nature machine intelligence

ARTICLES

https://doi.org/10.1038/s42256-020-0189-y

Check for updates

A novel machine learning framework for automated biomedical relation extraction from large-scale literature repositories

Lixiang Hong[1], Jinjian Lin[1], Shuya Li[1], Fangping Wan[1], Hui Yang[2], Tao Jiang[3,4,5], Dan Zhao[1]✉ and Jianyang Zeng[1,4]✉

Knowledge about the relations between biomedical entities (such as drugs and targets) is widely distributed in more than 30 million research articles and consistently plays an important role in the development of biomedical science. In this work, we propose a novel machine learning framework, named BERE, for automatically extracting biomedical relations from large-scale literature repositories. BERE uses a hybrid encoding network to better represent each sentence from both semantic and syntactic aspects, and employs a feature aggregation network to make predictions after considering all relevant statements. More importantly, BERE can also be trained without any human annotation via a distant supervision technique. Through extensive tests, BERE has demonstrated promising performance in extracting biomedical relations, and can also find meaningful relations that were not reported in existing databases, thus providing useful hints to guide wet-lab experiments and advance the biological knowledge discovery process.

Knowledge bases play an important role in the development of biomedical science. Most structured databases, such as DrugBank[1], CTD[2], SIDER[3] and BioGRID[4], are curated from a large number of scientific articles by human experts, who expend huge amounts of time and effort. Biomedical information extraction technology aims to shift this time-consuming and tedious burden to machines by developing efficient computational tools to extract meaningful facts from vast unstructured texts automatically[5,6]. After that, often with some human curation, the extracted data can be fed into the downstream tasks to facilitate the related biological knowledge discovery processes.

The information that biomedical researchers most care about generally falls into three types: biomedical entities, relations (interactions or associations between entities) and events (important facts or findings attached to at least one entity). In this work, we mainly focus on the second type—biomedical relations between entities described in the sentences. Such relations, like drug–drug interactions (DDIs) and drug–target interactions (DTIs), are generally significant and useful for many biomedical applications. For example, early detection of DDIs provides an effective way to prevent adverse drug reactions (ADRs)[7], while computational prediction of DTIs is a crucial step in the drug repositioning process, which aims to find novel targets of existing drugs[8–10]. Traditionally, such relations are sorted out through manual curation from the literature. With the rise of natural language processing (NLP) and machine learning techniques, automated biomedical relation extraction (BioRE) has been used to accelerate this process[11–13].

The BioRE task is often formulated as a classification of biomedical relations between entities from a set of sentences[14,15], with the supervision of relation annotated texts. However, collecting such labelled text data is often laborious. To alleviate this, distant supervision[16,17], in which all sentences mentioning the same pairs of entities are labelled by the relation facts reported in a certain knowledge base, has been proposed to expand the labelled datasets. It assumes that if two entities are involved in a relation, at least one sentence that mentions both entities suggests this relation. Based on this assumption, the related distantly supervised learning task can be then transformed into a multi-instance learning task[18]. More specifically, given a pair of entities with a bag of sentences that may be suggestive of the relation between the two entities, a label representing the relation between such a pair of entities is learned. So far, several distantly supervised datasets[16,17,19] have been constructed for relation extraction (RE). For example, the Riedel dataset[17] aligns Freebase relations with the *New York Times* corpus, and has been widely used as a benchmark dataset for evaluating different RE models. Despite successful application of distant supervision for a number of RE tasks[20–22], it remains unknown whether this new technique can be applied to extract meaningful biomedical relations that can yield useful insights to discover new biomedicine findings.

Recently, neural network-based RE models have become a popular tool for automatically extracting entity relations from unstructured texts[23–25]. These approaches often use models based on convolutional neural networks (CNNs) or recurrent neural network (RNNs) to learn semantic representations of each sentence, but often ignore the syntactic features of sentences. Models based on recursive neural networks (RvNNs)[26–28], by contrast, explicitly model syntactic features by recursively propagating information from the bottom and up through a sentence constituency-based parse tree (that is, a constituent structure that organizes words into nested phrases) and have achieved better prediction results than other methods. As concluded in ref. [29], recursive models are generally more suited to dealing with tasks (such as semantic relation

[1]Institute for Interdisciplinary Information Sciences, Tsinghua University, Beijing, China. [2]Silexon Co. Ltd, Nanjing, China. [3]Bioinformatics Division, TNLIST, MOE Key Laboratory of Bioinformatics and Center for Synthetic and Systems Biology, Tsinghua University, Beijing, China. [4]MOE Key Laboratory of Bioinformatics, Tsinghua University, Beijing, China. [5]Department of Computer Science and Engineering, University of California, Riverside, CA, USA. ✉e-mail: zhaodan2018@tsinghua.edu.cn; zengjy321@tsinghua.edu.cn

extraction) that require feature representations of long-distance associations between words of interest. Although these strategies have been shown to be useful in the RE task, they still have critical drawbacks, such as relying on external parsers to parse sentences and incompatibility with mini-batch training due to the variation of the employed tree structures[30].

Unlike the recursive models that encode parse trees explicitly, latent tree learning aims to understand sentence structures implicitly by learning how to parse sentences with indirect supervision from the prediction results of the downstream tasks, and has achieved great success in natural language inference and sentiment analysis tasks[31–33]. With latent tree learning, the parsing process can be carried out completely inside a neural network and thus tailored to the task of interest. Gumbel Tree-LSTM[33] is an example of latent tree learning, which is trained to select the most appropriate composition of feature vectors among adjacent words or phrases, one at a time, to construct the constituency-based parse trees. In addition, the self-attention mechanism has recently gained great popularity in image recognition and machine translation fields[34,35], mainly for its advantages in capturing long-range dependencies. Overall, both latent tree learning and self-attention techniques are suitable for capturing syntactic information and long-range dependencies in sentences. However, despite the advantages of the two techniques, they have rarely been used in the past to advance the RE task.

Inspired by the above observations, we propose a new machine learning framework, called BERE, for automated biomedical entity relation extraction from large-scale biomedical literature repositories. BERE applies latent tree learning and self-attention to fully exploit the semantic and syntactic information inside sentences, as well as the short- and long-range dependencies between words. BERE further adopts a scoring mechanism to evaluate the importance of each sentence in supporting a relation prediction. In addition, BERE employs a multi-instance learning framework with a distant supervision technique to greatly alleviate laborious human annotation and expand the training data to improve the prediction results. Through extensive tests on an existing single-sentence annotated DDI dataset and a distantly supervised DTI dataset, we have demonstrated that the proposed BERE framework outperforms the state-of-the-art models in biomedical relation extraction. Moreover, after applying a well-trained BERE model to widely distributed biomedical literature texts, we found that the extraction results can provide useful hints for discovering novel DTIs (not reported previously in current widely used databases). With experimental validation through wet-lab assays, we successfully identified two potential targets of the multi-target kinase inhibitor nintedanib. All these results suggest that BERE can serve as a powerful tool in biomedical relation extraction and provide useful assistance in the discovery of novel relations such as DTIs.

Results

Overview of BERE. The architecture of our proposed BERE framework is shown in Fig. 1a. Given a pair of entities (Entity1, Entity2) co-mentioned in a bag of sentences, BERE first represents each word (the representation is also called a word vector) in a sentence by concatenating its word embedding and part-of-speech (POS) embedding. Each word vector is then fed into a self-attention layer to capture long-range dependency, which is added back to the original word vector through a residual connection. Next, BERE uses a bidirectional gated recurrent unit (Bi-GRU) to encode the local contextual features of each word, followed by a Gumbel Tree-GRU, which uses a greedy-based strategy to find the best composition scheme (marked by green edges in Fig. 1a) among all feasible schemes (marked by red edges in Fig. 1a). Figure 1b gives an example of one operation in the Gumbel Tree-GRU layer. In step t, the example sentence is represented as a sequence of four vectors (indicating 'Entity1', 'correlates', 'with' and 'Entity2', respectively). Next, all adjacent vectors are composed through a shared Tree-GRU cell to obtain three candidate vectors, which are then individually assessed by a scoring function. In step $t+1$, the candidate vector with the highest score (that is, 'correlates with') is selected. Other vectors are copied directly from step t (that is, 'Entity1' and 'Entity2'). When all words are composed, the resulting final vector is basically the feature representation of the whole sentence. To capture the associations between target entities, BERE further embeds the contextual features of entities into a concatenated sentence representation. Finally, BERE uses an attention-based sentence aggregation scheme to calculate the weighted sum of concatenated features over the bag of sentences, which is then fed into a softmax classifier to predict the relation between Entity1 and Entity2. More details about BERE are provided in the Methods.

