Extracting Drug-Protein Interaction using an Ensemble of Biomedical Pre-trained Language Models through Sequence Labeling and Text Classification Techniques

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 text mining the interactions between drugs and proteins is significantly beneficial to drug discovery, drug repurposing, drug design, and bioinformatics knowledge graph mining. The DrugProt track of BioCreative VII aims to promote the development and evaluation of systems that are able to automatically detect in relations between chemical compounds/drugs and genes/proteins. This paper describes our method used to create our submissions to the task. We formulated the task of extracting the relation pairs of drugs and proteins using two separate frameworks: text classification and sequence labeling. The cutting-edge biomedical pre-trained language models are used for both frameworks. Then different ensemble methods are further used to improve the final performance. Our best submission achieves the F1-scores of 0.795 and 0.789 on the main test set and the additional large-scale test set, respectively.**

Keywords*— *relation extraction; drug-protein interaction; pre-trained models; sequence labeling

I. INTRODUCTION

Extracting the relations between drug/chemical and protein/ gene from the exponentially growing biomedical literature is crucial in various biomedical tasks such as drug discovery, drug repurposing, drug-induced adverse reactions, and bioinformatics knowledge graph mining (1-3). Manually curating the interaction between drug and protein from the biomedical literature is extremely expensive and time-consuming. Alternatively, automatic text mining methods could detect these relations efficiently. To accelerate the method development of extracting the relations between drug and protein, the BioCreative VII organized the DrugProt track (similar to the previous CHEMPROT task (4) of BioCreative VI) for the drug-protein relation extraction task (5). As a participant of this task, we developed two independent deep learning-based approaches based on the biomedical pre-trained language models (PLMs) (6-11). Our best submission achieves F1-scores of 0.795 and 0.789 on the main test set and the additional large-scale test set, respectively.

II. METHODS

In this track, the official corpus includes 3,500 abstracts for training, 750 abstracts for development, and 10,750 abstracts for testing which contains a subset of a total of 750 Gold Standard abstracts that will be used for evaluation purposes. Additionally, a large set of 2,366,081 PubMed records with pre-annotations of entity mentions of drugs and proteins is provided as the large-scale test set for the additional DrugProt Large Scale task (5). According to the characteristic of the corpus, only less than 1% of the relations are crossing multiple sentences in the training set. Therefore, we focus on the challenges of the relation extraction within sentences. Specifically, we formulated the task of extracting the relation pairs of drugs and proteins using two different frameworks: text classification and sequence labeling.

A. *Text Classification Framework*

In the text classification framework, every drug-protein pair in a sentence is a target for prediction. The output of the classification to this target is to predict the predefined relation types (the 13 types of interactions: INDIRECT-DOWNREGULATOR, INDIRECT-UPREGULATOR, DIRECT-REGULATOR, ACTIVATOR, INHIBITOR, AGONIST, ANTAGONIST, AGONIST-ACTIVATOR, AGONIST-INHIBITOR, PRODUCT-OF, SUBSTRATE, SUBSTRATE_PRODUCT-OF and PART-OF) or not a relation at all. In the other words, our system treated this task as a multi-class classification problem, which represents a sentence with a drug/protein pair using two different sequence representation ways. The first representation is to insert tags of "@DRUG$" and "@PROT$" in front of the drug and protein entities. We also treated the name text of drug and protein entity as the first sentence and use the [SEP] tag to concatenate it with the first way's sequence as the second representation. Since keeping the original entities in the sentences brings higher performance, our method does not replace the drug and protein entities to the tags. The examples in Table 1 illustrate the above two ways of text representation. Finally, we used the

TABLE I. EXAMPLES OF TEXT CLASSIFICATION

Entity Pair	Relation Type	Text Representation 1	Text Representation 2
icariin-PDE5	INHIBITOR	[CLS] The inhibitory effects of @DRUG$ icariin on @PROT$ PDE5 and PDE4 activities were investigated by the two-step radioisotope procedure with [(3)H]-cGMP/[(3)H]-cAMP.	[CLS] icariin and PDE5 [SEP] The inhibitory effects of @DRUG$ icariin on @PROT$ PDE5 and PDE4 activities were investigated by the two-step radioisotope procedure with [(3)H]-cGMP/[(3)H]-cAMP.
icariin-PDE4	INHIBITOR	[CLS] The inhibitory effects of @DRUG$ icariin on PDE5 and @PROT$ PDE4 activities were investigated by the two-step radioisotope procedure with [(3)H]-cGMP/[(3)H]-cAMP.	[CLS] icariin and PDE5 [SEP] The inhibitory effects of @DRUG$ icariin on PDE5 and @PROT$ PDE4 activities were investigated by the two-step radioisotope procedure with [(3)H]-cGMP/[(3)H]-cAMP.

first token [CLS] to represent the output of the whole sequence for the classification task. A softmax layer is connected at the end for the prediction of the relation type.

