A Novel Sequence Labeling Framework for Extracting Drug-Protein Relations from Biomedical Literature

Ling Luo, Po-Ting Lai, Chih-Hsuan Wei, Zhiyong Lu*

National Center for Biotechnology Information (NCBI), National Library of Medicine (NLM), National Institutes of Health (NIH), Bethesda, MD 20894, USA

*Corresponding author: zhiyong.lu@nih.gov

Abstract

Automatic extracting interactions between chemical compound/drug and gene/protein is significantly beneficial to drug discovery, drug repurposing, drug design, and biomedical knowledge graph construction. To promote the development and evaluation of systems that are able to automatically detect in relations between drug and protein, the DrugProt track was organized during the BioCreative VII challenge. This paper describes our approach to this challenging task. In addition to the conventional text classification framework that has been widely used in relation extraction tasks, we propose a novel sequence labeling framework to drug-protein relation extraction. We first comprehensively compared the cutting-edge biomedical pre-trained language models for both frameworks. Then, we explored several ensemble methods to further improve the final performance. In the challenge, our best submission (i.e., the ensemble of models in two frameworks via major voting) achieved the F1-score of 0.795 on the official test set. Further, we realized the sequence labeling framework is more efficient and achieves better performance than does the text classification framework. Finally, our ensemble of the sequence labeling models with majority voting achieves the best F1-score of 0.800 on the test set.

Database URL: https://github.com/lingluodlut/BioCreativeVII_DrugProt

Keywords: relation extraction, drug-protein interaction, pre-trained language models, sequence labeling.

1 Introduction

A significant number of associations and interactions between chemical/drug and gene/protein have been published in scientific literature. Such unstructured texts, however, are difficult to summarize in a structured format for automatic data analysis. Although a significant amount of manual curation has been involved in summarizing and depositing the information from the literature (e.g., CTD (1) and ChemProt (2)), the process is extremely time-consuming. Given the

rapid growth of the literature, developing computational methods to accelerate the efficiency of the knowledge extraction is critical (3, 4). Drugs and proteins play important roles in metabolism and the regulation of biological processes. Extracting the relations between drugs and proteins from the biomedical literature is crucial to various biomedical tasks such as drug discovery, drug repurposing, drug-induced adverse reactions, and biomedical knowledge graph construction. Such extraction is important not only for biological purposes but also to pharmacological and clinical research. In this regard, certain types of relations are highly relevant, including those among metabolic relations, antagonists, agonists, inhibitors, and activators.

To accelerate the advancement of a method to extract the relations between chemicals and proteins, the BioCreative VI organized the CHEMPROT task (5) for chemical-protein relations. The task requires the determination of whether an interaction exists between a chemical and protein pair in the given text (most are in a sentence), and identifying the type of interaction. Recently, in BioCreative VII, a related task was organized (i.e., BioCreative VII DrugProt (6)), which significantly increased the corpus from 2,432 to 5,000 abstracts. Further, DrugProt increased the number of granular relation categories from 5 to 13 to evaluate the performance of the system.

Compared to other biomedical entity relation extraction tasks (such as protein-protein interactions (7)), chemical-protein relation extraction was rarely discussed in the context of BioCreative, as no other publicly available corpus was available. In 2017, most of the participants in BioCreative VI applied deep learning-based neural network methods (e.g., convolutional- and recurrent- neural networks) to extract the relation pairs of chemicals and proteins (5). During the BioCreative challenge, the ensemble of support vector machine (SVM) and deep learning models achieved the best performance, with an F1-score of 0.641 (8). Later, transformer-based pre-trained language models, such as BERT (9), have shown the capability of the contextual representation at large volumes. Some biomedical versions of the transformer-based model (such as BioBERT (10), PubmedBERT(11), and BioM-ALBERT(12)) were proposed for the biomedical domain. These biomedical pre-trained models show promising results and significantly outperform previous state-of-the-art models. For example, BioM-ALBERT achieves an F1-score of 0.793 on the corpus of the BioCreative VI ChemProt task (12). In general, most of these methods treat the relation extraction task as a text classification problem. A given text with the highlighted chemical and protein is provided to the machine learning-based classifier to predict the predefined relation types (or no relation). This type of method is required to process all of the pairs between two entities, one by one, which is time-consuming, and it is difficult to handle large-scale data using advanced deep learning techniques. Moreover, these methods ignore the dependency between multiple relations, as they deconstruct the relation extraction into multiple independent relation classification subtasks.

To address these problems, we propose a novel sequence labeling framework for the drug-protein relation extraction task. In our framework, we convert the task to a sequence labeling problem. Different from the conventional classification framework, our framework can recognize

at once all possible entities associated with the target entity. It is more efficient and is able to fully exploit the dependencies of relations for improved performance. We also investigated multiple biomedical pre-trained language models (PLMs) for both frameworks. Further, we applied several ensemble methods to optimize the final performance. The experimental results show that the ensemble of our sequence labeling models with major voting achieves the highest F1-score of 0.800 on the test set.