Tests on the single-sentence annotated DDI'13 dataset. We performed extensive tests on the DDI'13 dataset (see Methods) to compare the performance of BERE with those of six other state-of-the-art DDI extraction methods: SCNN[36], CNN-bioWE[37], MCCNN[38], Joint AB-LSTM[39], RvNN[15] and Position-aware LSTM[40]. Among the CNN-based methods, SCNN, CNN-bioWE and MCCNN use syntax word embeddings, pretrained word embeddings and multi-channel word embeddings, respectively. Joint AB-LSTM and Position-aware LSTM are RNN-based methods, both incorporating an attention mechanism into a bidirectional LSTM network to enhance prediction. Similar to BERE, RvNN also propagates information through the constituency-based parse tree, but it requires pre-parsed sentences as inputs. We trained each model to classify the relation between a pair of drugs mentioned in a sentence into one of five DDI types and evaluated its performance on the test set using the micro-averaged F_1 score, as in previous works[15,36]. The F_1 score for a typical classification problem is defined as $(2PR)/(P+R)$, where P denotes the precision and R denotes the recall.

To enable batched computation, we padded or cropped each sentence to a fixed length, 60 words, which is longer than 85% of sentences in the DDI'13 dataset, to obtain a good trade-off between the efficiency and accuracy of our framework. We also applied the dropout mechanism[41] after input representation and before the classifier to alleviate overfitting. The learning rate in the training process was fine-tuned on the validation set using a grid search among {0.0001, 0.0002, …, 0.001}. More details on hyperparameter calibration are provided in Extended Data Fig. 2. Table 1 shows the performance of all the methods for DDI extraction on the DDI'13 dataset. Our proposed BERE model yielded an F_1 score of 73.9%, which outperformed all the other baseline approaches. Compared to RvNN, our method does not need any external parser to construct the parse trees and is also compatible with mini-batch training. According to the results of our ablation studies (Table 1), BERE still yielded a decent performance, even when part of the framework was removed. Overall, the ablation studies further demonstrated the effectiveness of each part in our framework.

Tests on the distantly supervised DTI dataset. To better verify the effectiveness of BERE on the distantly supervised dataset, we further made a comparison between BERE and other representative distant supervision-based RE methods on a distantly supervised DTI dataset (see Methods), in which each drug–target relation was supported by a bag of sentences. Among all the baseline methods, PCNN-AVE[21] and PCNN-ATT[23] adopt a similar CNN-based neural network to encode each sentence, but apply different sentence aggregation strategies (that is, an averaging strategy and an attention-based strategy, which represent each bag of sentences as a mean vector and a weighted sum vector of sentences inside the bag, respectively), BiGRU-ATT and BiGRU-2ATT[24] both apply an RNN-based neural network with an attention-based sentence aggregation strategy and BiGRU-2ATT employs another word-level attention to weigh the

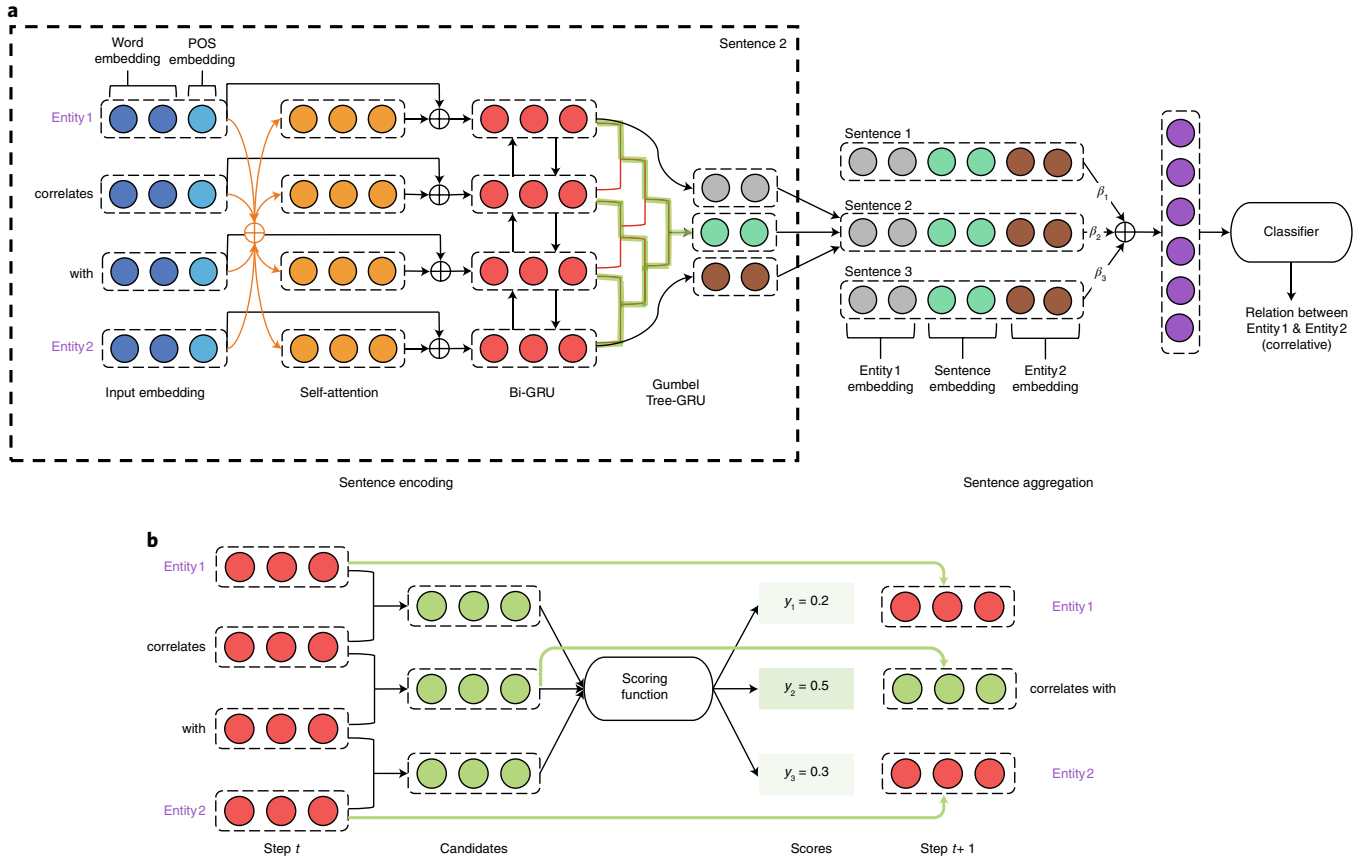


Fig. 1 | The architecture of BERE. a, Schematic overview of BERE. BERE first encodes each sentence in the bag through a self-attention layer to capture long-range dependencies. It then uses a Bi-GRU to capture local contextual features and applies a Gumbel Tree-GRU to organize words into nested phrases. Next, the contextual features of Entity1 and Entity2 are concatenated with the sentence features obtained by the Gumbel Tree-GRU to further capture the associations between entities. After that, an attention-based sentence aggregation strategy is employed to aggregate the concatenated features over the bag of sentences. Finally, BERE uses a softmax classifier to predict the relation between Entity1 and Entity2. The symbol ⊕ denotes vector addition. Green edges represent a constituency-based parse tree and red edges represent all possible constituent structures. **b**, An example of one operation step in the Gumbel Tree-GRU. At step t, the parent candidate for all adjacent feature vectors is computed through a shared Tree-GRU cell, which is then accessed by a scoring function to select the best composition scheme. At step $t+1$, the selected feature vector ('correlates with', with $y_2 = 0.5$) replaces the original vectors ('correlates' and 'with') and the remaining parts are copied directly from step t ('Entity1' and 'Entity2'). More details about the BERE framework are provided in the main text.

Table 1 | Test results on the DDI'13 dataset

Methods	*P* (%)	*R* (%)	*F* (%)
SCNN[36]	69.1	65.1	67.0
CNN-bioWE[37]	75.7	64.7	69.8
MCCNN[38]	75.9	65.2	70.2
Joint AB-LSTM[39]	73.4	69.6	71.5
RvNN[15]	74.4	69.3	71.7
Position-aware LSTM[40]	75.8	70.4	73.0
BERE (ours)	**76.8**	**71.3**	**73.9**
No self-attention	75.4	68.1	71.5
No Bi-GRU	71.3	68.0	69.6
No Gumbel Tree-GRU	71.0	69.1	70.0
No concatenated embedding	75.9	69.4	72.5

The top six rows show the results of the state-of-the-art baseline methods. The bottom five rows show the results of BERE and corresponding ablation studies. The ablation studies were carried out by removing each component from BERE. 'No self-attention' removed the self-attention layer, 'No Bi-GRU' removed the Bi-GRU component, 'No Gumbel Tree-GRU' removed the Gumbel Tree-GRU component and 'No concatenated embedding' removed entity embeddings in the concatenated representation of the sentence.

attribution of each word to the final prediction. We trained each model to classify the relation between a drug and target pair mentioned in a bag of sentences into one of six DTI types and evaluated its performance on the test set using the precision–recall curve, the area under the precision–recall curve (AUPRC) and the F_1 score, as in previous works[20,23]. The other settings were the same as in previous tests on the DDI'13 dataset (the hyperparameter settings are provided in Extended Data Fig. 2). Compared to all the baseline methods, BERE achieved the highest precision scores in most of the recall ranges (Fig. 2a). Overall, BERE yielded an AUPRC of 0.524 and F_1 score of 0.625, which were 6.7% and 4.6% better than the second best, respectively. The outperformance by BERE over PCNN-ATT and BiGRU-2ATT indicates that taking the sentence structure into account is helpful for relation extraction. In addition, the performance improvement of PCNN-ATT compared with PCNN implies an advantage of using the attention-based sentence aggregation strategy in distantly supervised relation extraction. The higher classification performance of BiGRU-2ATT compared with BiGRU-ATT also verifies the effectiveness of the word-level attention in the BioRE task.

