ELECTRA-DTA: A New Compound-protein Binding Affinity Prediction Model based on the Contextualized Sequence Encoding

junjie wang  
Nanjing Medical University  https://orcid.org/0000-0003-0732-051X

NaiFeng Wen  
Dalian Minzu University: Dalian Nationalities University  https://orcid.org/0000-0003-2725-8645

Chunyu Wang  
Harbin Institute of Technology  https://orcid.org/0000-0002-2965-9920

Lingling Zhao (✉️ zhaoll@hit.edu.cn)  
Harbin Institute of Technology

Liang Cheng  
Harbin Medical University  https://orcid.org/0000-0002-6665-6710

Research article

Keywords: deep learning, representative learning, drug-target affinity prediction, ELECTRA

Posted Date: October 5th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-934203/v1

License: ☕️ This work is licensed under a Creative Commons Attribution 4.0 International License. Read Full License
ELECTRA-DTA: A new compound-protein binding affinity prediction model based on the contextualized sequence encoding

Wang Junjie1, Wen NaiFeng2, Wang Chunyu3, Zhao Lingling3* and Cheng Liang4

*Correspondence: zhaoll@hit.edu.cn
3Faculty of Computing, Harbin Institute of Technology, Harbin, P. R. China
Full list of author information is available at the end of the article

Abstract

Motivation: Drug-target binding affinity (DTA) reflects the strength of the drug-target interaction; therefore, predicting the DTA can considerably benefit drug discovery by narrowing the search space and pruning drug-target (DT) pairs with low binding affinity scores. Representation learning using deep neural networks has achieved promising performance compared with traditional machine learning methods; hence, extensive research efforts have been made in learning the feature representation of proteins and compounds. However, such feature representation learning relies on a large-scale labelled dataset, which is not always available.

Results: We present an end-to-end deep learning framework, ELECTRA-DTA, to predict the binding affinity of drug-target pairs. This framework incorporates an unsupervised learning mechanism to train two ELECTRA-based contextual embedding models, one for protein amino acids and the other for compound SMILES string encoding. In addition, ELECTRA-DTA leverages a squeeze-and-excitation (SE) convolutional neural network block stacked over three fully connected layers to further capture the sequential and spatial features of the protein sequence and SMILES for the DTA regression task. Experimental evaluations show that ELECTRA-DTA outperforms various state-of-the-art DTA prediction models, especially with the challenging, interaction-sparse BindingDB dataset. In target selection and drug repurposing for COVID-19, ELECTRA-DTA also offers competitive performance, suggesting its potential in speeding drug discovery and generalizability for other compound- or protein-related computational tasks.

Keywords: deep learning; representative learning; drug-target affinity prediction; ELECTRA

Introduction

Drug discovery and development are laborious, time-consuming, expensive and challenging processes. One of the most important steps in developing a new drug or repurposing existing drugs is target identification and validation. A direct brute-force search is unrealistic because of the very large number of drug-like compounds and possible drug targets. Given current advances in computational methods and techniques, especially the application of machine learning in chemical and biological research fields, computer-aided methods may be a great opportunity to shorten the drug discovery process by significantly narrowing the search space.
Apart from the prediction of binary interaction relationships between compound-protein pairs, another crucial factor in candidate screening, the prediction of the binding affinity of compound-protein pairs, also called drug-target affinity (DTA), remains a challenge in drug discovery. Laboratory experiments conducted to measure the affinity value for a large-scale drug-target pool remain time consuming and expensive. Hence, computational methods to perform binding affinity prediction have received increasing attention in recent years, and much effort has been made to accurately quantify the strength of binding for compound-protein pairs based on machine learning or deep learning. Gradient boosting machines are used in quantitative structure-activity relationship studies for regression and classification problems. SimBoost [1] employs gradient boosting machines with novel feature engineering to extract new features from drugs, targets and drug-target pairs in training datasets, and then these features are used as inputs to models to predict the binding affinity for unknown pairs. Regularized least-square (RLS) is another efficient model with various applications. The KronRLS [2] model amends RLS with the Kronecker products of drug-drug and protein-protein interactions to speed up model training for DTA prediction and has achieved promising performance. The Kronecker product part of the KronRLS is a similarity-based method in which any similarity measure could be used.

Recently, inspired by successful applications in diverse research fields, deep learning approaches have also been intensively used in bioinformatics and cheminformatics, especially in drug discovery. The first deep learning-based DTA prediction model was DeepDTA [3], which uses simplified molecular input line entry system (SMILES), a one-dimensional representation of the drug compound chemical structure, as drug features, while the protein amino acid sequences are used to represent protein features. Furthermore, the drug SMILESs are labelled as encoded integer vectors and as protein sequences. The DeepDTA model uses a CNN with three 1D convolutional layers with pooling for drug embedding to learn latent features for each drug and an identical CNN for protein embedding. Then, each pair of drug-target feature vectors is concatenated and fed into fully connected layers for training and prediction. Another novel deep learning model for DTA is DeepAffinity [4], which represents drugs with SMILESs and proteins with structural property sequences. Because of the detailed structural information and higher resolution of sequences, DeepAffinity benefits DTA regression tasks. The drug SMILESs and protein structural sequences are both encoded into embedding representations by a recurrent neural network (RNN) auto-encoder model named seq2seq [5]. The seq2seq model maps raw sequences into vectors that are learned in an unsupervised fashion to capture dependencies in sequences of SMILESs or protein residues.

