Bioinformatics, 36, 2020, i490–i498
doi: 10.1093/bioinformatics/btaa430
ISMB 2020

PEDL: extracting protein–protein associations using deep language models and distant supervision

Leon Weber[1,2], Kirsten Thobe[2], Oscar Arturo Migueles Lozano[2], Jana Wolf[2,*] and Ulf Leser[1,*]

[1]Computer Science Department, Humboldt-Universität zu Berlin, Berlin 10099, Germany and [2]Group Mathematical Modelling of Cellular Processes, Max Delbrück Center for Molecular Medicine in the Helmholtz Association, Berlin 13125, Germany

*To whom correspondence should be addressed.

Abstract

Motivation: A significant portion of molecular biology investigates signalling pathways and thus depends on an up-to-date and complete resource of functional protein–protein associations (PPAs) that constitute such pathways. Despite extensive curation efforts, major pathway databases are still notoriously incomplete. Relation extraction can help to gather such pathway information from biomedical publications. Current methods for extracting PPAs typically rely exclusively on rare manually labelled data which severely limits their performance.

Results: We propose **P**PA **E**xtraction with **D**eep **L**anguage (PEDL), a method for predicting PPAs from text that combines deep language models and distant supervision. Due to the reliance on distant supervision, PEDL has access to an order of magnitude more training data than methods solely relying on manually labelled annotations. We introduce three different datasets for PPA prediction and evaluate PEDL for the two subtasks of predicting PPAs between two proteins, as well as identifying the text spans stating the PPA. We compared PEDL with a recently published state-of-the-art model and found that on average PEDL performs better in both tasks on all three datasets. An expert evaluation demonstrates that PEDL can be used to predict PPAs that are missing from major pathway databases and that it correctly identifies the text spans supporting the PPA.

Availability and implementation: PEDL is freely available at https://github.com/leonweber/pedl. The repository also includes scripts to generate the used datasets and to reproduce the experiments from this article.

Contact: leser@informatik.hu-berlin.de or jana.wolf@mdc-berlin.de

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Molecular biology explores chemical and physical interactions between key intermediates, mostly proteins, in cells. The biological function rarely depends on single interactions but on the complex interplay of many, for example in cellular signalling, metabolism or gene regulation. Techniques from network analysis are widely used to connect interactions between proteins to the functional organization of cells (Barabasi and Oltvai, 2004). A major challenge for building these networks is to gather all relevant information from the literature, since the quality of the model and the model predictions rely on completeness and correctness of the individual proteins and their interactions. It is important to not only have the knowledge that two proteins interact but also to know the exact type of interaction, such as kinase–substrate relation or gene–gene regulation. These functional protein–protein associations (PPAs) (Junge and Jensen, 2019) can be found in manually curated databases such as Reactome (Jassal *et al.*, 2019) or the Protein Interaction Database (PID) (Schaefer *et al.*, 2009). However, these databases are notoriously incomplete despite extensive curation efforts (Köksal *et al.*, 2018). For instance, we found that for a state-of-the-art model of p53 signalling (Hat *et al.*, 2016) 25% of the contained PPAs cannot be found, neither in Reactome nor in PID (see Supplementary Material S1 for details). Extracting PPAs from the biomedical literature has been a long-standing research goal. Early approaches focused on matching sentences to manually defined templates, usually leading to high-precision but low-recall results (Friedman *et al.*, 2001). Later methods used supervised machine learning to classify whether a sentence expresses a relation between a given pair of proteins, frequently relying on support-vector-machines (SVMs) with graph kernels (Miwa *et al.*, 2009; Tikk *et al.*, 2012). Similar techniques have been applied to biomedical event extraction, which aimed at not only extracting pairwise relations between two proteins but also complex biochemical reactions between proteins (Björne *et al.*, 2009; Miwa *et al.*, 2010). More recently, also approaches based on neural networks have been applied to sentence-wise supervised classification of protein–protein interactions (Peng and Lu, 2017) and to biomedical event extraction (Björne and Salakoski, 2018). None of these methods are capable of detecting relations between proteins mentioned in different sentences or make use of pre-trained language models that recently have led to large gains in other Natural Language Processing (NLP) tasks (Devlin *et al.*, 2019). Additionally, these models rely on manually annotated training data, which for PPA-extraction requires expert knowledge and thus is very costly. Consequently, the available manually labelled PPA datasets are

rather small, typically containing at most a few thousand sentences (Pyysalo *et al.*, 2008).

This data sparsity led to the introduction of distantly supervised approaches (Mintz *et al.*, 2009) for PPA prediction (Junge and Jensen, 2019; Poon *et al.*, 2014; Thomas *et al.*, 2011). However, both Thomas *et al.* (2011) and Poon *et al.* (2014) are based on non-neural models with manually defined features and Junge and Jensen (2019) use averaged word embeddings without leveraging multi-instance learning. Distantly supervised relation extraction methods generate noisy training data by aligning a knowledge base to a large collection of texts. To achieve this, a large knowledge base of relations (in our case PPAs) in the form (e_1, r, e_2) is connected to a text by first linking the entities from the knowledge base e_1, e_2 to the entities in the text. Initially, the core assumption of distant supervision was that every sentence that contains the entities e_1, e_2 expresses the relation r. This assumption can be relaxed through the use of multi-instance learning (Hoffmann *et al.*, 2011; Riedel *et al.*, 2010; Surdeanu *et al.*, 2012). Multi-instance learning explicitly models the assumption that *at least one* sentence expresses the relation between the entity pair in question by selecting only a subset of the sentences to generate the prediction. Originally, probabilistic graphical models were used to achieve this, but recently deep learning-based models in the form of piece-wise convolutional neural networks (Zeng *et al.*, 2015) with selective attention (Lin *et al.*, 2016) were successfully applied. An orthogonal line of work also uses auxiliary directly supervised training examples, achieving significant improvements for graphical models (Angeli *et al.*, 2014; Pershina *et al.*, 2014) and for neural networks (Beltagy *et al.*, 2019a; Liu *et al.*, 2017). However, all of these approaches only consider entity pairs that occur together in the same sentence, which severely limits recall (Quirk and Poon, 2017).

Accordingly, there is growing interest in using text that spans multiple sentences for distantly supervised biomedical relation extraction. Verga *et al.* (2018) used transformer-based models to predict all relations between chemicals, diseases and genes contained in one abstract but do not consider multiple abstracts simultaneously. Quirk and Poon (2017) used multi-instance learning to predict relations between drugs and genes that can be up to three sentences apart with an SVM-classifier on manually defined dependency graph features.