To further compare the performance of BERE with the performances of alternative approaches using other sentence aggregation strategies, we also performed another test, as shown in

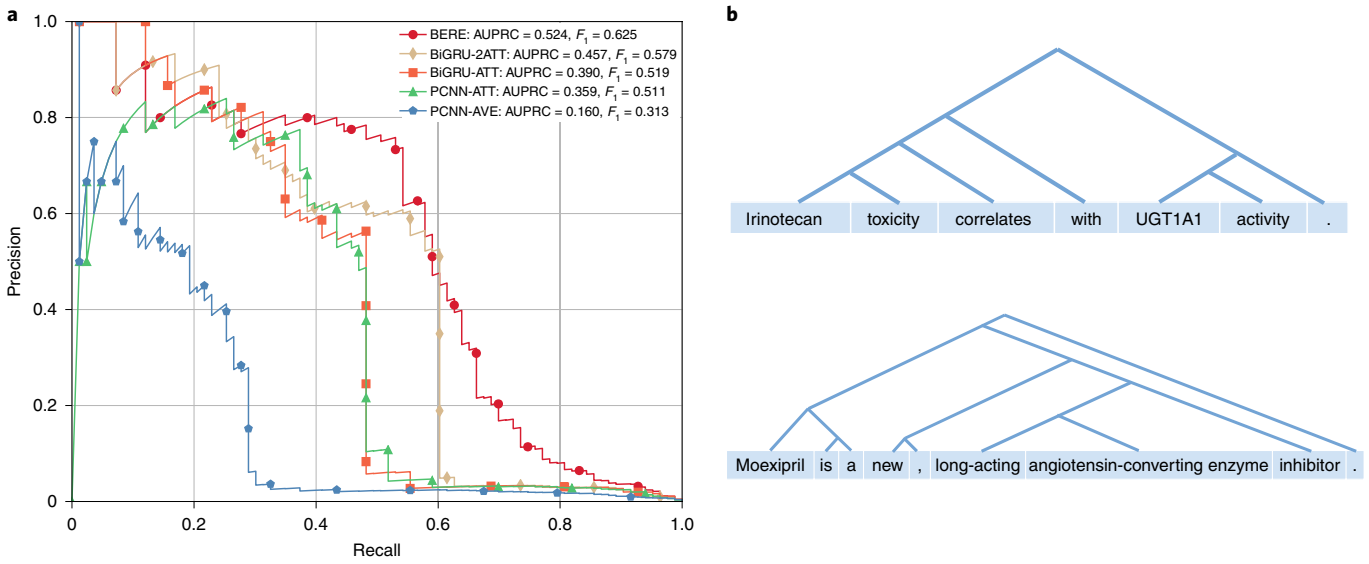


Fig. 2 | Test results on the distantly supervised DTI dataset. a, Comparisons of the precision–recall curves between BERE and other state-of-the-art baseline methods. The legend on the top right contains the AUPRC and F_1 score for each method. **b**, Examples of the parse trees constructed by BERE. For the sentence 'Irinotecan toxicity correlates with UGT1A1 activity', irinotecan is the drug and UGT1A1 is the target. For the sentence 'Moexipril is a new, long-acting angiotensin-converting enzyme inhibitor', moexipril is the drug and angiotensin-converting enzyme is the target.

Table 2 | Sentences with the highest scores in identifying the kinase targets of nintedanib produced by BERE

Target	Sentences with the highest scores	PMID
PLK1	Volasertib (BI 6727) is a potent **Plk-1** inhibitor which induces cell cycle arrest and apoptosis and was administered in combination with an angiokinase inhibitor **nintedanib** (BIBF 1120) in a phase I dose-escalation study.	25784931
mTOR	Furthermore, **nintedanib**, which blocks VEGFR2, RET, ERK1,2 and PI3K/AKT/FOXO1 like Vandetanib, also inhibits PI3K/AKT/**mTOR**, but may still have limited long-term anti-tumour effects on MTC due to the development of resistance.	30701022
AAK1	It is highly probable that this is the mechanism of the observed selective inhibition of BIKE over **AAK1** by **nintedanib**, although further crystal structures would be required to confirm our proposed binding mode.	26853940
ERBB2	**ERBB2** is also a target gene of the FDA approved drug **nintedanib**, which inhibits it.	28974751
JAK2	Since PDGFβ has been reported to induce the JAK2-STAT3 pathway by activating Src, **nintedanib** might inhibit **JAK2** by directly inhibiting PDGFβ and Src.	28798401
EGFR	In particular, agents that target the **EGFR** or the VEGFR, such as **nintedanib**, are associated with GI events in patients with NSCLC.	30643547
TGFBR1	Since **nintedanib** blocks EMT progression in NMuMG cells with an IC_{50} in the lower micromolar range and is able to block **TGFBR1** activity in biochemical assays in the submicromolar range, it is plausible that TGFBR inhibition contributes to its beneficial effects in vivo.	27036020
AXL	Multitargeted kinase inhibitors include a MET, RET, VEGFR, KIT, FLT-3, TIE-2, TRKB, **AXL** inhibitor (cabozantinib), and a VEGFR, FGFR, PDGFR, SRC, LCK, LYN, FLT-3 inhibitor (**nintedanib**).	28435296
ABL	**Nintedanib** is a tyrosine kinase inhibitor of VEGFR, PDGFR and FGFR, in addition to the Src and **Abl** tyrosine kinases.	28435296
RET	**Nintedanib** is a multitargeted angiokinase inhibitor against many growth factor receptors, including PDGFR, FGFR, VEGFR, as well as the proto-oncogenes **RET**, FTL3 and Src, with anti-angiogenic activity.	28798401

PMID stands for the PubMed Unique Identifier of the corresponding article to which the sentence belongs. The drug and target of interest are highlighted in bold.

Extended Data Fig. 1. In the alternatives of BERE, BERE-POOL uses the max-pooling[42] scheme to aggregate sentence features, while BERE-AVE employs the averaging strategy, as in PCNN-AVE. We found that BERE achieved better performance than its alternatives, indicating that the attention-based sentence aggregation strategy can successfully filter out those irrelevant sentences. The BERE-POOL method achieved a performance competitive with BERE, indicating that the max-pooling aggregation strategy can also make the model focus on the important features among sentences. However, unlike the aggregation strategy used in BERE, the max-pooling aggregation scheme cannot weigh the attribution of each sentence to the final prediction and thus may fail to select representative sentences to facilitate manual review. The performance of the BERE-AVE method (which considers each sentence equally) dropped significantly, which demonstrates that taking those irrelevant sentences into the prediction may degrade prediction performance.

Figure 2b provides two examples of tree structures built by BERE. Both examples demonstrate that BERE can parse a sentence in a human-like manner. Overall, all the above results show that BERE can yield satisfactory performance in automated relation extraction from a distantly supervised dataset. BERE was implemented on an NVIDIA GTX 1080 GPU; its training generally takes hours, for example, ~7 h on the DTI dataset and ~1 h on the DDI'13 dataset.

Table 3 | Sentences with the lowest scores in identifying the kinase targets of nintedanib produced by BERE

Target	Sentences with the lowest scores	PMID
PLK1	Quantification of CDK1, CDK4, CCNA2 and **PLK1** gene transcripts by RT-PCR in the primary lung-resident fibroblasts isolated from human IPF lung cultures and treated with vehicle or **nintedanib** (1 μM) for 16 h.	31156440
mTOR	It may also be important to determine if a Vandetanib/**mTOR** inhibitor combination or a **nintedanib** monotherapy is the most beneficial through future patient-centred studies.	30701022
AAK1	An interesting hit was **nintedanib**, a tyrosine kinase inhibitor in development for the treatment of idiopathic pulmonary fibrosis, which has a 10-fold higher affinity for BIKE than **AAK1**.	26853940
ERBB2	On the other hand, at least five additional LC patients can be treated with targeted inhibitors such as crizotinib (MET), **nintedanib** (FGFRs), trastuzumab (**ERBB2**) or buparlisib (PI3KCA).	29854313
JAK2	However, as shown in supplementary figure 4, **nintedanib** treatment had no obvious effect on either PTEN or SHP-2 but suppressed the phosphorylation of **JAK2** (Tyr1007/1008) and Src (Tyr 416).	28798401
EGFR	The CM was used to stimulate the growth and invasion of a panel of ADC and SCC cell lines that were selected based on their **EGFR** and KRAS wild-type status to mimic key genetic features of those patients that may be treated with **nintedanib**.	28898237
TGFBR1	On the other hand, inhibitors reported to target TGFBR, together with the multi-kinase inhibitors **nintedanib**, pazopanib and sorafenib revealed a significant correlation between their efficacy in blocking EMT and their inhibition of **TGFBR1** or 2.	27036020
AXL	Western blotting showed that crizotinib at 2 μM effectively suppressed the phosphorylation of MET in PC9-GR1, while BGB324 at the same concentration inhibited the activation of **AXL** in PC9-ER, and **nintedanib** the phosphorylation of FGFR substrate 2.	31000705
ABL	These data suggest that inhibition of TGFβ-signalling contributes to the therapeutic efficacy of **nintedanib** in IPF patients, either indirectly through c-**ABL** and/or ERK.	27036020
RET	Romidepsin had no effect on phosphorylation of **RET**, VEGFR2 or ERK1/2, while **nintedanib** alone or in combination with romidepsin lowered these signals.	30701022

PMID stands for the PubMed Unique Identifier of the corresponding article to which the sentence belongs. The drug and target of interest are highlighted in bold.