Our text classification models are formed by incorporating Biomedical PLMs with a softmax output layer. To select the biomedical PLMs with the best performance, we tried PubMedBERT (7), BioBERT (8), and BioELECTRA (11). Besides, both BioBERT and BioELECTRA have large versions of the pre-trained model. After testing those models, we chose PubMedBERT for our final submissions, which achieves the best performance (>77%) on the development set. We keep both text representations for final submission since the performances are similar.

B. *Sequence Labeling Framework*

Inspired by our previous works (12, 13), we proposed a novel sequence labeling framework to address the sentence-level biomedical relation extraction task. Different from the conventional text classification framework, the task is converted to a sequence labeling problem. Given a candidate source entity (e.g., drug entity of "icariin" in Fig.1) in a 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 candidate. For a sentence with *N* source entities and *M* target entities, the entire task can be deconstructed into *N* × *M* independent sentence classification subtasks. But our method can effectively narrow down to *N* sequence labeling subtasks. Besides, the sequence labeling framework is able to fully exploit the dependencies of source entities and relations. More details will be described below.

1) **Tagging Scheme**. Fig. 1 shows an example to tag a sentence with our tagging scheme according to the original gold standard annotations of the DrugProt dataset. To define the boundary of the entities, the "<Arg1>" and "</Arg1>" tags are inserted in the start and end of the candidate source entity. Besides, "<Drug>"/"<Prot>" and "</Drug>"/"</Prot>" tags are inserted in the start and end of the drug/protein mentions to aware the entity types and the boundaries. Each token is assigned a label that contributes to the extraction. The tokens can be divided into two types: (I) the target entities involve in the relations; (II) others. Concretely, the labels of type I consist of 13 relation types that are predefined according to the training sets. We used the label "O" to represent other tokens and entities which do not involve in a relation. As shown in Fig. 1, the input sentence contains three entities (i.e., the drug entity "icariin" and the protein entities "PDE5" and "PDE4") and two drug-protein relation triples (i.e., {icariin, INHIBITOR, PDE5} and {icariin, INHIBITOR, PDE4}). In the example, we set up the source entity to drug to predict the target entity of protein. We inserted the "<Arg1>" and "</Arg1>" tags in the start and end of the candidate source entity "icariin", and added the entity type tags of "<Prot></Prot>" to the protein entities of "PDE5" and "PDE4". Since the " PDE5" and "PDE4" participate the relation "inhibitor" with the candidate source entity "icariin", their labels are " INHIBITOR", and other tokens are "O".

2) **Model Architecture**. 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 field of NLP (14). A large array of pre-trained