2 Materials and methods

2.1 Dataset

In the BioCreative VII DrugProt track, the organizers released a large manually labeled corpus that included annotations of mentions of drugs (including chemical compounds and drugs) as well as proteins (including genes, proteins, and miRNA) and their relations. The DrugProt corpus for the DrugProt main task is split into training, development, and test sets. In addition, 10,000 background abstracts without annotations are available for specific use cases, e.g., fine-tuning language models. Further, around 2,400,000 PubMed records were annotated by their ML-based drug and protein recognizers, and this dataset is used as the DrugProt Large Scale subtask (6). Table 1 shows an overview of the DrugProt corpus.

Table 1. The overview of DrugProt corpus

Set	Number of abstracts	Number of entities		Number of relations
		Drug	Protein	
Training	3,500	46,274	43,255	17,274
Development	750	9,853	9,005	3,761
Test	750	9,434	9,515	3,491
Background	10,000	134,333	157,523	-
Large Scale	2,366,081	33,578,479	20,415,123	-

For the relations in the DrugProt corpus, a total of 13 relation types were used, including INDIRECT-DOWNREGULATOR, INDIRECT-UPREGULATOR, DIRECT-REGULATOR, ACTIVATOR, INHIBITOR, AGONIST, ANTAGONIST, AGONIST-ACTIVATOR, AGONIST-INHIBITOR, PRODUCT-OF, SUBSTRATE, SUBSTRATE_PRODUCT-OF, and PART-OF. Figure 1 shows a more granular overview of the annotations for each of the relation classes in the training and development merged set. We can see that the distribution of the relation classes is imbalanced. The most major relation type is INHIBITOR with over 6,000 instances, but some relation types have fewer than 50 instances (e.g., AGNONIST-INHIBITOR). According to the characteristic of the corpus, less than 1% of the relations cross multiple sentences in the training set. Therefore, we focus on the challenges of relation extraction within sentences. And we combined the relation that crosses multiple sentences into a new long sentence for training models.

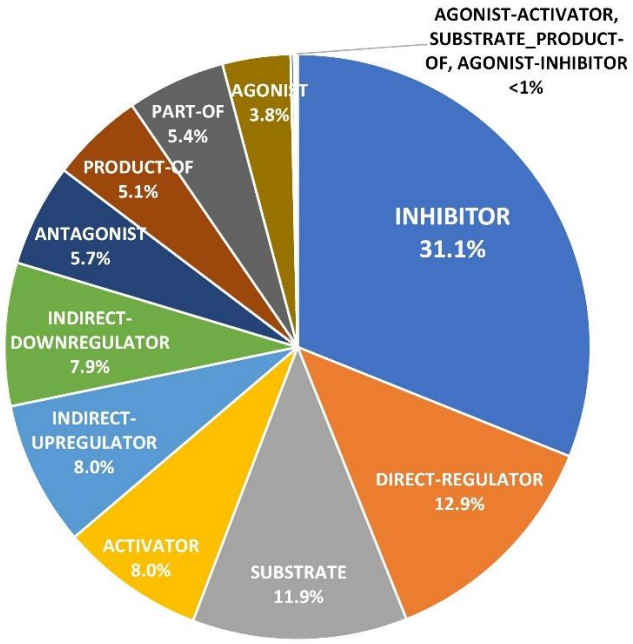

Figure 1. The distribution of the relation classes in the DrugProt in the training and development merged set.

2.2 Text classification framework

To provide a comparison with our proposed sequence labeling framework, we also built the corresponding neural network models in a conventional text classification framework with minimal architectural modification. The two frameworks are illustrated in Figure 2. We first introduce the basic text classification framework for this task as part of our methods. In this framework, the relation extraction task is treated as a multi-class classification problem. As shown in Figure 2A, we first need to generate all drug-protein entity pairs in the sentence as the input instances and then process all entity pairs one by one. For the input instances, we inserted “<Arg1></Arg1>” and “<Arg2></Arg2>” tag pairs in the front and end of the target entity pair, the “<Drug></Drug>” tag pairs to the other drugs, and “<Prot></Prot>” tag pairs to the other proteins. Given the input instance, we built a classifier to classify the relation (or lack) of entity pairs into a predefined relation type. We used the biomedical pre-trained language model (PLM) to encode the input text. Then the [CLS]’s output vector at the last hidden layer of the PLM is fed into a fully connected layer with ReLU (13) activation function. Finally, a softmax classification layer is used to classify the relation of an entity pair. In the experiments, we evaluated the five biomedical PLMs including PubMedBERT (11), BioBERT (10), BioRoBERTa (14), BioM-ELECTRA (12), and BioM-ALBERT (12) to this task.

2.3 Sequence labeling framework

Inspired by our previous work (15, 16), we proposed a novel sequence labeling framework to address the sentence-level biomedical relation extraction task. As shown in Figure 2B, we converted the task to a sequence labeling problem. Given a candidate source entity (e.g., drug entity of “icariin”) in the sentence, the goal of the model is to recognize all the corresponding target entities (e.g., protein entities of “PDE5” and “PDE4”) that are involved in the drug-protein relations with the target entity. Based on selecting different entity types as the source/target entity, there are two different labeling strategies to extract drug-protein relations, including (1) from drug to protein (D→P), selecting drug entity as the source entity to extract the corresponding target protein entity; and (2) from protein to drug (P→D), selecting protein entity as the source entity to extract the corresponding target drug entity. From the example shown in Figure 2, we can see that the text classification framework deconstructed input text into two independent sentence classification subtasks, while our sequence labeling framework can effectively narrow this down to one sequence labeling task. Therefore, our sequence lableling framework is able to fully exploit the dependencies of the source entity and all corresponding target entities in one instance and, thus, can be considered more efficient. Next, we use D→P as the example to describe the details of our method.