BERE provides useful insights into identifying potential DTIs. As mentioned previously, BERE can help biologists and clinicians better find and understand the relations between biomedical entities. In this section, we used a case study to show that BERE can help identify meaningful DTIs that were not reported in DrugBank. We also illustrate how to use BERE to select representative sentences supporting the extracted relations from biomedical literature.

We took the inhibitory capacity of nintedanib (a kinase inhibitor) against specific kinases as an example. We first used a dictionary-based named entity recognition (NER) approach (see Methods) to collect all sentences mentioning both nintedanib and a kinase from ~2.2 million PubMed Central (PMC) full-text articles (abstracts were removed to ensure that our model did not make a trivial extraction from the training data). Next, we applied a well-trained BERE model to extract the nintedanib–kinase interactions at the level of a bag of sentences and then checked whether the selected representative sentences made biological sense or not. More specifically, in each bag of sentences co-mentioning a certain nintedanib–kinase pair, we ranked the sentences according to the scores calculated by BERE (see Methods). Obviously, the sentence with the highest score plays a decisive role in selecting the relation and is naturally selected as a representative one. In Table 2, we list the top 10 extracted kinase targets of nintedanib (excluding those already reported in DrugBank) and their representative sentences. We find that, except PLK1 and AXL, all the predicted nintedanib targets can be directly supported or at least suggested by these representative sentences. To further examine the ability of BERE in distinguishing different sentences, we also list the sentences with the lowest scores in Table 3 and compare these with the results in Table 2. Most sentences in Table 2 with the highest scores indicate a true inhibitory effect of nintedanib on the targets, while those in Table 3 with the lowest scores are mostly irrelevant. This comparison result shows that BERE can successfully identify the 'important' evidence, in turn helping the model to better extract entity relations.

To go beyond relation extraction and further validate each predicted kinase target in a more rigorous way, we carefully inspected the relevant sentences and the corresponding articles or references. Among these targets identified by BERE (Table 2), we found that BERE extracted suggestive rather than true relations of nintedanib–JAK2 and nintedanib–EGFR; that is, the original articles did not show any direct evidence to support their interactions, but instead mentioned some indirect relations between them, which were different from those meaningless misclassifications. For example, in the paper mentioning the nintedanib–JAK2 relation[43], the authors only reported that the expression level of phosphorylated JAK2 could be reduced upon nintedanib treatment, and speculated that nintedanib might inhibit JAK2 by directly suppressing PDGFβ and Src, but no direct inhibition of JAK2 by nintedanib was observed. In the paper mentioning the nintedanib–EGFR pair[44], the interaction between EGFR and nintedanib was not examined experimentally, but the authors indicated that the EGFR upregulation can be blocked by nintedanib. Overall, all these articles only indicated that these two kinases are indirectly related to nintedanib. To show that these extraction results from our algorithm are meaningful, we further performed kinase activity assays (see Methods) to verify these interactions. The new wet-lab experiments revealed that nintedanib inhibits the kinase activities of JAK2 and EGFR effectively, with half-maximum inhibitory concentration (IC_{50}) values of 290.1 nM and 3,463.5 nM, respectively (Fig. 3). The above experimental validation confirmed the inhibitory capacity of nintedanib against JAK2 and EGFR, which thus implied that, in addition to mining the unstructured texts automatically from the rich literature data to accurately derive the real biomedical relations and hence expand the existing knowledge bases, BERE may also offer

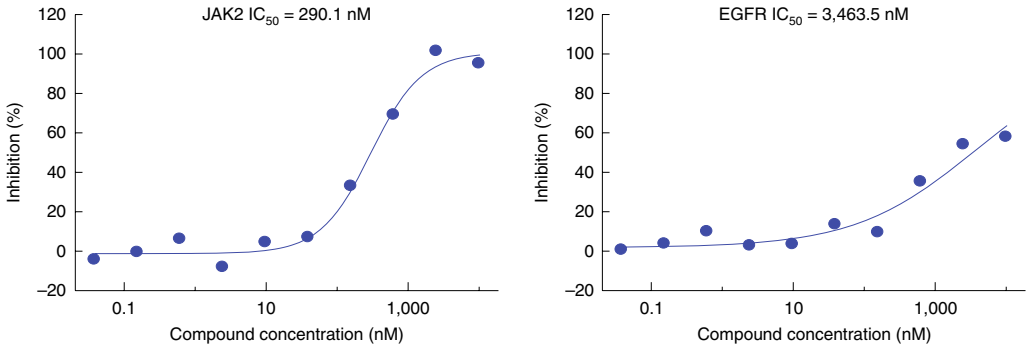


Fig. 3 | The in vitro inhibitory activity of nintedanib against JAK2 and EGFR. The inhibitory activity of nintedanib against JAK2 and EGFR kinases was measured by mobility shift assays with adenosine-5′-triphosphate (ATP) concentrations at Michaelis–Menten constant (K_m). The in vitro kinase assays were performed with the purified kinases/recombinant kinase domains in increasing concentrations of nintedanib (see Methods). IC_{50} values were calculated using the XLFit excel add-in[45].

useful insights into discovering novel interactions between drugs and targets to advance the drug development process.

Discussion

In this work, we propose BERE, a novel machine learning framework, to automatically extract biomedical relations from vast unstructured literature. By parsing sentences with latent tree learning, capturing short- and long-range dependencies through Bi-GRU and self-attention mechanisms and incorporating the local contextual features of entities into sentence encoding, BERE can fully exploit the sentence information from both semantic and syntactic aspects. Although this hybrid feature representation method may introduce more complexity to the model, the resulting overhead mainly lies in an increase in training time. Once BERE is well-trained, users can apply it to quickly extract the corresponding relations from the widely distributed texts in the literature.

We also looked into the details of the misclassifications produced by BERE on the distantly supervised DTI dataset. In particular, among the selected representative sentences of the extracted kinase targets of nintedanib, we observed that, for both sentences with PLK1 and AXL (Table 2), our model was confused with the referents of inhibitory effects, thereby causing misclassifications. Here, BERE mainly considers the referents of the relation in a sentence by embedding the local contextual features of the target entities into a concatenated sentence representation. However, it is inevitable that the relative positions of target entities may vary across different sentences. Therefore, the embedded local contextual features may be weakened by the interference of different sentences in the sentence aggregation layer. We speculate that an improved sentence aggregation strategy or a better feature representation of referent information will help overcome the information loss caused by the mutual interference among sentences.

Overall, through extensive tests on an existing single-sentence annotated DDI dataset, a proposed distantly supervised DTI dataset and a case study to identify potential drug–target interactions, we have demonstrated the promising performance of BERE in biomedical relation extraction. All these results suggest that BERE can not only serve as a powerful tool in biomedical relation extraction, but can also provide useful assistance in the discovery of potential relations such as DTIs.

Methods

Datasets. *The manually labelled, single-sentence annotated DDI'13 dataset.* We first tested BERE on the DDI'13 dataset[46], which is a BioRE corpus with drug names and DDIs annotated manually from 784 DrugBank texts and 233 MedLine abstracts. The DDIs were semantically labelled at the sentence level; that is, each sentence was marked with an individual label from set {NA, ADVICE, EFFECT, MECHANISM, INT}, which describes different types of interactions between drugs. The DDI'13 dataset has been widely used as a benchmark dataset for the BioRE task[36–39]. To further improve the data quality, this dataset had been preprocessed in different ways in previous works[15,36]. For a fair comparison, here we directly used the same dataset as in ref. [15], which was preprocessed by negative sentence filtering to adjust the proportion of negative samples. In this DDI'13 dataset, ~77% of the data were randomly selected for training and the remaining were used for testing. To fine-tune the learning rate of each model in the training process, we further held out 10% of training data as the validation set to determine the optimal values of hyperparameters. Definitions of individual labels and the basic statistics of the DDI'13 dataset are provided in Extended Data Fig. 3a.