Recently, deep language models have seen widespread success in NLP, including the biomedical domain (Beltagy *et al.*, 2019b). The often used two-step process of training these models can be regarded as a type of transfer learning (Pratt *et al.*, 1991): The first step is pre-training, in which a large model, typically with hundreds of million parameters, is trained on a huge corpus of texts with a language modelling task. In the second step, the pre-trained model is applied to the target task, either by fine-tuning the model parameters or using the model to generate contextualized embeddings (Peters *et al.*, 2019). BERT (Devlin *et al.*, 2019) is a highly successful deep language model based on the transformer architecture (Vaswani *et al.*, 2017) which allows to train very large models efficiently by leveraging GPUs. Originally, BERT was trained on a large collection of books and English Wikipedia, but recently two BERT models trained on biomedical abstracts and full texts have been released, BioBERT (Lee *et al.*, 2019) and SciBERT (Beltagy *et al.*, 2019b). As BERT uses WordPiece tokenization (Wu *et al.*, 2016), it learns a domain-dependent vocabulary that allows it to use sub-word information to relate similar words such as *TRAF2* and *TRAF3*. **P**PA **E**xtraction with **D**eep **L**anguage (PEDL) uses SciBERT as its pre-trained language model, because unlike BioBERT, its WordPiece vocabulary is optimized for scientific literature.

In this work, we propose PEDL models, a model that predicts functional PPAs from biomedical publications. We approach this problem by combining pre-trained language models with distant supervision. Specifically, we source a large number of protein pairs together with their PPAs from the PID database and find texts mentioning these pairs in a collection of roughly 24 million abstracts of biomedical publications and 3 million full texts. The resulting PPA extraction dataset is distantly supervised, i.e. it only contains annotations for relations between the proteins but it is not known whether a text span actually confirms the relation. Given a protein pair, PEDL takes the text spans mentioning both proteins as input and predicts which PPAs hold for this pair, if any. Importantly, in what we call evidence prediction, PEDL predicts not only the PPAs but also which text span expresses it. We augment the training data of PEDL with data which additionally contains annotations for evidence predictions, which we generate from those gold standard datasets (Kim *et al.*, 2011b) that include annotations for all PPA-types considered by us. Following Beltagy *et al.* (2019b), we call this type of data *directly supervised*. We compare the performance of PEDL to state-of-the-art approaches on three different datasets and find that, on average, it performs much better for both PPA and evidence prediction. In a manual evaluation of the top 10 predicted PPAs, conducted by three experts in Systems Biology, we find that PEDL can be used to predict PPAs that cannot be found in major pathway databases. Furthermore, the predicted evidence text spans actually express the relation and thus can be used for easy verification of the predicted PPAs, which is important for expert curation.

2 Materials and methods

In this work, we model PPA extraction following a multi-instance learning framework for relation extraction (Hoffmann *et al.*, 2011; Riedel *et al.*, 2010; Surdeanu *et al.*, 2012). Given two proteins p_1 and p_2, we aim to predict all PPAs $r \in R$ relating p_1 to p_2 by leveraging a corpus of biomedical literature. We focus on a set R of five PPAs which is a subset of the Simple Interaction Format relations available in Pathway Commons:

- *in-complex-with* is true for a protein pair (A, B), if A and B occur together in at least one protein complex.
- *controls-state-change-of* means that A regulates some change of B. This can be a post-translational modification such as phosphorylation or ubiquitination or a transfer between cellular compartments.
- *controls-phosphorylation-of* is a subset of *controls-state-change-of* and means that A phosphorylates B.
- *controls-transport-of* is a subset of *controls-state-change-of* and denotes that A controls the transfer of B to a cellular compartment.
- *controls-expression-of* implies that A modulates the expression of B.

Additionally, in what we term *evidence prediction*, we want the model to find the strongest possible evidence for these PPAs in the form of text expressing the relation between the proteins. This section describes how PEDL combines deep language models, distant supervision and auxiliary directly supervised data to approach these two tasks. A detailed graphical description of PEDL can be found in Figure 1.

2.1 PPA prediction as multi-instance learning

To predict relations between proteins p_1 and p_2, the first step of PEDL is to collect all text spans T, up to a given length, mentioning p_1 and p_2 together. This requires the use of named entity recognition (NER) (Weber *et al.*, 2020) and named entity normalization (NEN) (Wei *et al.*, 2019) as a pre-processing step.

For the two sub-tasks of relation prediction and evidence prediction, the model has to produce two vectors $r \in \mathbb{R}^{|R|}$ and $e \in \mathbb{R}^{|T|}$, where R is the set of considered PPAs and T is the set of spans for the pair. The vector r contains $|R|$ scores $\in [0, 1]$ reflecting the confidence of PEDL in each type of PPA. e contains $|T|$ scores $\in [0, 1]$, each modelling PEDL's confidence that the corresponding text span expresses a relation between p_1 and p_2.

For this, PEDL predicts a score-matrix $S \in \mathbb{R}^{|T| \times |R|}$ for each text span, representing the confidence of the model that a text span supports a given PPA. To achieve this, we first mark the entity pair in each text span by surrounding the first entity with the entity markers

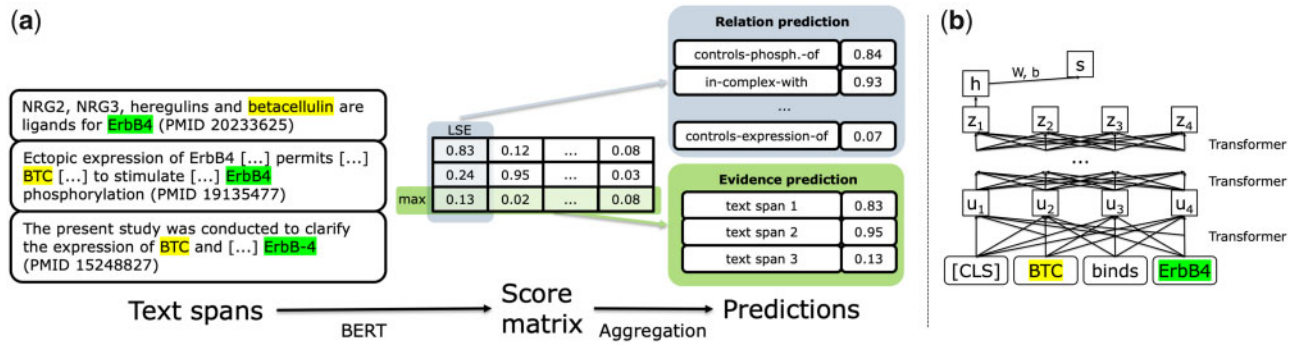

Fig. 1. (**a**) Overview of PEDL for the two tasks of relation prediction and evidence prediction. In this example, the model predicts relations for the protein pair BTC and ErbB4 given three text spans containing both proteins. First, the BERT component produces a score matrix containing a prediction for each text and relation type. The relation predictions are then generated by applying LSE column-wise to approximate the maximum score for a given PPA type across all spans. The evidence predictions are obtained by taking the row-wise maximum, which is the highest score assigned to this text span regardless of PPA type. (**b**) The generation of one row of the score matrix s. In each of BERT's 12 transformer layers, each token receives a 768 dimensional embedding (u_i for the first and z_i for the last layer). The embedding of the prepended [CLS] token is used to summarize the text span in the single vector h, which is then transformed to one row of the score matrix by the output layer (W, b)

<e1> and </e1> and the second entity with <e2>, </e2>. Then, each text span T_i is fed through BERT individually, to obtain the *[CLS]* embedding $h_i \in \mathbb{R}^{768}$ of the 768-dimensional final layer, which can be regarded as a summary of the whole text span. Finally, we use a single hidden layer to transform h_i to one row of the score matrix S_i containing logits reflecting the confidence of PEDL that the text span expresses a given PPA. See Figure 1 for a graphical description of this process.