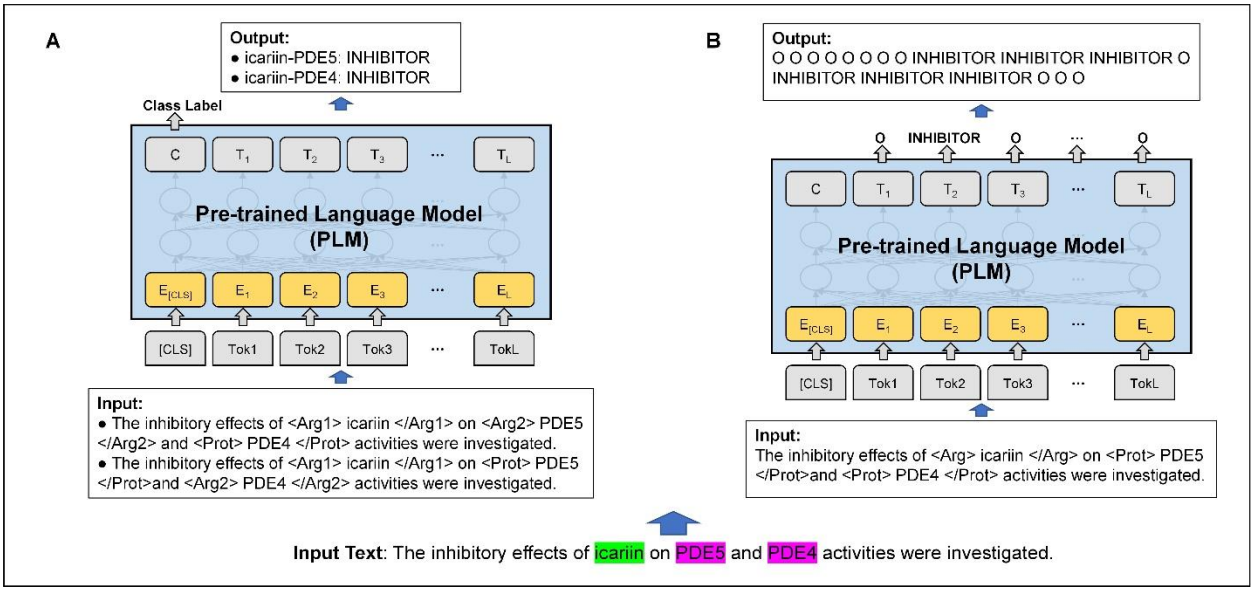

Figure 2. Example of relation extraction in text classification and sequence labeling frameworks. **(A)** Text classification framework. **(B)** Sequence labeling framework. In the example, the drug of “icariin” has INHIBITOR relations with the proteins of “PDE5” and “PDE4”.

(1) Tagging scheme. First, we designed a simple and effective tagging scheme for this task. For the input sequence, we insert “<Arg>” and “</Arg>” tags in the front and end of the source

entity to note the current drug entity. In addition, to note the other drug and protein entities in the text, we inserted “<Drug>” and “</Drug>” in the front and the end of the drugs, and “<Prot>” and “</Prot>” to the proteins. For the output sequence, each token is assigned a label that contributes to the extraction. The tokens involved in the relations are tagged by the corresponding relation type labels (e.g., INHIBITOR), which are predefined according to the training sets. We used the label “O” to represent other tokens that do not involve in a relation. Figure 2B shows an example of tagging a sentence with our tagging scheme, according to the original gold standard annotations of the DrugProt dataset. In this example, the input sentence contains three entities (i.e., the drug entity of “icariin” and the protein entities of “PDE5” and “PDE4”) and two drug-protein relation triples (i.e., {icariin, INHIBITOR, PDE5} and {icariin, INHIBITOR, PDE4}). We selected the drug of “icariin” as the source entity to predict the corresponding target protein entities associated with the drug. We inserted the “<Arg>” and “</Arg>” tags in the front and end of the source entity “icariin” and added the entity type tags of “<Prot></Prot>” to the protein entities of “PDE5” and “PDE4”. Because the “PDE5” and “PDE4” participate the relation of “INHIBITOR” with the source entity “icariin”, their labels are “INHIBITOR”, and other tokens are “O”.

(2) Model training. Based on our tagging scheme, we first transformed the training data into the training instances. Then we selected several cutting-edge deep learning models for this task. Recently, transformer-based pre-trained models have shown promising results in a broad range of natural language processing (NLP) tasks and are widely used in the sequence labeling tasks (such as named entity recognition) (17). A large array of pre-trained models that are pre-trained on PubMed abstracts and PMC full-text articles are available in the biomedical domain. With minimal architectural modification, biomedical PLMs can be applied to various downstream biomedical text mining tasks and can significantly outperform previous state-of-the-art models in terms of biomedical NLP tasks (18). Given an input instance, we used biomedical PLMs as the encoder to represent the text. To optimize the performance, we added a fully connected layer with the ReLU activation function to finalize the hidden representation of the biomedical PLM for each token. Then, we used a softmax classification layer over the output label set to predict the label probability score of each token. In the experiments, we evaluated the same five biomedical PLMs (i.e., PubMedBERT, BioBERT, BioRoBERTa, BioM-ELECTRA, and BioM-ALBERT) in the text classification framework to the sequence labeling method. The training objective is to minimize the cross-entropy loss. In addition to the standard categorical cross-entropy loss, we also explored sample weights in the loss function for handling class imbalance. Here the samples for class C are weighted by the equation: W_C = log (total number of the samples/number of samples in class C). Our experimental results, however, show that the weighted loss does not achieve better performance than the standard loss.