The aforementioned DTA models mainly focus on developing diverse neural network architectures to learn the hierarchical feature representations on given, known CP pairs with binding affinity values in a non-handcraft manner. These existing models usually take predefined compound molecular and protein descriptors as input features, especially SMILES strings and protein residue sequences. To be fed into deep networks such as CNNs or LSTMs, the raw sequence or SMILES strings are encoded by one-hot vectors, physicochemical property-aware encoding or static embeddings. Such mechanisms deploy fixed representations, which makes
the model characterize each amino acid or atom independently but do not consider
the contextual information and accordingly do not highlight those that are critical
to the whole protein or molecular compound. Applicable experimental structures of
protein complexes are abundant, but the number of available labelled drug-target
pairs remains limited. To address the issue of insufficient labelled data, unsuper-
vised encoding methods have recently been considered for proteins or compounds.
The idea behind this approach is to use protein sequences and compound SMILESs
or fingerprints as a codified language for human experts with limited words and
grammars. Similar to a linguist who extracts hidden knowledge from sentences of a
natural language, the structure and function of chemical compounds and protein
sequences can be processed deliberately to build novel solutions, such as DTA pre-
dictions, based on the level of understanding. State-of-the-art representation models
for NLP, such as BERT [6], have greatly improved downstream NLP tasks due to
their advantages. The recent transformer-based ELECTRA [7] model uses train-
ing data efficiently and shows better performance than BERT. ELECTRA employs
a generative-discriminative model to make better use of training data; a generator
model replaces some tokens in the corpus, and a co-trained discriminative model de-
tects these replacements. Because the corpus is well used in this fashion, ELECTRA
is more efficient than BERT with comparable model sizes, especially small models.
This paper proposes ELECTRA-based encoders for molecular SMILESs and pro-
tein sequences separately and accordingly yields a novel DTA prediction framework
incorporating an ELECTRA encoding layer, a CNN network and a regression block.
The proposed framework enhances the feature representation capability as follows.
First, the ELECTRA encoding layers used to represent each molecule can take ad-
vantage of existing chemical structure knowledge by pretraining on the PubChem
database to extract useful chemical information, and a similar prompt occurs for the
amino acid representation. Finally, the ELECTRA encoding provides the context-
related representation for protein sequences and compound SMILESs, which can
characterize the diversity of each atom or amino acid in a variety of sequences or
strings.

In summary, our method consists of two major building blocks: one for feature
representation and another for regression prediction. The first block, using two pre-
trained ELECTRA models, extracts local context information from drug SMILES
strings and raw protein sequences separately. Then, the learned representations for
both drugs and targets are passed into the second block, which uses a fully connected
neural network to predict the binding affinity as a regression task. Our framework
needs neither expert knowledge nor the 3D structure of the targets, so it is more
convenient than existing frameworks. Additionally, the proposed framework takes
advantage of the local chemical context information of atoms in drugs or amino
acids in proteins, which differentiates ELECTRA-DTA from existing deep learning
models. The main contributions of this paper are as follows:

1. We adopt ELECTRA, a state-of-the-art NLP model, to extract feature rep-
resentations from raw sequence data, taking advantage of the local context
information.
2. We leverage a squeeze-and-excitation convolution neural network block
stacked over three fully connected layers to capture the sequential and spatial
features of the protein sequence and SMILES for the DTA regression task.
3 We applied the ELECTRA-DTA model for target selection and drug repurposing for COVID-19 and again obtained competitive performance.

Methods
In this section, we first summarize the whole model. Then, we introduce the input representation for compounds and proteins. After that, we show the training for ELECTRA and the representation tensor of compounds and proteins. Finally, we provide the details of our training and prediction model.

Overview of the ELECTRA-DTA model
Figure 1 shows the overview of ELECTRA-DTA. It takes the SMILES strings of compounds and the amino acid sequences of proteins as the input. By incorporating a pre-trained ELECTRA model, ELECTRA-DTA encodes these sequences into feature tensors as internal representations. Then, the model exploits a CNN network to learn from known drug-protein pairs in a supervised manner. Finally, it outputs the binding affinity for new pairs. The design of ELECTRA-DTA includes three main steps:

1 The training of two ELECTRA models to encode all amino acids for protein sequences and characters in SMILESs for compounds separately;
2 The encoding of the whole sequence of compounds and proteins as feature tensors;
3 Model training and predicting the binding affinity with the proposed deep neural network.

Pretrained ELECTRA models for encoding protein sequences and compound SMILES strings
To encode the SMILESs/amino acid sequences, existing deep learning approaches such as DeepDTA and DeepCDA use label/one-hot encoding to represent each symbol in the SMILES/amino acid sequence. However, label/one-hot encoding often neglects the context of the symbol and thus cannot reveal the functionality of the symbol within the context.

Construction of the compound SMILES corpora
We need to train the ELECTRA model with an appropriate corpus for a specific task in advance to obtain an encoding layer for input sequences. Compound SMILES are linguistic constructs with a simple vocabulary (only atoms and chemical bond symbols) and a few grammar rules. Similar to natural language processing, compound SMILES strings are analogous to sentences, where each atom and bond symbol is a word (token). A corpus can then be naturally composed by collecting numerous compounds. For example, the SMILES string "CC(=)OC1≡C" can be listed as a sentence composed of the tokens 'C','C','≡','1',='C'. The corpus we constructed contains
all canonical SMILES extracted from PubChem [8, 9, 10] with a total number of 1114424.

To ensure the consistency of all SMILES from different sources, we use Open Babel [11] v3.1.0 to convert all SMILES strings in the corpora into a canonical format, which is also applied to SMILES in all datasets. Detailed information on the SMILES corpora we constructed is listed in Table 1.

Table 1 Statistics of the compound SMILES corpora

| No. of corpus | Average length of the corpus | Minimum length of the corpus | No. of vocabulary |
|---------------|------------------------------|-----------------------------|------------------|
| 1114424       | 47                           | 3                           | 72               |

Construction of the protein FASTA sequence corpora Similar to the compound SMILES, a 1-mer schema is exploited to extract words from the protein FASTA sequences and create a protein sequence corpora; in this way, each residue is a single word. We utilized the Swiss-Prot dataset from the UniProt knowledgebase [12] (UniProtKB), which included 563,552 manually annotated and reviewed protein sequences, to collect the proteins. UniProtKB/Swiss-Prot is a high-quality annotated and non-redundant protein sequence database. The detailed information on the protein sequence corpora we constructed is listed in Table 2.

Table 2 Statistics of protein sequence corpora

| No. of corpus | Average length of the corpus | Minimum length of the corpus | No. of vocabulary |
|---------------|------------------------------|-----------------------------|------------------|
| 1868198       | 382                          | 2                           | 30               |

Pretraining the ELECTRA models Our method exploits two separate ELECTRA models to separately encode the SMILESs and amino acid sequences into fixed dimensional vectors. In an effort to avoid confusion and awkward phrasing, we describe the procedure of training the ELECTRA model for the SMILESs. The procedure for the pre-trained ELECTRA model for the amino acids sequences is similar.

