Using BERT to identify drug-target interactions from whole PubMed

Jehad Aldahdooh
University of Helsinki
Markus Vähä-Koskela
University of Helsinki
Jing Tang
University of Helsinki
Ziaurrehman Tanoli (zia.rehman@helsinki.fi)
University of Helsinki

Research Article

Keywords: BERT, Bidirectional Encoder Representations from Transformers, BERT for biomedical data, drug target interaction prediction, mining drug target interactions, biomedical text mining, bioactivity data, drug repurposing

Posted Date: October 29th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-1015236/v1

Abstract

Background: Drug-target interactions (DTIs) are critical for drug repurposing and elucidation of drug mechanisms, and are manually curated by large databases, such as ChEMBL, BindingDB, DrugBank and DrugTargetCommons. However, the number of articles providing this data (~0.1 million) likely constitutes only a fraction of all articles on PubMed that contain experimentally determined DTIs. Finding such articles and extracting the experimental information is a challenging task, and there is a pressing need for systematic approaches to assist the curation of DTIs. To this end, we propose Bidirectional Encoder Representations from Transformers (BERT) to identify such articles. Because DTI data intimately depends on the type of assays used to generate it, we also aimed to incorporate functions to predict the assay format. Results: Our novel method identified ~2.1 million articles (along with drug and protein information) that are not previously included in public DTI databases. Using 10-fold cross-validation, we obtained ~99% accuracy for identifying articles containing quantitative drug-target profiles. The accuracy for the prediction of assay format is ~90%, which leaves room for improvement in future studies. Conclusion: The BERT model in this study is robust and the proposed pipeline can be used to identify previously overlooked articles containing quantitative DTIs. Overall, our method provides a significant advancement in machine-assisted DTI extraction and curation. We expect it to be a useful addition to drug mechanism discovery and repurposing.

1. Introduction

The average cost of developing a new drug ranges in billions of dollars, and it takes 9–15 years to bring a new drug to the market [1]. Hence, finding new uses for already approved drugs is of major interest to the pharmaceutical industry. This practice, termed drug repositioning or drug repurposing, is attractive because of its potential to speed up drug development, reduce costs, and provide treatments for unmet medical needs [2]. Central to drug discovery and repositioning are drug-target interactions (DTI), meaning the qualitative, quantitative, and relative interactions of compounds with the molecules that regulate cellular functions. DTIs are catalogued in public databases, which classify DTIs as binary (contains both active and inactive interactions), unary (only active interactions) or as quantitative (in terms of IC50, Kd, Ki etc.) [3]. The most well-known databases for quantitative bioactivity interactions are ChEMBL [4], BindingDB [5], PubChem [6], GtopDB [7] and DrugTargetCommons [8]. These resources contain experimental data for millions of compounds across thousands of protein targets. The quantitative DTI data in these databases is manually extracted from experimental articles. None of these drug-target databases provide target coverage for approved drugs at the whole proteome level, and only 11% of the human proteome are targeted by small molecules [9]. The combined non-overlapping articles covered by these five databases numbered less than 0.1 million, and contain around 3,000 protein targets with an average of 7.33 interactions per target [10]. Moreover, each data resource focuses only on specific journals for data curation. For instance, ChEMBL and DrugTargetCommons primarily focus on Medicinal chemistry, Nature biotechnology and a few other journals. However, there are more than 7000 journals and 32M articles on PubMed [11]. A large fraction of the skipped articles in drug-target databases may

contain experimentally tested DTIs. However, curating whole PubMed manually is not efficient. Therefore, there is a need to develop semi-automated text classifiers that can identify most relevant articles and dig further to extract actual DTIs reported in the full text of the articles.

Text classification is a well-known problem in natural language processing (NLP). The objective is to assign predefined categories to a given text sequence (in this case, it could be an abstract, title or full text for the article). One of the pre-processing step is to map textual data into numeric features [12] to make it understandable by the prediction model. Mapping of textual information into numeric features can be performed using pre-trained models on a large corpus of texts. Pre-trained language models on large text corpora are proven to be adequate for the task of text classification with a decrease in computational costs at runtime [13]. Among those are the word embedding based models, such as word2vec [14] and GloVe [15] and contextualized word embedding models, such as CoVe [16] and ELMo [17]. Others are sentence-level models, such as ULMFiT [18]. More recently, pre-trained language models are shown to be helpful in learning common language representations by utilizing a large amount of un-labelled data: e.g., OpenAI GPT [19] and BERT [20]. Bidirectional Encoder Representations from Transformers (BERT) is based on a multi-layer bidirectional Transformer [21] and is trained on large plain texts for masked word prediction and next sentence prediction tasks.

PubTator [22] and BEST [23] are currently the two most comprehensive web platforms that can automatically mine compounds and target proteins from PubMed or PubMed Central (PMC). However, these tools fail to capture the compound-target relationships (interactions), and the resulting output may or may not contain experimental data. To solve these shortcomings, we set out to construct a pipeline using a BERT-based text classifier to identify articles containing DTIs and extract the associated data. We trained several BERT models (i.e., BERT, SciBERT [13], BioBERT [24], BioMed-RoBERTa [25] and BlueBERT [26]) on known articles containing DTIs and used majority voting of five BERT models to predict new articles containing DTIs. We identified ~2.1M new articles that possibly have DTI in terms of quantitative bioactivities. The identified articles are further linked with mined compound and protein entities provided by PubTator. Furthermore, the proposed BERT model could predict the assay format used in the experiment with an accuracy of ~90%. The resulting predicted and integrated data is freely available at https://dataset.drugtargetcommons.org/. The script for models is freely available at: https://github.com/JehadAldahdooh/DTIs.

2. Materials And Methods

2.1. Compound and protein annotations for PubMed articles

We downloaded compounds and proteins identified from the full text of 24M documents (75% of PubMed) using PubTator's API [22]. We define here document as a merged text containing titles and abstracts for the articles. We extracted data in batches of 1000 documents using parallel requests to save time. After preprocessing, we truncated those documents where text size exceeds 512 words because the sequence limit for BERT tokenization is 512. Approximately a quarter of the articles at

PubTator missed the abstract information. We considered only those articles for which both abstract and title information is present in PubTator and therefore left with **~18.5M** documents.

