

BS6207 ADVANCED ARTIFICIAL INTELLIGENCE FOR BIOMEDICAL DATA SCIENCE

Proteins and Ligands Binding Classification

Ni Yuxin

Contents



Introduction



Preprocessing



Hyperparameter tuning



Results



Conclusion

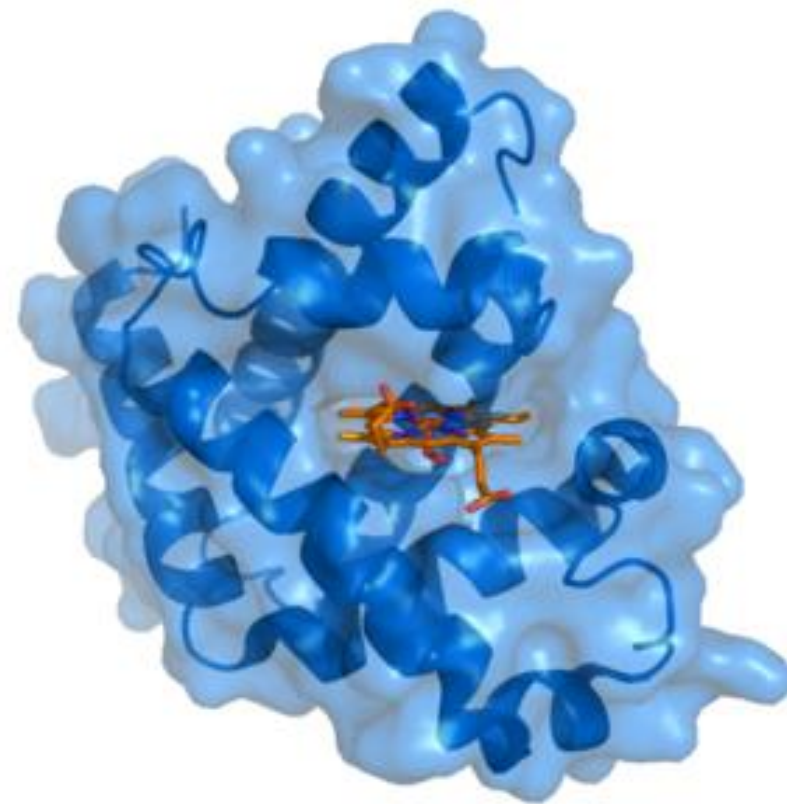
Introduction

A ligand is a substance that forms a complex with a biomolecule to serve a biological purpose.

Molecular recognition between proteins and ligands plays an important role in many biological processes, such as membrane receptor signaling and enzyme catalysis.

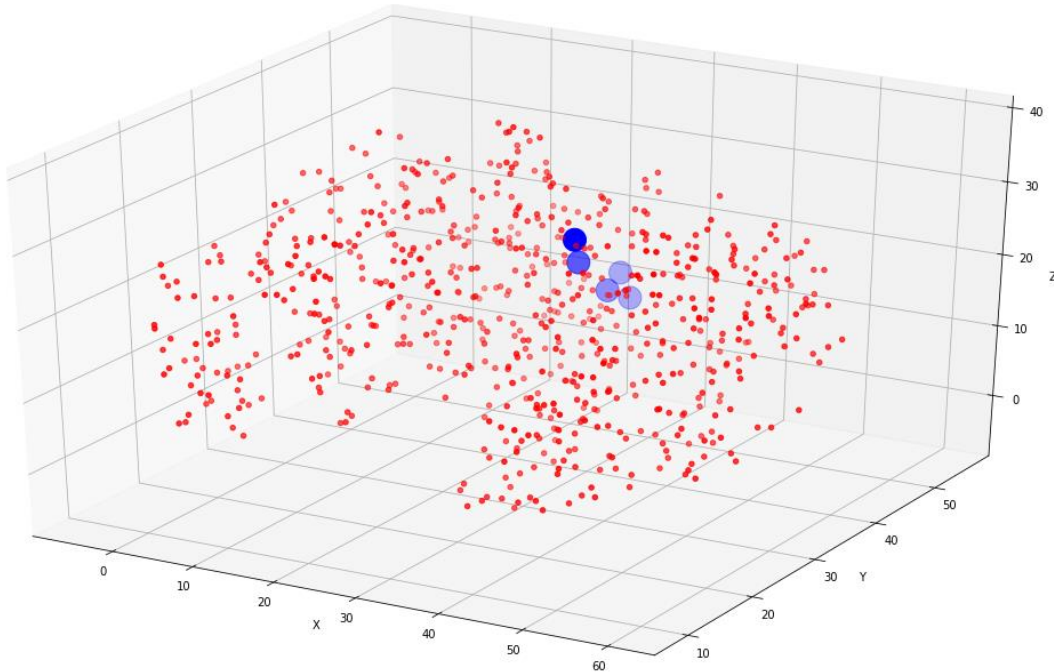
Target:

Use a deep learning model to predict protein and ligand binding.

