institute for Graduate Studies in
Science and Engineering in partial fulfillment of
the requirements for the degree of
Master of Science

Graduate Program in Computer Engineering

Boğaziçi University

2021

ACKNOWLEDGEMENTS

Firstly, I would like to state my thankfulness to Assoc. Prof. Arzucan Özgür and Assoc. Prof. Elif Özkırımlı for their endless patience, valuable advice, and continuous guidance throughout our study. It was an honor for me to work with them and gain a scientific perspective from them.

I am very grateful to my thesis committee members; Prof. Tunga Güngör and Prof. Nilgün Karalı for their valuable time and recommendations.

I would like to thank biomedical experts, Asu Büşra Temizer, Eren Can Ekşi, Prof. Hacer Karataş, İrem Şenman, Prof. Nilgün Karalı, Serhat Beyaz, and Yağmur Ersoy for their detailed evaluation of Vapur.

I am very thankful to Abdüllatif Köksal and Rıza Özçelik for their support, helpfulness, and their valuable comments on my thesis. Besides, I am grateful for their friendship and all the delightful time we shared.

Finally, I thank my family for their endless support and kindness. I am grateful to my mother and my father for their friendship and everything they have done for me. I am very thankful to my brother for always supporting me and making me happier.

ABSTRACT

RELATION EXTRACTION FOR CHEMICAL AND PROTEIN INTERACTIONS FROM BIOMEDICAL DOCUMENTS

The sharing of chemical-protein interactions (CPI) with the scientific communities plays a crucial role in understanding the mechanisms of diseases, as well as in facilitating drug discovery and drug repurposing studies. Significant amount of knowledge on CPI is published in unstructured documents. The goal of this thesis is to extract relations between chemicals and proteins from information provided in sentences. For this purpose, we focus on two tasks: (i) binary relation extraction and (ii) multi-class relation extraction from biomedical documents. The aim of the first task is to identify whether a sentence states a relation between a pair of biochemicals or not. On the other hand, the second task extends the first one by also aiming at identifying the type of the relation between the pair of biochemicals. For both tasks, we develop transformer-based models by utilising the BioBERT and SciBERT architectures. Furthermore, we investigate the effectiveness of different input representation approaches such as sentence and dependency tree-based representations. Our results demonstrate that BioBERT based model with whole sentence input representation achieves the best performance for both tasks on the benchmark ChemProt test data set with an F1-score of 77.8% for binary relation extraction and micro-averaged F1-score of 76.1% for multi-class relation extraction. Interestingly, the significantly shorter dependency tree based input representations achieve close F1-scores to whole sentence input representation. Finally, we introduce Vapur, which is a search engine for protein-chemical interactions extracted from COVID-19 related scientific publications. Vapur shows that our relation extraction models can be effectively used in real-world biomedical applications.

ÖZET

KİMYASALLAR VE PROTEİNLER ARASINDAKİ ETKİLEŞİMLER İÇİN BİYOMEDİKAL DOKÜMANLARDAN İLİŞKİ ÇIKARMA

Kimyasallar ve proteinler arasındaki etkileşimlerin (CPI) bilimsel topluluklar ile paylaşılması, hastalık mekanizmalarının anlaşılmasında, ilaç keşfindeki ve ilaçların yeniden kullanılmasındaki çalışmaların kolaylaştırılmasında önemli rol oynar. CPI hakkında önemli miktarda bilgi, düzenli yapısı olmayan dokümanlarda yayınlanmaktadır. Bu tezin amacı cümlelerde verilen bilgilerden kimyasallar ve proteinler arasındaki ilişkileri çıkarmaktır. Bu amaç için, biyomedikal dokümanlardan ikili ilişki çıkarma ve çok sınıflı ilişki çıkarma olmak üzere iki göreve odanlanmaktayız. İlk görevin amacı, bir cümlenin bir çift biyokimyasal arasındaki ilişkiyi ifade edip etmediğinin belirlenmesidir. İkinci ise görev bir çift biyokimyasal arasındaki ilişkinin tipini de belirlemeyi amaçlayarak ilk görevi genişletir. İki görevde de, BioBERT ve SciBERT mimarilerinden yararlanarak Dönüştürücü tabanlı modeller geliştiririz. Ek olarak, tüm cümle tabanlı ve bağlılık ağacı tabanlı temsillerinden oluşan farklı girdi temsilleri yaklaşımlarımızın etkisini araştırıyoruz. Bizim sonuçlarımız, ChemProt test veri seti üzerinde ikili ilişki çıkarma görevinde %77.8 F1 ölçütü ve çok sınıflı ilişki çıkarma görevinde %76.1 mikro-ortalamalı F1 ölçütü elde eden tüm cümle girdi temsilimiz ile eğittiğimiz BioBERT tabanlı modelimizin her iki görevde de en iyi performansı elde ettiğini göstermektedir. İlginç bir şekilde, önemli ölçüde daha kısa olan bağlılık ağacı tabanlı girdi temsilleri, tüm cümle girdi temsiline yakın F1 ölçütü elde eder. Son olarak, KOVID-19 ile ilgili bilimsel yayınlardan protein-kimyasal etkileşimleri çıkaran bir arama motoru olan Vapur'u tanıtıyoruz. Vapur, ilişki çıkarma modellerimizin gerçek yaşamdaki biyomedikal uygulamalarda etkin bir şekilde kullanılabildiğini göstermektedir.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF SYMBOLS

#	Number
<	A beginning tag of the special entities in our preprocessing steps
>	An end tag of the special entities in our preprocessing steps
&	And
±	Plus and minus symbol

LIST OF ACRONYMS/ABBREVIATIONS

BERN	A Neural Biomedical NER and Multi-type Normalization
BERT	Bidirectional Encoder Representations from Transformers
BiLSTM	Bidirectional Long Short-Term Memory
BioNLP	Biomedical Natural Language Processing
Books	BookCorpus
ChemProt	BioCreative VI: Chemical-Protein Interaction Track
CNN	Convolutional Neural Network
CORD-19	COVID-19 Open Research Dataset
COVID	Coronavirus Disease
COVID-19	Coronavirus Disease of 2019
CPI	Chemical Protein Interaction
CPR	Chemical Protein Relation
CRF	Conditional Random Fields
dev	Development Set
GAD	Genetic Association Database
GLUE	General Language Understanding Evaluation Benchmark
GPRO	Gene and Protein Related Objects
GRGT	Grammatical Relationship Graph for Triplets
ID	Identification
LDA	Linear Discriminant Analysis
LR	Learning Rate
LSTM	Long Short-Term Memory
MASS	Masked Seq2Seq Pre-training for Language Generation
MLM	Masked Language Model
NER	Named Entity Recognition
NLP	Natural Language Processing
NSP	Next Sentence Prediction
PMC	PubMed Central Full-Text Articles

PMID	PubMed Identifier
POS	Part of Speech
PPI	Protein-Protein Interaction
RdRp	RNA-dependent RNA Polymerase Enzyme
RE	Relation Extraction
RNN	Recurrent Neural Network
SARS-CoV-2	Severe Acute Respiratory Syndrome Coronavirus 2
SB	Sentence-based Input Representation
SGD	Stochastic Gradient Descent
SQuAD	Stanford Question Answering Dataset
SST	The Shortest Sequence Between Target Entity Pairs
SSTP	The Shortest Sequence Including Target Entity Pairs and Their Parent Node
std	Standard Deviation
SVM	Support Vector Machine
SWAG	The Situations With Adversarial Generations
UTF8	Unicode Transformation Format 8-bit
WHO	World Health Organization
Wiki	English Wikipedia
XGBoost	eXtreme Gradient Boosting

1. INTRODUCTION

There are various biomedical researches about the identification of the associations between chemicals and proteins, since the information included in chemical-protein interactions (CPI) plays important roles in crucial tasks such as drug discovery, understanding the mechanism of the diseases, and precision medicine. Thanks to biomedical researchers, a significant amount of knowledge on CPI has been obtained and published in the scientific literature. Since the relation information in the scientific literature is stored as unstructured text, it is not straightforward to extract the CPI information from literature manually. Even if the information is obtained from literature manually, information extraction of the CPI in the scientific documents is time-consuming due to a large amount of information. Therefore, manual extraction methods are neither efficient nor preferable. As a result, an automatic method for the CPI extraction from the scientific literature is crucial.

A significant number of studies have been developed to address similar problems such as protein-protein interaction and drug-target interaction from the biomedical literature. The studies have addressed the problems using various techniques including rule-based pattern matching, machine learning algorithms, and natural language processing methods. In contrast to these studies, there have been a couple of researches on the extraction of CPI from the scientific literature. There is a need to extract CPI information with time efficiency and deposit them into databases in order to facilitate effective querying of the information. Consequently, the structured CPI information can also be very beneficial as a database in other interactive studies such as drug discovery, disease, and medicine pairs. Therefore, we introduce an approach for CPI extraction from the scientific documents to address the problem and serve a purpose in biomedical studies.

This thesis aims to extract relations between chemical and protein entities from provided information in sentences. For this purpose, we focus on two tasks that are

binary relation extraction and multi-class relation extraction from biomedical documents. The first task aims to determine whether a sentence expresses that there is a relation between chemical and protein entities or not. In addition, we present Vapur as an application of relation extraction on Coronavirus Disease (COVID) literature using our binary relation extraction approach. Vapur is an online COVID-19 search engine so as to find relevant chemical and protein pairs from Coronavirus Disease of 2019 (COVID-19) literature and to retrieve the related documents to a user query. The objective of the second task is to determine the relation type specified on ChemProt data set if a sentence explicitly states that the specified pair of chemical and protein entities have a relation.

The following chapters of our thesis are organized as follows. Chapter 2 provides the related works on chemical-protein interactions, protein-protein interactions, relation extraction, and pretrained language models used in this thesis. Chapter 3 introduces detailed information about the data set and how it is used in this study. In addition, this chapter includes all preprocessing steps on the data set, sentence-based and dependency tree-based input representations as well as our relation extraction approaches that were used throughout this study. Our binary and multi-class relation extraction approaches are explained in this chapter. It provides the design of the experiments reported and discussed in this study. Chapter 4 contains the results and discussion of the experiments. In Chapter 5, we introduce Vapur as an application of relation extraction on COVID literature. This chapter includes detailed information about COVID-related data set, the workflow of the pipeline behind Vapur, the results of the experiments reported and discussed. We published Chapter 5 in [1]. Chapter 6 covers a summary of important findings and possible strategies for future works. Finally, supplementary information is reported in the Appendix.

2. RELATED WORK

2.1. Chemical-Protein Interactions

The BioCreative VI track 5 ChemProt task aimed to promote the improvement of the systems which can automatically classify the chemical-protein relations from PubMed abstracts. BioCreative VI provided an annotated chemical-protein interaction corpus which is called as ChemProt for this task. Therefore, there have been various studies on chemical-protein interactions extraction from the literature for ChemProt task at BioCreative VI and their top 20 results for the task are reported in Table 2.1.

An ensemble of three systems containing a support vector machine (SVM), a convolutional neural network (CNN), and a recurrent neural network (RNN) was introduced in [2]. The outputs of the three systems were combined using their majority voting or stacking. SVM was trained with various generated features such as the words surrounding the chemical and protein mentions of interest, bag-of-words between the chemical and protein mentions of interest in a sentence, the distance between two entity mentions in a sentence, and the existence of a keyword between two mentions. In CNN, each word was represented by concatenating embeddings of the words, part of the speech tags, name entities, dependencies, and positions relative to the two mentions of interest. Also, this study constructed a bidirectional Long-Short Term Memory (BiLSTM) with the same word representation as the input of CNN. As a result of the study, the majority voting was submitted with the highest score in the BioCreative task in 2017.

A study based on sentence structure analysis for chemical-protein interactions from the biomedical literature was presented in [3]. This study extracted CPI pairs which were defined as a chemical entity and a protein entity in the same sentence and CPI triplet which was described as the combination of a CPI pair with an interaction word in the sentence. After it constructed two different models for CPI pairs and

CPI triplets separately, they were combined into a final model. It generated features from the semantic pattern and the dependency graph of a sentence since the semantic pattern includes information of the interaction between chemicals and proteins in a sentence and the dependency graph gives information about the interconnection of the words in a sentence. These features were used in several machine learning classifiers such as Random Forest Classifier, Extremely Randomized Trees, Linear Discriminant Analysis (LDA), Logistic Regression, and eXtreme Gradient Boosting (XGBoost) [4]. Finally, this study applied the three-stage model building which is based on stacked generalization in the BioCreative task.

A CNN-based classification for chemical-protein interaction was developed in [5]. Firstly, this study converted each sentence into the form of the candidate relation. It produced two different approaches to represent the candidate relations, which were word embedding and distance embedding. Word embedding reveals the semantic aspect of the word in a sentence, whereas distance embedding indicates the relative distance between words and entities in a sentence. After creating a representation of the relation, it constructed a CNN model to predict the group of CPI.

2.2. Protein-Protein Interactions

Protein-protein interaction (PPI) extraction is a task to identify interactions between specified proteins in biomedical documents. The identification of the chemical-protein interactions is a similar task to the recognition of the protein-protein interactions since two tasks are depend on extracting predefined relations between specified entity pairs in biomedical papers. Therefore, we examine lots of studies on the extraction of protein-protein interactions.

Previous researches explore protein-protein interactions extraction from biomedical documents using traditional machine learning algorithms with handcrafted features. SVM has become a widely used classifier with bag-of-words features from sentences [6] as well as features based on linguistic knowledge such as lemma and part of speech

Table 2.1. Evaluation results of teams at The BioCreative VI track 5 ChemProt task. Each team has an opportunity to submit 5 prediction results on the Gold Standard data set. Micro-averaged F-score was selected as the main metric during the contest.

TEAM ID	**Run**	**Precision**	**Recall**	**F-score**
TEAM_430	RUN 5	0.7266	0.5735	0.641
TEAM_430	RUN 4	0.7311	0.5685	0.6397
TEAM_430	RUN 1	0.7437	0.5529	0.6343
TEAM_430	RUN 2	0.7283	0.5503	0.6269
TEAM_430	RUN 3	0.7426	0.5382	0.6241
TEAM_403	RUN 1	0.561	0.6784	0.6141
TEAM_417	RUN 3	0.6608	0.5662	0.6099
TEAM_417	RUN 4	0.6105	0.6006	0.6055
TEAM_417	RUN 5	0.6088	0.5989	0.6038
TEAM_424	RUN 2	0.6704	0.5194	0.5853
TEAM_424	RUN 1	0.676	0.5159	0.5852
TEAM_433	RUN 2	0.6352	0.5121	0.5671
TEAM_433	RUN 1	0.6276	0.4858	0.5477
TEAM_417	RUN 1	0.6373	0.4462	0.5249
TEAM_417	RUN 2	0.6337	0.4387	0.5185
TEAM_374	RUN 5	0.5738	0.4722	0.5181
TEAM_379	RUN 5	0.5301	0.4639	0.4948
TEAM_374	RUN 2	0.5156	0.467	0.4901
TEAM_379	RUN 2	0.4849	0.4913	0.4881
TEAM_379	RUN 4	0.5072	0.4306	0.4657

(POS) tags of tokens [7] to identify protein-protein interactions in documents. In addition, dependency tree-based methods have been used to extract interactions between proteins from biomedical documents. Grammatical Relationship Graph for Triplets (GRGT) is presented in [8] where triplets are defined as two protein entities and an interaction word between them in a sentence. This study focuses on the dependency graph to obtain sentence structure and the shortest paths among the pairs in triplets via Dijkstra's shortest path algorithm [9].

As deep learning methods have gained popularity in several NLP tasks, recent studies have been focused on neural network architectures to detect interactions between proteins in documents. Convolutional neural networks (CNN) have achieved notable results considering linguistic and semantic information in sentences such as word embedding, position features, part of speech tags, chunk, and dependency information of sentences [10], [11], [12]. In addition, BiLSTMs have promising results for PPI by capturing syntactic and semantic information of specified entities and their neighbors [13].

2.3. Relation Extraction

Relation extraction (RE) is one of the important topics in information extraction and is defined as the task of determining whether pairs of entities in a sequence of words have a semantic relation and classifying the relation into one of the predefined relation types. A wide range of NLP tasks including question answering [14] and text mining [15] makes use of the relation information in documents.

Traditional RE models address the task by applying feature-based methods [16], [17] and using pretrained word embeddings [18] with various neural network architectures [19], [20], [21], [22]. Recently, as transformers [23] such as BERT [24], GPT [25], RoBERTa [26], and MASS [27] have gained popularity on a range of NLP tasks including machine translation [23], document generation [28], and textual entailment [29], NLP researchers have widely used transformers-based models on relation extraction

[30], [31], [32], [33].

Keeping up to date with the growing body of biomedical literature is a big challenge and almost a kind of magic in scientific research. Efforts such as ChemProt [34] aim to promote research in this direction by presenting PubMed abstracts in which proteins, chemicals, and their relations are manually labeled. [35] and [36] developed sequential models for chemical and protein relation extraction on ChemProt, while [37] used an ensemble of deep models with SVM. However, all these models depend on the named entities to be recognized beforehand.

Named entity recognition and normalization are widely studied topics in biomedical text mining to extract and link entities. Conditional Random Fields (CRF) was a popular approach in early biomedical NER studies [38], [39], [40] and with the rise of deep learning, sequential models were also integrated [41]. Recently, transformers-based NER models attracted more attention, including BERN [42], which is a state-of-the-art biomedical named entity recognition and normalization tool that uses BioBERT [43] to identify and normalize the entities in a sentence. BERN is adopted by recent general-domain biomedical text mining studies [44] as well as a model to answer COVID-19 related questions based on CORD-19 [45].

BioBERT [43] is a domain-specific transformers-based language representation model which is pretrained on different combinations of text corpora: English Wikipedia (Wiki), BookCorpus (Books), PubMed abstracts (PubMed), and PubMed Central full-text articles (PMC). Due to the different combinations of text corpora, there are several pretrained versions of BioBERT trained. All of them use the original vocabulary and the initial weights of $\text{BERT}_{\text{BASE}}$ [24]. BioBERT is separately finetuned by utilizing the sentence classifier of BERT for biomedical relation extraction task on three different data sets which are GAD [46], EU-ADR [47], and ChemProt [48]. In each data set, target entities in a sentence are anonymized with predefined tags (e.g. @GENE$ and @DISEASE$). On ChemProt, BioBERT v1.1 achieves a higher micro-averaged F1 score than BERT, reported in [43]. On the other hand, BioBERT v1.0 (+ PMC)

outperforms BERT on EU-ADR, whereas BioBERT v1.0 (+ PubMed) had the second-best F score on GAD, stated in [43]. Finally, BioBERT obtained the highest F1 score on 2 out of 3 relation extraction data sets compared to BERT and other state-of-the-art models in 2019 [43].

SciBERT [49] is a transformers-based language representation model which is pretrained on papers from the computer science and the biomedical domain obtained from Semantic Scholar. Its corpus consists of the full text of papers where 18% of them is related to the computer science domain and others are from the biomedical domain. Although SciBERT is trained with the same configuration and architecture of $\text{BERT}_{\text{BASE}}$, SciBERT uses an in-domain vocabulary (SCIVOCAB) trained from scratch. During all experiments, there are four pretrained versions of SciBERT: i) case or uncased of text ii) vocabulary released with BERT or trained from the scientific corpus. Each version of SciBERT is finetuned by using the same architecture, optimization, and hyperparameters of $\text{BERT}_{\text{BASE}}$ for a range of NLP tasks such as named entity recognition, relation extraction, and text classification. In relation extraction task on ChemProt, SciBERT is finetuned with sentences having target entities that are encapsulated by special tokens.

2.4. Pretrained Language Models Used in This Thesis

The earliest works have widely used RNN architectures to represent input sequences in a variety of tasks including text classification, text generation, and machine translation. However, RNNs do not have the ability to solve the long-term dependency problem, since it mainly focuses on the previous token. Although LSTMs are designed to avoid the problem by forgetting the unimportant parts and remembering important parts of the input sequence using different gates, they have still a problem with the long sequence. The reason for this is that they compress all information of the input sequence into a fixed-length vector. To overcome this problem, the attention mechanism concentrates the most relevant information of the input sequence into a fixed-length vector instead of using the entire input sequence. Nevertheless, the at-

tention mechanism learns the dependency between inputs and outputs by focusing on global information. To handle the problem, a new encoder-decoder architecture based on a multi-head attention mechanism is designed and called Transformers. Since we used transformer-based models as pretrained language models in this thesis, we give detailed information about transformers in this section.

The Transformer is a sequence transduction model with multi-headed self-attention to represent the tokens with information of all other words in a sequence. The Transformer has a sequential encoder block with six different encoders and a decoder block with six different decoders. Although each encoder has an identical structure including a self-attention layer, normalization layers, and a feed-forward network layer, it has different weights of layers. While the bottom encoder takes the vector of the sum of positional embedding and token embedding of a token at each position as an input, other encoders take the output vectors of their previous encoder. The aim of the encoders is to create a new representation for each token by aggregating the information from other words in the sequence and focusing on different positions on the sequence with multi-head self-attentions. The output attention vectors of the final encoder feed each decoder. Likewise encoders, each decoder has different weights of the layers, although it has an identical structure consisting of a self-attention layer, an encoder-decoder attention layer, normalization layers, and a feed-forward network layer. The objective of the decoder block is to predict a new token in each iteration considering the output vectors from the encoder block by minimizing the loss score of the model. During the sequence prediction, the decoder block is fed by the embeddings of the earlier predicted tokens, their positional embeddings as well as the output vectors of the encoder block in each iteration. An encoder-decoder attention layer is used to help a decoder focus on appropriate positions in the input sequence [23].

As Transformers are suitable for different NLP tasks including machine translation tasks and sequence parsers and attain state-of-the-art results on these tasks [23], various transformer-based models such as BERT [24], GPT [25], RoBERTa [26], and XLNet [50] have been widely used on several NLP tasks. Since we use BERT-

based models on relation extraction tasks throughout this study, we firstly provide an overview of BERT and two other BERT-based models (SciBERT and BioBERT) used in our thesis.

2.4.1. BERT

BERT (Bidirectional Encoder Representations from Transformers) is a deep bidirectional language representation that is pretrained on unlabeled texts by fusing left and right context jointly. The pretrained BERT achieves state-of-the-art results by finetuning it with an output layer for a wide variety of NLP tasks with minimal task-specific architecture modifications.

In previous studies, unidirectional language models [18], [51] and a shallow concatenation of independently trained left-to-right and right-to-left language models [52] have been widely used in order to represent tokens in several NLP tasks such as question answering [53], sentiment analysis [54], and named entity recognition [55]. However, there are some limitations to represent tokens by previous language models. In unidirectional language models [25] using left-to-right architectures, each token can only be represented considering the previous ones. Since bidirectional language models using left-to-right and right-to-left language models represent a token with context-dependent features in task-specific architectures, these contextual representations are task-specific and feature-based, not deeply bidirectional. Unlike previous language models, after BERT is pretrained with two tasks which are masked word prediction by masked language model (MLM) and next sentence prediction (NSP), researchers finetune pretrained BERT on different NLP tasks [56].

The model architecture of BERT is multi-layer bidirectional Transformers described in [23]. There are two different BERT models with different configurations:

- $\text{BERT}_{\text{BASE}}$ has 12 layers, 768 hidden size, and 12 self-attention heads with 110M total parameters.

- $\text{BERT}_{\text{LARGE}}$ has 24 layers, 1024 hidden size, and 16 self-attention heads with 340M total parameters.

During pretraining, BERT takes the input as one token sequence which can be a single sentence or a pair of sentences. Each sequence is tokenized by WordPiece embedding [57]. The first token of a sequence is a special classification token ([CLS]) that has the fixed-length sequence representation for several classification tasks. Since sentence pairs can be aggregated into a single sequence, another special token ([SEP]) separates sentences. Each token is represented as the input embedding that is the sum of the token embeddings, segment embeddings, and position embeddings during pretraining. Pretraining corpus is collected from BooksCorpus [58] and English Wikipedia for two tasks, masked language model and next sentence prediction. For the first task, a masked language model is a bidirectional language model trained with masked words randomly selected as 15% of all tokens in a sequence and predicts only masked tokens instead of constructing the entire input.

A binarized next sentence prediction task aims to understand the relationship between two sentences, not captured directly by language models. A model is pretrained with two types of sentence pairs. The first type of pair consists of a sentence and its next sentence, whereas the second type of pair contains a sentence and a random sentence that does not follow it from the corpus. Each sentence pair has a special tag [CLS] that aggregates the sentence representation for the next sentence prediction task. This pretraining for next sentence prediction is effective on various NLP tasks such as Question Answering and Natural Language Inference to understand the relationship between sentences.

After BERT is pretrained with two essential tasks, it is finetuned for its all parameters end-to-end by using the self-attention mechanism in Transformers for several downstream NLP tasks with task-specific inputs and outputs. In text classification tasks, BERT takes the tokens of the text as input. At input for text classification, BERT takes the tokens of the sentences or texts as inputs and CLS representation

from BERT feeds an output layer. In addition, BERT is finetuned for other challenging NLP tasks such as the natural language understanding tasks on The General Language Understanding Evaluation (GLUE) benchmark [59], question answering on The Stanford Question Answering Dataset (SQuAD v1.1) [53], and the task of grounded commonsense inference on The Situations With Adversarial Generations (SWAG) [60]. BERT achieves state-of-the-art results on these tasks.

2.4.2. BioBERT

BioBERT is a BERT-based language representation model that is pretrained on large-scale biomedical corpora. BioBERT uses the same architecture of BERT and is initialized with the weights of BERT trained on English Wikipedia (Wiki), BookCorpus (Books). However, since BERT is trained in the general domain, its performance can be improved for several biomedical NLP tasks considering domain-specific nouns and terms on this domain. Therefore, it is pretrained on biomedical corpora such as PubMed abstracts (PubMed) and PubMed Central full-text articles (PMC) in order to have more adapted language representation on the biomedical domain. There are several pretraining strategies with different combinations of general domain corpora and biomedical domain corpora. Initially, BioBERT uses the initial weights of $\text{BERT}_{\text{BASE}}$ pretrained on English Wikipedia and BooksCorpus. BioBERT v1.0 (+ PubMed + PMC) is a version of BioBERT pretrained on both PubMed and PMC, whereas BioBERT v1.0 (+ PubMed) is pretrained on only PubMed and BioBERT v1.0 (+ PMC) is pretrained on only PMC. BioBERT v1.1 (+ PubMed) is pretrained on only PubMed with more steps than BioBERT v1.0 (+ PubMed). For tokenization, BioBERT uses WordPiece tokenization [57] that handles the out-of-vocabulary problem by representing a new word with frequent known sub-words in the vocabulary. BioBERT uses the same vocabulary of $\text{BERT}_{\text{BASE}}$ due to the compatibility of BioBERT with BERT and the ability to represent a new word in the biomedical domain from the original one of BERT.

To reveal the effectiveness of context-dependent language representation, BioBERT is separately finetuned on three different biomedical natural language processing tasks such as named entity recognition, relation extraction, and question answering. For an example study of relation extraction stated in [43], target entities are replaced with predefined tags such as @GENE$ or @DISEASE$ as a preprocessing step. Pretrained BioBERT utilizes a CLS token for the relation classification in a sequence in the same way as the sentence classifier of BERT. The sentence classifier of BERT is trained using a single output layer with a CLS token representation from BERT. BioBERT with the output layer is finetuned on preprocessed sentences with different hyperparameters. BioBERT outperforms previous models on relation extraction for biomedical data sets as recorded in [43].

2.4.3. SciBERT

SciBERT [49] is a pretrained language model based on BERT and trained on a large corpus of scientific papers collected from the full-text of the papers from the computer science domain and the biomedical domain. SciBERT is initialized with the weights of $\text{BERT}_{\text{BASE}}$ and uses its same architecture and configuration. However, unlike BioBERT, SciBERT uses SCIVOCAB which is constructed as a new WordPiece vocabulary on the scientific corpus. It produces cased and uncased vocabularies which are trained on lower-cased texts and original texts respectively. Therefore, there are four pretraining strategies with different combinations of the case of texts and vocabularies used. The first two models use BASEVOCAB built by $\text{BERT}_{\text{BASE}}$, whereas the last two models use SCIVOCAB.

To reveal the effectiveness of unsupervised pretraining on a large corpus of scientific papers and to observe the improvement of its performance on scientific NLP tasks, SciBERT is separately finetuned on several biomedical natural language processing tasks such as named entity recognition, relation classification, and dependency parsing. For relation extraction stated in [49], target entities are encapsulated by special tokens as a preprocessing step on ChemProt. Like BioBERT, a linear classification

layer is fed with SciBERT output of CLS token that is the fixed-length relation representation for a sentence. SciBERT with the output layer is finetuned on preprocessed sentences with different hyperparameters. SciBERT outperforms previous models for relation extraction on ChemProt data set. In summary, SciBERT with in-domain vocabulary has an effect on various natural language processing tasks in the scientific domain.

3. MATERIALS AND METHODS

3.1. Data Set

We used ChemProt corpus which was released for a contest in BioCreative VI: Chemical-Protein Interaction (ChemProt) Track [61]. The contest aims to extract the chemical-protein interactions (CPI) from scientific documents.

ChemProt has three different sets: training set, development set, and gold standard test set. For each set with the same format, ChemProt has three different files: abstracts, entities, and relations. For the test set, the abstracts and the entities files are provided in order to predict ChemProt relations, whereas the relations file is released to evaluate extracted relations during the prediction of relations for given entities.

Abstracts file contains UTF8-encoded set of PubMed records in a tab-separated format with three columns: Article identifier (PMID, PubMed identifier), a title of the article, and an abstract of the article. In this data set, PubMed abstracts are collected into training (1020 abstracts), development (612 abstracts), and test sets (800 abstracts).

Entities file includes the manually annotated chemical compounds and genes/protein (gene and protein related objects - GPRO as defined during BioCreative V). This file has entity records in a tab-separated format with six different fields: Article identifier (PMID), entity number of the entity mention, type of the entity mention, start character offset of the entity mention, end character offset of the entity mention, and text string of the entity mention. In this data set, there are three entity mention types: *CHEMICAL* is the chemical entity mention type, *GENE-Y* is the gene/protein entity mention type that can be normalized to a biological database identifier, and *GENE-N* is the gene/protein entity mention type that can not be normalized.

Relations file comprises chemicals, gene/proteins, and their annotated relations. This file has relation records in a tab-separated format with six different fields: Article identifier (PMID), Chemical-Protein relation (CPR) group, evaluation type, ChemProt relation (CPR), interactor argument 1 (Arg1: interactor term identifier), and interactor argument 2 (Arg2: interactor term identifier).

Since ChemProt Track focuses on a subset of key relation types that are important in biochemical and biomedical perspective, all annotated ChemProt relations (CPR) are grouped into 11 distinct Chemical-Protein relation (CPR) groups for CPIs in the data set and each CPI has a relation type from one of 11 CPR. However, only five groups which are CPR:3, CPR:4, CPR:5, CPR:6, and CPR:9 are evaluated in the track. Therefore, five groups are labeled with Y in order to consider them during the track evaluation, whereas the other six groups are labeled with N in the evaluation type column of the relations file. Table 3.1 provides information about relation groups in the data set [61].

ChemProt Track focuses on the accurate predictions of relations between only chemicals and genes/proteins. Relations between a chemical and another chemical or a gene/protein and another gene/protein are not evaluated in this data set. In addition, since there are rare entity pairs labeled with multiple relations (the same or different relation groups) in the same abstract, multiple relations for a given entity pair can be extracted from this data set. The training set has 98 chemical and protein entity pairs that have more than one relation type in the same abstract in the relations file. The development set has 86 chemical and protein entity pairs that have more than one relation label in the same abstract in the relations file.

Table 3.2 contains recurring relation numbers in the training and the development sets. In the training set, while 42 out of 98 have the same relation groups, others have two different relation groups for the same chemical and protein entity pairs in the same abstract. The development set has 24 relations out of 86 have the same relation groups. In addition, Table 3.3 illustrates duplicate relations for the same chemical and

Table 3.1. Grouped based on biological semantic classes. Only 5 groups (CPR:3, CPR:4, CPR:5, CPR:6, CPR:9) are labeled with *Y* in the evaluation column for ChemProt Track evaluation purpose.

Group	**Eval.**	**ChemProt relations**
CPR:0	N	UNDEFINED
CPR:1	N	PART_OF
CPR:2	N	REGULATOR | DIRECT_REGULATOR | INDIRECT_REGULATOR
CPR:3	Y	UPREGULATOR | ACTIVATOR | INDIRECT_UPREGULATOR
CPR:4	Y	DOWNREGULATOR | INHIBITOR | INDIRECT_DOWNREGULATOR
CPR:5	Y	AGONIST | AGONIST-ACTIVATOR | AGONIST-INHIBITOR
CPR:6	Y	ANTAGONIST
CPR:7	N	MODULATOR | MODULATOR-ACTIVATOR | MODULATOR-INHIBITOR
CPR:8	N	COFACTOR
CPR:9	Y	SUBSTRATE | PRODUCT_OF | SUBSTRATE_PRODUCT_OF
CPR:10	N	NOT

protein entity pairs in the sentences selected from the training set. In the first example sentence, there are two relation records with the same relation group (CPR:2) in the relations file for *triglycerides* and *PTP1B*. In the second example sentence, *cyclin E* and *Tomudex* have two relation records with two different relation groups (CPR:3, CPR:4) in the relations file for the same sentence.

Table 3.2. Duplicate relation statistics in the training and the development sets.

Relation Groups	**The number of relations in the training set**	**The number of relations in the development set**
CPR:2	30	29
CPR:3	27	11
CPR:4	21	21
CPR:5	2	1
CPR:6	7	10
CPR:8	2	0
CPR:9	7	14
CPR:10	2	0

Statistics of abstracts, entities, and relations are directly calculated from given files by ChemProt data set. Table 3.4 includes the statistical information of relation groups in ChemProt. On the other hand, since abstracts are split into sentences via GENIA [62], some relations in the relations file are filtered out due to the incorrect sentence splitting from GENIA. In addition, Table 3.4 shows the number of relation groups for the training, the development, and the test sets separately. We created three figures to explain the association between abstracts, entities, and relations file in the data set. Figure A.1 illustrates an example abstract in which its chemical entities and its protein entities are colored in blue and yellow, respectively. Figure A.2 illustrates the entities file given in the data set. The first column contains unique IDs of abstracts, while the second column includes a unique entity number for each entity in an abstract. An entity type for each entity mention is recorded in the third column. The fourth

Table 3.3. Examples of duplicate relations for the same chemical and protein entity pairs in the training and the development sets.