ELECTRA exploits two Transformer encoders as the base structure: one acts as a generator network, and the other acts as a discriminator network. The generator is typically a small, masked language model that produces an output distribution over the tokens. The tokens from a SMILES string are first masked, and some are replaced with a mask symbol [MASK] with a constant probability. Then, the masked tokens are fed into the generator for joint pre-training with the discriminator. The generator network first learns from the masked tokens and then fills the missing tokens with predicted values, but the predicted value may not be the same as the original value. However, the jointly trained discriminator network learns to resolve whether each token is the same as the original one. It uses another Transformer encoder to extract the contextual information as embedding representations, which are used to determine the probability of replacing the tokens, as shown in figure 2. After pre-training with SMILES strings, we obtain an ELECTRA-M model as an encoder that can encode each SMILES into a feature vector in downstream tasks. Similarly, we obtain an ELECTRA-P model pre-trained with protein sequences that also encode each protein sequence into a feature vector.
Input representation
Protein sequence and compound SMILES are fed into the entire framework as input; because the SMILES strings and protein sequences are of different lengths, we truncate them to fixed lengths for effective representation. As in [3], we also choose fixed lengths of 100 for SMILES and 1000 for protein sequences for the benchmark of datasets. We chose these maximum lengths based on the distributions of the datasets so that the maximum lengths cover at least 80% of the proteins and compounds in the datasets. Longer sequences are truncated to these lengths, while shorter sequences are padded with zeros to the fixed lengths.

Compound SMILES and protein sequence embedding
With the trained ELECTRA-M and ELECTRA-P models, the input compound SMILES strings and protein sequences are embedded into tensors separately. For individual compounds, the sequence of tokens from their SMILES strings, which represents atoms or structure indicators, is fed into the trained ELECTRA-M model to yield a compound encoding. Specifically, each token, which is one character, is converted to a vector of length $W_c$ by ELECTRA-M, and then a sequence of $N_c$ tokens is converted to a sequence of $N_c$ vectors that are finally concatenated into a $W_c \times N_c$ tensor as the compound representation. In the same way, one protein residue token is encoded into a vector of length $W_p$ by ELECTRA-P, and a protein sequence of length $N_p$ vectors that are concatenated into a $W_p \times N_p$ tensor, as shown in figure 3.

ELECTRA-DTA framework
In this work, the ELECTRA-DTA model incorporates two identical feature extractor networks and one regression network, as shown in figure 4. Each feature extractor network is a typical convolutional neural network equipped with two layers of stacked squeeze-and-excitation (SE) blocks in addition to one global max pooling filter layer. The feature extractor network takes the compound or protein tensor as input to learn latent features during the supervised training process and then produces feature vectors as a representation. Two feature vectors are then concatenated into a single vector, which is fed into the regression network for the prediction.
Inside the feature extractor network, the main blocks are two SE block layers. These are recently proposed CNN units that can be stacked for extremely effective generalization across different datasets, especially image processing. The SE block improves the joint encoding of image spatial and channel information by removing the spatial dependency with global average pooling to learn a channel-specific descriptor. This block is capable of feature recalibration by using global information to selectively emphasize informative features over the others. The SE block of the ELECTRA-DTA model is shown in Figure 5. We use one-dimensional convolution to project the input compound or protein tensor $X \in \mathbb{R}^{T \times 256}$ into the feature maps $U \in \mathbb{R}^{T \times C}$ to perform feature recalibration. This one-dimensional convolution is in the feature direction. Therefore, these feature maps $U$ are first squeezed and produce a channel (or feature) descriptor by global average pooling in the feature direction. The interaction information is accumulated in this descriptor. The squeezing operation is followed by an excitation operation with a fully connected layer and ReLU activation, which produces modulation weights from the features by a simple self-gating mechanism. The output of the SE block is in the form of the feature maps $U$ scaled by these modulation weights. The stacked SE block further enlarges this interaction information between features.

The regression network contains two tandem parts, as shown in Figure 4. The first tandem part consists of a stack of fully connected layers with a simple gating mechanism. The second tandem part has four fully connected layers of different sizes. The simple gating mechanism in the first part is adopted from [13] to better regulate information flow in consecutive training epochs. Each layer consists of two gates called a Carry and a Transform that separate information flow to two streams—one modified, the other untouched—to the next iteration. Finally, all information flows through four fully connected layers and then produces the predicted affinity score.

**Model validation**

**Datasets**

This study evaluated ELECTRA-DTA using three common benchmark datasets: the KIBA dataset [14], the Davis dataset [15], and the BindingDB dataset [16]. The Davis dataset consists of 442 proteins and 68 compounds forming 30,056 DT pairs, while the KIBA dataset contains 229 proteins and 2111 compounds forming 118,254 DT pairs. BindingDB is a web-accessible public dataset, and the filtered version is exploited in this study to maintain consistency with a previous study [4]. Table 3 provides the statistics of these datasets.

However, the versions of these datasets used here share one crucial problem: the same DT sequence has different or duplicate affinity values. We use the Davis dataset as an example to illustrate this phenomenon. The ABL1 proto-oncogene encodes a cytoplasmic and nuclear protein tyrosine kinase that has been implicated in the processes of cell differentiation, cell division, cell adhesion, and stress response [17]. The ABL1(Y253F) mutant is a Gleevec-resistant form
Table 3 The detailed statistics of the datasets, containing the number of proteins, compounds, interactions and average number of interactions per protein and per compound.

|          | No. Proteins | No. Compounds | No. Interactions | No. Interactions per Protein | No. Interactions per Compound | No. Non-inactive Interactions |
|----------|--------------|---------------|-----------------|-----------------------------|--------------------------------|-------------------------------|
| Davis    | 442          | 68            | 30056           | 68                          | 442                            | 27554                         |
| KIBA     | 229          | 2111          | 118254          | 516                         | 56                             | 93426                         |
| BindingDB| 1620         | 87461         | 144525          | 89                          | 2                              | 96917                         |

Duplicated samples may harm the effectiveness of the deep learning DTA model, while sample inputs (DT pairs) with different labels (binding affinity values) may be detrimental to training the deep learning DTA model. We thus removed these samples from the original datasets. Table [?] describes the details of the refined datasets.