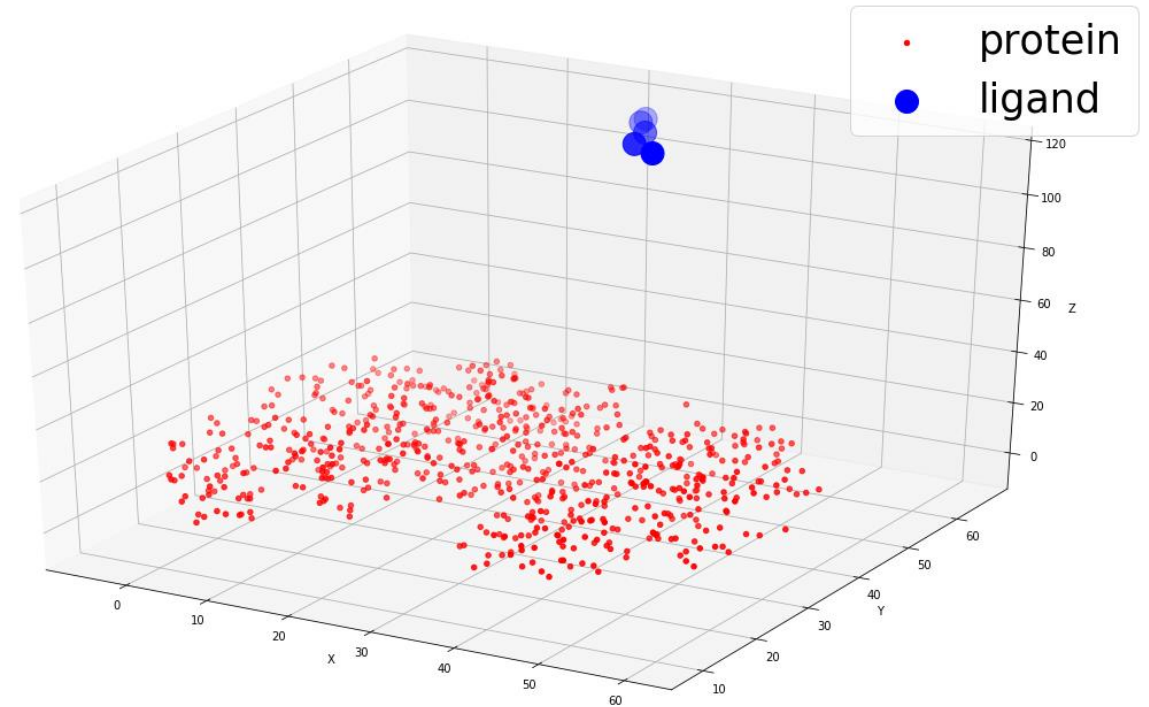


Protein ligand complex

Visualization



Binding: Protein 0001 and ligand 0001



Not Binding: Protein 0001 and ligand 0004

Contents



Introduction



Preprocessing



Hyperparameter tuning



Results



Conclusion

Representation of atoms

0001_pro_cg.pdb							X	Y	Z			TYPE
1	ATOM	2	CA	HIS	A	0	17.186	-28.155	-12.495	1.00	26.12	C
2	ATOM	5	CB	HIS	A	0	15.862	-28.669	-13.037	1.00	26.47	C
3	ATOM	12	CA	MET	A	1	16.156	-26.144	-9.429	1.00	28.80	C
4	ATOM	15	CB	MET	A	1	15.469	-24.766	-9.530	1.00	32.87	C
5	ATOM	20	CA	ASN	A	2	15.018	-27.739	-6.188	1.00	22.61	C
6	ATOM	23	CB	ASN	A	2	15.903	-27.912	-4.946	1.00	21.54	C
7	ATOM	28	CA	PRO	A	3	11.654	-26.110	-5.652	1.00	21.30	C
8	ATOM	31	CB	PRO	A	3	10.353	-26.736	-6.196	1.00	20.53	C
9	ATOM	35	CA	ILE	A	4	10.653	-23.899	-2.732	1.00	20.12	C

Atoms contains mainly **three characteristics**:

- XYZ coordinates.
- Atom type
- The molecule this atom belongs.

Type	Polar	Hydrophobic
One-hot encoding	1,0	0,1
	Protein	Ligand
Number	1	-1

Polar atoms in protein: (1,0,1)

Hydrophobic atoms in protein: (0,1,1)

Polar atoms in Ligand: (1,0,-1)

Hydrophobic in Ligand: (0,1,-1)

Then each **Atom contains 6 features**

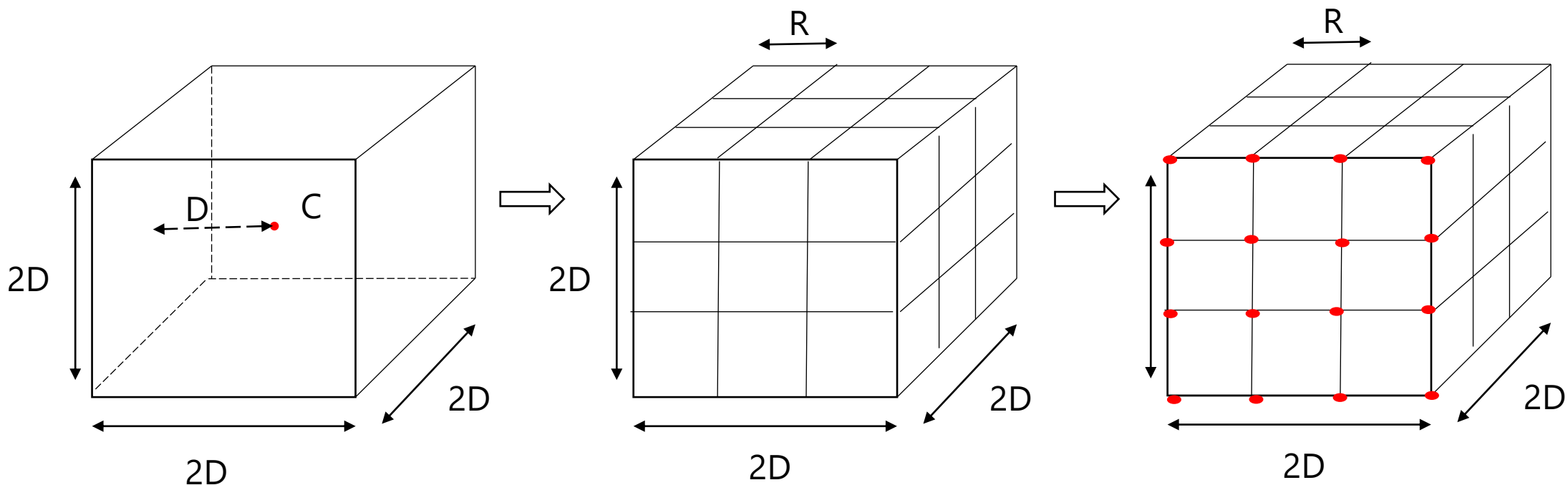
Representation of the space

Use three steps to decide the space:

- 1. the Ligand Center (C)
- 2. the maximum distance (D)
- 3. the resolution (R)

- Input dimension to CNN:

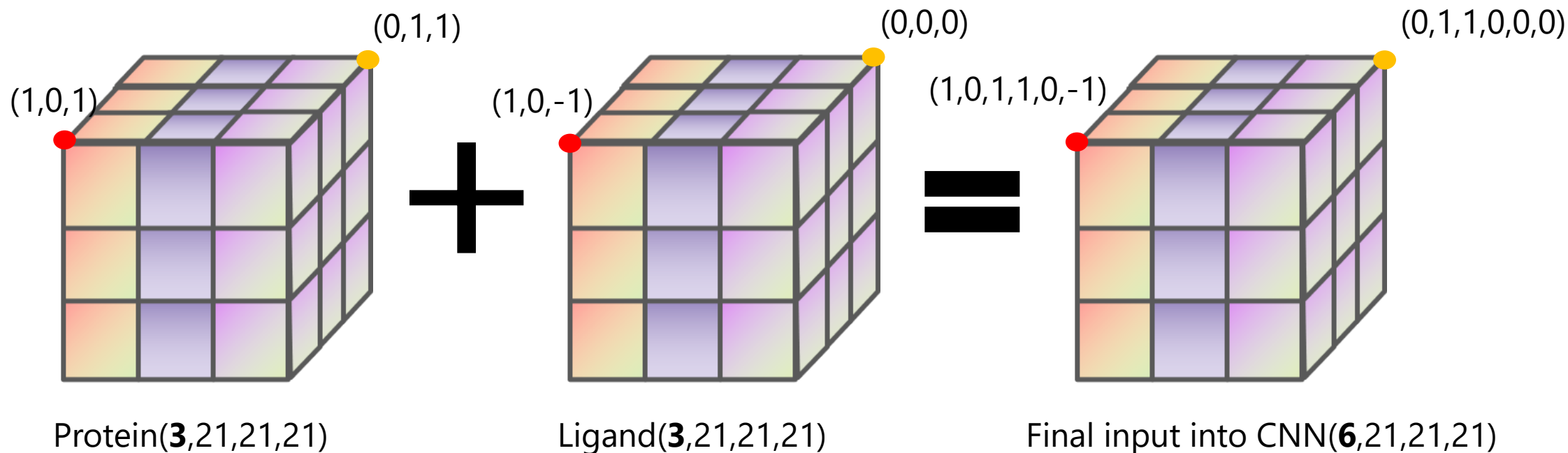
$$Nodes = \frac{2 * D}{R} + 1$$



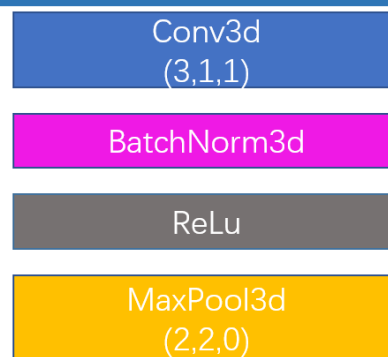
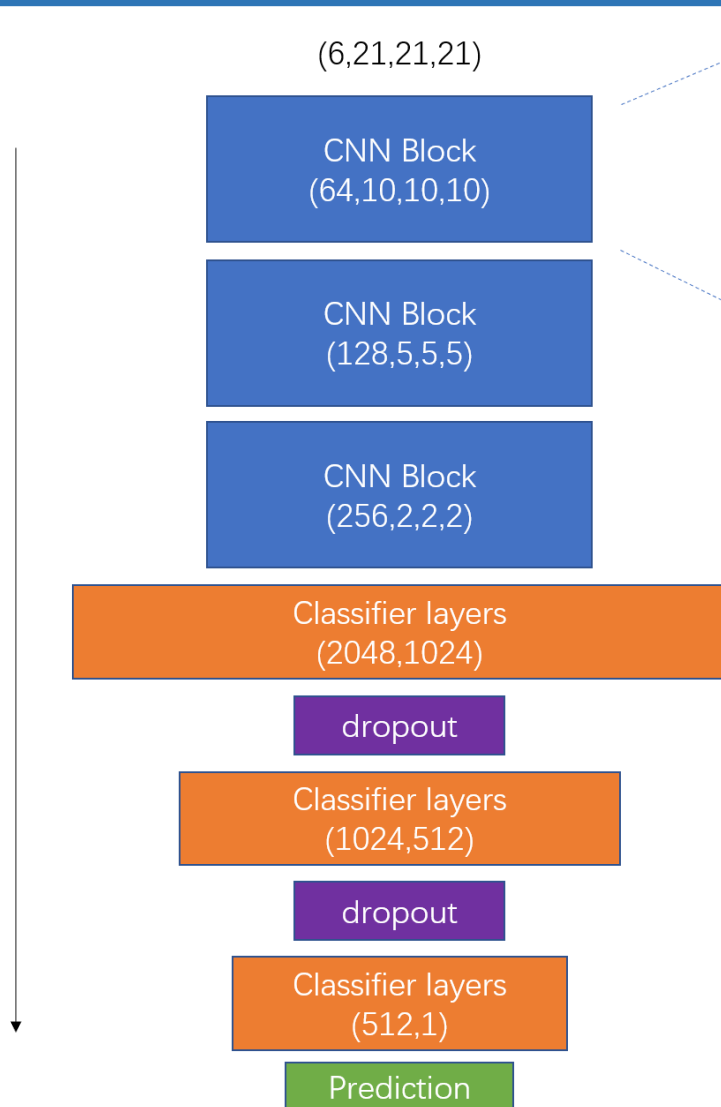
Representation of the proteins and ligand Binding

	Max distance	Resolution	Nodes
number	10	1	21

Round the atom to its **nearest node** in the cube



Loss and models



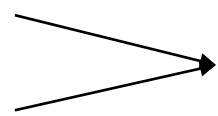
	Kernel size	stride	padding
Conv3d	3	1	1
MaxPooling3d	2	2	0

Loss: nn.BCEWithLogitsLoss()

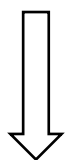
- Binary Cross Entropy Loss+ Sigmoid

Data splitting and training metrics

Protein: 3,000
Ligand: 3,000



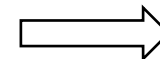
9,000,000 pairs



	Positive Data Pair	Negative Data Pair
Train Dataset	2,400	7,197,600
Validation Dataset	600	1,799,400
Total	3,000	8,997,000

