Title

Extracting Protein-Protein Interactions (PPIs) from Biomedical Literature using Attention-based Relational Context Information

ISBN

9781665480451

Authors

Park, Gilchan
McCorkle, Sean
Soto, Carlos
et al.

Publication Date

2022-12-20

DOI

10.1109/bigdata55660.2022.10021099

Extracting Protein-Protein Interactions (PPIs) from Biomedical Literature using Attention-based Relational Context Information

Gilchan Park*
Computational Science Initiative
Brookhaven National Laboratory
Upton, New York, USA
gpark@bnl.gov

Sean McCorkle
Computational Science Initiative
Brookhaven National Laboratory
Upton, New York, USA
mccorkle@bnl.gov

Carlos Soto
Computational Science Initiative
Brookhaven National Laboratory
Upton, New York, USA
csoto@bnl.gov

Ian Blaby
Joint Genome Institute
Lawrence Berkeley National Laboratory
Berkeley, California, USA
ikblaby@lbl.gov

Shinjae Yoo
Computational Science Initiative
Brookhaven National Laboratory
Upton, New York, USA
sjyoo@bnl.gov

2022 IEEE International Conference on Big Data (Big Data) | 978-1-6654-8045-1/22/$31.00 ©2022 IEEE | DOI: 10.1109/BigData55660.2022.10021099

***Abstract*—Because protein-protein interactions (PPIs) are crucial to understand living systems, harvesting these data is essential to probe disease development and discern gene/protein functions and biological processes. Some curated datasets contain PPI data derived from the literature and other sources (e.g., IntAct, BioGrid, DIP, and HPRD). However, they are far from exhaustive, and their maintenance is a labor-intensive process. On the other hand, machine learning methods to automate PPI knowledge extraction from the scientific literature have been limited by a shortage of appropriate annotated data. This work presents a unified, multi-source PPI corpora with vetted interaction definitions augmented by binary interaction type labels and a Transformer-based deep learning method that exploits entities' relational context information for relation representation to improve relation classification performance. The model's performance is evaluated on four widely studied biomedical relation extraction datasets, as well as this work's target PPI datasets, to observe the effectiveness of the representation to relation extraction tasks in various data. Results show the model outperforms prior state-of-the-art models. The code and data are available at:** https://github.com/BNLNLP/PPI-Relation-Extraction

***Index Terms*—protein–protein interactions, PPIs, relation extraction, RE, biomedical literature, attention, relation representation**

I. Introduction

Much effort in modern molecular biology either involves or is entirely focused on learning and understanding the functions and interactions of the millions of proteins that compose the basic building blocks of life. In particular, the prediction of protein structure and functions has been recognized as a paramount phase in some major issues of life science, such as the therapeutic approach for several diseases, which can ameliorate healthcare by accelerating drug discovery and development. The functions of most proteins currently are unknown with only a small fraction definitively established after extensive and labor-intensive lab work has been performed. These gold-standard protein function assignments have been extended computationally via DNA and amino acid sequence homology throughout the ever-expanding collection of protein sequences determined from genome sequencing. However, inference from homology often is inaccurate. Helpfully, clues about function can come from other sources, including interactions with proteins for which the function is known. While experiments that definitively determine interactions can be labor-intensive, several relatively high-throughput methods are in use, such as two-hybrid screening [1] and affinity purification followed by mass spectrometry [2]. Numerous databases, such as IntAct[1], STRING[2], DIP[3], BioGrid[4], HPRD[5], and MINT[6] are now dedicated to collecting and curating protein-protein interaction (PPI) results obtained using various techniques and from the scientific literature. Unfortunately, mining the literature requires manual effort and is slow. To remedy this, we aim to develop a machine learning (ML) model that effectively identifies statements of PPIs in scientific text.

Efforts to fully automate text knowledge extraction are widespread and ongoing with supervised learning approaches currently being the most favored. A key challenge in applying these methods to PPI extraction is a shortage of training

* Corresponding author.
The materials presented in this paper are based upon the work supported by the U.S. Department of Energy, Office of Science, Office of Biological and Environmental Research, DE-SC0012704.

[1] https://www.ebi.ac.uk/intact
[2] https://string-db.org
[3] https://dip.doe-mbi.ucla.edu/dip
[4] https://thebiogrid.org
[5] https://www.hprd.org
[6] https://mint.bio.uniroma2.it

data specifically annotated for this purpose. Several publicly available PPI training datasets suffer from biases of restricted biological focus (i.e., human-, medical-, or microbial-only) and also differences in the concept of what *defines* an interaction. For this work, we combine all of the aforementioned training sets, vet them for uniformity in interaction definition, and add interaction type labels. We also propose Transformer architecture-based models [3], which leverage entities' relational context information to build a relation representation that improves relation classification performances.

As detailed in this paper, our contribution is twofold:

1) We augment public PPI corpora with labels for protein types (*enzyme* and *structural*), which further delineate the functional role of proteins and consequently afford a helpful protein classification for the biology community. We also provide the interaction-typed PPI corpora for the community.
2) We present a Transformer-based relation prediction method that exploits entities' relational context information to build an improved relation representation. Our study shows the effectiveness of the proposed approach not only on the PPI datasets, but also four biomedical relation extraction datasets.

II. Related Work

There have been ongoing efforts to consolidate biological knowledge pertinent to PPIs from literature by creating machine-processable data and designing protein relation extraction methods.

A. PPI corpora

BioCreative VI [4] proposed a PPI relation extraction challenge task related to genetic mutations to foster the development of mining PPI information from biomedical literature. Bunescu et al. [5] annotated 1000 titles and abstracts from the MEDLINE repository that discuss human genes/proteins, the so-called AIMed corpus, which includes roughly 5000 protein names and 1000 protein interactions. Pyysalo et al. [6] created BioInfer (Bio Information Extraction Resource), containing 1100 sentences with named entities and their relationships tagged from abstracts of biomedical research articles. Fundel, Küffner, and Zimmer [7] tagged the sentences of 50 abstracts referenced by the Human Protein Reference Database (HPRD) with direct physical interactions, regulatory relations, and modifications between genes/proteins. The IEPA (Information Extraction Processing Assessment) corpus [8] was created to conduct a comparative study on the merits of different text processing units for interactions between biochemical entities. The Learning Language in Logic Workshop (LLL05) [9] designed the genic interaction extraction challenge task that aims to promote protein/gene interactions information extraction from biology abstracts in the MEDLINE bibliography database. The LLL challenge focused on gene interactions in *Bacillus subtilis*, a model bacterium, and many papers have been published about direct gene interactions involved in sporulation.

