

Data and text mining

DeepPurpose: a deep learning library for drug–target interaction prediction

Kexin Huang¹, Tianfan Fu², Lucas M. Glass³, Marinka Zitnik¹, Cao Xiao³ and Jimeng Sun^{4,*}

¹Harvard University, Boston, MA 02115, USA, ²Georgia Institute of Technology, Atlanta, GA 30332, USA, ³IQVIA, Cambridge, MA 02139, USA and ⁴University of Illinois at Urbana-Champaign, Urbana, IL 61801, USA

*To whom correspondence should be addressed.

Associate Editor: Jonathan Wren

Received on September 24, 2020; revised on November 2, 2020; editorial decision on November 13, 2020; accepted on November 19, 2020

Abstract

Summary: Accurate prediction of drug–target interactions (DTI) is crucial for drug discovery. Recently, deep learning (DL) models show promising performance for DTI prediction. However, these models can be difficult to use for both computer scientists entering the biomedical field and bioinformaticians with limited DL experience. We present DeepPurpose, a comprehensive and easy-to-use DL library for DTI prediction. DeepPurpose supports training of customized DTI prediction models by implementing 15 compound and protein encoders and over 50 neural architectures, along with providing many other useful features. We demonstrate state-of-the-art performance of DeepPurpose on several benchmark datasets.

Availability and implementation: <https://github.com/kexinhuang12345/DeepPurpose>.

Contact: jimeng@illinois.edu

Supplementary information: [Supplementary data](#) are available at *Bioinformatics* online.

1 Introduction

Drug–target interactions (DTI) characterize the binding of compounds to protein targets (Santos *et al.*, 2017). Accurate identification of molecular drug targets is fundamental for drug discovery and development (Rutkowska *et al.*, 2016; Zitnik *et al.*, 2019) and is especially important for finding effective and safe treatments for new pathogens, including SARS-CoV-2 (Velavan and Meyer, 2020).

Deep learning (DL) has advanced traditional computational modeling of compounds by offering an increased expressive power in identifying, processing and extrapolating complex patterns in molecular data (Lee *et al.*, 2019; Öztürk *et al.*, 2018). There are many DL models designed for DTI prediction (Lee *et al.*, 2019; Nguyen *et al.*, 2020; Öztürk *et al.*, 2018). However, to generate predictions, deploy DL models in practice, test and evaluate model performance, one needs considerable programming skills and extensive biochemical knowledge. Prevailing tools are designed for experienced interdisciplinary researchers. They are challenging to use by both computer scientists entering the biomedical field and domain bioinformaticians with limited experience in training and deploying DL models. Furthermore, each open-sourced tool has a different programming interface and is coded differently, which prevents easy integration of outputs from various methods for model ensembles (Yang *et al.*, 2019).

Here, we introduce DeepPurpose, a DL library for encoding and downstream prediction of proteins and compounds. DeepPurpose allows rapid prototyping via a programming framework that implements over 50 DL models, seven protein encoders and eight compound encoders. Empirically, we find that models implemented in DeepPurpose achieve state-of-the-art prediction performance on DTI benchmark datasets.

2 DeepPurpose library

DL models for DTI prediction can be formulated as an encoder-decoder architectures (Cho *et al.*, 2014). DeepPurpose library implements a unifying encoder-decoder framework, which makes the library uniquely flexible. By merely specifying an encoder's name, the user can automatically connect an encoder of interest with the relevant decoder. DeepPurpose then trains the corresponding encoder-decoder model in an end-to-end manner. Finally, the user accesses the trained model either programmatically or via a visual interface and uses the model for DTI prediction.

2.1 Module for encoding proteins and compounds

DeepPurpose takes the compound's simplified molecular-input line-entry system (SMILES) string and protein amino acid sequence pair



pre-trained deep model for prediction. This list can then be examined to identify promising compound candidates for further experiments. Second, DeepPurpose also supports user-friendly programming frameworks for other modeling tasks, including drug and protein property prediction, drug–drug interaction prediction and protein–protein interaction prediction (see [Supplementary Material](#)). Third, DeepPurpose provides an interface to many types of data, including public large binding affinity dataset ([Liu *et al.*, 2007](#)), bioassay data ([Kim *et al.*, 2019](#)) and a drug repurposing library ([Corsello *et al.*, 2017](#)).

