# Humboldt @ DrugProt: Chemical-Protein Relation Extraction with Pretrained Transformers and Entity Descriptions

Leon Weber, Mario Sänger, Samuele Garda, Fabio Barth, Christoph Alt, Ulf Leser
Computer Science Dept., Humboldt Universität zu Berlin, Berlin, Germany

***Abstract*—The detection of chemical-protein interactions is an important task with applications in drug design and biotechnology. The BioCreative VII - DrugProt shared task provides a benchmark for the automated extraction of such relations from scientific text. This article describes the Humboldt approach to solving it. We define the task as a relation classification problem, which we model with pretrained transformer language models and further use entity descriptions as an additional knowledge source. On the hidden test set of DrugProt, our model achieves 79.73% F1, yielding an improvement of over 17pp over the average score of all task participants.**

***Keywords*—relation extraction; transformers; entity descriptions**

## I. Introduction

With the rapid growth of biomedical literature, it is becoming increasingly difficult to obtain comprehensive information on any specific entity by only reading. One of the most important aspects of drugs are their interactions with other biomedical molecules, especially genes and proteins. Recognizing drug-protein relationships is crucial in various applications such as drug discovery (1), precision medicine (2), and curation of biomedical databases (3). Manual extraction of such relationships from the biomedical literature is costly and often prohibitively time-consuming. Alternatively, information extraction can help to automatically identify these relationships and make them more readily accessible. Extracting (biomedical) relationships from text has been investigated intensively over the last two decades (4). Methods employed hand-crafted features based on lexical or syntactic information (5), kernel-based learning (6), or various forms of neural networks (7-9). Most recently, a variety of approaches utilizing pretrained (transformer-based) language models have been introduced and achieved new state of the art performance across several domains (10, 11, 15). A plethora of approaches explored methods of enriching the training data. For example, Vashishth et al. (8) proposed a distantly-supervised method which applies Graph Convolution Networks to encode syntactic information from text and utilizes additional knowledge base data for improved relation extraction. Yuan et al. (16) extracts entities from PubMed abstracts and link them to UMLS to train an entity- and knowledge-aware language model.

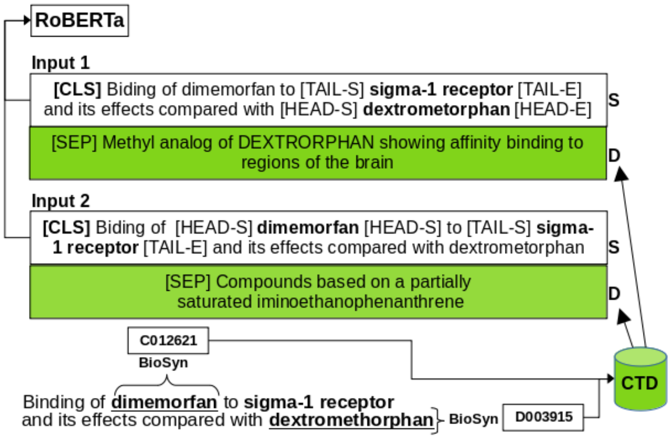
RoBERTa
Input 1
[CLS] Biding of dimemorfan to [TAIL-S] sigma-1 receptor [TAIL-E] and its effects compared with [HEAD-S] dextrometorphan [HEAD-E]
S
[SEP] Methyl analog of DEXTRORPHAN showing affinity binding to regions of the brain
D
Input 2
[CLS] Biding of [HEAD-S] dimemorfan [HEAD-S] to [TAIL-S] sigma-1 receptor [TAIL-E] and its effects compared with dextrometorphan
S
[SEP] Compounds based on a partially saturated iminoethanophenanthrene
D
C012621
BioSyn
Binding of dimemorfan to sigma-1 receptor and its effects compared with dextromethorphan
BioSyn
D003915
CTD

Fig. 1. Visual description of our approach. The model receives one example per valid entity pair in each sentence (S) enriched with chemical descriptions (D) derived from the CTD database. *[HEAD-S]* and *[HEAD-E]* mark start and end of the current head entity and *[TAIL-S]* and *[TAIL-E]* start and end of the current tail entity. Chemical mentions are linked with an inhouse BioSyn model.