Although the number of corpora and methods for PPI information extraction from biomedical text has increased as the interest in automatic mining systems has grown, the lack of consensus with respect to PPI annotation has hindered consolidation of heterogeneous datasets, thereby making it difficult for researchers to properly evaluate their methods on a standardized dataset for PPI extraction. Pyysalo et al. [10] have conducted a comparative analysis of the five PPI datasets—AIMed, BioInfer, HPRD50, IEPA, and LLL—and unified the PPI annotations to share with the community for clear and comparative method evaluation. To merge these diverse datasets, Pyysalo et al. [10] have found common categories across the five corpora and generated a unified PPI corpora composed of sentences tagged with undirected and untyped binary interactions (i.e., positive and negative). These unified versions of PPI datasets, hereafter called the *five benchmark PPI corpora*, have been widely used to evaluate various approaches on PPI extraction tasks [11]–[13]. In the biological literature, single sentences often discuss more than two proteins, and such statements are not all declarations of interactions between the proteins mentioned. These datasets include all identified protein/gene entity names found within each training sentence, as well as a pairwise evaluation of positive/negative interactions between each possible pairing.

However, some issues remain regarding the content and annotations in these benchmark PPI datasets (detailed in Section III-A). In this paper, we present an augmented, refined version of the five benchmark PPI corpora along with the BioCreative VI corpus that further specify positive interactions into two types of interactions: *enzyme* and *structural*. These interaction types are desirable to construct protein interaction networks.

B. PPI extraction methods

In the early stages of adopting ML approaches for the PPI extraction task, feature- and kernel-based approaches have been commonly used [12], [14]. In an attempt to capture syntactic and semantic information of sentences, Murugesan, Abdulkadhar, and Natarajan [15] developed a Distributed Smoothed Tree kernel (DSTK) composed of distributed lexical parse trees and semantic feature vectors and demonstrated that the shallow linguistic information helped enhance the PPI extraction capability with the model evaluation on the five benchmark PPI corpora.

With the recent success of deep learning in a number of applications, deep neural network models have emerged to tackle the PPI extraction task. Peng and Lu [16] have demonstrated their multichannel dependency-based convolutional neural network model (McDepCNN) effectively captures syntactic features of sentences by adding a separate channel for the dependency information of the sentence syntactic structure on the PPI task using AIMed and Bioinfer corpora. Attention mechanisms in natural language processing (NLP) have shed some light on solving long dependency issues between tokens in sequential data. The self-attention-based Transformer architecture [3] has proven to well preserve long-term dependencies

and establish effective contextual representations. NLP models built upon Transformer architecture, such as BERT [17], have achieved state-of-the-art (SOTA) results in various NLP tasks, including in biology domains [18]. Warikoo, Chang, and Hsu [13] have proposed a Lexically aware BERT model (LBERT) that generates syntactic contexts emphasized representations for sentence-level bio-entity relation extraction tasks taking *n*-gram parts-of-speech frames as an additional input embedding to deliver latent lexical properties, and the model outperformed the prior models on a PPI task with the five benchmark PPI corpora. Recently, Tang et al. [19] have built a PPI extraction model based on a domain-specifically pre-trained BERT and adversarial training, which showed significant improvement on the classification of the five benchmark PPI corpora.

III. Additional PPI Curation

This section details the further curation and enhancement of the aforementioned datasets.

A. Problems discovered during curation

In vetting the five benchmark PPI training corpora, we identified the following problems:

1) Bias due to restricted biological focus for each set: In particular, the AIMed and IEPA corpora are focused on human medical biochemistry and phenomena, including viral pathogens, whereas the set LLL is limited to a single bacterial species, *Bacillus subtilis*. These differences manifest in skew and distribution of protein/gene name frequency counts between the five sets, as well as other domain-specific terminology. In fact, the most frequently occurring protein in IEPA, *insulin*, accounts for 14% of the protein mentions in all of the IEPA positives, yet it does not occur in the AIMed positives set, where the most common protein, *p53*, accounts for only 1.75% of the protein names. These sets all sampled especially different populations in the literature. Combining all sets together helps to counter this bias, but, in the future, we plan to collect more training data to better address this issue.

2) Differences in notion of the definition of an interaction: The five sets largely restrict PPI-positive cases to clear statements of direct interaction between the two subjects. LLL further restricts positive PPI declarations to cases where a protein binds to DNA and causes or inhibits the transcription of the gene of another protein, or a statement of gene regulation—a markedly particular type of interaction.

We intentionally broaden our acceptance of a positive PPI indication. Our goal is to provide biologists with a tool to identify possible interactive connections between proteins directly from the scientific literature text. Because of the likelihood that claims of direct PPI will end up in future databases (if not there already), a less restrictive interpretation will allow a text mining system to report results of value that will not necessarily be found in a PPI database.

Along these lines, we did not distinguish between gene or protein for this work. In addition to direct binding between two proteins or a protein and itself (i.e., *dimers* and *multimers*), we also consider interacting cases where two proteins bound to a larger complex of other proteins without necessarily contacting each other directly.

The following details an example (from the BioCreative corpus) where a direct connection between proteins *PVA12* and *ORP3a* is made but is not declared an actual interaction.

> *The targeting of the oxysterol-binding protein* ***ORP3a*** *to the endoplasmic reticulum relies on the plant VAP33 homolog* ***PVA12****.*

On the other hand, we are mindful of the possibility of being too broad, which would result in too many PPI calls to be meaningful.

3) Confusion over PPI-negative annotations: This expanded threshold for PPI-positive impacts the public negative annotations. The following are two example cases (from AIMed corpus) where we disagree with the given negative labels.

> *In addition to this unique pathway,* ***FGFR3*** *also links to* ***GRB2****.*

A negative interaction between proteins *FGFR3* and *GRB2* was declared in the public set.