Article id	**Sentence**	**Relation Groups**
23481236	We studied accumulation of lipid metabolites [*triglycerides* (TAGs), diglycerides (DAGs)] and ceramides in relation to insulin signaling and expression and phosphorylation of *PTP1B* by preincubating rat skeletal muscle cells (L6 myotubes) with three saturated and three unsaturated free fatty acids (FFAs) (200 μM).	CPR2 & CPR2
10047461	The studies with dThyd rescue from *cyclin E*-cdk2 protein overexpression and growth inhibition by *Tomudex* indicate that increased cyclin E-cdk2 protein expression is associated with effective inhibition of thymidylate synthase and resultant dNTP pool imbalance.	CPR:3 & CPR:4
10947967	In contrast, *2,4-dioxo-5-acetamido-6-phenylhexanoic acid*, which is a competitive inhibitor with respect to ascorbate, exhibits a low degree of stereospecificity in binding to the ascorbate sites of both *PAM* and dopamine-beta-hydroxylase.	CPR:2 & CPR:4

column consists of the start offset of an entity, whereas the fifth column has the end offset of the entity in the abstract. The biomedical names of entities in abstracts are written in the last column. Figure A.3 illustrates the relations file given in the data set. The first column contains unique IDs of abstracts, whereas the second column includes the relation groups between proteins and chemicals in the entities file. The third column indicates the evaluation type of relations. ChemProt relations are written in the fourth column. The fifth column and the sixth column have entity numbers for interactor argument 1 and entity numbers for interactor argument 2 where argument 1 is a chemical and argument 2 is a gene/protein in relation, respectively.

Thanks to three figures, we clearly explain how to use three files in the data set in order to identify the relations between entities in abstracts. For example, *Estrogens* and *CYP450*, a chemical and a protein/gene respectively are in the second sentence of the abstract as shown in Figure A.1. Figure A.2 shows that the entity numbers of *Estrogens* and *CYP450* are T1 and T22 considering the start and end offsets of entities in the abstract, respectively. In Figure A.3, *Estrogens* is encoded as Arg1:T1 due to being a chemical, while *CYP450* is recorded as Arg2:T22 on account of being a protein/gene. For the specified entities, the fifth row of the figure has a relation record as PRODUCT-OF called CPR:9.

3.2. Preprocessing

Preprocessing steps involve sentence splitting, chemical and protein entity pair construction, adding entity markers, and relation label assignments for chemical and protein entity pairs.

Firstly, since we consider the relations that are present in a single sentence, we split the abstracts into sentences via GENIA Sentence Splitter [63]. In a sentence, we construct chemical and protein entity pairs which are labeled according to the relation annotations in the relations file. If there is no relation annotation for the specified entities in the file, the relation type of the entity pair is labeled as *0ther*.

Table 3.4. ChemProt summary statistics for training, development (dev), and test sets. The entity pairs are annotated with 11 relation groups in ChemProt and we refer to pairs whose relation information cannot be inferred from the context as *Other*.

Statistics	**Train Set**	**Dev Set**	**Test Set**
# abstracts	1020	612	800
# relations	18046	11294	15712
# CPIs	6437	3558	5744
# entities	25752	15567	20828
# chemical entities	13017	8004	10810
# unique chemical entities	3710	2517	3442
# protein entities	12735	7563	10018
# unique protein entities	4610	3018	3757
# duplicate entity pairs in relations	98	86	158
# of CPR:0	1	2	0
# of CPR:1	308	153	215
# of CPR:2	1652	780	1743
# of CPR:3	777	552	667
# of CPR:4	2260	1103	1667
# of CPR:5	173	116	198
# of CPR:6	235	199	293
# of CPR:7	29	19	25
# of CPR:8	34	2	25
# of CPR:9	727	457	644
# of CPR:10	241	175	267
# of *Other*	11664	7780	9987

Secondly, we focus on both multi-class relation extraction and binary relation extraction from sentences throughout this study. Therefore, although sentence splitting and entity pair construction are applied in the same way for both binary and multi-class relation extraction tasks, there are two different relation label assignments for entity pairs in two tasks.

3.2.1. Sentence Splitting

We split the abstracts into sentences via GENIA Sentence Splitter [63], since we extract the sentence-based relations between chemicals and proteins. GENIA Sentence Splitter is a widely used sentence splitter for biomedical texts. It is a classification model trained on GENIA Corpus [62] with a supervised learning method. After it identifies the candidate positions to split the text considering the several delimiters such as periods, commas, quotation marks, and parentheses, it decides whether each candidate has a real sentence splitter position. Apart from the delimiters, it takes into account several features such as capitalization of previous and next words, the existence of commas in previous and next words in order to have better sentence splitting.

3.2.2. Chemical and Protein Entity Pair Construction

After an abstract is split into sentences, the start and end offsets of each sentence in the abstract are determined. Given the start and end offsets of each entity in the abstract from the entities file in ChemProt, we decide whether protein/gene or chemical entities exist in the sentence. If a sentence contains at least a chemical and a protein/gene, it has a candidate relation between a chemical and a protein/gene for the model training and the prediction.

A chemical and a protein/gene in a sentence is called “Chemical and Protein Entity Pair”. We determine whether there is a relation between them according to the given relation labels in the relations file in ChemProt. If any chemical and protein entity pair in a sentence does not have a relation label in the relations file, its label is

assigned as *Other*.

In Figure 3.1, four sentences are given as examples in an example abstract. The first sentence has a chemical (*Tomudex*) and four proteins/genes (*p27(kip1)*, *cyclin E*, *cdk2*, and *kinase*). *Tomudex* and four different proteins/genes construct four different "Chemical and Protein Entity Pairs" in this example. Chemical and protein entity pairs constructed are colored in yellow in the figure. Each chemical and protein entity pair has a label (CPR:0 to CPR:9 or Other).

3.2.3. Entity Markers

We surrounded the target entities (chemicals and proteins) in the sentences with opening and closing tags to mark their location, following [64]. We tagged chemicals with <e1> and </e1> and proteins with <e2> and </e2> to encode entity type and location during training. When a sentence had multiple chemical - protein pairs, we considered each pair separately and created copies of the sentence with different tags. Table 3.5 provides an example of input and its two preprocessed forms. In the example, the raw sentence contains two chemicals (*gefitinib*, *erlotinib*) and one protein (EGFR), creating two protein - chemical pairs. Thus, we create two different forms of the sentence to encode each protein - chemical pair separately. We use <e1> and <e2> tags to enclose chemicals and proteins, respectively.

3.2.4. Relation Label Assignment for Binary Relation Extraction

The entity pairs with the first 10 relation groups (CPR:0 to CPR:9) are considered positive samples and labeled as 1 in binary relation extraction. On the other hand, CPR:10 is saved for *not relation*, since sentences indicate there is no relation between the pair of entities. Thus, the entity pairs in CPR:10 and Other class are considered negative samples. These entity pairs are labeled as 0 in the binary relation extraction task. Table 3.6 illustrates the relation label assignment for the binary relation extraction task. In our examples, an abstract with 14967461 PMID has *gefitinib*

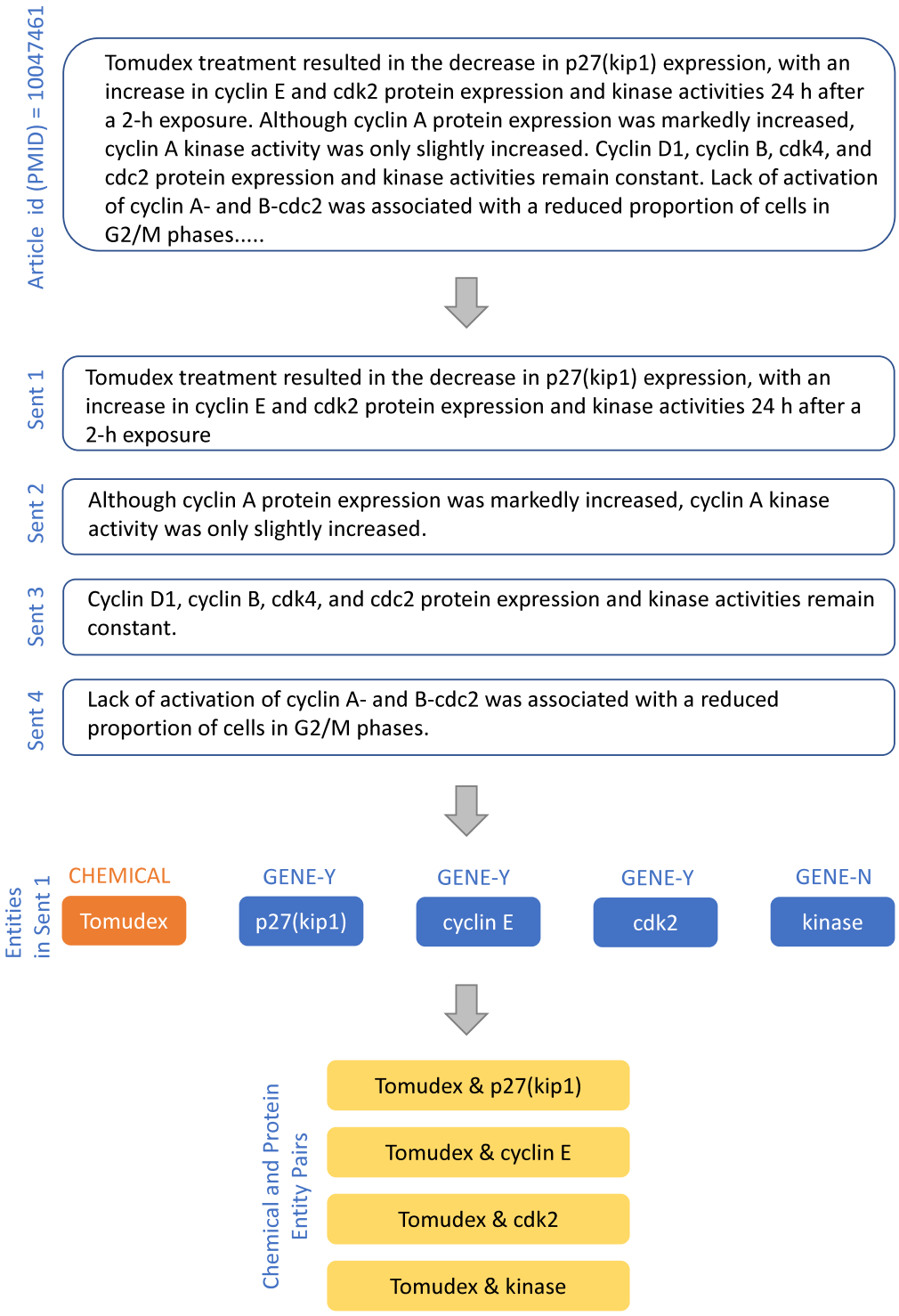

Figure 3.1. Examples of chemical and protein entity pair construction in an abstract with 10047461 PMID. Chemical (orange) and protein (blue) entity pairs constructed are colored in yellow.

Table 3.5. An example of preprocessing. We use <e1> and <e2> tags to enclose chemicals and proteins, respectively.

Raw Sentence	EGFR inhibitors currently under investigation include the small molecules gefitinib and erlotinib.
Preprocessed Form I	<e2>EGFR</e2> inhibitors currently under investigation include the small molecules <e1>gefitinib</e1> and erlotinib.
Preprocessed Form II	<e2>EGFR</e2> inhibitors currently under investigation include the small molecules gefitinib and <e1>erlotinib</e1>.

and *Epidermal growth factor receptor* as a chemical and a protein, respectively. These entities have a relation named CPR:4 in the relations file. Since entities in the first 10 relation groups (CPR:0 to CPR:9) are considered positive samples, the relation of the entities is labeled as 1 in the binary relation extraction task. On the other hand, the relations file has no record about the relations between *mitiglinide* and *SUR2A* in an abstract with 11716850 PMID. Therefore, the binary label of these entities is set 0 in the task.

Table 3.6. Examples of relation label assignment in binary relation extraction.

Article ID	**Chemical**	**Protein**	**Relation Group**	**Label**
14967461	gefitinib	Epidermal growth factor receptor	CPR:4	1
7678677	Alprenolol	beta 1-adrenoceptors	CPR:6	1
11511858	androgen	nuclear receptors	CPR:7	1
11716850	mitiglinide	SUR2A	CPR:10	0
16357751	methotrexate	tumor necrosis factor	Other	0

After the label assignment for chemical and protein entity pairs, duplicate entity pairs and their binary relation labels are arranged by the following strategy:

- If at least one of the relation groups for the same pair of entities in a sentence is one of the first 10 relation groups (CPR:0 to CPR:9), the relation group is kept once for the specified entity pair and assigned its relation label 1.

Table 3.7 includes the statistics of binary relation labels for binary relation extraction. As mentioned in sentence splitting, since abstracts are split into sentences via GENIA, some of the relations that are not in a sentence are filtered out. In this study, the relations in the same sentence with the same protein, chemical, and relation type are denoted as duplicate relations. After they are involved in training only once, we observe that negative entity pairs are more frequent than positive entity pairs. However, the training, the development, and the test sets have a similar distribution of relation labels.

Table 3.7. Positive and negative relation label statistics for binary relation extraction in ChemProt. Even though negative entity pairs are more frequent than positive ones, a similar distribution of relation labels exists in three data sets.

Statistics	**Train Set**	**Dev Set**	**Test Set**
# positive labels	6143	3339	5459
# negative labels	11903	7955	10253

3.2.5. Relation Label Assignment for Multi-class Relation Extraction

The relations of entity pairs in CPR:3, CPR:4, CPR:5, CPR:6, and CPR:9 are tagged identically to their relation groups in ChemProt for multi-class relation extraction. Since BioCreative VI: Chemical-Protein Interaction (ChemProt) Track does not evaluate the entity pairs in CPR:0, CPR:1, CPR:2, CPR:7, CPR:8, CPR:10, and *Other*, the entities are considered as joint samples and labeled as *CPR:X* in multi-class relation extraction. Table 3.8 illustrates the relation label assignment for multi-class relation extraction. In our examples, an abstract with 14967461 PMID has *gefitinib* and *Epidermal growth factor receptor* as a chemical and a protein, respectively. These

entities have a relation named CPR:4 in the relations file. Since the pairs of entities have relations in the first 10 relation groups (CPR:0 to CPR:9), their relation groups of the pairs are directly used as their relation labels in the multi-class relation extraction task. On the other hand, an abstract with 11716850 PMID has a CPR:10 relation with *mitiglinide* as a chemical and *SUR2A* as a protein. However, since CPR:10 is not in 5 relation groups that are evaluated in BioCreative VI: Chemical-Protein Interaction (ChemProt) Track, the relation labels of the entity pairs in CPR:10 are considered *CPR:X* in multi-class relation extraction.

Table 3.8. Examples of relation label assignment in multi-class relation extraction.

Article ID	**Chemical**	**Protein**	**Relation Group**	**Label**
14967461	gefitinib	Epidermal growth factor receptor	CPR:4	CPR:4
7678677	Alprenolol	beta 1-adrenoceptors	CPR:6	CPR:6
11511858	androgen	nuclear receptors	CPR:7	CPR:X
11716850	mitiglinide	SUR2A	CPR:10	CPR:X
16357751	methotrexate	tumor necrosis factor	Other	CPR:X

After the label assignment for chemical and protein entity pairs, duplicate entity pairs and their multi-class relation labels are arranged by the following strategy:

- If each of the relation groups for the same pair of entities in a sentence is one of CPR:3, CPR:4, CPR:5, CPR:6, and CPR:9, all relation groups with the specified entity pair are directly considered as relation labels.
- For the same pair of entities in a sentence, their relation groups can be one of the 5 relation groups in ChemProt evaluation (CPR:3, CPR:4, CPR:5, CPR:6, and CPR:9) as well as CPR:0, CPR:1, CPR:2, CPR:7, CPR:8 and Other. In this case, the specified pair of entities and their relation groups from 5 evaluated relation groups are kept and the others are removed from the sample set.
- If all relation groups for the same pair of entities in a sentence are any of CPR:0,

CPR:1, CPR:2, CPR:7, CPR:8, and Other, the specified pair of entities are retained only once and their relation label is assigned as *CPR:X*.

As a summary of the relation label assignments, Table 3.9 denotes the binary and multi-class relation extraction labels for the entity pairs in ChemProt.

Table 3.9. Binary and multi-class relation labels for Chemical - Protein Relations (CPR) in ChemProt.

Relations	**Binary Label**	**Multi-class Label**
CPR:0	1	CPR:X
CPR:1	1	CPR:X
CPR:2	1	CPR:X
CPR:3	1	CPR:3
CPR:4	1	CPR:4
CPR:5	1	CPR:5
CPR:6	1	CPR:6
CPR:7	1	CPR:X
CPR:8	1	CPR:X
CPR:9	1	CPR:9
CPR:10	0	CPR:X
Other	0	CPR:X

3.3. Sentence-based Input Representation

This study considers sentences having at least one chemical and protein entity pair, since it aims to extract the relations between target entities in sentences. Therefore, we ignore ones not having at least one chemical and protein entity pair. After abstracts are processed as mention in Section 3.2, chemical and protein entity pairs in sentences are constructed. Chemicals and proteins in sentences are marked with

special tags. In the end, sentences and their entity pairs have only one relation label. Sentences with the special tags and their relation labels feed different types of relation extraction models. Table 3.10 illustrates the input representation using entire sentences.

Table 3.10. Example inputs generated using entire sentences.

Preprocessed Sentences
<e1> Alprenolol </e1> and BAAM at 10(-7), 3 x 10(-7), and 10(-6) M inhibited the cardiac stimulation response slightly, which is indicative of membrane-stabilizing activity independent of <e2> beta-adrenoceptor </e2> blockade.
Dopamine D(2) receptor-induced <e2> COX-2 </e2> -mediated production of <e1> prostaglandin E(2) </e1> in D(2)-transfected Chinese hamster ovary cells without simultaneous administration of a Ca(2+)-mobilizing agent.

3.4. Dependency Tree-based Input Representations

As with sentence-based input representation, we consider sentences having at least one chemical and protein entity pair in these representations. We create two different dependency tree-based input representations: (i) Shortest Subsequence of Chemical and Protein Entity Pair, (ii) Shortest Subsequence Including Chemical and Protein Entity Pair and Their Parent Node. Firstly, we provide an overview of the background of the dependency parser and tree.

3.4.1. Dependency Parser and Tree

Dependency grammars are grammar formalisms that are based on dependency relations between pairs of words in a sentence. Words and sets of directed grammatical relations among the words describe the syntactic structure of a sentence. Relations among words in a sentence are represented by the labeled or unlabeled arcs from heads

to dependents. Since labels represent fixed grammatical relations among words, this kind of grammatical description for the syntactic structure of a sentence is called a typed dependency structure [65].

The notion of grammatical relation provides the basis for the binary relations that involve dependency structures. The arguments of a relation comprise a head and a dependent. Since the head word of a constituent is a central word, the remaining words in the constituent are dependents on their head. Figure 3.2 illustrates dependency analysis using graphical way in the dependency parsing. Whereas nodes in the dependency parsing stand for the words, arcs represent the binary dependency relations between words. A root node implies the head of the whole structure.

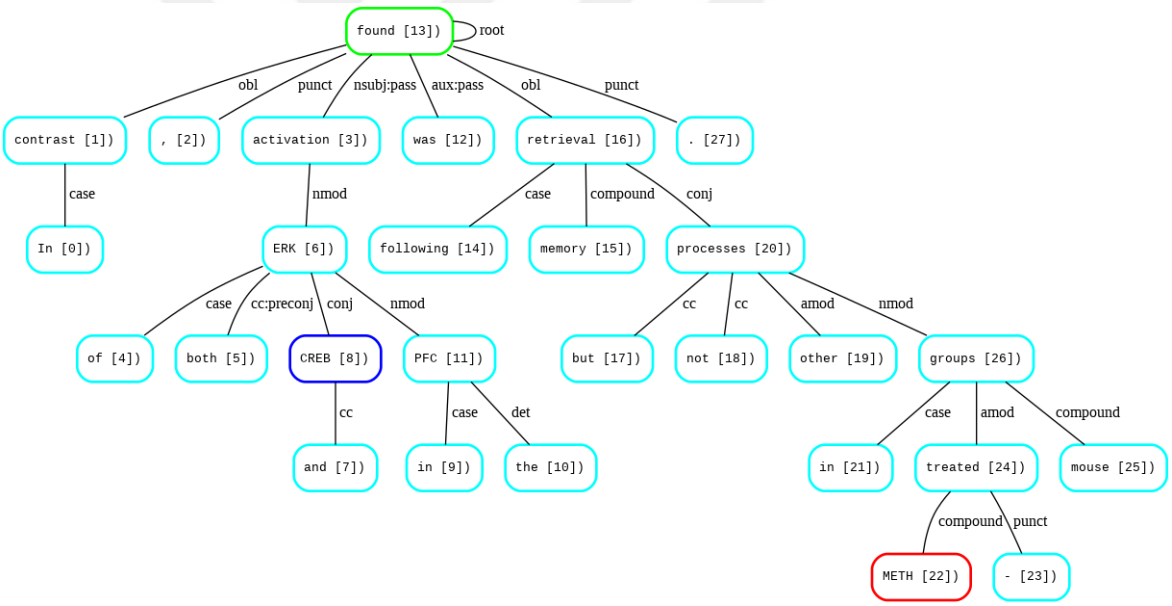

Figure 3.2. Dependency tree of a sentence in ChemProt. We used Stanza library [66] derived from The Stanford CoreNLP Natural Language Processing Toolkit [67] to parse a sentence. The green node indicates the root node of the sentence in the tree.

Dependency structures are acyclic directed graphs whose structure are G = (V, A) including a set of vertices V and a set of arcs A [68]. Vertices correspond to words in a sentence, while arcs correspond to grammatical relations between the vertices in V. The graph is called Dependency Tree under the following constraints [65]:

- The dependency tree includes the entire input.
- There is a single root node having no incoming arcs.
- There is no cycle between any two nodes.
- Apart from the root node, each vertex has only one incoming arc. In other words, each vertex has only one parent node.
- Each vertex is reachable from the root node by following a unique path.

Dependency trees have been widely used in a range of NLP tasks such as named entity recognition [69], sentiment classification [70], and question answering [71], [72]. In relation extraction, dependency tree patterns of target entities are used as features that feed different machine learning algorithms [73] and neural network architectures [74]. In biomedical relation extraction, since sentences of biomedical texts are long and complex, studies benefit from dependency trees of sentences so as to obtain helpful and understandable relations between words in sentences with annotated edges [75].

3.4.2. Shortest Subsequence Between Chemical and Protein Entity Pair

As biomedical texts generally have long and complex sentences, they may contain some words that do not contribute to extracting relation information from sentences. Due to the redundancy in sentences, relation extraction models may not capture the relation information from sentences. We investigate whether relation extraction models have higher results with simplified input sequences. For this purpose, we simplify a sentence by taking the shortest subsequence that includes two nodes of target entities in the dependency tree. Nodes linked with target entities are traced up to extracting the minimal subtree in the dependency tree. Other tokens out of the subsequence are removed from the input sentence. We apply the following steps:

- Find the shortest path between specified entities
- Get a minimum and a maximum node id in the shortest path
- Take the nodes having ids between the minimum and the maximum node id

Table 3.11 illustrates the input representation derived from the shortest subsequence of chemical and protein entity pair. Dependency trees of our examples are presented in the Appendix A. For the first example in the table, *dicoumarol* is a chemical with node id 6 and *NQO1* is a protein/gene with node id 13 in Figure B.1. The shortest path between the entities has nodes with ids of 6, 10, 12, and 13. The minimum id is 6, whereas the maximum id is 13 in the shortest path. Finally, we take the nodes with ids of 6, 7, 8, 9, 10, 11, 12, and 13 inside dashes in the figure. At the end, we have a shortest subsequence between entities with an important keyword which is *inhibitor* in this example. For the second sentence in the table, *AMPK* and *malonyl CoA* are target entities that are a protein/gene and a chemical, respectively in the sentence from an abstract with 16642960 PMID. In Figure B.3, the node of *AMPK* is colored in blue and the node of *malonyl CoA* is colored in red in the dependency tree of the sentence. Although the node of *AMPK* is the parent node of *malonyl CoA*, a subsequence in the sentence can not be created with only these two nodes. These two nodes and the child nodes of *malonyl CoA* (*or, decrease, and, /*) construct the shortest subsequence of the sentence. For the last sentence in the table, target entities are *AspRSs* (chemical) and *Gly* (protein/gene) in the sentence from an abstract with 165997625 PMID. As shown in Figure B.4, in the dependency tree of the sentence, *AspRSs* and *Gly* are colored in blue and red, respectively. First of all, since *allowed* and *placing* are in the path between two target entities, the shortest subsequence includes two of them as well as target entities. Although *Gly* is a parent node of *269* and *position* is the child node of *placing*, two different sub-trees that have parent nodes as *position* and *269* do not contribute the shortest subsequence of the entities. The reason for this is that *AspRSs*, *allowed*, *placing*, and *Gly* are sufficient for the shortest path between the target entities. Other nodes are removed from the sentence.

3.4.3. Shortest Subsequence Including Chemical and Protein Entity Pair and Their Parent Node

We trim a sentence by taking the shortest subsequence that includes not only two target entities but also their common parent node in its dependency tree. Firstly,

Table 3.11. Example inputs generated by the shortest subsequence between a chemical and protein entity pair in a sentence.

Preprocessed Sentences	**The Shortest Subsequence Between Target Entities**
The mNQO activity was insensitive to <e1> dicoumarol </e1> , a potent inhibitor of cytosolic <e2> NQO1 </e2> .	<e1> dicoumarol </e1> , a potent inhibitor of cytosolic <e2> NQO1 </e2>
The possibility is also raised that pharmacological agents and other factors that activate <e2> AMPK </e2> and/or decrease <e1> malonyl CoA </e1> could be therapeutic targets.	<e2> AMPK </e2> and / or decrease <e1> malonyl CoA </e1>
<e2> Cyclooxygenase </e2> (COX)-2 and membrane-bound <e1> prostaglandin E </e1> synthase-1 (mPGES-1) were induced by treatment with lipopolysaccharide (LPS) but not with PS liposomes.	<e2> Cyclooxygenase </e2> (COX) - 2 and membrane - bound <e1> prostaglandin E </e1> synthase
Sequence alignments of various <e2> AspRSs </e2> allowed placing <e1> Gly </e1> -269 at a position occupied by Asp-220, the residue contacting G73 in the crystallographic structure of E. coli AspRS-tRNA(Asp) complex.	<e2> AspRSs </e2> allowed placing <e1> Gly </e1>

two nodes of target entities in a sentence are involved in the shortest subsequence, as explained in Section 3.4.2. Then, the parent of this sub-tree is marked. We extend the previous approach by adding the parent node id. We find the minimum and maximum node id in the shortest path and the parent node. Other steps are directly applied in the same way as the previous approach. Finally, a subsequence that includes their parent node and the shortest sequence of target entities is determined in the dependency tree of a sentence. We remove the remaining tokens out of the new subsequence from the input sentence.

We explain this approach with the following examples with their dependency trees presented in the Appendix A. For the first example in Table 3.12, *stathmin* and *CCNU* are target entities, while *inhibition* is the parent node. By considering the parent node, we take *inhibition* keyword in the dependency tree inside dashes in Figure B.2. For the second sentence in Table 3.12, *activate* is the common parent node of *AMPK* and *malonyl CoA* in the dependency tree, as shown in Figure B.3. The shortest subsequence of target entities is constructed as in Section 3.4.2. In addition to it, a new subsequence consists of *activate* as a parent node and the shortest subsequence containing *AMPK*, *and*, */*, *or*, *decrease*, and *malonyl CoA*. We remove the remaining linked nodes out of the new subsequence from the sentence. By adding the parent node into the shortest sequence, we examine the importance of the parent node in the relation representation. For the last sentence in Table 3.12, *alignments* is the common parent node of *AspRSs* and *Gly* in the dependency tree, as shown in Figure B.4. The shortest sequence of target entities forms of *AspRSs*, *allowed*, *placing*, and *Gly* as explained in Section 3.4.2. A new subsequence starts with *alignments* as a common parent node. The sequence continues with its minimal sub-tree having the shortest subsequence of target entities. Therefore, it involves *of* and *various* that are linked to *AspRSs* in order to obtain a part of the sentence. Finally, we ignore two different sub-trees that begin with the node of *269* and the node of *position*.

Table 3.12. Example inputs generated by the shortest subsequence, including target entities and their parent node in a sentence.

Preprocessed Sentences	**The Shortest Subsequence Including Target Entities and Their Parent Node**
The direct inhibition of <e2> stathmin </e2> by <e1> CCNU </e1> is likely a contributing factor.	inhibition of <e2> stathmin </e2> by <e1> CCNU </e1>
The possibility is also raised that pharmacological agents and other factors that activate <e2> AMPK </e2> and/or decrease <e1> malonyl CoA </e1> could be therapeutic targets.	activate <e2> AMPK </e2> and / or decrease <e1> malonyl CoA </e1>
<e2> Cyclooxygenase </e2> (COX)-2 and membrane-bound <e1> prostaglandin E </e1> synthase-1 (mPGES-1) were induced by treatment with lipopolysaccharide (LPS) but not with PS liposomes.	<e2> Cyclooxygenase </e2> (COX) - 2 and membrane - bound <e1> prostaglandin E </e1> synthase - 1 (m PGES - 1) were induced
Sequence alignments of various <e2> AspRSs </e2> allowed placing <e1> Gly </e1> -269 at a position occupied by Asp-220, the residue contacting G73 in the crystallographic structure of E. coli AspRS-tRNA(Asp) complex.	alignments of various <e2> AspRSs </e2> allowed placing <e1> Gly </e1>

3.5. Relation Extraction

Our Chemical-Protein Interaction Classification is a task of whether chemical and protein entities in a sentence have a specified relation in ChemProt data set. Throughout this study, we extract the relation information of target entities from sentences. The relations of target entities in ChemProt are categorized into 11 distinct relation groups. The first 10 relation groups (CPR:0 to CPR:9) indicate the diverse types of biochemical relations. However, the last relation group (CPR:10) is declared as *not relation*, since the sentence directly expresses that there is no relation between the specified entity pairs.

In this study, Chemical-Protein Interactions Classification and Relation Extraction are used to state the same purpose. This study focuses on 2 different relation extraction tasks on ChemProt:

- Binary Relation Extraction aims to determine whether chemical and protein entities in a sentence are biochemically related. The entity pairs in the first 10 relation groups (CPR:0 to CPR:9) are treated as positive samples. The entity pairs in the last relation group (CPR:10) and Other are considered as negative samples, since the sentences for these two groups do not express that there is a relation between the target entity pairs. Therefore, relation label assignment for binary relation extraction is applied to the actual relation groups in ChemProt during training.
- Multi-class Relation Extraction aims to extract the specified relation types between chemical and protein entities in a sentence if there is a biochemical relation between them. The specified relation groups are CPR:3, CPR:4, CPR:5, CPR:6, and CPR:9 in The BioCreative VI track 5 ChemProt task. We treat the entity pairs in the remaining relation groups (CPR:1, CPR:2, CPR:7, CPR:8, CPR:10) and *Other* as joint samples. Thus, relation label assignment for multi-class relation extraction is applied to the actual relation groups in ChemProt during training.

This study focuses on the binary and multi-class relation extractions in ChemProt using transformer-based architectures. In this section, we give detailed information about our binary and multi-class relation extraction models.

Transformer-based approaches have been widely used and achieved the state-of-the-art results on relation extraction tasks [43], [49]. In this study, we use transformer-based models for both binary and multi-class relation extractions in ChemProt. Although the two relation extraction tasks have two different objectives and outputs, their solution designs are similar in our approaches.

We present the training pipeline of the relation extraction in Figure 3.3. The workflow of the relation extraction starts with splitting abstracts from ChemProt to sentences by GENIA Sentence Splitter [62]. Chemical and protein entity pairs in a sentence are constructed by given the locations of the entities in ChemProt. A sentence with a chemical and protein entity pair is labeled with its relation group from ChemProt. For the two different relation extraction tasks (binary and multi-class), relation label assignments are applied to specify the relation label. In addition, we insert predefined entity markers into sentences having chemical and entity pairs to identify the locations of the entities. Afterward, one of the three different input representations is selected. If we choose one of the dependency tree based input representations, the sentence is modified, explained in Section 3.4. Each input is tokenized by a tokenizer which is constructed by WordPiece vocabulary on different corpora. After tokenization, a transformer-based model is fine-tuned by learning the entity start and end markers from scratch and creates a fixed-length relation representation for the starting tag CLS. A single linear layer with softmax activation is trained with the vector of CLS tag in an input. The combined architecture predicts the relation label for the input. We calculate Cross-Entropy Loss as a loss function during the prediction of the relation label. An optimizer with its parameters is used to minimize the loss.

For both binary and multi-class relation extraction experiments, BioBERT and SciBERT are used as transformer-based models, since they are pretrained on biomedical

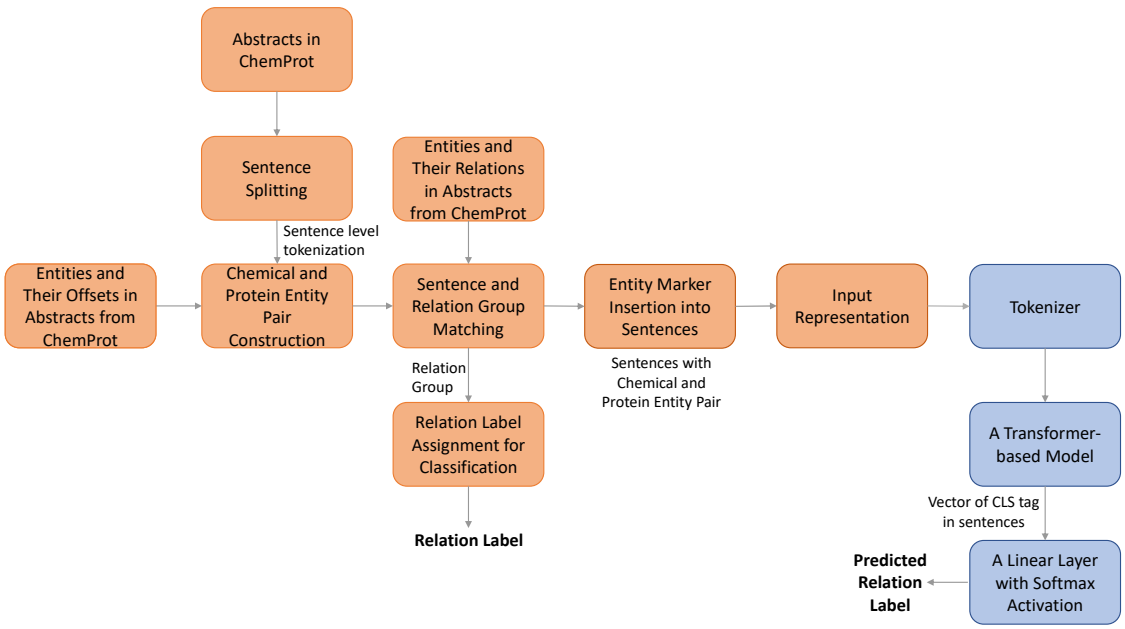

Figure 3.3. The training pipeline behind our transformer-based relation extraction model on ChemProt. Our preprocessing steps are colored orange. A Transformer-based model, its tokenizer, and an output layer are colored blue.

documents. When we conduct the experiments with BioBERT, a tokenizer is the same as a tokenizer in $\text{BERT}_{\text{BASE}}$. On the other hand, a tokenizer derived from SciVOCAB is used during the training of the relation extraction tasks with SciBERT. The linear layer with softmax activation has 2 output neurons for the binary relation extraction task, while it has 6 output neurons for the multi-class relation extraction task. Lastly, Cross-Entropy Loss is selected as a loss function to calculate the loss between the actual relation label and the predicted label. Technical details about BioBERT and SciBERT are as follows:

- The initial weights of BioBERT v.1.1 which has 12 layers, 768 hidden size, and 110M parameters are used. (accessed by `https://huggingface.co/monologg/biobert_v1.1_pubmed`)
- The initial weights of SciBERT which has 12 layers, 768 hidden size, and 110M parameters are used. (accessed by `https://huggingface.co/allenai/scibert_scivocab_cased`)

3.5.1. Binary Relation Extraction

A binary relation extraction model is composed of a transformer-based model and a binary classification layer. Inputs are tokenized by using WordPiece tokenization. Biochemical entities are enclosed with predefined tags (<e1> and <e2>). In addition, the transformer-based model uses *CLS* and *SEP* tags as fixed-length relation representation and sentence separation, respectively. The binary classification layer is trained with the vector of *CLS* token from the transformer-based model finetuned by adding entity markers. Figure 3.4 shows our binary relation extraction model trained with ChemProt data set. We use the training pipeline for relation extraction, shown in Figure 3.3, with the selected input representation. Among our all experiments with BioBERT and SciBERT, BioBERT finetuned by entire sentences with the following parameters outperforms our other binary relation extraction models:

- After the hyperparameter tuning, AdamW with learning rate 3e-5 and weight decay 0.3 is selected as the best parameters in our study.

