Using BERT to identify drug-target interactions from whole PubMed

Aldahdooh, Jehad

2022-06-21

Aldahdooh , J , Vähä-Koskela , M , Tang , J & Tanoli , Z 2022 , ' Using BERT to identify drug-target interactions from whole PubMed ' , BMC Bioinformatics , vol. 23 , no. 1 , 245 . https://doi.org/10.1186/s12

http://hdl.handle.net/10138/346168
https://doi.org/10.1186/s12859-022-04768-x

BMC Bioinformatics

RESEARCH

Open Access

Using BERT to identify drug-target interactions from whole PubMed

Jehad Aldahdooh[1,3], Markus Vähä-Koskela[2], Jing Tang[1*] and Ziaurrehman Tanoli[1,4*]

*Correspondence:
jing.tang@helsinki.fi; zia.rehman@helsinki.fi

[1] Research Program in Systems Oncology, Faculty of Medicine, University of Helsinki, Helsinki, Finland
[2] Institute for Molecular Medicine Finland, University of Helsinki, Helsinki, Finland
[3] Doctoral Programme in Computer Science, University of Helsinki, Helsinki, Finland
[4] BioICAWtech, Helsinki, Finland

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 curated articles likely constitutes only a fraction of all the articles 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 applied 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 0.6 million articles (along with drug and protein information) which 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 F1 micro for the prediction of assay format is 88%, 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.

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

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 drugs 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], 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 [10]. 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 [11].

To overcome the limited coverage of DTI profiles in the public databases, several in-silico DTI prediction studies are proposed. For instance, the IDG-DREAM Challenge is based on crowdsourcing-based AI and ML methods to predict target activities for kinase inhibitors [12]. Thafar et al., predicted new DTIs using graph embedding and similarity based approaches [13]. Similarly, Zheng et al., used multiple kernels into a tripartite heterogeneous drug–target–disease interaction spaces to predict DTIs [14]. Several other computational approaches have been developed over the past decade, providing systematic means for predicting potential DTIs [15–17]. These in-silico methods provide a deeper understanding of the factors affecting DTI prediction and have opened novel strategies for computational drug repurposing.

Another alternative strategy is the curation of DTIs from experiment-based articles, adapted by several major databases such as ChEMBL, BindingDB and DrugTargetCommons. However, each 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 32 M articles on PubMed [18]. A large fraction of the non-curated articles may contain experimentally tested DTIs. However, curating the whole PubMed manually is not efficient. Therefore, there is a need to develop semi-automated text classifiers to identify the most relevant 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 numerical features [19], to make it understandable by the prediction model. Mapping of textual information into numerical 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 [20]. Among those are the word embedding based models, such as word2vec [21] and GloVe [22], as well as contextualized word embedding models, such as CoVe [23] and ELMo [24]. Others are sentence-level models, such as ULMFiT [25]. 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 [26] and BERT [27]. Bidirectional Encoder Representations from

Transformers (BERT) is based on a multi-layer bidirectional Transformer and is trained on large plain texts for masked word prediction and next sentence prediction tasks.

PubTator [28] and BEST [29] are currently the two most comprehensive web platforms that can automatically mine drug and target proteins from PubMed or PubMed Central (PMC). However, these tools did not capture the DTIs, 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 from PubTator. We trained several BERT models (i.e., BERT, SciBERT [20], BioBERT [30], BioMed-RoBERTa [31] and BlueBERT [32]) on known articles containing DTIs and used majority voting of five BERT models to predict 0.6 M new articles. The identified articles are further linked with mined drug and protein entities provided by PubTator. Furthermore, the BERT models predicted the assay format used in the experiment with an F1 micro of 88%. The resulting predicted and integrated datasets are freely available at https://dataset.drugtargetcommons.org/. The script for generating these models is freely available at: https://github.com/JehadAldahdooh/DTIs.

Materials and methods

Drug and protein annotations for PubMed articles

We downloaded drug and protein annotations for the abstracts of 24 M documents (75% of the PubMed) using PubTator's API [28]. We define here document as a merged text containing titles and abstracts for the articles. Approximately a quarter of the articles in PubTator missed the abstract information. We considered only those articles for which both abstract and title information is present in PubTator, after which **18.5 M** documents remained.

Known articles for drug-target bioactivity data

Data used for the model training contains 28,075 positive examples (articles containing drug-target bioactivity data) and 28,075 negative examples (other biological articles), which is available at: https://dataset.drugtargetcommons.org/Training_DTIs_data/. We considered only those articles in the positive dataset that contain both drug and protein annotations in PubTator. Drug-target articles are extracted from DrugTargetCommons and ChEMBL (27th release), whereas data for other biological documents is extracted from DisGeNET [33]. We used DisGeNET as a negative dataset mainly because it is a comprehensive and manually curated database for disease and gene associations. Trained models are then used to predict documents that are likely to contain DTIs. Finally, the predicted documents likely to contain DTIs are associated with drug and protein entities as identified by PubTator.