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

Each epoch: Random choose negative pairs



	Positive Data Pair	Negative Data Pair
Train	2,400	2,400
Valid	600	600

Contents



Introduction



Preprocessing



Hyperparameter tuning

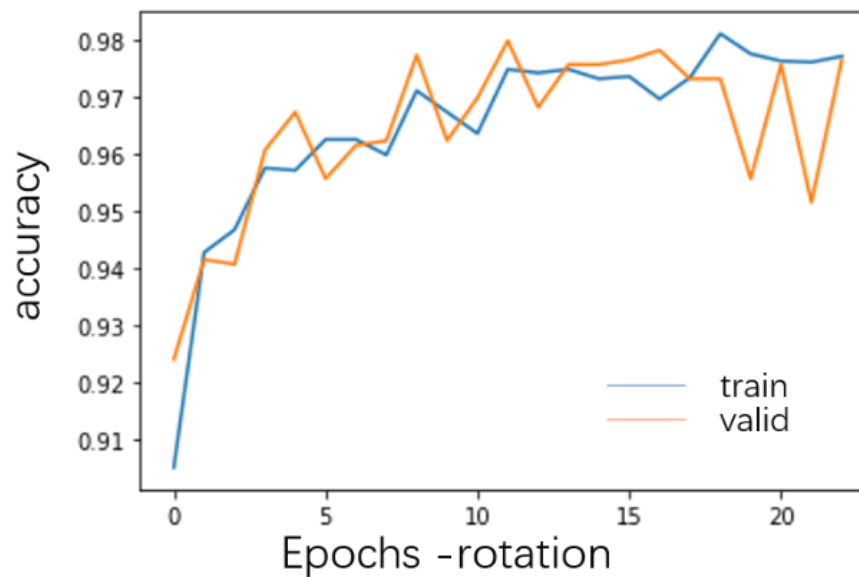
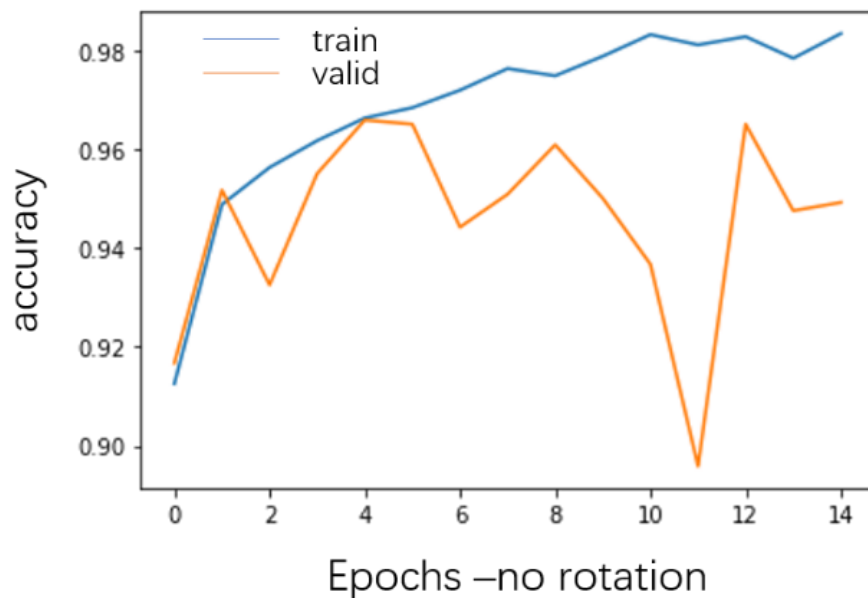


Results



Conclusion

Rotation of training data

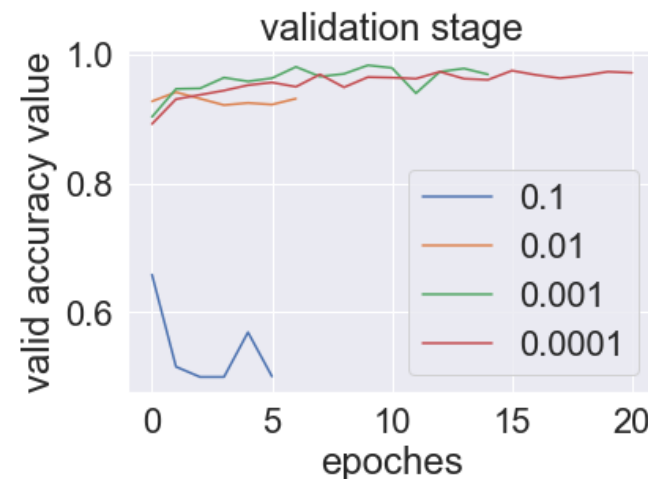
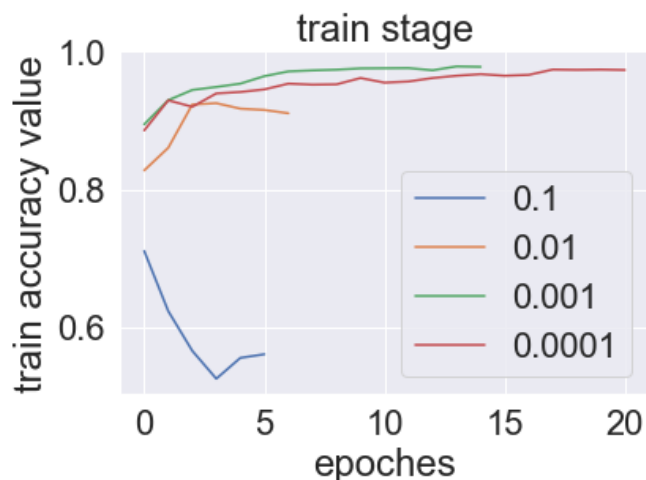