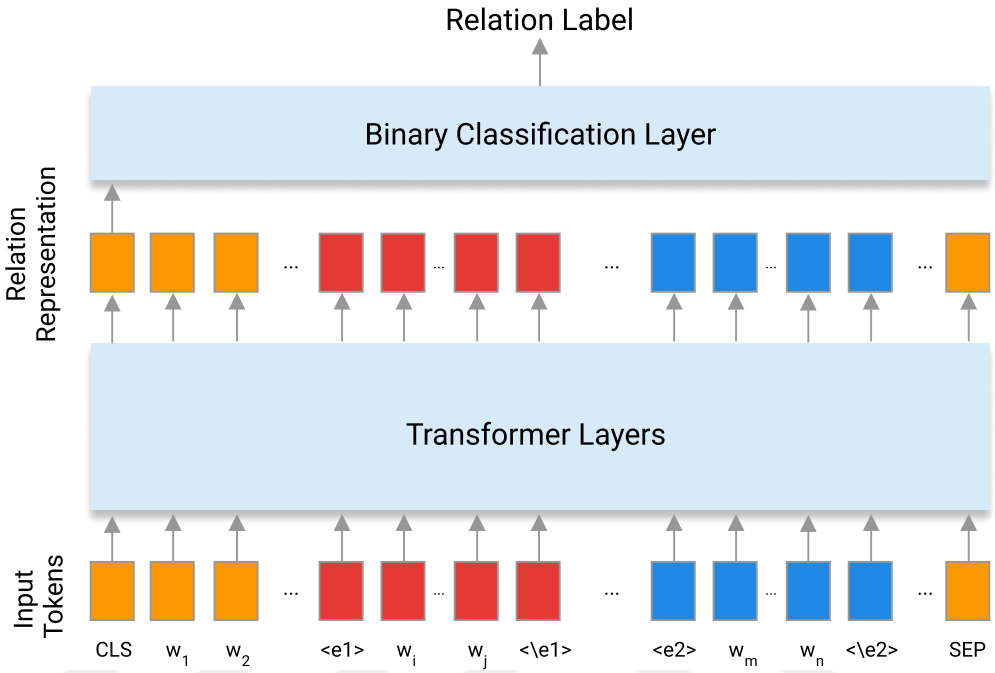

Figure 3.4. Binary Relation Extraction Model. The model is composed of a transformer-based model and a binary classification layer.

3.5.2. Multi-class Relation Extraction

Our multi-class relation extraction model is composed of a transformer-based model and a linear layer with softmax activation. Inputs are tokenized by using WordPiece tokenization. Biochemical entities are enclosed with predefined tags (<e1> and <e2>). In addition, the transformer-based model uses *CLS* and *SEP* tags as fixed-length relation representation and sentence separation, respectively. The linear classification layer is trained with the vector of *CLS* token from the transformer-based model finetuned by adding entity markers. Unlike the binary relation extraction model, the linear layer has 6 neurons, where there are 5 relation classes and 1 joint class. Relation groups CPR:3, CPR:4, CPR:5, CPR:6, and CPR:9 are considered relation classes, while we called the remaining relation groups (CPR:1, CPR:2, CPR:7, CPR:8, CPR:10) and Other as CPR:X. Figure 3.5 shows our multi-class relation extraction model. We use the training pipeline for relation extraction, shown in Figure 3.3, with the se-

lected input representation. Among our all experiments with BioBERT and SciBERT, BioBERT finetuned by entire sentences with the following parameters outperforms our other multi-class relation extraction models:

- After the hyperparameter tuning, AdamW with learning rate 3e-5 and weight decay 0.1 is selected as the best parameters in our study.

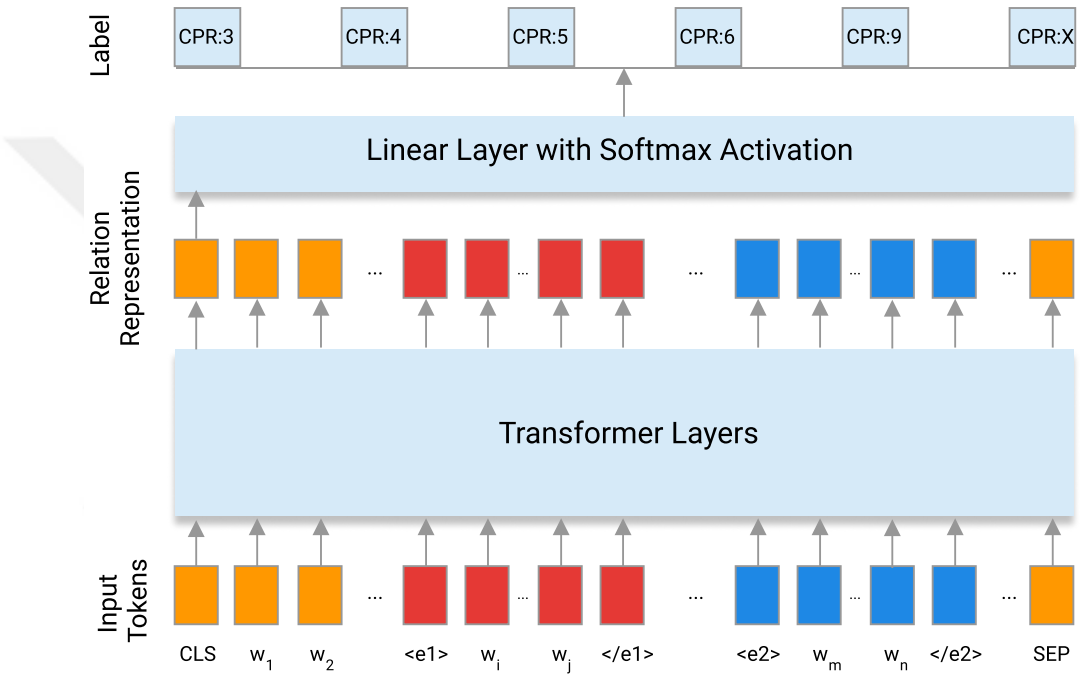

Figure 3.5. Multi-class Relation Extraction Model. The model is composed of a transformer-based model and a linear layer with softmax activation.

3.6. Experiment Design

We conducted a hyper-parameter search with different optimizers, learning rates, and weight decays. We trained a model 10 times per parameter combination and selected the best one based on the F1-score and micro-averaged F1-score on the development set for binary and multi-class relation extraction, respectively. In binary relation extraction, we selected the best setting for relation extraction and computed mean precision, recall, and F1-score as well as the standard deviation on ChemProt development (dev) and test sets. In multi-class relation extraction, we selected the best setting for relation extraction and computed mean micro precision, micro recall,

micro-averaged F1-score, and macro-averaged F1-score as well as the standard deviation on ChemProt development and test sets. On the other hand, Cross-Entropy Loss is selected as a loss function during the model training.

The parameter search strategy is applied to find the optimal parameters minimized Cross-Entropy Loss for binary and multi-class relation extraction models. In transformer-based models, we set the batch size to 16 and select the Cross-Entropy Loss function as a loss function during all model training experiments. Learning rate, weight decay, and optimizer are searched to reach the minimum Cross-Entropy Loss value on ChemProt development and test sets by taking the best model out of 10 models trained with different parameter combinations. Table 3.13 indicates all parameters kept constant and searched with their search space during both binary and multi-class relation extraction experiments.

Table 3.13. Parameters of BioBERT and SciBERT. Batch size and loss function are kept constant for experiments of two tasks. We search for the best combination of the learning rate, weight decay, and an optimizer with their search space.

Network Parameter	**Value**
Epochs	4
Batch Size	16
Loss function	Cross-Entropy Loss
Learning Rate	1e-5, 2e-5, 3e-5, 4e-5, 5e-5
Weight Decay	0.01, 0.02, 0.03, 0.1, 0.2, 0.3
Momentum	0.99
Optimizer	AdamW [76], Adam [77], SGD

We assess our models with different evaluation metrics throughout this study. The automatically extracted relations are compared with the manually annotated (gold standard) relations in ChemProt. We use precision, recall and F1-score to evaluate our binary relation extraction approaches. Precision is the proportion of correctly retrieved

relations to all retrieved relations from our relation extraction approaches. Recall is defined as the proportion of correctly retrieved relations to all manually annotated relations. F1-score is calculated as the harmonic mean of precision and recall. On the other hand, we use micro precision, micro recall, micro-averaged F1-score, and macro-averaged F1-score to evaluate our multi-class relation extraction approaches on five relation groups (CPR:3, CPR:4, CPR:5, CPR:6, CPR:9). Micro precision is defined as the proportion of the sum of correctly retrieved relations for five relation groups to all retrieved relations belonging to five relation groups, whereas micro recall is defined as the sum of correctly retrieved relations for five relation groups to all manually annotated relations belonging to five relation groups. Micro-averaged F1-score is the harmonic mean of micro precision and micro recall values. Macro precision is defined as the mean of five precision values of five relation groups. Macro recall is defined as the average of five recall values of five relation groups. Macro-averaged F1-score is performed by taking average of F1-score for five relation classes.

In BioCreative VI: Chemical-Protein Interaction (ChemProt) Track, micro-averaged F1-score on five relation groups is the main metric to evaluate predicted relations against manually annotated relations, since the size of the data set of five relation groups is different from each other. As micro-averaged F1-score depends on the data set size on relation groups, it can be high even if multi-class relation extraction approaches perform poorly on a rare relation group. In addition, we report macro-averaged F1-score to evaluate our multi-class relation extraction approaches on five relation groups regardless of their data set size.

4. EXPERIMENTS AND RESULTS

Throughout this study, we conducted experiments using BioBERT and SciBERT with three input representations: (i) the sentence-based input representation, (ii) the shortest sequence between chemical and protein pairs, and (iii) the shortest sequence including chemical and protein pairs and their parent nodes. In this chapter, we present the scores tables of our binary and multi-class relation extraction experiments. At the end of this chapter, we discuss the three input representations with relation extraction models. In addition, we provide an examination of the relation extraction models with BioBERT and SciBERT. In this section, we make discussion on the model results on the development set.

4.1. Binary Relation Extraction Experiments with BioBERT

4.1.1. Binary Relation Extraction with Sentence-based Input Representation

- Development Set Results

Table 4.1. The mean and standard deviation (std) of evaluation metrics in the development set with parameters used in binary relation extraction with BioBERT using the entire sentence.

Optimizer	**Learning Rate**	**Weight Decay**	**Precision**	**Recall**	**F1-Score**
AdamW	1e-5	0.1	0.733±0.012	0.805±0.018	0.767±0.004
AdamW	1e-5	0.2	0.776±0.037	0.784±0.048	0.778±0.009
AdamW	1e-5	0.3	0.748±0.027	0.777±0.043	0.761±0.01
AdamW	2e-5	0.1	0.761±0.049	0.772±0.078	0.762±0.022
AdamW	2e-5	0.2	0.763±0.026	0.803±0.026	0.782±0.01

Table 4.1. The mean and std of evaluation metrics in the dev set. (cont.)

Optimizer	**Learning Rate**	**Weight Decay**	**Precision**	**Recall**	**F1-Score**
AdamW	2e-5	0.3	0.762±0.017	0.811±0.03	0.785±0.008
AdamW	3e-5	0.1	0.779±0.029	0.784±0.053	0.78±0.018
AdamW	3e-5	0.2	0.766±0.02	0.799±0.029	0.781±0.007
AdamW	3e-5	0.3	0.763±0.012	0.812±0.018	0.787±0.005
Adam	1e-5	0.01	0.455±0.005	0.459±0.022	0.457±0.009
Adam	1e-5	0.02	0.469±0.008	0.448±0.018	0.458±0.007
Adam	1e-5	0.03	0.468±0.006	0.448±0.012	0.458±0.007
Adam	2e-5	0.01	0.478±0.012	0.432±0.033	0.453±0.014
Adam	2e-5	0.02	0.481±0.007	0.436±0.017	0.457±0.009
Adam	2e-5	0.03	0.501±0.013	0.428±0.027	0.461±0.012
Adam	3e-5	0.01	0.512±0.009	0.43±0.019	0.467±0.008
Adam	3e-5	0.02	0.512±0.008	0.419±0.022	0.46±0.011
Adam	3e-5	0.03	0.511±0.005	0.416±0.021	0.458±0.012
Adam	4e-5	0.01	0.525±0.011	0.401±0.036	0.453±0.021
Adam	4e-5	0.02	0.531±0.016	0.396±0.035	0.452±0.019
Adam	4e-5	0.03	0.533±0.016	0.415±0.039	0.465±0.019
Adam	5e-5	0.01	0.529±0.014	0.438±0.038	0.478±0.019
Adam	5e-5	0.02	0.526±0.032	0.451±0.077	0.48±0.038
Adam	5e-5	0.03	0.502±0.035	0.474±0.194	0.505±0.05
SGD	1e-5	0.01	0.734±0.023	0.238±0.113	0.343±0.148
SGD	1e-5	0.02	0.722±0.01	0.291±0.031	0.414±0.031
SGD	1e-5	0.03	0.732±0.01	0.274±0.048	0.396±0.051
SGD	2e-5	0.01	0.719±0.015	0.429±0.128	0.525±0.108
SGD	2e-5	0.02	0.714±0.014	0.462±0.029	0.56±0.023
SGD	2e-5	0.03	0.718±0.012	0.427±0.051	0.534±0.044
SGD	3e-5	0.01	0.719±0.019	0.548±0.052	0.62±0.033
SGD	3e-5	0.02	0.734±0.022	0.507±0.072	0.596±0.046

Table 4.1. The mean and std of evaluation metrics in the dev set. (cont.)

Optimizer	**Learning Rate**	**Weight Decay**	**Precision**	**Recall**	**F1-Score**
SGD	3e-5	0.03	0.723±0.019	0.482±0.095	0.572±0.076
SGD	4e-5	0.01	0.733±0.03	0.578±0.059	0.643±0.03
SGD	4e-5	0.02	0.732±0.018	0.536±0.088	0.614±0.062
SGD	4e-5	0.03	0.729±0.024	0.533±0.063	0.613±0.039
SGD	5e-5	0.01	0.728±0.018	0.622±0.033	0.671±0.02
SGD	5e-5	0.02	0.726±0.019	0.61±0.051	0.661±0.026
SGD	5e-5	0.03	0.725±0.019	0.537±0.062	0.614±0.04

- Test Set Results

Table 4.2. The mean and standard deviation (std) of evaluation metrics in the test set with parameters used in binary relation extraction with BioBERT using the entire sentence.

Optimizer	**Learning Rate**	**Weight Decay**	**Precision**	**Recall**	**F1-Score**
AdamW	1e-5	0.1	0.78±0.01	0.751±0.022	0.765±0.011
AdamW	1e-5	0.2	0.8±0.038	0.714±0.075	0.751±0.031
AdamW	1e-5	0.3	0.792±0.018	0.719±0.048	0.752±0.021
AdamW	2e-5	0.1	0.815±0.026	0.726±0.052	0.766±0.02
AdamW	2e-5	0.2	0.804±0.021	0.74±0.032	0.77±0.011
AdamW	2e-5	0.3	0.806±0.018	0.74±0.032	0.771±0.013
AdamW	3e-5	0.1	0.813±0.019	0.724±0.051	0.765±0.024
AdamW	3e-5	0.2	0.802±0.014	0.752±0.039	0.775±0.017
AdamW	3e-5	0.3	0.804±0.014	0.753±0.022	0.778±0.006
Adam	1e-5	0.01	0.511±0.006	0.455±0.022	0.481±0.01
Adam	1e-5	0.02	0.517±0.007	0.439±0.015	0.475±0.008

Table 4.2. The mean and std of evaluation metrics in the test set. (cont.)

Optimizer	**Learning Rate**	**Weight Decay**	**Precision**	**Recall**	**F1-Score**
Adam	1e-5	0.03	0.519±0.003	0.44±0.012	0.477±0.006
Adam	2e-5	0.01	0.522±0.009	0.423±0.029	0.466±0.015
Adam	2e-5	0.02	0.525±0.002	0.421±0.019	0.467±0.012
Adam	2e-5	0.03	0.534±0.01	0.398±0.027	0.456±0.016
Adam	3e-5	0.01	0.55±0.006	0.404±0.016	0.465±0.009
Adam	3e-5	0.02	0.55±0.008	0.4±0.022	0.463±0.013
Adam	3e-5	0.03	0.551±0.006	0.399±0.014	0.463±0.009
Adam	4e-5	0.01	0.559±0.01	0.388±0.036	0.457±0.023
Adam	4e-5	0.02	0.558±0.01	0.378±0.038	0.449±0.024
Adam	4e-5	0.03	0.562±0.01	0.388±0.044	0.457±0.028
Adam	5e-5	0.01	0.559±0.006	0.412±0.04	0.473±0.026
Adam	5e-5	0.02	0.562±0.017	0.415±0.08	0.472±0.053
Adam	5e-5	0.03	0.547±0.027	0.448±0.189	0.512±0.067
SGD	1e-5	0.01	0.795±0.043	0.247±0.114	0.359±0.15
SGD	1e-5	0.02	0.779±0.011	0.301±0.03	0.433±0.03
SGD	1e-5	0.03	0.783±0.009	0.288±0.046	0.419±0.049
SGD	2e-5	0.01	0.758±0.023	0.423±0.115	0.532±0.099
SGD	2e-5	0.02	0.755±0.01	0.457±0.025	0.569±0.019
SGD	2e-5	0.03	0.761±0.008	0.426±0.045	0.545±0.038
SGD	3e-5	0.01	0.751±0.014	0.529±0.045	0.62±0.029
SGD	3e-5	0.02	0.763±0.02	0.493±0.062	0.596±0.042
SGD	3e-5	0.03	0.759±0.023	0.475±0.089	0.578±0.074
SGD	4e-5	0.01	0.762±0.022	0.552±0.054	0.638±0.032
SGD	4e-5	0.02	0.761±0.017	0.518±0.08	0.612±0.057
SGD	4e-5	0.03	0.753±0.028	0.526±0.061	0.616±0.039

Table 4.2. The mean and std of evaluation metrics in the test set. (cont.)

Optimizer	**Learning Rate**	**Weight Decay**	**Precision**	**Recall**	**F1-Score**
SGD	5e-5	0.01	0.758±0.016	0.59±0.028	0.663±0.018
SGD	5e-5	0.02	0.752±0.019	0.584±0.043	0.656±0.022
SGD	5e-5	0.03	0.751±0.021	0.532±0.055	0.62±0.035

4.1.2. Binary Relation Extraction with The Shortest Sequence Between Chemical and Protein Entity Pairs

- Development Set Results

Table 4.3. The mean and standard deviation (std) of evaluation metrics in the development set with parameters used in binary relation extraction with BioBERT using the shortest subsequence between chemical and protein entity pairs.

Optimizer	**Learning Rate**	**Weight Decay**	**Precision**	**Recall**	**F1-Score**
AdamW	1e-5	0.1	0.767±0.022	0.745±0.037	0.755±0.01
AdamW	1e-5	0.2	0.749±0.021	0.768±0.029	0.758±0.006
AdamW	1e-5	0.3	0.758±0.025	0.759±0.035	0.757±0.007
AdamW	2e-5	0.1	0.755±0.022	0.778±0.022	0.766±0.006
AdamW	2e-5	0.2	0.759±0.02	0.773±0.019	0.765±0.004
AdamW	2e-5	0.3	0.757±0.027	0.771±0.035	0.763±0.009
AdamW	3e-5	0.1	0.768±0.02	0.766±0.025	0.766±0.006
AdamW	3e-5	0.2	0.773±0.033	0.754±0.054	0.761±0.013
AdamW	3e-5	0.3	0.757±0.023	0.772±0.037	0.764±0.01

Table 4.3. The mean and std of evaluation metrics in the dev set. (cont.)

Optimizer	**Learning Rate**	**Weight Decay**	**Precision**	**Recall**	**F1-Score**
Adam	1e-5	0.01	0.788±0.023	0.668±0.055	0.721±0.026
Adam	1e-5	0.02	0.772±0.02	0.658±0.043	0.71±0.019
Adam	1e-5	0.03	0.786±0.032	0.592±0.088	0.67±0.049
Adam	2e-5	0.01	0.759±0.021	0.693±0.05	0.723±0.022
Adam	2e-5	0.02	0.752±0.049	0.517±0.281	0.69±0.029
Adam	2e-5	0.03	0.757±0.047	0.589±0.195	0.632±0.191
Adam	3e-5	0.01	0.746±0.079	0.563±0.257	0.655±0.132
Adam	3e-5	0.02	0.752±0.07	0.598±0.114	0.654±0.061
Adam	3e-5	0.03	0.699±0.059	0.612±0.094	0.645±0.04
Adam	4e-5	0.01	0.734±0.068	0.491±0.271	0.66±0.037
Adam	4e-5	0.02	0.706±0.087	0.496±0.231	0.595±0.09
Adam	4e-5	0.03	0.745±0.072	0.198±0.262	0.587±0.045
Adam	5e-5	0.01	0.665±0.033	0.4±0.347	0.663±0.013
Adam	5e-5	0.02	0.733±0.109	0.229±0.263	0.541±0.088
Adam	5e-5	0.03	0.734±0.077	0.205±0.213	0.376±0.205
SGD	1e-5	0.01	0.753±0.016	0.432±0.076	0.545±0.067
SGD	1e-5	0.02	0.75±0.014	0.428±0.115	0.534±0.123
SGD	1e-5	0.03	0.748±0.011	0.433±0.054	0.546±0.045
SGD	2e-5	0.01	0.772±0.016	0.552±0.082	0.64±0.059
SGD	2e-5	0.02	0.765±0.01	0.538±0.084	0.628±0.062
SGD	2e-5	0.03	0.771±0.011	0.554±0.065	0.642±0.049
SGD	3e-5	0.01	0.771±0.011	0.614±0.08	0.681±0.054
SGD	3e-5	0.02	0.771±0.029	0.576±0.091	0.654±0.063
SGD	3e-5	0.03	0.779±0.032	0.575±0.073	0.658±0.044
SGD	4e-5	0.01	0.77±0.042	0.578±0.215	0.697±0.046
SGD	4e-5	0.02	0.77±0.02	0.641±0.05	0.699±0.038
SGD	4e-5	0.03	0.767±0.018	0.559±0.199	0.686±0.023

Table 4.3. The mean and std of evaluation metrics in the dev set. (cont.)

Optimizer	**Learning Rate**	**Weight Decay**	**Precision**	**Recall**	**F1-Score**
SGD	5e-5	0.01	0.794±0.025	0.592±0.091	0.673±0.057
SGD	5e-5	0.02	0.773±0.019	0.645±0.042	0.702±0.02
SGD	5e-5	0.03	0.781±0.032	0.609±0.065	0.682±0.037

- Test Set Results

Table 4.4. The mean and standard deviation (std) of evaluation metrics in the test set with parameters used in binary relation extraction with BioBERT using the shortest subsequence between chemical and protein entity pairs.

Optimizer	**Learning Rate**	**Weight Decay**	**Precision**	**Recall**	**F1-Score**
AdamW	1e-5	0.1	0.791±0.022	0.703±0.04	0.743±0.015
AdamW	1e-5	0.2	0.777±0.017	0.724±0.027	0.749±0.008
AdamW	1e-5	0.3	0.782±0.022	0.715±0.039	0.746±0.013
AdamW	2e-5	0.1	0.782±0.018	0.733±0.021	0.756±0.005
AdamW	2e-5	0.2	0.785±0.018	0.724±0.022	0.753±0.006
AdamW	2e-5	0.3	0.788±0.025	0.726±0.039	0.755±0.012
AdamW	3e-5	0.1	0.795±0.018	0.717±0.028	0.754±0.009
AdamW	3e-5	0.2	0.803±0.027	0.707±0.054	0.75±0.021
AdamW	3e-5	0.3	0.786±0.018	0.721±0.042	0.751±0.019
Adam	1e-5	0.01	0.809±0.023	0.628±0.047	0.705±0.023
Adam	1e-5	0.02	0.798±0.018	0.618±0.041	0.695±0.02
Adam	1e-5	0.03	0.808±0.028	0.56±0.082	0.657±0.05
Adam	2e-5	0.01	0.794±0.021	0.652±0.04	0.715±0.018
Adam	2e-5	0.02	0.784±0.041	0.488±0.266	0.682±0.034
Adam	2e-5	0.03	0.787±0.05	0.552±0.184	0.62±0.19

Table 4.4. The mean and std of evaluation metrics in the test set. (cont.)

Optimizer	**Learning Rate**	**Weight Decay**	**Precision**	**Recall**	**F1-Score**
Adam	3e-5	0.01	0.782±0.066	0.531±0.247	0.649±0.145
Adam	3e-5	0.02	0.783±0.063	0.566±0.104	0.647±0.064
Adam	3e-5	0.03	0.73±0.052	0.571±0.094	0.633±0.051
Adam	4e-5	0.01	0.773±0.051	0.456±0.252	0.649±0.047
Adam	4e-5	0.02	0.744±0.073	0.466±0.222	0.588±0.104
Adam	4e-5	0.03	0.762±0.067	0.182±0.242	0.561±0.057
Adam	5e-5	0.01	0.712±0.033	0.377±0.327	0.666±0.014
Adam	5e-5	0.02	0.759±0.101	0.212±0.247	0.522±0.106
Adam	5e-5	0.03	0.767±0.083	0.187±0.197	0.357±0.203
SGD	1e-5	0.01	0.778±0.017	0.43±0.071	0.55±0.063
SGD	1e-5	0.02	0.769±0.014	0.424±0.108	0.537±0.117
SGD	1e-5	0.03	0.774±0.013	0.438±0.049	0.558±0.042
SGD	2e-5	0.01	0.785±0.017	0.532±0.07	0.631±0.05
SGD	2e-5	0.02	0.782±0.01	0.516±0.073	0.619±0.054
SGD	2e-5	0.03	0.788±0.013	0.53±0.054	0.632±0.041
SGD	3e-5	0.01	0.788±0.014	0.588±0.074	0.67±0.051
SGD	3e-5	0.02	0.79±0.03	0.554±0.078	0.647±0.056
SGD	3e-5	0.03	0.798±0.029	0.554±0.068	0.65±0.043
SGD	4e-5	0.01	0.79±0.038	0.553±0.205	0.687±0.042
SGD	4e-5	0.02	0.79±0.017	0.611±0.044	0.688±0.034
SGD	4e-5	0.03	0.788±0.018	0.532±0.189	0.675±0.02
SGD	5e-5	0.01	0.814±0.025	0.564±0.086	0.661±0.056
SGD	5e-5	0.02	0.8±0.019	0.615±0.039	0.694±0.021
SGD	5e-5	0.03	0.796±0.027	0.581±0.057	0.669±0.034

4.1.3. Binary Relation Extraction with The Shortest Subsequence Including Chemical and Protein Entity Pair and Their Parent Node

- Development Set Results

Table 4.5. The mean and std of evaluation metrics in the development set with parameters used in binary relation extraction with BioBERT using the shortest subsequence including chemical and protein entity pairs and their parent node.

Optimizer	**Learning Rate**	**Weight Decay**	**Precision**	**Recall**	**F1-Score**
AdamW	1e-5	0.1	0.759±0.023	0.761±0.036	0.759±0.009
AdamW	1e-5	0.2	0.751±0.02	0.775±0.03	0.762±0.006
AdamW	1e-5	0.3	0.759±0.028	0.765±0.039	0.761±0.009
AdamW	2e-5	0.1	0.762±0.027	0.77±0.035	0.765±0.005
AdamW	2e-5	0.2	0.757±0.024	0.782±0.036	0.768±0.009
AdamW	2e-5	0.3	0.761±0.025	0.77±0.037	0.764±0.008
AdamW	3e-5	0.1	0.769±0.023	0.769±0.029	0.768±0.006
AdamW	3e-5	0.2	0.761±0.014	0.779±0.014	0.769±0.007
AdamW	3e-5	0.3	0.765±0.041	0.769±0.071	0.763±0.023
Adam	1e-5	0.01	0.763±0.017	0.721±0.024	0.741±0.006
Adam	1e-5	0.02	0.776±0.023	0.649±0.045	0.705±0.021
Adam	1e-5	0.03	0.765±0.023	0.64±0.047	0.695±0.022
Adam	2e-5	0.01	0.757±0.036	0.687±0.065	0.717±0.022
Adam	2e-5	0.02	0.751±0.034	0.637±0.05	0.687±0.021
Adam	2e-5	0.03	0.736±0.07	0.612±0.113	0.657±0.039
Adam	3e-5	0.01	0.743±0.067	0.65±0.107	0.683±0.035
Adam	3e-5	0.02	0.736±0.045	0.594±0.079	0.652±0.034
Adam	3e-5	0.03	0.669±0.079	0.622±0.135	0.629±0.053

Table 4.5. The mean and std of evaluation metrics in the dev set. (cont.)

Optimizer	**Learning Rate**	**Weight Decay**	**Precision**	**Recall**	**F1-Score**
Adam	4e-5	0.01	0.691±0.047	0.678±0.065	0.68±0.012
Adam	4e-5	0.02	0.688±0.049	0.491±0.274	0.641±0.049
Adam	4e-5	0.03	0.658±0.039	0.369±0.321	0.632±0.021
Adam	5e-5	0.01	0.713±0.051	0.485±0.268	0.648±0.043
Adam	5e-5	0.02	0.68±0.066	0.408±0.299	0.614±0.059
Adam	5e-5	0.03	0.666±0.01	0.153±0.247	0.576±0.032
SGD	1e-5	0.01	0.748±0.022	0.389±0.096	0.504±0.098
SGD	1e-5	0.02	0.744±0.015	0.431±0.042	0.545±0.033
SGD	1e-5	0.03	0.739±0.009	0.437±0.026	0.548±0.021
SGD	2e-5	0.01	0.763±0.011	0.574±0.073	0.652±0.052
SGD	2e-5	0.02	0.758±0.013	0.543±0.088	0.629±0.064
SGD	2e-5	0.03	0.759±0.01	0.501±0.185	0.641±0.045
SGD	3e-5	0.01	0.774±0.02	0.627±0.077	0.689±0.046
SGD	3e-5	0.02	0.779±0.008	0.588±0.066	0.668±0.045
SGD	3e-5	0.03	0.769±0.017	0.601±0.05	0.673±0.03
SGD	4e-5	0.01	0.766±0.018	0.668±0.025	0.714±0.02
SGD	4e-5	0.02	0.778±0.019	0.619±0.08	0.686±0.053
SGD	4e-5	0.03	0.766±0.012	0.607±0.063	0.676±0.042
SGD	5e-5	0.01	0.78±0.021	0.598±0.221	0.715±0.041
SGD	5e-5	0.02	0.771±0.012	0.519±0.271	0.63±0.223
SGD	5e-5	0.03	0.771±0.028	0.573±0.214	0.693±0.044

- Test Set Results

Table 4.6. The mean and standard deviation (std) of evaluation metrics in the test set with parameters used in binary relation extraction with BioBERT using the shortest subsequence including chemical and protein entity pairs and their parent node.

Optimizer	**Learning Rate**	**Weight Decay**	**Precision**	**Recall**	**F1-Score**
AdamW	1e-5	0.1	0.783±0.019	0.714±0.038	0.746±0.015
AdamW	1e-5	0.2	0.775±0.017	0.728±0.03	0.75±0.01
AdamW	1e-5	0.3	0.781±0.02	0.719±0.044	0.747±0.018
AdamW	2e-5	0.1	0.79±0.021	0.719±0.041	0.752±0.014
AdamW	2e-5	0.2	0.785±0.019	0.736±0.041	0.759±0.015
AdamW	2e-5	0.3	0.791±0.02	0.717±0.041	0.751±0.016
AdamW	3e-5	0.1	0.8±0.018	0.718±0.032	0.756±0.012
AdamW	3e-5	0.2	0.793±0.01	0.728±0.016	0.759±0.008
AdamW	3e-5	0.3	0.796±0.033	0.716±0.074	0.75±0.033
Adam	1e-5	0.01	0.781±0.016	0.675±0.023	0.724±0.007
Adam	1e-5	0.02	0.793±0.02	0.608±0.044	0.687±0.023
Adam	1e-5	0.03	0.78±0.021	0.607±0.045	0.681±0.023
Adam	2e-5	0.01	0.781±0.033	0.645±0.059	0.704±0.021
Adam	2e-5	0.02	0.773±0.032	0.596±0.047	0.671±0.022
Adam	2e-5	0.03	0.75±0.057	0.576±0.113	0.641±0.048
Adam	3e-5	0.01	0.763±0.057	0.608±0.108	0.668±0.046
Adam	3e-5	0.02	0.748±0.045	0.557±0.077	0.633±0.039
Adam	3e-5	0.03	0.686±0.073	0.591±0.137	0.62±0.063

Table 4.6. The mean and std of evaluation metrics in the test set. (cont.)

Optimizer	**Learning Rate**	**Weight Decay**	**Precision**	**Recall**	**F1-Score**
Adam	4e-5	0.01	0.716±0.039	0.637±0.056	0.671±0.012
Adam	4e-5	0.02	0.704±0.043	0.465±0.26	0.63±0.053
Adam	4e-5	0.03	0.673±0.035	0.348±0.302	0.621±0.021
Adam	5e-5	0.01	0.729±0.042	0.459±0.253	0.636±0.047
Adam	5e-5	0.02	0.7±0.057	0.384±0.284	0.603±0.071
Adam	5e-5	0.03	0.678±0.006	0.143±0.23	0.558±0.023
SGD	1e-5	0.01	0.783±0.028	0.392±0.091	0.514±0.094
SGD	1e-5	0.02	0.776±0.016	0.432±0.041	0.553±0.032
SGD	1e-5	0.03	0.773±0.01	0.436±0.029	0.557±0.024
SGD	2e-5	0.01	0.776±0.009	0.547±0.064	0.639±0.047
SGD	2e-5	0.02	0.774±0.017	0.523±0.072	0.621±0.052
SGD	2e-5	0.03	0.777±0.01	0.483±0.175	0.633±0.035
SGD	3e-5	0.01	0.785±0.026	0.591±0.069	0.671±0.041
SGD	3e-5	0.02	0.792±0.011	0.559±0.053	0.654±0.036
SGD	3e-5	0.03	0.784±0.017	0.571±0.047	0.659±0.028
SGD	4e-5	0.01	0.775±0.014	0.63±0.019	0.695±0.014
SGD	4e-5	0.02	0.791±0.017	0.587±0.074	0.671±0.052
SGD	4e-5	0.03	0.786±0.011	0.58±0.049	0.666±0.032
SGD	5e-5	0.01	0.79±0.021	0.565±0.207	0.697±0.038
SGD	5e-5	0.02	0.798±0.024	0.493±0.257	0.618±0.219
SGD	5e-5	0.03	0.782±0.026	0.545±0.203	0.679±0.044

4.2. Binary Relation Extraction Experiments with SciBERT

4.2.1. Binary Relation Extraction with Sentence-based Input Representation

- Development Set Results

Table 4.7. The mean and standard deviation (std) of evaluation metrics in the development set with parameters used in binary relation extraction with SciBERT using the entire sentence.