(3) Model prediction. In the prediction phase, the input text was first split into sentences and tokenized using the Stanza tool (19). Only the sentences with both drug and protein entities were used to extract the relations. For each iteration, we sequentially selected a drug entity as the current source entity and assigned the tag pairs by the tagging scheme to distinguish the source drug from the other drug and protein entities. Then we use the trained model to tag the tokens of

the sentence to extract the corresponding target genes associated with the source drug. If a conflict of the relation types exists between the tokens of the target gene entity, the label of the first token of the target gene would be selected as the final relation type.

2.4 Model ensemble

To further optimize the performance, three ensemble alternatives (i.e., majority voting, voting with random search, and voting with backward search) were investigated to combine the results of the different models in our experiments. For majority voting, we selected the relation class that is predicted by more than half of all models. In addition, we searched backward and randomly to find a subset of our models that might achieve higher performance on the development set, rather than using all models. In the random search, we randomly generated a combination of our models each time, and we kept the best performance on the development set until the number of combinations reached our predefined value. In the backward search, we combined the results of all models. Then we iteratively removed the models to achieve higher performance until we found the combination of the models with the highest performance on the development set.

3 Results and discussion

3.1 Experimental setting

In our experiments, we downloaded five biomedical PLMs (i.e., PubMedBERT[1], BioBERT[2], BioRoBERTa[3], BioM-ELECTRA[4], and BioM-ALBERT[5]) and evaluated them in both frameworks. Note that, with the exception of PubMedBERT, which has only a base version, the other PLMs have base and large versions. The large PLMs were used as our final selection, as they achieved better performance than the base models in our experiments. For the parameter setting, we used PLMs with the default parameter settings and set the main hyper-parameters as follows: learning rate of 1e-5, batch size of 16, and last fully connected layer size of 128. The number of training epochs was chosen by the early stopping strategy (20) (50 epochs at most). Specially, we explored three methods: (1) DevES, whereby the training set is used for model training, and the development set is used for early stopping according to the overall F1-score; (2) ValES, for which we merged the official training and development sets, then randomly selected 350 abstracts as our validation set for early stopping according to the overall F1-score, and remain abstracts were used as the training set; and (3) TrainES, for which we merged the official training and development sets for model training, then chosen number of training epochs by early stopping according to the training loss score. Our models are implemented using the open-source deep learning libraries: HuggingFace's Transformer (21) and TensorFlow (22). The performance of the

[1] https://huggingface.co/microsoft/BiomedNLP-PubMedBERT-base-uncased-abstract
[2] https://huggingface.co/dmis-lab/biobert-large-cased-v1.1
[3] https://dl.fbaipublicfiles.com/biolm/RoBERTa-large-PM-M3-Voc-hf.tar.gz
[4] https://huggingface.co/sultan/BioM-ELECTRA-Large-Discriminator
[5] https://huggingface.co/sultan/BioM-ALBERT-xxlarge

models is evaluated on the test set using the official evaluation script with micro-averaged scores: precision (P), recall (R), and F1-score (F1).

3.2 Performance of individual models in sequence labeling framework

Table 2 presents the results of individual models with different PLMs and different labeling strategies in the sequence labeling framework on the test set. All models were trained using the standard categorical cross-entropy loss with the DevES way for fair comparisons.

Table 2. Performance of individual models in sequence labeling framework on the test set

PLM	D→P			P→D		
	P	R	F1	P	R	F1
PubmedBERT	0.759	0.787	0.772	0.748	0.776	0.762
BioBERT	**0.782**	0.748	0.765	0.759	0.780	0.770
BioRoBERTa	0.765	0.771	0.768	0.740	**0.805**	0.771
BioM-ELECTRA	0.753	**0.797**	**0.774**	0.761	0.799	**0.780**
BioM-ALBERT	0.751	0.791	0.770	**0.770**	0.788	0.779

A comparison of the two sequence labeling strategies (D→P and P→D) shows that most PLMs with P→D achieve better F1-scores than the ones with D→P with the exception of PubmedBERT. The average F1-score of P→D is slightly higher than that of D→P (0.772 vs. 0.770). Further, in a comparison of the different PLMs in the same sequence labeling strategy, BioM-ELECTRA achieves the highest F1-scores in the two strategies (0.774 and 0.780 with D→P and P→D, respectively). Other models also can obtain similar performance.

3.3 Comparison of sequence labeling and text classification frameworks

To explore the effectiveness of our sequence labeling framework, we comprehensively compared the performance of individual models in classification and sequence labeling frameworks on the test. All models are trained using the standard categorical cross-entropy loss using the DevES method. We selected the results of the model with a better F1-score in two sequence labeling strategies as the sequence labeling result. The performance comparison is shown in Figure 3.

