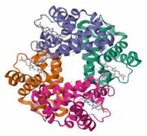


DeepProtein

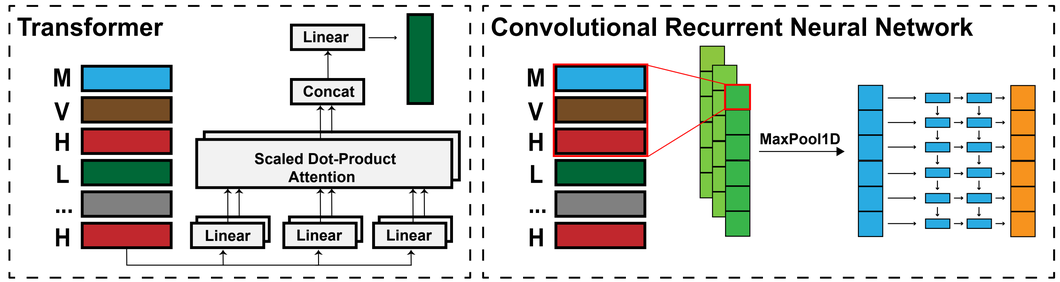
A. Database

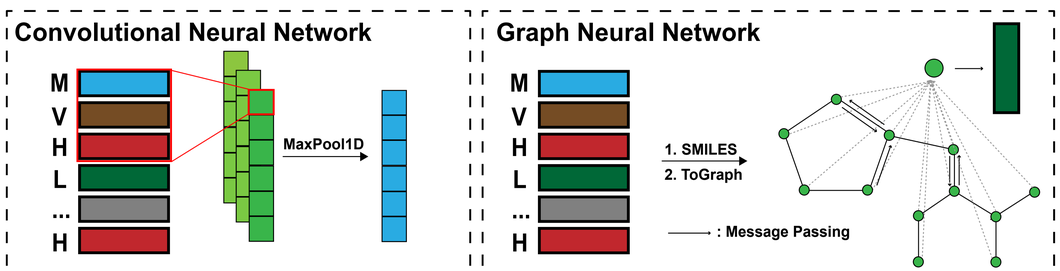
Extraction

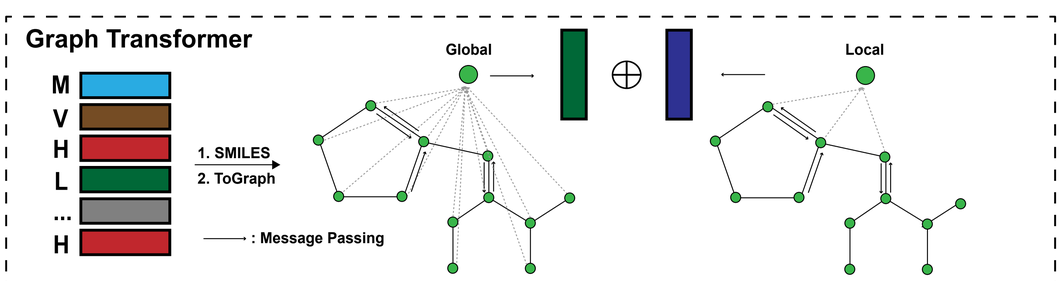
Protein

Name:
human deoxyhaemoglobin
Sequence:
MVHLTPEEKSAVTALWGKVNVD
.....
VQAAYQKVVAGVANALAHKYH

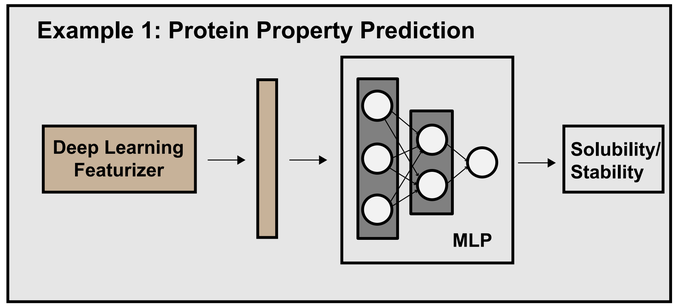
B. Deep Learning Featurizers in DeepProtein

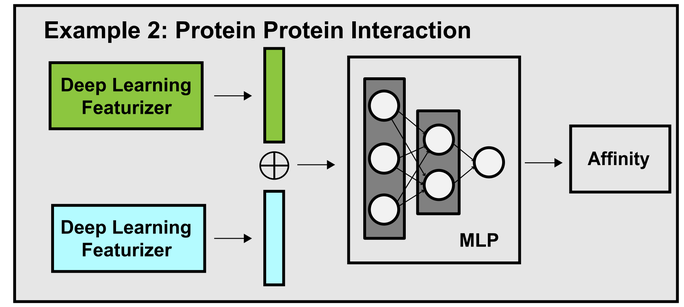






C. Prediction Tasks





D. DeepProtein 20 Lines of Codes

```
>>> import os, sys, argparse, torch, wandb

>>> from DeepProtein.dataset import *
>>> import DeepProtein.utils as utils
>>> import DeepProtein.ProteinPred as models

>>> path = os.getcwd()
>>> train_beta = Beta_lactamase(path + '/DeepProtein/data', 'train')
>>> valid_beta = Beta_lactamase(path + '/DeepProtein/data', 'valid')
>>> test_beta = Beta_lactamase(path + '/DeepProtein/data', 'test')

>>> train_protein_processed, train_target, train_protein_idx = collate_fn(train_beta)
>>> valid_protein_processed, valid_target, valid_protein_idx = collate_fn(valid_beta)
>>> test_protein_processed, test_target, test_protein_idx = collate_fn(test_beta)

>>> target_encoding = 'CNN'
```

```
>>> train, _, _ = utils.data_process(X_target =train_protein_processed, y =train_target, \
    target_encoding =target_encoding, split_method ='random', frac=[0.99998, 1e-5, 1e-5])

>>> _, val, _ = utils.data_process(X_target =valid_protein_processed, y =valid_target, \
    target_encoding =target_encoding, split_method ='random', frac=[1e-5, 0.99998, 1e-5])

>>> _, _, test = utils.data_process(X_target =test_protein_processed, y =test_target, \
    target_encoding =target_encoding,split_method ='random', frac=[1e-5, 1e-5, 0.99998])

>>> config = generate_config(target_encoding =target_encoding, \
                             cls_hidden_dims =[1024, 1024], \
                             train_epoch =20, \
                             LR=0.0001, \
                             batch_size =32, \
                             )
>>> config['multi'] = False
>>> torch.manual_seed(args.seed)
>>> model = models.model_initialize( **config)
>>> model.train(train, val, test, compute_pos_enc = False)
```

E. Prediction Performance (Comparison) and loss

Model # train/valid/test	PPI Affinity (R^2 ↑) 2127 / 212 / 343	Yeast (PR-AUC ↑) 1668 / 131 / 373	Human PPI (PR-AUC ↑) 6844 / 277 / 227
CNN	0.493 ± 0.015	51.93 ± 0.92	70.37 ± 1.22
CNN-RNN	$\mathbf{0.584 \pm 0.026}$**	53.28 ± 0.85	70.45 ± 2.68
Transformer	0.425 ± 0.021	53.79 ± 1.07	59.36 ± 4.00
GCN	0.394 ± 0.006	$\mathbf{58.98 \pm 0.72}$	$\mathbf{82.21 \pm 1.13}$
GAT	0.230 ± 0.001	53.72 ± 0.39	77.63 ± 3.13
NeuralFP	0.100 ± 0.054	57.00 ± 1.51	80.11 ± 1.25