Optimizer	**Learning Rate**	**Weight Decay**	**Precision**	**Recall**	**F1-Score**
AdamW	1e-5	0.1	0.75±0.025	0.781±0.045	0.763±0.015
AdamW	1e-5	0.2	0.749±0.018	0.783±0.028	0.765±0.008
AdamW	1e-5	0.3	0.747±0.019	0.775±0.041	0.76±0.013
AdamW	2e-5	0.1	0.755±0.026	0.787±0.048	0.769±0.014
AdamW	2e-5	0.2	0.765±0.03	0.766±0.068	0.763±0.025
AdamW	2e-5	0.3	0.745±0.006	0.805±0.011	0.774±0.005
AdamW	3e-5	0.1	0.762±0.038	0.765±0.071	0.76±0.026
AdamW	3e-5	0.2	0.765±0.029	0.77±0.063	0.765±0.022
AdamW	3e-5	0.3	0.759±0.019	0.775±0.026	0.766±0.008
Adam	1e-5	0.01	0.765±0.025	0.681±0.067	0.718±0.032
Adam	1e-5	0.02	0.735±0.03	0.668±0.075	0.697±0.036
Adam	1e-5	0.03	0.734±0.026	0.588±0.12	0.644±0.076
Adam	2e-5	0.01	0.747±0.031	0.693±0.067	0.716±0.03
Adam	2e-5	0.02	0.714±0.05	0.637±0.106	0.664±0.062
Adam	2e-5	0.03	0.687±0.07	0.628±0.137	0.64±0.055

Table 4.7. The mean and std of evaluation metrics in the dev set. (cont.)

Optimizer	**Learning Rate**	**Weight Decay**	**Precision**	**Recall**	**F1-Score**
Adam	3e-5	0.01	0.719±0.068	0.669±0.131	0.679±0.071
Adam	3e-5	0.02	0.656±0.047	0.646±0.093	0.644±0.032
Adam	3e-5	0.03	0.656±0.053	0.585±0.108	0.608±0.048
Adam	4e-5	0.01	0.699±0.031	0.607±0.058	0.647±0.024
Adam	4e-5	0.02	0.672±0.078	0.53±0.169	0.565±0.122
Adam	4e-5	0.03	0.701±0.093	0.41±0.19	0.478±0.163
Adam	5e-5	0.01	0.658±0.051	0.615±0.095	0.628±0.043
Adam	5e-5	0.02	0.638±0.057	0.53±0.117	0.565±0.073
Adam	5e-5	0.03	0.665±0.079	0.442±0.179	0.496±0.15
SGD	1e-5	0.01	0.708±0.009	0.492±0.056	0.579±0.043
SGD	1e-5	0.02	0.71±0.011	0.499±0.022	0.586±0.013
SGD	1e-5	0.03	0.711±0.022	0.459±0.084	0.552±0.061
SGD	2e-5	0.01	0.716±0.023	0.602±0.068	0.652±0.043
SGD	2e-5	0.02	0.712±0.02	0.572±0.056	0.633±0.038
SGD	2e-5	0.03	0.713±0.021	0.534±0.073	0.607±0.05
SGD	3e-5	0.01	0.716±0.034	0.588±0.081	0.641±0.048
SGD	3e-5	0.02	0.737±0.03	0.553±0.103	0.624±0.068
SGD	3e-5	0.03	0.718±0.029	0.532±0.119	0.604±0.093
SGD	4e-5	0.01	0.734±0.032	0.637±0.055	0.68±0.027
SGD	4e-5	0.02	0.721±0.036	0.59±0.109	0.642±0.061
SGD	4e-5	0.03	0.717±0.023	0.623±0.054	0.665±0.025
SGD	5e-5	0.01	0.741±0.035	0.623±0.055	0.674±0.029
SGD	5e-5	0.02	0.725±0.047	0.598±0.133	0.643±0.084
SGD	5e-5	0.03	0.705±0.035	0.589±0.064	0.639±0.037

- Test Set Results

Table 4.8. The mean and standard deviation (std) of evaluation metrics in the test set with parameters used in binary relation extraction with SciBERT using the entire sentence.

Optimizer	**Learning Rate**	**Weight Decay**	**Precision**	**Recall**	**F1-Score**
AdamW	1e-5	0.1	0.788±0.024	0.725±0.042	0.754±0.017
AdamW	1e-5	0.2	0.785±0.015	0.727±0.032	0.754±0.013
AdamW	1e-5	0.3	0.802±0.032	0.717±0.073	0.753±0.034
AdamW	2e-5	0.1	0.795±0.025	0.736±0.048	0.763±0.018
AdamW	2e-5	0.2	0.805±0.027	0.711±0.065	0.752±0.029
AdamW	2e-5	0.3	0.79±0.007	0.755±0.007	0.772±0.004
AdamW	3e-5	0.1	0.785±0.021	0.719±0.038	0.75±0.016
AdamW	3e-5	0.2	0.802±0.025	0.717±0.057	0.755±0.024
AdamW	3e-5	0.3	0.799±0.017	0.728±0.029	0.761±0.013
Adam	1e-5	0.01	0.796±0.03	0.64±0.067	0.706±0.035
Adam	1e-5	0.02	0.761±0.029	0.628±0.072	0.685±0.037
Adam	1e-5	0.03	0.759±0.032	0.557±0.116	0.633±0.078
Adam	2e-5	0.01	0.778±0.027	0.659±0.068	0.71±0.037
Adam	2e-5	0.02	0.744±0.047	0.615±0.106	0.664±0.069
Adam	2e-5	0.03	0.716±0.06	0.603±0.134	0.641±0.065
Adam	3e-5	0.01	0.749±0.06	0.64±0.126	0.677±0.076
Adam	3e-5	0.02	0.693±0.04	0.623±0.089	0.65±0.038
Adam	3e-5	0.03	0.689±0.052	0.555±0.098	0.606±0.05

Table 4.8. The mean and std of evaluation metrics in the test set. (cont.)

Optimizer	**Learning Rate**	**Weight Decay**	**Precision**	**Recall**	**F1-Score**
Adam	4e-5	0.01	0.734±0.02	0.573±0.056	0.641±0.031
Adam	4e-5	0.02	0.71±0.074	0.492±0.153	0.557±0.123
Adam	4e-5	0.03	0.733±0.084	0.382±0.18	0.466±0.167
Adam	5e-5	0.01	0.695±0.046	0.579±0.083	0.625±0.043
Adam	5e-5	0.02	0.674±0.052	0.483±0.109	0.551±0.08
Adam	5e-5	0.03	0.702±0.069	0.402±0.16	0.482±0.149
SGD	1e-5	0.01	0.735±0.013	0.468±0.046	0.571±0.035
SGD	1e-5	0.02	0.739±0.011	0.471±0.023	0.575±0.015
SGD	1e-5	0.03	0.742±0.023	0.438±0.075	0.546±0.057
SGD	2e-5	0.01	0.74±0.018	0.559±0.06	0.635±0.038
SGD	2e-5	0.02	0.738±0.016	0.533±0.046	0.618±0.031
SGD	2e-5	0.03	0.742±0.025	0.499±0.064	0.593±0.046
SGD	3e-5	0.01	0.744±0.026	0.55±0.075	0.629±0.05
SGD	3e-5	0.02	0.762±0.035	0.517±0.094	0.609±0.065
SGD	3e-5	0.03	0.741±0.025	0.494±0.104	0.587±0.086
SGD	4e-5	0.01	0.762±0.03	0.594±0.051	0.665±0.026
SGD	4e-5	0.02	0.75±0.036	0.556±0.1	0.632±0.059
SGD	4e-5	0.03	0.736±0.025	0.583±0.057	0.648±0.03
SGD	5e-5	0.01	0.771±0.037	0.579±0.054	0.658±0.03
SGD	5e-5	0.02	0.755±0.044	0.559±0.123	0.63±0.084
SGD	5e-5	0.03	0.732±0.026	0.557±0.061	0.63±0.038

4.2.2. Binary Relation Extraction with The Shortest Sequence Between Chemical and Protein Entity Pairs

- Development Set Results

Table 4.9. The mean and standard deviation (std) of evaluation metrics in the development set with parameters used in binary relation extraction with SciBERT using the shortest subsequence between chemical and protein entity pairs.

Optimizer	**Learning Rate**	**Weight Decay**	**Precision**	**Recall**	**F1-Score**
AdamW	1e-5	0.1	0.715±0.154	0.686±0.141	0.699±0.142
AdamW	1e-5	0.2	0.762±0.024	0.742±0.043	0.751±0.012
AdamW	1e-5	0.3	0.747±0.023	0.756±0.041	0.75±0.013
AdamW	2e-5	0.1	0.751±0.016	0.77±0.027	0.76±0.008
AdamW	2e-5	0.2	0.753±0.024	0.748±0.036	0.749±0.011
AdamW	2e-5	0.3	0.756±0.045	0.749±0.076	0.748±0.027
AdamW	3e-5	0.1	0.752±0.024	0.765±0.042	0.758±0.014
AdamW	3e-5	0.2	0.761±0.023	0.748±0.041	0.753±0.011
AdamW	3e-5	0.3	0.753±0.026	0.762±0.034	0.756±0.006
Adam	1e-5	0.01	0.767±0.007	0.72±0.019	0.742±0.008
Adam	1e-5	0.02	0.758±0.014	0.701±0.026	0.728±0.014
Adam	1e-5	0.03	0.755±0.052	0.675±0.099	0.705±0.047
Adam	2e-5	0.01	0.776±0.033	0.666±0.069	0.713±0.029
Adam	2e-5	0.02	0.743±0.031	0.67±0.044	0.703±0.014
Adam	2e-5	0.03	0.75±0.042	0.623±0.066	0.677±0.025
Adam	3e-5	0.01	0.728±0.022	0.685±0.036	0.705±0.015
Adam	3e-5	0.02	0.728±0.051	0.621±0.088	0.663±0.033
Adam	3e-5	0.03	0.703±0.04	0.62±0.068	0.655±0.021

Table 4.9. The mean and std of evaluation metrics in the dev set. (cont.)

Optimizer	**Learning Rate**	**Weight Decay**	**Precision**	**Recall**	**F1-Score**
Adam	4e-5	0.01	0.742±0.073	0.607±0.125	0.654±0.058
Adam	4e-5	0.02	0.709±0.076	0.589±0.125	0.628±0.058
Adam	4e-5	0.03	0.699±0.069	0.562±0.112	0.61±0.064
Adam	5e-5	0.01	0.699±0.061	0.627±0.101	0.653±0.027
Adam	5e-5	0.02	0.652±0.032	0.617±0.048	0.633±0.023
Adam	5e-5	0.03	0.682±0.041	0.517±0.068	0.584±0.033
SGD	1e-5	0.01	0.754±0.024	0.549±0.052	0.633±0.033
SGD	1e-5	0.02	0.753±0.02	0.553±0.079	0.634±0.05
SGD	1e-5	0.03	0.754±0.017	0.544±0.074	0.629±0.05
SGD	2e-5	0.01	0.749±0.022	0.637±0.053	0.688±0.039
SGD	2e-5	0.02	0.753±0.028	0.633±0.072	0.684±0.04
SGD	2e-5	0.03	0.751±0.04	0.622±0.071	0.676±0.032
SGD	3e-5	0.01	0.758±0.025	0.688±0.066	0.719±0.029
SGD	3e-5	0.02	0.772±0.018	0.657±0.034	0.709±0.015
SGD	3e-5	0.03	0.761±0.013	0.642±0.06	0.695±0.034
SGD	4e-5	0.01	0.761±0.02	0.681±0.062	0.717±0.034
SGD	4e-5	0.02	0.746±0.048	0.686±0.063	0.711±0.023
SGD	4e-5	0.03	0.765±0.07	0.624±0.151	0.669±0.096
SGD	5e-5	0.01	0.769±0.03	0.695±0.078	0.726±0.038
SGD	5e-5	0.02	0.756±0.027	0.693±0.046	0.721±0.022
SGD	5e-5	0.03	0.758±0.038	0.672±0.073	0.708±0.035

- Test Set Results

Table 4.10. The mean and standard deviation (std) of evaluation metrics in the test set with parameters used in binary relation extraction with SciBERT using the shortest subsequence between chemical and protein entity pairs.

Optimizer	**Learning Rate**	**Weight Decay**	**Precision**	**Recall**	**F1-Score**
AdamW	1e-5	0.1	0.798±0.027	0.681±0.046	0.733±0.016
AdamW	1e-5	0.2	0.79±0.027	0.7±0.042	0.741±0.013
AdamW	1e-5	0.3	0.779±0.022	0.717±0.042	0.746±0.016
AdamW	2e-5	0.1	0.778±0.016	0.723±0.023	0.749±0.006
AdamW	2e-5	0.2	0.782±0.021	0.706±0.041	0.741±0.015
AdamW	2e-5	0.3	0.79±0.037	0.708±0.076	0.743±0.036
AdamW	3e-5	0.1	0.78±0.023	0.71±0.043	0.742±0.016
AdamW	3e-5	0.2	0.79±0.022	0.705±0.039	0.744±0.014
AdamW	3e-5	0.3	0.785±0.02	0.717±0.036	0.748±0.013
Adam	1e-5	0.01	0.79±0.009	0.677±0.015	0.729±0.005
Adam	1e-5	0.02	0.78±0.019	0.665±0.026	0.717±0.011
Adam	1e-5	0.03	0.784±0.05	0.635±0.095	0.695±0.05
Adam	2e-5	0.01	0.803±0.034	0.625±0.066	0.699±0.034
Adam	2e-5	0.02	0.777±0.028	0.633±0.047	0.696±0.02
Adam	2e-5	0.03	0.772±0.034	0.592±0.064	0.667±0.031
Adam	3e-5	0.01	0.763±0.017	0.651±0.034	0.701±0.013
Adam	3e-5	0.02	0.757±0.049	0.583±0.085	0.652±0.039
Adam	3e-5	0.03	0.731±0.034	0.591±0.063	0.651±0.025

Table 4.10. The mean and std of evaluation metrics in the test set. (cont.)

Optimizer	**Learning Rate**	**Weight Decay**	**Precision**	**Recall**	**F1-Score**
Adam	4e-5	0.01	0.768±0.06	0.576±0.118	0.647±0.064
Adam	4e-5	0.02	0.74±0.062	0.56±0.124	0.624±0.068
Adam	4e-5	0.03	0.722±0.058	0.534±0.114	0.601±0.077
Adam	5e-5	0.01	0.734±0.053	0.599±0.101	0.652±0.039
Adam	5e-5	0.02	0.684±0.027	0.589±0.049	0.632±0.026
Adam	5e-5	0.03	0.702±0.039	0.487±0.065	0.571±0.036
SGD	1e-5	0.01	0.779±0.022	0.527±0.052	0.627±0.036
SGD	1e-5	0.02	0.779±0.021	0.53±0.071	0.628±0.049
SGD	1e-5	0.03	0.778±0.016	0.528±0.067	0.626±0.047
SGD	2e-5	0.01	0.775±0.019	0.607±0.042	0.681±0.033
SGD	2e-5	0.02	0.777±0.03	0.608±0.07	0.679±0.041
SGD	2e-5	0.03	0.773±0.04	0.6±0.071	0.671±0.035
SGD	3e-5	0.01	0.78±0.028	0.655±0.061	0.709±0.028
SGD	3e-5	0.02	0.794±0.018	0.63±0.031	0.702±0.014
SGD	3e-5	0.03	0.785±0.013	0.62±0.057	0.691±0.034
SGD	4e-5	0.01	0.783±0.019	0.651±0.055	0.709±0.031
SGD	4e-5	0.02	0.769±0.049	0.657±0.063	0.705±0.022
SGD	4e-5	0.03	0.786±0.063	0.6±0.146	0.664±0.1
SGD	5e-5	0.01	0.79±0.031	0.659±0.077	0.714±0.041
SGD	5e-5	0.02	0.783±0.024	0.661±0.041	0.715±0.019
SGD	5e-5	0.03	0.783±0.04	0.639±0.065	0.7±0.033

4.2.3. Binary Relation Extraction with The Shortest Subsequence Including Chemical and Protein Entity Pair and Their Parent Node

- Development Set Results

Table 4.11. The mean and std of evaluation metrics in the development set with parameters used in binary class relation extraction with BioBERT using the shortest subsequence including chemical and protein entity pairs and their parent node.

Optimizer	**Learning Rate**	**Weight Decay**	**Precision**	**Recall**	**F1-Score**
AdamW	1e-5	0.1	0.745±0.02	0.773±0.038	0.758±0.011
AdamW	1e-5	0.2	0.759±0.019	0.751±0.039	0.754±0.013
AdamW	1e-5	0.3	0.741±0.007	0.786±0.014	0.763±0.004
AdamW	2e-5	0.1	0.752±0.018	0.769±0.041	0.759±0.015
AdamW	2e-5	0.2	0.755±0.024	0.769±0.038	0.761±0.008
AdamW	2e-5	0.3	0.753±0.035	0.761±0.039	0.755±0.008
AdamW	3e-5	0.1	0.76±0.032	0.751±0.051	0.754±0.013
AdamW	3e-5	0.2	0.776±0.05	0.718±0.093	0.739±0.037
AdamW	3e-5	0.3	0.757±0.022	0.763±0.046	0.759±0.016
Adam	1e-5	0.01	0.754±0.052	0.719±0.062	0.733±0.02
Adam	1e-5	0.02	0.781±0.033	0.663±0.068	0.714±0.033
Adam	1e-5	0.03	0.77±0.026	0.673±0.056	0.716±0.024
Adam	2e-5	0.01	0.769±0.029	0.686±0.049	0.723±0.016
Adam	2e-5	0.02	0.723±0.023	0.698±0.031	0.709±0.008
Adam	2e-5	0.03	0.725±0.025	0.655±0.04	0.687±0.016
Adam	3e-5	0.01	0.736±0.062	0.675±0.096	0.696±0.04
Adam	3e-5	0.02	0.715±0.063	0.623±0.125	0.653±0.06
Adam	3e-5	0.03	0.698±0.062	0.624±0.109	0.649±0.047

Table 4.11. The mean and std of evaluation metrics in the dev set. (cont.)

Optimizer	**Learning Rate**	**Weight Decay**	**Precision**	**Recall**	**F1-Score**
Adam	4e-5	0.01	0.743±0.071	0.601±0.134	0.649±0.068
Adam	4e-5	0.02	0.674±0.048	0.629±0.092	0.643±0.042
Adam	4e-5	0.03	0.707±0.05	0.53±0.109	0.595±0.066
Adam	5e-5	0.01	0.712±0.094	0.591±0.159	0.622±0.078
Adam	5e-5	0.02	0.693±0.08	0.536±0.148	0.583±0.089
Adam	5e-5	0.03	0.68±0.062	0.515±0.12	0.572±0.079
SGD	1e-5	0.01	0.748±0.015	0.585±0.049	0.655±0.033
SGD	1e-5	0.02	0.751±0.015	0.56±0.047	0.641±0.034
SGD	1e-5	0.03	0.752±0.009	0.564±0.039	0.644±0.025
SGD	2e-5	0.01	0.754±0.02	0.664±0.05	0.705±0.025
SGD	2e-5	0.02	0.763±0.033	0.605±0.093	0.669±0.063
SGD	2e-5	0.03	0.759±0.022	0.62±0.027	0.682±0.017
SGD	3e-5	0.01	0.76±0.016	0.657±0.065	0.703±0.036
SGD	3e-5	0.02	0.763±0.016	0.674±0.049	0.714±0.025
SGD	3e-5	0.03	0.76±0.037	0.649±0.058	0.697±0.028
SGD	4e-5	0.01	0.776±0.023	0.669±0.078	0.715±0.044
SGD	4e-5	0.02	0.764±0.024	0.674±0.029	0.715±0.012
SGD	4e-5	0.03	0.755±0.022	0.673±0.032	0.711±0.014
SGD	5e-5	0.01	0.765±0.043	0.686±0.102	0.716±0.057
SGD	5e-5	0.02	0.764±0.018	0.678±0.047	0.717±0.024
SGD	5e-5	0.03	0.769±0.039	0.641±0.087	0.693±0.048

- Test Set Results

Table 4.12. The mean and std of evaluation metrics in the test set with parameters used in binary class relation extraction with BioBERT using the shortest subsequence including chemical and protein entity pairs and their parent node.

Optimizer	**Learning Rate**	**Weight Decay**	**Precision**	**Recall**	**F1-Score**
AdamW	1e-5	0.1	0.771±0.018	0.731±0.042	0.749±0.017
AdamW	1e-5	0.2	0.784±0.02	0.707±0.042	0.742±0.017
AdamW	1e-5	0.3	0.765±0.009	0.744±0.011	0.754±0.005
AdamW	2e-5	0.1	0.779±0.013	0.727±0.04	0.751±0.018
AdamW	2e-5	0.2	0.777±0.02	0.723±0.04	0.748±0.014
AdamW	2e-5	0.3	0.779±0.029	0.715±0.036	0.745±0.01
AdamW	3e-5	0.1	0.784±0.024	0.709±0.051	0.743±0.021
AdamW	3e-5	0.2	0.797±0.036	0.674±0.092	0.725±0.048
AdamW	3e-5	0.3	0.78±0.018	0.712±0.043	0.744±0.019
Adam	1e-5	0.01	0.774±0.041	0.676±0.063	0.718±0.019
Adam	1e-5	0.02	0.798±0.03	0.618±0.064	0.693±0.035
Adam	1e-5	0.03	0.784±0.023	0.62±0.056	0.69±0.029
Adam	2e-5	0.01	0.789±0.025	0.643±0.05	0.707±0.023
Adam	2e-5	0.02	0.749±0.019	0.661±0.034	0.702±0.01
Adam	2e-5	0.03	0.745±0.018	0.623±0.043	0.677±0.023
Adam	3e-5	0.01	0.764±0.053	0.637±0.098	0.687±0.051
Adam	3e-5	0.02	0.732±0.049	0.59±0.125	0.641±0.071
Adam	3e-5	0.03	0.719±0.058	0.589±0.105	0.638±0.053

Table 4.12. The mean and std of evaluation metrics in the test set. (cont.)

Optimizer	**Learning Rate**	**Weight Decay**	**Precision**	**Recall**	**F1-Score**
Adam	4e-5	0.01	0.762±0.059	0.567±0.129	0.636±0.077
Adam	4e-5	0.02	0.696±0.04	0.595±0.089	0.635±0.049
Adam	4e-5	0.03	0.716±0.05	0.503±0.102	0.581±0.067
Adam	5e-5	0.01	0.735±0.082	0.56±0.152	0.615±0.085
Adam	5e-5	0.02	0.704±0.075	0.5±0.152	0.562±0.106
Adam	5e-5	0.03	0.688±0.053	0.483±0.117	0.554±0.085
SGD	1e-5	0.01	0.775±0.012	0.563±0.044	0.651±0.031
SGD	1e-5	0.02	0.776±0.017	0.542±0.045	0.637±0.035
SGD	1e-5	0.03	0.777±0.008	0.546±0.037	0.641±0.025
SGD	2e-5	0.01	0.78±0.024	0.629±0.049	0.694±0.026
SGD	2e-5	0.02	0.786±0.025	0.581±0.089	0.663±0.065
SGD	2e-5	0.03	0.783±0.017	0.598±0.029	0.678±0.016
SGD	3e-5	0.01	0.781±0.02	0.627±0.059	0.694±0.034
SGD	3e-5	0.02	0.78±0.017	0.644±0.048	0.704±0.026
SGD	3e-5	0.03	0.783±0.033	0.629±0.057	0.695±0.029
SGD	4e-5	0.01	0.796±0.024	0.637±0.072	0.704±0.043
SGD	4e-5	0.02	0.787±0.024	0.647±0.029	0.709±0.013
SGD	4e-5	0.03	0.782±0.018	0.643±0.034	0.705±0.016
SGD	5e-5	0.01	0.787±0.04	0.653±0.097	0.707±0.06
SGD	5e-5	0.02	0.79±0.021	0.649±0.042	0.712±0.02
SGD	5e-5	0.03	0.795±0.03	0.608±0.084	0.684±0.052

4.3. Multi-class Relation Extraction Experiments with BioBERT

4.3.1. Multi-class Relation Extraction with Sentence-based Input Representation

- Development Set Results

Table 4.13. The mean and standard deviation (std) of evaluation metrics in the development set with parameters used in multi-class relation extraction with BioBERT using the entire sentence. LR is the learning rate in the table.

Optimizer	**LR**	**Weight Decay**	**Micro Precision**	**Micro Recall**	**Micro-averaged F1-score**	**Macro-averaged F1-score**
AdamW	1e-5	0.1	0.753±0.014	0.784±0.025	0.767±0.008	0.793±0.01
AdamW	1e-5	0.2	0.77±0.011	0.797±0.014	0.783±0.006	0.807±0.008
AdamW	1e-5	0.3	0.791±0.024	0.785±0.027	0.787±0.006	0.809±0.009
AdamW	2e-5	0.1	0.735±0.023	0.794±0.023	0.763±0.01	0.789±0.011
AdamW	2e-5	0.2	0.767±0.025	0.801±0.014	0.783±0.008	0.81±0.009
AdamW	2e-5	0.3	0.783±0.011	0.789±0.012	0.786±0.009	0.811±0.012
AdamW	3e-5	0.1	0.749±0.018	0.794±0.014	0.771±0.006	0.797±0.006
AdamW	3e-5	0.2	0.768±0.014	0.79±0.02	0.779±0.009	0.804±0.014
AdamW	3e-5	0.3	0.756±0.027	0.811±0.016	0.782±0.016	0.805±0.02
Adam	1e-5	0.01	0.723±0.04	0.742±0.038	0.731±0.028	0.744±0.04
Adam	1e-5	0.02	0.721±0.026	0.688±0.031	0.703±0.019	0.718±0.022
Adam	1e-5	0.03	0.69±0.015	0.635±0.063	0.66±0.04	0.659±0.076
Adam	2e-5	0.01	0.72±0.013	0.738±0.016	0.729±0.006	0.745±0.008
Adam	2e-5	0.02	0.688±0.066	0.657±0.109	0.662±0.045	0.674±0.051
Adam	2e-5	0.03	0.639±0.059	0.652±0.08	0.639±0.026	0.65±0.047
Adam	3e-5	0.01	0.659±0.039	0.733±0.023	0.693±0.015	0.708±0.016
Adam	3e-5	0.02	0.639±0.059	0.652±0.08	0.639±0.026	0.65±0.047

Table 4.13. The mean and std of evaluation metrics in the dev set. (cont.)

Optimizer	**LR**	**Weight Decay**	**Micro Precision**	**Micro Recall**	**Micro-averaged F1-score**	**Macro-averaged F1-score**
Adam	3e-5	0.03	0.548±0.025	0.611±0.054	0.576±0.021	0.588±0.024
Adam	4e-5	0.01	0.648±0.035	0.665±0.07	0.653±0.027	0.663±0.05
Adam	4e-5	0.02	0.604±0.079	0.534±0.089	0.556±0.031	0.548±0.063
Adam	4e-5	0.03	0.54±0.067	0.435±0.186	0.496±0.05	0.534±0.14
Adam	5e-5	0.01	0.624±0.05	0.54±0.206	0.605±0.034	0.651±0.085
Adam	5e-5	0.02	0.578±0.045	0.386±0.178	0.478±0.078	0.524±0.147
Adam	5e-5	0.03	0.509±0.061	0.397±0.162	0.464±0.042	0.519±0.13
SGD	1e-5	0.01	0.509±0.023	0.179±0.029	0.264±0.031	0.601±0.136
SGD	1e-5	0.02	0.529±0.037	0.152±0.076	0.248±0.077	0.598±0.16
SGD	1e-5	0.03	0.558±0.103	0.139±0.095	0.2±0.134	0.533±0.12
SGD	2e-5	0.01	0.569±0.029	0.355±0.125	0.426±0.121	0.494±0.073
SGD	2e-5	0.02	0.591±0.037	0.332±0.115	0.414±0.11	0.52±0.091
SGD	2e-5	0.03	0.568±0.039	0.355±0.099	0.431±0.094	0.516±0.082
SGD	3e-5	0.01	0.611±0.065	0.518±0.127	0.557±0.107	0.537±0.082
SGD	3e-5	0.02	0.637±0.037	0.505±0.092	0.561±0.076	0.532±0.063
SGD	3e-5	0.03	0.615±0.059	0.482±0.088	0.539±0.08	0.521±0.072
SGD	4e-5	0.01	0.654±0.068	0.645±0.113	0.646±0.086	0.651±0.097
SGD	4e-5	0.02	0.636±0.074	0.55±0.169	0.583±0.136	0.613±0.109
SGD	4e-5	0.03	0.655±0.031	0.546±0.126	0.588±0.105	0.584±0.075
SGD	5e-5	0.01	0.664±0.062	0.689±0.081	0.676±0.07	0.693±0.076
SGD	5e-5	0.02	0.681±0.051	0.523±0.186	0.573±0.14	0.593±0.122
SGD	5e-5	0.03	0.628±0.074	0.515±0.203	0.598±0.085	0.611±0.115

- Test Set Results

Table 4.14. The mean and standard deviation (std) of evaluation metrics in the test set with parameters used in multi-class relation extraction with BioBERT using the entire sentence. LR is the learning rate in the table.

Optimizer	**LR**	**Weight Decay**	**Micro Precision**	**Micro Recall**	**Micro-averaged F1-score**	**Macro-averaged F1-score**
AdamW	1e-5	0.1	0.76±0.01	0.738±0.023	0.749±0.008	0.766±0.009
AdamW	1e-5	0.2	0.775±0.014	0.749±0.01	0.761±0.006	0.778±0.006
AdamW	1e-5	0.3	0.79±0.025	0.735±0.025	0.761±0.005	0.776±0.008
AdamW	2e-5	0.1	0.745±0.026	0.746±0.021	0.745±0.012	0.762±0.013
AdamW	2e-5	0.2	0.77±0.02	0.752±0.016	0.761±0.006	0.777±0.007
AdamW	2e-5	0.3	0.783±0.012	0.738±0.01	0.759±0.008	0.774±0.008
AdamW	3e-5	0.1	0.754±0.014	0.746±0.012	0.749±0.006	0.766±0.006
AdamW	3e-5	0.2	0.776±0.011	0.745±0.022	0.76±0.009	0.775±0.011
AdamW	3e-5	0.3	0.763±0.022	0.759±0.015	0.761±0.011	0.776±0.013
Adam	1e-5	0.01	0.715±0.034	0.693±0.033	0.703±0.018	0.718±0.029
Adam	1e-5	0.02	0.715±0.021	0.648±0.028	0.679±0.014	0.699±0.019
Adam	1e-5	0.03	0.695±0.015	0.611±0.057	0.649±0.037	0.652±0.078
Adam	2e-5	0.01	0.715±0.016	0.702±0.013	0.708±0.007	0.729±0.009
Adam	2e-5	0.02	0.69±0.057	0.621±0.102	0.645±0.047	0.664±0.051
Adam	2e-5	0.03	0.652±0.06	0.61±0.078	0.623±0.027	0.638±0.049
Adam	3e-5	0.01	0.665±0.026	0.696±0.024	0.68±0.006	0.702±0.006
Adam	3e-5	0.02	0.612±0.042	0.61±0.072	0.606±0.02	0.623±0.022
Adam	3e-5	0.03	0.567±0.025	0.574±0.053	0.569±0.019	0.587±0.021

Table 4.14. The mean and std of evaluation metrics in the test set. (cont.)

Optimizer	**LR**	**Weight Decay**	**Micro Precision**	**Micro Recall**	**Micro-averaged F1-score**	**Macro-averaged F1-score**
Adam	4e-5	0.01	0.649±0.037	0.623±0.055	0.633±0.019	0.65±0.039
Adam	4e-5	0.02	0.611±0.066	0.506±0.083	0.544±0.034	0.544±0.068
Adam	4e-5	0.03	0.554±0.065	0.409±0.173	0.487±0.05	0.526±0.146
Adam	5e-5	0.01	0.627±0.048	0.508±0.192	0.588±0.031	0.643±0.087
Adam	5e-5	0.02	0.594±0.05	0.365±0.168	0.467±0.083	0.515±0.151
Adam	5e-5	0.03	0.521±0.045	0.379±0.157	0.458±0.045	0.514±0.135
SGD	1e-5	0.01	0.56±0.024	0.215±0.028	0.309±0.027	0.579±0.112
SGD	1e-5	0.02	0.588±0.061	0.182±0.086	0.293±0.077	0.678±0.109
SGD	1e-5	0.03	0.594±0.062	0.167±0.112	0.236±0.156	0.573±0.14
SGD	2e-5	0.01	0.605±0.022	0.365±0.108	0.445±0.101	0.49±0.095
SGD	2e-5	0.02	0.617±0.031	0.346±0.1	0.435±0.094	0.517±0.1
SGD	2e-5	0.03	0.593±0.03	0.366±0.087	0.447±0.08	0.502±0.101
SGD	3e-5	0.01	0.63±0.063	0.514±0.109	0.564±0.094	0.524±0.073
SGD	3e-5	0.02	0.653±0.033	0.497±0.085	0.562±0.07	0.529±0.064
SGD	3e-5	0.03	0.632±0.053	0.473±0.069	0.54±0.064	0.496±0.055
SGD	4e-5	0.01	0.669±0.067	0.615±0.095	0.639±0.077	0.634±0.098
SGD	4e-5	0.02	0.651±0.068	0.538±0.144	0.585±0.12	0.601±0.108
SGD	4e-5	0.03	0.66±0.033	0.525±0.111	0.58±0.093	0.563±0.074
SGD	5e-5	0.01	0.682±0.061	0.663±0.075	0.671±0.065	0.683±0.077
SGD	5e-5	0.02	0.699±0.051	0.508±0.166	0.572±0.125	0.581±0.122
SGD	5e-5	0.03	0.637±0.069	0.494±0.188	0.589±0.07	0.6±0.117

4.3.2. Multi-class Relation Extraction with The Shortest Sequence Between Chemical and Protein Entity Pairs

- Development Set Results

Table 4.15. The mean and std of evaluation metrics in the dev set with parameters used in multi-class relation extraction with BioBERT using the shortest subsequence between chemical and protein entity pairs. LR is the learning rate in the table.