The relation prediction r for each PPA type is generated by aggregating the scores for the PPA over all spans, i.e. column-wise. Correspondingly, the evidence prediction e for an individual sentence is produced by aggregating the scores of all PPA predictions for this sentence, i.e. row-wise. Finally, both vectors are normalized by applying the sigmoid function. In preliminary experiments, we used maximum for both aggregations, but found that the resulting sparse gradient flow hampered optimization. Thus, we use the smooth approximation of maximum LogSumExp as aggregation function for PPA predictions, because it allows for gradient flow through all sentences and empirically works well in end-to-end training of transformer models (Verga *et al.*, 2018). Putting everything together, the formulae for predicting PPAs and evidence are the following:

$$\begin{aligned} h_i &= \mathrm{BERT}_{[\mathrm{CLS}]}(T_i) \\ S_{ij} &= (W \cdot h_i)_j + b_j \\ e_i &= \sigma(\max_j(S_{ij})) \\ r_j &= \sigma(\log \sum_i \exp(S_{ij})), \end{aligned} \tag{1}$$

where log and exp denote element-wise application of logarithm and exponentiation, $W \in \mathbb{R}^{768 \times |R|}$ and $b \in \mathbb{R}^{|R|}$ are trainable parameters, and σ is the element-wise sigmoid function. Alternatively, S_{ij} can be directly used as an evidence score *per relation*.

For the training of PEDL, we assume that two types of data are available: *Distantly supervised* data which only has labels for relation prediction and *directly supervised* data which has labels for both relation and evidence prediction. Furthermore, we assume that both types of data share the same label space. The *directly supervised* data is used to give the model additional guidance on how text spans expressing PPAs look like. To achieve this, we combine both types of data using a multi-task learning framework. We introduce one loss term each type of data: $\mathcal{L}_{distant}$ for the distantly supervised and $\mathcal{L}_{direct}$ for the directly supervised data. The loss for the directly supervised data is composed of a loss term for relation prediction and another term for evidence prediction: $\mathcal{L}_{direct} = \mathcal{L}_{direct}^{relation} + \mathcal{L}_{direct}^{evidence}$. The loss for the distantly supervised data only consists of the loss term for the relation prediction task, because labels for evidence predictions are not available for this type of data: $\mathcal{L}_{distant} = \mathcal{L}_{distant}^{relation}$. The total loss for the batch is then a weighted average of the direct and distant losses:

$$\mathcal{L}_{total} = \alpha \cdot \mathcal{L}_{direct} + (1 - \alpha) \cdot \mathcal{L}_{distant} \tag{2}$$

where $\alpha \in [0, 1]$ is a hyperparameter controlling the relative importance of the direct loss and will be tuned on the development set of each considered dataset separately. At each optimization step, we sample a batch from the distant and one from the directly supervised data.

Since we model PPA prediction as a multi-label task, all losses are computed with binary cross entropy. Note, that the only parameters of PEDL are those of BERT and one output layer (W, b). We optimize these parameters with Adam (Kingma and Ba, 2015). The detailed hyperparameter settings can be found in Supplementary Material S2. One training step on one batch (16 protein pairs with up to 100 text spans each) takes ~9.5 s on four *RTX 2080 Ti* GPUs.

2.2 Data

The training of PEDL requires *distantly* and *directly supervised* data. To obtain the distantly supervised data, we follow the standard approach for creating a multi-instance learning dataset (Riedel *et al.*, 2010). First, we collect all protein pairs and the relations between each pair from a large knowledge base, where we opt for the PID data base (Schaefer *et al.*, 2009), due to its very high curation standards. We gather our data from the Simple Interaction Format version of PID provided by PathwayCommons (https://www.pathwaycommons.org/archives/PC2/v11/PathwayCommons11.pid.hgnc.txt.gz) (Cerami *et al.*, 2011). Then, for each protein pair p_1 and p_2, we collect all text spans up to the length of 300 characters that mention p_1 and p_2 together. To estimate the probability that a protein pair is related by none of the considered PPAs, we also require negative pairs which are not related by any PPA. We generate such negative examples by randomly sampling $10 \cdot |\mathrm{PID}|$ pairs, where $|\mathrm{PID}|$ is the number of pairs obtained from PID.

As a text corpus, we use all 24 377 760 *PubMed* abstracts available through PubTator Central (ftp://ftp.ncbi.nlm.nih.gov/pub/lu/PubTatorCentral/bioconcepts2pubtatorcentral.offset.gz, Version of 2019/08/19) (Wei *et al.*, 2019) and 2 986 273 full texts available in the PubmedCentral BioC text mining collection (ftp://ftp.ncbi.nlm.nih.gov/pub/wilbur/BioC-PMC/, Version of 2019/05/24) (Comeau *et al.*, 2019). We use the NER and NEN annotations from PubTator Central for both abstracts and full texts. We transform the Entrez ids provided by PubTator Central to Uniprot identifiers with the mapping provided by Uniprot (ftp://ftp.uniprot.org/pub/databases/uniprot/current_release/knowledgebase/idmapping/by_organism/HUMAN_9606_idmapping.dat.gz) to relate them to the Uniprot identifiers from PID. Additionally, we expand the identified proteins with all homologous proteins obtained from the HomoloGene

Table 1. Statistics of the datasets BioNLP 2011, BioNLP 2013 and PID

	Relations						Pairs		Texts (Avg.)	
	expr.	phosph.	State	Transport	Complex	Total	pos.	neg.	pos.	neg.
BioNLP 2011	245	44	136	38	278	741	615	1845	19.69	4.97
BioNLP 2013	179	104	160	43	441	927	730	2190	17.44	4.85
PID	2376	2714	8425	1020	5799	20 622	16 369	54 261	53.60	16.32

Note: Relations gives the total number of protein pairs for the five considered relations *controls-expression-of* (expr.), *controls-phosphorylation-of* (phosph.), *controls-state-change-of* (state), *controls-transport-of* (transport) and *in-complex-with* (complex). Pairs denote the total number of protein pairs with at least one relation (pos.) and without any relation (neg.). Texts states the average number of text spans per protein pair for pairs with at least one relation (pos.) and without any relation (neg.).

database (ftp://ftp.ncbi.nih.gov/pub/HomoloGene/build68/homologene.data), to increase the number of text spans per protein pair, considering only the taxa *Homo Sapiens*, *Rattus norvegicus*, *Mus musculus*, *Oryctolagus cuniculus* and *Cricetulus longicaudatus*. For protein pairs which occur together in more than 100 texts, we randomly sample 100 texts and discard the rest. Finally, we discard all (positive and negative) protein pairs which did not co-occur at least once. Detailed statistics of the resulting dataset can be found in Table 1.