2.2. Known articles for compound-target bioactivity data

Data used for the model's training contains 66,521 positive examples (articles containing compound-target bioactivity data) and 66,521 negative examples (other biological articles) is available at: https://dataset.drugtargetcommons.org/DT_dataset.csv. Compound-target articles are extracted from DrugTargetCommons and ChEMBL (27^{th} release), whereas data for other biological documents is extracted from DisGeNET [27]. DisGeNET is a comprehensive resource for cataloguing gene-disease associations. Abstracts and titles (for articles) are extracted using PubMed's API. Trained models are then used to predict documents (abstract + title) likely to contain DTIs. Positively predicted documents are associated with compound and protein entities as identified by PubTator.

2.3. Assay formats for compound-target bioactivity data

Furthermore, we trained our models to predict the assay format most likely used in the documents. Assay format annotations are extracted from DrugTargetCommons for 28,102 documents with 14,109 focusing on cell-based assays and 13,993 having other assay formats (e.g., biochemical (93), cell-free (66), organism-based (12,845), tissue-based (424) and physiochemical (565)). We merged several assay formats under the 'other' category to avoid data imbalance problems while predicting assay format (available at https://dataset.drugtargetcommons.org/assay_format_data_label.csv).

2.4. Proposed method

BERT base is a masked language model (MLM) with 12 layers of architecture, pre-trained on > 2.5B words from English Wikipedia. We used BERT base and other BERT models (SciBERT, BioBERT, BioMed-RoBERTa and BlueBERT) to identify new documents on PubMed containing DTIs. The five BERT models are merged using majority voting to predict the label for the documents. Majority voting is a technique in machine learning used to combine the prediction power of several models. We adapted the majority voting technique to have more confidence in the prediction of true labels. SciBERT is an MLM pre-trained model trained on 1.14M full-texts from Semantic Scholar corpus with 82% from the biomedical domain [28]. SciBERT uses a different vocabulary (SCIVOCAB), whereas BERT, in general, is based on BASEVOCAB. In this study, we adapted uncased SciBERT. BioBERT is an MLM pre-trained language model based on the BERT representation for the biomedical domain. We used BioBERT-v1.1, pre-trained on PubMed for 200K steps and 270K steps on PMC. The model is pre-trained using the same hyper-parameter settings as for the original BERT model. BioMed-RoBERTa is MLM pre-trained language model based on the RoBERTa [25]. Finally, BlueBERT is pre-trained on ~4B words extracted from PubMed.

We used the BERT representations for the classification task by fine-tuning the BERT variants with minimal changes applied during the training phase. All BERT models used in this analysis comprised 12 layers of transformer encoder with hidden state dimensions equal to 768 and having >110M parameters

as adopted in [21]. In our architecture, we have used the embedding vector of the BERT [CLS] token from the last hidden layer as a representation of each textual sequence. The special [SEP] token is used to separate the text sequences. It is further processed by 2 fully connected layers and a SoftMax activation function.

The BERT variants are fine-tuned using NVIDIA Tesla V100 SXM2 32 GB GPU, with a batch size of 32, maximum sequence length: 512, a learning rate of 2e-5 and maximum epoch size of 3. We used Adam with b_1=0.9 and b_2=0.999, slanted triangular learning rates as in [18], warm-up portion to 0.1 and ensured that GPU memory is fully utilized. The model architecture for all BERT models in this study is shown in Figure 1.

Next, we divided the overall workflow into three modules:

1. To identify whether a PubMed's article is likely to contain bioactivity data for drug and protein pair or not.
2. Extract drug and protein information by taking advantage of already extracted entities by PubTator.
3. Predict assay format for positively identified articles.

For module 1, we used the fine-tuned BERT models to predict whether PubMed's article contains a compound-target relationship or not. BERT models are trained on 66,521 positive and 66,521 negative documents (abstracts + title of the article) as explained in the previous section. We pre-processed all the BERT models by assigning tokens for each word in the documents and converted all words into lower case. We then padded for cases where document length <512. Finally, each document (article) is mapped into 768 numeric features with minor differences in the five models. After training, BERT models are merged in majority voting to identify new articles possibly containing bioactivity DTIs. For module 2, we then matched and linked positively predicted documents with identified drug and protein entities using the PubTator dataset. Finally, for module 3, using the same architecture, we tried to predict assay formats for the positively predicted documents (cell-based vs other assays). We emphasized on assay format field because assay formats are critical in defining scores for DTIs [29]. We organized the final output in terms of PubMed id for the article, predicted assay format, drugs, and proteins, and which is freely available at https://dataset.drugtargetcommons.org/. The workflow for the proposed research is shown in Figure 2.

3. Results And Discussions

3.1. Ten-fold cross-validation results using BERT models

To feed the PubMed documents into the BERT pipeline, we first converted linguistic units (text snippets, words and phrases that carry meaning) in 66,521 positive and 66,521 negative documents into tokens. We truncated those documents for which document length >512 (maximum limit by BERT). Each BERT

model generated a numeric feature set of length 768 for each document. BERT text classifier is trained using 10-fold cross-validation, and the performances are shown in Table 1.

Table 1: Ten-fold cross-validation results for different BERT models. The last column shows accuracy% for independent testing of negative class.

BERT model	Cross validation performance		Independent testing performance
	Accuracy %	F1 %	Accuracy %
BERT	99.2 ±0.1	99.1 ±0.1	99.6 ±0.0
SciBERT	99.3 ±0.1	99.3 ±0.1	99.7 ±0.1
BioBERT	99.3 ±0.1	99.3 ±0.1	99.7 ±0.1
BioMed-RoBERTa	**99.7 ±0.0**	**99.6 ±0.0**	**99.8 ±0.0**
BlueBERT	99.3 ±0.1	99.3 ±0.1	99.6 ±0.1

Our analyses showed that all BERT models reached accuracies higher than 99%. In comparison, the traditional machine learning algorithms including Support Vector Machine and Naïve Bayes yielded accuracies of 98.87% and 98.29%, respectively. The accuracy of independent BERT testing on 7,785 articles of only the negative class (other biological documents) also showed > 99% accuracy. Our findings demonstrate that BERT models can successfully detect drug-target like documents and distinguish true negatives with great precision.

To examine the word distributions in two types of documents, we also analyzed the top frequently occurring words in positive and negative documents. As shown in Figure 3, the most frequently occurring words in drug target documents are ‘compounds’, ‘activity’ and ‘potent’, whereas the most frequent words for other biological documents are ‘expression’, ‘patients’, and ‘gene’. Word distribution analysis can demonstrate developing a simple model based on word frequencies to identify drug-target or other biological documents.