Optimizer	**LR**	**Weight Decay**	**Micro Precision**	**Micro Recall**	**Micro-averaged F1-score**	**Macro-averaged F1-score**
AdamW	1e-5	0.1	0.743±0.028	0.739±0.027	0.74±0.006	0.763±0.008
AdamW	1e-5	0.2	0.734±0.034	0.75±0.035	0.741±0.01	0.766±0.007
AdamW	1e-5	0.3	0.744±0.017	0.75±0.013	0.747±0.005	0.77±0.006
AdamW	2e-5	0.1	0.744±0.019	0.76±0.015	0.752±0.011	0.774±0.013
AdamW	2e-5	0.2	0.743±0.032	0.758±0.026	0.749±0.012	0.77±0.013
AdamW	2e-5	0.3	0.748±0.036	0.755±0.026	0.751±0.013	0.767±0.019
AdamW	3e-5	0.1	0.756±0.019	0.755±0.016	0.755±0.008	0.773±0.012
AdamW	3e-5	0.2	0.745±0.022	0.765±0.012	0.755±0.01	0.772±0.013
AdamW	3e-5	0.3	0.764±0.038	0.745±0.039	0.753±0.014	0.773±0.014
Adam	1e-5	0.01	0.742±0.025	0.709±0.034	0.724±0.019	0.748±0.017
Adam	1e-5	0.02	0.736±0.023	0.66±0.045	0.695±0.026	0.715±0.03
Adam	1e-5	0.03	0.703±0.052	0.599±0.072	0.643±0.047	0.659±0.053
Adam	2e-5	0.01	0.716±0.038	0.607±0.083	0.652±0.044	0.67±0.061
Adam	2e-5	0.02	0.687±0.043	0.652±0.061	0.665±0.025	0.688±0.048
Adam	2e-5	0.03	0.681±0.03	0.632±0.027	0.655±0.005	0.688±0.013
Adam	3e-5	0.01	0.673±0.036	0.675±0.038	0.672±0.008	0.703±0.009
Adam	3e-5	0.02	0.651±0.058	0.625±0.053	0.634±0.01	0.665±0.016
Adam	3e-5	0.03	0.614±0.042	0.618±0.054	0.612±0.016	0.642±0.017

Table 4.15. The mean and std of evaluation metrics in the dev set. (cont.)

Optimizer	**LR**	**Weight Decay**	**Micro Precision**	**Micro Recall**	**Micro-averaged F1-score**	**Macro-averaged F1-score**
Adam	4e-5	0.01	0.653±0.043	0.631±0.045	0.64±0.011	0.655±0.041
Adam	4e-5	0.02	0.601±0.076	0.609±0.084	0.595±0.028	0.628±0.018
Adam	4e-5	0.03	0.601±0.03	0.509±0.182	0.581±0.016	0.632±0.092
Adam	5e-5	0.01	0.626±0.073	0.613±0.096	0.609±0.042	0.637±0.04
Adam	5e-5	0.02	0.604±0.064	0.436±0.237	0.567±0.028	0.648±0.127
Adam	5e-5	0.03	0.56±0.039	0.51±0.185	0.561±0.007	0.62±0.092
SGD	1e-5	0.01	0.549±0.028	0.292±0.073	0.374±0.073	0.456±0.069
SGD	1e-5	0.02	0.535±0.041	0.297±0.065	0.375±0.061	0.455±0.069
SGD	1e-5	0.03	0.547±0.026	0.318±0.022	0.401±0.018	0.454±0.026
SGD	2e-5	0.01	0.614±0.072	0.45±0.123	0.514±0.108	0.56±0.088
SGD	2e-5	0.02	0.635±0.03	0.505±0.082	0.559±0.066	0.587±0.086
SGD	2e-5	0.03	0.586±0.084	0.417±0.128	0.483±0.12	0.596±0.057
SGD	3e-5	0.01	0.67±0.033	0.512±0.187	0.614±0.039	0.631±0.095
SGD	3e-5	0.02	0.652±0.035	0.556±0.079	0.598±0.059	0.578±0.061
SGD	3e-5	0.03	0.653±0.055	0.549±0.114	0.593±0.098	0.62±0.051
SGD	4e-5	0.01	0.678±0.045	0.529±0.212	0.626±0.081	0.659±0.11
SGD	4e-5	0.02	0.677±0.063	0.606±0.075	0.639±0.069	0.614±0.089
SGD	4e-5	0.03	0.661±0.016	0.545±0.192	0.632±0.014	0.64±0.091
SGD	5e-5	0.01	0.701±0.061	0.604±0.217	0.685±0.051	0.718±0.095
SGD	5e-5	0.02	0.695±0.05	0.642±0.063	0.667±0.056	0.677±0.059
SGD	5e-5	0.03	0.669±0.033	0.612±0.052	0.638±0.04	0.639±0.044

- Test Set Results

Table 4.16. The mean and std of evaluation metrics in the test set with parameters used in multi-class relation extraction with BioBERT using the shortest subsequence between chemical and protein entity pairs. LR is the learning rate in the table.

Optimizer	**LR**	**Weight Decay**	**Micro Precision**	**Micro Recall**	**Micro-averaged F1-score**	**Macro-averaged F1-score**
AdamW	1e-5	0.1	0.739±0.026	0.706±0.036	0.721±0.01	0.734±0.009
AdamW	1e-5	0.2	0.73±0.031	0.72±0.032	0.724±0.007	0.737±0.008
AdamW	1e-5	0.3	0.734±0.02	0.717±0.011	0.725±0.009	0.739±0.009
AdamW	2e-5	0.1	0.737±0.017	0.719±0.015	0.728±0.009	0.741±0.009
AdamW	2e-5	0.2	0.737±0.029	0.723±0.024	0.729±0.007	0.74±0.01
AdamW	2e-5	0.3	0.739±0.029	0.718±0.023	0.728±0.01	0.74±0.013
AdamW	3e-5	0.1	0.746±0.018	0.718±0.013	0.732±0.007	0.742±0.01
AdamW	3e-5	0.2	0.734±0.017	0.727±0.015	0.73±0.007	0.743±0.007
AdamW	3e-5	0.3	0.747±0.037	0.706±0.036	0.725±0.012	0.738±0.011
Adam	1e-5	0.01	0.721±0.022	0.671±0.038	0.695±0.021	0.715±0.022
Adam	1e-5	0.02	0.728±0.026	0.633±0.035	0.676±0.018	0.697±0.021
Adam	1e-5	0.03	0.701±0.052	0.58±0.075	0.63±0.05	0.63±0.07
Adam	2e-5	0.01	0.711±0.04	0.587±0.085	0.638±0.046	0.652±0.063
Adam	2e-5	0.02	0.687±0.037	0.626±0.066	0.651±0.031	0.669±0.045
Adam	2e-5	0.03	0.684±0.021	0.604±0.027	0.641±0.01	0.667±0.016
Adam	3e-5	0.01	0.681±0.038	0.646±0.04	0.661±0.007	0.687±0.009
Adam	3e-5	0.02	0.659±0.045	0.584±0.059	0.615±0.018	0.647±0.021
Adam	3e-5	0.03	0.62±0.039	0.579±0.053	0.596±0.018	0.625±0.02

Table 4.16. The mean and std of evaluation metrics in the test set. (cont.)

Optimizer	**LR**	**Weight Decay**	**Micro Precision**	**Micro Recall**	**Micro-averaged F1-score**	**Macro-averaged F1-score**
Adam	4e-5	0.01	0.665±0.036	0.599±0.037	0.628±0.013	0.645±0.04
Adam	4e-5	0.02	0.611±0.067	0.561±0.085	0.576±0.037	0.616±0.023
Adam	4e-5	0.03	0.598±0.028	0.469±0.169	0.556±0.022	0.615±0.098
Adam	5e-5	0.01	0.641±0.06	0.578±0.095	0.598±0.05	0.628±0.046
Adam	5e-5	0.02	0.609±0.063	0.4±0.217	0.544±0.023	0.637±0.13
Adam	5e-5	0.03	0.566±0.035	0.457±0.169	0.532±0.015	0.604±0.096
SGD	1e-5	0.01	0.547±0.027	0.317±0.07	0.395±0.065	0.472±0.076
SGD	1e-5	0.02	0.534±0.029	0.323±0.069	0.396±0.061	0.458±0.068
SGD	1e-5	0.03	0.54±0.024	0.341±0.025	0.417±0.019	0.45±0.026
SGD	2e-5	0.01	0.609±0.058	0.452±0.1	0.516±0.085	0.548±0.078
SGD	2e-5	0.02	0.621±0.021	0.495±0.077	0.548±0.059	0.567±0.081
SGD	2e-5	0.03	0.581±0.065	0.418±0.111	0.482±0.101	0.559±0.054
SGD	3e-5	0.01	0.661±0.034	0.498±0.182	0.601±0.033	0.612±0.1
SGD	3e-5	0.02	0.643±0.034	0.539±0.075	0.584±0.055	0.568±0.042
SGD	3e-5	0.03	0.636±0.052	0.525±0.098	0.573±0.084	0.575±0.058
SGD	4e-5	0.01	0.674±0.039	0.512±0.201	0.614±0.072	0.633±0.118
SGD	4e-5	0.02	0.671±0.058	0.587±0.063	0.626±0.059	0.595±0.083
SGD	4e-5	0.03	0.655±0.02	0.527±0.186	0.618±0.016	0.623±0.097
SGD	5e-5	0.01	0.688±0.059	0.586±0.208	0.668±0.044	0.692±0.095
SGD	5e-5	0.02	0.686±0.041	0.62±0.059	0.651±0.05	0.652±0.058
SGD	5e-5	0.03	0.66±0.033	0.593±0.05	0.624±0.037	0.618±0.05

4.3.3. Multi-class Relation Extraction with The Shortest Subsequence Including Chemical and Protein Entity Pair and Their Parent Node

- Development Set Results

Table 4.17. The mean and std of evaluation metrics in the dev set with parameters used in multi-class relation extraction with BioBERT using the shortest subsequence including chemical and protein entity pairs and their parent node.

Optimizer	**LR**	**Weight Decay**	**Micro Precision**	**Micro Recall**	**Micro-averaged F1-score**	**Macro-averaged F1-score**
AdamW	1e-5	0.1	0.731±0.027	0.762±0.027	0.746±0.007	0.769±0.01
AdamW	1e-5	0.2	0.734±0.012	0.759±0.022	0.746±0.007	0.773±0.008
AdamW	1e-5	0.3	0.737±0.032	0.751±0.03	0.743±0.007	0.768±0.008
AdamW	2e-5	0.1	0.745±0.031	0.769±0.028	0.756±0.006	0.779±0.007
AdamW	2e-5	0.2	0.748±0.021	0.767±0.011	0.757±0.009	0.783±0.008
AdamW	2e-5	0.3	0.74±0.014	0.774±0.013	0.757±0.008	0.779±0.009
AdamW	3e-5	0.1	0.755±0.02	0.762±0.02	0.758±0.006	0.78±0.006
AdamW	3e-5	0.2	0.752±0.028	0.76±0.025	0.756±0.018	0.771±0.038
AdamW	3e-5	0.3	0.753±0.011	0.772±0.013	0.762±0.006	0.786±0.005
Adam	1e-5	0.01	0.735±0.045	0.704±0.077	0.715±0.033	0.734±0.036
Adam	1e-5	0.02	0.718±0.025	0.676±0.031	0.696±0.021	0.707±0.046
Adam	1e-5	0.03	0.716±0.033	0.617±0.037	0.662±0.025	0.668±0.05
Adam	2e-5	0.01	0.737±0.034	0.712±0.033	0.723±0.008	0.746±0.01
Adam	2e-5	0.02	0.675±0.042	0.68±0.03	0.676±0.012	0.7±0.012
Adam	2e-5	0.03	0.697±0.053	0.595±0.092	0.634±0.046	0.654±0.051
Adam	3e-5	0.01	0.692±0.038	0.673±0.05	0.68±0.016	0.705±0.02
Adam	3e-5	0.02	0.668±0.098	0.532±0.219	0.61±0.056	0.656±0.099
Adam	3e-5	0.03	0.611±0.064	0.61±0.08	0.603±0.021	0.625±0.031

Table 4.17. The mean and std of evaluation metrics in the dev set. (cont.)

Optimizer	**LR**	**Weight Decay**	**Micro Precision**	**Micro Recall**	**Micro-averaged F1-score**	**Macro-averaged F1-score**
Adam	4e-5	0.01	0.644±0.019	0.669±0.008	0.656±0.009	0.687±0.01
Adam	4e-5	0.02	0.606±0.061	0.605±0.06	0.6±0.019	0.623±0.048
Adam	4e-5	0.03	0.607±0.041	0.498±0.186	0.575±0.03	0.61±0.11
Adam	5e-5	0.01	0.643±0.062	0.604±0.083	0.615±0.026	0.639±0.029
Adam	5e-5	0.02	0.608±0.06	0.416±0.26	0.537±0.122	0.64±0.136
Adam	5e-5	0.03	0.563±0.049	0.471±0.181	0.536±0.031	0.597±0.104
SGD	1e-5	0.01	0.551±0.044	0.297±0.037	0.385±0.039	0.476±0.096
SGD	1e-5	0.02	0.574±0.07	0.226±0.114	0.3±0.142	0.423±0.049
SGD	1e-5	0.03	0.559±0.074	0.232±0.103	0.309±0.123	0.467±0.055
SGD	2e-5	0.01	0.647±0.041	0.491±0.078	0.557±0.066	0.587±0.052
SGD	2e-5	0.02	0.637±0.049	0.461±0.097	0.531±0.087	0.583±0.064
SGD	2e-5	0.03	0.612±0.028	0.44±0.099	0.507±0.084	0.53±0.072
SGD	3e-5	0.01	0.661±0.038	0.532±0.142	0.58±0.104	0.573±0.082
SGD	3e-5	0.02	0.657±0.055	0.507±0.142	0.562±0.106	0.581±0.073
SGD	3e-5	0.03	0.662±0.016	0.576±0.016	0.616±0.013	0.555±0.018
SGD	4e-5	0.01	0.705±0.019	0.638±0.069	0.668±0.045	0.674±0.053
SGD	4e-5	0.02	0.691±0.029	0.602±0.08	0.641±0.057	0.626±0.058
SGD	4e-5	0.03	0.646±0.062	0.501±0.196	0.597±0.078	0.625±0.111
SGD	5e-5	0.01	0.699±0.075	0.607±0.141	0.643±0.106	0.644±0.13
SGD	5e-5	0.02	0.704±0.015	0.668±0.02	0.685±0.01	0.705±0.012
SGD	5e-5	0.03	0.665±0.053	0.595±0.064	0.627±0.055	0.625±0.073

- Test Set Results

Table 4.18. The mean and std of evaluation metrics in the test set with parameters used in multi-class relation extraction with BioBERT using the shortest subsequence including chemical and protein entity pairs and their parent node.

Optimizer	**LR**	**Weight Decay**	**Micro Precision**	**Micro Recall**	**Micro-averaged F1-score**	**Macro-averaged F1-score**
AdamW	1e-5	0.1	0.723±0.025	0.724±0.02	0.722±0.006	0.737±0.008
AdamW	1e-5	0.2	0.733±0.013	0.725±0.02	0.728±0.007	0.744±0.008
AdamW	1e-5	0.3	0.733±0.032	0.717±0.029	0.724±0.006	0.74±0.007
AdamW	2e-5	0.1	0.741±0.032	0.722±0.028	0.73±0.005	0.746±0.006
AdamW	2e-5	0.2	0.741±0.015	0.726±0.013	0.733±0.004	0.749±0.003
AdamW	2e-5	0.3	0.732±0.017	0.733±0.014	0.732±0.007	0.749±0.009
AdamW	3e-5	0.1	0.749±0.015	0.719±0.016	0.734±0.006	0.749±0.005
AdamW	3e-5	0.2	0.732±0.017	0.733±0.014	0.732±0.007	0.749±0.009
AdamW	3e-5	0.3	0.743±0.014	0.726±0.011	0.734±0.007	0.749±0.006
Adam	1e-5	0.01	0.725±0.046	0.673±0.068	0.694±0.027	0.713±0.025
Adam	1e-5	0.02	0.708±0.023	0.656±0.029	0.68±0.019	0.69±0.044
Adam	1e-5	0.03	0.712±0.031	0.601±0.038	0.651±0.024	0.662±0.05
Adam	2e-5	0.01	0.717±0.027	0.678±0.035	0.696±0.008	0.716±0.006
Adam	2e-5	0.02	0.673±0.04	0.663±0.032	0.666±0.008	0.689±0.005
Adam	2e-5	0.03	0.692±0.053	0.568±0.095	0.616±0.052	0.632±0.061
Adam	3e-5	0.01	0.681±0.033	0.644±0.043	0.66±0.012	0.683±0.019
Adam	3e-5	0.02	0.657±0.095	0.503±0.207	0.587±0.056	0.637±0.102
Adam	3e-5	0.03	0.613±0.058	0.573±0.075	0.586±0.026	0.611±0.035

Table 4.18. The mean and std of evaluation metrics in the test set. (cont.)

Optimizer	**LR**	**Weight Decay**	**Micro Precision**	**Micro Recall**	**Micro-averaged F1-score**	**Macro-averaged F1-score**
Adam	4e-5	0.01	0.641±0.011	0.631±0.016	0.636±0.005	0.665±0.005
Adam	4e-5	0.02	0.59±0.054	0.565±0.058	0.573±0.022	0.6±0.047
Adam	4e-5	0.03	0.591±0.038	0.458±0.169	0.543±0.026	0.588±0.114
Adam	5e-5	0.01	0.639±0.055	0.573±0.078	0.598±0.03	0.624±0.033
Adam	5e-5	0.02	0.599±0.071	0.38±0.238	0.506±0.117	0.621±0.144
Adam	5e-5	0.03	0.564±0.051	0.432±0.163	0.514±0.021	0.593±0.105
SGD	1e-5	0.01	0.559±0.032	0.32±0.032	0.406±0.031	0.476±0.092
SGD	1e-5	0.02	0.589±0.074	0.248±0.117	0.325±0.141	0.44±0.052
SGD	1e-5	0.03	0.57±0.078	0.253±0.107	0.33±0.124	0.462±0.069
SGD	2e-5	0.01	0.636±0.033	0.483±0.062	0.548±0.052	0.541±0.04
SGD	2e-5	0.02	0.627±0.037	0.454±0.087	0.524±0.075	0.555±0.068
SGD	2e-5	0.03	0.612±0.023	0.44±0.088	0.508±0.073	0.53±0.041
SGD	3e-5	0.01	0.652±0.034	0.522±0.128	0.572±0.093	0.555±0.073
SGD	3e-5	0.02	0.651±0.048	0.499±0.134	0.555±0.098	0.541±0.079
SGD	3e-5	0.03	0.653±0.015	0.556±0.012	0.601±0.009	0.549±0.028
SGD	4e-5	0.01	0.693±0.018	0.616±0.063	0.651±0.042	0.646±0.053
SGD	4e-5	0.02	0.685±0.023	0.584±0.07	0.629±0.05	0.611±0.053
SGD	4e-5	0.03	0.648±0.059	0.491±0.189	0.591±0.071	0.611±0.111
SGD	5e-5	0.01	0.691±0.061	0.584±0.129	0.627±0.096	0.622±0.124
SGD	5e-5	0.02	0.705±0.013	0.648±0.021	0.675±0.01	0.691±0.014
SGD	5e-5	0.03	0.662±0.047	0.579±0.059	0.617±0.049	0.616±0.055

4.4. Multi-class Relation Extraction Experiments with SciBERT

4.4.1. Multi-class Relation Extraction with Sentence-based Input Representation

- Development Set Results

Table 4.19. The mean and standard deviation (std) of evaluation metrics in development set with parameters used in multi-class relation extraction with SciBERT using entire sentence. LR is the learning rate in the table.

Optimizer	**LR**	**Weight Decay**	**Micro Precision**	**Micro Recall**	**Micro-averaged F1-score**	**Macro-averaged F1-score**
AdamW	1e-5	0.1	0.727±0.018	0.773±0.021	0.749±0.006	0.778±0.007
AdamW	1e-5	0.2	0.721±0.011	0.761±0.028	0.74±0.013	0.767±0.018
AdamW	2e-5	0.1	0.767±0.029	0.75±0.043	0.757±0.012	0.782±0.017
AdamW	2e-5	0.2	0.737±0.022	0.78±0.018	0.758±0.015	0.785±0.019
AdamW	3e-5	0.1	0.753±0.023	0.774±0.017	0.763±0.01	0.785±0.013
AdamW	3e-5	0.2	0.745±0.021	0.779±0.019	0.761±0.008	0.788±0.01
Adam	1e-5	0.01	0.701±0.049	0.725±0.039	0.711±0.023	0.732±0.027
Adam	1e-5	0.02	0.692±0.02	0.68±0.038	0.685±0.016	0.701±0.027
Adam	1e-5	0.03	0.657±0.046	0.643±0.069	0.646±0.032	0.664±0.032
Adam	2e-5	0.01	0.707±0.015	0.718±0.037	0.712±0.017	0.726±0.024
Adam	2e-5	0.02	0.685±0.038	0.635±0.056	0.657±0.021	0.652±0.044
Adam	2e-5	0.03	0.647±0.04	0.633±0.056	0.637±0.017	0.638±0.042
Adam	3e-5	0.01	0.662±0.054	0.708±0.049	0.681±0.021	0.692±0.026
Adam	3e-5	0.02	0.626±0.042	0.622±0.056	0.62±0.016	0.633±0.017
Adam	3e-5	0.03	0.622±0.054	0.568±0.071	0.588±0.028	0.589±0.06

Table 4.19. The mean and std of evaluation metrics in the dev set. (cont.)

Optimizer	**LR**	**Weight Decay**	**Micro Precision**	**Micro Recall**	**Micro-averaged F1-score**	**Macro-averaged F1-score**
Adam	4e-5	0.01	0.666±0.039	0.65±0.065	0.654±0.021	0.654±0.051
Adam	4e-5	0.02	0.584±0.012	0.617±0.02	0.6±0.009	0.614±0.01
Adam	4e-5	0.03	0.546±0.031	0.545±0.037	0.544±0.013	0.556±0.012
Adam	5e-5	0.01	0.629±0.026	0.631±0.038	0.629±0.012	0.639±0.017
Adam	5e-5	0.02	0.566±0.075	0.539±0.106	0.539±0.044	0.546±0.047
Adam	5e-5	0.03	0.549±0.069	0.452±0.102	0.483±0.048	0.494±0.064
SGD	1e-5	0.01	0.495±0.065	0.204±0.17	0.253±0.184	0.488±0.065
SGD	1e-5	0.02	0.512±0.063	0.196±0.167	0.277±0.177	0.527±0.149
SGD	1e-5	0.03	0.457±0.171	0.269±0.177	0.356±0.163	0.568±0.128
SGD	2e-5	0.01	0.549±0.097	0.391±0.251	0.461±0.194	0.579±0.144
SGD	2e-5	0.02	0.584±0.075	0.484±0.127	0.525±0.106	0.56±0.062
SGD	2e-5	0.03	0.544±0.088	0.308±0.189	0.406±0.147	0.589±0.129
SGD	3e-5	0.01	0.616±0.07	0.614±0.137	0.611±0.106	0.666±0.037
SGD	3e-5	0.02	0.607±0.085	0.431±0.261	0.515±0.195	0.648±0.1
SGD	3e-5	0.03	0.562±0.094	0.419±0.254	0.492±0.188	0.631±0.106
SGD	4e-5	0.01	0.579±0.088	0.553±0.238	0.594±0.118	0.664±0.106
SGD	4e-5	0.02	0.604±0.096	0.546±0.164	0.569±0.135	0.606±0.113
SGD	4e-5	0.03	0.52±0.128	0.446±0.225	0.469±0.191	0.594±0.1
SGD	5e-5	0.01	0.631±0.056	0.595±0.223	0.585±0.21	0.643±0.11
SGD	5e-5	0.02	0.604±0.086	0.46±0.282	0.581±0.122	0.676±0.138
SGD	5e-5	0.03	0.589±0.055	0.537±0.169	0.552±0.122	0.564±0.12

- Test Set Results

Table 4.20. The mean and standard deviation (std) of evaluation metrics in the test set with parameters used in multi-class relation extraction with SciBERT using the entire sentence. LR is the learning rate in the table.

Optimizer	**LR**	**Weight Decay**	**Micro Precision**	**Micro Recall**	**Micro-averaged F1-score**	**Macro-averaged F1-score**
AdamW	1e-5	0.1	0.743±0.017	0.732±0.018	0.737±0.006	0.757±0.007
AdamW	1e-5	0.2	0.739±0.013	0.723±0.023	0.731±0.008	0.75±0.01
AdamW	2e-5	0.1	0.771±0.025	0.714±0.038	0.74±0.014	0.759±0.011
AdamW	2e-5	0.2	0.748±0.018	0.735±0.015	0.741±0.011	0.76±0.011
AdamW	3e-5	0.1	0.758±0.024	0.729±0.013	0.743±0.011	0.762±0.013
AdamW	3e-5	0.2	0.747±0.019	0.739±0.018	0.743±0.007	0.762±0.007
Adam	1e-5	0.01	0.713±0.048	0.689±0.034	0.699±0.016	0.721±0.02
Adam	1e-5	0.02	0.711±0.015	0.658±0.047	0.682±0.023	0.7±0.033
Adam	1e-5	0.03	0.676±0.049	0.626±0.073	0.645±0.034	0.663±0.037
Adam	2e-5	0.01	0.716±0.016	0.684±0.03	0.699±0.013	0.72±0.02
Adam	2e-5	0.02	0.691±0.037	0.606±0.058	0.643±0.024	0.651±0.046
Adam	2e-5	0.03	0.658±0.03	0.605±0.058	0.628±0.025	0.636±0.052
Adam	3e-5	0.01	0.672±0.048	0.675±0.056	0.67±0.019	0.691±0.023
Adam	3e-5	0.02	0.628±0.041	0.586±0.059	0.602±0.016	0.628±0.012
Adam	3e-5	0.03	0.623±0.056	0.528±0.063	0.567±0.028	0.571±0.054

Table 4.20. The mean and std of evaluation metrics in the test set. (cont.)

Optimizer	**LR**	**Weight Decay**	**Micro Precision**	**Micro Recall**	**Micro-averaged F1-score**	**Macro-averaged F1-score**
Adam	4e-5	0.01	0.672±0.034	0.613±0.063	0.638±0.025	0.651±0.049
Adam	4e-5	0.02	0.589±0.011	0.571±0.015	0.58±0.008	0.603±0.009
Adam	4e-5	0.03	0.543±0.03	0.496±0.037	0.517±0.016	0.528±0.021
Adam	5e-5	0.01	0.637±0.024	0.599±0.037	0.616±0.01	0.639±0.018
Adam	5e-5	0.02	0.568±0.072	0.497±0.104	0.517±0.042	0.535±0.047
Adam	5e-5	0.03	0.542±0.058	0.417±0.101	0.46±0.058	0.468±0.074
SGD	1e-5	0.01	0.463±0.174	0.218±0.174	0.303±0.178	0.533±0.139
SGD	1e-5	0.02	0.525±0.062	0.207±0.167	0.295±0.181	0.527±0.146
SGD	1e-5	0.03	0.55±0.071	0.277±0.177	0.334±0.194	0.505±0.083
SGD	2e-5	0.01	0.586±0.088	0.388±0.233	0.474±0.18	0.572±0.144
SGD	2e-5	0.02	0.608±0.075	0.477±0.109	0.532±0.095	0.549±0.063
SGD	2e-5	0.03	0.522±0.201	0.3±0.179	0.41±0.146	0.572±0.139
SGD	3e-5	0.01	0.634±0.076	0.592±0.123	0.61±0.101	0.656±0.039
SGD	3e-5	0.02	0.634±0.089	0.408±0.243	0.51±0.191	0.637±0.102
SGD	3e-5	0.03	0.577±0.095	0.403±0.241	0.488±0.186	0.622±0.105
SGD	4e-5	0.01	0.611±0.087	0.53±0.221	0.598±0.11	0.651±0.113
SGD	4e-5	0.02	0.625±0.094	0.526±0.146	0.568±0.125	0.599±0.115
SGD	4e-5	0.03	0.551±0.129	0.443±0.204	0.483±0.181	0.602±0.1
SGD	5e-5	0.01	0.661±0.051	0.573±0.212	0.585±0.205	0.649±0.092
SGD	5e-5	0.02	0.624±0.077	0.44±0.265	0.578±0.109	0.67±0.138
SGD	5e-5	0.03	0.623±0.056	0.518±0.151	0.557±0.113	0.572±0.116

4.4.2. Multi-class Relation Extraction with The Shortest Sequence Between Chemical and Protein Entity Pairs

- Development Set Results

Table 4.21. The mean and std of evaluation metrics in the dev set with parameters used in multi-class relation extraction with SciBERT using the shortest subsequence between chemical and protein entity pairs. LR is the learning rate in the table.

Optimizer	**LR**	**Weight Decay**	**Micro Precision**	**Micro Recall**	**Micro-averaged F1-score**	**Macro-averaged F1-score**
AdamW	1e-5	0.1	0.721±0.019	0.74±0.011	0.73±0.007	0.759±0.009
AdamW	1e-5	0.2	0.719±0.028	0.739±0.017	0.728±0.011	0.755±0.01
AdamW	1e-5	0.3	0.729±0.017	0.734±0.023	0.731±0.006	0.758±0.006
AdamW	2e-5	0.1	0.735±0.025	0.734±0.032	0.733±0.008	0.758±0.013
AdamW	2e-5	0.2	0.722±0.039	0.745±0.033	0.732±0.014	0.757±0.014
AdamW	2e-5	0.3	0.735±0.022	0.748±0.02	0.741±0.006	0.762±0.007
AdamW	3e-5	0.1	0.723±0.038	0.753±0.036	0.736±0.017	0.761±0.013
AdamW	3e-5	0.2	0.733±0.041	0.743±0.033	0.737±0.015	0.758±0.014
AdamW	3e-5	0.3	0.719±0.051	0.742±0.03	0.728±0.021	0.754±0.024
Adam	1e-5	0.01	0.712±0.047	0.697±0.029	0.703±0.024	0.737±0.021
Adam	1e-5	0.02	0.707±0.031	0.684±0.014	0.695±0.017	0.727±0.019
Adam	1e-5	0.03	0.687±0.04	0.642±0.05	0.661±0.028	0.684±0.055
Adam	2e-5	0.01	0.736±0.012	0.704±0.015	0.719±0.007	0.746±0.008
Adam	2e-5	0.02	0.686±0.064	0.646±0.086	0.658±0.044	0.687±0.035
Adam	2e-5	0.03	0.667±0.036	0.632±0.024	0.648±0.008	0.673±0.011
Adam	3e-5	0.01	0.692±0.013	0.674±0.015	0.683±0.008	0.711±0.011
Adam	3e-5	0.02	0.67±0.072	0.604±0.09	0.626±0.035	0.645±0.04
Adam	3e-5	0.03	0.663±0.092	0.553±0.114	0.587±0.047	0.59±0.064

Table 4.21. The mean and std of evaluation metrics in the dev set. (cont.)

Optimizer	**LR**	**Weight Decay**	**Micro Precision**	**Micro Recall**	**Micro-averaged F1-score**	**Macro-averaged F1-score**
Adam	4e-5	0.01	0.683±0.043	0.613±0.056	0.643±0.018	0.667±0.03
Adam	4e-5	0.02	0.616±0.015	0.598±0.018	0.606±0.005	0.631±0.007
Adam	4e-5	0.03	0.643±0.061	0.522±0.078	0.568±0.031	0.586±0.064
Adam	5e-5	0.01	0.662±0.073	0.593±0.089	0.616±0.043	0.642±0.039
Adam	5e-5	0.02	0.613±0.08	0.548±0.115	0.563±0.056	0.58±0.074
Adam	5e-5	0.03	0.603±0.066	0.529±0.083	0.554±0.025	0.586±0.013
SGD	1e-5	0.01	0.601±0.024	0.46±0.164	0.552±0.022	0.615±0.102
SGD	1e-5	0.02	0.571±0.076	0.448±0.107	0.499±0.093	0.576±0.08
SGD	1e-5	0.03	0.597±0.018	0.472±0.055	0.526±0.042	0.56±0.052
SGD	2e-5	0.01	0.626±0.039	0.558±0.092	0.588±0.069	0.601±0.067
SGD	2e-5	0.02	0.656±0.025	0.582±0.058	0.616±0.041	0.631±0.046
SGD	2e-5	0.03	0.646±0.026	0.586±0.029	0.614±0.019	0.616±0.034
SGD	3e-5	0.01	0.625±0.056	0.582±0.067	0.602±0.058	0.619±0.08
SGD	3e-5	0.02	0.637±0.081	0.574±0.114	0.603±0.099	0.639±0.075
SGD	3e-5	0.03	0.641±0.081	0.588±0.058	0.613±0.067	0.645±0.064
SGD	4e-5	0.01	0.673±0.036	0.655±0.049	0.663±0.033	0.691±0.054
SGD	4e-5	0.02	0.652±0.072	0.591±0.14	0.616±0.119	0.676±0.063
SGD	4e-5	0.03	0.666±0.042	0.615±0.057	0.638±0.038	0.665±0.059
SGD	5e-5	0.01	0.691±0.036	0.677±0.028	0.683±0.02	0.72±0.021
SGD	5e-5	0.02	0.628±0.113	0.47±0.17	0.531±0.157	0.617±0.097
SGD	5e-5	0.03	0.66±0.073	0.521±0.206	0.615±0.084	0.671±0.101

- Test Set Results

Table 4.22. The mean and std of evaluation metrics in the test set with parameters used in multi-class relation extraction with SciBERT using the shortest subsequence between chemical and protein entity pairs. LR is the learning rate in the table.

Optimizer	**LR**	**Weight Decay**	**Micro Precision**	**Micro Recall**	**Micro-averaged F1-score**	**Macro-averaged F1-score**
AdamW	1e-5	0.1	0.708±0.017	0.71±0.007	0.709±0.006	0.726±0.008
AdamW	1e-5	0.2	0.705±0.03	0.708±0.017	0.706±0.009	0.723±0.01
AdamW	1e-5	0.3	0.713±0.015	0.696±0.025	0.704±0.009	0.723±0.007
AdamW	2e-5	0.1	0.726±0.028	0.699±0.026	0.711±0.005	0.73±0.004
AdamW	2e-5	0.2	0.708±0.039	0.705±0.025	0.705±0.011	0.725±0.012
AdamW	2e-5	0.3	0.724±0.021	0.705±0.023	0.714±0.007	0.731±0.008
AdamW	3e-5	0.1	0.71±0.045	0.71±0.038	0.708±0.018	0.73±0.014
AdamW	3e-5	0.2	0.723±0.044	0.702±0.026	0.711±0.013	0.728±0.011
AdamW	3e-5	0.3	0.712±0.044	0.695±0.03	0.702±0.02	0.722±0.017
Adam	1e-5	0.01	0.71±0.046	0.667±0.028	0.686±0.019	0.712±0.018
Adam	1e-5	0.02	0.708±0.031	0.652±0.016	0.679±0.019	0.706±0.021
Adam	1e-5	0.03	0.686±0.041	0.62±0.053	0.649±0.029	0.668±0.052
Adam	2e-5	0.01	0.718±0.01	0.664±0.015	0.69±0.005	0.715±0.006
Adam	2e-5	0.02	0.685±0.05	0.617±0.096	0.642±0.057	0.675±0.039
Adam	2e-5	0.03	0.674±0.027	0.616±0.03	0.643±0.007	0.669±0.011
Adam	3e-5	0.01	0.689±0.012	0.64±0.013	0.663±0.006	0.692±0.01
Adam	3e-5	0.02	0.678±0.076	0.578±0.097	0.613±0.041	0.635±0.048
Adam	3e-5	0.03	0.661±0.077	0.523±0.119	0.569±0.06	0.583±0.075

Table 4.22. The mean and std of evaluation metrics in the test set. (cont.)