Assay formats for drug-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, 12,845 having organism based and 1,148 as other assay formats (e.g., biochemical (93), cell-free (66), tissue-based (424) and physiochemical (565)). The training data for assays is available at https://dataset.drugtargetcommons.org/Training_assay_data/.

Proposed methods

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 articles on PubMed likely to contain DTIs. SciBERT is an MLM pre-trained model trained on 1.14 M full-texts from Semantic Scholar corpus with 82% from the biomedical domain [34]. 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 200 K steps and 270 K steps on PMC. The model is pre-trained using the same hyper-parameter settings as for the original BERT model. BioMed-RoBERTa is a MLM pre-trained language model based on the RoBERTa [31]. Finally, BlueBERT is pre-trained on approximately 4B words extracted from PubMed.

To fit the training data into the BERT models, we preprocessed it by applying the tokenization to break up the text into tokens. We used the class AutoTokenizer from the HuggingFace Transformers package [35]. It allows to instantiate a tokenizer for the selected BERT model and format the text by adding the special [CLS] token at the beginning of each text and [SEP] token at the end of the sentences. It also pads or truncates the resulting vectors to a standardized length limit of BERT model (512 tokens at a time).

We used the BERT representations for the classification task by fine-tuning the BERT variants with minimal changes applied during the training phase. All the BERT models used in this analysis comprised of 12 layers of transformer encoder with hidden state dimensions equal to 768 and having > 110 M parameters as adopted in [36]. 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. It is further processed by two 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, a maximum sequence length of 512, a learning rate of 2e-5 for DTIs task, 5e-5 for assay classification task, a maximum epoch size of 3 for DTI prediction task, and 9 for assay prediction task. We used Adam with $\beta_1 = 0.9$ and $\beta_2 = 0.999$, slanted triangular learning rates as in [25], warm-up portion to 0.1, and ensured that GPU memory is fully utilized. The model architecture for all the BERT models in this study is shown in Fig. 1.

Next, we divided the overall workflow into three modules:

1. To identify whether a PubMed article is likely to contain bioactivity data for drug-target interactions.
2. To extract drug and protein information by taking advantage of already extracted entities by PubTator.
3. To 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 drug-target relationship or not. The BERT models are trained on 28,075 positive and 28,075 negative documents as explained in the previous section. Each

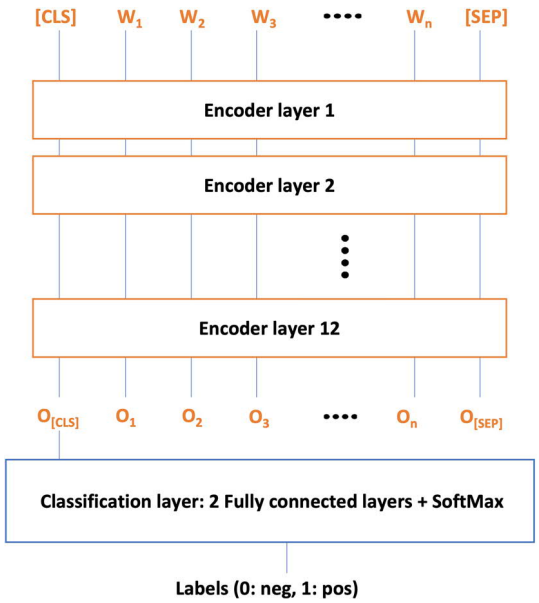

Fig. 1 Architecture for all the BERT models, where Wi represents input word token and O_i represents contextual embeddings at the output layer. The $O_{[CLS]}$ is first token of output sequence and contains class label

document is mapped into 768 numerical features with minor differences in the architecture of the five models. After training, individual BERT models are merged in a majority voting to identify new articles possibly containing DTIs. We used the majority voting because different BERT models performed differently on the external test datasets (Table 2) and the majority voting may reduce the risk of false positives. For module 2, we then matched and linked positively predicted documents with annotated drug and protein entities using the PubTator dataset. Finally, for module 3, using the same model architecture, we tried to predict assay formats (cell-based, organism based or other assays) for the positively predicted documents in module 2. We emphasized on the assay format prediction because assay formats are critical in defining the confidence scores for DTIs [37]. We reported these predicted articles in https://dataset.drugtargetcommons.org/New_predictions/. The workflow of the proposed strategy is shown in Fig. 2.