3.2. Identify new drug-target articles and associate drug and protein pairs using PubTator dataset

After successfully training the BERT models, we tried to identify new articles on PubMed that possibly contain bioactivity data for drug-target pairs. For this purpose, we used ~18.5M documents downloaded using PubTator’s API. Each BERT model has its strength, and we merged the predictions from the five BERT models in majority voting to determine whether a article classification is positive or negative. Table 2 shows the number of positively and negatively predicted documents out of ~18.5M documents. The fourth column (articles containing drugs or proteins on PubTator) in Table 2 shows how many among positively predicted articles (using BERT models) have either drug or protein entity extracted by PubTator. Finally, the last column indicates the number of articles for which PubTator identified both drugs and the

proteins. These two columns, validate those articles that are identified as drug-target articles using BERT model.

Table 2:Prediction of drug-target like documents from 18.5M articles on PubMed. The fourth column shows the number of documents that contain either drug or protein entities as identified by PubTator. In contrast, the fifth column indicates the number of documents that contain both drugs and protein entities.

BERT model	Predicted as drug-target articles	Predicted as other articles	Articles containing drugs or proteins on PubTator	Articles containing both drugs and proteins on PubTator
BERT	2,394,207	16,164,382	2,239,986	513,972
SciBERT	**2,564,440**	15,994,149	**2,386,665**	**562,051**
BioBERT	1,914,147	16,644,442	1,821,758	459,349
BioMed-RoBERTa	1,371,113	17,187,476	1,311,936	315,756
BlueBERT	2,443,832	16,114,757	2,304,304	530,033
Majority voting	2,129,731	16,428,858	2,019,050	467,638

Superior performance on unseen articles shows how well the predictions by BERT models can be generalized. Using the majority voting of BERT models, 94.8% (2,019,050) of the articles identified as drug-target like (positive) containing either drugs or proteins entity identified by PubTator. Out of the ~2.1M positively predicted documents, 21.9% (467,638) contain both drug and protein entities at PubTator. The result is likely an underestimation, as drug or protein entities (or both) may have been deposited as supplementary data, which is not captured by PubTator's back-end algorithm. This means that even though the article is positively predicted our workflow might not capture drugs or proteins in some cases, leaving the task for manual curators to check the supplementary material. Indeed, many high throughput articles do not mention drug or protein names in the article's main text but instead provide these in the supplement (e.g. Davis et al., 2011) [31]. Of the BERT models, SciBERT identified more drug-target like documents compared to other models, with at least 562,051 articles containing both drugs and proteins in the PubTator dataset. This could be because SciBERT is additionally trained for biomedical applications, whereas other models are designed for general text classification applications.

We also analyzed the top journals, and yearly distribution for ~2.1M predicted articles. This will give an idea for manual curators as to what journals (and year of publications) are suitable for extracting DTIs. Using the Bio Entrez package in python, we obtained journal names and year of publication for ~2.1M

articles. We extracted journal names and year information only for ~0.2M out of ~2.1M articles (though PubMed IDs are present in Entrez). Journal names and years of publications for most of the articles are missing from the Entrez database. However, we still can have an idea of the general trend based on captured information. Figure 4A shows the top 15 journals containing drug-target like articles, whereas Figure 4B shows year wise frequencies for the articles. As shown in Figure 5A, Journal of Medicinal Chemistry, Biological Chemistry, and Bioorganic & Medicinal Chemistry Letters are present among the top 15 journals. These three journals are among the leading journals for bioactivity data extraction in ChEMBL [32]. Furthermore, most drug-target articles are from recent years, with the year of 2020 containing the most significant number of articles.

There are 66,521 DTI articles that are used to train BERT models. Among these, 60,995 are overlapping with ~18.5M articles in PubTator dataset. We also tried to analyze the overlap of 60,995 DTI articles with ~2.1M articles that are predicted as DTI articles using BERT models. Figure 5 shows that 99.6% (60,775) DTI articles are present among the list of predicted DTI articles. This means our analysis identified 2,068,956 additional articles containing DTIs. These newly identified articles, along with PubMed IDs, drugs and proteins involved in the article, are freely available at https://dataset.drugtargetcommons.org/. The output of our analysis can be used as a starting point to further extract the quantitative drug-target bioactivity values from the identified articles. We hope that our output will significantly ease the job of manual curators as we are providing the actual PubMed ID, drugs, and protein entities, as well as assay formats for ~2.1M identified articles.

3.3. Predict assay format for drug-target articles

After successfully identifying DTIs, the next task is to predict the assay format most likely adapted in the identified articles. For that purpose, we separately trained each BERT model on assay format dataset with 14,109 documents based on cell-based and 13,993 having other assay formats. We used the same fine-tuning settings as for drug-target article identification task. Figure 6 shows 10-fold cross-validation performances for BERT models in terms of accuracy% and F1%. All BERT models have accuracy >88% with BioBERT slightly outperforming other models. This shows that BERT models can successfully identify drug-target articles as well as assay formats.

After successful training of BERT models to predict assay formats (on known articles), we applied trained models for predicting assay formats in those articles, which are positively predicted using majority voting. There are ~2.1M articles that are predicted as drug-target like document using majority voting as mentioned in Table 2 (last row). Therefore, we tried to predict assay formats for those ~2.1M articles. As shown in Table 3, >26.5% of the drug-target like articles (564,425 out of 2,129,731) are likely based on cell-based assays. It is impossible to computationally validate these predictions now. However, in the next release of DrugTargetCommons, we will validate the predicted assay formats using manual curation of ~2.1M identified articles.

Table 3: BERT models to predict assay formats for ~2.1M newly predicted drug-target like articles.

BERT models	Cell based assay	Another assay
BERT	864,319	1,265,412
SciBERT	541,322	1,588,409
BioBERT	474,283	1,655,448
BioMed-RoBERTa	1,013,497	1, 6,234
BlueBERT	476,716	1,653,015
Majority voting	564,425	1,565,306

4. Conclusions

Most DTI resources are compound centric and lack DTI profiles at complete proteomic level. In compound centric approaches, thousands of the compounds are tested across a particular target protein. Only 11% of the human proteome are targeted by small molecules or drugs, whereas one in three proteins is still being investigated [33]. The combined non-overlapping experimental articles (on PubMed) are less than 0.1M. Curating quantitative bioactivity values reported in an article cannot be fully automated. However, semi-automated NLP based methods can assist in identifying related articles and easing the workload for the manual curators. BERT is recently proposed as a state-of-the-art model for several NLP tasks, including text classification. Therefore, in this research, we investigated several models of BERT to identify new articles possibly containing DTIs.

