Chiral Molecule Classification Project

Name: Yichuan Peng E-mail: pyc617633844@sjtu.edu.cn

Scikit-Learn Related Method, XGBoost Algorithm And some Transfer Learning Models

In this part, I have used the dataset `data_3d_aug.pkl` . In my research, I use `E3FP (Extended-3-Point Fingerprint)` for molecular representation. A 1024×1 vector is used to save the fingerprint.

K-Nearest-Neighbor Algorithm And SVM

Failed attempt, these classifiers categorize all molecules into one class of central chiral molecules, resulting in poor performance.

XGBoost Algorithm

XGBoost demonstrates a superior classification performance with the following accuracy rates:

Learning rate	Accuracy(%)	Central(%)	Axial(%)	Helical(%)	Planar(%)
0.00005	96.90	99.46	40.94	14.29	25.53
0.0001	97.21	99.8	40.16	10.71	27.66
0.0002	96.69	99.52	37.8	0	12.77
0.0005	96.77	99.74	29.92	7.14	14.89
0.0008	96.84	99.58	37.01	7.14	21.27
0.001	97.08	99.76	43.31	3.57	12.77
0.0015	96.44	99.24	29.92	7.14	31.91
0.002	97.02	99.8	44.09	0	2.13

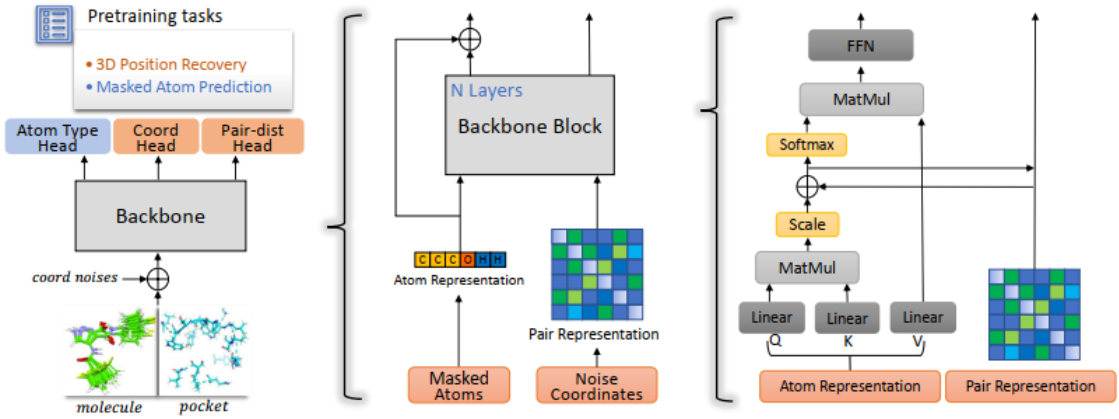
Some Other Transfer Learning Methods

Methods	Accuracy(%)	Center(%)	Axial(%)	Helical(%)	Planar(%)
MLP	95.42	98.48	24.51	0	18.09
ResNet	96.53	99.81	16.21	0	21.28
IRMv1	95.57	98.58	26.88	0	18.09
REx	96.51	98.7	24.51	0	15.96
DANN	96.24	98.10	32.41	0	12.78

It can be observed that these models exhibit poor classification performance for helical chiral molecules.

Uni-Mol

Uni-mol choose Transformer as the backbone model in Uni-Mol, as it fully connects the nodes/atoms and thus can learn the possible long-range interactions. As is shown in the figure, the Uni-Mol backbone is a Transformer based model. It has two inputs, atom types and atom coordinates. And two representations (atom and pair) are maintained in the model. The atom representation is initialized from atom types, by the Embedding layer; The pair representation is initialized by invariant spatial positional encoding calculated from atom coordinates. In particular, based on pair-wise Euclidean distances among atoms, the pair representation is invariant to global rotation and translation. The two representations communicate with each other in self-attention module.



Author has provided pre-train models. I used Uni-Mol for the finetuning task. And the Result are as follows:

train test fraction	accuracy(%)	Center(%)	Axial(%)	Helical(%)	Planar(%)
7:3	99.35	99.94	89.74	70.0	76.92
3:7	97.39	99.71	68.66	59.26	7.14
5:5	97.10	99.77	65.32	27.27	50.0