Table 4 The detailed statistics of the refined datasets, containing the number of proteins, compounds, interactions and average number of interactions per protein and per compound.

|          | No. Proteins | No. Compounds | No. Interactions | No. Interactions per Protein | No. Interactions per Compound | No. Non-inactive Interactions |
|----------|--------------|---------------|-----------------|-----------------------------|--------------------------------|-------------------------------|
| Davis    | 361          | 68            | 24548           | 68                          | 61                             | 22899                         |
| KIBA     | 229          | 2052          | 117184          | 511.720                     | 57.107                         | 92641                         |
| BindingDB| 1615         | 129109        | 144525          | 79.944                      | 1.541                          | 87622                         |

It can be observed that these datasets differ significantly in terms of the number of interactions per protein or compound. For the Davis and KIBA datasets, the average number of interactions per protein and compound is higher than 50, whereas BindingDB contains much fewer interactions for both compounds and proteins, suggesting that the connections between compounds and proteins are sparse, which would make the training of the prediction models challenging. In addition, all of these datasets have skewed distributions.

Evaluation metrics

The performance of all models was measured by the concordance index [19] (CI), the mean squared error (MSE), Pearson correlation coefficient, the $r^2_m$ index [20, 21]

of ABL1. ABL1 and ABL1(Y253F) share same amino acid sequence, and their binding affinity values for the drug fedratinib (Pubchem ID: 16722836, SMILES: CC1=CN=C(N=C1NC2=CC(=CC=C2)S(=O)(=O)NC(C)(C)C)NC3=CC(=C(C3)OCCN4CCCC4) are 6.7447 and 7.2518, respectively. The MET(Y1235D) recombinant human protein (956–end) was expressed in insect cells. MET receptor tyrosine kinase is a proto-oncogene that encodes a transmembrane growth factor receptor that is a heterodimer of two disulfide-linked chains of 50 kDa (alpha) and 145 kDa (beta) [18]. They also share the same amino acid sequence, but their binding affinity values for the drug Bms-345541 (PubChem ID: 9926054, SMILES: CC1=CC2=C(C=Cl)N=C(C3=NC=C(N23)C)NCCN.Cl) are both 5.0.
and the area under precision recall (AUPR) curve score. CI indicates the ranking performance of the models and can be calculated by equations (1).

\[
CI = \frac{1}{Z} \sum_{\delta x > \delta y} h (b_x - b_y), \quad \text{where } h(m) = \begin{cases} 
1, & \text{if } m > 0 \\
0.5, & \text{if } m = 0 \\
0, & \text{if } m < 0 
\end{cases} \tag{1}
\]

The \( r_m^2 \) index defines the possibility of an acceptable model and is calculated by equation (2), where \( r_0^2 \) and \( r_m^2 \) are the squared correlation coefficients with and without intercepts, respectively.

\[
r_m^2 = r^2 \ast \left( 1 - \sqrt{r^2 - r_0^2} \right) \tag{2}
\]

Generally, a model with an \( r_m^2 \) index greater than 0.5 on a test set can be considered acceptable. The AUPR curve score is generally adopted for binary prediction; therein, we converted the regressions on these datasets into their binary forms by thresholding similarly to DeepDTA. The Pearson correlation (R) is a metric that measures the correlation between two continuous variables.

There are three variant ELECTRA models: ELECTRA small, with 14M parameters, ELECTRA base, with 110M parameters, and ELECTRA large, with 335M parameters. Due to limited computing resources, we only pretrain two ELECTRA small models, one for compounds and another for proteins. We use code from simpletransformers to train these two ELECTRA small models. Unlike with the NLP ELECTRA models, we use the atom tokenization method for the compound SMILESs and protein amino acid sequences. The pre-training task of the ELECTRA small models takes 6 days with a batch size of 256, a learning rate of 1e-5, and an epoch of 100 by using 8 NVIDIA GeForce RTX 2080 Ti GPUs. The feature extraction block comprises 2 convolutional layers, one with 256 and the other with 512 channels. All convolutional layers apply 3 convolution kernels and are activated through rectified linear unit (ReLU) activation function. The regression block has 4 FC layers with 1024, 1024, 512, and 1 node. Each of the first two FC layers is followed by a dropout layer with a rate of 0.4. These 4 FC layers also use ReLU as the activation function. The Adam optimizer was used to train the parameters of the feature extraction block and regression block with the default learning rate of 0.001 for each DTA dataset.

Results

Results of predictive performance of random splitting settings

Four baseline approaches were selected for comparison with the proposed ELECTRA-DTA, including two classical machine learning-based prediction models, KronRLS and SimBoost, and two state-of-the-art deep learning-based models, DeepDTA and DeepCDA [22], when using the original datasets. It should be noted that to avoid potential errors from the implementation and training of these models, the results of the baseline approaches reported in the following tables are taken directly from their original publications. As the authors for DeepCDA did not provide details...
on the hyperparameters of the network, we were unable to reproduce their methods. Therefore, we compared our method with another state-of-the-art method: AttentionDTA [23].

Given the output of ELECTRA layer \( \ell_k \), we used two different input settings for the feature extractor network:

1. ELECTRA-DTA-AVG: Average of all token vectors from 12 Transformer layers.
2. ELECTRA-DTA-LAST: Vector representation of the last Transformer layer.

To quantify the prediction performance in an unbiased and consistent manner, we carried out 5-fold cross-validation similar to that used for DeepDTA, DeepCDA and SimBoost; that is, we split the data arbitrarily into six equivalent parts in which one part is selected as the independent test set. Any one of the remaining parts is used as the validation set and the others as the training set. The purpose of this division is to determine the hyper-parameters via 5-fold cross validation. Tables 5, 6 and 7 report the results of all the evaluated models and their average CI, MSE, \( R \), \( r_m^2 \) and AUPR curve scores with the original and refined KIBA, Davis and BindingDB datasets, respectively.

