Literature discussions

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

Dongdong Zhang

Song Xia

December 21st 2020

What we know about DeepPurpose

- Solve drug-target interaction (DTI) problems, drug properties predictions
- Implementing 15 compound and protein encoders and over 50 neural architectures
- By merely specifying an encoder’s name, users can automatically connect a encoder of interest with the relevant decoder
- DeepPurpose then trains the corresponding encoder-decoder model in an end-to-end manner

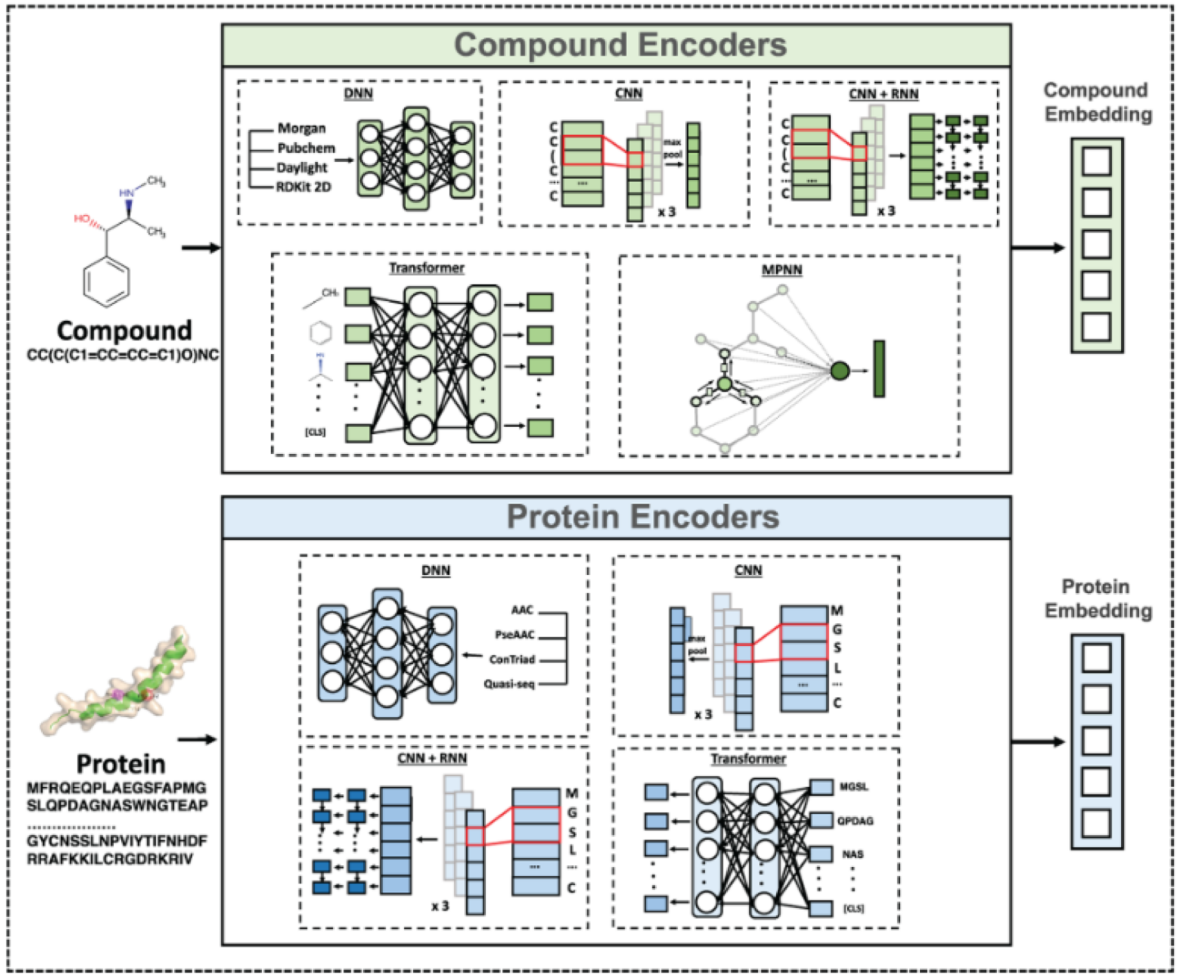
Overview of the encoders

Drug Encodings	Description
Morgan	Extended-Connectivity Fingerprints
Pubchem	Pubchem Substructure-based Fingerprints
Daylight	Daylight-type fingerprints
rdkit_2d_normalized	Normalized Descriptastorus
ESPF	Explainable Substructure Partition Fingerprint
CNN	Convolutional Neural Network on SMILES
CNN_RNN	A GRU/LSTM on top of a CNN on SMILES
Transformer	Transformer Encoder on ESPF
MPNN	Message-passing neural network

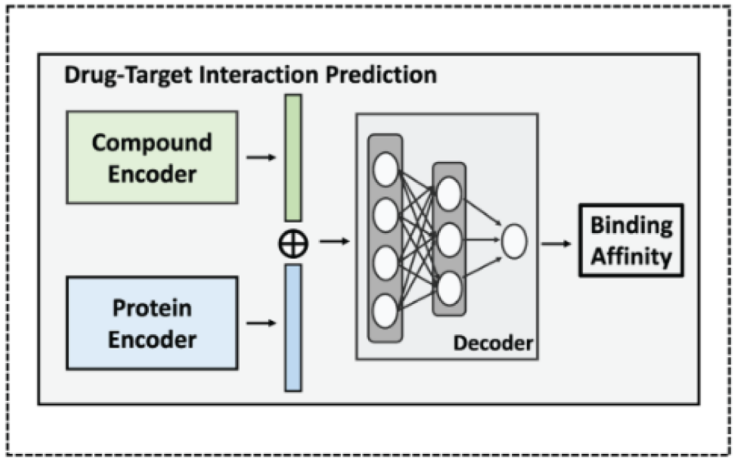
Target Encodings	Description
AAC	Amino acid composition up to 3-mers
PseudoAAC	Pseudo amino acid composition
Conjoint_triad	Conjoint triad features
Quasi-seq	Quasi-sequence order descriptor
ESPF	Explainable Substructure Partition Fingerprint
CNN	Convolutional Neural Network on target seq
CNN_RNN	A GRU/LSTM on top of a CNN on target seq
Transformer	Transformer Encoder on ESPF