Next, we describe the generation of the directly supervised data, which we need for two different purposes. First, we use it as additional training data as explained above and second, it allows us to perform experiments with known relations for text spans, which then lets us evaluate the performance for evidence prediction without manual inspection of the predictions. To perform these experiments, we actually need *two* distinct directly supervised datasets, one for evaluation and one as additional training data for PEDL. To generate the directly supervised data, we transform sentence-level event extraction data from the BioNLP-shared tasks (Kim *et al.*, 2011a; Nédellec *et al.*, 2013) into multi-instance learning data. We transform the BioNLP event structures into pairwise relations between proteins with the same five relation types as for the *distant* data. The details of this transformation can be found in Supplementary Material S3. Then, akin to the generation of the *distant* data, we normalize all protein mentions, collect all pairs of co-occurring proteins and sample non-interacting proteins as negative examples. We normalize protein mentions by querying MyGeneInfo (Xin *et al.*, 2016) for the human uniprot id. Tokenization and sentence splitting are performed with the *en_core_sci_sm* model of SciSpacy (Neumann *et al.*, 2019). We perform this transformation for the *Genia* (Kim *et al.*, 2011b) and *epigenetics* (Ohta *et al.*, 2011) datasets from *BioNLP 2011* as well as the *Genia* (Kim *et al.*, 2013) and *Pathway Curation* (Ohta *et al.*, 2013) tasks from *BioNLP 2013*. These BioNLP datasets were specifically selected since they were the only ones containing annotations for all considered PPA types. Finally, we aggregate the protein pairs of both 2011 and 2013 tasks, respectively. This yields two multi-instance learning datasets with the additional information of which text spans express relations between the proteins. Detailed statistics of both datasets can be found in Table 1.

In preliminary experiments on the *PID* dataset, we found that the predictions of PEDL seemed to almost exclusively rely on the protein names appearing in the text span. While this led to good performance for relation prediction, this is most likely an artefact of the PID database, because if two proteins are related by a given PPA, then frequently, all members of the respective protein families are related by the same PPA. Ultimately, we are interested in predicting PPAs that are *not* contained in PID, and thus, we performed all further experiments on entity blinded data, which prevents PEDL from inferring family membership. To achieve this, we replaced all protein names recognized by the *en_ner_jnlpba_md* model of SciSpacy with dummy identifiers.

2.3 Baselines

We compare PEDL to the two competitor methods *comb-dist* (Beltagy *et al.*, 2019a) and *EVEX* (Van Landeghem *et al.*, 2013), representing the state-of-the-art for distantly supervised relation extraction (*comb-dist*) and for sentence-level relation extraction applied on whole PubMed (*EVEX*).

comb-dist is a recently published multi-instance learning method for distantly supervised relation extraction. It set a new state-of-the-art on a standard benchmark for distantly supervised relation extraction (Riedel *et al.*, 2010) strongly outperforming competitor methods by additionally integrating directly supervised data. As a base model, comb-dist uses a piece-wise convolutional neural network with selective attention and pre-trained word embeddings. comb-dist was not developed for biomedical applications and has never been applied to such data as far as we know. In all experiments with comb-dist, we use the (selective) attention distribution over the text spans as evidence predictions. A detailed discussion of the differences between PEDL and comb-dist is provided in Supplementary Material S6. For word embeddings, we equip comb-dist with *wikipedia-pubmed-PMC* (http://evexdb.org/pmresources/vec-space-models/wikipedia-pubmed-and-PMC-w2v.bin) embeddings of Pyysalo *et al.* (2013), because they performed well in our earlier work (Habibi *et al.*, 2017). The hyperparameter settings of comb-dist for each task are provided in Supplementary Material S2.

EVEX is a database of text-mined biological events, accompanied by inferred pairwise PPAs and has annotations for whether an event was deemed speculative or negated. The database was created by applying a state-of-the-art biomedical event extraction tool (Björne, 2014) to a large collection of PubMed abstracts and PMC full texts. Since the EVEX database was last updated in 2013, we compare PEDL with EVEX on a modified test data of PID in which we only use texts published prior to 2013 to ensure a fair comparison. We apply a straight-forward mapping of EVEX's types of PPAs to the five considered in our work (see Supplementary Material S4) and remove all relations with a detected negation, but retain speculative relations.

2.4 Evaluation details

We use the three datasets *PID*, *BioNLP 2011* and *BioNLP 2013* in three different experimental settings *E1*, *E2* and *E3*.

- *E1*: *PID* is the distantly supervised data and the union of both BioNLP datasets are the directly supervised auxiliary training data.
- *E2*: *BioNLP 2011* is the distantly supervised data (disregarding evidence annotations during training) and *BioNLP 2013* is the directly supervised auxiliary training data.
- *E3*: *BioNLP 2013* is the distantly supervised data and *BioNLP 2011* is the directly supervised auxiliary training data.

In both *E2* and *E3*, we report the average of five runs with different seeds to compensate for the small dataset sizes. Note that results from the BioNLP shared tasks are not comparable to *E2* and *E3* because we do perform multi-instance learning (and not sentential prediction) and the label spaces are different. We use the directly supervised data only during training and remove all protein pairs occurring in the development and test set from the directly supervised data to prevent knowledge leaks. We split each dataset into train, development and test set by randomly dividing protein pairs with their associated text in a 60:10:30 ratio. For relation

prediction, we compare models by plotting their precision–recall (PR) curves. These curves are computed by ranking all PPAs by the predicted confidence score of the model and computing the resulting (micro-averaged) precision and recall for all possible threshold values. We also report the average precision (AP) which is an approximation of the area under the PR-curve. We use mean average precision (mAP) and precision at ten (P@10) to evaluate evidence predictions, both for the automated evaluation in *E2* and *E3*, as well as for the manual evaluation by domain experts in *E1*. mAP averages the individual APs of evidence predictions for each protein pair and P@10 is defined the mean precision of the top ten predictions.

3 Results

We evaluate PEDL, a method for predicting PPA-relations between proteins and the evidence for these relations, on three different datasets. The results are compared to two competitor methods: comb-dist, a recently published state-of-the-art multi-instance relation extraction method, and EVEX, a large data base of PPAs that was generated by applying biomedical event extraction to a large collection of abstracts and full texts.

Table 2. Results on the two BioNLP datasets (E2 and E3)

	BioNLP '11		BioNLP '13	
	r-AP	e-mAP	r-AP	e-mAP
comb-dist	65.4(2.6)	75.86(1.6)	70.68(2.6)	79.35(0.9)
– direct	62.33(1.8)	54.38(26.9)	70.06(2.1)	54.64(27.2)
PEDL	65.59(4.9)	82.36(1.2)	76.75(2.0)	84.67(1.6)
– direct	60.65(4.1)	64.64(4.1)	71.03(3.0)	75.14(2.1)

Note: r-AP is the AP for relation prediction and e-mAP the mAP for evidence prediction. All results are averages of five runs with different random seeds, with standard deviations given in brackets. '- direct' shows scores without directly supervised data. The best scores are displayed in bold.