The results of the models in the text classification framework show a similar trend to that of the models in the sequence labeling framework. BioM-ELECTRA achieves a slight better performance than the other PLMs in the text classification framework, with an F1-score of 0.773. Compared with the text classification framework, all models in our sequence labeling framework achieve better performance (an average improvement of 0.68% in F1-score). The main potential reason is that, compared to the text classification framework, our sequence labeling framework can learn more dependency information about the source entity and all corresponding target entities. The text classification framework ignores the dependency, as all entity pairs are classified independently.

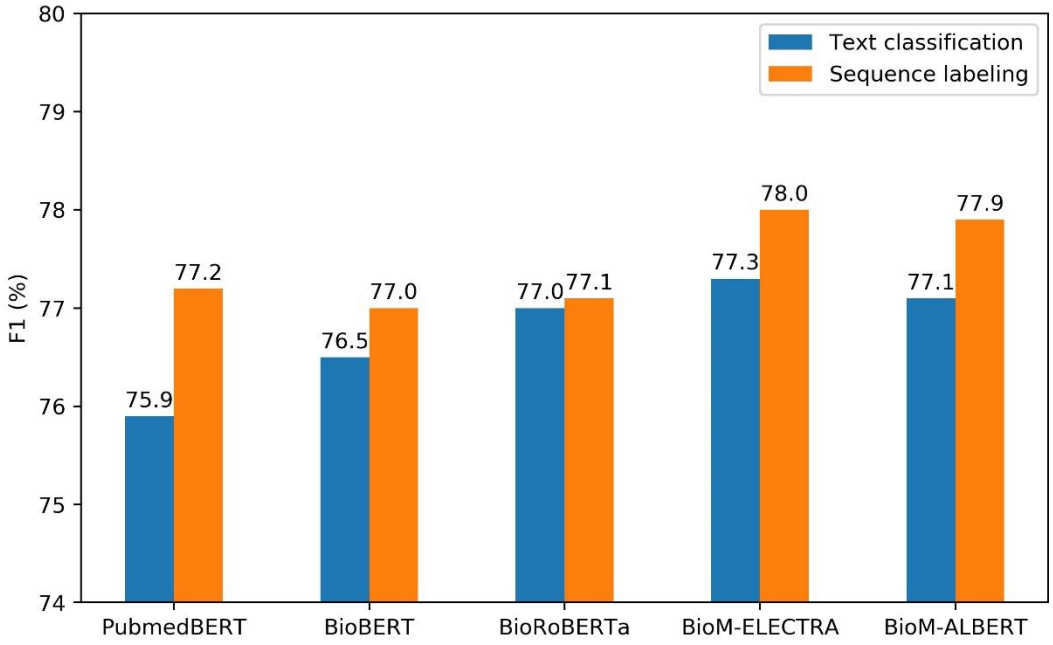

Figure 3. Performance comparison of the text classification and sequence labeling frameworks

Moreover, we recorded the prediction processing time of the PLMs in the different frameworks per 100 abstracts to compare the processing times of the two frameworks as shown in Table 3. All models are tested on the same GPU (Tesla V100-SXM2-32GB) and CPU (Intel(R) Xeon(R) Gold 6226 CPU @ 2.70GHz, 24Cores). The results show that our sequence labeling framework is more efficient than the text classification framework when tested on the GPU and CPU. The main reason is that the model in the text classification framework is required to process all the pairs between drug and protein entities one by one, but our sequence labeling framework can recognize all possible target entities associate with the source entity at once. For example, for a sentence with *N* drug entities and *M* protein entities, the relation extraction task is deconstructed into $N \times M$ sentence classification subtasks in a conventional classification framework. Our sequence labeling (D→P) can effectively narrow this down to *N* sequence labeling subtasks, and our sequence labeling (P→D) can narrow this down to M sequence labeling subtasks. As shown in Table 1, the numbers of drug and protein entities are approximate in the DrugProt corpus. Therefore, the processing time of the two sequence labeling strategies is similar and it is ~2 times faster than the text classification framework. This indicates that selecting the type of entities with a lower number of entities as the source entity is more efficient in our sequence labeling framework.

Table 3. Processing time comparison (Seconds per 100 abstracts)

Model	Framework	GPU	CPU
PubmedBERT	Text classification	32	289
	Sequence labeling (D→P)	22	123
	Sequence labeling (P→D)	25	145
BioBERT/BioRoBERTa/BioM-ELECTRA	Text classification	72	897
	Sequence labeling (D→P)	38	417
	Sequence labeling (P→D)	42	465
BioM-ALBERT	Text classification	301	2770
	Sequence labeling (D→P)	141	1245
	Sequence labeling (P→D)	162	1435

Note: The processing times of the BioBERT, BioRoBERTa, and BioM-ELECTRA are almost the same, as their model architectures and parameters are similar.