Results and discussions

Ten-fold cross-validation results using BERT models

The BERT text classifiers are trained using tenfold cross-validation. Our analyses showed that all the BERT models reached accuracies higher than 99%. Furthermore, we tested BERT models on three independent datasets i.e. DrugProt [38] (a positive dataset), Medline (a negative dataset) used by Papadatos et al., [39], and non-overlapping articles from ChEMBL (a positive dataset). As shown in Table 1, BioBERT achieved an accuracy of 71.5% on the DrugProt dataset, while BlueBERT was able to correctly identify negative articles from Medline with 100% accuracy, and SciBERT successfully identified positive articles from ChEMBL with 93.2% accuracy. Using manual curation, we also validated 100 DTIs articles (from 0.316 M articles at PubTator that are predicted as DrugTarget articles and contain both drug and protein entities). We confirmed that all the articles

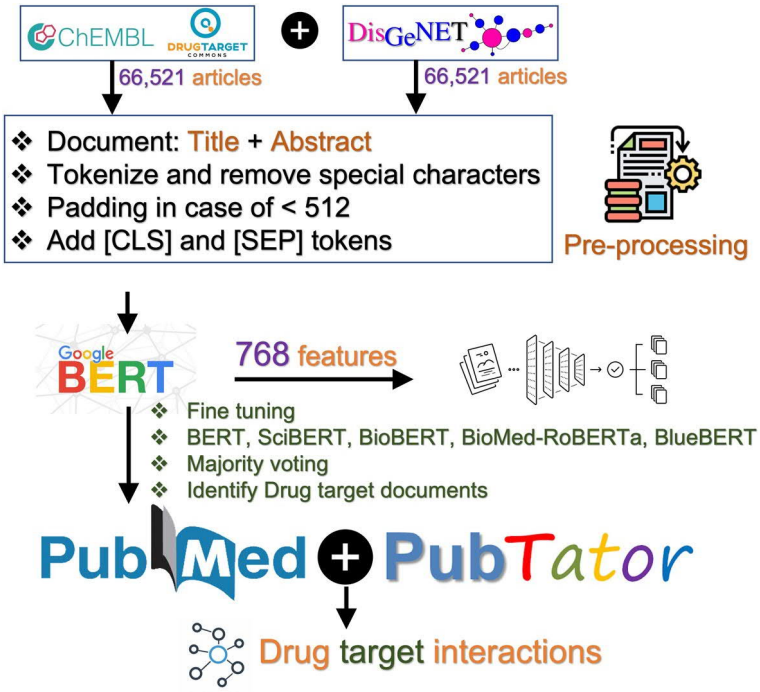

Fig. 2 Workflow for identifying new articles containing drug-target bioactivity data

Table 1 Accuracy of BERT models on three independent datasets. DrugProt is the dataset containing 2788 positive articles based on DTIs (positive class) and 1215 from negative articles class. Medline is a completely negative dataset, and ChEMBL is a completely positive dataset containing DTIs

Dataset	Articles	BERT	SciBERT	BioBERT	BioMed-RoBERTa	BlueBERT	Majority voting
DrugProt	4003	68	65.9	**71.5**	71.4	67.5	69.6
Medline	55,056	99.7	98.6	75.2	99.9	100	100
ChEMBL	876	89.6	93.2	91.2	83.4	88.7	90.3

Bold values indicate the top results for a dataset

contain relationship words (such as inhibition or binding) in the abstracts of the articles. These 100 articles are provided in Additional file 1. As the model was primarily trained on the subset of PubTator showing 99% accuracy, that is why we obtained 100% accuracy on those 100 articles. Articles in DrugProt datasets are more complex and are different from PubTator (0.31 M articles). Especially, DrugProt focusses on several types of relationships (including substrate, up regulators, down regulators, and others) which are not in the scope of current study. Therefore, we did not include those types of articles in model training, resulting in slightly reduced accuracy for the DrugProt dataset. However, high performance at Medline and ChEMBL datasets depicts the generalizability of BERT models to identify drug-target like articles with great precision.