Furthermore, we tried to predict assay formats most likely used in the articles. Assays formats, along with actual bioactivities, are critical in defining scores for DTIs. Using majority voting based on BERT models, we identified 2,129,731 articles from which 467,638 are confirmed to have both drug and protein entities in the PubTator dataset. Most of these ~2.1M articles are not reported in any of the manually curated bioactivity databases as the combined non-overlapping articles curated by commonly used DTI databases are around 0.1M. These identified DTIs (along with annotations) are freely available at https://dataset.drugtargetcommons.org/. We hope that the identified articles and drug and protein entities will ease the job of manual curators and improve protein target coverage across investigational and approved compounds. Lastly, increased target coverage for investigational and approved drugs will enhance the understanding of drug mechanism of action and open new drug repurposing opportunities. The manual curation team of DrugTargetCommons will take advantage of newly identified articles and curate more bioactivity data in their next release.

Declarations

Ethics approval, consent to participate and consent for publication: Not applicable

Availability of data and material: Newly identified articles, extracted drug/protein entities and predicted assay formats are freely available at https://dataset.drugtargetcommons.org/.

Competing interests: Authors have no competing interests

Funding: This work was supported by the EU H2020 (EOSC-LIFE, No. 824087), the European Research Council (DrugComb, No. 716063) and the Academy of Finland (No. 317680).

Authors' contributions: Z.T and J.A performed data analysis, Z.T wrote the manuscript, J.A J.T and M.V reviewed manuscript.

Acknowledgements: We thank CSC, Finland for providing us IT services.

References

[1] M. Dickson, J.P. Gagnon, The cost of new drug discovery and development, Discov. Med. 4 (2009) 172–179.

[2] A.F. Shaughnessy, Old drugs, new tricks, BMJ. 342 (2011) d741.

[3] Z. Tanoli, U. Seemab, A. Scherer, K. Wennerberg, J. Tang, M. Vähä-Koskela, Exploration of databases and methods supporting drug repurposing: a comprehensive survey, Brief. Bioinform. (2020).

[4] A. Gaulton, A. Hersey, M. Nowotka, A.P. Bento, J. Chambers, D. Mendez, P. Mutowo, F. Atkinson, L.J. Bellis, E. Cibrián-Uhalte, E. Al., The ChEMBL database in 2017, Nucleic Acids Res. 45 (2016) D945–D954.

[5] M.K. Gilson, T. Liu, M. Baitaluk, G. Nicola, L. Hwang, J. Chong, E. Al., BindingDB in 2015: A public database for medicinal chemistry, computational chemistry and systems pharmacology, Nucleic Acids Res. 44 (2016) D1045–D1053.

[6] Y. Wang, S.H. Bryant, T. Cheng, J. Wang, A. Gindulyte, B.A. Shoemaker, P.A. Thiessen, S. He, J. Zhang, PubChem BioAssay: 2017 update, Nucleic Acids Res. 45 (2016) D955–D963.

[7] S.P.H. Alexander, D. Fabbro, E. Kelly, A. Mathie, J.A. Peters, E.L. Veale, J.F. Armstrong, E. Faccenda, S.D. Harding, A.J. Pawson, The concise guide to pharmacology 2019/20: catalytic receptors, Br. J. Pharmacol. 176 (2019) S247–S296.

[8] Z. Tanoli, Z. Alam, M. Vähä-Koskela, B. Ravikumar, A. Malyutina, A. Jaiswal, J. Tang, K. Wennerberg, T. Aittokallio, Drug Target Commons 2.0: a community platform for systematic analysis of drug–target interaction profiles, Database. 2018 (2018) 1–13.

[9] D.-T. Nguyen, S. Mathias, C. Bologa, S. Brunak, N. Fernandez, A. Gaulton, A. Hersey, J. Holmes, L.J. Jensen, A. Karlsson, G. Liu, A. Ma'ayan, G. Mandava, S. Mani, S. Mehta, J. Overington, J. Patel, A.D. Rouillard, S. Schürer, T. Sheils, A. Simeonov, L.A. Sklar, N. Southall, O. Ursu, D. Vidovic, A. Waller, J. Yang,

A. Jadhav, T.I. Oprea, R. Guha, Pharos: Collating protein information to shed light on the druggable genome., Nucleic Acids Res. 45 (2017) D995–D1002. https://doi.org/10.1093/nar/gkw1072.

[10] Z. Tanoli, J. Aldahdooh, F. Alam, Y. Wang, U. Seemab, M. Fratelli, P. Pavlis, M. Hajduch, F. Bietrix, P. Gribbon, A. Zaliani, M.D. Hall, M. Shen, K. Brimacombe, E. Kulesskiy, J. Saarela, K. Wennerberg, M. Vähä-Koskela, J. Tang, Minimal information for chemosensitivity assays (MICHA): a next-generation pipeline to enable the FAIRification of drug screening experiments, Brief. Bioinform. (2021). https://doi.org/10.1093/bib/bbab350.

[11] J. White, PubMed 2.0, Med. Ref. Serv. Q. 39 (2020) 382–387.

[12] C. Sun, X. Qiu, Y. Xu, X. Huang, How to fine-tune bert for text classification?, in: China Natl. Conf. Chinese Comput. Linguist., Springer, 2019: pp. 194–206.

[13] I. Beltagy, K. Lo, A. Cohan, Scibert: A pretrained language model for scientific text, ArXiv Prepr. ArXiv1903.10676. (2019).

[14] T. Mikolov, I. Sutskever, K. Chen, G.S. Corrado, J. Dean, Distributed representations of words and phrases and their compositionality, in: Adv. Neural Inf. Process. Syst., 2013: pp. 3111–3119.

[15] J. Pennington, R. Socher, C.D. Manning, Glove: Global vectors for word representation, in: Proc. 2014 Conf. Empir. Methods Nat. Lang. Process., 2014: pp. 1532–1543.

[16] B. McCann, J. Bradbury, C. Xiong, R. Socher, Learned in translation: Contextualized word vectors, ArXiv Prepr. ArXiv1708.00107. (2017).

[17] M.E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, L. Zettlemoyer, Deep contextualized word representations, ArXiv Prepr. ArXiv1802.05365. (2018).

[18] J. Howard, S. Ruder, Universal language model fine-tuning for text classification, ArXiv Prepr. ArXiv1801.06146. (2018).

[19] A. Radford, K. Narasimhan, T. Salimans, I. Sutskever, Improving language understanding by generative pre-training, (2018).

[20] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, ArXiv Prepr. ArXiv1810.04805. (2018).

[21] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, in: Adv. Neural Inf. Process. Syst., 2017: pp. 5998–6008.