> *After a brief historical incursion regarding renal artery stenosis (RAS) of renal origin, we present the main extrarenal* ***angiotensin****-forming enzymes, starting with* ***isorenin****,* ***tonin****, and D and* ***G cathepsin*** *and ending with the conversion enzyme and* ***chymase****.*

In this case, negative interactions are annotated between *angiotensin* and each of *isorenin*, *tonin*, *G cathepsin*, and *chymase*, respectively, even though they are declared as forming angiotensin.

The following shows an example of a negative PPI sentence where we agree with the given label and have included in our curated set (from AIMed corpus).

> *The molar ratio of serum* ***retinol-binding protein (RBP)*** *to* ***transthyretin (TTR)*** *is not useful to assess vitamin A status during infection in hospitalized children.*

To reduce confusion in our initial models regarding updated positive and negative relabels, we consider only those negatively labeled sentences where no positive pairs were declared in a sentence. Then, we manually examine each case to make sure we agree, disregarding (for now) those where we differ. For the same reason, in this work, we also disregard negative pair cases in sentences with both positive and negative annotations.

B. Interaction Type Annotation

PPIs aid with biological engineering. Notably, structure and protein subunit complex knowledge is critical to protein engineering, and transient interactions (e.g., chaperone to client protein) knowledge is needed for engineering at a broader scale. To make the public PPI corpora more useful for this purpose, we have added interaction type labels for the positively defined pairs in the unified datasets and the BioCreative set. In determining the interaction type labels, we first considered top-level protein function categories from IntAct's molecular interaction ontology but discovered we lacked enough training examples to provide sufficient statistics in each of the 28 categories to properly train a model (not all interaction types occur with equal frequency). We then tried to reduce the number of categories by making them coarser, first lowering to roughly 10 then three types. However, we found that making assignments in this manner proves too complicated with only questionable scientific value.

We finally decided on a simple binary classification with interactions being declared either *enzyme* or *structural* for our first pass because *enzyme* or *structural* accurately delineates the functional role of almost all proteins and consequently provides a concise but meaningful protein classification. The *structural* label is applied to protein assemblages of large, permanent cellular components, such as cell walls, histones, golgi apparatus, microtubules, membranes, and inter-cellular structures. All other interactions are classified as *enzyme*. Type is determined by examining the given function for each protein/gene, where it can be obtained from any of several online protein databases, such as Uniprot, NCBI, and GeneCards, and from the sentence context itself. For the five sentence-based datasets, interaction type labels are applied for positively identified protein pairs. An example of a structural interaction label for the proteins *alpha-syntrophin* and *utrophin* (from BioInfer corpus) follows:

> *Absence of* ***alpha-syntrophin*** *leads to structurally aberrant neuromuscular synapses deficient in* ***utrophin****.*

The remaining non-structural interactions are considered *enzymatic*, a label applied to nominal enzyme activity (proteins that catalyze chemical reactions of metabolites in reaction pathways) and proteins that activate other proteins (*kinases*). In this work, we also applied said label to all proteins that activated, inhibited, signaled, and formed temporary complexes with other proteins, as well as those that bind to DNA to regulate gene expression, chaperones which help proteins fold, and those that destroy proteins (proteases). The following is an example of an enzyme-labeled PPI between *JAK2* and *Ref-1* (from AIMed corpus):

> *The cytokine-activated tyrosine kinase* ***JAK2*** *activates* ***Raf-1*** *in a p21ras-dependent manner.*

This process of adding type labels proved to be the most difficult and labor-intensive aspect of the training data curation with thousands of gene names and symbols that required external lookups in addition to an equally large host of specialized biological jargon and acronyms (chemical names, cell lines, experimental conditions, etc.) that required research to differentiate from proteins and establish the context necessary for understanding each sentence. Importantly, because this annotation effort is informed by resources and knowledge external to the text in question, it encodes specialized domain knowledge that makes the PPI type classification task more challenging, increasing pressure on ML models to capture sufficiently informative context adequately to make a class determination.

Appendix A shows the annotation process. Two domain experts have performed the PPI annotation and reached a high inter-annotator agreement as seen in Appendix B. The definition of an interaction and the annotation rules were carefully determined ahead of time, according to domain expertise. Some of the rules are shown in Appendix C, and the complete rules can be found in our GitHub repository.

IV. Methodology

We have adopted a Transformer-based approach for the PPI classification task. In particular, we improve a relation representation exploiting the relational context information of an entity pair.

A. Relation Representation augmented with Attention-based Context Information

In a relation classification task, the `[CLS]` token is frequently used to represent a relation representation, which is a special classification token in BERT employed to capture the overall information of an input sequence. Another popular method is the entity mention pooling approach that concatenates a pair of two max-pooled entity embeddings in the last hidden state of BERT. To explicitly indicate target tokens for a relation, entity markers can be used in input, which are additional special input tokens indicating which tokens need focus for relation learning. Soares, Fitzgerald, Ling, and Kwiatkowski [20] have conducted the comparative study between marker-free and marker-embed representations showing the marker embedded approach outperforms marker-free representations on several supervised relation extraction tasks. Specifically, the concatenation of the entity start markers achieves the best performance.

We additionally improve the relation representation built upon a pair of entities or entity start markers by adding relational context information of entities. The rationale is that additional tokens for relational context can serve a crucial role in determining the relation of the entities. For instance, the word *activates* in *"A activates B"* and *Interaction* in *"Interaction between A and B"* are important clues for the effector-effectee relation. To find the most relevant tokens for relation information, we leverage entity tokens' attention probabilities generated in the last hidden layer in BERT. We

sum two entities' attention probabilities and retrieve additional tokens by the probability scores. The retrieved tokens are max-pooled then added to the final relation representation.

$$e^{attn} = max(\sum_{h}^{H} P_{attn}(e_h))$$

$$rc(e_1^{attn}, e_2^{attn}) = max(\sum_{i}^{N} e_{1,i}^{attn} + e_{1,i}^{attn})$$

$$\mathbf{x^r} = e_1^r \oplus rc(e_1^{attn}, e_2^{attn}) \oplus e_2^r,$$

where P_{attn} denotes attention probabilities of a token. H is the number of heads in the model. rc stands for relation context, and N is the number of tokens to be attentive. N also is a hyper-parameter and is set prior to model training. In this study, N is set as 20% of an input length, which was empirically determined using validation sets of biomedical relation extraction benchmarks (see Appendix D). $\mathbf{x^r}$ is the final relation representation for a classifier, which is the linking of entity embeddings (e_1^r, e_2^r) (mention pooling or entity start marker) and a max-pooled relation context embedding. When selecting tokens for relation context, we only account for alphanumerical tokens and exclude entity tokens and special tokens (besides entity markers). If a token is a part of a word (tokens with "##"), the entire word is included. Figure 1 illustrates the construction of a relation representation for a sentence with entity start markers, and the mention pooling approach is depicted in Appendix E.