The newly constructed, distantly supervised DTI dataset. Existing BioRE datasets, such as DDI'13[46], CDR[47] and CPR[48], are all semantically annotated, a process that requires tremendous levels of time and effort to be expended in human curation, and they are thus often limited in size. To address this issue and further enhance the learning capacity of our model, we constructed a much larger DTI dataset that automatically annotates drug and target names by NER and the relations between drugs and targets at a bag of sentences level by the distant supervision technique[16,17]. This newly constructed DTI dataset was divided into four sets, for training, validation, test and prediction purposes, respectively. Among these, the first three are labelled sets, which were annotated by automatically aligning drug–target pairs in sentences from nearly 20 million PubMed abstracts against the DTI facts in DrugBank[1]. The last set is an unlabelled set, in which the positions of drugs and targets in sentences were simply located from ~2.2 million PMC articles (except the abstracts) by NER. We evaluated the performance of BERE and other baseline methods mainly on the labelled sets, and applied a well-trained BERE model to the unlabelled set to predict potential new DTIs. In the labelled sets, the relation between each drug–target pair was annotated from the set {NA, substrate, inhibitor, agonist/antagonist, unknown, other} (the meanings of individual labels will be explained in the following and in Extended Data Fig. 3b). The construction of this dataset consisted of four steps: (1) text preprocessing, (2) named entity recognition, (3) distant annotation and (4) postprocessing. We first performed sentence segmentation and word tokenization on the texts using spaCy[49]. A dictionary-based NER scheme was then used to match sentences to the names of drugs and targets. The name dictionary was collected from DrugBank, with ambiguous names (for example, common words) removed to improve recognition accuracy. Next, each bag of sentences that co-mentioned a certain drug–target pair was annotated with a DTI fact in DrugBank. For any drug–target pair, if their relation was absent in DrugBank, it was marked *NA*. If a drug behaves as a substrate, inhibitor or agonist/antagonist of its target partner, the corresponding drug–target pair was marked with substrate, inhibitor or agonist/antagonist, respectively. If there exists a certain relation between a drug–target pair, but the action mechanism is unknown according to DrugBank, the corresponding pair was marked with unknown. Other relations with fewer occurrences (including inducer, binder, potentiator, ligand and so on, accounting for 10% of the total samples) were labelled with other. Finally, we removed those sentences that were too long (more than 64 words) or repeated the same DTI too many times (more than 64 times) to further improve the quality of the dataset and control the number of sentences in a bag. More details about this DTI dataset constructed by the distant supervision approach are provided in Extended Data Fig. 3b.

Input representation. The inputs to BERE are the vector representations of words in sentences. In particular, the ith word in an input sentence is represented by a d-dimensional vector $\mathbf{e}_i$, which concatenates a word embedding representing its semantic meaning and a POS embedding encoding the corresponding POS (for example, noun, verb or adjective). Here, we used the word embeddings from the published materials[50], which were pretrained from the PubMed and PMC texts[51]

using the word2vec tools[52] and then fine-tuned by the relation extraction task described below. POS embeddings representing the grammatical meanings of words were initialized randomly and then updated during the training process of BERE.

Encoding short- and long-range dependencies between words. Learning short- and long-range dependencies between words in sentences has always been a key point in natural language processing tasks. BERE uses Bi-GRU and self-attention mechanisms to encode short- and long-range dependencies, respectively.

Self-attention. Self-attention has been commonly used to capture long-range dependencies between distant words in sentences or articles[12,53], by correlating the features at each position with the features at all positions in the input sequence. Here, we adopt a multi-layer (or multi-head) self-attention[35] with residual connection (that is, a shortcut connecting input and output directly)[54]. More specifically, given a sentence S consisting of M words, it can be represented as a sequence of vectors:

$$S = (\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_M) \tag{1}$$

where $\mathbf{e}_i$ is a d-dimensional vector representing the encoded features of the ith word. The output $\mathbf{g}_i \in \mathbb{R}^{d_k}$ (d_k is a hyperparameter that controls the output dimension) of a single-layer (or single-head) self-attention for the ith word is a weighted sum of all word vectors:

$$\mathbf{g}_i = \sum_{j=1}^{M} \frac{\alpha_{ij}}{\sqrt{d_k}} \mathbf{W}_v \mathbf{e}_j \tag{2}$$

where $\mathbf{W}_v \in \mathbb{R}^{d_k \times d}$ stands for the learned weight matrix, $\frac{1}{\sqrt{d_k}}$ is a scaling factor that controls the magnitude of the dot product and α_{ij} is a scalar value representing the attention weight between $\mathbf{e}_i$ and $\mathbf{e}_j$, which is calculated by

$$\alpha_{ij} = \frac{\exp(v_{ij})}{\sum_{j=1}^{M} \exp(v_{ij})} \tag{3}$$

where

$$v_{ij} = \mathbf{e}_j^T \mathbf{W}_k^T \mathbf{W}_q \mathbf{e}_i \tag{4}$$

and $\mathbf{W}_k \in \mathbb{R}^{d_k \times d}$ and $\mathbf{W}_q \in \mathbb{R}^{d_k \times d}$ are the corresponding learned weight matrices. Next, the multi-layer self-attention is computed by calculating the single-layer self-attention T times with different learned weight matrices. These independent attention outputs are then concatenated and projected through another learned weight matrix $\mathbf{W}_g \in \mathbb{R}^{d \times Td_k}$ once again to obtain the final output $\mathbf{g}'_i \in \mathbb{R}^d$:

$$\mathbf{g}'_i = \mathbf{W}_g \left[\mathbf{g}_i^1; \mathbf{g}_i^2; \ldots; \mathbf{g}_i^T\right] \tag{5}$$

where $\mathbf{g}_i^t$ stands for the tth layer of the self-attention output of the ith word. Finally, the residual connection is computed between input and output:

$$\mathbf{e}'_i = \gamma \mathbf{g}'_i + \mathbf{e}_i \tag{6}$$

where $\mathbf{e}'_i$ is the updated representation of the ith word and γ is a learned parameter that controls the contribution of the self-attention output in equation (6).

Bi-GRU. To organize words into nested phrases, the dependencies between adjacent words need to be extracted. The Bi-GRU[55], which has been widely used to process the ordered sequences from both backward and forward states, is suitable for capturing such dependencies. Given a sequence of vectors $S' = (\mathbf{e}'_1, \mathbf{e}'_2, \ldots, \mathbf{e}'_M)$ that represents the updated word features, the output of Bi-GRU is denoted by

$$H = (\mathbf{h}_1, \mathbf{h}_2, \ldots, \mathbf{h}_M) \tag{7}$$

where $\mathbf{h}_i$ is a $2u$-dimensional vector, which is obtained by concatenating the u-dimensional GRU states of the ith word from both directions.

Encoding syntactic structures of sentences. Gumbel Tree-LSTM[33] is a kind of latent tree learning method for implicitly learning the syntactic features of sentences, such as constituent structures, which organize words into nested phrases. Here, we adopt a variant, Gumbel Tree-GRU, to learn constituency parsing by finding the best composition scheme among all feasible solutions for words in a sentence through a greedy-based strategy. More specifically, in each step, the following two operations are conducted (Fig. 1b):

- Every two adjacent vectors in a sequence are composed into a single vector as a candidate encoded by a shared Tree-GRU cell.
- A scoring function is used to assess each candidate. The best candidate is selected and other word vectors are copied from the previous step directly.

Therefore, for a sentence with M words, after $M-1$ steps, only one vector remains, denoted by $\mathbf{h}^{\text{root}}$, which represents the sentence features from both syntactic and semantic levels. After that, the contextual features of the two entities from the previous layer are also incorporated into sentence encoding to strengthen the semantic representations of the words close to the target entities:

$$\mathbf{h}^{\text{concat}} = [\mathbf{h}^{\text{Entity1}}; \mathbf{h}^{\text{root}}; \mathbf{h}^{\text{Entity2}}] \tag{8}$$

where $\mathbf{h}^{\text{concat}} \in \mathbb{R}^{6u}$ is the concatenated feature representation, $\mathbf{h}^{\text{root}} \in \mathbb{R}^{2u}$ represents the sentence embedding and $\mathbf{h}^{\text{Entity1}} \in \mathbb{R}^{2u}$ and $\mathbf{h}^{\text{Entity2}} \in \mathbb{R}^{2u}$ are the feature embeddings of Entity1 and Entity2 (that is, the corresponding output vectors in equation (7)), respectively.