[22] C.-H. Wei, H.-Y. Kao, Z. Lu, PubTator: a web-based text mining tool for assisting biocuration, Nucleic Acids Res. 41 (2013) W518–W522.

[23] S. Lee, D. Kim, K. Lee, J. Choi, S. Kim, M. Jeon, S. Lim, D. Choi, S. Kim, A.-C. Tan, BEST: next-generation biomedical entity search tool for knowledge discovery from biomedical literature, PLoS One. 11 (2016) e0164680.

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

[25] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov, Roberta: A robustly optimized bert pretraining approach, ArXiv Prepr. ArXiv1907.11692. (2019).

[26] Y. Peng, S. Yan, Z. Lu, Transfer learning in biomedical natural language processing: an evaluation of BERT and ELMo on ten benchmarking datasets, ArXiv Prepr. ArXiv1906.05474. (2019).

[27] J. Piñero, À. Bravo, N. Queralt-Rosinach, A. Gutiérrez-Sacristán, J. Deu-Pons, E. Centeno, J. García-García, F. Sanz, L.I. Furlong, DisGeNET: a comprehensive platform integrating information on human disease-associated genes and variants, Nucleic Acids Res. 45 (2017) D833–D839.

[28] W. Ammar, D. Groeneveld, C. Bhagavatula, I. Beltagy, M. Crawford, D. Downey, J. Dunkelberger, A. Elgohary, S. Feldman, V. Ha, Construction of the literature graph in semantic scholar, ArXiv Prepr. ArXiv1805.02262. (2018).

[29] Z. Tanoli, Z. Alam, A. Ianevski, K. Wennerberg, M. Vähä-Koskela, T. Aittokallio, Interactive visual analysis of drug–target interaction networks using Drug Target Profiler, with applications to precision medicine and drug repurposing, Brief. Bioinform. (2018). https://doi.org/10.1093/bib/bby119.

[30] M.I. Davis, J.P. Hunt, S. Herrgard, P. Ciceri, L.M. Wodicka, G. Pallares, M. Hocker, D.K. Treiber, P.P. Zarrinkar, Comprehensive analysis of kinase inhibitor selectivity, Nat. Biotechnol. 29 (2011) 1046–1051.

[31] T. Anastassiadis, S.W. Deacon, K. Devarajan, H. Ma, J.R. Peterson, Comprehensive assay of kinase catalytic activity reveals features of kinase inhibitor selectivity, Nat. Biotechnol. 29 (2011) 1039–1045.

[32] G. Papadatos, G.J.P. van Westen, S. Croset, R. Santos, S. Trubian, J.P. Overington, A document classifier for medicinal chemistry publications trained on the ChEMBL corpus, J. Cheminform. 6 (2014) 1–8.

[33] T.I. Oprea, C.G. Bologa, S. Brunak, A. Campbell, G.N. Gan, A. Gaulton, S.M. Gomez, R. Guha, A. Hersey, J. Holmes, Unexplored therapeutic opportunities in the human genome, Nat. Rev. Drug Discov. 17 (2018) 317–332.

Figures

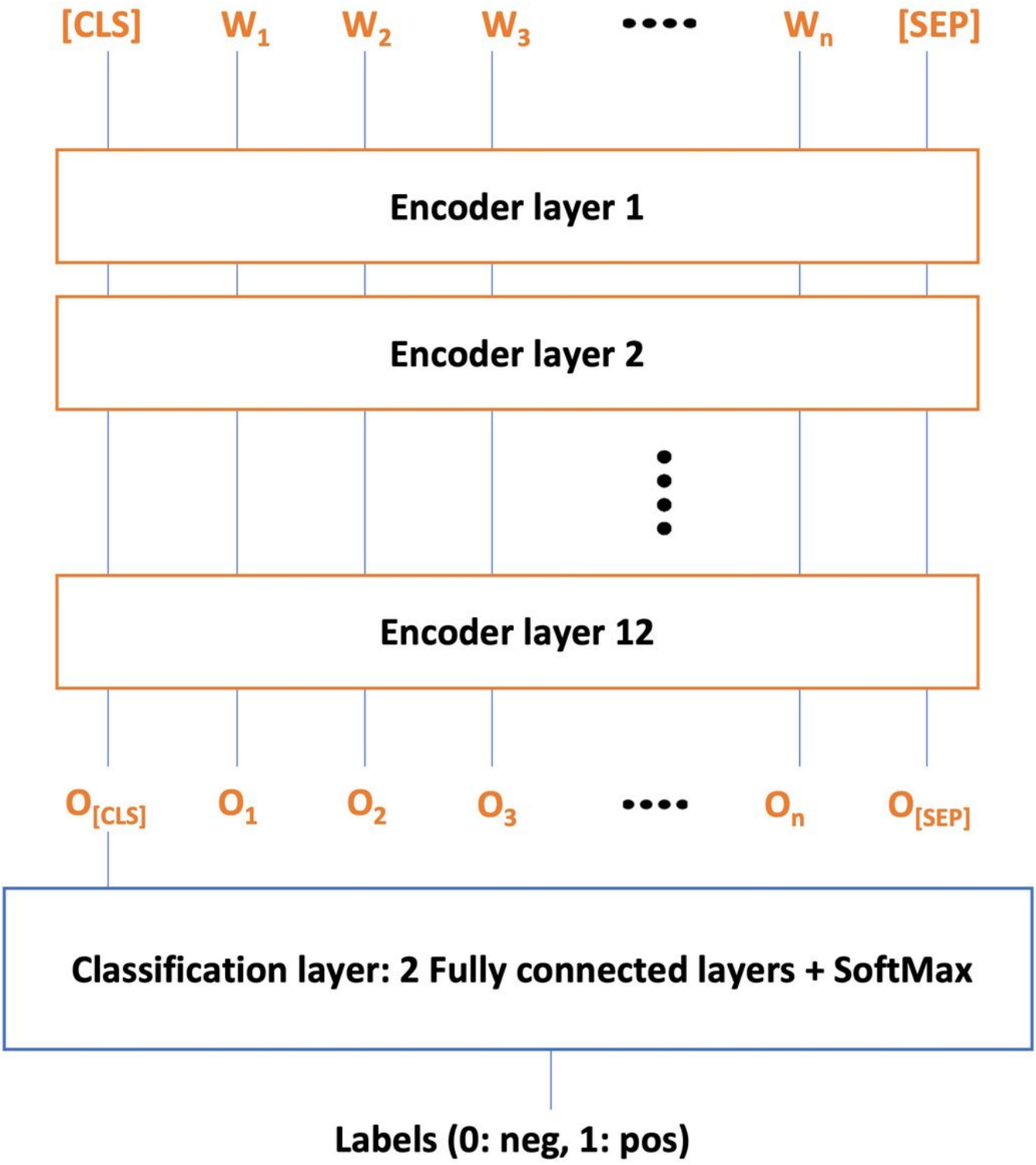

Figure 1

Architecture for all BERT variants, where Wi represents input word token and Oi represents contextual embeddings at the output layer. The O[CLS] is first token of output sequence and contains class label.