B. Model Architecture

Our Transformer-based relation extraction model performs a sequence classification task using a logistic regression with softmax to determine the probability of relation class (e.g., $c \in \{$*enzyme*, *structural*, *negative*$\}$) as follows:

$$P(c|X) = softmax(W\mathbf{x^r}), \tag{1}$$

where X and $\mathbf{x^r}$ denote examples and relation representations, respectively. The model parameters are optimized using a categorical cross entropy.

$$-\sum_{c} \delta(X, c) \log P(c|X), \tag{2}$$

where $\delta(X, c)$ indicates whether the class of X is correctly predicted ($\delta(X, c) = 1$) or not ($= 0$). Algorithm 1 illustrates the model training procedure.

V. Experimental Setup

We first demonstrate the effectiveness of the proposed approach on four well-known relation extraction benchmark datasets in the biomedical domain. Then, the method is evaluated on the five PPI benchmark corpora and our PPI corpus with interaction types by comparing the performance with SOTA models.

Algorithm 1 Training a PPI model

Initialize: Load a pre-trained BERT model and set the max epoch and mini-batch size.
Output: Refined BERT model for PPI classification task using an attention-based relation representation.

Given relation extraction samples, define entity spans and add entity tags when using markers.
for s in $S_{relation}$ **do**
  $D \leftarrow define\_entity\_span\_and\_add\_marker(s)$
end for
while *epoch* **to** $epoch_{max}$ **do**
  // b is a mini-batch.
  for b **in** D **do**
    for each (*e1: entity 1, e2: entity 2*) $\in b$ **do**
      Generate attention-based relation representations.
      $R \leftarrow e1\_emb \oplus relation\_context \oplus e2\_emb$
    end for
    Produce logits.
    $logits = relation\_classifier(R)$
    Compute loss.
    $\mathcal{L} = CrossEntropyLoss(logits, labels)$
    Compute gradient and update parameters.
    $\theta = \theta - \eta\nabla\theta$
  end for
end while

A. Datasets

In this study, we use four biomedical relation extraction (RE) datasets: ChemProt [21], DDI [22], GAD [23], and EU-ADR [24]. There are various versions of the ChemProt, DDI, and GAD datasets. Here, we adopt the recent and widely used benchmark data, the Biomedical Language Understanding and Reasoning Benchmark (BLURB) provided by [25]. We also use the EU-ADR data in BioBERT [26]. The ChemProt, DDI, and GAD datasets consist of a train/validation/test set, while the EU-ADR contains 10-fold sets for cross validation. In all of the data, target entities are anonymized with pre-defined tags, including `@GENE$`, `@CHEMICAL$`, `@DRUG$`, and `@DISEASE$`. In ChemProt and DDI, additional tags, `@CHEM-GENE$` and `@DRUG-DRUG$`, are used for overlapping entities. When entity markers are used, `@CHEM-GENE$` and `@DRUG-DRUG$` are surrounded by the `[E1-E2]` tag. Descriptions of each data follow, and Table I displays the number of data samples.

1) ChemProt contains chemical-protein interactions extracted from 1,820 PubMed abstracts, and the task is evaluated using five high-level relation interaction classes: CPR:3 (UPREGULATOR), CPR:4 (DOWN-REGULATOR), CPR:5 (AGONIST), CPR:6 (ANTAGONIST), and CPR:9 (SUBSTRATE).
2) DDI consists of drug-drug relations four relation classes (Advice, Effect, Mechanism, Int) based on 792 texts from DrugBank and 233 Medline abstracts.

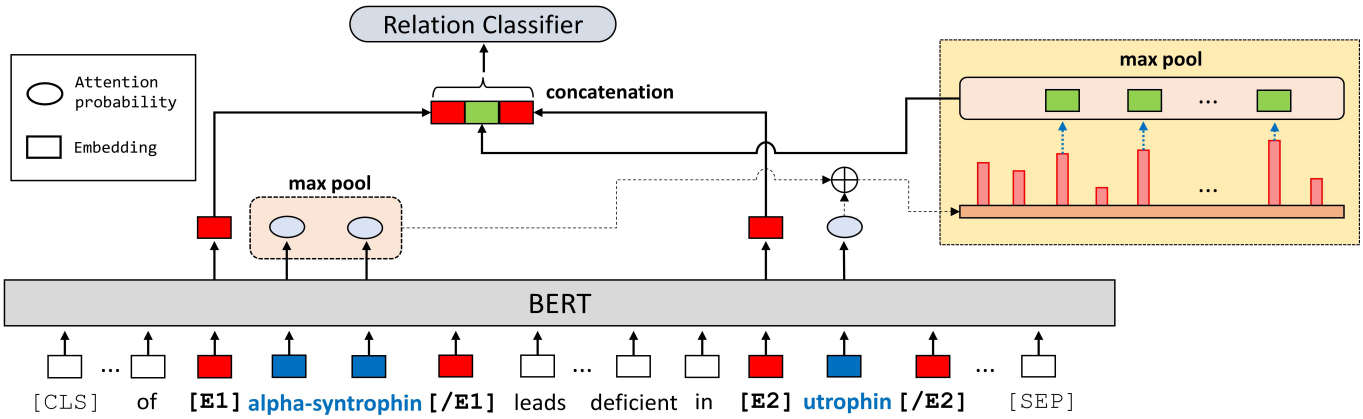


Fig. 1. The relation representation consists of entity start markers and the max-pooled of relational context, which is a series of tokens chosen by attention probability of the entities. The relation representation based on mention pooling is depicted in Appendix E. $\oplus$ denotes element-wise addition. The example sentence is *Absence of alpha-syntrophin leads to structurally aberrant neuromuscular synapses deficient in utrophin*. (Source: BioInfer corpus).