Table 3. APs for relation prediction on the PID data (E1) for the PPA types *controls-expression-of* (expr.), *controls-phosphorylation-of* (phosph.), *controls-state-change-of* (state), *controls-transport-of* (transport) and *in-complex-with* (complex)

	expr.	phosph.	State	Transport	Complex	Total
comb-dist	42.77	38.38	49.14	5.87	47.86	44.78
PEDL	**46.45**	**40.26**	**52.70**	**18.21**	**49.70**	**46.02**
count	694	817	2532	288	1668	5999

Note: Total gives the AP for all PPA types as a micro-average. The best score per relation-type is displayed in bold. Count denotes the number of protein pairs with this type of PPA in the test set. Note that total is computed on a ranking of predictions including all PPA types, which leads to the fact that the difference between both models is smaller than every distance of the individual PPAs. EVEX cannot be compared in this setting, because it does not consider texts published after 2013.

3.1 Prediction of PPAs

At first, we investigate the results of PEDL for predicting PPAs between pairs of proteins. The results for the BioNLP datasets (E2 and E3) can be found in Table 2 and results for PID (*E1*) in Table 3. In terms of AP, PEDL performs better than the competitor methods on two of the three considered datasets and comparable on the third. On BioNLP 2013 (*E3*), PEDL achieves an AP score that is 6.07 pp higher than that of comb-dist, while on PID (*E1*, mixing predictions for all PPA types) it is 1.24 pp higher. If one considers predictions for each type of PPA on PID individually, the difference between both models is considerably larger. PEDL performs better than comb-dist on all five types with differences ranging from 1.84 pp for *in-complex-with* to 12.34 for *controls-transport-of*, with an average of 4.66 pp. On BioNLP 2011 (*E2*), the difference in AP of both models is marginal.

It is instructive to compare the PR-curves of PEDL, comb-dist and EVEX for relation prediction on the PID data (E1, see Fig. 2). We compare with the results of EVEX only on abstracts and full texts published prior to 2013 to account for the fact that EVEX was last updated in 2013. Both models strongly outperform EVEX on the *before 2013* data, both in terms of recall and precision. The difference in recall is especially pronounced, because EVEX only generates positive predictions for fewer than 37% of the PPAs in the PID test set. PEDL performs better than comb-dist in the mid-precision regime but a little worse for low precisions when provided all articles and full texts. On the *before 2013* subset, PEDL performs equal to comb-dist in the high-precision regime but worse for mid-to-low precision values, leading to 40.54% AP for PEDL and 44.24% AP for comb-dist (see Section 4.2 for a discussion of this).

3.2 Evidence prediction

In most biomedical applications, extracted PPAs are not accepted per se, but undergo confirmation through experts. The reason is the far from perfect performance of state-of-the-art approaches, and the fact that even a correctly extracted text needs not express biological truth, for instance due to weak experimental evidence. Therefore, it is important that methods not only predict the correct PPA, but also the text spans on which the model's PPA prediction is based (which we call evidence prediction). Table 2 gives results for evidence prediction on the BioNLP datasets, where both PEDL and comb-dist achieve high mAP scores for evidence prediction. PEDL outperforms comb-dist on both datasets with 6.38 pp for BioNLP '11 and 5.32 pp for BioNLP'13.

In contrast, PID is a distantly supervised dataset and does not have annotations to evaluate evidence predictions. For the predictions of comb-dist and PEDL, two domain experts evaluated the top ten evidence predictions for the top 10 predictions of each PPA-type, amounting to 500 evaluated evidence predictions (the annotation

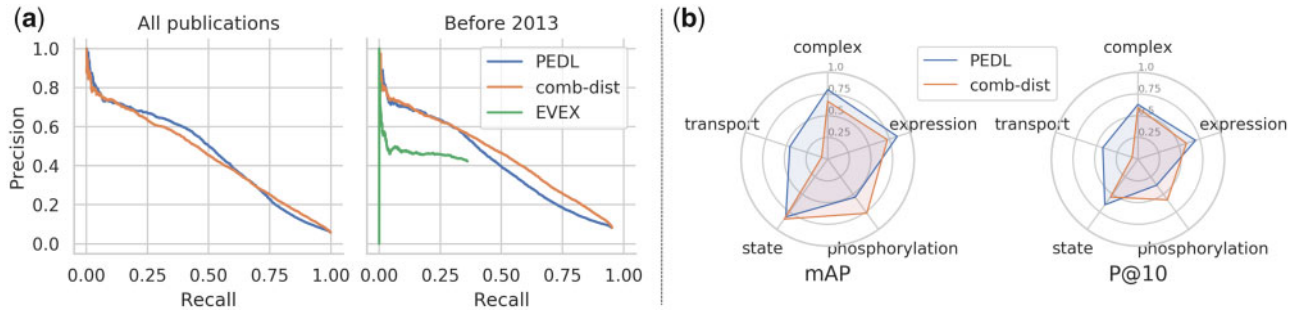

Fig. 2. (**a**) PR curve for the PID data. The left plot shows results for all available abstracts and full texts. The right plot displays the results using only abstracts and full texts published prior to 2013, which allows a fair comparison with EVEX. These results are based on a ranking that includes all types of PPA. The improvement of PEDL over comb-dist is larger for rankings of only one type of PPA (see Table 3 for numbers and explanation). (**b**) Results from the manual evaluation of evidence prediction on PID

guidelines can be found in Supplementary Material S5). Note, that for this evaluation, we directly use the rows of the score matrix as evidence score per relation for PEDL. This refinement is not possible for comb-dist, because the attention distribution is computed independently of the relation type. This allows PEDL to rank the evidence specifically for one PPA type, while comb-dist only predicts whether there is a relation between the proteins at all. The results of this analysis show that PEDL performs better than comb-dist for predicting evidence for the three PPA-types *controls-transport-of*, *in-complex-with* and *controls-expression-of* (see Fig. 2). The results for *controls-state-change-of* are comparable and worse for *controls-phosphorylation-of*. The improvement over comb-dist is especially striking in the case of *controls-transport-of*, for which comb-dist produces almost no correct evidence predictions and PEDL achieves a mAP of 46%. The results in terms of P@10 are similar, with PEDL additionally achieving better results for *controls-state-change-of*. Moreover, the variability in performance across different PPA types is much larger for comb-dist than for PEDL. On average, PEDL achieves a 7.66 pp higher mAP and a 8.14 pp higher P@10 than comb-dist.

3.3 Analysis of new predictions

We also evaluated PEDL in a realistic application scenario, where three experts in systems biology manually analyzed the top 10 predictions that are not contained in the aforementioned PathwayCommons versions of neither Reactome nor PID. The results are summarized in Table 4, where we provide all predictions considered biologically justified by all experts together with the highest ranking true evidence text span. In the evaluation, 6 out of 10 are predicted correctly, while one prediction is wrong due to errors in the protein normalization pre-processing step, and the other three are errors of PEDL. It can be further observed, that for all correct predictions but one, the highest ranking text span (columns *Text span* and *t*) actually expresses the PPA and either states the finding of the PPA or refers to an earlier publication reporting it.