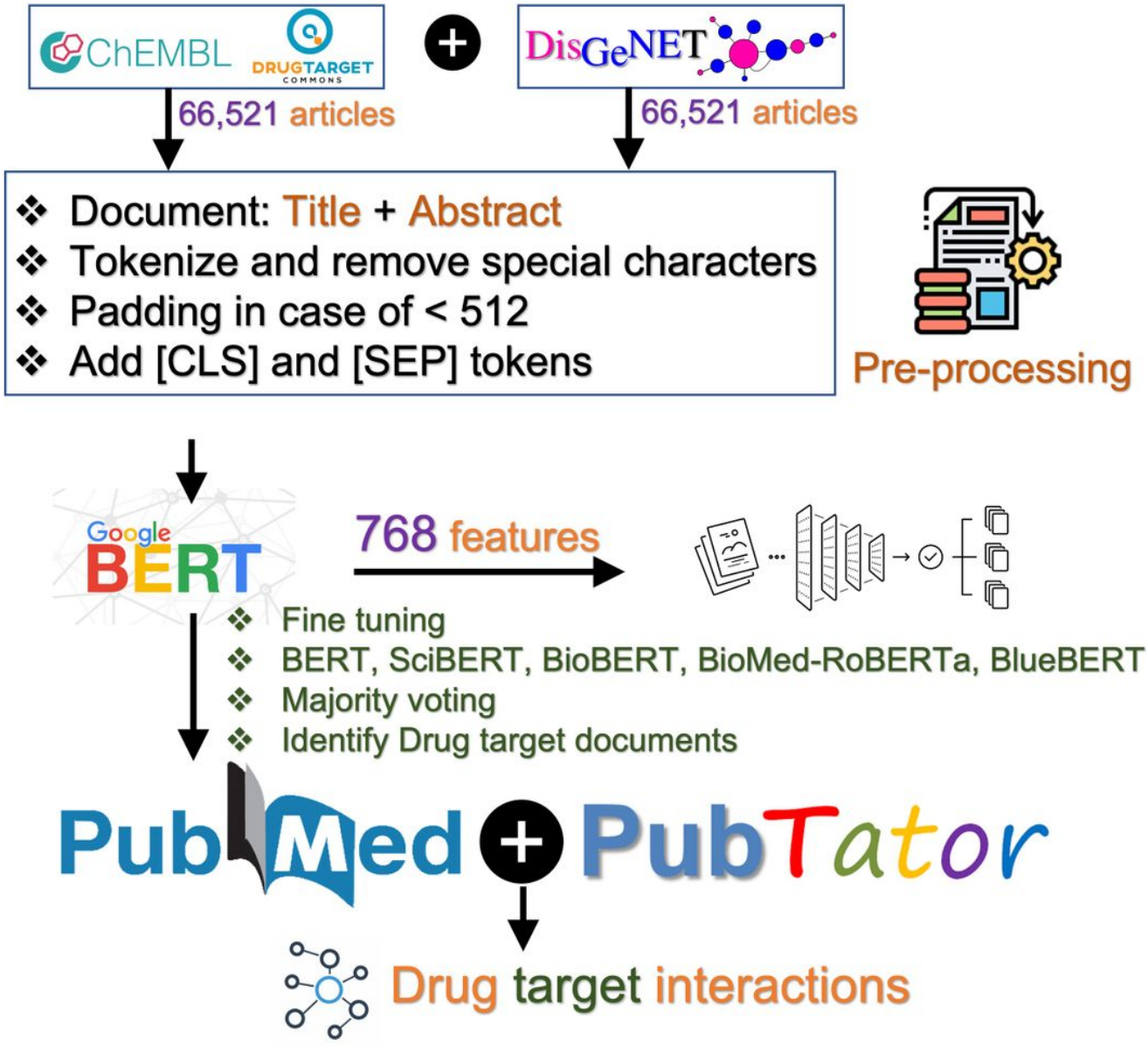

Figure 2

Workflow for identifying new articles containing drug-target bioactivity data. The drug and protein entities for articles are integrated from the PubTator dataset. The final output contains predictions for ~18.5M articles that possibly have DTIs