Rotation on training data avoid overfitting

Learning rate, dropout tuning

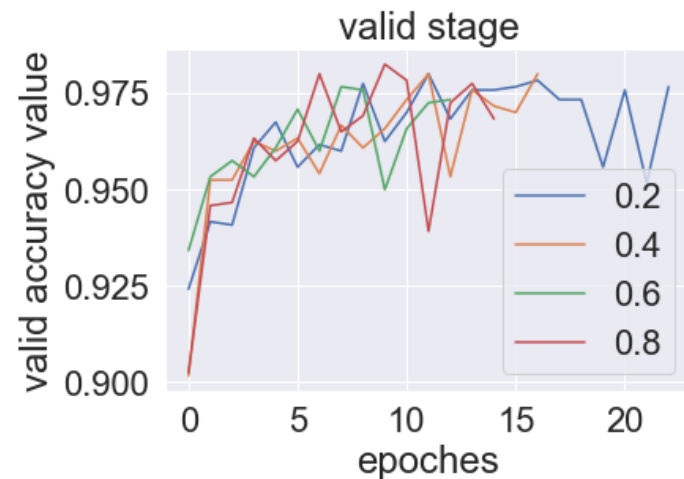
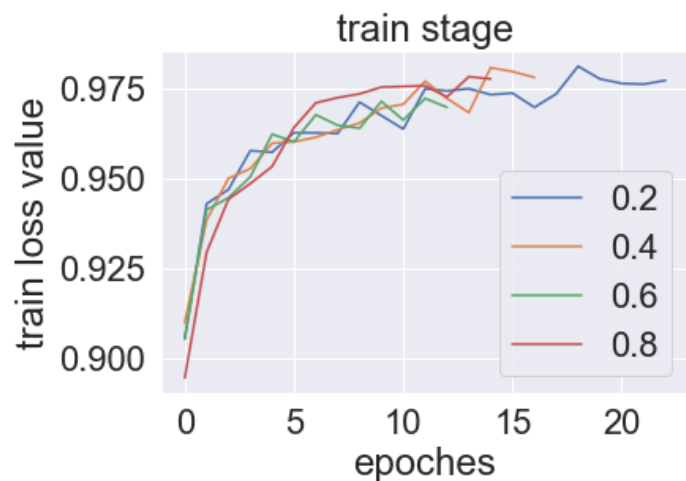
Learning rate of 0.001 got the best performance

Learning rate

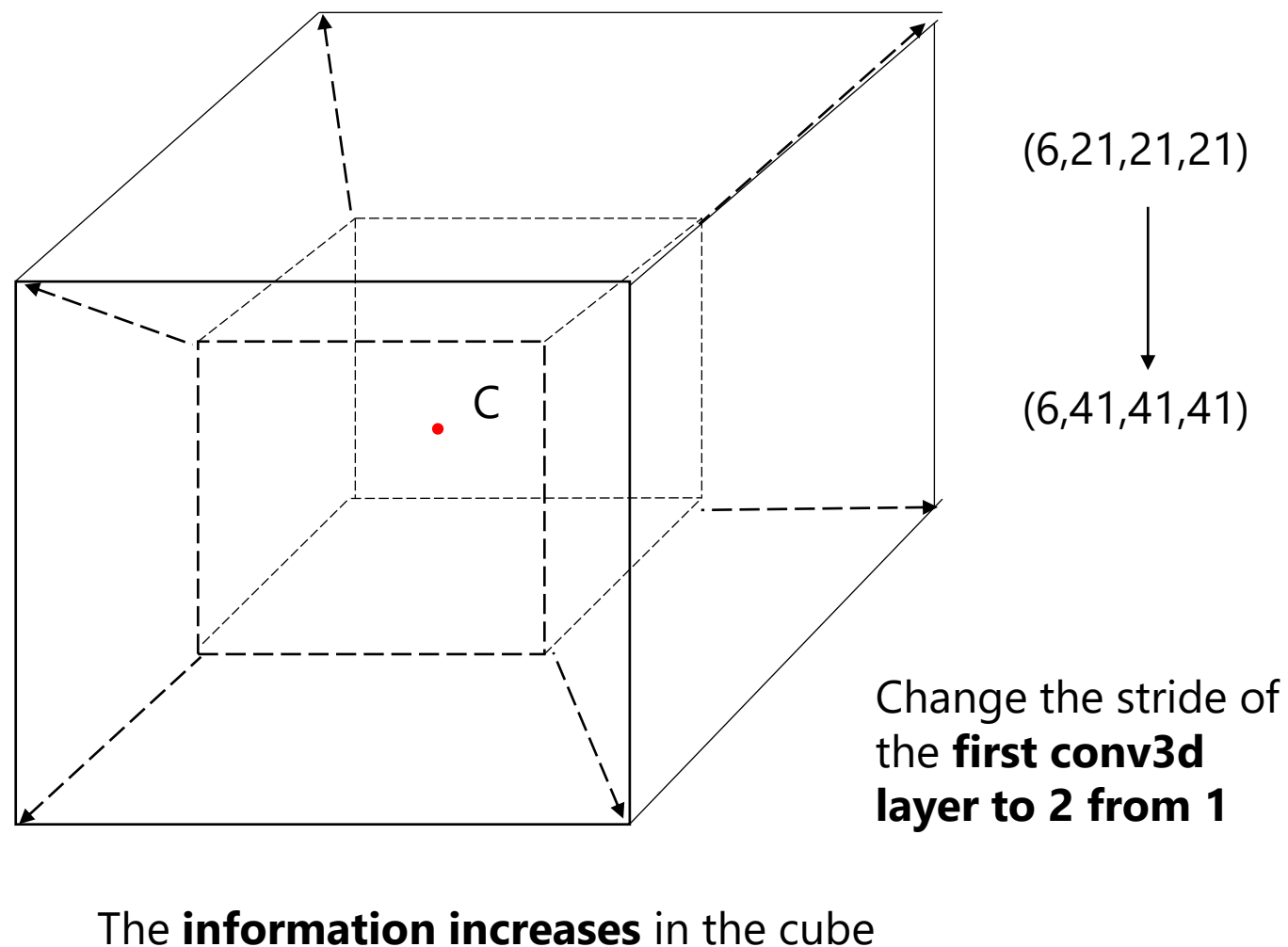
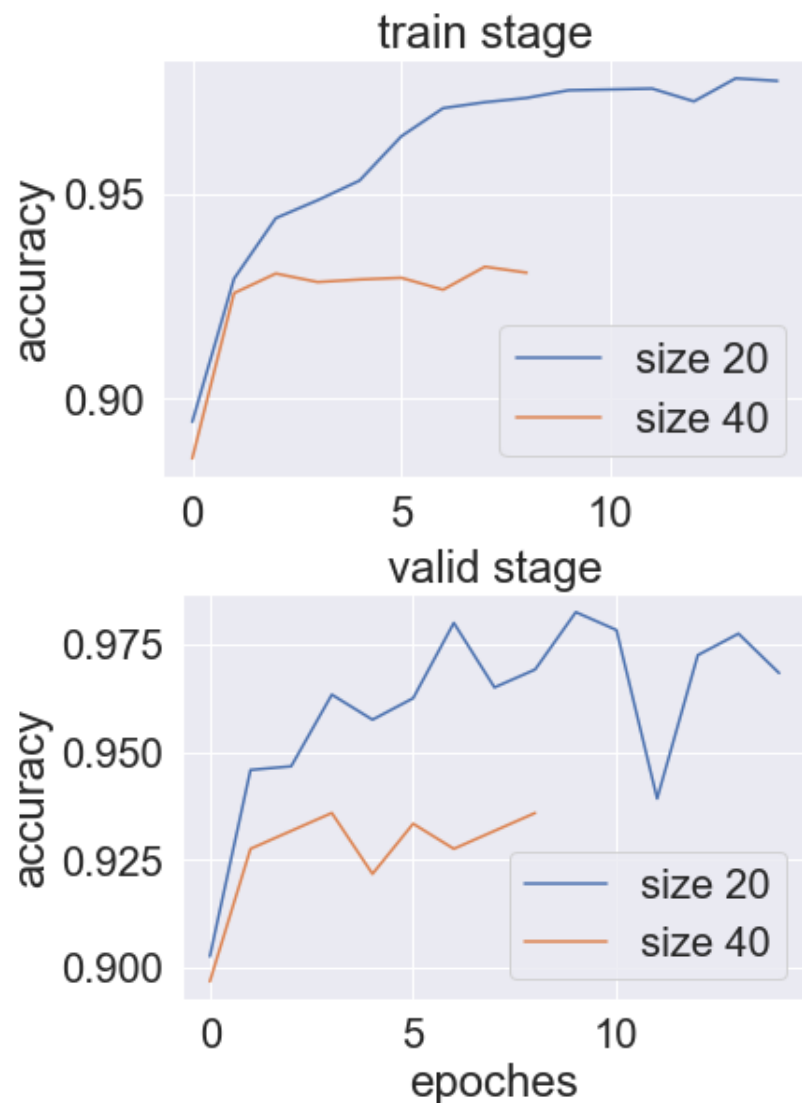