We also compared the top frequently occurring words in both positive and negative documents. As shown in Fig. 3, the most frequently occurring words in drug target documents are 'compounds', 'activity' and 'potent', whereas the most frequent words for

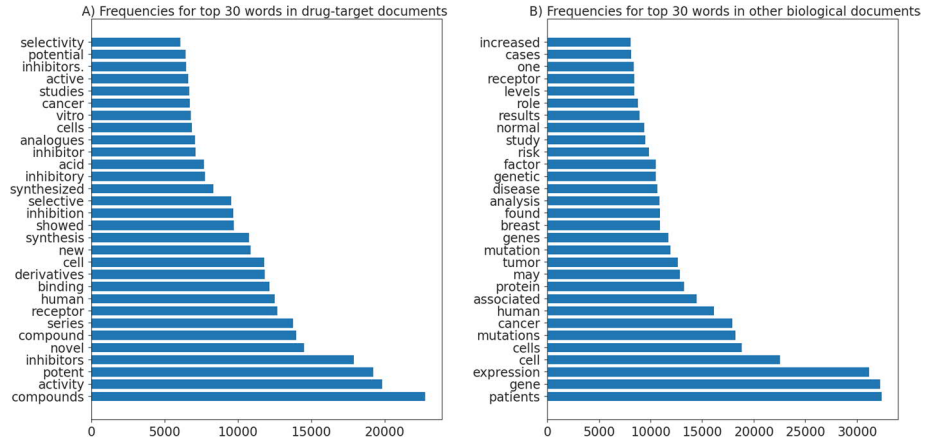

Fig. 3 Top word frequencies for **A** Drug-target documents, and **B** Other biological documents

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

BERT model	Predicted as drug-target articles	Articles containing drugs or proteins on PubTator	Articles containing both drugs and proteins on PubTator
BERT	688,206	682,150	342,902
SciBERT	594,999	589,999	321,831
BioBERT	636,091	630,132	340,638
BioMed-RoBERTa	725,748	720,030	**385,015**
BlueBERT	570,284	564,220	297,834
Majority voting	597,844	592,789	316,794

Bold value indicates the top result for a dataset

other biological documents are 'patients', 'gene', and 'expression'. The word distribution analysis can demonstrate developing a simple model based on word frequencies to identify drug-target or other biological documents. Simple model can have good time complexity but at the cost of lower accuracy.

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 18.5 M documents downloaded from PubTator. Table 2 shows the number of positively predicted documents out of these documents. The third column (articles containing drugs or proteins on PubTator) shows how many among positively predicted articles have either drug or protein entities annotated by PubTator. Finally, the last column indicates the number of articles for which PubTator annotated both drug and the protein entities. These two columns validate those articles that are identified as drug-target articles.

We obtained a superior performance on unseen articles. For example, using the majority voting, 99% (597,844) of the articles were identified as drug-target like (positive) containing either drugs or proteins entity identified by PubTator. Out of these positively predicted documents, 53% (316,794) contain both drug and protein entities according to PubTator. The result (53%) is likely an underestimation, as drug and protein entities may not appear in the main text of an article but may be deposited as supplementary data, which are not captured by PubTator's back-end algorithm. It is also possible that drug, and protein entities are present in the main text but were not captured by PubTator. This means that even though the article is positively predicted by our model, we might not be able to capture drug or protein entities in some cases, leaving the task for manual curators to check the supplementary material. Indeed, many high throughput drug-target profiling articles do not mention drug or protein names in the main text but instead provide these in the supplementary material, e.g. [40]. Of the BERT models, BioMed-RoBERTa identified more drug-target like documents compared to the other models, with at least 385,015 articles containing both drugs and proteins in PubTator.

We also analyzed the publication journals and years for these predicted articles. We found that Journal of Medicinal Chemistry, Bioorganic & Medicinal Chemistry Letters and Biological Chemistry are the top three journals based on our prediction (Fig. 4A). These three journals are also among the leading journals for bioactivity data extraction in ChEMBL [39]. Furthermore, most drug-target articles are from recent years, with the year of 2020 containing the most significant number of articles (Fig. 4B).