Tree-GRU. Tree-structured GRU (Tree-GRU), a type of RvNN[56,57], is commonly applied to propagate information through constituency-based parse trees. In this work, we use a weight shared Tree-GRU cell[58] to compose every two feature vectors of adjacent words or phrases into a larger one. In particular, for each none-leaf node in the tree, it receives inputs from both its children. Suppose $\mathbf{h}_j$ and $\mathbf{h}_{j+1}$ are its left and right children, respectively, then the updating formulae for their parent $\mathbf{h}$ are

$$\mathbf{i} = \sigma(\mathbf{W}_i \mathbf{h}_j + \mathbf{U}_i \mathbf{h}_{j+1} + \mathbf{b}_i) \tag{9}$$

$$\mathbf{f} = \sigma(\mathbf{W}_f \mathbf{h}_j + \mathbf{U}_f \mathbf{h}_{j+1} + \mathbf{b}_f) \tag{10}$$

$$\mathbf{r} = \sigma(\mathbf{W}_r \mathbf{h}_j + \mathbf{U}_r \mathbf{h}_{j+1} + \mathbf{b}_r) \tag{11}$$

$$\tilde{\mathbf{h}} = \tanh(\mathbf{W}_h(\mathbf{r} \odot \mathbf{h}_j) + \mathbf{U}_h(\mathbf{r} \odot \mathbf{h}_{j+1}) + \mathbf{b}_h) \tag{12}$$

$$\mathbf{h} = \mathbf{i} \odot \tilde{\mathbf{h}} + f \odot (\mathbf{h}_j + \mathbf{h}_{j+1}) \tag{13}$$

where $\mathbf{W}_i \in \mathbb{R}^{2u \times 2u}$, $\mathbf{W}_f \in \mathbb{R}^{2u \times 2u}$, $\mathbf{W}_r \in \mathbb{R}^{2u \times 2u}$, $\mathbf{W}_h \in \mathbb{R}^{2u \times 2u}$, $\mathbf{U}_i \in \mathbb{R}^{2u \times 2u}$, $\mathbf{U}_f \in \mathbb{R}^{2u \times 2u}$, $\mathbf{U}_r \in \mathbb{R}^{2u \times 2u}$ and $\mathbf{U}_h \in \mathbb{R}^{2u \times 2u}$ are the learned weight matrices, $\mathbf{b}_i \in \mathbb{R}^{2u}$, $\mathbf{b}_f \in \mathbb{R}^{2u}$, $\mathbf{b}_r \in \mathbb{R}^{2u}$ and $\mathbf{b}_h \in \mathbb{R}^{2u}$ are the learned bias vectors, $\odot$ indicates the element-wise product and $\sigma(\cdot)$ is the sigmoid function.

Scoring function. A scoring function is used to assess all proposed composition candidates of adjacent feature vectors in a sentence and select the best one. In step t, the candidates computed by equations (9) to (13) can be represented as ($\mathbf{q}_1$, $\mathbf{q}_2$, …, $\mathbf{q}_{M-t}$), where $\mathbf{q}_i \in \mathbb{R}^{2u}$ is the feature vector of the ith candidate. We then calculate the unnormalized score π_i for the ith candidate by

$$\pi_i = \mathbf{W}_\pi \tanh(\mathbf{V}_\pi \mathbf{q}_i) \tag{14}$$

where $\mathbf{W}_\pi \in \mathbb{R}^{1 \times d_q}$ and $\mathbf{V}_\pi \in \mathbb{R}^{d_q \times 2u}$ are the learned weight matrices and d_q is a hyperparameter. If we simply sample the best candidate according to π, the computational graph would not be differentiable. Thus, the model could not be trained using the standard backpropagation algorithm. The Gumbel–Softmax estimator[59] aims to solve this problem by adding a sampled Gumbel noise g to the logarithm of π, which transfers the non-differentiable sampling operation from π to g. More specifically, given the unnormalized probabilities π_1, …, π_{M-t}, Gumbel–Softmax generates a set of normalized scores by

$$y_i = \frac{\exp((\log(\pi_i) + g_i)/\tau)}{\sum_{j=1}^{k} \exp((\log(\pi_j) + g_j)/\tau)}, \text{ for } i = 1, \ldots, M - t \tag{15}$$

where $g_i \sim \text{Gumbel}(0,1)$ and can be computed by $g_i = -\log(-\log(u_i))$ according to $u_i \sim \text{Uniform}(0,1)$, and τ is the temperature parameter. If the temperature approaches zero, a sample from the Gumbel–Softmax distribution will resemble the one-hot vector. In practice, we set τ as a learnable parameter with initial value 1. After that, we select the candidate with the highest y value and make the backpropagation differentiable at the same time (now sampling of g is no longer in the computational graph). In the validation and test phases, we directly select the candidate with the highest π, instead of using equation (15).

Feature aggregation over a bag of sentences. To reduce the influence of irrelevant sentences and make the model focus on the important evidence in predicting a relation, BERE uses an attention-based sentence aggregation strategy[23] to score and then aggregate the features over a bag of sentences describing a certain entity pair (also see Fig. 1). In detail, a bag of N sentences, denoted by G, can be represented as

$$G = \left(\mathbf{h}_1^{\text{concat}}, \mathbf{h}_2^{\text{concat}}, \ldots, \mathbf{h}_N^{\text{concat}}\right) \tag{16}$$

where $\mathbf{h}_i^{\text{concat}}$ is a $6u$-dimensional vector representing the extracted features of the ith sentence. The corresponding weight β_i for the ith sentence is calculated as

$$\beta_i = \frac{\exp(\mathbf{W}_s \mathbf{h}_i^{\text{concat}})}{\sum_{j=1}^{N} \exp(\mathbf{W}_s \mathbf{h}_j^{\text{concat}})} \tag{17}$$

where $\mathbf{W}_s \in \mathbb{R}^{1\times 6u}$ is the learned weight matrix. The final aggregated features of the bag of sentences are calculated by the weighted sum of all $\mathbf{h}_i^{\text{concat}}$, $i = 1, \cdots, N$:

$$\mathbf{B} = \sum_{i=1}^{N} \beta_i \mathbf{h}_i^{\text{concat}} \tag{18}$$

Classification and optimization. Finally, given the aggregated features **B** for a bag of sentences, we calculate the probability of a possible relation label r between entity pair (*Entity*1, *Entity*2) using a softmax classifier:

$$p(r|\mathbf{B}) = \text{softmax}(\mathbf{W}_2(\Phi_{\text{relu}}(\mathbf{W}_1\mathbf{B} + \mathbf{b}_1)) + \mathbf{b}_2) \tag{19}$$

where $\mathbf{W}_1 \in \mathbb{R}^{d_s\times 6u}$ and $\mathbf{W}_2 \in \mathbb{R}^{1\times d_s}$ are the learned weight matrices, $\mathbf{b}_1 \in \mathbb{R}^{d_s}$ and $\mathbf{b}_2 \in \mathbb{R}^1$ are the learned bias vectors, Φ_{relu} is the rectified linear unit function[60] (that is, $\Phi_{\text{relu}}(x) = \max(x, 0)$) and d_s is a hyperparameter.

We define the objective function using the following cross-entropy loss:

$$J(\theta) = -\frac{1}{c}\sum_{i=1}^{c} t_i \log\left(\text{p}(\text{r}_\text{i}|\mathbf{B})\right) \tag{20}$$

where θ indicates all the learned parameters in BERE, $t_i \in \{0, 1\}$ is the ground truth of relation r_i, and c is the number of relation types. We minimize $J(\theta)$ using an Adam optimizer[61] with mini-batch training.

Kinase activity assays and reagents. Biochemical kinase activity assays were performed at ChemPartner (Shanghai) following the manufacturer's instructions (compound screening service, ChemPartner). The inhibitory activities of nintedanib against JAK2 and EGFR kinases were measured by mobility shift assays with ATP concentrations at K_m. In brief, nintedanib was threefold serially diluted in 100% dimethylsulfoxide to create 10-point titrations at a starting concentration of 10 μM (as shown in Fig. 3). All substrate/kinase mixtures were diluted to a 2.5× working concentration with a kinase buffer (50 mM HEPES pH 7.5, 0.0015% Brij-35). All reagents were mixed and incubated at 28 °C for 1 h, and then reactions were stopped by adding stop buffer (100 mM HEPES pH 7.5, 0.015% Brij-35, 0.2% coating reagent #3, 50 mM EDTA). The reactions were measured on Caliper (LabChip EZ Reader). IC_{50} values were calculated by XLFit excel add-in version 5.4.0.8[45].

Nintedanib was obtained from MedChemExpress (cat. no. HY-50904), the recombinant kinase protein EGFR was obtained from Eurofins (cat. no. 14-531), recombinant kinase protein JAK2 was obtained from Carna (cat. no. 08-045) and Eu-anti-P-4E-BP1 (Thr37/46) was obtained from Promega (cat. no. TRF0216-M). All other chemicals were obtained from Sigma-Aldrich.

Data availability

The DDI and DTI datasets used in this work can be found at https://github.com/haiya1994/BERE. The full dataset for discovering potential DTIs is available from the corresponding authors upon request.

Code availability

The source code of BERE can be downloaded from the GitHub repository at https://github.com/haiya1994/BERE or the Zenodo repository at https://doi.org/10.5281/zenodo.3757058. All other code may be obtained from the corresponding authors upon request.