3) GAD (The Genetic Association Database corpus) contains a set of gene-disease binary associations, which was semi-automatically collected from PubMed abstracts.
4) EU-ADR features a list of binary associations between drugs, diseases, genes, and proteins annotated on Medline abstracts.

TABLE I
STATISTICS OF BIOMEDICAL RELATION EXTRACTION DATASETS. EU-ADR CONSISTS OF 10-FOLD SETS FOR CROSS VALIDATION.

	Train	**Dev**	**Test**	Total
ChemProt	18,035	11,268	15,745	45,048
DDI	25,296	2,496	5,716	33,508
GAD	4,261	535	534	5,330
EU-ADR	*NA*	*NA*	*NA*	355

TABLE II
FIVE PPI BENCHMARK CORPORA FOR *positive* AND *negative* CLASSES.

Data \ **Class**	**Positive**	**Negative**
AIMed	1,000	4,834
BioInfer	2,534	7,132
HPRD50	163	270
IEPA	335	482
LLL	164	166
TOTAL	4,196	12,884

The five PPI benchmark corpora include AIMed [5], BioInfer [6], HPRD50 [7], IEPA [8], and LLL [9]. We adopt the unified version of PPI benchmark datasets provided by [10] that has been used in the SOTA models. In the datasets, the PPI relations are tagged with either *positive* or *negative*. The corpus statistics is described in Table II. Our PPI annotations with interaction types (*enzyme*, *structural*, or *negative*) are the expanded version of the five benchmark corpora and the BioCreative VI protein interaction dataset [4]. Table III displays the corpora statistics. The annotation work in all corpora has been carried out in a sentence boundary as engaged in the five PPI benchmark corpora.

TABLE III
INTERACTION TYPED PPI CORPORA FOR *enzyme*, *structural*, AND *negative* CLASSES. † ANNOTATIONS USING THE PPI DATA FROM BIOCREATIVE VI TRACK 4: MINING PROTEIN INTERACTIONS AND MUTATIONS FOR PRECISION MEDICINE (PM). THE SIGNIFICANT REDUCTION FROM THE ORIGINAL DATA IN NEGATIVE SAMPLES IS EXPLAINED IN III-A3.

Data \ **Class**	**Enzyme**	**Structural**	**Negative**
BioCreative VI†	378	83	0
AIMed	548	182	1,371
BioInfer	604	1,465	2,148
HPRD50	103	34	87
IEPA	271	2	224
LLL	163	0	0
TOTAL	2,067	1,766	3,830

B. Implementation details

We use domain-specific pre-trained BERT models on biomedical literature, including BioBERT [26] and PubMedBERT [25], which has demonstrated excellent performance in biomedical NLP applications. We use PyTorch (version 1.10.2) and the HuggingFace's Transformers package (version 4.17.0) [28], while the pre-trained models used are obtained from the HuggingFace model repository[7]. The model architecture and weight initialization follow the pre-trained models, and the hyper-parameters are tuned with the range: epoch number (3–20), batch size (8, 16), and learning rate (1e-5, 3e-5, 5e-5) with Adam. For objective comparisons, we endeavor to adopt the same models and hyperparameters used in the SOTA systems (if available) to reproduce the identical results with their relation representation. The hyperparameter details can be found in our GitHub repository. We use a dense layer

[7] https://huggingface.co/models

TABLE IV
F1 SCORES ON THE TEST SETS FOR CHEMPROT, DDI, GAD, AND 10-FOLD CV FOR EU-ADR. IN THE DATASETS, TARGET ENTITIES ARE ANONYMIZED WITH PRE-DEFINED TAGS (E.G., @GENE$, @CHEMICAL$, @DRUG$). *Mention* IS A CONCATENATION OF THE CONTEXTUAL EMBEDDINGS OF THE ENTITY MENTIONS. *Entity Start (markers)* ARE [E1] AND [E2]. (**BOLD**: BEST SCORE IN OUR METHOD; UNDERLINE: BEST SCORE IN SOTA)

		ChemProt	DDI	GAD	EU-ADR
SOTA					
KeBioLM [27]		77.5	81.9	84.3	-
PubMedBERT [25]		77.2	83.6	84.1	-
BioBERT [26] (PyTorch version)		-	-	82.4	85.1
Ours					
Input	*Representation*				
Entity Anonymization	[CLS]	77.9	81.7	82.1	85.1
	Mention	78.8	80.0	83.0	84.2
	Mention + Relation Context	**80.1**	81.3	**85.0**	**86.0**
Entity Anonymization + Markers	[CLS]	78.7	82.6	83.5	85.6
	Entity Start	76.5	80.7	82.6	85.0
	Entity Start + Relation Context	79.2	**83.6**	84.5	85.5

with linear activation as a post-Transformer layer and train the model on the machine, Tesla V100-SXM2-32GB $\times$ 2.

VI. RESULTS AND DISCUSSION

A. Evaluation on biomedical RE datasets

We use the BioBERT large-cased model for the ChemProt, the PubMedBERT-uncased-fulltext model for DDI and GAD, and the BioBERT base-cased model for EU-ADR. We compare our model's performance with the SOTA results, including KeBioLM [27] for ChemProt and GAD, PubMedBERT [25] for DDI, and BioBERT [26] (Version[8] as our model was built on PyTorch) for EU-ADR. KeBioLM and PubMedBERT use the combinations of entity mentions, and BioBERT uses the [CLS] token for relation classification. We measure the performance by the same metrics used in the SOTA systems. The results demonstrate that our proposed representation of the entity mention augmented with the relation context achieved SOTA results for ChemProt, GAD, EU-ADR, while the combination of entity start markers with the relation context produced comparable performance for DDI (shown in Table IV). The relation context improves the predictions in all cases. Notably, its significance is clearly shown in EU-ADR, where we have replicated the result obtained in the SOTA model ([CLS]: 85.1 F1 score) using the same model, input (without markers), representation, and adding the relation context to the mention pooling, which produced a superior result over the [CLS] token.