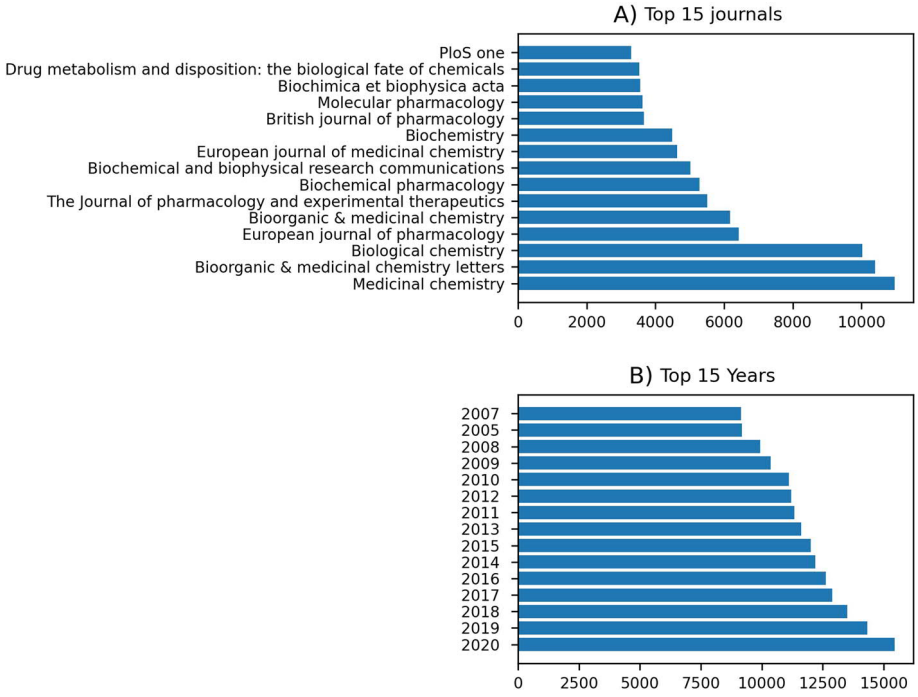

Fig. 4 **A** Top 15 journals for the articles that are predicted as drug-target based articles, **B** Top 15 years for articles predicted as drug-target articles

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 newly identified DTI articles.

Predict assay format for drug-target articles

After successfully identifying DTI articles, the next task is to predict the assay format that was reported. For that purpose, we separately trained each BERT model with 14,109 articles based on cell-based assay, 12,845 articles based on organism-based assay, and 1148 with the other assay formats. We used the same fine-tuning settings as for the drug-target article identification task and used F1 macro and F1 micro metrics to evaluate the performance of the models. We observed better accuracy improvements when using the weighted cross entropy (each class with a different weight based on effective number of samples) defined as:

$$\mathrm{WCE}\left(\mathrm{y},\tilde{\mathrm{y}}\right) = -\sum_{n}^{\mathrm{N}} \mathrm{w}_{y_n}.\mathrm{y}_n \log\left(\tilde{\mathrm{y}}_n\right)$$

where y is the target and w is the weight and the predicted class label $\tilde{y}$ is the index of the maximum predicted probability score among the three classes.

Table 3 shows tenfold cross-validation performances for the BERT models. We found that BioMed-RoBERTa outperforms the other models, with F1 micro of 88.1 ± 0.5, and F1 macro of 87.8 ± 0.5. The superior performance of BioMed-RoBERTa could be due to the additional pre-training over more data consisting of 2.68 M full-text papers from S2ORC [41] and the additional pre-training for longer steps.

After successful finetuning of BERT models to predict assay formats (of known articles), we used the best model (i.e., BioMed-RoBERTa) to predict assay formats for 597,844 articles (identified as drug-target articles using majority voting in Table 2). BioMed-RoBERTa predicted that 243,828 (out of 597,844) articles are cell based, 220,357 articles are organism-based articles, and 133,659 articles are others assays.

Discussions and conclusions

More than 80% of the approved drugs target only two protein classes: enzymes or receptors [42]. There are 25 000 genes in humans, but only 600 disease-modifying protein drug targets exist [43]. Therefore, target identification has recently shifted to other macromolecules, such as RNAs. Due to their involvement in gene regulation, miRNAs have been identified as high-value targets for therapy. There are

Table 3 The tenfold cross validation results for identifying assay formats

BERT model	F1 macro	F1 micro
BERT	81.0 ± 1	81.5 ± 1
SciBERT	85.2 ± 1.1	85.6 ± 1.1
BioBERT	86 ± 1.3	86.4 ± 1.3
BioMed-RoBERTa	**87.8 ± 0.5**	**88.1 ± 0.5**
BlueBERT	87.0 ± 1	87.5 ± 1

Bold values indicate the top results for a dataset

approximately 2000 miRNAs in humans (www.mirbase.org). They regulate 30% of all genes which are crucial in many biological processes [44, 44]. Therefore, traditionally 'undruggable' proteins can be targeted via their miRNA gene regulators, enabling the treatment of incurable diseases [46]. Recently in-silico methods have been developed to predict drugs for miRNA. For instance, Chen et al. proposed a bounded nuclear norm regularization method [47]. Niu et al. adapted graph neural network-based method to predict drug resistance for miRNAs [48]. Several more methods are published on in-silico drug associations with miRNA [49–51].

However, in this study, our focus is mainly on protein targets due to (1) insufficient miRNA targets available in the public databases and '(2) Lack of miRNAs annotations at PubTator [28], which is the main source of our pipeline. Therefore, we omitted miRNAs and focused only on protein targets in the present study. However due to the growing importance of miRNAs as emerging drug targets, in future, we aim to also include miRNA in text-mining based drug-target relationship extraction and combine it with machine learning-based prediction method to identify novel drug associations with miRNA.