### Table 5 Comparison of all baseline approaches and ELECTRA-DTA on the KIBA datasets

| Dataset          | Model     | CI    | MSE   | \( R \) | \( r_m^2 \) | AUPR  |
|------------------|-----------|-------|-------|---------|------------|-------|
| Original KIBA    | KronRLS   | 0.782 | 0.411 | -       | 0.342      | 0.635 |
|                  | SimBoost  | 0.836 | 0.222 | -       | 0.529      | 0.700 |
|                  | DeepDTA   | 0.863 | 0.194 | 0.848   | 0.173      | 0.728 |
|                  | WideDTA   | 0.875 | 0.179 | -       | -          | -     |
|                  | DeepCDA   | 0.899(0.002) | 0.176 | 0.855 | 0.682(0.008) | 0.812(0.005) |
|                  | ELECTRA-DTA-AVG | 0.886(0.002) | 0.172 | 0.873 | 0.723(0.016) | 0.786(0.003) |
|                  | ELECTRA-DTA-LAST | 0.897(0.003) | 0.152 | 0.896 | 0.766(0.085) | 0.798(0.063) |
| Refined KIBA     | DeepDTA   | 0.892(0.026) | 0.152 | 0.896 | 0.766(0.085) | 0.798(0.063) |
|                  | Attention-DTA | 0.880(0.001) | 0.158 | 0.883 | 0.742(0.015) | 0.795(0.003) |
|                  | ELECTRA-DTA-LAST | 0.884(0.001) | 0.189 | 0.860 | 0.710(0.152) | 0.790(0.103) |
|                  | ELECTRA-DTA-AVG | 0.894(0.001) | 0.148 | 0.890 | 0.769(0.018) | 0.802(0.010) |

### Table 6 Comparison of all baseline approaches and ELECTRA-DTA on the Davis datasets

| Dataset          | Model     | CI    | MSE   | \( R \) | \( r_m^2 \) | AUPR  |
|------------------|-----------|-------|-------|---------|------------|-------|
| Original Davis   | KronRLS   | 0.871 | 0.379 | -       | 0.407      | 0.661 |
|                  | SimBoost  | 0.872 | 0.282 | -       | 0.644      | 0.709 |
|                  | DeepDTA   | 0.876(0.004) | 0.261 | 0.846 | 0.643(0.017) | 0.710(0.010) |
|                  | WideDTA   | 0.886 | 0.262 | -       | -          | -     |
|                  | DeepCDA   | 0.891(0.003) | 0.246 | 0.857 | 0.649(0.009) | 0.739(0.006) |
|                  | ELECTRA-DTA-AVG | 0.897(0.003) | 0.238 | 0.844 | 0.671(0.032) | 0.706(0.010) |
|                  | ELECTRA-DTA-LAST | 0.896(0.003) | 0.233 | 0.846 | 0.663(0.022) | 0.709(0.016) |
| Refined Davis    | DeepDTA   | 0.882(0.016) | 0.191 | 0.843 | 0.690(0.035) | 0.695(0.03) |
|                  | Attention-DTA | 0.888(0.007) | 0.195 | 0.836 | 0.897(0.005) | 0.877(0.022) |
|                  | ELECTRA-DTA-LAST | 0.901(0.004) | 0.194 | 0.841 | 0.668(0.039) | 0.685(0.02) |
|                  | ELECTRA-DTA-AVG | 0.902(0.003) | 0.198 | 0.842 | 0.637(0.048) | 0.686(0.039) |
Table 7: Comparison of all baseline approaches and the ELECTRA-DTA on the BindingDB Dataset

| Dataset                     | Model                | CI     | MSE   | R     | $r^2_m$ | AUPR  |
|-----------------------------|----------------------|--------|-------|-------|---------|-------|
| Original BindingDB Dataset  | DeepDTA              | 0.812(0.002) | 0.832 | 0.824 | 0.623(0.02) | 0.443(0.01) |
|                             | DeepCDA              | 0.822(0.001) | 0.844 | 0.808 | 0.621(0.002) | 0.459(0.003) |
|                             | ELECTRA-DTA-AVG      | 0.832(0.004) | 0.852 | 0.693 | 0.045(0.012) | 0.807(0.002) |
|                             | ELECTRA-DTA-LAST     | 0.829(0.001) | 0.848 | 0.707 | 0.653(0.023) | 0.801(0.002) |
| refined BindingDB Dataset   | DeepDTA              | 0.826(0.001) | 0.703 | 0.845 | 0.669(0.004) | 0.795(0.003) |
|                             | AttentionDTA         | 0.804(0.003) | 0.844 | 0.811 | 0.619(0.009) | 0.764(0.004) |
|                             | ELECTRA-DTA-LAST     | 0.838(0.001) | 0.861 | 0.641 | 0.690(0.02)  | 0.812(0.002) |
|                             | ELECTRA-DTA-AVG      | 0.841(0.002) | 0.864 | 0.626 | 0.690(0.024) | 0.816(0.003) |

For the original KIBA dataset, the proposed ELECTRA-DTA-AVG and DeepCDA achieve the best performance metrics, with ELECTRA-DTA-AVG outperforming DeepCDA by 0.016, 0.025 and 0.063 for MSE, R, and the $r^2_m$ index, respectively, while DeepCDA achieved the best CI (0.889 vs 0.885) and AUPR curve score. Both models achieved much better prediction performance than DeepDTA, SimBoost and KronRLS. As seen from Table 5, the proposed ELECTRA-DTA-AVG achieved values of 0.894, 0.148, 0.890, 0.769, and 0.802 for CI, MSE, R, $r^2_m$ index and AUPR curve score, respectively, for the refined KIBA dataset. Except for the R metric, our proposed method outperformed DeepDTA and AttentionDTA. With respect to the MSE metric, the proposed ELECTRA-DTA-LAST performed the worst, with a value of 0.189, among all models.

For the original Davis dataset as in Table 6, ELECTRA-DTA-AVG outperformed the next-best model in terms of the CI and $r^2_m$ index by 0.006 and 0.035, respectively. Additionally, the MSE error was 0.238, which was lower than that of the state-of-the-art methods. It should also be noted that DeepCDA performed better than ELECTRA-DTA in terms of the AUPR curve score. For the refined Davis dataset, the proposed ELECTRA-DTA-AVG and ELECTRA-DTA-LAST models had the highest CIs (demonstrating improvements of 0.014 and 0.013, respectively). Our methods did not generate the best results for the other metrics, but the differences were small.