Dropout value of 0.8 can reach the highest faster than other

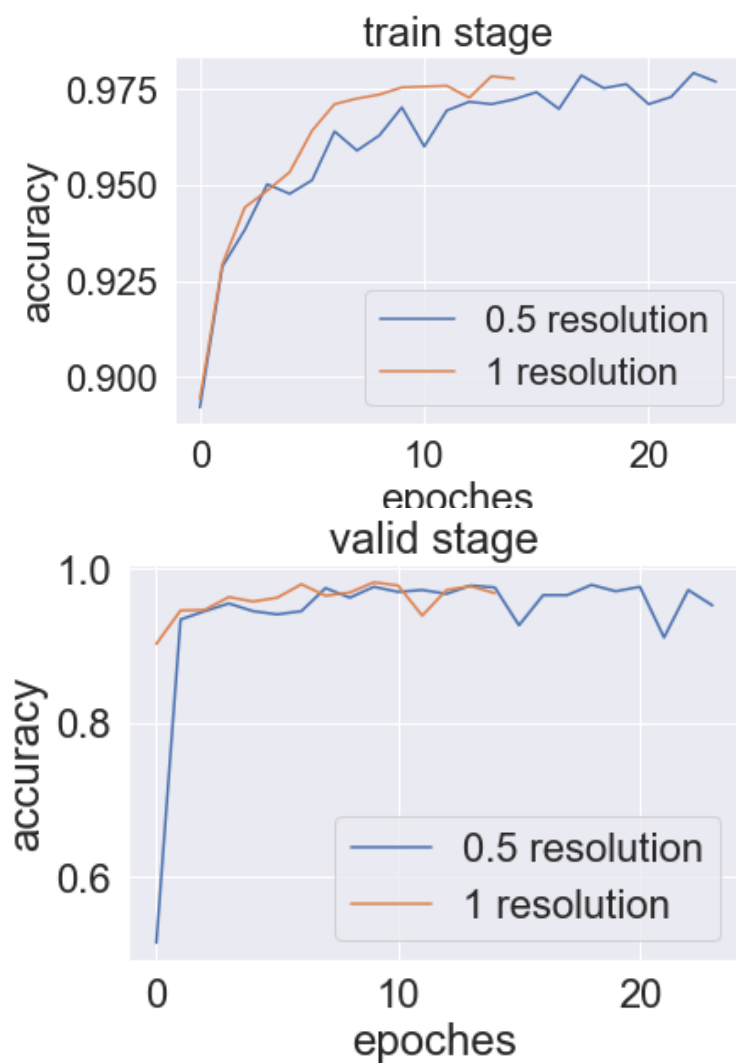
dropout



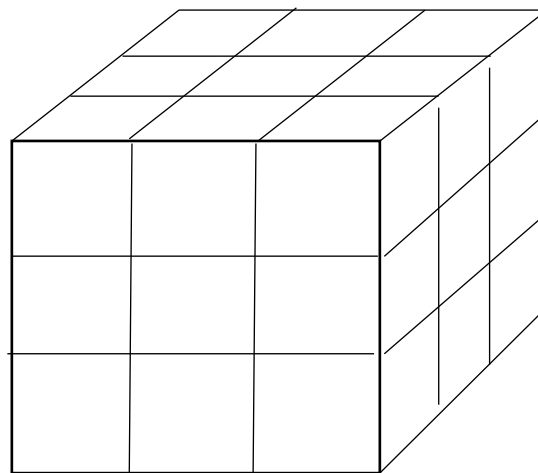
Max distance tuning



Resolution tuning



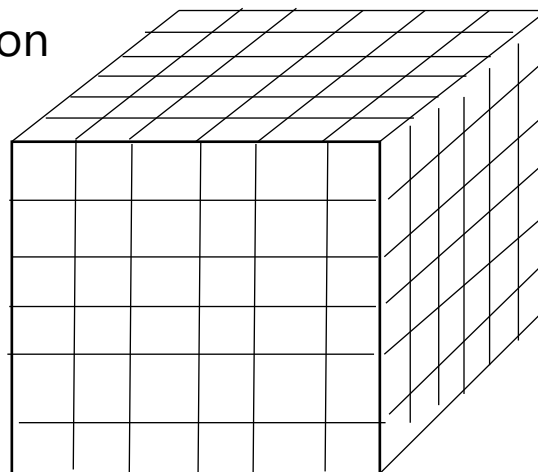
1 resolution



(6,21,21,21)



0.5 resolution



(6,41,41,41)

The **information stays same** in the cube, cube is more precise

Contents



Introduction



Preprocessing



Hyperparameter tuning



Results



Conclusion

Confusion matrix result

Hyper-parameters:

Lr=0.001; max distance=10; resolution=1; dropout = 0.8;

Final accuracy = 98.25%

	Valid Data	test Data
Protein data	600	824
Possible combination	600 X 824	824 X 824
Total	494,400	678,976

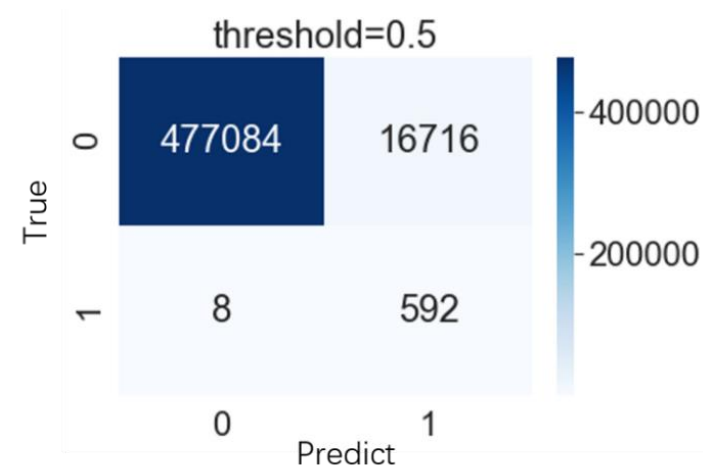
Confusion matrix result

Hyper-parameters:

Lr=0.001; max distance=10; resolution=1; dropout = 0.8;

Final accuracy = 98.25%

	Valid Data	test Data
Protein data	600	824
Possible combination	600 X 824	824 X 824
Total	494,400	678,976



	precision	recall	f1-score	support
0.0	1.00	0.97	0.98	493800
1.0	0.03	0.99	0.07	600
accuracy			0.97	494400
macro avg	0.52	0.98	0.52	494400
weighted avg	1.00	0.97	0.98	494400

Metrics for ranking result

- Whether binding ligand **exists** in top 10 list of target proteins

$$\text{Success rate} = \frac{n_s}{N} \times 100\%$$

- Rank position matters!** Evaluate the ranking of the binding ligand in target proteins

$$nDCG_p = \frac{DCG_p}{IDCG_p}$$

$$DCG_p = \sum_{i=1}^p \frac{2^{rel_i} - 1}{\log_2(i + 1)}$$

$$IDCG_p = \frac{2^1 - 1}{\log_2(1+1)} = 1$$