3.4 Performance of different model combinations

To further improve performance, we explored three alternatives (i.e., majority voting, voting with random search, and voting with backward search), as described in the "Model Ensemble" section, to combine the individual model results. During the DrugProt challenge, we submitted the following five runs as our final submissions:

- Run 1: The ensemble of all sequence labeling models, including four types (the combinations of the D→P and P→D with standard loss and weighted loss) for each PLM, trained by ValES via majority voting.
- Run 2: The ensemble of all sequence labeling models trained by DevES via majority voting.
- Run 3: The ensemble of the sequence labeling and text classification models trained by DevES via voting with backward search.
- Run 4: The ensemble of the sequence labeling and text classification models trained by DevES via voting with random search.
- Run 5: The ensemble of all sequence labeling and text classification models trained by DevES and ValES via majority voting.

During the DrugProt challenge, Run 5 (i.e., the ensemble of all models) achieves slightly better performance and obtains the highest overall F1-score on the test set in our five submissions. Comparing the different ensemble methods, although the voting with random search and backward search ensemble method achieves better performance on the development set, they cannot achieve better performance than the simple major voting on the test set. We found that, compared with DevES, the models with the early stopping strategy of ValES achieve worse performance. One possible reason is that the size of our validation set (350 abstracts) is small so that the performance of the models trained with the ValES is unstable. Therefore, at post-

challenge, we trained the models with TrainES instead of ValES to make full use of the development set. Moreover, we observed that the model trained using the weighted loss function cannot obtain a better performance than the model using the standard loss. Thus, we removed the model trained using the weighted loss and recombined the models with majority voting based on the following configurations:

- Run 6: The ensemble of all text classification models trained by DevES via majority voting.
- Run 7: The ensemble of all sequence labeling models trained by DevES via majority voting.
- Run 8: The ensemble of all text classification models trained by DevES and TrainES via majority voting.
- Run 9: The ensemble of all sequence labeling models trained by DevES and TrainES via majority voting.
- Run 10: The ensemble of all text classification and sequence labeling models with DevES and TrainES via majority voting.

Table 4. Performance of different model combinations on the test set

Method	Framework	Early stopping	Ensemble	P	R	F1
Single	BioM-ELECTRA in TC	DevES	-	0.774	0.773	0.773
Single	BioM-ELECTRA in SL (standard loss)	DevES	-	0.761	0.799	0.780
Run 1	SL (standard and weighted loss)	ValES	Majority voting	0.782	0.799	0.791
Run 2	SL (standard and weighted loss)	DevES	Majority voting	0.793	0.795	0.794
Run 3	SL (standard and weighted loss) + TC	DevES	Backward search	0.785	0.803	0.794
Run 4	SL (standard and weighted loss) + TC	DevES	Random search	0.790	0.798	0.794
Run 5	SL (standard and weighted loss) + TC	DevES + ValES	Majority voting	0.785	**0.805**	0.795
Run 6	TC	DevES	Majority voting	0.784	0.782	0.783
Run 7	SL (standard loss)	DevES	Majority voting	0.809	0.781	0.795
Run 8	TC	DevES + TrainES	Majority voting	0.801	0.775	0.788
Run 9	SL (standard loss)	DevES + TrainES	Majority voting	**0.802**	0.797	**0.800**
Run 10	SL (standard loss) + TC	DevES + TrainES	Majority voting	0.801	0.793	0.797

Note: SL denotes sequence labeling framework; TC denotes text classification framework.

From the results shown in Table 4, as compared with the best results of individual models, it is clear that the ensemble of models with majority voting (i.e., Runs 6 and 7) improves performance (corresponding improvements of 1.5% and 1.0% in sequence labeling and text

classification frameworks, respectively). When the models trained with TrainES are added into ensemble, the performance is further improved. Compared with the ensemble of text classification models, the ensemble of sequence label models achieves better performance. Finally, we found that Run 9 (i.e., the ensemble of only sequence labeling models trained by DevES and TrainES with standard loss) achieves the best performances in all runs. The ensemble of all models (Run 10) cannot achieve further improvement.

3.5 Performance comparison with other existing methods

To further demonstrate the effectiveness of our approach, we compare our methods with the top three teams' methods in the BioCreative VII DrugProt challenge. Table 5 shows the detailed granular results by relation type (F1-score for each relation type) and the overall results of the methods on the test set. Among these methods, both Humboldt (23) and KU-AZ (24) teams' methods are the ensemble of PLMs based on the text classification framework. The results of NLM-NCBI team (25) are the results of our Run 5 (i.e., our best submission during the challenge). Ours (single) is the best result of our single model (i.e., the BioM-ELECTRA (P->D) in the sequence labeling framework), and Ours (ensemble) is our best result of the ensemble at post-challenge (i.e., Run 9, the ensemble of only sequence labeling models with standard loss trained by DevES and TrainES).