In target centric drug discovery, a large number of compounds are tested across a particular target protein, resulting in the lack of DTI profiles at the proteomics level for many compounds. Curating quantitative drug-target bioactivity values reported in an article is therefore a critical task for establishing a more comprehensive drug-target profiles. Semi-automated NLP based methods can assist in identifying such articles and easing the workload for the data 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 likely containing DTIs.

Furthermore, we developed these models to predict the assay formats most likely used in the articles. Assays formats are critical in evaluating the quality for DTIs. We found that BioMed-RoBERTa performed slightly better than the other models for both drug-target article identification and assay format prediction.

Using the majority voting based on BERT models, we identified 597,844 articles from which 316,794 are confirmed to have both drug and protein entities in PubTator. Most of these 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 0.1 M. 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 these newly identified articles and curate bioactivity data. Meanwhile, we will try to extend our recently published method on drug target relationship extraction [52] to automatically identify DTI relationships from these articles.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12859-022-04768-x.

Additional file 1. Pmids for 100 articles that are manually validated to contain drug-target interactions.

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

Author contributions
ZT and JA performed data analysis, ZT, JA and JT wrote the manuscript, MV reviewed the manuscript. All authors read and approved the final manuscript.

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).

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

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
Authors have financial competing interests.

Received: 25 October 2021 Accepted: 3 June 2022
Published online: 21 June 2022

References

1. Dickson M, Gagnon JP. The cost of new drug discovery and development. Discov Med. 2009;4:172–9.
2. Shaughnessy AF. Old drugs, new tricks. BMJ. 2011;342: d741.
3. Tanoli Z, Seemab U, Scherer A, Wennerberg K, Tang J, Vähä-Koskela M. Exploration of databases and methods supporting drug repurposing: a comprehensive survey. Brief Bioinform. 2020;22(2):1656–78.
4. Gaulton A, Hersey A, Nowotka M, Bento AP, Chambers J, Mendez D, Mutowo P, Atkinson F, Bellis LJ, Cibrián-Uhalte E, Al E. The ChEMBL database in 2017. Nucleic Acids Res. 2016;45:D945–54.
5. Gilson MK, Liu T, Baitaluk M, Nicola G, Hwang L, Chong J, Al E, BindingDB in,. A public database for medicinal chemistry, computational chemistry and systems pharmacology. Nucleic Acids Res. 2015;44(2016):D1045–53.
6. Wang Y, Bryant SH, Cheng T, Wang J, Gindulyte A, Shoemaker BA, Thiessen PA, He S, Zhang J, BioAssay PubChem. Update. Nucleic Acids Res. 2017;45(2016):D955–63.
7. Alexander SPH, Fabbro D, Kelly E, Mathie A, Peters JA, Veale EL, Armstrong JF, Faccenda E, Harding SD, Pawson AJ. The concise guide to pharmacology 2019/20: catalytic receptors. Br J Pharmacol. 2019;176:S247–96.
8. Tanoli Z, Alam Z, Vähä-Koskela M, Ravikumar B, Malyutina A, Jaiswal A, Tang J, Wennerberg K, Aittokallio T. Drug Target Commons 2.0: a community platform for systematic analysis of drug–target interaction profiles. Database. 2018;1:1–13.
9. Tang J, Tanoli Z-R, Ravikumar B, Alam Z, Rebane A, Vähä-Koskela M, Peddinti G, van Adrichem AJ, Wakkinen J, Jaiswal A, Karjalainen E. Drug target commons: a community effort to build a consensus knowledge base for drug-target interactions, cell. Chem Biol. 2018;25:224–9. https://doi.org/10.1016/J.CHEMBIOL.2017.11.009.
10. Nguyen D-T, Mathias S, Bologa C, Brunak S, Fernandez N, Gaulton A, Hersey A, Holmes J, Jensen LJ, Karlsson A, Liu G, Ma'ayan, Mandava G, Mani S, Mehta S, Overington J, Patel J, Rouillard AD, Schürer S, Sheils T, Simeonov A, Sklar LA, Southall N, Ursu O, Vidovic D, Waller A, Yang J, Jadhav A, Oprea TI, Guha R. Pharos: Collating protein information to shed light on the druggable genome. Nucleic Acids Res. 2017;45(D1):D995–1002. https://doi.org/10.1093/nar/gkw1072.
11. Tanoli Z, Aldahdooh J, Alam F, Wang Y, Seemab U, Fratelli M, Pavlis P, Hajduch M, Bietrix F, Gribbon P, Zaliani A, Hall MD, Shen M, Brimacombe K, Kulesskiy E, Saarela J, Wennerberg K, Vähä-Koskela M, Tang J. 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.
12. Cichońska A, Ravikumar B, Allaway RJ, Wan F, Park S, Isayev O, Li S, Mason M, Lamb A, Tanoli Z, Jeon M, Kim S, Popova M, Capuzzi S, Zeng J, Dang K, Koytiger G, Kang J, Wells CI, Willson TM, Tan M, Huang C-H, Shih ESC, Chen T-M, Chih-Hsun W, Fang W-Q, Chen J-Y, Hwang M-J, Wang X, Guebila MB, Shamsaei B, Singh S, Nguyen T, Karimi M, Di W, Wang Z, Shen Y, Öztürk H, Ozkirimli E, Özgür A, Lim H, Xie L, Kanev GK, Kooistra AJ, Westerman BA, Terzopoulos P, Ntagiantas K, Fotis C, Alexopoulos L, Boeckaerts D, Stock M, De Baets B, Briers Y, Luo Y, Hailin H, Peng J, Dogan T, Rifaioglu AS, Atas H, Atalay RC, Atalay V, Martin MJ, Jeon M, Lee J, Yun S, Kim B, Chang B, Turu G, Misák Á, Szalai B,