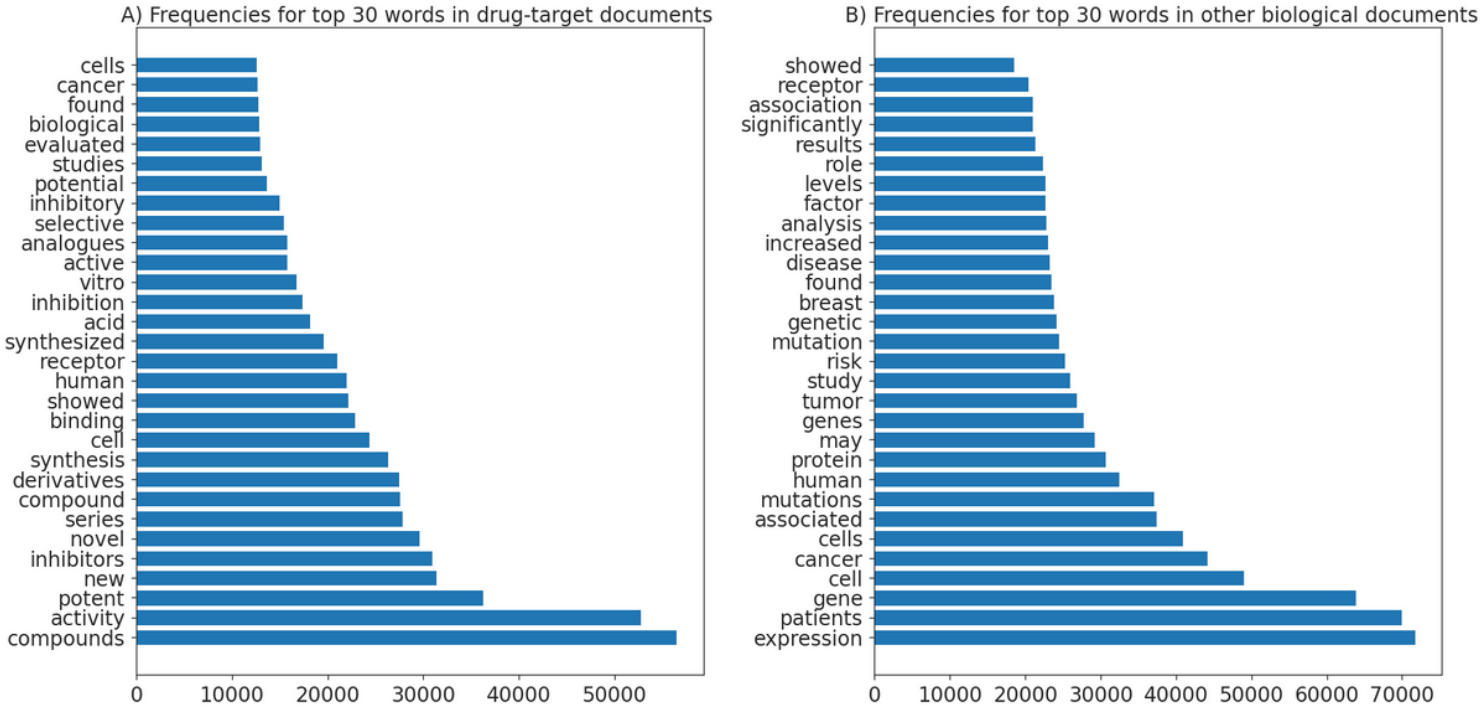

Figure 3

Top 30 words frequencies, A) Drug-target like documents, B) Other biological documents.

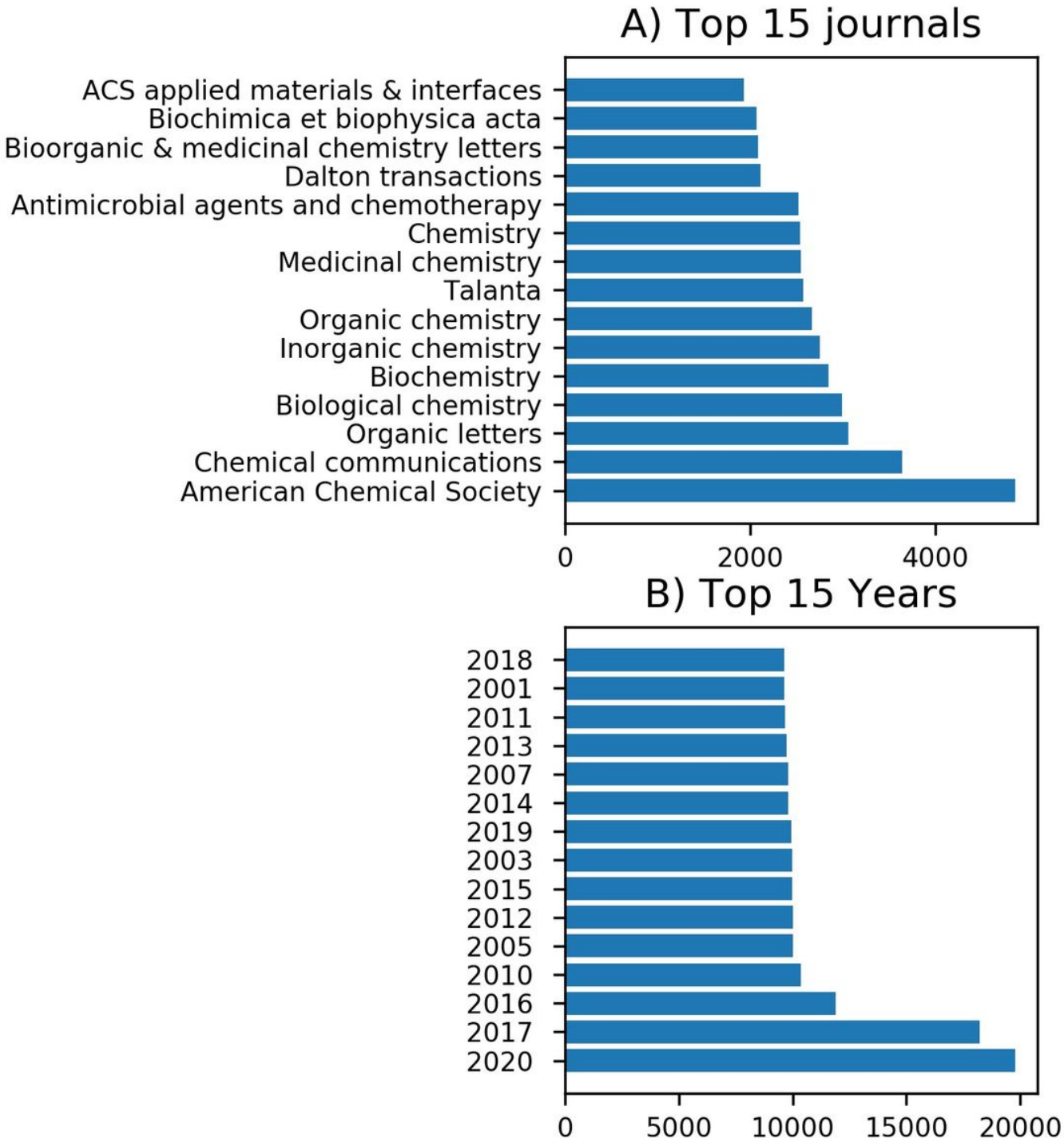

Figure 4

Statistics for ~0.2M articles which are predicted as drug-target like articles using majority voting, A) Top 15 journals and B) Top 15 years for drug-target like articles.