rel is 1 if the ligand is binding ligand, 0 if the ligand is not

i is the position of the ligand

eval_predictions.txt - 记事本

文件(F) 编辑(E) 格式(O) 查看(V) 帮助(H)

pro_id	lig1_id	lig2_id	lig3_id	lig4_id	lig5_id	lig6_id	lig7_id	lig8_id	lig9_id	lig10_id
1977	1977	2448	1457	1761	2411	1485	8562	2951	2159	1224
2827	2827	1236	1680	1594	1186	0820	1947	1869	1232	0477
1963	1963	2638	1075	1485	2512	0503	1769	1796	1756	0920
0750	0459	0750	1288	2473	0021	0064	1169	1148	2011	2498
1957	1957	0993	1654	2199	1669	1180	2706	0209	1073	0971
2683	2126	2683	2423	1072	1085	1765	0691	2262	0240	2548
2745	2745	0534	1947	0820	1784	2220	2021	1886	1903	1896
2403	2403	2093	0230	2267	1264	2129	1045	1110	0416	2276
2155	2290	0495	2663	2621	1392	1238	2429	2155	0434	2039
2739	2739	0608	0227	1644	1174	2656	1705	0604	1828	0084
1778	1778	1758	1305	1872	0946	2776	2165	2489	2487	2764
2359	2359	1556	0850	1974	1011	2445	1477	0676	0085	2655
0840	0738	0840	1368	2082	2675	2222	1127	1412	2157	0159
2709	0273	2709	1798	1475	1360	2134	1480	1885	2959	1141
2859	1391	1860	0082	1387	1954	2859	2869	0541	0601	2667
0527	2829	0527	0403	0518	0582	0928	0609	1012	1102	0087
2906	1504	1360	2642	2589	1993	0760	2819	1981	2906	0545
2898	0288	2966	0962	2725	2898	1259	2612	2563	2129	1077
0179	2284	0424	0179	2138	1656	2508	1068	1346	1828	0282
0859	0504	0859	0427	1799	1825	1601	1441	2727	0912	1687
0897	0897	1645	2104	2153	1857	2909	0524	0599	0834	1152
0027	1205	1563	0449	0027	0149	1523	2559	0054	0674	0982
1432	1463	2645	1508	1432	0563	1475	2945	1844	1384	0245