Table 7 summarizes the evaluation metrics of the various methods for the original and refined BindingDB datasets. ELECTRA-DTA-AVG outperformed the baseline approaches for all evaluation metrics. For the original dataset, ELECTRA-DTA-AVG obtained a 0.007 increase in the CI, while the MSE value of our method was 0.109 lower than that of DeepCDA. The most substantial improvement was observed for the AUPR curve score, which increased from 0.459 for DeepCDA to 0.829. For the refined BindingDB dataset, our two ELECTRA-based models exceeded DeepDTA and AttentionDTA for all evaluation metrics, with ELECTRA-DTA-AVG yielding the best performance. Among all evaluation metrics, MSE has the most noticeable difference, which decreasing from 0.844 for AttentionDTA to 0.626. The improvement in the CI was distinct, achieving the highest value (0.841) over the baselines. The $r^2_m$ index, R and the AUPR curve score improved substantially relative to those of AttentionDTA and DeepDTA. ELECTRA-DTA-LAST suffered from a higher MSE error than ELECTRA-DTA-AVG. This higher MSE indicates
that the use of the only last embedding layer of ELECTRA is insufficient for obtaining a more accurate representation for difficult datasets. It should be noted that ELECTRA-DTA achieved better results in the BindingDB dataset in all metrics than the baseline models, likely due to differences in the dataset distribution. According to an analysis of these datasets, the BindingDB dataset has sparser samples, thereby making it more difficult to train prediction models. This suggests that our model has more robustness and reliability.

Interestingly, we observed that the CIs for DeepDTA and ELECTRA-DTA increased significantly from the original to the refined datasets. The reason for this is that different proteins or drugs have identical representations in the original datasets, and consequently, in the deep learning-based methods, the network becomes confused as to how to learn an effective representation for these proteins or drugs.

According to the reported results in Tables 5, 6 and 7, ELECTRA-DTA and DeepCDA consistently performed the other DTA models in terms of all metrics, with ELECTRA-DTA-AVG being responsible for 10 of the 15 best scores for the refined datasets, indicating that our model has more reliable and accurate predictive performance than the other models. This experiment also indicated that the performance of ELECTRA-DTA-AVG is superior to that of ELECTRA-DTA-LAST, especially in terms of the CI, over all refined datasets. This shows that the output from each layer in ELECTRA is useful in solving the DTA problem.

Results of predictive performance of warm/cold splitting settings

One of the key difficulties in DTA prediction is generalization of the model and the discovery of the binding affinity for unseen drugs or targets. Therefore, in these experiments, three splitting schemes were used for all refined datasets:

1. Warm splitting: Every drug or target in the test set is encountered in the training set.
2. Cold-drug splitting: Every drug in the test set is absent in the training set.
3. Cold-target splitting: Every target in the test set is absent from the training set.
4. Blind splitting: the target and the drug are both absent in the training set.

The cold-drug, cold-target and blinding splits deliver realistic and more difficult appraisal schemes for the DTA problem. For these splitting schemes, the targets and proteins are both split at an 8/2 ratio for the training/testing set.

Figure 6 reports a seemingly contradictory phenomenon: all methods perform better in the cold-target setting with the Davis dataset but worse with the KIBA dataset than in the cold-drug setting. We see that the prediction performance of the deep learning-based methods is correlated to the data distribution of the dataset. Combined with the data in Table 4, it can be seen that when the number of drugs in the training set is large and the number of targets is small, the cold-target splitting scheme tends to yield a more challenging problem than the cold-drug splitting scheme; on the other hand, when there are many targets and few drugs, the cold-target predictions appear to be better than the cold-drug predictions. We hypothesize that the reason behind such contradictory results could be that the models become more stable with greater knowledge on entities (drugs or targets).
in the training sets. In addition, we also observe an interesting phenomenon in which the gap between the cold-drug and cold-target predictions is different over the KIBA and Davis datasets. The cold-drug and cold-target prediction CIs are similar in the KIBA dataset, even though it has 2052 drugs and only 229 targets. Along this line, one might conclude that the model needs many more types of drugs than targets to learn their chemical representations. As expected, the performance of the warm splitting scheme tends to be better than that of the cold-drug and cold-target splitting schemes.

Additionally, we realized that the blinding split schemes that had fewer drug-target pairs turned out to be the most challenging for the deep learning based models. We can observed that all the models performance decreased drastically while our methods usually exhibited a relatively stable performance on the three datasets. We reckon that the reason is due the proposed pretrained ELECTRA model can extract the abundant information.

**Case Study on Drug Repurposing for COVID-19**

We conducted a case study to showcase the use of our proposed ELECTRA-DTA for biomedical researchers for the repurposing of existing drugs. Here, we took key proteins from SARS-CoV-2 as our targets and selected potential drugs from among the 82 known antiviral drugs approved or in development for treating infections caused by HIV, hepatitis B virus (HBV), hepatitis C virus (HCV) and influenza [24]. SARS-CoV-2 is an enveloped, positive-sense, single-stranded RNA beta-coronavirus [25] with a similar structure to the severe acute respiratory syndrome coronavirus (SARS-CoV) and Middle East respiratory syndrome coronavirus (MERS-CoV).

Three categories of proteins are encoded by the SARS-CoV-2 genome, including non-structural proteins (such as 3-chymotrypsin-like protease (3CL Pro), papain-like protease, helicase, and RNA-dependent RNA polymerase (RdRp)), structural proteins (such as spike glycoprotein) and accessory proteins. The four non-structural proteins and the spike glycoprotein mentioned above are essential for the viral life cycle and virus–cell receptor interactions; consequently, these five proteins were recognized as attractive targets. Motivated by this knowledge, we used 3CL Pro and RdRp as examples and then applied the ELECTRA-DTA model trained on the BindingDB database to report the binding affinity scores of each drug candidate and targeted protein. Drug candidates were ranked according to their predicted binding affinity scores with 3CL Pro and RdRp (as shown in supplementary table S1). Additionally, the top 10 drugs with the highest binding affinity for 3CL Pro or RdRp are also provided, for a total of 14 drugs, as listed in Table 8.