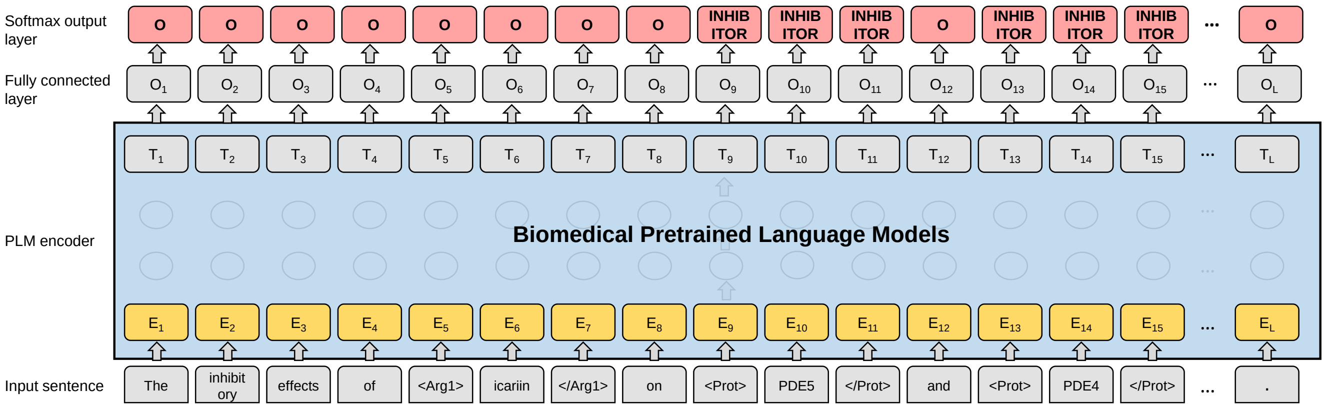
Softmax output layer
Fully connected layer
PLM encoder
Input sentence
O O O O O O O O INHIBITOR INHIBITOR INHIBITOR O INHIBITOR INHIBITOR INHIBITOR ... O
O1 O2 O3 O4 O5 O6 O7 O8 O9 O10 O11 O12 O13 O14 O15 ... OL
T1 T2 T3 T4 T5 T6 T7 T8 T9 T10 T11 T12 T13 T14 T15 ... TL
Biomedical Pretrained Language Models
E1 E2 E3 E4 E5 E6 E7 E8 E9 E10 E11 E12 E13 E14 E15 ... EL
The inhibitory effects of <Arg1> icariin </Arg1> on <Prot> PDE5 </Prot> and <Prot> PDE4 </Prot> ... .

Fig. 1. The overview of our sequence labeling framework

models which pre-trained on PubMed abstracts and PMC full-text articles are available to the biomedical domain. With minimal architectural modification, Biomedical PLMs can be applied to various downstream biomedical text mining tasks and significantly outperforms previous state-of-the-art models to the biomedical NLP tasks (6). The architecture of our model is illustrated in Fig. 1. To optimize the performance, we feed the final hidden representation of the Biomedical PLM for each token into a fully connected layer with ReLU (15) activation function. Then, we use a softmax classification layer over the output label set to predict the label probability score of each token. Similar to the text classification framework, we evaluated the five biomedical PLMs including PubMedBERT (7), BioBERT (8), BioRoBERTa (9), BioM-ELECTRA (10), and BioM-ALBERT (10) to the sequence labeling method on the development set.

3) **Relation Extraction**. During the model development, we can set up the source entity to drug (or protein) to predict the target entity of protein (or drug). In addition to the standard categorical cross-entropy loss function, we also applied 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*). Therefore, we trained four models (the combinations of the "drug to protein" and "protein to drug" with standard loss and weighted loss) for each kind of PLM. In the test phase, the input text is split into sentences and tokenized. The sentences with both drug and protein entities are tagged by our trained models. If there is a relation type conflict to the tokens of the entity, the label of the first token of the target entity is chosen to be the relation type.

C. Model Ensemble

For each individual model, we tuned the hyper-parameters on the development set by random search (16). Our models are implemented using the open-source deep learning libraries Hugging Face (17) and TensorFlow (18). To further optimize the performance, three ensemble alternatives (majority voting, voting with random search, and voting with backward search) are investigated in our experiments. For the majority voting, we select the relations that are predicted by more than half of all models. In addition, we search backward and random to find a subset of our approaches that might achieve higher performance on the development set than using all models. In random search, we randomly generate a combination of our models every time, and we keep the best performance on the development set until the number of combinations reaches our predefined value. In backward search, we first combine the results of all models, then remove a model which can bring higher performance. We iteratively removed the models until we found the combination of the models with the highest performance on the development set.

III. Results and Discussion

During the DrugProt task, we submitted five runs as our final submissions. Our submitted five runs in the main task are based on the following configurations.

- Run 1: we merged the official training and development sets, then randomly selected 350 abstracts as our development set for early stopping strategy (19) and remain articles were used as the training set. Only sequence labeling models are ensembled with the majority voting.
- Run 2: we used the official training set only for model training, and the number of training epochs is chosen by early stopping strategy according to the performance on the development set. Only sequence labeling models are ensembled with simple majority voting.
- Run 3: In the sequence labeling framework, we first used the data augmentation technologies (including synonym substitution, swap word randomly, back translation) to increase the number of the lower resource relation types (i.e., AGONIST-ACTIVATOR, AGONIST-INHIBITOR, SUBSTRATE_PRODUCT-OF), then trained the sequence labeling models. After that, both the text classification and sequence labeling models are ensembled by voting with backward search.
- Run 4: the text classification and sequence labeling models without data augmentation are ensembled by voting with random search.
- Run 5: all our text classification and sequence labeling models are ensembled together by majority voting.