Since 2003, the BioCreative[1] initiative hosts challenges to foster the development and evaluation of text mining approaches in the biomedical domain and has hosted a successful shared task on chemical-protein relation extraction before (17). Track 1 (DrugProt) of the 2021 BioCreative VII challenge (18) explores the recognition of chemical-protein relations in scientific abstracts. The organizers compiled a manually annotated corpus of abstracts labeled with all chemicals and gene/protein mentions as well as binary relationships between them, categorized into 13 different types of interactions. Participants of the challenge were asked to develop methods which, given the abstract text and annotations of the mentioned chemicals and genes/proteins, detect all binary relations and their type. In this paper, we describe the Humboldt contribution to the challenge. We define the task as a relation classification problem, which we model with pretrained transformer language models and use entity

[1] https:// biocreative.bioinformatics.udel.edu/

descriptions as an additional knowledge source. Our code and model are publicly available.[2]

## II. Method

### A. Chemical-Protein Relation Extraction as Relation Classification

We model chemical-protein relation extraction as sentence-level relation classification. To this end, we first split the abstract into sentences using *segtok*[3]. Next, we generate one example for each chemical-protein mention pair, that co-occurs in the same sentence. We mark the head entity (chemical) and the tail entity (gene) by inserting marker tokens into the sentence and then treat the task as a sentence classification problem. For an example, see Figure 1, where *[HEAD-S]* and *[HEAD-E]* mark start and end of the current head entity and *[TAIL-S]* and *[TAIL-E]* start and end of the current entity tail. For classifying the resulting example, we embed the text with a RoBERTa-large model (19). Then, we take the embedding of the *[CLS]* token to which we apply dropout (20). Finally, we feed the resulting embedding through a linear layer to arrive at our predictions. We use a cross entropy loss for training and initialize the model by using the weights from the RoBERTa-large-PM-M3-Voc model[4] of Lewis et al. (21), which was trained on the union of 22 million PubMed abstracts, 3.4 million PMC full texts and data from 60 thousand MIMIC-III reports. Finally, we ensemble our models by training ten models with different random seeds and then average the predicted probabilities for a given example. We train our model on the union of training and development set which increases the size of the total training data from 17,274 to 21,035 relations. We implement our model with the huggingface transformers framework[5].

### B. Entity Descriptions

We hypothesized that enriching the input with external textual descriptions of the head and tail entities could provide additional useful information to the model. For instance, the given chemical might inhibit the family of proteins to which the tail belongs or the protein in question may catalyze a reaction in which the chemical is involved. Such information can be found in chemical / protein databases, often in the form of text. We found in preliminary experiments on the development set that providing only a description of the chemical led to the largest improvement. Thus, we enriched the input only with chemical descriptions, which we created by gathering the first sentence of the Definition field of the CTD (22) chemicals vocabulary[6]. To match these descriptions with the entity mentions in the to-be-analyzed texts, we perform Named Entity Normalization (NEN) using BioSyn (23), the state-of-the-art method for this task. We train the BioSyn model with its default hyper-parameters for 20 epochs on the train and test split of the BioCreative V CDR dataset (24) and use it to link every mention to its CTD identifier. If the predicted chemical identifier has no associated CTD definition, we use the definition of the chemical's parent in the CTD hierarchy. This allows us to assign a description to every chemical in the challenge data set. For an example of chemical descriptions, see Figure 1.

### C. Hyperparameters

We select hyperparameters by performing an exhaustive grid search on the development set for the following values (best are marked bold):

- Learning rate: 5e-5, **3e-5**, 1e-5, 5e-6
- Epochs: 1, **3**, 5, 10

We use Adam (25) with a linear decay learning rate schedule and 10% warmup (26). We set the maximum length to 256 subword tokens and first truncate the chemical description before truncating the input sentence. The dropout rate is set to 0.1.