Overview of this framework

A. Molecular Encoding Module



B. Drug-Target Interaction Prediction



C. DeepPurpose 10 Lines Framework

```
>>> from DeepPurpose import DTI
>>> from DeepPurpose.utils import *
>>> from DeepPurpose.dataset import *
>>>
>>> X_drug, X_target, y = load_process_DAVIS(SAVE_PATH, binary=False)
>>>
>>> drug_encoding, target_encoding = 'CNN', 'CNN'
>>> train, val, test = data_process(X_drug, X_target, y, drug_encoding, \
                                    target_encoding, split_method='random', \
                                    frac=[0.7,0.1,0.2], random_seed = 1)
>>>
>>> config = generate_config(drug_encoding, target_encoding, \
                             cls_hidden_dims = [1024,1024,512], \
                             train_epoch = 100, LR = 0.001, batch_size = 256, \
                             cnn_drug_filters = [32,64,96], \
                             cnn_drug_kernels = [4,8,12], \
                             cnn_target_filters = [32,64,96], \
                             cnn_target_kernels = [4,8,12])
>>>
>>> model = DTI.model_initialize(**config)
>>> model.train(train, val, test)
```

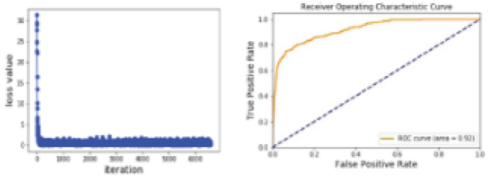
D. SOTA Prediction Performance

Dataset 1: DAVIS	Model	MSE	Concordance Index
Baselines	KronRLS	0.329 (0.019)	0.847 (0.006)
	GraphDTA	0.263 (0.015)	0.864 (0.007)
	DeepDTA	0.262 (0.022)	0.870 (0.003)
DeepPurpose	CNN+CNN	0.254 (0.018)	0.879 (0.011)
	MPNN+CNN	0.271 (0.012)	0.858 (0.007)
	MPNN+AAC	0.242 (0.009)	0.881 (0.005)
	CNN+Trans	0.282 (0.009)	0.852 (0.006)
	Morgan+CNN	0.271 (0.012)	0.858 (0.007)
	Morgan+AAC	0.258 (0.012)	0.861 (0.008)
	Daylight+AAC	0.277 (0.014)	0.861 (0.008)

Dataset 2: KIBA	Model	MSE	Concordance Index
Baselines	KronRLS	0.852 (0.014)	0.688 (0.003)
	GraphDTA	0.183 (0.003)	0.862 (0.005)
	DeepDTA	0.196 (0.008)	0.864 (0.002)
DeepPurpose	CNN+CNN	0.196 (0.005)	0.856 (0.004)
	MPNN+CNN	0.222 (0.006)	0.825 (0.003)
	MPNN+AAC	0.178 (0.002)	0.872 (0.001)
	CNN+Trans	0.240 (0.013)	0.818 (0.004)
	Morgan+CNN	0.229 (0.008)	0.825 (0.004)
	Morgan+AAC	0.233 (0.009)	0.823 (0.004)
	Daylight+AAC	0.252 (0.014)	0.808 (0.008)

E. More Functionalities

- **Training Monitoring and Metrics Auto-generation**



- **Repurposing and Screening Ranked List Generation**

Drug Repurposing Result for SARS-CoV2 3CL Protease

Rank	Drug Name	Target Name	Binding Score
1	Sofosbuvir	SARS-CoV2 3CL Protease	360.22
2	Daclatasvir	SARS-CoV2 3CL Protease	424.06
3	Velpatasvir	SARS-CoV2 3CL Protease	423.78
4	Elbasvir	SARS-CoV2 3CL Protease	768.33

- **Supporting Drug Property, Protein Function, Drug-Drug Interaction Prediction, and Protein-Protein Interaction Prediction, all less than 10 lines of codes.**

Compound properties: FreeSol benchmark

Encoder	Test RMSE (kcal/mol)
Morgan	2.01
CNN	1.81
MPNN	1.56
Transformer	3.78

By Dongdong

100 epochs
3 encoder layers
128 in dimension

Drug-target Interaction benchmark

Number	Drug Encoder	Target Encoder	Testing MSE	Pearson Correlation	Concordance Index
1	Transformer	Transformer	0.75	0.31	0.64
2	MPNN	Transformer	0.70	0.41	0.71
3	Morgan	Transformer	**0.50**	**0.62**	0.79
4	Morgan	CNN	0.52	0.62	**0.80**
5	MPNN	CNN	0.58	0.54	0.76
6	Transformer	CNN	0.75	0.31	0.65

By Song

Pros and cons

- Pros:
  - Easily call for different encoding methods for compounds and proteins
  - Training deployment is also very convenient
  - Good to compare traditional Fingerprint-based QSAR with GNN and NLP

- Cons:
  - Only one type of MPNN model provided
  - Only the very fundamental transformer framework provided
  - Not good for customizing and designing new models