4 Discussion

4.1 Importance of directly supervised data

The results given in Table 2 allow for interesting observations regarding the importance of directly supervised data. On the BioNLP datasets, the incorporation of directly supervised data improves results for both relation and evidence prediction. The improvement is much more pronounced for the evidence prediction task than for relation prediction. This supports our hypothesis, that we can improve evidence prediction specifically by including directly supervised data. Compared to comb-dist, PEDL has a much larger gain from directly supervised data in the relation prediction task (5.33 pp versus 1.85 pp). For BioNLP 2011, comb-dist even outperforms PEDL in relation prediction when no directly supervised data is available. This might partly be because the inclusion of directly supervised data stabilizes PEDL's training process. In preliminary experiments on the PID dataset, we observed that without access to directly supervised data the model failed to converge, while setting the whole score matrix to zero. We attribute this to the fact that usually only a few of the (max.) 100 text spans actually express the annotated relation and think that the directly supervised data compensates for the resulting label imbalance for evidence prediction.

Notably, PEDL achieves strong results for evidence prediction even without access to directly supervised data. This suggests that the constraint of only aggregating (logit-)scores, and not high-dimensional embeddings as in comb-dist's selective attention, is more appropriate for evidence prediction in absence of directly supervised data. These scores also have a clear interpretation as the confidence of PEDL that the given text span supports a given PPA. The lower (average) performance of comb-dist in this setting can be attributed to strong performance drops for some random seeds (min. 24.03 versus max. 76.61), indicating a notable instability of the model. We furthermore found that running comb-dist with the most recent versions of PyTorch (1.4.0) and AllenNLP (0.9.0) leads to a performance drop of 1 to 5 pp. for both relation prediction and evidence prediction.

4.2 Comparison to EVEX

The comparison of the two distantly supervised methods to EVEX (cf. Fig. 2) is instructive, because it allows to compare methods trained only on directly supervised data to models with access to both types of data. Especially striking is the difference in recall between EVEX and the distantly supervised models, where EVEX only contains predictions for 36.15% of the positive protein pairs, while PEDL and comb-dist produce predictions for 95.1% and 95.33% of the protein pairs. This might be partially attributed to the advancements in NER and Normalization that were achieved since 2013—which we implicitly incorporate by using PubTator Central—but also stresses the importance of predicting relations for proteins that occur in different sentences. Recall that EVEX only considers single sentences.

The importance of using multiple sentences will be further discussed in the next section. Notably, the increased recall does not come at the price of reduced precision, as both PEDL and comb-dist strongly outperform EVEX in all precision regimes. Together with the encouraging results of the evidence prediction, this indicates that distant supervision is a promising paradigm to train accurate classifiers for PPA prediction.

A related interesting observation is that PEDL performs markedly worse on the *before 2013* subset of the data, whereas comb-dist almost retains its performance. We hypothesized that the reason for this behaviour lies in the fact that PEDL does not model the semantic interactions between text spans via attention, making it more

Table 4. Evaluation results for the top-10 predictions that cannot be found either in Reactome or in PID

k	PPA	Text span (source PMID)	t	Evidence
1	IGF-II *in-complex-with* VN	'We have previously reported that IGF-II binds the extracellular matrix protein vitronectin (VN) [. . .] ' (12746303)	1	Upton *et al.* (1999)
2	hnRNP-A1 *controls-expression-of* IL10	'These results suggest that hnRNP-A1 promotes transcription of human IL10.' (19349988)	1	Noguchi *et al.* (2009)
4	NCOR1 *controls-expression-of* PSA	'ChIP-reChIP assays revealed that NCOR and [. . .] p300 are present in distinct AR complexes on the promoter of PSA gene [. . .]' (23518348)	4	Qi *et al.* (2013)
5	ets-2 *controls-expression-of* BRCA1	'Conditional overproduction of ets-2 in MCF-7 cells resulted in repression of endogenous BRCA1 mRNA expression.' (12637547)	1	Baker *et al.* (2003)
6	c-Rel *controls-expression-of* Bcl-X	'We further demonstrate [. . .] that introduction of two downstream c-Rel target genes, Bcl-X [. . .]' (15922711)	1	Chen *et al.* (2000)/ Lee *et al.* (1999)
8	C/EBP-beta *controls-expression-of* COX-2	'C/EBP-beta is a transcription factor [. . .] capable of inducing COX-2 expression [. . .]' (19124115)	1	Kim and Fischer (1998)/ Zhu *et al.* (2002)

Note: The rank of the prediction is given by *k*. We provide the highest ranking evidence text span that actually expresses the relation and its rank in PEDL (*t*), as well as manually sourced literature evidence that provides strong biological evidence for the existence of the PPA. Note that this evidence need not be identical to the evidence span predicted by the model.

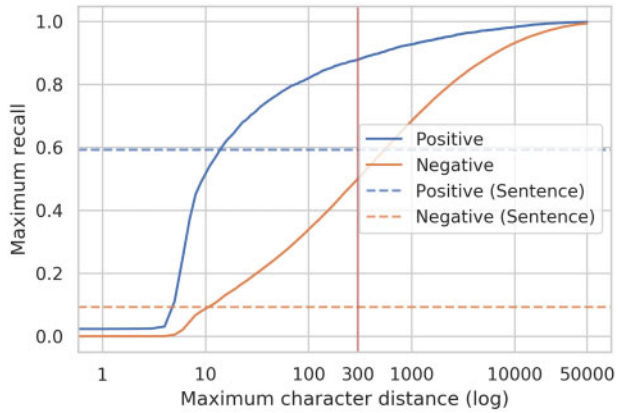

Fig. 3. Maximum possible recall for a given maximum character distance between the protein mentions. 'Positive' refers to protein pairs with at least one PPA in PID and 'Negative' to pairs without any. The dashed lines indicate the maximum recall that is possible for sentence level approaches. The red vertical line indicates our choice for the maximum distance between pairs

susceptible to violations of the at-least-once assumption. To validate this, we inspect the top 10 predictions of PEDL for true PPAs with the largest drop in ranking between the full and the *before 2013* data. We found that for nine of the ten PPAs, none of the texts published prior to 2013 contains any mention of the PPA. Additionally, no text published prior to 2013 contained any mention of the associated protein pair for 5% of all true PPAs, which limits PEDL's maximum recall to 95% for the *before 2013* data.

4.3 Importance of using multiple sentences

We investigate the effect of considering protein mentions across sentences by measuring the fraction of protein-pairs in PID that are at most d characters away from each other in at least one text for different values of d. Additionally, we report this quantity considering only single sentences, again using the *en_core_sci_sm* model of SciSpacy to split the text into sentences. The results are depicted in Figure 3. It can be observed that considering only protein mentions that occur within the same sentence has a strong limiting effect on maximum recall. Using $d = 300$, PEDL can predict PPAs for 87.9% of the positive pairs in PID, which is a large gain over the 59.24% that would be achievable if we considered only single sentences. This, however, comes at the price of more included negative protein pairs. PEDL predicts PPAs for 50.01% of the considered negative pairs, whereas a sentence-level approach would predict PPAs for only 9.25%. This highlights the importance of using a strong machine learning model to rank the predicted PPAs instead of relying on simple co-occurrence statistics in the high-recall regime.