## III. Results

### A. Description of the submitted runs

We submitted five different runs for the shared task evaluation:

- **run-1**: Ensemble of ten differently seeded RoBERTa-large-PM-M3-Voc models trained on the union of training and development set with entity descriptions.
- **run-2**: Single RoBERTa-large-PM-M3-Voc trained on the union of training and development set with entity descriptions.
- **run-3**: Ensemble of ten differently seeded RoBERTa-large-PM-M3-Voc models trained on the union of training and development set without entity descriptions.
- **run-4**: Single RoBERTa-large-PM-M3-Voc model trained on the union of training and development set without entity descriptions.
- **run-5**: Ensemble of ten differently seeded RoBERTa-large-PM-M3-Voc models with entity descriptions trained on only the training set.

### B. Main results

Table 1 shows the main results on the test set of our submitted runs. Our best performing model is *run1*, an ensemble of ten differently seeded RoBERTa-large-PM-M3-Voc models with CTD chemical descriptions. It achieves a micro-averaged F1 score of 79.73%. When compared to the average score of all DrugProt participants of 61.96% this corresponds to an improvement of 17.77 percentage points (pp). Ablating the entity descriptions (*run-3*) leads to a

[2] https://github.com/leonweber/drugprot

[3] https://github.com/fnl/segtok

[4] https://github.com/facebookresearch/bio-lm

[5] https://github.com/huggingface/transformers

[6] http://ctdbase.org/reports/CTD_chemicals.csv.gz

TABLE I. RESULTS ON DRUGPROT TEST SET

| | *Precision (%)* | *Recall (%)* | *F1 (%)* |
|---|---|---|---|
| run-1 | 79.61 | 79.86 | **79.73** |
| run-2 | 76.25 | **80.49** | 78.31 |
| run-3 | **81.51** | 76.53 | 78.94 |
| run-4 | 76.16 | 80.00 | 78.03 |
| run-5 | 79.15 | 79.83 | 79.49 |

decrease of 0.79 pp F1, while taking only the best single model with entity descriptions instead of the 10x ensemble (*run-2*) leads to a decrease in F1 of 1.42 pp. An ablation of both ensembling and chemical descriptions leads to 1.7 pp lower F1 (*run-4*). When trained only on the training data without the addition of the development data, the F1 score of the ensemble decreases by 0.24 pp (*run-5*). We therefore conclude that both ensembling and entity descriptions have a positive effect on accuracy and that improvement is more pronounced for ensembling. These findings are further supported by our experiments on the development set for which the results are summarized in Table II. In this setting, in which we trained our model on the training set and evaluated it on the development set, ensembling leads to a gain of 0.9 pp F1 and ablating the entity descriptions causes a drop of 0.9 pp in F1. Surprisingly, using the development set as additional training data led to only a very modest gain even though it increased the size of the training data by over 20%. This might indicate that the amount of training data is not the only limiting factor.

### C. *Results by Relation Type*

Table 2 shows the results of our best submission (*run-1*) for each relation type. There is strong variability across different relation types with three relation types having an F1 score of zero, while the maximum F1 score is above 91%. The F1 scores correlate strongly with the number of training instances per relation type (Pearson's R 0.56). All three relation types with an F1 score of zero have very few training examples (10 to 27). However, for the other classes there seem to be additional factors influencing performance. For instance, the *Substrate* relation type has 2,003 training examples, but the model achieves an F1 score of only 68.18%. We leave a more detailed error analysis for future work.