Optimizer	**LR**	**Weight Decay**	**Micro Precision**	**Micro Recall**	**Micro-averaged F1-score**	**Macro-averaged F1-score**
Adam	4e-5	0.01	0.696±0.041	0.584±0.059	0.631±0.025	0.656±0.036
Adam	4e-5	0.02	0.633±0.014	0.568±0.019	0.598±0.005	0.631±0.004
Adam	4e-5	0.03	0.644±0.06	0.492±0.076	0.55±0.034	0.58±0.066
Adam	5e-5	0.01	0.669±0.059	0.558±0.09	0.6±0.051	0.634±0.043
Adam	5e-5	0.02	0.625±0.076	0.522±0.116	0.554±0.061	0.58±0.085
Adam	5e-5	0.03	0.599±0.066	0.495±0.087	0.532±0.032	0.58±0.02
SGD	1e-5	0.01	0.611±0.026	0.457±0.163	0.554±0.016	0.563±0.12
SGD	1e-5	0.02	0.573±0.075	0.447±0.101	0.5±0.089	0.554±0.079
SGD	1e-5	0.03	0.599±0.018	0.471±0.047	0.527±0.036	0.547±0.051
SGD	2e-5	0.01	0.628±0.037	0.552±0.083	0.586±0.063	0.581±0.06
SGD	2e-5	0.02	0.653±0.025	0.57±0.055	0.608±0.04	0.608±0.044
SGD	2e-5	0.03	0.646±0.027	0.567±0.03	0.603±0.017	0.588±0.033
SGD	3e-5	0.01	0.633±0.049	0.578±0.065	0.603±0.053	0.6±0.08
SGD	3e-5	0.02	0.637±0.08	0.567±0.102	0.599±0.092	0.626±0.068
SGD	3e-5	0.03	0.642±0.075	0.574±0.043	0.605±0.054	0.613±0.062
SGD	4e-5	0.01	0.675±0.033	0.643±0.046	0.658±0.03	0.674±0.045
SGD	4e-5	0.02	0.657±0.07	0.579±0.13	0.612±0.112	0.66±0.058
SGD	4e-5	0.03	0.665±0.046	0.592±0.055	0.624±0.038	0.642±0.056
SGD	5e-5	0.01	0.692±0.03	0.66±0.028	0.675±0.014	0.698±0.018
SGD	5e-5	0.02	0.636±0.101	0.463±0.157	0.529±0.146	0.606±0.095
SGD	5e-5	0.03	0.662±0.066	0.504±0.197	0.604±0.077	0.654±0.103

4.4.3. Multi-class Relation Extraction with The Shortest Subsequence Including Chemical and Protein Entity Pair and Their Parent Node

- Development Set Results

Table 4.23. The mean and std of evaluation metrics in the dev set with parameters used in multi-class relation extraction with SciBERT using the shortest subsequence including chemical and protein entity pairs and their parent node.

Optimizer	**LR**	**Weight Decay**	**Micro Precision**	**Micro Recall**	**Micro-averaged F1-score**	**Macro-averaged F1-score**
AdamW	1e-5	0.1	0.729±0.026	0.728±0.043	0.727±0.013	0.755±0.015
AdamW	1e-5	0.2	0.735±0.017	0.738±0.025	0.736±0.006	0.762±0.008
AdamW	1e-5	0.3	0.733±0.016	0.734±0.03	0.733±0.009	0.758±0.011
AdamW	2e-5	0.1	0.738±0.041	0.741±0.043	0.738±0.012	0.76±0.015
AdamW	2e-5	0.2	0.739±0.019	0.753±0.024	0.745±0.008	0.771±0.009
AdamW	2e-5	0.3	0.733±0.028	0.749±0.015	0.74±0.014	0.768±0.015
AdamW	3e-5	0.1	0.745±0.019	0.748±0.016	0.746±0.006	0.771±0.005
AdamW	3e-5	0.2	0.751±0.033	0.746±0.043	0.747±0.012	0.771±0.012
AdamW	3e-5	0.3	0.743±0.03	0.738±0.032	0.74±0.015	0.76±0.02
Adam	1e-5	0.01	0.719±0.026	0.697±0.049	0.706±0.02	0.732±0.027
Adam	1e-5	0.02	0.717±0.017	0.674±0.041	0.694±0.017	0.719±0.018
Adam	1e-5	0.03	0.681±0.041	0.668±0.046	0.672±0.02	0.693±0.027
Adam	2e-5	0.01	0.719±0.02	0.699±0.026	0.708±0.011	0.731±0.009
Adam	2e-5	0.02	0.717±0.045	0.635±0.04	0.671±0.011	0.693±0.017
Adam	2e-5	0.03	0.669±0.04	0.624±0.038	0.644±0.012	0.661±0.017
Adam	3e-5	0.01	0.685±0.022	0.691±0.018	0.688±0.008	0.712±0.01
Adam	3e-5	0.02	0.659±0.066	0.592±0.097	0.613±0.049	0.633±0.048
Adam	3e-5	0.03	0.635±0.063	0.59±0.053	0.607±0.018	0.617±0.024

Table 4.23. The mean and std of evaluation metrics in the dev set. (cont.)

Optimizer	**LR**	**Weight Decay**	**Micro Precision**	**Micro Recall**	**Micro-averaged F1-score**	**Macro-averaged F1-score**
Adam	4e-5	0.01	0.685±0.034	0.64±0.019	0.661±0.011	0.68±0.005
Adam	4e-5	0.02	0.632±0.048	0.583±0.051	0.603±0.014	0.62±0.02
Adam	4e-5	0.03	0.657±0.069	0.487±0.108	0.546±0.052	0.56±0.064
Adam	5e-5	0.01	0.651±0.043	0.585±0.06	0.613±0.024	0.626±0.044
Adam	5e-5	0.02	0.587±0.027	0.58±0.025	0.583±0.008	0.604±0.02
Adam	5e-5	0.03	0.625±0.03	0.488±0.059	0.545±0.033	0.551±0.06
SGD	1e-5	0.01	0.586±0.025	0.486±0.055	0.53±0.043	0.551±0.062
SGD	1e-5	0.02	0.588±0.051	0.468±0.084	0.519±0.071	0.543±0.062
SGD	1e-5	0.03	0.518±0.063	0.301±0.148	0.4±0.101	0.567±0.138
SGD	2e-5	0.01	0.592±0.106	0.511±0.2	0.578±0.095	0.623±0.117
SGD	2e-5	0.02	0.638±0.031	0.546±0.101	0.584±0.072	0.597±0.05
SGD	2e-5	0.03	0.641±0.032	0.557±0.067	0.593±0.043	0.585±0.053
SGD	3e-5	0.01	0.639±0.096	0.603±0.129	0.619±0.114	0.672±0.07
SGD	3e-5	0.02	0.648±0.026	0.628±0.053	0.637±0.038	0.653±0.055
SGD	3e-5	0.03	0.67±0.072	0.52±0.191	0.55±0.195	0.593±0.086
SGD	4e-5	0.01	0.62±0.104	0.532±0.168	0.564±0.134	0.62±0.102
SGD	4e-5	0.02	0.649±0.059	0.618±0.117	0.629±0.088	0.661±0.092
SGD	4e-5	0.03	0.65±0.058	0.589±0.083	0.614±0.055	0.637±0.059
SGD	5e-5	0.01	0.671±0.071	0.541±0.237	0.629±0.122	0.682±0.113
SGD	5e-5	0.02	0.657±0.047	0.58±0.208	0.65±0.045	0.688±0.095
SGD	5e-5	0.03	0.647±0.064	0.501±0.221	0.592±0.115	0.647±0.115

- Test Set Results

Table 4.24. The mean and std of evaluation metrics in the test set with parameters used in multi-class relation extraction with SciBERT using the shortest subsequence including chemical and protein entity pairs and their parent node.

Optimizer	**LR**	**Weight Decay**	**Micro Precision**	**Micro Recall**	**Micro-averaged F1-score**	**Macro-averaged F1-score**
AdamW	1e-5	0.1	0.722±0.029	0.695±0.035	0.707±0.008	0.726±0.008
AdamW	1e-5	0.2	0.725±0.02	0.696±0.019	0.709±0.003	0.731±0.003
AdamW	1e-5	0.3	0.718±0.018	0.698±0.022	0.708±0.005	0.727±0.006
AdamW	2e-5	0.1	0.723±0.035	0.697±0.041	0.708±0.011	0.729±0.01
AdamW	2e-5	0.2	0.724±0.018	0.707±0.019	0.715±0.006	0.735±0.008
AdamW	2e-5	0.3	0.718±0.028	0.703±0.021	0.71±0.015	0.731±0.013
AdamW	3e-5	0.1	0.731±0.02	0.707±0.014	0.719±0.006	0.739±0.004
AdamW	3e-5	0.2	0.734±0.033	0.702±0.041	0.716±0.012	0.736±0.009
AdamW	3e-5	0.3	0.73±0.024	0.7±0.03	0.714±0.011	0.731±0.014
Adam	1e-5	0.01	0.716±0.026	0.659±0.043	0.685±0.019	0.708±0.022
Adam	1e-5	0.02	0.717±0.015	0.642±0.041	0.676±0.021	0.705±0.018
Adam	1e-5	0.03	0.687±0.038	0.646±0.048	0.664±0.021	0.689±0.028
Adam	2e-5	0.01	0.709±0.017	0.66±0.025	0.683±0.01	0.713±0.007
Adam	2e-5	0.02	0.708±0.036	0.599±0.046	0.647±0.017	0.68±0.016
Adam	2e-5	0.03	0.669±0.03	0.597±0.044	0.629±0.015	0.656±0.021
Adam	3e-5	0.01	0.679±0.014	0.656±0.027	0.667±0.008	0.696±0.008
Adam	3e-5	0.02	0.659±0.056	0.556±0.1	0.593±0.059	0.621±0.057
Adam	3e-5	0.03	0.626±0.053	0.553±0.055	0.583±0.019	0.599±0.034

Table 4.24. The mean and std of evaluation metrics in the test set. (cont.)

Optimizer	**LR**	**Weight Decay**	**Micro Precision**	**Micro Recall**	**Micro-averaged F1-score**	**Macro-averaged F1-score**
Adam	4e-5	0.01	0.68±0.025	0.605±0.021	0.639±0.007	0.665±0.006
Adam	4e-5	0.02	0.627±0.041	0.544±0.047	0.58±0.016	0.595±0.032
Adam	4e-5	0.03	0.648±0.073	0.449±0.103	0.517±0.052	0.543±0.069
Adam	5e-5	0.01	0.65±0.034	0.564±0.06	0.601±0.027	0.614±0.051
Adam	5e-5	0.02	0.585±0.021	0.544±0.024	0.563±0.006	0.594±0.017
Adam	5e-5	0.03	0.604±0.033	0.45±0.057	0.512±0.033	0.526±0.064
SGD	1e-5	0.01	0.609±0.025	0.493±0.051	0.544±0.041	0.531±0.057
SGD	1e-5	0.02	0.597±0.034	0.426±0.08	0.495±0.066	0.504±0.066
SGD	1e-5	0.03	0.542±0.063	0.311±0.147	0.416±0.097	0.57±0.137
SGD	2e-5	0.01	0.6±0.104	0.505±0.194	0.578±0.087	0.599±0.136
SGD	2e-5	0.02	0.65±0.03	0.54±0.097	0.586±0.071	0.593±0.05
SGD	2e-5	0.03	0.651±0.023	0.544±0.068	0.59±0.046	0.574±0.054
SGD	3e-5	0.01	0.646±0.096	0.59±0.118	0.615±0.108	0.658±0.07
SGD	3e-5	0.02	0.658±0.022	0.616±0.048	0.636±0.032	0.643±0.053
SGD	3e-5	0.03	0.67±0.07	0.502±0.182	0.54±0.189	0.578±0.078
SGD	4e-5	0.01	0.627±0.101	0.521±0.151	0.561±0.123	0.607±0.1
SGD	4e-5	0.02	0.653±0.057	0.596±0.106	0.62±0.083	0.628±0.114
SGD	4e-5	0.03	0.652±0.052	0.576±0.077	0.607±0.045	0.615±0.069
SGD	5e-5	0.01	0.674±0.062	0.525±0.224	0.621±0.11	0.671±0.11
SGD	5e-5	0.02	0.66±0.044	0.56±0.2	0.64±0.0386	0.678±0.092
SGD	5e-5	0.03	0.644±0.051	0.487±0.211	0.582±0.103	0.635±0.111

4.5. Discussion

In ChemProt test set, our best scores are 77.8% F1-Score and 76.1% micro-averaged F1-Score with the BioBERT-based models trained with our sentence-based input representation in the binary and multi-class relation extraction, respectively. Although there are no published studies for binary relation extraction task on ChemProt data set apart from this thesis, several works for multi-class relation extraction task on the same data set and their results were published in BioNLP literature. A SciBERT-based model achieves the state-of-the-art result for multi-class relation extraction on ChemProt test set. It has 83.64% micro-averaged F1-score and outperforms BioBERT-based models on the same task in [49]. However, the best micro-averaged F1-score of our SciBERT-based models is 74.3% on the test set. The reasons can be caused by using different preprocessing steps, parameter combinations, and distinct combined architecture. The highest score of the BioBERT-based model on the state-of-the-art is 76.68% in micro-averaged F1-score in [43], whereas the best micro-averaged F1-score of our BioBERT-based models is 76.1%. On the other hand, we contribute to NLP researches by investigating the effectiveness of dependency tree-based input representations on transformers-based model performance. This thesis demonstrates that dependency tree-based input representations as our novel representations achieve close F1-scores to whole sentence input representation. Lastly, we present Vapur to show that biomedical researchers can effectively use our relation extraction models in their studies.

Throughout this section, we separately evaluate our binary and multi-class relation extraction tasks by analysing our different input representations for each task to explore the limitations of the representations. We selected the best three models for each relation extraction task separately to examine their performances on their sequences, since they have the highest performance in each representation. For each model, we used the best model out of 10 models trained with different random seeds to predict the development set pairs and report the mean and standard deviation of the accuracy by each relation group. In addition, we determined whether the order of

target entities in sentences influences the model performance in each representation. Finally, we analyzed the importance of the sentence length in the model performance with each representation. For these analyses, we listed three models for binary relation extraction:

- The parameters of BioBERT with sentence-based input representation are 3e-5 learning rate, AdamW optimizer, and 0.3 weight decay.
- The parameters of BioBERT with the shortest sequence between target entities are 3e-5 learning rate, AdamW optimizer, and 0.1 weight decay.
- The parameters of BioBERT with the shortest subsequence including target entities and their parent nodes are 3e-5 learning rate, AdamW optimizer, and 0.2 weight decay.

We listed three models for multi-class relation extraction for the analyses:

- The parameters of BioBERT with sentence-based input representation are 3e-5 learning rate, AdamW optimizer, and 0.3 weight decay.
- The parameters of BioBERT with the shortest sequence between target entities are 3e-5 learning rate, AdamW optimizer, and 0.1 weight decay.
- The parameters of BioBERT with the shortest subsequence including target entities and their parent nodes are 3e-5 learning rate, AdamW optimizer, and 0.1 weight decay.

We conduct several significant tests with 95% confidence interval and 9 degrees of freedom for our best models with different input representations and publish them in Appendices. Their t value of selected best binary models with SB and SST is written in Table D.1, whereas their t value of selected best binary models with SB and SSTP is recorded in Table D.2. On the other hand, we write t value of selected best multi-class models with SB and SST as well as t value of selected best multi-class models with SB and SSTP in Table D.3 and Table D.4, respectively. Their differences are statistically significant based on paired t-test.

Firstly, we show the length of input sequences from our three input representations in Figure 4.1. Our dependency tree-based input representations generally have fewer words than the sentence-based input representation. In Figure 4.1, blue bars show that sentence-based input representation has more samples in the range of 20 and 60 words in sentences from the data set. Orange and green bars indicate that dependency tree-based input representations provide more sequences in the range of 4 and 40 words in sentences in ChemProt. Therefore, since it is possible to lose necessary words for relations in sentences via dependency tree-based input representations, the models trained with the sentence-based input representation always have better performance among all experiments during this study. We use three abbreviations for our input representations during the discussion section: SB for the sentence-based input representation, SST for the shortest sequence between target entity pairs, and SSTP for the shortest sequence between target entity pairs and their parent node. For the binary relation extraction, Table 4.25 includes the best F1-scores of our input representations in binary relation extraction task. Our dependency tree-based input representations achieve close F1-scores to whole sentence input representation despite 29.5% and 23.4% compression ratio on the input sequences. Table 4.26 contains the best F1-scores of our input representations in multi-class relation extraction task. The dependency tree-based input representations perform close micro-averaged F1-scores to sentence-based input representation.

Table 4.25. Our best scores from BioBERT based binary models with different input representations on the dev set.

Input Type	**Compression Ratio**	**Precision**	**Recall**	**F1-Score**
SB	-	0.763	0.812	0.787
SST	29.5%	0.768	0.766	0.766
SSTP	23.4%	0.761	0.779	0.769

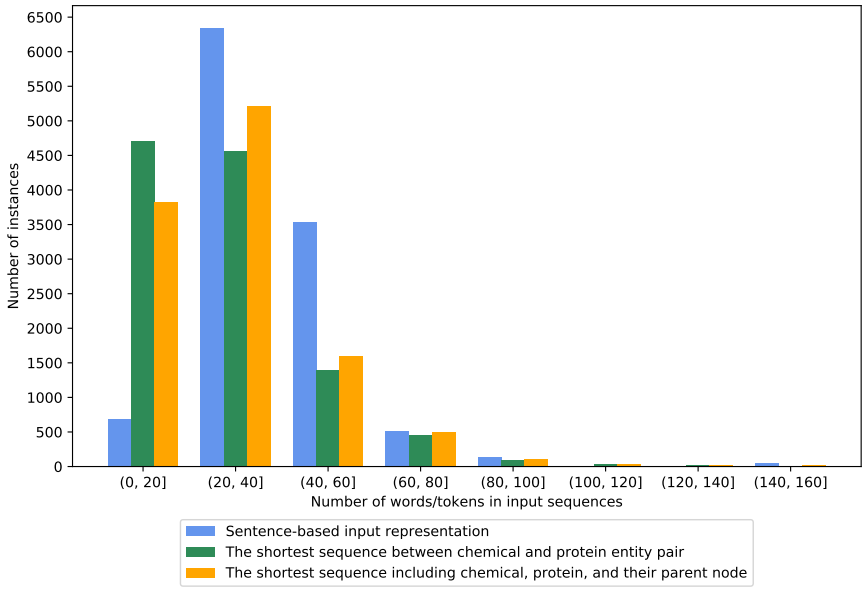

Figure 4.1. Histogram of the number of words in our sentence-based and dependency tree-based input representations. We calculated the number of words in sentences by white space.

Table 4.26. Our best scores from BioBERT based multi-class models with different input representations on the dev set.

Input Type	**Compression Ratio**	**Micro-averaged Precision**	**Micro-averaged Recall**	**Micro-averaged F1-Score**
SB	-	0.791	0.785	0.787
SST	29.5%	0.756	0.755	0.755
SSTP	23.4%	0.755	0.762	0.758

Apart from input representations, we found that BioBERT with the best parameters on each representation has always better performance than SciBERT with its best parameters on different input representations. In addition, models with AdamW have usually higher F1-scores and micro-averaged F1-scores than the models with Adam and SGD in binary and multi-class relation extraction tasks, respectively.

4.5.1. Discussion on Binary Relation Extraction

We wrote accuracy scores on the development set for three models finetuned with the sentence-based input and dependency tree-based representations in Table 4.27. BioBERT finetuned with SB has the highest accuracy score. BioBERT-based models with SST and SSTP have accuracy scores very close to the best one during our experiments. Furthermore, we examined whether the order of target entities in sentences affects the model performance. Table 4.27 shows that BioBERT-based models with SST and SSTP perform better when chemical entities precede protein/gene entities. In addition, we reported accuracy scores per CPR for the different representations in Table 4.28. BioBERT finetuned with SB has the highest accuracy on CPR:1, CPR:2, CPR:3, CPR:4, CPR:5, CPR:6, CPR:7, CPR:9, and CPR:10. BioBERT finetuned with SST has the highest accuracy on Other class. However, BioBERT-based models fine-tuned with SST and SB have an accuracy close to the other model on Other. It indicates that the BioBERT-based model with SB has the highest accuracy on almost all relation groups.

Secondly, we analyzed whether the length of sentences affects the model performance. In Figure 4.2, the three models do not produce significantly different accuracy scores within the same range of sentence lengths. The length of sentences is measured by counting the number of words by splitting raw sentences with white space. We do not have any sentences with the number of words in the range of 100 and 120 on the dev set. However, the three models give better accuracy scores on long sentences, although the data set has fewer sentences having words in the range of 120 and 180. On the other hand, Figure 4.3 indicates the accuracy scores of the three models and the distance between entity markers in a sentence. The distance between entity markers is calculated by counting words between them. As the distance between entity markers increases from 20 to 60, the three models have better accuracy scores. Nevertheless, we do not have sufficient sequences that have the distance between entity markers in a range of 60 and 140.

Table 4.27. Accuracy scores calculated considering the order of target entities in our binary relation extraction models. 55% of the total input sequences in the dev set have chemical entities that come before protein/gene entities in sentences.

Property of data	**Accuracy for SB**	**Accuracy for SST**	**Accuracy for SSTP**
Chemical entities come first	0.87±0.004	0.865±0.005	0.864±0.006
Protein/gene entities come first	0.87±0.004	0.859±0.005	0.860±0.006
Total data set	0.87±0.0025	0.862±0.0032	0.862±0.0046

Table 4.28. Accuracy for each relation group in our binary relation extraction models.

Relation Group	**Accuracy for SB**	**Accuracy for SST**	**Accuracy for SSTP**
CPR:0	0.450 ± 0.158	0.450 ± 0.158	0.450 ± 0.158
CPR:1	0.710 ± 0.055	0.593 ± 0.093	0.609 ± 0.078
CPR:2	0.697 ± 0.033	0.659 ± 0.028	0.676 ± 0.027
CPR:3	0.842 ± 0.018	0.804 ± 0.026	0.815 ± 0.022
CPR:4	0.906 ± 0.010	0.880 ± 0.021	0.891 ± 0.014
CPR:5	0.939 ± 0.013	0.854 ± 0.040	0.875 ± 0.023
CPR:6	0.928 ± 0.026	0.874 ± 0.026	0.863 ± 0.013
CPR:7	0.732 ± 0.094	0.632 ± 0.124	0.700 ± 0.136
CPR:8	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000
CPR:9	0.718 ± 0.039	0.642 ± 0.054	0.659 ± 0.038
CPR:10	0.831 ± 0.033	0.810 ± 0.022	0.812 ± 0.056
Other	0.896 ± 0.009	0.904 ± 0.014	0.899 ± 0.008

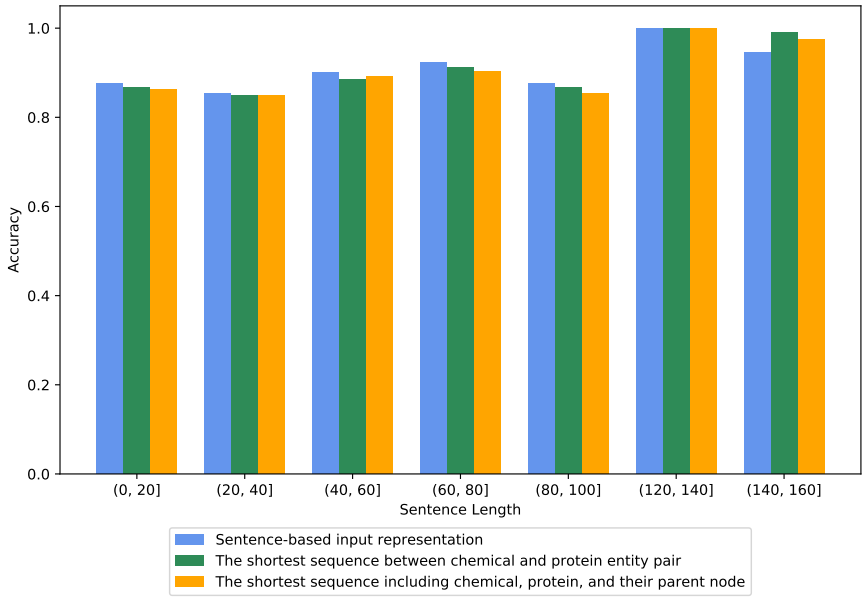

Figure 4.2. Accuracy scores for different range of sentence length with our binary models.

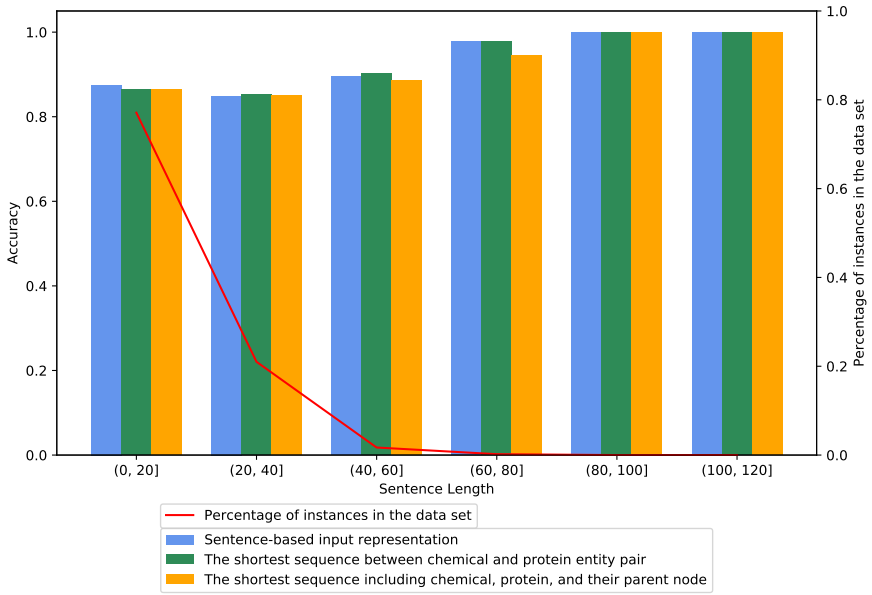

Figure 4.3. Accuracy scores for the distance between entity markers in sequences with our binary models.

4.5.2. Discussion on Multi-class Relation Extraction

We reported accuracy scores on the development set for the models finetuned with the sentence-based input and dependency tree-based representations in Table 4.29. We obtained the highest accuracy by BioBERT finetuned with SB on the data set. Table 4.29 indicates that BioBERT-based models with SST and SSTP perform better when chemical entities precede protein/gene entities. However, the BioBERT-based model with SB has better performance when protein/gene entities come first in sentences. Besides, we calculated the accuracy scores per CPR for three input representations in Table 4.30. BioBERT finetuned with SB has the highest accuracy on CPR:3, CPR:4, CPR:5, CPR:6, CPR:7, and CPR:9. BioBERT finetuned with SST has the highest accuracy scores on CPR:10 and Other classes. It indicates that the BioBERT-based model with SB has the highest accuracy on five relation groups that are evaluated in the BioCreative Task.

Secondly, Figure 4.4 indicates that the three models do not produce significantly different accuracy scores within the same range of sentence lengths. The length of sentences is measured by counting the number of words by splitting raw sentences with white space. We do not have any sentences with the number of words in the range of 100 and 120 on the development set. However, the three models have lower accuracy scores in the sentences having words in a range of 120 and 140. On the other hand, Figure 4.5 indicates the accuracy scores of the three models and the distance between entity markers in a sentence. As the distance between entity markers increases from 20 to 60, the three models have better accuracy scores. We have slightly better accuracy scores on dependency tree-based input representations in a range of 40 and 60 words in a sentence. Nevertheless, we do not have enough sequences that have the distance between entity markers in a range of 60 and 140.

Table 4.29. Accuracy scores calculated considering the order of target entities in our multi-class relation extraction models.

Property of data	**Accuracy for SB**	**Accuracy for SST**	**Accuracy for SSTP**
Chemical entities come first	0.902±0.005	0.903±0.003	0.902±0.003
Protein/gene entities come first	0.904±0.004	0.898±0.006	0.900±0.005
Total data set	0.903±0.0036	0.901±0.0036	0.901±0.0029

Table 4.30. Accuracy for each group in our multiclass relation extraction models.

Relation Group	**Accuracy for SB**	**Accuracy for SST**	**Accuracy for SSTP**
CPR:0	0.500 ± 0.000	0.500 ± 0.000	0.500 ± 0.000
CPR:1	0.992 ± 0.002	0.994 ± 0.003	0.996 ± 0.004
CPR:2	0.876 ± 0.013	0.873 ± 0.023	0.876 ± 0.017
CPR:3	0.739 ± 0.014	0.692 ± 0.033	0.704 ± 0.028
CPR:4	0.840 ± 0.014	0.816 ± 0.014	0.811 ± 0.023
CPR:5	0.879 ± 0.041	0.702 ± 0.083	0.785 ± 0.048
CPR:6	0.868 ± 0.021	0.788 ± 0.055	0.823 ± 0.029
CPR:7	0.868 ± 0.080	0.653 ± 0.106	0.611 ± 0.127
CPR:8	1.000 ± 0.000	1.000 ± 0.000	1.000 ± 0.000
CPR:9	0.695 ± 0.047	0.684 ± 0.041	0.683 ± 0.037
CPR:10	0.830 ± 0.031	0.885 ± 0.024	0.878 ± 0.018
Other	0.939 ± 0.007	0.948 ± 0.006	0.946 ± 0.007

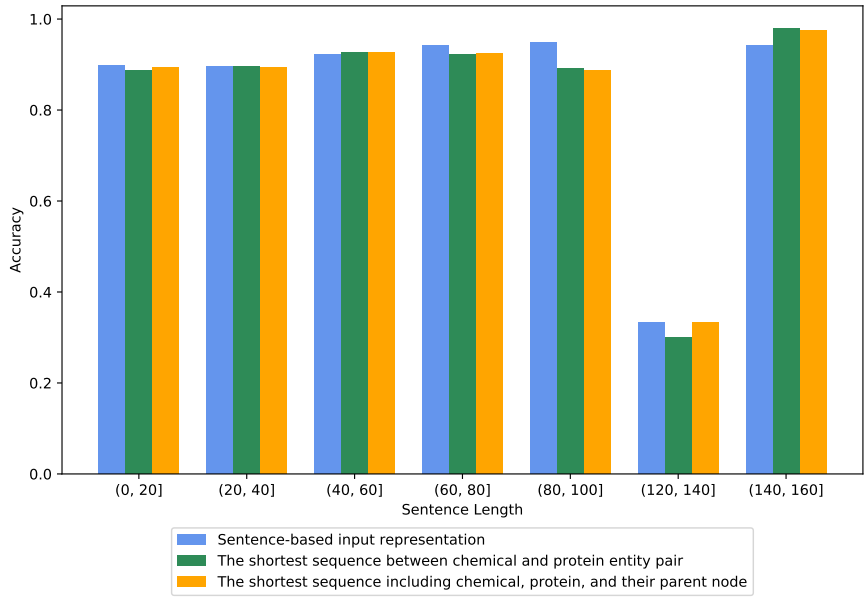

Figure 4.4. Accuracy scores for different range of sentence length with our multi-class models.

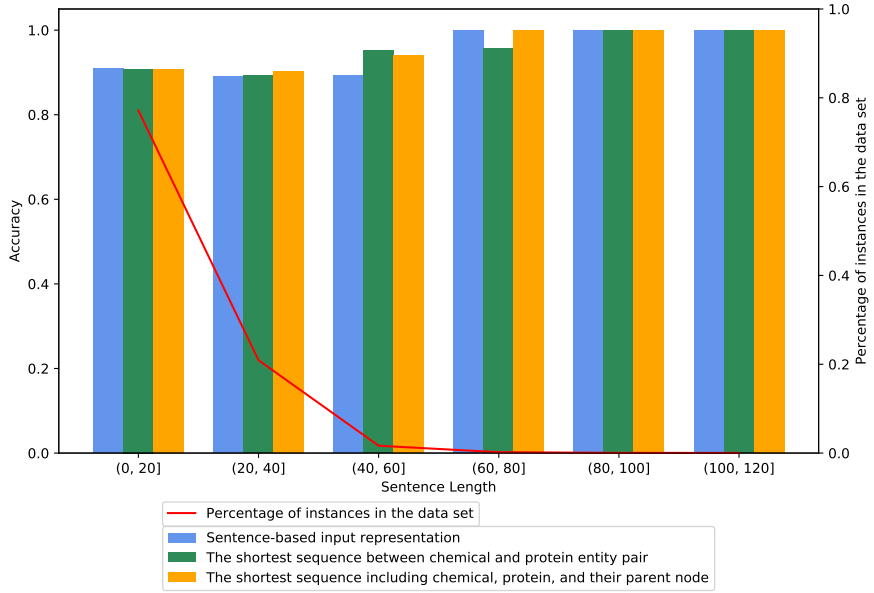

Figure 4.5. Accuracy scores for the distance between entity markers in sequences with our multi-class models.

5. Vapur: AN APPLICATION OF RELATION EXTRACTION ON COVID LITERATURE

Coronavirus Disease of 2019 (COVID-19) is a public health problem that was declared as a Public Health Emergency International Concern by World Health Organization (WHO). Coronavirus Disease of 2019 (COVID-19) outbreak had severe impacts on human health all around the world since December 2019, but also triggered an exceptional amount of scientific work. As of September 2020, PubMed recorded over 45K articles [78] related to COVID-19 since its inception, including works on diagnosis [79], drug repurposing [80], and text-mining [81]. As the body of literature keeps growing in the form of unstructured text, it also becomes more and more challenging for researchers to find the relevant information they need. Furthermore, the publications include domain-specific named entities and relations that challenge the general-purpose search engines. Therefore, it is of critical importance to build a search engine that can find relevant documents in this terminology-rich and domain-specific literature.

Biomedical named entity recognition (NER) and relation extraction can be utilized to semantically structure publications around the biochemically related entities. When named entities and their relations are extracted, a document can be expressed as a set of triplets of the form $(Entity1, Entity2, Relation)$. This formulation can be converted to an inverted index from the related entities to the publications in order to enable retrieving relevant documents to a query by entity and relation matching. If the same entities are referenced with different words (e.g. ACE2, Angiotensin-converting enzyme 2, Q9BYF1), named entity normalization can be used to identify different mentions of the same entity. Enhanced with named entity normalization, the inverted index can retrieve the documents that contain biochemical relations of a free-text query, as grouped by the related biomolecules.

In this work, we present Vapur, an online search engine to find related protein - compound pairs in the COVID-19 anthology. Vapur is empowered with a biochemical relation-based inverted index that is created through named entity recognition, named entity normalization, and relation extraction on CORD-19 abstracts [82]. Thanks to the underlying biochemical domain-specific tools and relation extraction model, Vapur identifies biochemically related entities to a free-text query and retrieves the publications that mention the relation. The design of Vapur offers a novel approach to explore COVID-19 literature with a focus on biomolecules and their relations. We present Vapur at `https://tabilab.cmpe.boun.edu.tr/vapur/` and publicly share the code and models at `https://github.com/boun-tabi/vapur`.