Received: 15 December 2019; Accepted: 7 May 2020;
Published online: 8 June 2020

References

1. Wishart, D. S. et al. DrugBank: a comprehensive resource for in silico drug discovery and exploration. *Nucleic Acids Res.* **34**, D668–D672 (2006).
2. Mattingly, C. J., Colby, G. T., Forrest, J. N. & Boyer, J. L. The Comparative Toxicogenomics Database (CTD). *Environ. Health Perspect.* **111**, 793–795 (2003).
3. Kuhn, M., Letunic, I., Jensen, L. J. & Bork, P. The SIDER database of drugs and side effects. *Nucleic Acids Res.* **44**, D1075–D1079 (2015).
4. Oughtred, R. et al. BioGRID: a resource for studying biological interactions in yeast. *Cold Spring Harbor Protoc.* **2016**, pdb.top080754 (2016).
5. Wang, S. et al. Annotating gene sets by mining large literature collections with protein networks. In *Proceedings of the Pacific Symposium on Biocomputing* 601–613 (World Scientific, 2018).
6. Wang, S. et al. Deep functional synthesis: a machine learning approach to gene functional enrichment. Preprint at https://doi.org/10.1101/824086 (2019).
7. Magro, L., Moretti, U. & Leone, R. Epidemiology and characteristics of adverse drug reactions caused by drug–drug interactions. *Expert Opin. Drug Saf.* **11**, 83–94 (2012).
8. Yang, F., Xu, J. & Zeng, J. Drug–target interaction prediction by integrating chemical, genomic, functional and pharmacological data. In *Proceedings of the Pacific Symposium on Biocomputing 2014* 148–159 (World Scientific, 2014).
9. Luo, Y. et al. A network integration approach for drug–target interaction prediction and computational drug repositioning from heterogeneous information. *Nat. Commun.* **8**, 573 (2017).
10. Wan, F., Hong, L., Xiao, A., Jiang, T. & Zeng, J. NeoDTI: neural integration of neighbor information from a heterogeneous network for discovering new drug–target interactions. *Bioinformatics* **35**, 104–111 (2018).
11. Percha, B. & Altman, R. B. A global network of biomedical relationships derived from text. *Bioinformatics* **34**, 2614–2624 (2018).
12. Verga, P., Strubell E. & McCallum, A. Simultaneously self-attending to all mentions for full-abstract biological relation extraction. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* Vol. 1, 872–884 (ACL, 2018).
13. Zhang, Y. et al. A hybrid model based on neural networks for biomedical relation extraction. *J. Biomed. Inform.* **81**, 83–92 (2018).
14. Yu, K. et al. Automatic extraction of protein–protein interactions using grammatical relationship graph. *BMC Med. Inform. Decis. Mak.* **18**, 42 (2018).
15. Lim, S., Lee, K. & Kang, J. Drug drug interaction extraction from the literature using a recursive neural network. *PLoS ONE* **13**, e0190926 (2018).
16. Mintz, M., Bills, S., Snow, R. & Jurafsky, D. Distant supervision for relation extraction without labeled data. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP* Vol. 2, 1003–1011 (ACL, 2009).
17. Riedel, S., Yao, L. & McCallum, A. Modeling relations and their mentions without labeled text. In *Proceedings of the Joint European Conference on Machine Learning and Knowledge Discovery in Databases* 148–163 (Springer, 2010).
18. Dietterich, T. G., Lathrop, R. H. & Lozano-Pérez, T. Solving the multiple instance problem with axis-parallel rectangles. *Artif. Intell.* **89**, 31–71 (1997).
19. Jat, S., Khandelwal, S. & Talukdar, P. Improving distantly supervised relation extraction using word and entity based attention. In *Proceedings of the 6th Workshop on Automated Knowledge Base Construction* (2017).
20. Vashishth, S., Joshi, R., Prayaga, S. S., Bhattacharyya, C. & Talukdar, P. RESIDE: improving distantly-supervised neural relation extraction using side information. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing* 1257–1266 (ACL, 2018).
21. Zeng, D., Liu, K., Chen, Y. & Zhao, J. Distant supervision for relation extraction via piecewise convolutional neural networks. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing* 1753–1762 (ACL, 2015).
22. Quirk, C. & Poon, H. Distant supervision for relation extraction beyond the sentence boundary. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics* Vol. 1, 1171–1182 (ACL, 2017).
23. Lin, Y., Shen, S., Liu, Z., Luan, H. & Sun, M. Neural relation extraction with selective attention over instances. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics* Vol. 1, 2124–2133 (ACL, 2016).
24. Zhou, P. et al. Attention-based bidirectional long short-term memory networks for relation classification. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics* Vol. 2, 207–212 (ACL, 2016).
25. Sun, X. et al. Drug–drug interaction extraction via recurrent hybrid convolutional neural networks with an improved focal loss. *Entropy* **21**, 37 (2019).
26. Socher, R. et al. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing* 1631–1642 (ACL, 2013).
27. Iyyer, M., Boyd-Graber, J., Claudino, L., Socher, R. & DauméIII, H. A neural network for factoid question answering over paragraphs. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)* 633–644 (ACL, 2014).
28. Hashimoto, K., Miwa, M., Tsuruoka, Y. & Chikayama, T. Simple customization of recursive neural networks for semantic relation classification. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing* 1372–1376 (ACL, 2013).
29. Li, J., Luong, M. T., Jurafsky, D. & Hovy, E. When are tree structures necessary for deep learning of representations? In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing* 2304–2314 (ACL, 2015).
30. Bowman, S. R. et al. A fast unified model for parsing and sentence understanding. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics* Vol. 1, 1466–1477 (ACL, 2016).
31. Yogatama, D., Blunsom, P., Dyer, C., Grefenstette, E. & Ling, W. Learning to compose words into sentences with reinforcement learning. In *Proceedings of the 5th Interational Conference on Learning Representations* (2017).

32. Maillard, J., Clark, S. & Yogatama, D. Jointly learning sentence embeddings and syntax with unsupervised Tree-LSTMs. *Nat. Lang. Eng.* **25**, 433–449 (2019).
33. Choi, J., Yoo, K. M. & Lee, S.-g. Learning to compose task-specific tree structures. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence* 5094–5101 (AAAI, 2018).
34. Wang, X., Girshick, R., Gupta, A. & He, K. Non-local neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* 7794–7803 (IEEE, 2018).
35. Vaswani, A. et al. Attention is all you need. In *Proceedings of Advances in Neural Information Processing Systems* 5998–6008 (NIPS, 2017).
36. Zhao, Z., Yang, Z., Luo, L., Lin, H. & Wang, J. Drug drug interaction extraction from biomedical literature using syntax convolutional neural network. *Bioinformatics* **32**, 3444–3453 (2016).
37. Liu, S., Tang, B., Chen, Q. & Wang, X. Drug-drug interaction extraction via convolutional neural networks. *Comput. Math. Methods Med.* **2016**, 6918381 (2016).
38. Quan, C., Hua, L., Sun, X. & Bai, W. Multichannel convolutional neural network for biological relation extraction. *Biomed Res. Int.* **2016**, 1850404 (2016).
39. Sahu, S. K. & Anand, A. Drug–drug interaction extraction from biomedical texts using long short-term memory network. *J. Biomed. Inform.* **86**, 15–24 (2018).
40. Zhou, D., Miao, L. & He, Y. Position-aware deep multi-task learning for drug–drug interaction extraction. *Artif. Intell. Med.* **87**, 1–8 (2018).
41. Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I. & Salakhutdinov, R. Dropout: a simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.* **15**, 1929–1958 (2014).
42. Tolias, G., Sicre, R. & Jégou, H. Particular object retrieval with integral max-pooling of CNN activations. In *Proceedings of the 4th International Conference on Learning Representations* (2016).
43. Liu, C. Y. et al. The tyrosine kinase inhibitor nintedanib activates SHP-1 and induces apoptosis in triple-negative breast cancer cells. *Exp. Mol. Med.* **49**, e366 (2017).
44. Kato, M. et al. Gastrointestinal adverse effects of nintedanib and the associated risk factors in patients with idiopathic pulmonary fibrosis. *Sci. Rep.* **9**, 12062 (2019).
45. XLFit 5.4.0.8 (IDBS, 2014); https://www.idbs.com/excelcurvefitting/xlfit-product/
46. Herrero-Zazo, M., Segura-Bedmar, I., Martínez, P. & Declerck, T. The DDI corpus: an annotated corpus with pharmacological substances and drug–drug interactions. *J. Biomed. Inform.* **46**, 914–920 (2013).
47. Li, J. et al. BioCreative V CDR task corpus: a resource for chemical disease relation extraction. *Database* **2016**, baw068 (2016).
48. Krallinger, M. et al. Overview of the BioCreative VI chemical-protein interaction track. In *Proceedings of the Sixth BioCreative Challenge Evaluation Workshop* Vol. 1, 141–146 (2017).
49. Honnibal, M. & Montani, I. spaCy 2.0.18 (2018); https://spacy.io/
50. Pyysalo, S., Ginter, F., Moen, H., Salakoski, T. & Ananiadou, S. Word vectors (NLPLab, 2013); http://bio.nlplab.org/
51. Pyysalo, S., Ginter, F., Moen, H., Salakoski, T. & Ananiadou, S. Distributional semantics resources for biomedical text processing. In *Proceedings of the 5th International Symposium on Languages in Biology and Medicine* 39–44 (2013).
52. Mikolov, T., Chen, K., Corrado, G. & Dean, J. Efficient estimation of word representations in vector space. In *Proceedings of the 1st International Conference on Learning Representations* (2013).
53. Tan, Z., Wang, M., Xie, J., Chen, Y. & Shi, X. Deep semantic role labeling with self-attention. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence* 16725 (AAAI, 2018).
54. He, K., Zhang, X., Ren, S. & Sun, J. J. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* 770–778 (IEEE, 2016).
55. Cho, K., Van Merriënboer, B., Bahdanau, D. & Bengio, Y. On the properties of neural machine translation: Encoder-decoder approaches. In *Proceedings of SSST-8, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation* 103–111 (ACL, 2014).
56. Socher, R., Lin, C. C., Manning, C. & Ng, A. Y. Parsing natural scenes and natural language with recursive neural networks. In *Proceedings of the 28th International Conference on Machine Learning* (*ICML-11*) 129–136 (ACM, 2011).
57. Tai, K. S., Socher, R. & Manning, C. D. Improved semantic representations from tree-structured long short-term memory networks. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing* Vol. 1, 1556–1566 (ACL, 2015).
58. Kokkinos, F. & Potamianos, A. Structural attention neural networks for improved sentiment analysis. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics* Vol. 2, 586–591 (ACL, 2017).
59. Jang, E., Gu, S. & Poole, B. Categorical reparameterization with gumbel-softmax. In *Proceedings of the 5th International Conference on Learning Representations* (2017).
60. Nair, V. & Hinton, G. E. Rectified linear units improve restricted Boltzmann machines. In *Proceedings of the 27th International Conference on Machine Learning* (*ICML-10*) 807–814 (ACM, 2010).
61. Kingma, D. P. & Ba, J. Adam: A method for stochastic optimization. In *Proceedings of the 3rd International Conference on Learning Representations* (2015).