Table 2 shows the overall results (overall precision, recall, and F1 score) of our runs on the official development and main test sets. Table 3 shows the detailed granular results by relation type (F1-score for each relation type) and overall results of our submissions on the test sets as reported by the organizer. Run 5 (i.e., the ensemble of all models) achieves the highest overall F1 score on the main test set.

TABLE II. Overall Results on the Development and Test Sets

	The development set				The main test set		
	P	**R**	**F1**		**P**	**R**	**F1**
Run1	-	-	-		0.782	0.799	0.791
Run2	0.813	0.811	0.812		**0.793**	0.795	0.794
Run3	0.811	**0.825**	0.818		0.785	0.803	0.794
Run4	**0.819**	0.819	**0.819**		0.790	0.798	0.794
Run5	-	-	-		0.785	**0.805**	**0.795**

Note that, Run 1 and 5 use the development set for training, so we do not evaluate their performance on the development set.

For the additional DrugProt large-scale task, we did not use all models to predict the results, since some large PLMs are computationally expensive on the large-scale test set. Instead, we selected four efficient models (i.e., PubMedBERT sequence labeling model from protein to drug with standard loss, BioM-ELECTRA sequence labeling model from drug to protein with weighted loss, BioRoBERTa sequence labeling model from drug to protein with weighted loss, and PubMedBERT text classification model with text representation 2) according to the performances on the development set, then used different combinations of them with the simple majority voting to generate our submissions. Each model took ~5 days for predicting the whole large-scale test set on one NVIDIA Tesla

TABLE III. DETAILED GRANULAR RESULTS FOR OUR ENSEMBLE SYSTEM ON THE TEST SETS

	The main test set						**The large-scale test set**				
	Run1	**Run2**	**Run3**	**Run4**	**Run5**		**Run1**	**Run2**	**Run3**	**Run4**	**Run5**
ACTIVATOR	**0.830**	0.826	0.827	0.815	**0.830**		0.814	**0.816**	0.803	0.804	0.802
AGONIST	0.808	**0.851**	0.837	0.837	0.843		0.848	**0.856**	0.822	0.882	0.853
AGONIST-ACTIVATOR	0.000	0.000	0.000	0.000	0.000		0.000	0.000	0.000	0.000	0.000
AGONIST-INHIBITOR	1.000	1.000	1.000	1.000	1.000		1.000	1.000	1.000	1.000	1.000
ANTAGONIST	**0.930**	0.920	0.921	0.920	0.927		0.911	0.914	**0.918**	0.913	**0.918**
DIRECT-REGULATOR	0.712	0.700	0.708	**0.714**	0.711		0.700	**0.719**	0.707	0.705	0.716
INDIRECT-DOWNREGULATOR	0.759	0.759	**0.763**	0.758	0.757		0.759	**0.776**	0.772	0.751	0.751
INDIRECT-UPREGULATOR	0.765	0.767	0.774	**0.777**	0.775		0.764	**0.779**	0.755	0.769	0.765
INHIBITOR	0.867	0.873	0.867	**0.874**	0.867		0.865	0.867	**0.871**	0.866	0.867
PART-OF	0.766	**0.777**	0.760	0.763	0.770		**0.746**	0.739	0.734	0.724	0.743
PRODUCT-OF	0.680	0.684	**0.688**	0.675	0.683		**0.660**	0.645	0.648	0.654	0.649
SUBSTRATE	0.699	0.712	**0.727**	0.717	0.722		0.688	0.698	0.699	0.701	**0.708**
SUBSTRATE_PRODUCT-OF	0.000	0.000	0.000	0.000	0.000		0.000	0.000	0.000	0.000	0.000
Overall-Precision	0.782	**0.793**	0.785	0.790	0.785		**0.778**	0.773	0.775	0.768	0.748
Overall-Recall	0.799	0.803	0.803	0.798	**0.805**		0.789	0.805	0.796	0.799	**0.826**
Overall-F1	0.791	0.794	0.794	0.794	**0.795**		0.784	**0.789**	0.785	0.783	0.785

V100 SXM2 GPU. Our submitted five runs for this task are based on the following configurations.