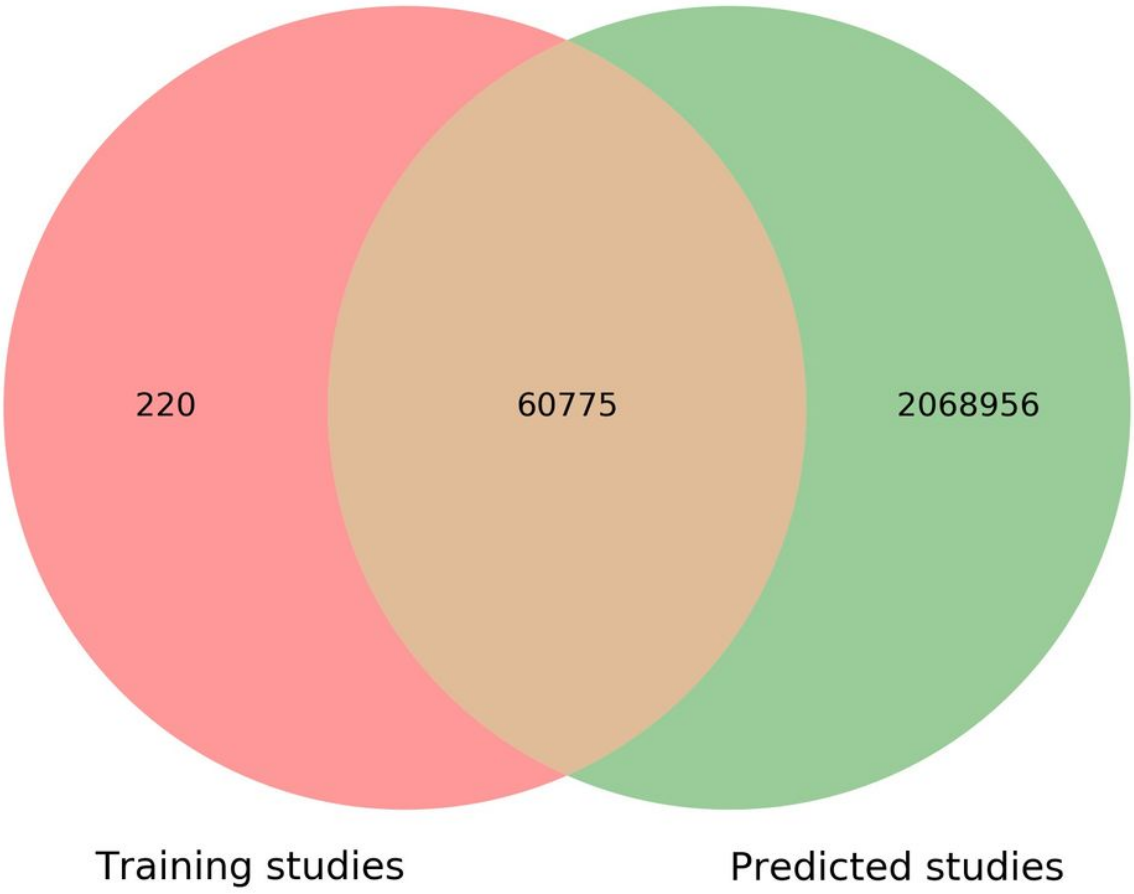

Figure 5

Overlap between training and predicted articles containing quantitative DTIs.

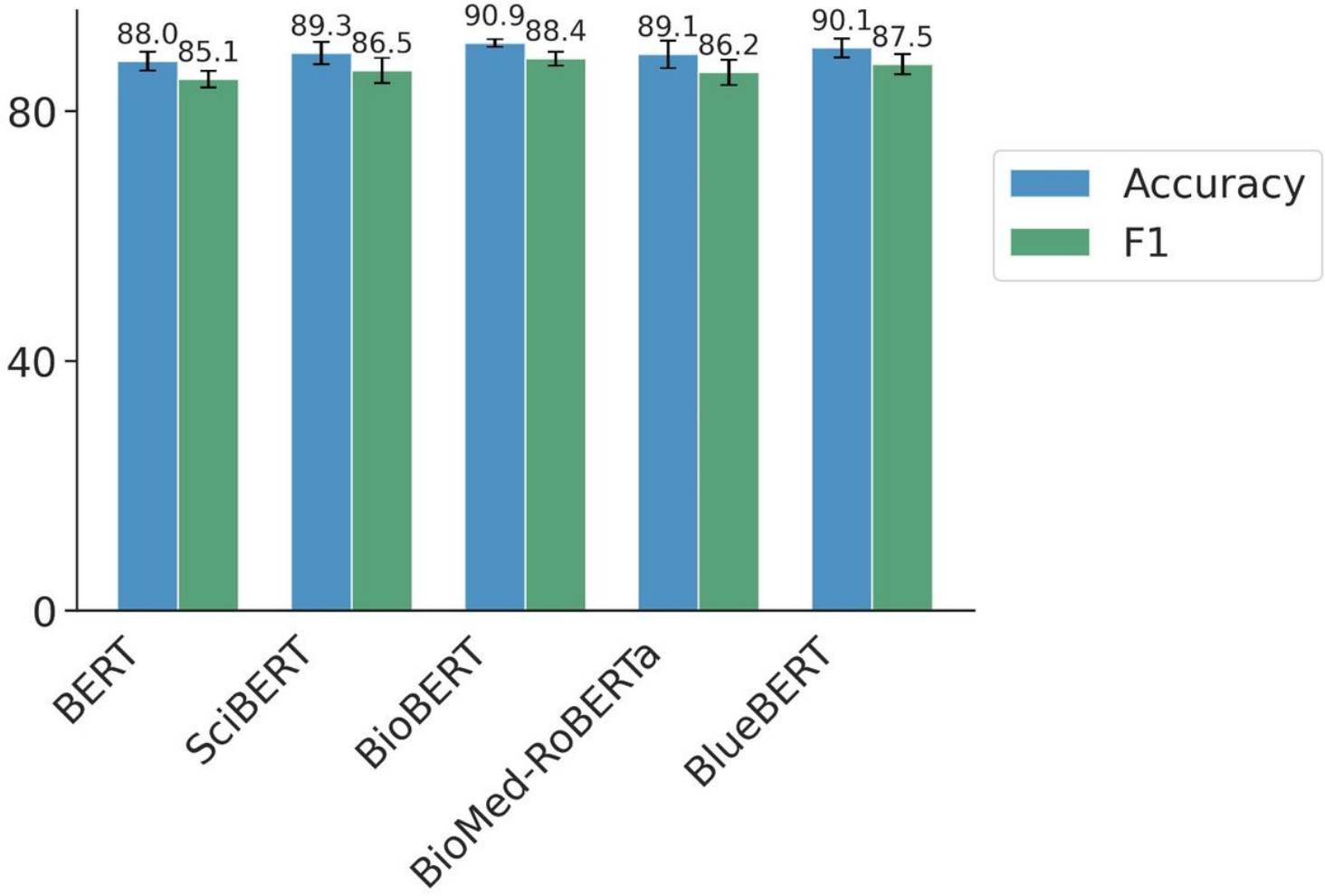

Figure 6

10-fold cross- validation performance for the task of assay format prediction.