Acknowledgements

We thank Z. Liu, T. Yang and H. Hu for their helpful discussions about this work. This work was supported in part by the National Natural Science Foundation of China (grants 61872216, 81630103 and 31900862), the Turing AI Institute of Nanjing and the Zhongguancun Haihua Institute for Frontier Information Technology.

Author contributions

L.H., D.Z. and J.Z. conceived the concept. L.H. designed the methodology and performed experiments. L.H., J.L., S.L., T.J. and D.Z. analysed the results. H.Y. contributed to wet-lab experiments. L.H. and J.Z. wrote the paper. S.L., F.W., T.J. and D.Z. contributed to revision of the manuscript.

Competing interests

J.Z. is founder and CTO of Silexon AI Technology Co. Ltd and has an equity interest.

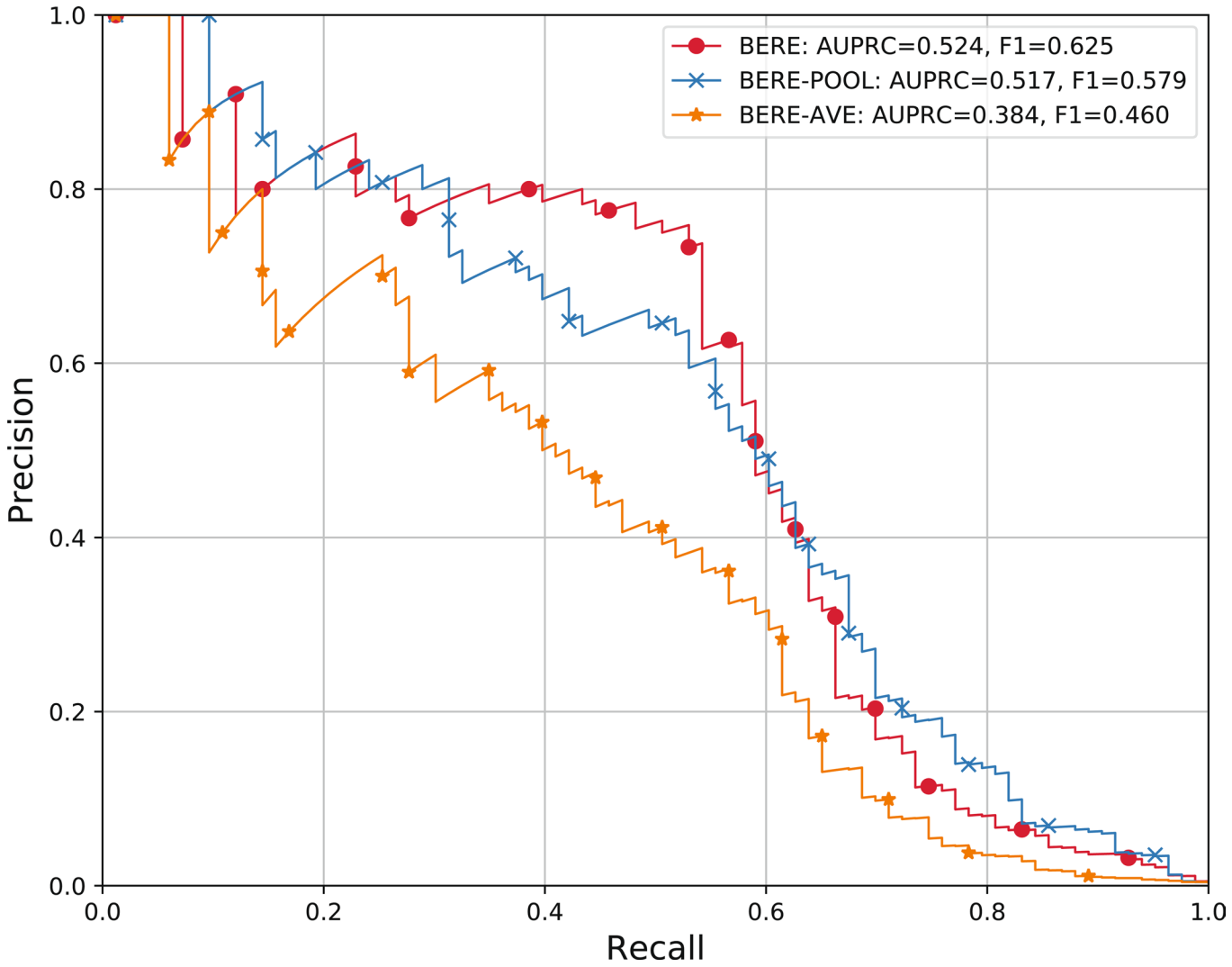
Additional information

Extended data is available for this paper at https://doi.org/10.1038/s42256-020-0189-y.

Correspondence and requests for materials should be addressed to D.Z. or J.Z.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Extended Data Fig. 1 | Comparison of the precision-recall curve between BERE and its alternatives with other sentence aggregation strategies on the distantly supervised DTI dataset. BERE+POOL and BERE+AVE adopt a max-pooling strategy and an average strategy to aggregate sentence representations, respectively. The legend on the top right contains area under precision-recall curve (AUPRC) and F_1-score for each method.

Hyper-parameters in BERE	The DDI'13 dataset	The distantly supervised DTI dataset
Word embedding dimension	200	200
POS embedding dimension	50	50
Hidden dimension u in Bi-GRU	250	250
T in self-attention	10	10
d_k in self-attention	25	25
d_q in scoring function	50	50
d_s in classifier	150	150
Learning rate λ	0.0007	0.0001
Dropout probability	0.5	0.5
Batch size	128	64
Maximum epoch	50	10

Extended Data Fig. 2 | The hyperparameter settings of BERE on different test datasets. The learning rates were determined using a grid search among {0.0001, 0.0002, …, 0.001}. Other hyper-parameters were set empirically.

(a) The DDI'13 dataset

Data split	Total	NA	ADVICE	EFFECT	MECHANISM	INT
Training	11556	8088	734	1434	1131	169
Validation	1285	899	80	158	129	19
Test	3020	2049	221	357	301	92

(b) The distantly supervised DTI dataset

Data split	Total	NA	substrate	inhibitor	agonist/ antagonist	unknown	other
Training	472k	464k	1710	2612	855	2534	604
Validation	4769	4686	12	20	11	37	3
Test	4817	4734	18	19	10	26	10
Unlabelled	666k	/	/	/	/	/	/

Extended Data Fig. 3 | The basic statistics of the datasets used in our tests. (a) The numbers of sentences annotated with five different types of DDI relations in the DDI'13 dataset. *NA* means no interaction. *ADVICE* means the recommended concomitant medication usage. *EFFECT* means that there exists a certain pharmacodynamic effect between two drugs. *MECHANISM* means that there exists a certain pharmacokinetic mechanism between two drugs. *INT* means that a DDI occurs without any additional information. (b) The numbers of bags of sentences annotated with six different types of DTI relations in the distantly supervised DTI dataset. *NA* means no interaction.*Substrate* means that the drug is what the target (that is, enzyme) acts upon. *Inhibitor* means that the drug binds to the target (that is, enzyme) and impede with the functioning of the target. *Agonist/Antagonist* means that the drug binds to the target (that is, receptor) and activates/blocks it to produce a biological response. *Unknown* means that there exists a certain relation between a drug–target pair, but the action mechanism is unknown in DrugBank. *Other* is a unified name of all the other types of interactions with fewer occurrences. The unlabelled set, which was mainly used for prediction, was collected from the PMC articles after excluding abstracts.