5.1. System Description

Vapur is an online search engine with a focus on finding related proteins and chemicals in the COVID-19 literature. Vapur is able to retrieve relevant documents to a query as categorized by the biochemically related entities thanks to its relation-oriented inverted index. In order to obtain the index, Vapur first identifies and normalizes the named entities in CORD-19 abstracts using a pre-trained model, BERN [42]. Afterward, Vapur determines if the entity pairs in the same sentence are related to each other by the binary relation extraction model we trained on the ChemProt data set [34]. The result of the relation extraction model is a list of related entities for each abstract. This list is then used to construct the inverted index that represents biochemical relations as entity pairs and maps each relation to the documents in which the relation was mentioned. Vapur is publicly available at `https://tabilab.cmpe.boun.edu.tr/vapur/` and Figure 5.1 illustrates its workflow. We first split CORD-19 abstracts to sentences and use BERN to detect and normalize the entities in the text. We then identify the biochemical relations with the relation extraction model that we trained and reform the output as an inverted index of relations. Vapur leverages this inverted index to retrieve relevant publications to the query as categorized by related entities.

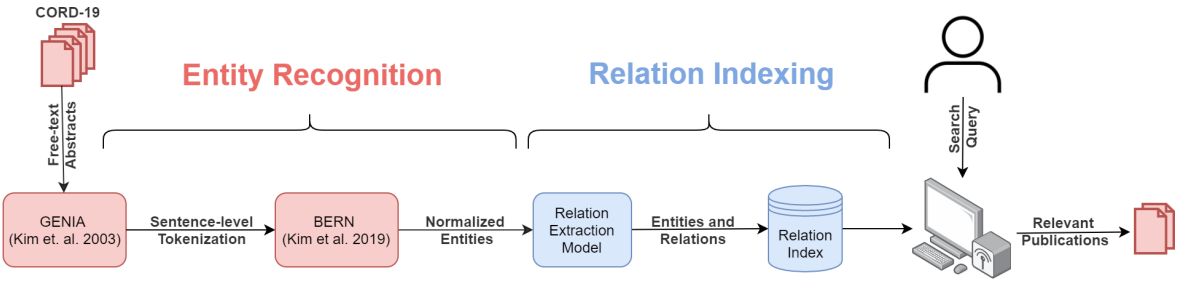

Figure 5.1. The workflow of the pipeline behind Vapur.

5.2. COVID-19 Specific Search Engines

CORD-19 [82] is a data set comprising scientific publications related to COVID-19 and it has provided a valuable resource for text mining studies [83], [84], [85], including COVID-19 specific search engine development such as SemViz [86] and KD-COVID (available at `http://kdcovid.nl/about.html`).

SemViz aims to present subtle relations between entities by semantic visualization of a CORD-19 based knowledge-graph. KD-COVID, on the other hand, identifies the most similar sentence in CORD-19 to a query and retrieves the corresponding publication. KD-COVID also provides links to large protein, disease, and protein-disease relation databases. Vapur is different from these approaches, since it is based on an inverted index from biochemically related entities to publications. While most other systems depend on existing knowledge bases to provide information about related entities, Vapur uses NLP techniques to automatically extract the most up-to-date information from the scientific literature.

5.3. Use Case Scenario

We consulted with domain experts to build scenarios in which Vapur can help researchers. We present a use case that highlights how the relation-based index of Vapur and the similar entity prediction can extend the scope of a research. We illustrate the scenario in Figure 5.2. When *Favipiravir* is entered as the query, Vapur shows the query and its unique ids in *1*. It lists contextually similar biomolecules to Favipiravir

in *2*. It displays related biomolecules found by the relation extraction model in *3*. For example, the first result is RNA-dependent RNA polymerase for Favipiravir query. Finally, it shows sentences in related papers to provide information about the relations as in *4*.

Relation between Favipiravir and RdRp. A medicinal chemist working on COVID-19 drug development is interested in "Favipiravir", one of the drugs clinically used in COVID-19 therapy. She searches Favipiravir on Vapur and the results highlight RNA-dependent RNA polymerase (RdRp) enzyme as the most frequently co-mentioned protein that is identified by Vapur as being related to Favipiravir. Vapur displays sentences from the publications stating that Favipiravir inhibits RdRp and also proposes "Examorelin" as a contextually similar molecule to Favipiravir. The researcher decides to extend her research to Examorelin, another drug known to bind RdRp.

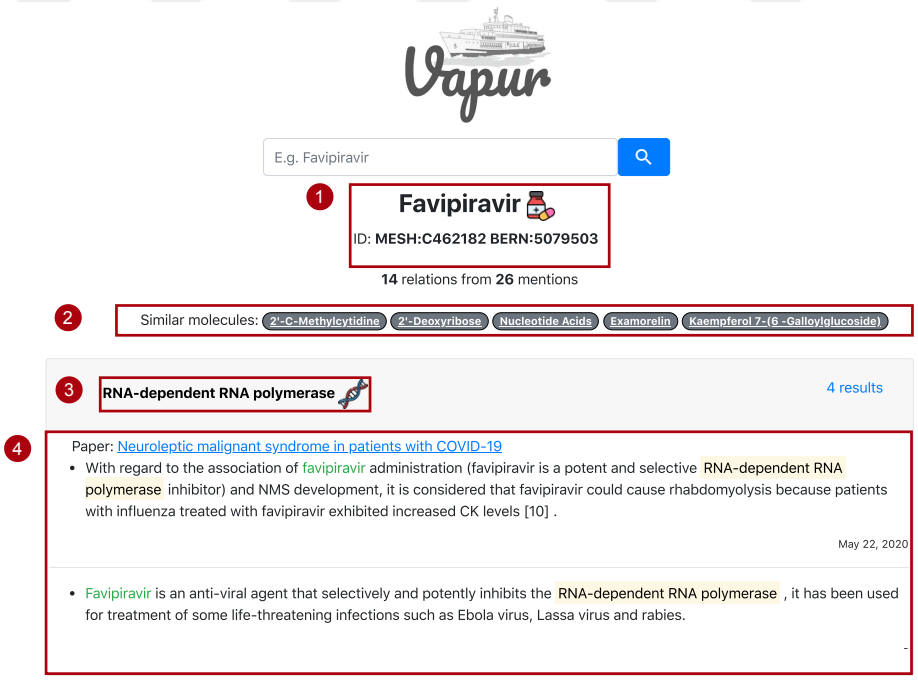

Figure 5.2. A search scenario on Vapur.

5.4. CORD-19 Data Set

We create a relation-oriented search engine for COVID-19 Open Research Dataset (CORD-19) [82] abstracts. CORD-19 is a regularly updated data set that contains studies related to COVID-19 and SARS-CoV-2. We use the CORD-19 snapshot of August 23, 2020, that contains ≈ 233K documents for which ≈ 143K unique abstracts are provided. Vapur indexes these abstracts, but is also able to return the linked full-paper.

5.5. Methods

5.5.1. Named Entity Recognition and Normalization

We used BERN [42], a neural NER architecture with an integrated normalizer that can perform in a multi-type setting, to identify and normalize the entities in CORD-19 abstracts. BERN is an ensemble of existing tools and models [39], [38], [87], [88], [40], [43] and outputs a set of IDs for each recognized entity, given a sentence.

In this study, we first tokenized the abstracts with GENIA [62] and identified 1.23M sentences. We used BERN to extract named entities and their IDs in these sentences and recognized 1.58M entities in total, in which 171K are chemicals, 318K are proteins, and others are diseases and species. With normalization, we computed the number of unique chemicals and proteins as 20K and 70K, respectively.

5.5.2. Relation Extraction

We propose a BioBERT-based model to identify related entities in CORD-19 abstracts. To this end, we first preprocessed the sentences to explicitly encode the entities in the input and then fine-tuned BioBERT with the preprocessed sentences in ChemProt. We used the binary labels in Section 3.1 as outputs.

Preprocessing. We surrounded the named entities in the sentences with opening and closing tags to mark their location. We tagged chemicals with <e1> and </e1> and proteins with <e2> and </e2> to encode entity type and location during training. When a sentence had multiple chemical - protein pairs, we considered each pair separately and created copies of the sentence with different tags. Table 3.10 provides an example of input and its two preprocessed forms.

Model. We trained a binary model that decides if two entities in the same sentence are biochemically related. The goal of training a binary model instead of a multi-class one is to identify biochemical relations beyond the classes in ChemProt. By formulating the problem as binary classification, we can potentially identify the biochemically related molecules in CORD-19, whose biochemical relation type cannot be categorized under any of the ChemProt classes.

We used BioBERT with the addition of a single-layer binary log-softmax classifier and trained the classifier with the inputs and outputs from the previous step using the cross-entropy loss. BioBERT creates a fixed-length relation representation for the starting tag *CLS* and we used this representation to vectorize the sentences. Figure 3.4 illustrates our relation extraction model.

Experimental setup. We conducted a hyper-parameter search with different optimizers, learning rates, and weight decays. We trained a model 10 times per parameter combination and selected the best one based on the F-score on the development set. We found that the AdamW optimizer with a 0.00003 learning rate and 0.1 weight decay yields the best results. We selected the best setting for relation extraction and computed mean precision, recall, and F1-score as well as the standard deviation on the ChemProt development (dev) and test sets.

5.6. Results and Discussion

In order to assess the performance of Vapur, we evaluated the performance of its components both as separate modules and together as an end-to-end pipeline. Since BERN, the named entity recognition and normalization module, has already been shown to be successful [42], we focus on automatic evaluation of the relation extraction model and expert evaluation of Vapur in this work.

5.6.1. Error Analysis of Relation Extraction Model

We built the relation extraction model by fine-tuning a binary classifier on BioBERT and then computed precision, recall, and F1-Score on the dev and test sets of ChemProt. We also fine-tuned the same classifier using English BERT based model and report the results in Table 5.1.

Table 5.1 demonstrates that BioBERT obtained higher scores in terms of all three metrics on both folds, indicating that BioBERT is superior to BERT as a pre-trained language model for relation extraction on ChemProt. We relate this with the fact that BioBERT is trained with a more domain-related text and BERT observed texts from a wider range.

We further investigated the performance of our relation extraction model by computing the test accuracy per CPR category and illustrate the results in Table 5.2. We observe that the relation extraction model achieved the highest accuracy in the *Other* category, indicating that the model can successfully identify whether the context is sufficient to deduce a relation between the entities or not. We relate the high performance for the *Other* category by using a contextual model for relation representation, BioBERT. Context-awareness enables Vapur to eliminate the documents that mention the queried entity in irrelevant contexts and to retrieve a document only if it contains relation information for the query.

Table 5.1. BERT- and BioBERT-based fine-tuning results for binary relation extraction. We computed precision, recall, and F1 on the dev and test folds of ChemProt.

Fold	**Model**	**Precision**	**Recall**	**F1-Score**
Dev	BERT	0.718±0.027	0.737±0.026	0.727±0.010
	BioBERT	0.742±0.036	0.829±0.035	0.782±0.012
Test	BERT	0.759±0.023	0.710±0.026	0.733±0.007
	BioBERT	0.791±0.026	0.766±0.036	0.777±0.010

Table 5.2. Test performance of the relation extraction model by CPR label. We used the best model out of 10 models trained with different random seeds to predict test set pairs and report mean and standard deviation of the test accuracy by label.

Relation ID	**Test Set Accuracy**
CPR:1	0.661 ± 0.053
CPR:2	0.667 ± 0.043
CPR:3	0.855 ± 0.015
CPR:4	0.874 ± 0.027
CPR:5	0.850 ± 0.045
CPR:6	0.856 ± 0.050
CPR:7	0.890 ± 0.044
CPR:8	0.550 ± 0.066
CPR:9	0.617 ± 0.060
CPR:10	0.828 ± 0.053
Other	0.892 ± 0.020

5.6.2. User Evaluation of Vapur

We evaluated Vapur from two different perspectives. We first analyzed 41 sample sentences in which Vapur identified a biochemical relation as a first step to discover the limitations of the complete pipeline. Then, we asked six biologists/chemists to use Vapur and rate its different aspects to demonstrate the success and usefulness of Vapur for future research.

Our inspection of 41 sample sentences indicated that most of the incorrect relation labels were due to incorrect entity assignment by BERN. In some cases, parts of the protein sequence such as *N-terminal*, *carboxyl terminal* or residue names such as *Asp238* are recognized as compounds. Table 5.3 illustrates sample sentences with incorrectly labeled entities. The entity types of *Alanine, carboxyl*, and *Ben* were incorrectly predicted as compounds by BERN during the named entity recognition step. Consequently, the relation extraction model incorrectly identified a biochemical relation. Other examples that were manually checked by a domain expert are presented in the Appendices. A domain expert manually checked sentences and their chemical and protein entities recognized by BERN. Identified chemical and protein columns have chemicals and proteins recognized by BERN, respectively. A domain expert reports the chemical and protein label columns as 1 if the recognition from BERN is correct and 0 otherwise. A domain expert sets the relation label as 1 if the entity pairs have a biochemical relation in the related sentence.

In order to evaluate the real-life usefulness of Vapur, we asked six domain experts to use Vapur for five COVID-19 related queries (totalling up to 30) of their own. They each filled in a questionnaire where for each query they indicated (i) if each of the top three search results is related to the query, (ii) if similar molecules predicted by Vapur are in fact useful, and (iii) if the extracted sentences are useful. They also rated the ease of use of Vapur between 1 (very difficult) and 5 (very easy) and assessed its usefulness for future research on COVID-19.

Table 5.3. Sample sentences with incorrectly labeled entities.

Sentence in CORD-19	**Incorrectly Labeled Entity**
<e1>Alanine</e1> substitution of either Arg-76 or Tyr-94 in the N-terminal domain of <e2>IBV N protein</e2> led to a significant decrease in its RNA-binding activity and a total loss of the infectivity of the viral RNA to Vero cells.	Alanine
<e2>Rat microsomal aldehyde dehydrogenase</e2> (msALDH) has no amino-terminal signal sequence, but instead it has a characteristic hydrophobic domain at the <e1>carboxyl</e1> terminus (Miyauehi, K., R.	carboxyl
Also, the <e2>protease Factor Xa</e2>, a target of <e1>Ben</e1>-HCl abundantly expressed in infected cells, was able to cleave the recombinant and pseudoviral S protein into S1 and S2 subunits, and the cleavage was inhibited by Ben-HCl.	Ben

The expert evaluations demonstrated that 27 out of 30 (90%) top search results and 76 out of 90[1] (84%) top three search results are biochemically related to the query, suggesting that Vapur successfully retrieves biochemically related entities to the query. During these experiments, Vapur retrieved at least one related document for each query. It returned only one result for two of the queries, which caused 4 of the 14 unsuccessful cases. These evaluations suggest that the inverted index of Vapur spans a comprehensive range of entities and contains sufficient number of documents to find a biochemically related result for each of the tested queries.

[1]Note that for 30 queries there are a total of 30x3=90 top three search results.

The evaluation further showed that for 22 out of the 30 queries (73%) the extracted sentences are useful for research. Besides, the experts unanimously expressed that Vapur is very easy to use (5/5) and useful for future research.

Overall, both manual inspection and expert evaluations showed that Vapur can successfully find biochemically related proteins and chemicals in CORD-19 and can help future research on COVID-19.

6. CONCLUSION

We present BioBERT and SciBERT-based models to extract a relation between the pair of biochemicals from a sentence and to identify the type of the relation between the pair of biochemicals. We propose our new dependency tree-based input representations with acceptable F1-scores in case of any hardware constraints. Finally, we introduce Vapur, a search engine for protein – chemical interactions extracted from COVID-19 related scientific publications. Our relation extraction models can be effectively used in real-world biomedical applications.

We evaluated our binary and multi-class relation extraction models on ChemProt and observed that our binary and multi-class BioBERT-based models with the sentence-based input representation have the highest performance among others. On ChemProt test set, we achieved 77.8% F1-Score and 76.1% micro-averaged F1-Score with the models in the binary and multi-class relation extraction, respectively. In addition, although we use the shorter inputs produced by our dependency tree-based input representations to finetune the models, the difference between the highest F1-Score of these models and the highest F1-Score of the models trained with the sentence-based input representation is less than 2% on ChemProt data set in the binary task. In experiments of multi-class relation extraction, our micro-average F1-scores decrease by around 3% by using our dependency tree-based input representations.

We introduce Vapur, an online search engine that uses a relation extraction model to find related biochemical entities in COVID-19 literature. Vapur retrieves documents related to a query by its biochemical entities thanks to its relation extraction model. For sample annotated sentences on CORD-19, a domain expert evaluated that Vapur is successful to find the relations between biochemical entities when the entities are precisely recognized. Lastly, six domain experts who used Vapur with 5 different queries stated that Vapur retrieves documents related to queries. They expressed that Vapur is user-friendly and beneficial for future studies. Even though Vapur re-

trieves documents from CORD-19, our approach can shed light on applications in any biomedical domain with various databases. We believe that domain-specific search engines will pay more attention to future research since COVID-19 is unlikely to be a recent global health problem. We make Vapur and the pipeline behind it available at `https://github.com/boun-tabi/vapur` and `https://tabilab.cmpe.boun.edu.tr/vapur/`, respectively. As future work for Vapur, we plan to extend Vapur to be able to find relations over full papers, to recognize entities by more robust named entity recognizers and normalizers, and index various databases on different biomedical domains. In addition, we will collect the questionnaire from users who search a query on Vapur in order to evaluate the correctness of relations related to queries. In future works for our relation extraction pipeline, although we used relation information from a sentence for specified entities, it will be good to have relation information from multiple sentences. In addition, it can be better to train the models with full-texts from scientific documents. Finally, we believe that the combination of models with our different input representations can give better results.

REFERENCES

1. Köksal, A., H. Dönmez, R. Özçelik, E. Ozkirimli and A. Özgür, "Vapur: A Search Engine to Find Related Protein - Compound Pairs in COVID-19 Literature", *Proceedings of the 1st Workshop on NLP for COVID-19 (Part 2) at EMNLP 2020*, Dec. 2020.

2. Peng, Y., A. Rios, R. Kavuluru and Z. Lu, "Extracting chemical–protein relations with ensembles of SVM and deep learning models", *Database*, Vol. 2018, 2018.

3. Lung, P.-Y., Z. He, T. Zhao, D. Yu and J. Zhang, "Extracting chemical–protein interactions from literature using sentence structure analysis and feature engineering", *Database*, Vol. 2019, 2019.

4. Chen, T., T. He, M. Benesty, V. Khotilovich, Y. Tang, H. Cho *et al.*, "Xgboost: extreme gradient boosting", *R Package Version 0.4-2*, Vol. 1, No. 4, 2015.

5. Yüksel, A., H. Öztürk, E. Ozkirimli, A. Özgür *et al.*, "CNN based chemical-protein interactions classification", *Proceedings of the BioCreative VI Workshop, Bethesda, MD*, pp. 184–186, 2017.

6. Mitsumori, T., M. Murata, Y. Fukuda, K. Doi and H. Doi, "Extracting protein-protein interaction information from biomedical text with svm", *IEICE Transactions on Information and Systems*, Vol. 89, No. 8, pp. 2464–2466, 2006.

7. Giuliano, C., A. Lavelli and L. Romano, "Exploiting shallow linguistic information for relation extraction from biomedical literature", *11th Conference of the European Chapter of the Association for Computational Linguistics*, 2006.

8. Yu, K., P.-Y. Lung, T. Zhao, P. Zhao, Y.-Y. Tseng and J. Zhang, "Automatic extraction of protein-protein interactions using grammatical relationship graph", *BMC Medical Informatics and Decision Making*, Vol. 18, No. 2, p. 42, 2018.

9. Dijkstra, E. W. *et al.*, "A note on two problems in connexion with graphs", *Numerische Mathematik*, Vol. 1, No. 1, pp. 269–271, 1959.

10. Zeng, D., K. Liu, S. Lai, G. Zhou and J. Zhao, "Relation classification via convolutional deep neural network", *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pp. 2335–2344, 2014.

11. Hua, L. and C. Quan, "A shortest dependency path based convolutional neural network for protein-protein relation extraction", *BioMed Research International*, Vol. 2016, 2016.

12. Peng, Y. and Z. Lu, "Deep learning for extracting protein-protein interactions from biomedical literature", *arXiv preprint arXiv:1706.01556*, 2017.

13. Hsieh, Y.-L., Y.-C. Chang, N.-W. Chang and W.-L. Hsu, "Identifying protein-protein interactions in biomedical literature using recurrent neural networks with long short-term memory", *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pp. 240–245, 2017.

14. Xu, K., S. Reddy, Y. Feng, S. Huang and D. Zhao, "Question answering on freebase via relation extraction and textual evidence", *arXiv preprint arXiv:1603.00957*, 2016.

15. Aggarwal, C. C. and C. Zhai, *Mining text data*, Springer Science & Business Media, 2012.

16. Kambhatla, N., "Combining lexical, syntactic, and semantic features with maximum entropy models for information extraction", *Proceedings of the ACL Interactive Poster and Demonstration Sessions*, pp. 178–181, 2004.

17. Li, Q. and H. Ji, "Incremental joint extraction of entity mentions and relations", *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 402–412, 2014.

18. Mikolov, T., I. Sutskever, K. Chen, G. Corrado and J. Dean, "Distributed representations of words and phrases and their compositionality", *arXiv preprint arXiv:1310.4546*, 2013.

19. Zhang, D. and D. Wang, "Relation classification via recurrent neural network", *arXiv preprint arXiv:1508.01006*, 2015.

20. Xu, Y., L. Mou, G. Li, Y. Chen, H. Peng and Z. Jin, "Classifying relations via long short term memory networks along shortest dependency paths", *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pp. 1785–1794, 2015.

21. Miwa, M. and M. Bansal, "End-to-end relation extraction using lstms on sequences and tree structures", *arXiv preprint arXiv:1601.00770*, 2016.

22. Nguyen, T. H. and R. Grishman, "Combining neural networks and log-linear models to improve relation extraction", *arXiv preprint arXiv:1511.05926*, 2015.

23. Vaswani, A., N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser and I. Polosukhin, "Attention is all you need", *arXiv preprint arXiv:1706.03762*, 2017.

24. Devlin, J., M.-W. Chang, K. Lee and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding", *arXiv preprint arXiv:1810.04805*, 2018.

25. Radford, A., J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever *et al.*, "Language models are unsupervised multitask learners", *OpenAI blog*, Vol. 1, No. 8, p. 9, 2019.

26. Liu, Y., M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer and V. Stoyanov, "Roberta: A robustly optimized bert pretraining approach", *arXiv preprint arXiv:1907.11692*, 2019.

27. Song, K., X. Tan, T. Qin, J. Lu and T.-Y. Liu, "Mass: Masked sequence to sequence pre-training for language generation", *arXiv preprint arXiv:1905.02450*, 2019.

28. Liu, P. J., M. Saleh, E. Pot, B. Goodrich, R. Sepassi, L. Kaiser and N. Shazeer, "Generating wikipedia by summarizing long sequences", *arXiv preprint arXiv:1801.10198*, 2018.

29. Williams, A., N. Nangia and S. Bowman, "A Broad-Coverage Challenge Corpus for Sentence Understanding through Inference", *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pp. 1112–1122, Jun. 2018.

30. Soares, L. B., N. FitzGerald, J. Ling and T. Kwiatkowski, "Matching the blanks: Distributional similarity for relation learning", *arXiv preprint arXiv:1906.03158*, 2019.

31. Köksal, A. and A. Özgür, "The RELX Dataset and Matching the Multilingual Blanks for Cross-Lingual Relation Classification", *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 340–350, Nov. 2020.

32. Alt, C., M. Hübner and L. Hennig, "Fine-tuning Pre-Trained Transformer Language Models to Distantly Supervised Relation Extraction", *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 1388–1398, Jul. 2019.

33. Papanikolaou, Y., I. Roberts and A. Pierleoni, "Deep Bidirectional Transformers for Relation Extraction without Supervision", *Proceedings of the 2nd Workshop on Deep Learning Approaches for Low-Resource NLP (DeepLo 2019)*, pp. 67–75, Nov. 2019.

34. Krallinger, M., O. Rabal, S. A. Akhondi, M. P. Pérez, J. Santamaría, G. Rodríguez *et al.*, "Overview of the BioCreative VI chemical-protein interaction Track",

Proceedings of the Sixth BioCreative Challenge Evaluation Workshop, Vol. 1, pp. 141–146, 2017.

35. Liu, S., F. Shen, R. Komandur Elayavilli, Y. Wang, M. Rastegar-Mojarad, V. Chaudhary and H. Liu, "Extracting chemical–protein relations using attention-based neural networks", *Database*, Vol. 2018, 10 2018.

36. Lim, S. and J. Kang, "Chemical–gene relation extraction using recursive neural network", *Database*, Vol. 2018, 06 2018.

37. Peng, Y., A. Rios, R. Kavuluru and Z. Lu, "Chemical-protein relation extraction with ensembles of SVM, CNN, and RNN models", *arXiv preprint arXiv:1802.01255*, 2018.

38. Leaman, R., C.-H. Wei and Z. Lu, "tmChem: a high performance approach for chemical named entity recognition and normalization", *Journal of Cheminformatics*, Vol. 7, No. S1, p. S3, 2015.

39. Wei, C.-H., H.-Y. Kao and Z. Lu, "GNormPlus: an integrative approach for tagging genes, gene families, and protein domains", *BioMed Research International*, 2015.

40. Wei, C.-H., L. Phan, J. Feltz, R. Maiti, T. Hefferon and Z. Lu, "tmVar 2.0: integrating genomic variant information from literature with dbSNP and ClinVar for precision medicine", *Bioinformatics*, Vol. 34, No. 1, pp. 80–87, 2018.

41. Sachan, D. S., P. Xie, M. Sachan and E. P. Xing, "Effective use of bidirectional language modeling for transfer learning in biomedical named entity recognition", *Machine Learning for Healthcare Conference*, pp. 383–402, 2018.

42. Kim, D., J. Lee, C. H. So, H. Jeon, M. Jeong, Y. Choi, W. Yoon, M. Sung and J. Kang, "A neural named entity recognition and multi-type normalization tool for biomedical text mining", *IEEE Access*, Vol. 7, pp. 73729–73740, 2019.

43. Lee, J., W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So and J. Kang, "BioBERT: a pre-trained biomedical language representation model for biomedical text mining", *Bioinformatics*, Vol. 36, No. 4, pp. 1234–1240, 2020.

44. Jang, Y. E., I. Jang, S. Kim, S. Cho, D. Kim, K. Kim, J. Kim, J. Hwang, S. Kim, J. Kim, J. Kang, B. Lee and S. Lee, "ChimerDB 4.0: an updated and expanded database of fusion genes", *Nucleic Acids Research*, Vol. 48, No. D1, pp. D817–D824, 11 2019.

45. Lee, J., S. S. Yi, M. Jeong, M. Sung, W. Yoon, Y. Choi, M. Ko and J. Kang, "Answering Questions on COVID-19 in Real-Time", *arXiv preprint arXiv:2006.15830*, 2020.

46. Bravo, À., J. Piñero, N. Queralt-Rosinach, M. Rautschka and L. I. Furlong, "Extraction of relations between genes and diseases from text and large-scale data analysis: implications for translational research", *BMC Bioinformatics*, Vol. 16, No. 1, pp. 1–17, 2015.

47. Van Mulligen, E. M., A. Fourrier-Reglat, D. Gurwitz, M. Molokhia, A. Nieto, G. Trifiro, J. A. Kors and L. I. Furlong, "The EU-ADR corpus: annotated drugs, diseases, targets, and their relationships", *Journal of Biomedical Informatics*, Vol. 45, No. 5, pp. 879–884, 2012.

48. Krallinger, M., O. Rabal, S. A. Akhondi, M. P. Pérez, J. Santamaría, G. P. Rodríguez *et al.*, "Overview of the BioCreative VI chemical-protein interaction Track", *Proceedings of the Sixth BioCreative Challenge Evaluation Workshop*, Vol. 1, pp. 141–146, 2017.

49. Beltagy, I., K. Lo and A. Cohan, "SciBERT: Pretrained Language Model for Scientific Text", *EMNLP*, 2019.

50. Yang, Z., Z. Dai, Y. Yang, J. Carbonell, R. Salakhutdinov and Q. V. Le, "Xlnet:

Generalized autoregressive pretraining for language understanding", *arXiv preprint arXiv:1906.08237*, 2019.

51. Pennington, J., R. Socher and C. D. Manning, "Glove: Global vectors for word representation", *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1532–1543, 2014.

52. Peters, M. E., M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee and L. Zettlemoyer, "Deep contextualized word representations", *arXiv preprint arXiv:1802.05365*, 2018.

53. Rajpurkar, P., J. Zhang, K. Lopyrev and P. Liang, "Squad: 100,000+ questions for machine comprehension of text", *arXiv preprint arXiv:1606.05250*, 2016.

54. Socher, R., A. Perelygin, J. Wu, J. Chuang, C. D. Manning, A. Y. Ng and C. Potts, "Recursive deep models for semantic compositionality over a sentiment treebank", *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pp. 1631–1642, 2013.

55. Sang, E. F. and F. De Meulder, "Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition", *arXiv preprint cs/0306050*, 2003.

56. Dönmez, H., A. Köksal, E. Ozkirimli and A. Ozgür, "BOUN-REX at CLEF-2020 ChEMU Task 2: Evaluating Pretrained Transformers for Event Extraction", *11th Conference and Labs of the Evaluation Forum (CLEF-2020)*, 2020.

57. Wu, Y., M. Schuster, Z. Chen, Q. V. Le, M. Norouzi, W. Macherey, M. Krikun, Y. Cao, Q. Gao, K. Macherey *et al.*, "Google's neural machine translation system: Bridging the gap between human and machine translation", *arXiv preprint arXiv:1609.08144*, 2016.

58. Zhu, Y., R. Kiros, R. Zemel, R. Salakhutdinov, R. Urtasun, A. Torralba and

S. Fidler, "Aligning Books and Movies: Towards Story-Like Visual Explanations by Watching Movies and Reading Books", *2015 IEEE International Conference on Computer Vision (ICCV)*, pp. 19–27, 2015.

59. Wang, A., A. Singh, J. Michael, F. Hill, O. Levy and S. R. Bowman, "GLUE: A multi-task benchmark and analysis platform for natural language understanding", *arXiv preprint arXiv:1804.07461*, 2018.

60. Zellers, R., Y. Bisk, R. Schwartz and Y. Choi, "Swag: A large-scale adversarial dataset for grounded commonsense inference", *arXiv preprint arXiv:1808.05326*, 2018.

61. "ChemProt corpus: BioCreative VI", `https://biocreative.bioinformatics.udel.edu/news/corpora/chemprot-corpus-biocreative-vi/`, accessed in December 2019.

62. Kim, J.-D., T. Ohta, Y. Tateisi and J. Tsujii, "GENIA corpus—a semantically annotated corpus for bio-textmining", *Bioinformatics*, Vol. 19, 2003.

63. Sætre, R., K. Yoshida, A. Yakushiji, Y. Miyao, Y. Matsubayashi and T. Ohta, "AKANE system: protein-protein interaction pairs in BioCreAtIvE2 challenge, PPI-IPS subtask", *Proceedings of the Second Biocreative Challenge Workshop*, Vol. 209, p. 212, 2007.

64. Baldini Soares, L., N. FitzGerald, J. Ling and T. Kwiatkowski, "Matching the Blanks: Distributional Similarity for Relation Learning", *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 2895–2905, Jul. 2019.

65. Jurafsky, D. and J. H. Martin, *Speech and Language Processing*, 2020.

66. Qi, P., Y. Zhang, Y. Zhang, J. Bolton and C. D. Manning, "Stanza: A Python Natural Language Processing Toolkit for Many Human Languages", *Proceedings of*

the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations, pp. 101–108, Jul. 2020.

67. Manning, C., M. Surdeanu, J. Bauer, J. Finkel, S. Bethard and D. McClosky, "The Stanford CoreNLP Natural Language Processing Toolkit", *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pp. 55–60, Jun. 2014.

68. McDonald, R. and J. Nivre, "Analyzing and integrating dependency parsers", *Computational Linguistics*, Vol. 37, No. 1, pp. 197–230, 2011.

69. Jie, Z., A. Muis and W. Lu, "Efficient dependency-guided named entity recognition", *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 31, 2017.

70. Nakagawa, T., K. Inui and S. Kurohashi, "Dependency tree-based sentiment classification using CRFs with hidden variables", *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pp. 786–794, 2010.

71. Heilman, M. and N. A. Smith, "Tree edit models for recognizing textual entailments, paraphrases, and answers to questions", *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pp. 1011–1019, 2010.

72. Cui, H., R. Sun, K. Li, M.-Y. Kan and T.-S. Chua, "Question answering passage retrieval using dependency relations", *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 400–407, 2005.

73. Xu, H., C. Hu and G. Shen, "Discovery of dependency tree patterns for relation extraction", *Proceedings of the 23rd Pacific Asia Conference on Language, Information and Computation, Volume 2*, pp. 851–858, 2009.

74. Jin, L., L. Song, Y. Zhang, K. Xu, W.-y. Ma and D. Yu, "Relation extraction exploiting full dependency forests", *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34, pp. 8034–8041, 2020.

75. Fundel, K., R. Küffner and R. Zimmer, "RelEx—Relation extraction using dependency parse trees", *Bioinformatics*, Vol. 23, No. 3, pp. 365–371, 2007.

76. Loshchilov, I. and F. Hutter, "Decoupled weight decay regularization", *arXiv preprint arXiv:1711.05101*, 2017.

77. Kingma, D. P. and J. Ba, "Adam: A method for stochastic optimization", *arXiv preprint arXiv:1412.6980*, 2014.

78. Chen, Q., A. Allot and Z. Lu, "Keep up with the latest coronavirus research", *Nature*, Vol. 579, No. 7798, p. 193, 2020.

79. Dorche, M. S., P. Huot, M. Osherov, D. Wen, A. Saveriano, P. Giacomini, J. P. Antel and A. Mowla, "Neurological complications of coronavirus infection; a comparative review and lessons learned during the COVID-19 pandemic", *Journal of the Neurological Sciences*, p. 117085, 2020.

80. Gao, K., D. D. Nguyen, J. Chen, R. Wang and G.-W. Wei, "Repositioning of 8565 Existing Drugs for COVID-19", *The Journal of Physical Chemistry Letters*, Vol. 11, No. 13, pp. 5373–5382, 2020.

81. Tarasova, O., S. Ivanov, D. A. Filimonov and V. Poroikov, "Data and Text Mining Help Identify Key Proteins Involved in the Molecular Mechanisms Shared by SARS-CoV-2 and HIV-1", *Molecules*, Vol. 25, No. 12, p. 2944, 2020.

82. Wang, L. L., K. Lo, Y. Chandrasekhar, R. Reas, J. Yang, D. Eide, K. Funk, R. Kinney, Z. Liu, W. Merrill *et al.*, "CORD-19: The Covid-19 Open Research Dataset", *ArXiv*, 2020.

83. Su, D., Y. Xu, T. Yu, F. B. Siddique, E. J. Barezi and P. Fung, "CAiRE-COVID: A Question Answering and Multi-Document Summarization System for COVID-19 Research", *arXiv preprint arXiv:2005.03975*, 2020.

84. Esteva, A., A. Kale, R. Paulus, K. Hashimoto, W. Yin, D. Radev and R. Socher, "Co-search: Covid-19 information retrieval with semantic search, question answering, and abstractive summarization", *arXiv preprint arXiv:2006.09595*, 2020.