Among these, saquinavir, an antiviral protease inhibitor, was predicted to have the highest binding affinity with 3CL Pro with ELECTRA-DTA. In Singapore, saquinavir has been used to treat patients with SARS-CoV-2. When RdRp was used as the target, daclatasvir had the strongest binding affinity value. Some studies have supported daclatasvir as a potential drug for treating COVID-19. Darunavir and cobicistat were both among the top-ranked drugs. In an ongoing clinical trial in China, darunavir and cobicistat were assessed for their efficacy and safety. Lopinavir has been reported to inhibit SARS-CoV-2 at a half-maximum efficacy concentration (EC50)—the level of a drug that induces a response halfway between baseline and
maximum after a specified exposure time—of 26.36 \( \mu \)M. Lopinavir has also been listed as an emergency drug for COVID-19 [26]. Administration of lopinavir as an emergency drug in China was shown to increase the eosinophil count among COVID-19 patients [27].

Overall, there is promising evidence from the literature or clinical trials on 12 recommendations out of the 14 highest-affinity drugs, which illustrates the potential of ELECTRA-DTA for recommending high-quality repurposing candidates. Consequently, the abovementioned drugs could be regarded as candidates to treat COVID-19 infections, but further study on the efficiency of these drugs is also necessary. It should be noted that our ELECTRA-DTA model only uses target and drug sequence information without any details on current ongoing clinical trials. Information from clinical trials was only used to affirm the efficacy of our approach.

Discussion
This study presented ELECTRA-DTA, a DTA prediction approach with two trained ELECTRA models for encoding protein amino acid sequences and compound SMILESs separately in an unsupervised manner. For compound SMILES strings, all characters found in different SMILESs were treated as the vocabulary, and selected compounds from the PubChem dataset were used as a corpus to obtain vector representations of each character, subsequently yielding a vector sequence for each compound SMILES. The unsupervised training on a large-scale sample space incorporated additional contextual information, including potential substructures or function groups. The ELECTRA-based embedding mechanism and the design of our network architecture ensured a better predictive ability than that of state-of-the-art baseline methods; moreover, they provide a feasible encoding module that can be stacked over other feature representation networks as the embedding layer for protein sequences or compound SMILESs in various learning tasks. In addition, they can also effectively train a new embedding model by constructing domain-specific corpora for specific downstream tasks.

Acknowledgements
We thank the anonymous reviewers very much for their effort in evaluating our paper.

Funding
This work is supported by the National Natural Science Foundation of China (NSFC, Grant no. 62171164, 62102191 and 61872114), and the National key Research & Development plan of Ministry of Science and Technology of China (Grant no. 2018YFC1314900, 2018YFC1314901).

Availability of data and materials
Source code and datasets are available at Github.

Competing interests
The authors declare that they have no competing interests.

Authors’ contributions
Lingling Zhao, Junjie Wang and Liang Cheng substantially contributed to the conception and design of the study and acquisition of data; NaiFeng Wen analysed and interpreted the data; Junjie Wang, Lingling Zhao and Chunyu Wang drafted the article.

Author details
1 Department of Medical Informatics, School of Biomedical Engineering and Informatics, Nanjing Medical University, Nanjing, P. R. China. 2 School of Mechanical and Electrical Engineering, Dalian Minzu University, Dalian, P. R. China. 3 Faculty of Computing, Harbin Institute of Technology, Harbin, P. R. China. 4 College of Bioinformatics Science and Technology, Harbin Medical University, Harbin, P. R. China.
References

1. He, T., Heidemeyer, M., Ban, F., Cherkasov, A., Ester, M.: Simboost: a read-across approach for predicting drug–target binding affinities using gradient boosting machines. Journal of Cheminformatics 9(1), 24–24 (2017)

2. Pahikkala, T., Airola, A., Pietilä, S., Shakyawar, S., Szewcja, A., Tang, J., Aittokallio, T.: Toward more realistic drug–target interaction predictions. Briefings in bioinformatics 16(2), 325–337 (2015)

3. Öztürk, H., Özgür, A., Ozkirim, E.: Deepdt: deep drug–target binding affinity prediction. Bioinformatics 34(17), 821–829 (2018)

4. Karimi, M., Wu, D., Wang, Z., Shen, Y.: Deepaffinity: interpretable deep learning of compound–protein affinity through unified recurrent and convolutional neural networks. Bioinformatics 35(18), 3329–3338 (2019)

5. Sutskever, I., Vinyals, O., Le, Q.V.: Sequence to sequence learning with neural networks. In: Advances in Neural Information Processing Systems 27, vol. 27, pp. 3104–3112 (2014)

6. Devlin, J., Chang, M.-W., Lee, K., Toutanova, K.N.: Bert: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pp. 4171–4186 (2018)

7. Clark, K., Luong, M.-T., Le, Q.V., Manning, C.D.: Electra: Pre-training text encoders as discriminators rather than generators. In: ICLR 2020: Eighth International Conference on Learning Representations (2020)

8. Bolton, E.E., Wang, Y., Thiessen, P.A., Bryant, S.H.: Pubchem: Integrated platform of small molecules and biological activities. vol. 4, pp. 237–241 (2008)

9. Kim, S., Thiessen, P.A., Bolton, E.E., Chen, J., Fu, G., Gindulyte, A., Han, L., He, J., He, S., Shoemaker, B.A., Wang, J., Yu, B., Zhang, J.-J., Bryant, S.H.: Pubchem substance and compound databases. Nucleic Acids Research 44, 1202–1213 (2016)

10. Kim, S., Chen, J., Bengtsson, A., Welleda, T., Zheng, J., Blackwell, G., Shoemaker, B.A., Thiessen, P.A., Yu, B., Zhang, J., Aslava, L., Zhang, J., Bolton, E.E.: Pubchem 2019 update: improved access to chemical data. Nucleic Acids Research 47 (2019)

11. O’Boyle, N.M., Banck, M., James, C.A., Morley, C., Vandermeersch, T., Hutchison, G.R.: Open babel: An open chemical toolbox. Journal of Cheminformatics 3(1), 33–33 (2011)