5 Conclusion

We propose PEDL, a method for predicting PPAs and their textual evidence by integrating deep language models, distant supervision and auxiliary directly supervised data. We compare PEDL on three different datasets with two state-of-the-art methods and find that, on average, it outperforms them in most cases and performs comparably in the remaining ones. A manual evaluation of the predicted PPAs shows that PEDL can be used to identify PPAs that are missing in major pathway data bases. Furthermore, we demonstrate that the predicted evidence text spans actually express the relation and thus can be used to quickly verify the predicted PPAs.

Owing to the incorporation of BERT, the method proposed in this article has very high runtime requirements which make it unsuitable for predicting PPAs between all possible pairs. This problem could be solved using recently published model distillation techniques for BERT (Sanh *et al.*, 2019). We only address pairwise PPA prediction in which a relation holds between exactly two proteins. Actual biochemical reactions are much more complex than that, as they can have multiple reactants, products and regulators, which can also be protein complexes or completely different molecules (Berg *et al.*, 2019). It would be worthwhile to study whether biomedical event extraction (Ohta *et al.*, 2013) can be combined with distant supervision to predict such complex biochemical reactions. Finally, PEDL could also be used to predict evidence for known PPAs, for instance those from the distantly supervised training data, which we did not investigate in this work.

Acknowledgements

The authors thank Mareike Simon for manual evaluation of the newly predicted PPAs. L.W. acknowledges the support of the Helmholtz Einstein International Berlin Research School in Data Science (HEIBRiDS). The authors acknowledge the support of NVIDIA Corporation with the donation of the Titan X Pascal GPU used for this research.

Funding

This work was supported by the Helmholtz Society through the research training group HEIBRIDS. J.W. acknowledges funding by the German Federal Ministry of Education and Research BMBF [e:med project 031L0189D].

Conflict of Interest: none declared.

References

Angeli,G. *et al.* (2014) Combining distant and partial supervision for relation extraction. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Association for Computational Linguistics, Doha, Qatar, pp. 1556–1567.

Baker,K.M. *et al.* (2003) Ets-2 and components of mammalian SWI/SNF form a repressor complex that negatively regulates the BRCA1 promoter. *J. Biol. Chem.*, **278**, 17876–17884.

Barabasi,A.-L. and Oltvai,Z.N. (2004) Network biology: understanding the cell's functional organization. *Nat. Rev. Genet.*, **5**, 101–113.

Beltagy,I. *et al.* (2019a) Combining distant and direct supervision for neural relation extraction. 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)*, Association for Computational Linguistics, Minneapolis, Minnesota, pp. 1858–1867.

Beltagy,I. *et al.* (2019b) SciBERT: a pretrained language model for scientific text. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, Association for Computational Linguistics, Hong Kong, China, pp. 3613–3618.

Berg,J.M. *et al.* (2019) *Biochemistry*, 9th edn. WH Freeman, New York.

Björne,J. (2014) *Biomedical Event Extraction with Machine Learning*. Ph.D. thesis, University of Turku.

Björne,J. and Salakoski,T. (2018) Biomedical event extraction using convolutional neural networks and dependency parsing. In *Proceedings of the BioNLP 2018 Workshop*, Association for Computational Linguistics, Melbourne, Australia, pp. 98–108.

Björne,J. *et al.* (2009) Extracting complex biological events with rich graph-based feature sets. In *Proceedings of the BioNLP 2009 Workshop Companion Volume for Shared Task*, Association for Computational Linguistics, Boulder, Colorado, pp. 10–18.

Cerami,E.G. *et al.* (2011) Pathway Commons, a web resource for biological pathway data. *Nucleic Acids Res.*, **39**, D685–D690.

Chen,C. *et al.* (2000) The Rel/NF-kappaB family directly activates expression of the apoptosis inhibitor Bcl-x(L). *Mol. Cell. Biol.*, **20**, 2687–2695.

Comeau,D.C. *et al.* (2019) PMC text mining subset in BioC: about three million full-text articles and growing. *Bioinformatics*, **35**, 3533–3535.

Devlin,J. *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)*, Association for Computational Linguistics, Minneapolis, Minnesota, pp. 4171–4186.

Friedman,C. *et al.* (2001) GENIES: a natural-language processing system for the extraction of molecular pathways from journal articles. *Bioinformatics*, **17**, S74–S82.

Habibi,M. *et al.* (2017) Deep learning with word embeddings improves biomedical named entity recognition. *Bioinformatics*, **33**, i37–i48.

Hat,B. *et al.* (2016) Feedbacks, bifurcations, and cell fate decision-making in the p53 system. *PLoS Comput. Biol.*, **12**, e1004787.

Hoffmann,R. *et al.* (2011) Knowledge-based weak supervision for information extraction of overlapping relations. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, Association for Computational Linguistics, Portland, Oregon, USA, pp. 541–550.

Jassal,B. *et al.* (2019) The reactome pathway knowledgebase. *Nucleic Acids Res.*, **48**, D498–D503.

Junge,A. and Jensen,L.J. (2019) CoCoScore: context-aware co-occurrence scoring for text mining applications using distant supervision. *Bioinformatics,* **36**, 264–271.

Kim,J.-D. *et al.* (2011a) Overview of BioNLP shared task 2011. In *Proceedings of BioNLP Shared Task 2011 Workshop*, Association for Computational Linguistics, Portland, Oregon, USA, pp. 1–6.

Kim,J.-D. *et al.* (2011b) Overview of Genia event task in BioNLP shared task 2011. In *Proceedings of BioNLP Shared Task 2011 Workshop*, Association for Computational Linguistics, Portland, Oregon, USA, pp. 7–15.

Kim,J.-D. *et al.* (2013) The Genia event extraction shared task, 2013 edition – overview. In *Proceedings of the BioNLP Shared Task 2013 Workshop*, Association for Computational Linguistics, Sofia, Bulgaria, pp. 8–15.

Kim,Y. and Fischer,S.M. (1998) Transcriptional regulation of cyclooxygenase-2 in mouse skin carcinoma cells regulatory role of CCAAT/enhancer-binding proteins in the differential expression of cyclooxygenase-2 in normal and neoplastic tissues. *J. Biol. Chem.*, **273**, 27686–27694.

Kingma,D.P. and Ba,J. (2015) Adam: a method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7–9, 2015, Conference Track Proceedings.*

Köksal,A.S. *et al.* (2018) Synthesizing signaling pathways from temporal phosphoproteomic data. *Cell Rep.*, **24**, 3607–3618.

Lee,H.H. *et al.* (1999) NF-kappaB-mediated up-regulation of BCL-x and Bfl-1/A1 is required for CD40 survival signaling in b lymphocytes. *Proc. Natl. Acad. Sci. USA*, **96**, 9136–9141.

Lee,J. *et al.* (2019) BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, **36**, 1234–1240.