Hunyady L, Lienhard M, Prasse P, Bachmann I, Ganzlin J, Barel G, Herwig R, Oršolić D, Lučić B, Stepanić V, Šmuc T, Oprea TI, Schlessinger A, Drewry DH, Stolovitzky G, Wennerberg K, Guinney J, Aittokallio T. Crowdsourced mapping of unexplored target space of kinase inhibitors. Nat Commun. 2021. https://doi.org/10.1038/s41467-021-23165-1.
13. Thafar MA, Olayan RS, Ashoor H, Albaradei S, Bajic VB, Gao X, Gojobori T, Essack M. DTiGEMS+: drug–target interaction prediction using graph embedding, graph mining, and similarity-based techniques. J Cheminform. 2020;12:1–17.
14. Zheng Y, Wu Z. A machine learning-based biological drug-target interaction prediction method for a tripartite heterogeneous network. ACS Omega. 2021;6:3037–45.
15. Sachdev K, Gupta MK. A comprehensive review of feature based methods for drug target interaction prediction. J Biomed Inform. 2019;93: 103159.
16. Anusuya S, Kesherwani M, Priya KV, Vimala A, Shanmugam G, Velmurugan D, Gromiha MM. Drug-target interactions: prediction methods and applications. Curr Protein Pept Sci. 2018;19:537–61.
17. Bagherian M, Sabeti E, Wang K, Sartor MA, Nikolovska-Coleska Z, Najarian K. Machine learning approaches and databases for prediction of drug–target interaction: a survey paper. Brief Bioinform. 2021;22:247–69.
18. White J. PubMed 2.0. Med Ref Serv Q. 2020;39(4):382–7. https://doi.org/10.1080/02763869.2020.1826228.
19. Sun C, Qiu X, Yige X, Huang X. How to fine-tune bert for text classification? In: Sun M, Huang X, Ji H, Liu Z, Liu Y, editors. Chinese computational linguistics: 18th China National Conference, CCL 2019, Kunming, China, October 18–20, 2019, Proceedings. Cham: Springer; 2019. p. 194–206. https://doi.org/10.1007/978-3-030-32381-3_16.
20. Beltagy I, Lo K, Cohan A, Scibert A. A pretrained language model for scientific text. 2019. ArXiv Prepr. ArXiv1903.10676.
21. Mikolov T, Sutskever I, Chen K, Corrado GS, Dean J. Distributed representations of words and phrases and their compositionality, in: Adv. Neural Inf. Process. Syst., 2013: pp. 3111–3119.
22. Pennington J, Socher R, Manning CD. Glove: global vectors for word representation, in: Proc. 2014 Conf. Empir. Methods Nat. Lang. Process., 2014: pp. 1532–1543.
23. McCann B, Bradbury J, Xiong C, Socher R Learned in translation: Contextualized word vectors. 2017. ArXiv Prepr. ArXiv1708.00107.
24. Peters ME, Neumann M, Iyyer M, Gardner M, Clark C, Lee K, Zettlemoyer L. Deep contextualized word representations. 2018. ArXiv Prepr. ArXiv1802.05365.
25. Howard J, Ruder S. Universal language model fine-tuning for text classification. 2018. ArXiv Prepr. ArXiv1801.06146.
26. Radford A, Narasimhan K, Salimans T, Sutskever I. Improving language understanding by generative pre-training, 2018.
27. Devlin J, Chang M-W, Lee K, Toutanova K. Bert: Pre-training of deep bidirectional transformers for language understanding. 2018. ArXiv Prepr. ArXiv1810.04805
28. Wei C-H, Kao H-Y, Lu Z. PubTator: a web-based text mining tool for assisting biocuration. Nucleic Acids Res. 2013;41:W518–22.
29. Lee S, Kim D, Lee K, Choi J, Kim S, Jeon M, Lim S, Choi D, Kim S, Tan A-C. BEST: next-generation biomedical entity search tool for knowledge discovery from biomedical literature. PLoS ONE. 2016;11: e0164680.
30. Lee J, Yoon W, Kim S, Kim D, Kim S, So CH, Kang J. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics. 2020;36:1234–40.
31. 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).
32. Peng Y, Yan S, Lu Z. Transfer learning in biomedical natural language processing: an evaluation of BERT and ELMo on ten benchmarking datasets. 2019. ArXiv Prepr. ArXiv1906.05474.
33. Piñero J, Bravo À, Queralt-Rosinach N, Gutiérrez-Sacristán A, Deu-Pons J, Centeno E, García-García J, Sanz F, Furlong LI. DisGeNET: a comprehensive platform integrating information on human disease-associated genes and variants. Nucleic Acids Res. 2017;45:D833–9.
34. 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).
35. T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, Transformers: State-of-the-art natural language processing, in: Proc. 2020 Conf. Empir. Methods Nat. Lang. Process. Syst. Demonstr., 2020: pp. 38–45.
36. 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.
37. Tanoli Z, Alam Z, Ianevski A, Wennerberg K, Vähä-Koskela M, Aittokallio T. 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.
38. A. Miranda, F. Mehryary, J. Luoma, S. Pyysalo, A. Valencia, M. Krallinger, Overview of DrugProt BioCreative VII track: quality evaluation and large scale text mining of drug-gene/protein relations, in: Proc. Seventh BioCreative Chall. Eval. Work., 2021.
39. Papadatos G, van Westen GJP, Croset S, Santos R, Trubian S, Overington JP. A document classifier for medicinal chemistry publications trained on the ChEMBL corpus. J Cheminform. 2014;6:1–8.
40. Anastassiadis T, Deacon SW, Devarajan K, Ma H, Peterson JR. Comprehensive assay of kinase catalytic activity reveals features of kinase inhibitor selectivity. Nat Biotechnol. 2011;29:1039–45.
41. K. Lo, L.L. Wang, M. Neumann, R. Kinney, D.S. Weld, S2ORC: The semantic scholar open research corpus, ArXiv Prepr. ArXiv1911.02782. (2019).
42. Overington JP, Al-Lazikani B, Hopkins AL. How many drug targets are there? Nat Rev Drug Discov. 2006;5:993–6.
43. Hopkins AL, Groom CR. The druggable genome. Nat Rev Drug Discov. 2002;1:727–30.
44. Dykxhoorn DM, Novina CD, Sharp PA. Killing the messenger: short RNAs that silence gene expression. Nat Rev Mol Cell Biol. 2003;4:457–67.
45. Fabian MR, Sonenberg N. The mechanics of miRNA-mediated gene silencing: a look under the hood of miRISC. Nat Struct Mol Biol. 2012;19:586–93.