12. Bateman, A., Martin, M.J., O’Donovan, C., Magrane, M., Apweiler, R., Alpi, E., Antunes, R., Arganika, J., Bely, B., Bingley, M., Bonilla, C., Britto, R., Bursteinas, B., Chavali, G., Cibrian-Uhalte, E., Silva, A.D., Giorgi, M.D., Dogan, T., Fazzini, F., Gane, P., Castro, L.G., Garvity, P., Hatan-Ellis, E., Hita, R., Huntley, R., Legge, D., Liu, W., Luo, J., Macdougall, A., Mutowu, P., Nightingale, A., Orchard, S., Pichler, K., Poggioi, D., Purdor, S., Pureza, L., Qi, G., Rosanno, S., Saiedi, R., Sawford, T., Shypetsyna, A., Turner, E., Volynkin, V., Wardell, T., Watkins, X., Zelning, H., Cowley, A., Figiurea, L., Li, W., McWilliam, H., Lopez, R., Xeranos, I., Bouguerelet, L., Bridge, A., Poux, S., Redaschi, N., Aimo, L., Argoud-Puy, G., Aucshinloss, A., Axelsen, K., Bansal, P., Baratin, D., Blatter, M.C., Boeckmann, B., Bolleman, J., Boutet, E., Brezua, L., Casal-Casas, C., Castro, E.D., Coudert, E., Cuche, B., Doche, M., Donev, D., Duvada, S., Estreicher, A., Famiglietti, L., Feueraman, M., Gasteiger, E., Gehant, S., Gerritsen, V., Greco, D., Gromov-Gumowski, N., Hinz, U., Hulo, C., Jurgf, F., Keller, G., Lara, V., Lemercier, P., Lieberherr, D., Lombardot, T., Martin, X., Masson, P., Morgat, A., Neto, T., Nouisipel, N., Paesano, S., Pedruzzi, I., Pilbout, S., Pozzato, M., Pruss, M., Rivoire, C., Roechert, B., Schneider, M., Sigrist, C., Soneson, K., Stahl, S., Stutz, A., Sundaram, S., Tognolli, M., Verbregue, L., Veltine, A., Wu, C.H., Aghighi, C.N., Arminksi, L., Chen, C., Chen, Y., Ravelli, J.S., Huang, H., Laiho, K., McGeary, P., Natale, D.A., Suszek, B.E., Vinayaka, C.R., Wang, Q., Wang, Y., Yeh, L.S., Yerramulla, M.S., Zhang, J.: Uniprot: A hub for protein information. Nucleic Acids Research 43 (2015)

13. Rohanian, O., Taslimipoor, S., Kouchaki, S., Ha, L.A., Mitkov, R.: Bridging the gap: Attending to discontinuity in identification of multword expressions. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pp. 2692–2698 (2019)

14. Tang, J., Shakyawar, S., Xu, T., Hintsanen, P., Wennerberg, K., Aittokallio, T.: Making sense of drug–target binding affinities using gradient boosting machines. Journal of Cheminformatics 9(1), 24–24 (2017)

15. Davis, M.I., Hunt, J.P., Herrgard, S., Ciceri, P., Wodicka, L.M., Pallares, G., Hocker, M., Treiber, D.K., Tybulewicz, V.L.J., Crawford, C.E., Jackson, P.K., Bronco, P.K., Mulligan, R.C.: Neonatal lethality and lymphopenia in mice with a homozygous disruption of the c-abl proto-oncogene. Cell 65(7), 1153–1163 (1991)

16. Cristiani, C., Rusconi, L., Pergo, R., Schiering, N., Kalisz, H.M., Knapp, S., Isauchi, A.: Regulation of the wild-type and y1235d mutant met kinase activation. Biochemistry 44(43), 4110–4119 (2005)

17. Gonen, M., Heller, G.: Concordance probability and discriminatory power in proportional hazards regression. Biometrika 92(4), 965–970 (2005)

18. Pratim Roy, P., Paul, S., Mitra, I., Roy, K.: On two novel parameters for validation of predictive qsar models. Molecules 14(5), 1660–1701 (2009)

19. Roy, K., Chakraborty, P., Mitra, I., Ojha, P.K., Kar, S., Das, R.N.: Some case studies on application of “rm2” metrics for judging quality of quantitative structure–activity relationship predictions: emphasis on scaling of response data. Journal of computational chemistry 34(12), 1071–1082 (2013)

20. Abbasi, K., Razzazong, P., Poso, A., Amanlou, M., Ghasemi, J.B., Masoudi-Nejad, A.: Deepcat: Deep cross-domain compound-protein affinity prediction through lstm and convolutional neural networks. Bioinformatics 36(17), 4633–4642 (2020)

21. Zhao, Q., Xiao, F., Yang, M., Li, Y., Wang, J.: Attentiondta: prediction of drug–target binding affinity using attention model. In: 2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), pp. 64–69 (2019)
24. Hosseini, F.S., Amanlou, M.: Anti-hcv and anti-malaria agent, potential candidates to repurpose for coronavirus infection: Virtual screening, molecular docking, and molecular dynamics simulation study. Life sciences 258, 118205 (2020)

25. Pant, S., Singh, M., Ravichandiran, V., Murty, U., Srivastava, H.K.: Peptide-like and small-molecule inhibitors against covid-19. Journal of Biomolecular Structure and Dynamics, 1–10 (2020)

26. Hu, F., Jiang, J., Yin, P.: Prediction of potential commercially inhibitors against sars-cov-2 by multi-task deep model. arXiv preprint arXiv:2003.00728 (2020)

27. Feng, Z., Chen, M., Xue, Y., Liang, T., Chen, H., Zhou, Y., Nolin, T.D., Smith, R.B., Xie, X.-Q.: Mccs: a novel recognition pattern-based method for fast track discovery of anti-sars-cov-2 drugs. Briefings in bioinformatics 22(2), 946–962 (2021)

Additional Files

Additional file 1 — table 8.docx
Ranked drug candidates for the 3CL protease and RdRp proteins of SARS-CoV-2 with predicted affinity.

Additional file 2 — duplicate-davis.txt
The removed samples for the Davis Dataset.

Additional file 3 — duplicate-kiba.txt
The removed samples for the KIBA Dataset.

Additional file 3 — duplicate-bindingdb.txt
The removed samples for the BindingDB Dataset.
Figure 1

Overview of ELECTRA-DTA
The principle of ELECTRA pre-training

Figure 2

Compound SMILES and protein sequence embedding

Figure 3
Figure 4

Architecture of the ELECTRA-DTA model

Figure 5

Structure of the squeeze-and-excitation block
Figure 6 reports a seemingly contradictory phenomenon.

**Supplementary Files**

This is a list of supplementary files associated with this preprint. Click to download.

- duplicatedavis.txt
- duplicatekiba.txt
- duplicatebindingdb.txt