TABLE II. RESULTS ON DRUGPROT DEVELOPMENT SET

| | *Precision (%)* | *Recall (%)* | *F1 (%)* |
|---|---|---|---|
| Best single model | 78.9 | 79.5 | 79.2 |
| Single model without entity descriptions | 77.1 | 79.6 | 78.3 |
| 10x Ensemble | **80.4** | **79.7** | **80.1** |

TABLE III. DETAILED TEST SET RESULTS FOR RUN-1

| | *Precision (%)* | *Recall (%)* | *F1 (%)* | *# instances in train + dev* |
|---|---|---|---|---|
| Activator | 83.23 | 80.24 | 81.71 | 1,674 |
| Agonist | 85.11 | 79.21 | 82.05 | 789 |
| Agonist-Inhibitor | 0.00 | 0.00 | 0.00 | 15 |
| Antagonist | 87.95 | 95.42 | 91.54 | 1,190 |
| Direct-Regulator | 75.82 | 70.16 | 72.88 | 2,705 |
| Indirect-Downregulator | 74.93 | 84.54 | 79.44 | 1,661 |
| Indirect-Upregulator | 75.09 | 79.42 | 77.19 | 1,680 |
| Inhibitor | 88.01 | 88.01 | 88.01 | 6,538 |
| Part-Of | 71.21 | 80.26 | 75.46 | 1,142 |
| Product-Of | 67.33 | 75.14 | 71.02 | 1,078 |
| Substrate | 72.07 | 64.68 | 68.18 | 2,497 |
| Substrate_Product-Of | 0.00 | 0.00 | 0.00 | 27 |
| Agonist-Activator | 0.00 | 0.00 | 0.00 | 10 |

## IV. CONCLUSION

We described our contribution to the DrugProt shared task in which we model chemical-protein relation extraction as a relation classification problem at the sentence level. We propose a model that builds on ensembled pretrained transformers and additional textual descriptions of chemicals taken form the CTD database. The proposed model achieves an F1 score of 79.73% on the hidden DrugProt test set which is an improvement of over 17 percentage points over the average score of all task participants. Our analysis indicates that both ensembling and entity descriptions improve results and that the number of training examples strongly influences performance for the different relation types. In future work, we want to integrate the proposed chemical-protein relation extraction model into our standalone tool for biomedical relation extraction (9) and explore generative approaches for chemical-protein relation extraction (28), as the intrinsic few/zero-shot capabilities of such generative models might improve results for relation types with few annotated examples.

## ACKNOWLEDGMENT

Leon Weber acknowledges the support of the Helmholtz Einstein International Berlin Research School in Data Science (HEIBRiDS). Samuele Garda is supported by the Deutsche Forschungsgemeinschaft as part of the research unit "Beyond the Exome".

## REFERENCES

1. Zheng, S., Dharssi, S., Wu, M., Li, J., & Lu, Z. (2019). Text mining for drug discovery. *Bioinformatics and Drug Discovery*, 231-252.

2. Dugger, S. A., Platt, A., & Goldstein, D. B. (2018). Drug development in the era of precision medicine. *Nature reviews Drug discovery*, *17*(3), 183-196.

3. Griffith, M., Spies, N. C., Krysiak, K., McMichael, J. F., Coffman, A. C., Danos, A. M., ... & Griffith, O. L. (2017). CIViC is a community knowledgebase for expert crowdsourcing the clinical interpretation of variants in cancer. *Nature genetics*, *49*(2), 170-174.

4. Zhou, D., Zhong, D., & He, Y. (2014). Biomedical relation extraction: from binary to complex. *Computational and mathematical methods in medicine*, *2014*.

5. Giuliano, C., Lavelli, A., & Romano, L. (2006). Exploiting shallow linguistic information for relation extraction from biomedical literature. In *11th Conference of the European Chapter of the Association for Computational Linguistics*.

6. Tikk, D., Thomas, P., Palaga, P., Hakenberg, J., & Leser, U. (2010). A comprehensive benchmark of kernel methods to extract protein–protein interactions from literature. *PLoS computational biology*, *6*(7), e1000837.