B. Evaluation on PPI datasets

We adopt BioBERT for the evaluation on the PPI data that achieved greater improvements on the performances in the recent PPI extraction works [13], [19]. To compare the performance of the proposed approach with SOTA works, we evaluate our model using a 10-fold cross-validation (CV) manner and a micro F1 performance metric as adopted in the SOTA models. Table V displays the evaluation results on the five benchmark PPI corpora, showing our models produce the best performances and outperform the SOTA models on the overall classification as described in the average F1 scores. Unlike the entity anonymized inputs, the inputs with entity markers perform better than the original inputs across all data, while using the [CLS] token in the original input performs the worst. This finding also has been observed in earlier works [20], [25], implying the significance of explicit indication for target entities, such as markers or entity anonymization, with its type. The relation context constantly improves the performances, although a slight degradation occurred for the combination with entity mention in the LLL data, and the representation of entity start markers augmented with relation context achieves the best predictions.

In addition, we examine the model's ability on our PPI corpora with interaction types. In this experiment, we combine the six corpora where some datasets contain only single class or highly skewed samples so the model can be trained on more balanced data. The model evaluation also is carried out in a 10-fold CV manner, and Table VI reflects the micro F1 scores of each representation. The results demonstrate that the models yield consistent predictions with the best 87.8 F1 score compared to the previous experiments, and the representations augmented with relation context continually generate satisfactory outcomes. Through the observation of enhanced results on various relation extraction tasks, we can conclude that contextual representations that target entities are attentive and able to effectively provide additional information to determine the relations of entity pairs.

[8] https://github.com/dmis-lab/biobert-pytorch

TABLE V

F1 SCORES VIA 10-FOLD CV ON THE PPI CLASSIFICATION WITH THE FIVE BENCHMARK PPI CORPORA. *Mention* IS A CONCATENATION OF THE CONTEXTUAL EMBEDDINGS OF THE ENTITY MENTIONS. *Entity Start (markers)* ARE `[E1]` AND `[E2]`. OUR METHODS USE THE BIOBERT BASE-CASED MODEL. (**BOLD**: BEST SCORE IN OUR METHOD; UNDERLINE: BEST SCORE IN SOTA)

		AIMed	BioInfer	HPRD50	IEPA	LLL	*Avg.*
SOTA							
DSTK [15]		71.0	76.3	80.0	80.2	89.2	79.3
DeepResCNN [29]		77.6	86.9	77.7	75.5	83.2	80.2
LBERT [13]		74.0	72.8	85.5	83.7	86.0	80.4
ADVBERT [19]		83.9	90.3	84.8	84.9	88.7	86.5
Ours							
Input	*Representation*						
Original	`[CLS]`	83.2	79.1	65.3	68.0	62.4	71.6
	Mention	90.6	88.0	83.4	85.2	84.9	86.4
	Mention + Relation Context	90.8	88.2	84.5	85.9	84.6	86.8
Entity Markers	`[CLS]`	91.8	90.9	83.1	82.9	85.2	86.8
	Entity Start	91.4	90.9	87.3	86.4	88.8	89.0
	Entity Start + Relation Context	**92.0**	**91.3**	**88.2**	**87.4**	**89.4**	**89.7**

TABLE VI

F1 SCORES VIA 10-FOLD CV ON THE TYPED PPI CORPORA. THE BIOBERT BASE-CASED MODEL IS USED.

		Typed PPI
Input	*Representation*	
Original	`[CLS]`	84.7
	Mention	85.9
	Mention + Relation Context	86.4
Entity Markers	`[CLS]`	85.9
	Entity Start	86.9
	Entity Start + Relation Context	**87.8**

VII. CONCLUSION

In this work, we have augmented existing PPI corpora annotated with interaction types, which is expected to be beneficial for extracting more PPI information from scientific publications. We also have presented a Transformer architecture-based model for relation extraction. Specifically, we have improved a relation representation by adding relational context information based on entities' attention probabilities. Our models outperform SOTA models and offer proof about the effectiveness of additional relational context embedding on the biomedical relation extraction benchmarks and PPI corpora.

We will continue to improve our PPI annotations by resolving identified problems, including debiasing the training data. More examples are needed from across biological subject areas (plants, environmental, microbiomes, etc). Our goal is to provide a tool that works across all subfields of biology. Granularity in type classifications also needs to be increased, which will require more training data and manual annotation. Finally, statements of interaction that span two (or more) sentences also will require added attention in the future.

REFERENCES

[1] A. Brückner, C. Polge, N. Lentze, N. Auerbach, and U. Schlattner, "Yeast two-hybrid, a powerful tool for systems biology," *International Journal of Molecular Sciences*, no. 10, pp. 2763–2788, 2009.

[2] W. Dunham, M. Mullin, and A. Gingras, "Affinity-purification coupled to mass spectrometry: basic principles and strategies," *Proteomics*, vol. 12, no. 10, pp. 1576–90, 2012.

[3] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in neural information processing systems*, 2017, pp. 5998–6008.

[4] R. Islamaj Doğan, S. Kim, A. Chatr-Aryamontri, C.-H. Wei, D. C. Comeau, R. Antunes, S. Matos, Q. Chen, A. Elangovan, N. C. Panyam *et al.*, "Overview of the biocreative vi precision medicine track: mining protein interactions and mutations for precision medicine," *Database*, vol. 2019, 2019.

[5] R. Bunescu, R. Ge, R. J. Kate, E. M. Marcotte, R. J. Mooney, A. K. Ramani, and Y. W. Wong, "Comparative experiments on learning information extractors for proteins and their interactions," *Artificial intelligence in medicine*, vol. 33, no. 2, pp. 139–155, 2005.

[6] S. Pyysalo, F. Ginter, J. Heimonen, J. Björne, J. Boberg, J. Järvinen, and T. Salakoski, "Bioinfer: a corpus for information extraction in the biomedical domain," *BMC bioinformatics*, vol. 8, no. 1, pp. 1–24, 2007.

[7] K. Fundel, R. Küffner, and R. Zimmer, "Relex—relation extraction using dependency parse trees," *Bioinformatics*, vol. 23, no. 3, pp. 365–371, 2007.

[8] J. Ding, D. Berleant, D. Nettleton, and E. Wurtele, "Mining medline: abstracts, sentences, or phrases?" in *Biocomputing 2002*. World Scientific, 2001, pp. 326–337.