- Run 1: the ensemble result of all models other than the BioM-ELECTRA sequence labeling model.
- Run 2: the ensemble result of all models other than the PubMedBERT sequence labeling model.
- Run 3: the ensemble result of all models other than the BioRoBERTa sequence labeling model.
- Run 4: the ensemble result of all models other than the PubMedBERT text classification model.
- Run 5: ensemble result of all four models.

Similar results are observed on the large-scale test set, and the best submission achieves an F1-score of 0.785.

IV. CONCLUSION

In this paper, we present our method based on the pre-trained language models in the BioCreative VI DrugProt task. In addition to the classic text classification framework, we propose a novel sequence labeling framework to extract the relations of drugs and proteins. Then different ensemble methods are further used to optimize the final performance. The results show that our method can effectively extract the drug-protein relations from biomedical literature.

ACKNOWLEDGMENT

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

REFERENCES

1. Peng Y, Wei C-H, Lu Z. (2016) Improving chemical disease relation extraction with rich features and weakly labeled data. *Journal of cheminformatics*, **8**(1):1-12.

2. Singhal A, Simmons M, Lu Z. (2016) Text mining for precision medicine: automating disease-mutation relationship extraction from biomedical literature. *Journal of the American Medical Informatics Association*, **23**(4):766-772.

3. Lai P-T, Lu Z. (2020) BERT-GT: cross-sentence n-ary relation extraction with BERT and Graph Transformer. *Bioinformatics*, **36**(24):5678-5685.

4. Krallinger M, Rabal O, Akhondi SA, Pérez MP, Santamaría J, Rodríguez GP, Tsatsaronis G, Intxaurrondo A. (2017) Overview of the BioCreative VI chemical-protein interaction Track. In: *Proceedings of the sixth BioCreative challenge evaluation workshop*, 141-146.

5. Miranda A, Mehryary F, Luoma J, Pyysalo S, Valencia A, Krallinger M. (2021) Overview of DrugProt BioCreative VII track: quality evaluation and large scale text mining of drug-gene/protein relations. In: *Proceedings of the seventh BioCreative challenge evaluation workshop*: Citeseer.

6. 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*, 58-65.

7. Gu Y, Tinn R, Cheng H, Lucas M, Usuyama N, Liu X, Naumann T, Gao J, Poon H. (2020) Domain-specific language model pretraining for biomedical natural language processing. *arXiv preprint arXiv:15779*.

8. Lee J, Yoon W, Kim S, Kim D, Kim S, So CH, Kang J. (2020) BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, **36**(4):1234-1240.

9. Lewis P, Ott M, Du J, Stoyanov V. (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*, 146-157.

10. 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*, 221-227.

11. raj Kanakarajan K, Kundumani B, Sankarasubbu M. (2021) BioELECTRA: Pretrained Biomedical text Encoder using Discriminators. In: *Proceedings of the 20th Workshop on Biomedical Language Processing*, 143-154.

12. Li Z, Yang Z, Xiang Y, Luo L, Sun Y, Lin H. (2020) Exploiting sequence labeling framework to extract document-level relations from biomedical texts. *BMC bioinformatics*, **21**(1):1-14.

13. Luo L, Yang Z, Cao M, Wang L, Zhang Y, Lin H. (2020) A neural network-based joint learning approach for biomedical entity and relation extraction from biomedical literature. *Journal of biomedical informatics*, **103**:103384.

14. Qiu X, Sun T, Xu Y, Shao Y, Dai N, Huang X. (2020) Pre-trained models for natural language processing: A survey. *Science China Technological Sciences*:1-26.

15. Agarap AFJapa. (2018) Deep learning using rectified linear units (relu).

16. Bergstra J, Bengio Y. (2012) Random search for hyper-parameter optimization. *The Journal of Machine Learning Research*, **13**(1):281-305.

17. Wolf T, Chaumond J, Debut L, Sanh V, Delangue C, Moi A, Cistac P, Funtowicz M, Davison J, Shleifer S. (2020) Transformers: State-of-the-art natural language processing. In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 38-45.

18. Abadi M, Barham P, Chen J, Chen Z, Davis A, Dean J, Devin M, Ghemawat S, Irving G, Isard M. (2016) Tensorflow: A system for large-scale machine learning. In: *12th {USENIX} symposium on operating systems design and implementation ({OSDI} 16)*, 265-283.

19. Prechelt L. (1998) Automatic early stopping using cross validation: quantifying the criteria. *Neural Networks*, **11**(4):761-767.