The functionality of DeepPurpose is modularized into six key steps where a single line of code can invoke each step: (i) Load the dataset from a local file or load a DeepPurpose benchmark dataset. (ii) Specify the names of compound and protein encoders. (iii) Split the dataset into training, validation and testing sets using `data_process` function, which implements a variety of data-split strategies. (iv) Create a configuration file and specify model parameters. If needed, DeepPurpose can automatically search for optimal values of hyperparameters. (v) Initialize a model using the configuration file. Alternatively, the user can load a pre-trained model or a previously saved model. (vi) Finally, train the model using `train` function and monitor the progress of training and performance metrics. DeepPurpose is OS-agnostic and uses the Jupyter Notebook interface. It can be run in the cloud or locally. All datasets, models, documentation, installation instructions and tutorials are provided at <https://github.com/kexinhuang12345/DeepPurpose>.

To demonstrate the use of DeepPurpose, we compare DeepPurpose with KronRLS (Pahikkala *et al.*, 2015), a popular DTI method, and GraphDTA (Nguyen *et al.*, 2020) and DeepDTA (Öztürk *et al.*, 2018), state-of-the-art DL methods. We find that many DeepPurpose models achieve comparable prediction performance on two benchmark datasets, DAVIS (Davis *et al.*, 2011) and KIBA (He *et al.*, 2017) (Fig. 1D). A complete script to generate the results is provided in [Supplementary Material](#).

4 DeepPurpose with interactive web interface

In addition to rapid model prototyping, DeepPurpose also provides utility functions to load a pre-trained model and make predictions for a new drug and target inputs. This functionality allows domain scientists to examine predictions quickly, modify the inputs based on predictions, and iterate on the process until finding a drug or target with desired properties. We leverage Gradio (Abid *et al.*, 2019) to create a web interface programmatically. We use a user-trained DeepPurpose model in the backend and create a custom web interface in fewer than ten code lines. This web interface takes the SMILES and amino acid sequence as the input and returns prediction scores with less than 1-second latency. We provide examples in the [Supplementary Material](#).

Financial Support: none declared. M.Z. and K.H. are supported, in part, by NSF grant nos. IIS-2030459 and IIS-2033384, and by the Harvard Data Science Initiative. T.F. and J.S. was in part supported by the NSF SCH-2014438, IIS-1418511, CCF-1533768, IIS-1838042, the NIH award NIH R01 1R01NS107291-01 and R56HL138415.

Conflict of Interest: none declared.

References

- Abid, A. *et al.* (2019) Gradio: Hassle-free sharing and testing of ml models in the wild. *arXiv preprint arXiv:1906.02569*.
- Cho, K. *et al.* (2014) On the properties of neural machine translation: encoder-decoder approaches. In: *Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation at ACL*, Doha, Qatar, pp. 103–111.
- Corsello, S.M. *et al.* (2017) The drug repurposing hub: a next-generation drug library and information resource. *Nat. Med.*, **23**, 405–408.
- Davis, M.I. *et al.* (2011) Comprehensive analysis of kinase inhibitor selectivity. *Nat. Biotechnol.*, **29**, 1046–1051.
- He, T. *et al.* (2017) SimBoost: a read-across approach for predicting drug–target binding affinities using gradient boosting machines. *J. Cheminf.*, **9**, 24.
- Kim, S. *et al.* (2019) Pubchem 2019 update: improved access to chemical data. *Nucleic Acids Res.*, **47**, D1102–D1109.
- Lee, I. *et al.* (2019) DeepConv-DTI: prediction of drug–target interactions via deep learning with convolution on protein sequences. *PLOS Comput. Biol.*, **15**, e1007129.
- Liu, T. *et al.* (2007) BindingDB: a web-accessible database of experimentally determined protein–ligand binding affinities. *Nucleic Acids Res.*, **35**, D198–D201.
- Nguyen, T. *et al.* (2020) GraphDTA: Predicting drug–target binding affinity with graph neural networks. *bioRxiv*.
- Öztürk, H. *et al.* (2018) DeepDTA: deep drug–target binding affinity prediction. *Bioinformatics*, **34**, i821–i829.
- Pahikkala, T. *et al.* (2015) Toward more realistic drug–target interaction predictions. *Brief. Bioinf.*, **16**, 325–337.
- Rutkowska, A. *et al.* (2016) A modular probe strategy for drug localization, target identification and target occupancy measurement on single cell level. *ACS Chem. Biol.*, **11**, 2541–2550.
- Santos, R. *et al.* (2017) A comprehensive map of molecular drug targets. *Nat. Rev. Drug Discov.*, **16**, 19–34.
- Velavan, T.P. and Meyer, C.G. (2020) The COVID19 epidemic. *Trop. Med. Int. Health*, **25**, 278–280.
- Yang, K. *et al.* (2019) Analyzing learned molecular representations for property prediction. *J. Chem. Inf. Model.*, **59**, 3370–3388.
- Zitnik, M. *et al.* (2019) Machine learning for integrating data in biology and medicine: principles, practice, and opportunities. *Inf. Fus.*, **50**, 71–91.