[9] C. Nédellec, "Learning language in logic-genic interaction extraction challenge," in *4. Learning language in logic workshop (LLL05)*. ACM-Association for Computing Machinery, 2005.

[10] S. Pyysalo, A. Airola, J. Heimonen, J. Björne, F. Ginter, and T. Salakoski, "Comparative analysis of five protein-protein interaction corpora," in *BMC bioinformatics*, vol. 9, no. 3. BioMed Central, 2008, pp. 1–11.

[11] D. Tikk, P. Thomas, P. Palaga, J. Hakenberg, and U. Leser, "A comprehensive benchmark of kernel methods to extract protein–protein interactions from literature," *PLoS Comput Biol*, vol. 6, no. 7, p. e1000837, 2010.

[12] Q.-C. Bui, S. Katrenko, and P. M. Sloot, "A hybrid approach to extract protein–protein interactions," *Bioinformatics*, vol. 27, no. 2, pp. 259–265, 2011.

[13] N. Warikoo, Y.-C. Chang, and W.-L. Hsu, "Lbert: Lexically aware transformer-based bidirectional encoder representation model for learning universal bio-entity relations," *Bioinformatics*, vol. 37, no. 3, pp. 404–412, 2021.

[14] W. A. Baumgartner, Z. Lu, H. L. Johnson, J. G. Caporaso, J. Paquette, A. Lindemann, E. K. White, O. Medvedeva, K. B. Cohen, and L. Hunter, "Concept recognition for extracting protein interaction relations from biomedical text," *Genome biology*, vol. 9, no. 2, pp. 1–15, 2008.

[15] G. Murugesan, S. Abdulkadhar, and J. Natarajan, "Distributed smoothed tree kernel for protein-protein interaction extraction from the biomedical literature," *PLoS One*, vol. 12, no. 11, p. e0187379, 2017.

[16] Y. Peng and Z. Lu, "Deep learning for extracting protein-protein interactions from biomedical literature," *BioNLP 2017*, p. 29, 2017.

[17] J. D. M.-W. C. Kenton and L. K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of NAACL-HLT*, 2019, pp. 4171–4186.

[18] J. Vig, A. Madani, L. R. Varshney, C. Xiong, N. Rajani *et al.*, "Bertology meets biology: Interpreting attention in protein language models," in *International Conference on Learning Representations*, 2020.

[19] Z. Tang, X. Guo, Z. Bai, L. Diao, S. Lu, and L. Li, "A protein-protein interaction extraction approach based on large pre-trained language model and adversarial training," *KSII Transactions on Internet and Information Systems (TIIS)*, vol. 16, no. 3, pp. 771–791, 2022.

[20] L. B. Soares, N. Fitzgerald, J. Ling, and T. Kwiatkowski, "Matching the blanks: Distributional similarity for relation learning," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2019, pp. 2895–2905.

[21] Y. Peng, A. Rios, R. Kavuluru, and Z. Lu, "Extracting chemical–protein relations with ensembles of svm and deep learning models," *Database*, vol. 2018, 2018.

[22] M. Herrero-Zazo, I. Segura-Bedmar, P. Martínez, and T. Declerck, "The ddi corpus: An annotated corpus with pharmacological substances and drug–drug interactions," *Journal of biomedical informatics*, vol. 46, no. 5, pp. 914–920, 2013.

[23] À. Bravo, 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.

[24] E. M. Van Mulligen, 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.

[25] Y. Gu, R. Tinn, H. Cheng, M. Lucas, N. Usuyama, X. Liu, T. Naumann, J. Gao, and H. Poon, "Domain-specific language model pretraining for biomedical natural language processing," *ACM Transactions on Computing for Healthcare (HEALTH)*, vol. 3, no. 1, pp. 1–23, 2021.

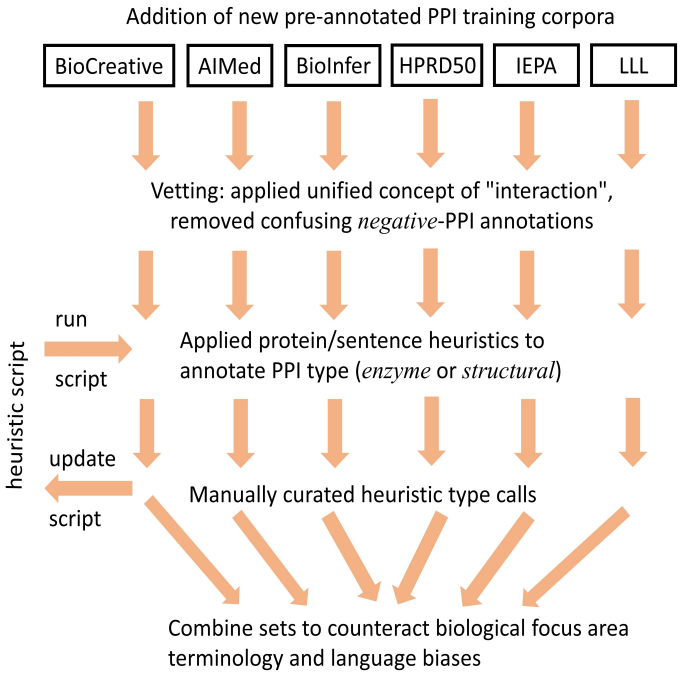
[26] J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So, and J. Kang, "Biobert: a pre-trained biomedical language representation model for biomedical text mining," *Bioinformatics*, vol. 36, no. 4, pp. 1234–1240, 2020.

[27] Z. Yuan, Y. Liu, C. Tan, S. Huang, and F. Huang, "Improving biomedical pretrained language models with knowledge," in *Proceedings of the 20th Workshop on Biomedical Language Processing*, 2021, pp. 180–190.

[28] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz *et al.*, "Transformers: State-of-the-art natural language processing," in *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations*, 2020, pp. 38–45.

[29] H. Zhang, R. Guan, F. Zhou, Y. Liang, Z.-H. Zhan, L. Huang, and X. Feng, "Deep residual convolutional neural network for protein-protein interaction extraction," *IEEE Access*, vol. 7, pp. 89 354–89 365, 2019.