46. Schmidt MF. Drug target miRNAs: chances and challenges. Trends Biotechnol. 2014;32:578–85.
47. X. Chen, C. Zhou, C.-C. Wang, Y. Zhao, Predicting potential small molecule–miRNA associations based on bounded nuclear norm regularization, Brief. Bioinform. 22 (2021) bbab328.
48. Y. Niu, C. Song, Y. Gong, W. Zhang, MiRNA-Drug Resistance Association Prediction Through the Attentive Multimodal Graph Convolutional Network., Front. Pharmacol. 12 (2021) 799108.
49. P. Pandey, P.K. Srivastava, S.P. Pandey, Prediction of plant miRNA targets, in: Plant MicroRNAs, Springer, 2019: pp. 99–107.
50. Xu P, Wu Q, Rao Y, Kou Z, Fang G, Liu W, Han H. Predicting the influence of MicroRNAs on drug therapeutic effects by random walking. IEEE Access. 2020;8:117347–53.
51. Qu J, Chen X, Sun Y-Z, Zhao Y, Cai S-B, Ming Z, You Z-H, Li J-Q. In Silico prediction of small molecule-miRNA associations based on the HeteSim algorithm. Mol Ther Acids. 2019;14:274–86.
52. Aldahdooh J, Tanoli Z, Tang J. R-BERT-CNN: Drug-target interactions extraction from biomedical literature, BioCreative Challenge VII Track 1 submission. In: Proceedings of the seventh BioCreative challenge evaluation workshop. 2021.

Publisher's Note