85. Wang, X., X. Song, Y. Guan, B. Li and J. Han, "Comprehensive named entity recognition on cord-19 with distant or weak supervision", *arXiv preprint arXiv:2003.12218*, 2020.

86. Tu, J., M. Verhagen, B. Cochran and J. Pustejovsky, "Exploration and discovery of the COVID-19 literature through semantic visualization", *arXiv preprint arXiv:2007.01800*, 2020.

87. D'Souza, J. and V. Ng, "Sieve-based entity linking for the biomedical domain", *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pp. 297–302, 2015.

88. Wei, C.-H., R. Leaman and Z. Lu, "Beyond accuracy: creating interoperable and scalable text-mining web services", *Bioinformatics*, Vol. 32, No. 12, pp. 1907–1910, 2016.

APPENDIX A: SAMPLE DATA IN CHEMPROT

Update on the use of aromatase inhibitors in breast cancer. Estrogens are biosynthesised from androgens by the CYP450 enzyme complex called aromatase. Aromatase is expressed in the ovary, placenta, brain, bone, adipose tissue and breast tissue. In breast cancer, intratumoural aromatase is the source for local estrogen production in the tissue. Inhibition of aromatase is an important approach for reducing growth stimulatory effects of estrogens in estrogen-dependent breast cancer. The potent and selective third-generation aromatase inhibitors anastrozole, letrozole and exemestane were introduced to the market as endocrine therapy in postmenopausal patients failing anti-estrogen therapy alone, or multiple hormonal therapies. Anastrozole and letrozole are both non-steroidal aromatase inhibitors that compete with the substrate for binding to the enzyme active site. Exemestane is a mechanism-based steroidal inhibitor that mimics the substrate, is converted by the enzyme to a reactive intermediate, and results in inactivation of aromatase. These third-generation aromatase inhibitors are currently approved as first-line therapy for the treatment of postmenopausal women with metastatic estrogen-dependent breast cancer. The use of an aromatase inhibitor as initial therapy, or after treatment with tamoxifen, is now recommended as adjuvant hormonal therapy for postmenopausal women with hormone-dependent breast cancer. Several clinical studies of aromatase inhibitors focus on the use of these agents in the adjuvant setting, for the treatment of early breast cancer. Recently published results show improved responses with these agents compared with tamoxifen.

Figure A.1. An example abstract in ChemProt with 17020418 PMID (PubMed identifier). It has 12 protein/gene entities and 15 chemical entities marked in blue and yellow, respectively.

17020418	T1	CHEMICAL	60	69	Estrogens
17020418	T2	CHEMICAL	1197	1205	estrogen
17020418	T3	CHEMICAL	1309	1318	tamoxifen
17020418	T4	CHEMICAL	1663	1672	tamoxifen
17020418	T5	CHEMICAL	311	319	estrogen
17020418	T6	CHEMICAL	94	103	androgens
17020418	T7	CHEMICAL	438	447	estrogens
17020418	T8	CHEMICAL	451	459	estrogen
17020418	T9	CHEMICAL	548	559	anastrozole
17020418	T10	CHEMICAL	561	570	letrozole
17020418	T11	CHEMICAL	677	685	estrogen
17020418	T12	CHEMICAL	733	744	Anastrozole
17020418	T13	CHEMICAL	749	758	letrozole
17020418	T14	CHEMICAL	772	781	steroidal
17020418	T15	CHEMICAL	906	915	steroidal
17020418	T16	GENE-Y	1073	1082	aromatase
17020418	T17	GENE-Y	1245	1254	aromatase
17020418	T18	GENE-Y	1459	1468	aromatase
17020418	T19	GENE-Y	277	286	aromatase
17020418	T20	GENE-Y	360	369	aromatase
17020418	T21	GENE-Y	527	536	aromatase
17020418	T22	GENE-N	111	117	CYP450
17020418	T23	GENE-Y	782	791	aromatase
17020418	T24	GENE-Y	140	149	aromatase
17020418	T25	GENE-Y	151	160	Aromatase
17020418	T26	GENE-Y	1039	1048	aromatase
17020418	T27	GENE-Y	21	30	aromatase

Figure A.2. All entities and their related information from the example abstract with 17020418 PMID. Chemicals (yellow) whose entity numbers are colored green have relations with protein/genes (blue) whose entity numbers are highlighted in purple.

17020418	CPR:4	Y	INHIBITOR	Arg1:T10	Arg2:T21
17020418	CPR:4	Y	INHIBITOR	Arg1:T12	Arg2:T23
17020418	CPR:4	Y	INHIBITOR	Arg1:T13	Arg2:T23
17020418	CPR:4	Y	INHIBITOR	Arg1:T9	Arg2:T21
17020418	CPR:9	Y	PRODUCT-OF	Arg1:T1	Arg2:T22
17020418	CPR:9	Y	PRODUCT-OF	Arg1:T1	Arg2:T24
17020418	CPR:9	Y	PRODUCT-OF	Arg1:T5	Arg2:T19
17020418	CPR:9	Y	SUBSTRATE	Arg1:T6	Arg2:T24

Figure A.3. Relations between chemicals and proteins/genes from the example abstract with 17020418 PMID. In each row, an entity number marked in green is a chemical having a relation with a protein whose entity number is colored purple.

APPENDIX B: EXAMPLES OF DEPENDENCY TREES

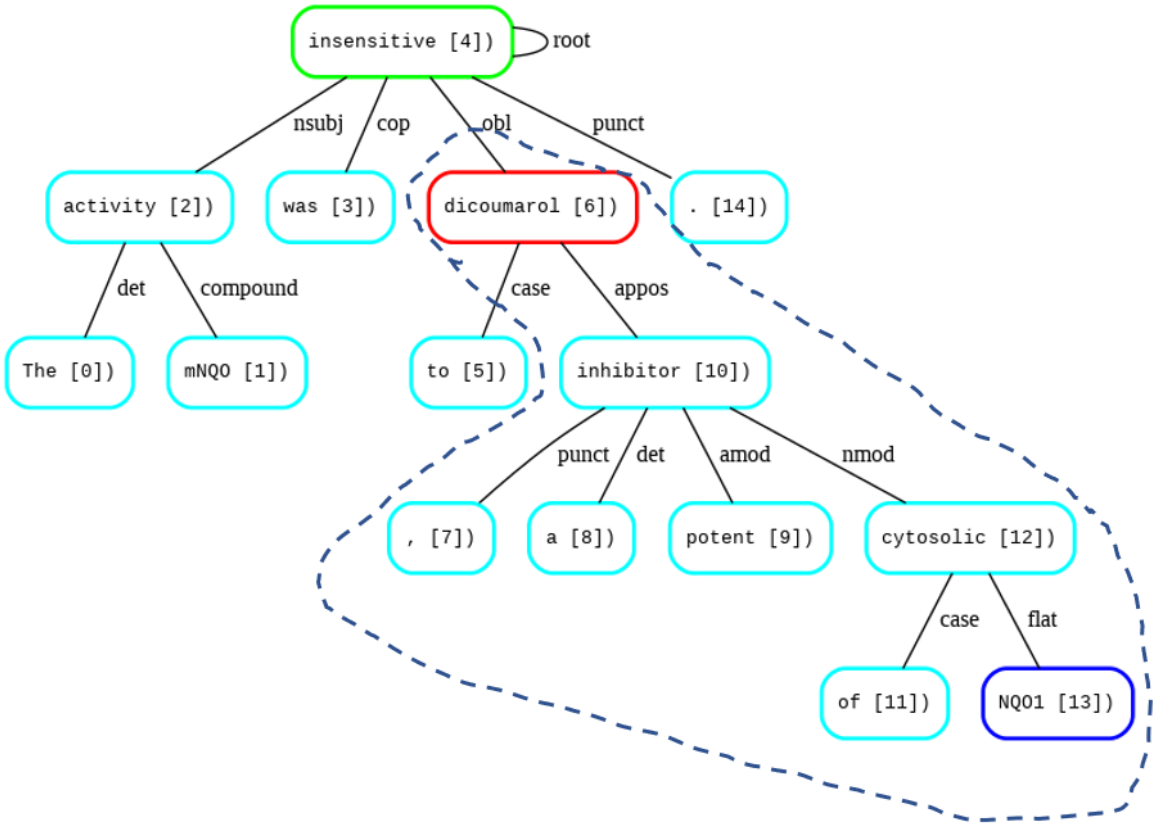

Figure B.1. Dependency tree of the first sentence in Table 3.11. Since *dicoumarol* is a chemical and *NQO1* is a protein/gene, their nodes are colored in red and blue in the dependency tree, respectively. They have a CPR:4 relation.

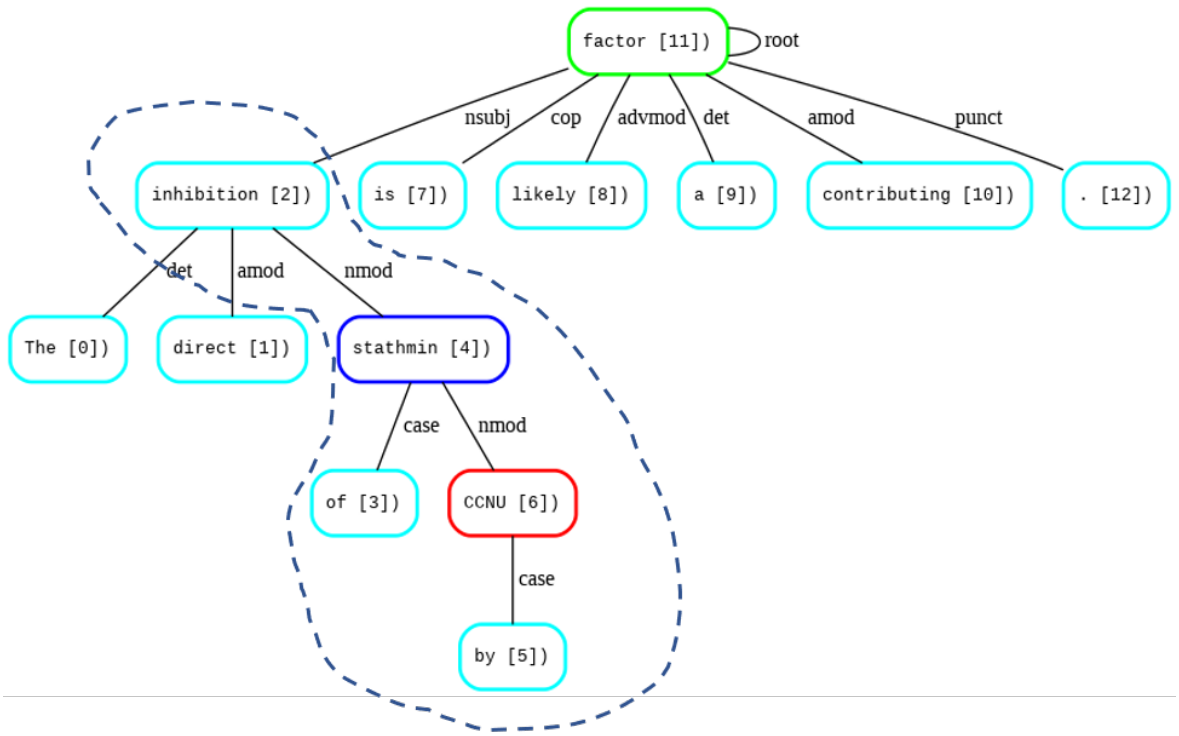

Figure B.2. Dependency tree of the first sentence in Table 3.12. Since *CCNU* is a chemical and *stathmin* is a protein/gene, their nodes are colored in red and blue in the dependency tree, respectively. They have a CPR:4 relation.

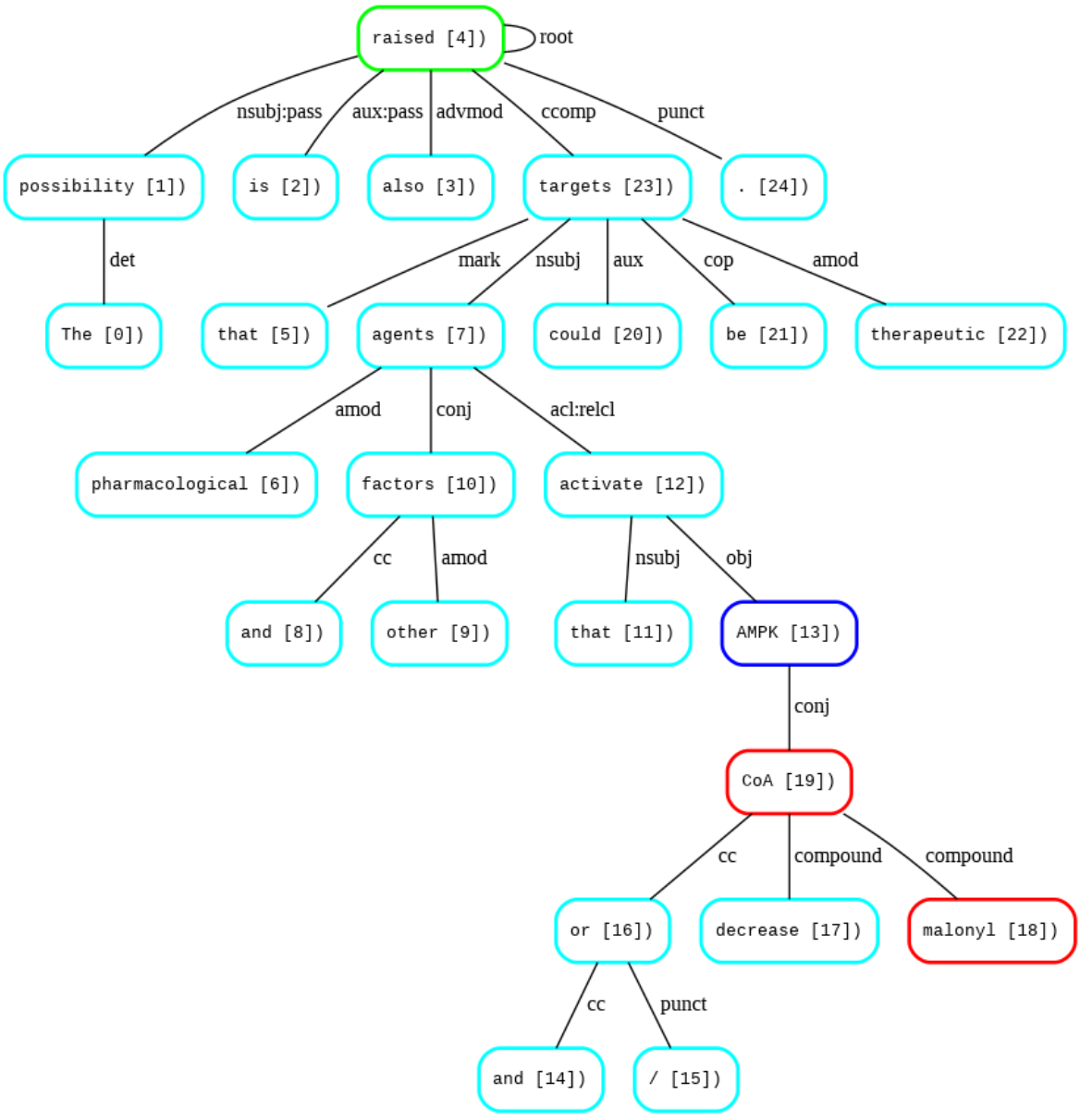

Figure B.3. Dependency tree of the second sentence in Table 3.11. The sentence is from an abstract with 16642960 PMID. Since *AMPK* is a protein/gene and *malonyl CoA* is a chemical, their nodes are colored in blue and red in the tree, respectively.

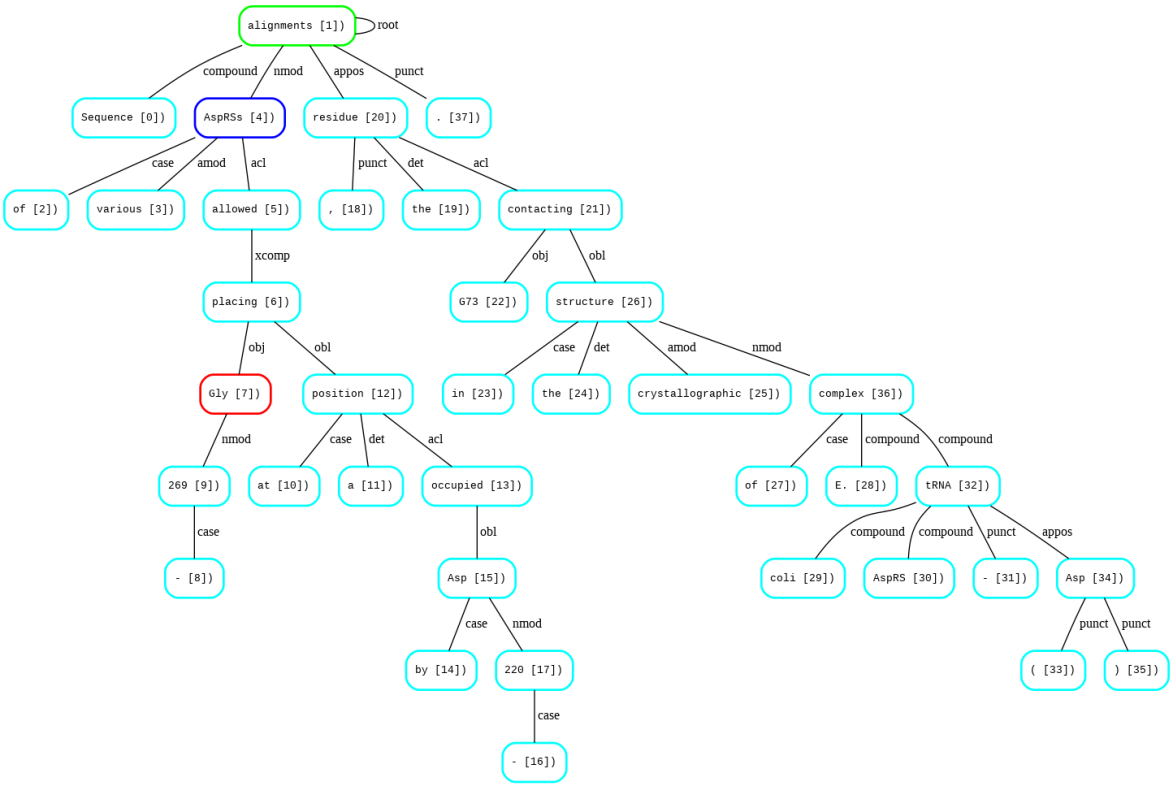

Figure B.4. Dependency tree of the last sentence in Table 3.11. The sentence is from an abstract with 165997625 PMID. Since *AspRSs* is a protein/gene and *Gly* is a chemical, their nodes are colored in blue and red in the dependency tree, respectively.

APPENDIX C: SAMPLE SENTENCES FROM CORD-19 EVALUATED BY A DOMAIN EXPERT

Table C.1. Sample sentences from CORD-19 predicted to have a biochemical relation by our model.

Sentence	**Identified Chemical**	**Identified Protein**	**Relation Label**	**Chemical Label**	**Protein Label**
The N-terminal domain of the coronavirus nucleocapsid (N) protein adopts a fold resembling a right hand with a flexible, positively charged b-hairpin and a hydrophobic palm.	N	coronavirus nucleocapsid (N) protein	0	0	1
Alanine substitution of either Arg-76 or Tyr-94 in the N-terminal domain of IBV N protein led to a significant decrease in its RNA-binding activity and a total loss of the infectivity of the viral RNA to Vero cells.	Alanine	IBV N protein	0	0	1

Table C.1. Sample sentences from CORD-19 predicted by our model (cont.)

Sentence	**Identified Chemical**	**Identified Protein**	**Relation Label**	**Chemical Label**	**Protein Label**
Rat microsomal aldehyde dehydrogenase (msALDH) has no amino-terminal signal sequence, but instead it has a characteristic hydrophobic domain at the carboxyl terminus (Miyauehi, K., R.	carboxyl	msALDH	0	0	1
The IMPDH inhibitor merimepodib (MMPD) is an investigational antiviral drug that acts as a noncompetitive inhibitor of IMPDH.	merimepodib	IMPDH	1	1	1

Table C.1. Sample sentences from CORD-19 predicted by our model (cont.)

Sentence	**Identified Chemical**	**Identified Protein**	**Relation Label**	**Chemical Label**	**Protein Label**
EMA/PMA treatments significantly decreased amplifiable hg/cont and significantly increased the number of PVDA positive probes and their signal intensity compared to untreated spiked lung samples.	EMA	hg/cont	0	1	0
EMA/PMA treatments significantly decreased amplifiable hg/cont and significantly increased the number of PVDA positive probes and their signal intensity compared to untreated spiked lung samples.	PMA	hg/cont	0	1	0

Table C.1. Sample sentences from CORD-19 predicted by our model (cont.)

Sentence	**Identified Chemical**	**Identified Protein**	**Relation Label**	**Chemical Label**	**Protein Label**
Thus, EMA/PMA treatments offer a new approach to lower the amplifiable hg/cont in clinical samples and increase the success of PVDA and HTS to identify viruses.	EMA	hg/cont	0	1	0
A role for glycosylation in cell-surface tetherin expression is supported by tunicamycin treatment, which inhibits the first step of N-linked glycosylation and impairs both cell-surface expression and antiviral activity.	tunicamycin	tetherin	1	1	1

Table C.1. Sample sentences from CORD-19 predicted by our model (cont.)

Sentence	**Identified Chemical**	**Identified Protein**	**Relation Label**	**Chemical Label**	**Protein Label**
The effects of cholesterol depletion and PI3K/AKT signaling pathway activation during S. agalactiae-human umbilical vein endothelial cells (HUVEC) interaction were analysed by pre-treatment with methyl-β-cyclodextrin (MβCD) or LY294002 inhibitors, immunofluorescence and immunoblot analysis.	MbCD	PI3K	1	1	1

Table C.1. Sample sentences from CORD-19 predicted by our model (cont.)

Sentence	**Identified Chemical**	**Identified Protein**	**Relation Label**	**Chemical Label**	**Protein Label**
Inhibition of complex-type glycosylation with kifunensine, an inhibitor of the oligosaccharide processing enzyme mannosidase 1, had no effect on either the cell-surface expression or antiviral activity of tetherin.	kifunensine	oligosaccharide processing enzyme mannosidase 1	1	1	1
Conversely, results from cell-based small molecule screening studies have shown that the antibiotic hexachlorophene can down-regulate b-catenin in colon cancer cells.	hexachlorophene	b-catenin	1	1	1

Table C.1. Sample sentences from CORD-19 predicted by our model (cont.)

Sentence	**Identified Chemical**	**Identified Protein**	**Relation Label**	**Chemical Label**	**Protein Label**
The CT of tetherin contains an "STS" sequence that is implicated in ubiquitylation, and a highly-conserved tyrosine-based motif, "YxxY", that is essential for clathrin-dependent endocytosis of tetherin, and activation of nuclear factor-κB (NF-κB) [16] [17] [18] [19] .	tyrosine	tetherin	0	0	1
Here we report that hexachlorophene also counteracts the elevated bcatenin levels in EBV-infected B lymphomas.	hexachlorophene	bcatenin	1	1	1

Table C.1. Sample sentences from CORD-19 predicted by our model (cont.)

Sentence	**Identified Chemical**	**Identified Protein**	**Relation Label**	**Chemical Label**	**Protein Label**
Our results suggest that Siah-1 is targeted by both LMP-1 and hexachlorophene with opposite effects.	hexachlorophene	Siah-1	1	1	1
The hexachlorophene modulation of Siah-1 and b-catenin is independent of p53 and results in reduced expression of cyclin-D1 and c-Myc (target genes of b-catenin), leading to the growth arrest of B lymphoma cells.	hexachlorophene	Siah-1	1	1	1

Table C.1. Sample sentences from CORD-19 predicted by our model (cont.)

Sentence	**Identified Chemical**	**Identified Protein**	**Relation Label**	**Chemical Label**	**Protein Label**
Alanine substitution of either Arg-76 or Tyr-94 in the N-terminal domain of IBV N protein led to a significant decrease in its RNA-binding activity and a total loss of the infectivity of the viral RNA to Vero cells.	Alanine	IBV N protein	0	0	1
Rat microsomal aldehyde dehydrogenase (msALDH) has no amino-terminal signal sequence, but instead it has a characteristic hydrophobic domain at the carboxyl terminus (Miyauehi, K., R.	carboxyl	Rat microsomal aldehyde dehydrogenase	0	0	1

Table C.1. Sample sentences from CORD-19 predicted by our model (cont.)

Sentence	**Identified Chemical**	**Identified Protein**	**Relation Label**	**Chemical Label**	**Protein Label**
Also, the protease Factor Xa, a target of Ben-HCl abundantly expressed in infected cells, was able to cleave the recombinant and pseudoviral S protein into S1 and S2 subunits, and the cleavage was inhibited by Ben-HCl.	Ben	protease Factor Xa	0	0	1
Since the 3a protein forms ion channels, we were interested to see any conformational changes occurring in the Cyot3a upon calcium binding, using fluorescence spectroscopy and circular dichroism.	calcium	3a protein	1	1	1

Table C.1. Sample sentences from CORD-19 predicted by our model (cont.)

Sentence	**Identified Chemical**	**Identified Protein**	**Relation Label**	**Chemical Label**	**Protein Label**
These studies clearly indicate a significant change in the conformation of the Cyto3a protein after binding with calcium.	calcium	Cyto3a protein	1	1	1
Our results strongly suggest that the cytoplasmic domain of the 3a protein of SARS-CoV binds calcium in vitro, causing a change in protein conformation.	calcium	3a protein	1	1	1
Further, the drug hexamethylene amiloride (HMA), but not amiloride, inhibited in vitro ion channel activity of some synthetic coronavirus E proteins, and also viral replication.	hexamethylene amiloride	synthetic coronavirus E proteins	1	1	1

Table C.1. Sample sentences from CORD-19 predicted by our model (cont.)

Sentence	**Identified Chemical**	**Identified Protein**	**Relation Label**	**Chemical Label**	**Protein Label**
The high prevalence of variants in the G6PD gene found in this analysis suggests that it may be a significant interaction factor in clinical trials of chloroquine and hydrochloroquine for treatment of COVID-19 in Africans.	chloroquine	G6PD gene	1	1	1
In an effort to circumvent resistance to rapamycin -an mTOR inhibitor -we searched for novel rapamycindownstream-targets that may be key players in the response of cancer cells to therapy.	rapamycin	mTOR	1	1	1

Table C.1. Sample sentences from CORD-19 predicted by our model (cont.)

Sentence	**Identified Chemical**	**Identified Protein**	**Relation Label**	**Chemical Label**	**Protein Label**
We found that rapamycin, at nM concentrations, increased phosphorylation of eukaryotic initiation factor (eIF) 2a in rapamycin-sensitive and estrogen-dependent MCF-7 cells, but had only a minimal effect on eIF2a phosphorylation in the rapamycininsensitive triple-negative MDA-MB-231 cells.	rapamycin	eukaryotic initiation factor (eIF) 2a	1	1	1

Table C.1. Sample sentences from CORD-19 predicted by our model (cont.)

Sentence	**Identified Chemical**	**Identified Protein**	**Relation Label**	**Chemical Label**	**Protein Label**
Addition of salubrinal -an inhibitor of eIF2a dephosphorylationdecreased expression of a surface marker associated with capacity for self renewal, increased senescence and induced clonogenic cell death, suggesting that excessive phosphorylation of eIF2a is detrimental to the cells' survival.	salubrinal	eIF2a dephosphorylation-decreased	1	1	1

Table C.1. Sample sentences from CORD-19 predicted by our model (cont.)

Sentence	**Identified Chemical**	**Identified Protein**	**Relation Label**	**Chemical Label**	**Protein Label**
Treating cells with salubrinal enhanced radiation-induced increase in eIF2a phosphorylation and clonogenic death and showed that irradiated cells are more sensitive to increased eIF2a phosphorylation than non-irradiated ones.	salubrinal	eIF2a	1	1	1

Table C.1. Sample sentences from CORD-19 predicted by our model (cont.)

Sentence	**Identified Chemical**	**Identified Protein**	**Relation Label**	**Chemical Label**	**Protein Label**
Similar to salubrinal -the phosphomimetic eIF2a variant -S51D -increased sensitivity to radiation, and both abrogated radiation-induced increase in breast cancer type 1 susceptibility gene, thus implicating enhanced phosphorylation of eIF2a in modulation of DNA repair.	salubrinal	eIF2a variant	1	1	1
In addition to its effect on radiation, salubrinal enhanced eIF2a phosphorylation and clonogenic death in response to the histone deacetylase inhibitor -vorinostat.	salubrinal	eIF2a	1	1	1

Table C.1. Sample sentences from CORD-19 predicted by our model (cont.)

Sentence	**Identified Chemical**	**Identified Protein**	**Relation Label**	**Chemical Label**	**Protein Label**
Finally, the catalytic competitive inhibitor of mTOR -Ku-0063794 -increased phosphorylation of eIF2a demonstrating further the involvement of mTOR activity in modulating eIF2a phosphorylation.	Ku-0063794	mTOR	1	1	1
Moreover, glycyrrhizin treatment still enhanced IFN-γ and reduced IL-4 levels in glycyrrhizin-treated mice.	glycyrrhizin	IFN-g	1	1	1
While determining the 5' ends of C. elegans actin mRNAs, we have discovered a 22 nucleotide spliced leader sequence.	nucleotide	C. elegans actin mRNAs	0	0	0

Table C.1. Sample sentences from CORD-19 predicted by our model (cont.)

Sentence	**Identified Chemical**	**Identified Protein**	**Relation Label**	**Chemical Label**	**Protein Label**
The actin mRNA leader sequence is identical to the first 22 nucleotides of a novel 100 nucleotide RNA transcribed adjacent, and in the opposite orientation, to the 5S ribosomal gene.	nucleotides	actin mRNA leader sequence	0	0	0
The evidence suggests that the actln mRNA leader sequence is acquired from this novel nucleotide transcript by an intermolecular frans-splicing mechanism.	nucleotide	actln mRNA leader sequence	0	0	0

Table C.1. Sample sentences from CORD-19 predicted by our model (cont.)

Sentence	**Identified Chemical**	**Identified Protein**	**Relation Label**	**Chemical Label**	**Protein Label**
We found that the aspartic acid at position 95, previously believed to be required for binding of PSGs to cells, is not required for PSG1 activity but that the amino acids implicated in the formation of a salt bridge within the N-domain are essential for PSG1 function.	aspartic acid	PSGs	0	0	1
In vitro expression of a construct containing the Lb gene fused to a portion of the VP4 and 3D genes demonstrated cis cleavage activity that could be blocked by the thiol protease inhibitor E-64.	E-64	Lb gene	1	1	1

Table C.1. Sample sentences from CORD-19 predicted by our model (cont.)

Sentence	**Identified Chemical**	**Identified Protein**	**Relation Label**	**Chemical Label**	**Protein Label**
In contrast, conversion of LC3-I/LC3-II could be significantly inhibited by 4-PBA, an ER stress inhibitor, indicating that ORF3-induced autophagy is dependent on ER stress response.	4-PBA	LC3-I	1	1	1

Table C.1. Sample sentences from CORD-19 predicted by our model (cont.)

Sentence	**Identified Chemical**	**Identified Protein**	**Relation Label**	**Chemical Label**	**Protein Label**
All the studied molecules could bind to the active site of the SARS-CoV-2 protease (PDB: 6Y84), out of which rutin (a natural compound) has the highest inhibitor efficiency among the 33 molecules studied, followed by ritonavir (control drug), emetine (anti-protozoal), hesperidin (a natural compound), lopinavir (control drug) and indinavir (anti-viral drug).	rutin	SARS-CoV-2 protease	1	1	1

Table C.1. Sample sentences from CORD-19 predicted by our model (cont.)

Sentence	**Identified Chemical**	**Identified Protein**	**Relation Label**	**Chemical Label**	**Protein Label**
Using degenerate PCR primers complementary to the most conserved genome regions of adenoviruses, the complete nucleotide sequence of the penton base gene, and partial nucleotide sequences of the DNA polymerase, hexon, and pVII genes were obtained.	nucleotide	penton base gene	0	0	1
Estradiol is connected with CD4+ T cell numbers and increases T-reg cell populations, affecting immune responses to infection.	Estradiol	CD4+	1	1	1

Table C.1. Sample sentences from CORD-19 predicted by our model (cont.)

Sentence	**Identified Chemical**	**Identified Protein**	**Relation Label**	**Chemical Label**	**Protein Label**
It is known that estradiol exerts a protective effect on endothelial function, activating the generation of nitric oxide (NO) via endothelial nitric oxide synthase.	estradiol	endothelial nitric oxide synthase	1	1	1

APPENDIX D: SIGNIFICANT TEST RESULTS

- Significant test for binary models with sentence-based input representation and shortest subsequence between target entities

Table D.1. Significant test for our best two BioBERT-based models with sentence-based input representation and shortest subsequence between target entities.

Iteration	**F1-score for SB**	**F1-Score for SST**
0	0.781	0.766
1	0.772	0.745
2	0.779	0.758
3	0.785	0.764
4	0.767	0.764
5	0.781	0.750
6	0.768	0.746
7	0.783	0.757
8	0.782	0.739
9	0.779	0.748
Mean	0.778	0.754
Standard Deviation	0.0061	0.0089
t value	6.696	

- Significant test for binary models with sentence-based input representation and shortest subsequence including target entities and their parent node

Table D.2. Significant test for our best two BioBERT-based models with sentence-based input representation and shortest subsequence including target entities and their parent.

Iteration	**F1-score for SB**	**F1-Score for SSTP**
0	0.781	0.747
1	0.772	0.767
2	0.779	0.759
3	0.785	0.757
4	0.767	0.766
5	0.781	0.764
6	0.768	0.765
7	0.783	0.744
8	0.782	0.764
9	0.779	0.754
Mean	0.778	0.759
Standard Deviation	0.0061	0.0077
t value	5.811	

- Significant test for multi-class models with sentence-based input representation and shortest subsequence between target entities

Table D.3. Significant test for our best two multi-class BioBERT-based models with sentence-based input representation and shortest subsequence between target entities.

Iteration	**Micro-averaged F1-score for SB**	**Micro-averaged F1-Score for SST**
0	0.765	0.732
1	0.756	0.730
2	0.735	0.735
3	0.769	0.725
4	0.770	0.732
5	0.765	0.727
6	0.754	0.731
7	0.765	0.720
8	0.773	0.744
9	0.757	0.740
Mean	0.761	0.732
Standard Deviation	0.0105	0.0066
t value	7.08	

- Significant test for multi-class models with sentence-based input representation and shortest subsequence including target entities and their parent node

Table D.4. Significant test for our best two multi-class BioBERT-based models with sentence-based input representation and shortest subsequence including target entities and their parent.

Iteration	**Micro-averaged F1-score for SB**	**Micro-averaged F1-Score for SST**
0	0.765	0.741
1	0.756	0.734
2	0.735	0.733
3	0.769	0.735
4	0.770	0.739
5	0.765	0.727
6	0.754	0.742
7	0.765	0.729
8	0.773	0.735
9	0.757	0.722
Mean	0.761	0.734
Standard Deviation	0.0105	0.006
t value	6.751	