7. Zhao, Z., Yang, Z., Luo, L., Lin, H., & Wang, J. (2016). Drug drug interaction extraction from biomedical literature using syntax convolutional neural network. *Bioinformatics*, *32*(22), 3444-3453.

8. Vashishth, S., Joshi, R., Prayaga, S. S., Bhattacharyya, C., & Talukdar, P. (2018). Reside: Improving distantly-supervised neural relation extraction using side information. *arXiv preprint arXiv:1812.04361*.

9. Weber, L., Thobe, K., Migueles Lozano, O. A., Wolf, J., & Leser, U. (2020). PEDL: extracting protein–protein associations using deep language models and distant supervision. *Bioinformatics*, *36* (Supplement_1), 490-498.

10. Alt, C., Hübner, M., & Hennig, L. (2019). Fine-tuning pre-trained transformer language models to distantly supervised relation extraction. *arXiv preprint arXiv:1906.08646*.

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

12. Thomas, P., Solt, I., Klinger, R., & Leser, U. (2011). Learning protein–protein interaction extraction using distant supervision. In *Proceedings of Workshop on Robust Unsupervised and Semisupervised Methods in Natural Language Processing,* (pp. 25-32).

13. Smirnova, A., & Cudré-Mauroux, P. (2018). Relation extraction using distant supervision: A survey. *ACM Computing Surveys (CSUR)*, *51*(5), 1-35.

14. Ye, Z. X., & Ling, Z. H. (2019). Distant supervision relation extraction with intra-bag and inter-bag attentions. *arXiv preprint arXiv:1904.00143*.

15. Phan, L. N., Anibal, J. T., Tran, H., Chanana, S., Bahadroglu, E., Peltekian, A., & Altan-Bonnet, G. (2021). SciFive: a text-to-text transformer model for biomedical literature. *arXiv preprint arXiv:2106.03598*.

16. Yuan, Z., Liu, Y., Tan, C., Huang, S., & Huang, F. (2021). Improving Biomedical Pretrained Language Models with Knowledge. *arXiv preprint arXiv:2104.10344*.

17. Krallinger, M., Rabal, O., Akhondi, S.A., Pérez, M.P., Santamaría, J., Rodríguez, G.P., Tsatsaronis, G. and Intxaurrondo, A. (2017), October. Overview of the BioCreative VI chemical-protein interaction Track. In *Proceedings of the sixth BioCreative challenge evaluation workshop* (Vol. 1, pp. 141-146).

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

19. Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., ... & Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

20. Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, *15*(1), 1929-1958.

21. Lewis, P., Ott, M., Du, J., & Stoyanov, V. (2020, November). Pretrained Language Models for Biomedical and Clinical Tasks: Understanding and Extending the State-of-the-Art. In *Proceedings of the 3rd Clinical Natural Language Processing Workshop* (pp. 146-157).

22. Mattingly, C. J., Rosenstein, M. C., Colby, G. T., Forrest Jr, J. N., & Boyer, J. L. (2006). The Comparative Toxicogenomics Database (CTD): a resource for comparative toxicological studies. *Journal of Experimental Zoology Part A: Comparative Experimental Biology*, *305*(9), 689-692.

23. Sung, M., Jeon, H., Lee, J., & Kang, J. (2020). Biomedical entity representations with synonym marginalization. *arXiv preprint arXiv:2005.00239*.

24. Li, J., Sun, Y., Johnson, R. J., Sciaky, D., Wei, C. H., Leaman, R., ..., & Lu, Z. (2016). BioCreative V CDR task corpus: a resource for chemical disease relation extraction. *Database*, *2016*.

25. Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

26. Xiong, R., Yang, Y., He, D., Zheng, K., Zheng, S., Xing, C., Zhang, H., Lan, Y., Wang, L. & Liu, T. (2020). On layer normalization in the transformer architecture. *International Conference on Machine Learning* (pp. 10524-10533)

27. Du, X., Rush, A. M., & Cardie, C. (2021). Template Filling with Generative Transformers. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 909-914).