Lin,Y. *et al.* (2016) Neural relation extraction with selective attention over instances. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Association for Computational Linguistics, Berlin, Germany, pp. 2124–2133.

Liu,T. *et al.* (2017) A soft-label method for noise-tolerant distantly supervised relation extraction. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, Association for Computational Linguistics, Copenhagen, Denmark, pp. 1790–1795.

Mintz,M. *et al.* (2009) Distant supervision for relation extraction without labeled data. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, Association for Computational Linguistics, Suntec, Singapore, pp. 1003–1011.

Miwa,M. *et al.* (2009) A rich feature vector for protein–protein interaction extraction from multiple corpora. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, Association for Computational Linguistics, Singapore, pp. 121–130.

Miwa,M. *et al.* (2010) Event extraction with complex event classification using rich features. *J. Bioinf. Comput. Biol.*, **08**, 131–146.

Nédellec,C. *et al.* (2013) Overview of BioNLP shared task 2013. In *Proceedings of the BioNLP Shared Task 2013 Workshop*, Association for Computational Linguistics, Sofia, Bulgaria, pp. 1–7.

Neumann,M. *et al.* (2019) ScispaCy: fast and robust models for biomedical natural language processing. In *Proceedings of the 18th BioNLP Workshop and Shared Task*, Association for Computational Linguistics, Florence, Italy, pp. 319–327.

Noguchi,E. *et al.* (2009) A Crohn's disease-associated NOD2 mutation suppresses transcription of human IL10 by inhibiting activity of the nuclear ribonucleoprotein hnRNP-A1. *Nat. Immunol.*, **10**, 471–479.

Ohta,T. *et al.* (2011) Overview of the epigenetics and post-translational modifications (EPI) task of BioNLP shared task 2011. In *Proceedings of BioNLP Shared Task 2011 Workshop*, Association for Computational Linguistics, Portland, Oregon, USA, pp. 16–25.

Ohta,T. *et al.* (2013) Overview of the pathway curation (PC) task of BioNLP shared task 2013. In *Proceedings of the BioNLP Shared Task 2013 Workshop*, Association for Computational Linguistics, Sofia, Bulgaria, pp. 67–75.

Peng,Y. and Lu,Z. (2017) Deep learning for extracting protein–protein interactions from biomedical literature. In: *BioNLP 2017*, Association for Computational Linguistics, Vancouver, Canada, pp. 29–38.

Pershina,M. *et al.* (2014) Infusion of labeled data into distant supervision for relation extraction. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, Association for Computational Linguistics, Baltimore, Maryland, pp. 732–738.

Peters,M.E. *et al.* (2019) To tune or not to tune? adapting pretrained representations to diverse tasks. In *Proceedings of the 4th Workshop on Representation Learning for NLP (RepL4NLP-2019)*, Association for Computational Linguistics, Florence, Italy, pp. 7–14.

Poon,H. *et al.* (2014) Distant supervision for cancer pathway extraction from text. In: *Proceedings of the Pacific Symposium on Biocomputing, Kohala Coast, Hawaii, USA*, pp. 120–131.

Pratt,L.Y. *et al.* (1991) Direct transfer of learned information among neural networks. In *Proceedings of the Ninth National Conference on Artificial Intelligence - Volume 2*, AAAI'91. AAAI Press, Anaheim, USA, pp. 584–589.

Pyysalo,S. *et al.* (2008) Comparative analysis of five protein–protein interaction corpora. *BMC Bioinformatics*, **9**, S6.

Pyysalo,S. *et al.* (2013) Distributional semantics resources for biomedical text processing. In *Proceedings of LBM 2013*, Database Center for Life Science, Tokyo, pp. 39–44.

Qi,J. *et al.* (2013) The E3 ubiquitin ligase Siah2 contributes to castration–resistant prostate cancer by regulation of androgen receptor transcriptional activity. *Cancer Cell*, **23**, 332–346.

Quirk,C. and Poon,H. (2017). Distant supervision for relation extraction beyond the sentence boundary. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, Association for Computational Linguistics, Valencia, Spain, pp. 1171–1182.

Riedel,S. *et al.* (2010) Modeling relations and their mentions without labeled text. In: Balcázar,J.L. *et al.* (eds.) *Machine Learning and Knowledge Discovery in Databases*. Springer, Berlin, Heidelberg, pp. 148–163.

Sanh,V. *et al.* (2019) Distilbert, a Distilled Version of BERT: Smaller, Faster, Cheaper and Lighter. In: *The 5th Workshop on Energy Efficient Machine Learning and Cognitive Computing*, VancouverBC, Canada.

Schaefer,C.F. *et al.* (2009) PID: the Pathway Interaction Database. *Nucleic Acids Res.*, **37**, D674–679.

Surdeanu,M. *et al.* (2012) Multi-instance multi-label learning for relation extraction. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, Association for Computational Linguistics, Jeju Island, Korea, pp. 455–465.

Thomas,P. *et al.* (2011) Learning protein–protein interaction extraction using distant supervision. In *Proceedings of Workshop on Robust Unsupervised and Semisupervised Methods in Natural Language Processing*, Association for Computational Linguistics, Hissar, Bulgaria, pp. 25–32.

Tikk,D. *et al.* (2012) A detailed error analysis of 13 kernel methods for protein–protein interaction extraction. *BMC Bioinformatics, 14, 12.*

Upton,Z. *et al.* (1999) Identification of vitronectin as a novel insulin-like growth factor-II binding protein. *Endocrinology*, **140**, 2928–2931.

Van Landeghem,S. *et al.* (2013) Large-scale event extraction from literature with multi-level gene normalization. *PLoS One*, **8**, e55814–12.

Vaswani,A. *et al.* (2017) Attention is all you need. In: Guyon,I. *et al.* (eds.) *Advances in Neural Information Processing Systems*, *Annual Conference on Neural Information Processing Systems 2017*, Long Beach, CA, USA, **pp.** 5998–6008.

Verga,P. *et al.* (2018) Simultaneously self-attending to all mentions for full-abstract biological relation extraction. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, Association for Computational Linguistics, New Orleans, Louisiana, pp. 872–884.

Weber,L. *et al.* (2020) HUNER: improving biomedical NER with pretraining. *Bioinformatics*, **36**, 295–302.

Wei,C.-H. *et al.* (2019) PubTator central: automated concept annotation for biomedical full text articles. *Nucleic Acids Res.*, **47**, W587–W593.

Wu,Y. *et al.* (2016) Google's neural machine translation system: Bridging the gap between human and machine translation. *preprint arXiv:1609.08144.*

Xin,J. *et al.* (2016) High-performance web services for querying gene and variant annotation. *Genome Biol.*, **17**, 91.

Zeng,D. *et al.* (2015). Distant supervision for relation extraction via piecewise convolutional neural networks. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, Association for Computational Linguistics, Lisbon, Portugal, pp. 1753–1762.

Zhu,Y. *et al.* (2002) Dynamic regulation of cyclooxygenase-2 promoter activity by isoforms of CCAAT/enhancer-binding proteins. *J. Biol. Chem.*, **277**, 6923–6928.