Table 5. Detailed granular results on the test set

Relation Type	No. of Relations	Humboldt(23)	NLM-NCBI(25)	KU-AZ(24)	Ours (single)	Ours (ensemble)
INHIBITOR	6,538	**0.880**	0.867	0.878	0.861	0.876
DIRECT-REGULATOR	2,705	**0.729**	0.711	0.667	0.694	0.706
SUBSTRATE	2,497	0.682	0.722	0.683	0.711	**0.734**
ACTIVATOR	1,674	0.817	0.830	-	0.788	**0.832**
INDIRECT-UPREGULATOR	1,680	0.772	0.775	-	0.759	**0.781**
INDIRECT-DOWNREGULATOR	1,661	**0.794**	0.757	-	0.755	0.759
ANTAGONIST	1,190	0.915	**0.927**	-	0.912	0.924
PRODUCT-OF	1,078	**0.710**	0.683	-	0.663	0.682
PART-OF	1,142	0.755	**0.770**	-	0.745	0.769
AGONIST	789	0.821	0.843	-	0.819	**0.853**
AGONIST-ACTIVATOR	39	0.000	0.000	-	0.000	0.000
SUBSTRATE_PRODUCT-OF	27	0.000	0.000	-	**0.333**	0.000
AGONIST-INHIBITOR	15	0.000	**1.000**	**-**	**1.000**	**1.000**
Overall-Precision	-	0.796	0.785	0.797	0.761	**0.802**
Overall-Recall	-	0.799	**0.805**	0.782	0.799	0.797
Overall-F1	-	0.797	0.795	0.789	0.780	**0.800**

Note: No. of relations denotes the number of relations in the training and development merged set.

Both the Humboldt and KU-AZ teams explored the effect of the addition database (i.e., Comparative Toxicogenomics Database, CTD) for this task. The Humboldt team used the chemical definitions in the CTD chemical vocabulary to enrich the model input. Their experimental results show that the chemical descriptions can cause an improvement of 0.79% in the F1-score (23). The KU-AZ team selected the chemical-gene interactions dump from CTD to refine data augmentation. However, the overall performance of models pre-trained on augmented datasets is not better than the performance of models without the augmented dataset (24). Compared with the results of the single model, all ensemble methods are able to improve performance. Among these methods, our best ensemble of sequence labeling models achieves slightly better performance than the result of the top 1 team. Our method, without any additional datasets or post-processing, achieves an overall F1-score of 0.800. The detailed granular results by relation type show that the F1-score of our method on the ANTAGONIST relation is higher than 0.900, and F1-scores of most relation types with enough training data are over 0.750. Notably, relation types such as AGONIST-ACTIVATOR and SUBSTRATE_PRODUCT-OF seem to be more difficult to detect due to the small size of training data. Although PRODUCT-OF, DIRECT-REGULATOR, and SUBSTRATE have large training data support, they have relatively lower metrics. These relation types seem to be more complicated to be detected for the task.

Moreover, we participated in the additional DrugProt large-scale subtask. During the challenge, we did not use all models to predict the results, as some large PLMs are computationally expensive on the large-scale test set. Instead, we selected four efficient models (i.e., three sequence labeling models: PubMedBERT, BioM-ELECTRA, and BioRoBERTa; and a text classification model: PubMedBERT), and then used different combinations of them with simple majority voting to generate our submissions. Each model took ~5 days to predict the whole large-scale test set on one GPU. Finally, our best submission (i.e., the ensemble result of all models other than the PubMedBERT sequence labeling model) achieves the highest F1-score, 0.789, in all participating systems.

3.6 Error analysis

Although our sequence labeling method exhibits promising results for the DrugProt task, there is still room for improvement. To understand the causes of errors, we manually analyzed the error cases in which our ensemble of sequence labeling models with major voting on the development set, as the test gold standard annotation was not released. According to our error analysis, there are 604 false positives and 802 false negatives. Among these, there are 92 cases that correctly identify the association of the entity pair but assign the wrong relation type. More errors are caused by the fact that the gold standard relations cannot be identified from the text. We manually analyzed a random sample of the development set for errors. The errors are due mainly to several reasons. Table 1 shows that the relation type distributions are very different in different relation types. The model obtains lower performance on the relation types with small numbers, as it is difficult for the model to learn enough information from only a few dozen training instances. We also observed that pairs with a long distance between entities were more difficult to extract

correctly by our method. For example, in the last sentence of PMID: 10727715, there are 30 words between the drug of “pindolol” and the protein of “5-HT(1A)”. Our method does not extract the ANTAGONIST relation for them, as the long dependency information is difficult to capture. Moreover, some relations of entity pairs in the sentence are complicated, particularly because there are several relations in one sentence. For example, in PMID:9990013, there are three drug entities, three protein entities, and seven relations. Unfortunately, our method cannot always handle the cases well.

4 Conclusions

In this paper, we present our method, which is based on the pre-trained language models for the BioCreative VI DrugProt task. In addition to the classical text classification framework, we propose a novel sequence labeling framework to extract the relations of drugs and proteins. The experimental results show that our proposed framework is more efficient and is able to fully exploit the dependencies of relations for improved performance. Moreover, the different model combinations by different ensemble methods are further explored to optimize the final performance. The results show that the ensemble of only sequence labeling models with major voting achieves the best performance.

Our sequence labeling framework without any additional knowledge bases or post-processing exhibits promising results for the DrugProt task. In future work, we will investigate whether external resources (e.g., existing knowledge bases, dependency parser information) can be used to further improve our method. In addition, our sequence labeling framework can be easily adapted to other biomedical entity relation extractions. We will test the generalize ability of our framework to other biomedical relations such as drug-drug interactions.

Acknowledgements

The authors thank the organizers of the BioCreative VII DrugProt task. This research was supported by the Intramural Research Program of the National Library of Medicine (NLM), National Institutes of Health.