APPENDIX A
ANNOTATION PROCESS DIAGRAM



APPENDIX B
INTER-ANNOTATOR AGREEMENT

We measured the inter-annotator agreement scores to observe the discrepancy between the annotators in the PPI relation types. The annotated data statistics can be found in Table III. As seen in Table VII, the two annotators achieved a high inter-annotator agreement.

TABLE VII
INTER-ANNOTATOR AGREEMENT STATISTICS BETWEEN THE TWO ANNOTATORS FOR THE THREE PPI TYPES.

Relation type	A1	A2
enzyme		
A1	*NA*	0.92
A2	0.92	*NA*
structural		
A1	*NA*	0.90
A2	0.90	*NA*
negative		
A1	*NA*	0.95
A2	0.95	*NA*

APPENDIX C
ANNOTATION RULE EXAMPLES

1) Proteins/Genes ending in –in or –ins are pre-identified as structural (actin, catenin, ...). Exceptions include:
   a) Toxin
   b) Beta-catenin (can be gene regulator OR structural as it is a dual-function gene)
   c) Calreticulin – multifunction; mostly enzyme.

TABLE VIII
F1 SCORES ON THE VALIDATION SET FOR CHEMPROT, DDI, GAD, AND EU-ADR WITH DIFFERENT SIZES OF RELATION CONTEXT: 10%, 20%, AND 30% OF AN INPUT LENGTH (EXCEPT FOR TOKENS TO BE IGNORED).

	ChemProt	**DDI**	**GAD**	**EU-ADR**	*Avg.*
	10%/20%/30%	10%/20%/30%	10%/20%/30%	10%/20%/30%	10%/20%/30%
Mention + Relation Context	82.2/**82.3**/81.7	85.1/**87**/85.1	83.9/**84.6**/84.2	**86.3**/86.2/85.8	84.4/**85.0**/84.2
Entity Start + Relation Context	81.8/**83.4**/82.9	86.6/**86.8**/83.4	83.7/**84.4/84.4**	88.5/**88.9**/88.6	85.2/**85.9**/84.8

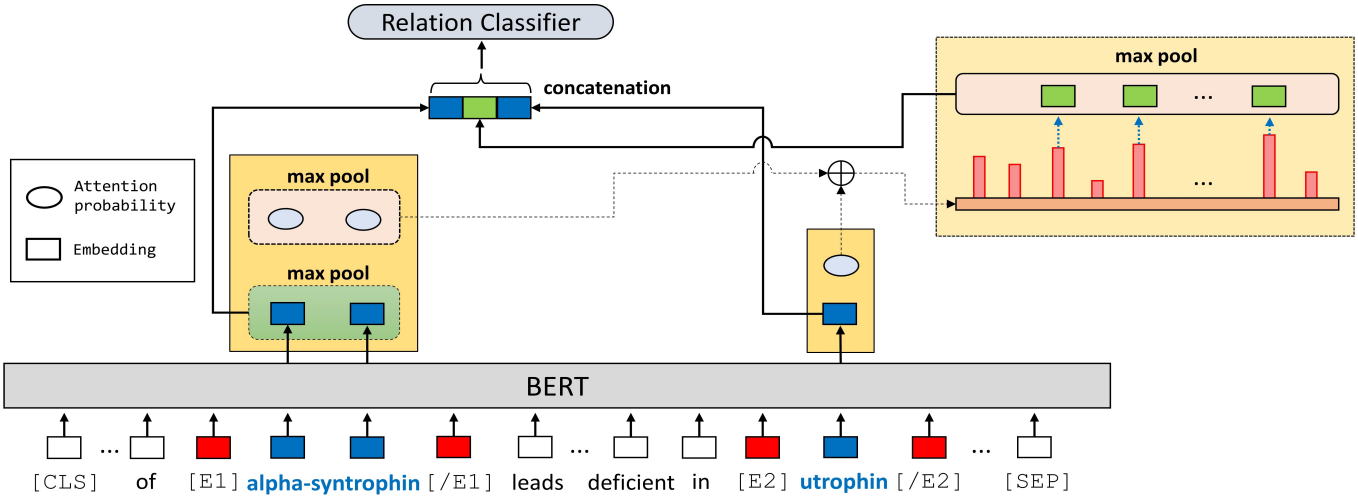


Fig. 2. The relation representation consists of the max-pooled of two entity contextualized embeddings and the max-pooled of relational context, which is a series of tokens chosen by attention probability of the entities. ⊕ denotes element-wise addition. The example sentence is *Absence of alpha-syntrophin leads to structurally aberrant neuromuscular synapses deficient in utrophin*. (Source: BioInfer corpus).

2) Histones and nucleosomes are not considered structural because their "structure" is mutable and controls regulation.
3) Proteins/Genes ending in –ase are preidentified as enzymes.
4) Proteins/Genes containing inhibitor, activator, transcription factor, repressor, enhancer, or regulator are preidentified as enzymes.

APPENDIX D
EVALUATION ON DIFFERENT RELATION CONTEXT SIZES

To find an appropriate size of attentive context of target entities, we evaluated different sizes of relation context using the biomedical relation extraction benchmark datasets: ChemProt, DDI, GAD, and EU-ADR. We leveraged 10%, 20%, and 30% of a sequence length for a number of attentive tokens of target entities and compared them on the respective validation set of the datasets. When selecting tokens for relation context, we only account for the alphanumerical tokens and exclude entity tokens (e.g., `[CLS]`; `[SEP]`) and special tokens (besides entity markers). Because the EU-ADR is a 10-fold cross validation set, we split a training set in each fold in a 9:1 ratio, i.e., 90% of the data are used for training the model, while 10% are used for validating the model. Without using a test set, the average scores of cross validations on train/validation sets were measured. Table VIII demonstrates the F1 scores of different sizes of relation context, and 20% of an input length—except for tokens to be ignored—showed the best performances on both entity mention use and entity start marker use in representation.

APPENDIX E
RELATION REPRESENTATION USING MENTION POOLING

Figure 2 illustrates the construction of a relation representation for a sentence using mention pooling. As in the entity start marker method, input sentences are tagged with entity markers. The rectangles and ovals represent the tokens' embeddings and attention probabilities, respectively.