References

1. Davis A.P., Grondin C.J., Johnson R.J.*, et al*. (2021) Comparative toxicogenomics database (CTD): update 2021. *Nucleic Acids Res*, **49**, D1138-D1143.
2. Kringelum J., Kjaerulff S.K., Brunak S.*, et al*. (2016) ChemProt-3.0: a global chemical biology diseases mapping. *Database*, **2016**, bav123.
3. Lai P.-T., Lu Z. (2020) BERT-GT: cross-sentence n-ary relation extraction with BERT and Graph Transformer. *Bioinformatics*, **36**, 5678-5685.
4. Peng Y., Wei C.-H., Lu Z. (2016) Improving chemical disease relation extraction with rich features and weakly labeled data. *J Cheminform*, **8**, 1-12.

5. Krallinger M., Rabal O., Akhondi S.A.*, et al*. (2017) Overview of the BioCreative VI chemical-protein interaction Track. In: *Proceedings of the BioCreative VI challenge evaluation workshop.* Bethesda, MD USA, 141-146.
6. Miranda A., Mehryary F., Luoma J.*, et al*. (2021) Overview of DrugProt BioCreative VII track: quality evaluation and large scale text mining of drug-gene/protein relations. In: *Proceedings of the BioCreative VII challenge evaluation workshop.* Online, 11-21.
7. Peng Y., Lu Z. (2017) Deep learning for extracting protein-protein interactions from biomedical literature. In: *BioNLP 2017*, 29-38.
8. Peng Y., Rios A., Kavuluru R.*, et al*. (2018) Extracting chemical–protein relations with ensembles of SVM and deep learning models. *Database*, **2018**, bay073.
9. Devlin J., Chang M.-W., Lee K.*, et al*. (2019) BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers).* Minneapolis, Minnesota, Association for Computational Linguistics, 4171-4186.
10. Lee J., Yoon W., Kim S.*, et al*. (2020) BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, **36**, 1234-1240.
11. Gu Y., Tinn R., Cheng H.*, et al*. (2021) Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare*, **3**, 1-23.
12. Alrowili S., Vijay-Shanker K. (2021) BioM-Transformers: Building Large Biomedical Language Models with BERT, ALBERT and ELECTRA. In: *Proceedings of the 20th Workshop on Biomedical Language Processing.* Online, Association for Computational Linguistics, 221-227.
13. Agarap A.F. (2018) Deep learning using rectified linear units (relu). *arXiv preprint arXiv:08375*.
14. Lewis P., Ott M., Du J.*, et al*. (2020) Pretrained Language Models for Biomedical and Clinical Tasks: Understanding and Extending the State-of-the-Art. In: *Proceedings of the 3rd Clinical Natural Language Processing Workshop.* Online, Association for Computational Linguistics, 146-157.
15. Luo L., Yang Z., Cao M.*, et al*. (2020) A neural network-based joint learning approach for biomedical entity and relation extraction from biomedical literature. *J Biomed Inform*, **103**, 103384.
16. Li Z., Yang Z., Xiang Y.*, et al*. (2020) Exploiting sequence labeling framework to extract document-level relations from biomedical texts. *BMC Bioinformatics*, **21**, 1-14.
17. Qiu X., Sun T., Xu Y.*, et al*. (2020) Pre-trained models for natural language processing: A survey. *Science China Technological Sciences*, 1-26.
18. Peng Y., Yan S., Lu Z. (2019) Transfer Learning in Biomedical Natural Language Processing: An Evaluation of BERT and ELMo on Ten Benchmarking Datasets. In: *Proceedings of the 18th BioNLP Workshop and Shared Task.* Florence, Italy, Association for Computational Linguistics, 58-65.
19. Zhang Y., Zhang Y., Qi P.*, et al*. (2021) Biomedical and clinical English model packages for the Stanza Python NLP library. *J Am Med Inform Assoc*, **28**, 1892-1899.
20. Prechelt L. (1998) Automatic early stopping using cross validation: quantifying the criteria. *Neural Netw*, **11**, 761-767.
21. Wolf T., Chaumond J., Debut L.*, et al*. (2020) Transformers: State-of-the-art natural language processing. In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations.* Online, Association for Computational Linguistics, 38-45.

22. Abadi M., Barham P., Chen J.*, et al*. (2016) Tensorflow: A system for large-scale machine learning. In: *Proceedings of the 12th USENIX Symposium on Operating Systems Design and Implementation (OSDI '16).* Savannah, GA, USA, 265-283.
23. Weber L., Sänger M., Garda S.*, et al*. (2021) Humboldt@ DrugProt: Chemical-Protein Relation Extraction with Pretrained Transformers and Entity Descriptions. In: *Proceedings of the BioCreative VII challenge evaluation workshop.* Online, 22-25.
24. Yoon W., Yi S., Jackson R.*, et al*. (2021) Using Knowledge Base to Refine Data Augmentation for Biomedical Relation Extraction. In: *Proceedings of the BioCreative VII challenge evaluation workshop.* Online, 31-35.
25. Luo L., Lai P.-T., Wei C.-H.*, et al*. (2021) Extracting Drug-Protein Interaction using an Ensemble of Biomedical Pre-trained Language Models through Sequence Labeling and Text Classification Techniques. In: *Proceedings of the BioCreative VII challenge evaluation workshop.* Online, 26-30.