Metrics for ranking result

- Whether binding ligand exist in top 10 list of target proteins

$$\text{Success rate} = \frac{n_s}{N} \times 100\%$$

$$nDCG_{10} = \frac{2^1 - 1}{\log_2(1+1)} = 1$$

- Rank position matters! Evaluate the ranking of the binding ligand in target proteins

$$nDCG_p = \frac{DCG_p}{IDCG_p}$$

$$DCG_p = \sum_{i=1}^p \frac{2^{rel_i} - 1}{\log_2(i + 1)}$$

$$IDCG_p = \frac{2^1 - 1}{\log_2(1+1)} = 1$$

eval_predictions.txt - 记事本

文件(F) 编辑(E) 格式(O) 查看(V) 帮助(H)

pro_id	lig1_id	lig2_id	lig3_id	lig4_id	lig5_id	lig6_id	lig7_id	lig8_id	lig9_id	lig10_id
1977	1977	2440	1457	1761	2411	1485	0562	2951	2159	1224
2827	2827	1236	1680	1594	1186	0820	1947	1869	1232	0477
1963	1963	2638	1075	1485	2512	0503	1769	1796	1756	0920
0750	0459	0750	1288	2473	0021	0064	1169	1148	2011	2498
1957	1957	0993	1654	2199	1669	1180	2706	0209	1073	0971
2683	2126	2683	2423	1072	1085	1765	0691	2262	0240	2548
2745	2745	0534	1947	0820	1784	2220	2021	1886	1903	1896
2403	2403	2093	0230	2267	1264	2129	1045	1110	0416	2276
2155	2290	0495	2663	2621	1392	1238	2429	2155	0434	2039
2739	2739	0608	0227	1644	1174	2656	1705	0604	1828	0084
1778	1778	1758	1305	1872	0946	2776	2165	2489	2487	2764
2359	2359	1556	0850	1974	1011	2445	1477	0676	0085	2655
0840	0738	0840	1368	2082	2675	2222	1127	1412	2157	0159
2709	0273	2709	1798	1475	1360	2134	1480	1885	2959	1141
2859	1391	1860	0082	1387	1954	2859	2869	0541	0601	2667
0527	2829	0527	0403	0518	0582	0928	0609	1012	1102	0087
2906	1504	1360	2642	2589	1993	0760	2819	1981	2906	0545
2898	0288	2966	0962	2725	2898	1259	2612	2563	2129	1077
0179	2284	0424	0179	2138	1656	2508	1068	1346	1828	0282
0859	0504	0859	0427	1799	1825	1601	1441	2727	0912	1687
0897	0897	1645	2104	2153	1857	2909	0524	0599	0834	1152
0027	1205	1563	0449	0027	0149	1523	2559	0054	0674	0982
1432	1463	2645	1508	1432	0563	1475	2945	1844	1384	0245

Metrics for ranking result

- Whether binding ligand exist in top 10 list of target proteins

$$\text{Success rate} = \frac{n_s}{N} \times 100\%$$

- Rank position matters! Evaluate the ranking of the binding ligand in target proteins

$$nDCG_p = \frac{DCG_p}{IDCG_p}$$

$$DCG_p = \sum_{i=1}^p \frac{2^{rel_i} - 1}{\log_2(i + 1)}$$

$$IDCG_p = \frac{2^1 - 1}{\log_2(1+1)} = 1$$

$$nDCG_{10} = \frac{2^1 - 1}{\log_2(9 + 1)} = 0.301$$

eval_predictions.txt - 记事本

文件(F) 编辑(E) 格式(O) 查看(V) 帮助(H)

pro_id	lig1_id	lig2_id	lig3_id	lig4_id	lig5_id	lig6_id	lig7_id	lig8_id	lig9_id	lig10_id
1977	1977	2440	1457	1761	2411	1485	0562	2951	2159	1224
2827	2827	1236	1680	1594	1186	0820	1947	1869	1232	0477
1963	1963	2638	1075	1485	2512	0503	1769	1796	1756	0920
0750	0459	0750	1288	2473	0021	0064	1169	1148	2011	2498
1957	1957	0993	1654	2199	1669	1180	2706	0209	1073	0971
2683	2126	2683	2423	1072	1085	1765	0691	2262	0240	2548
2745	2745	0534	1947	0820	1784	2220	2021	1886	1903	1896
2403	2403	2093	0230	2267	1264	2129	1045	1110	0416	2276
2155	2290	0495	2663	2621	1392	1238	2429	2155	0434	2039
2739	2739	0608	0227	1644	1174	2656	1705	0604	1828	0084
1778	1778	1758	1305	1872	0946	2776	2165	2489	2487	2764
2359	2359	1556	0850	1974	1011	2445	1477	0676	0085	2655
0840	0738	0840	1368	2082	2675	2222	1127	1412	2157	0159
2709	0273	2709	1798	1475	1360	2134	1480	1885	2959	1141
2859	1391	1860	0082	1387	1954	2859	2869	0541	0601	2667
0527	2829	0527	0403	0518	0582	0928	0609	1012	1102	0087
2906	1504	1360	2642	2589	1993	0760	2819	1981	2906	0545
2898	0288	2966	0962	2725	2898	1259	2612	2563	2129	1077
0179	2284	0424	0179	2138	1656	2508	1068	1346	1828	0282
0859	0504	0859	0427	1799	1825	1601	1441	2727	0912	1687
0897	0897	1645	2104	2153	1857	2909	0524	0599	0834	1152
0027	1205	1563	0449	0027	0149	1523	2559	0054	0674	0982
1432	1463	2645	1508	1432	0563	1475	2945	1844	1384	0245

Metrics for ranking result

- Whether binding ligand exist in top 10 list of target proteins

$$\text{Success rate} = \frac{n_s}{N} \times 100\%$$

- Rank position matters! Evaluate the ranking of the binding ligand in target proteins

$$nDCG_p = \frac{DCG_p}{IDCG_p}$$

$$DCG_p = \sum_{i=1}^p \frac{2^{rel_i} - 1}{\log_2(i + 1)}$$

$$IDCG_p = \frac{2^1 - 1}{\log_2(1+1)} = 1$$

eval_predictions.txt - 记事本

文件(F) 编辑(E) 格式(O) 查看(V) 帮助(H)

pro_id	lig1_id	lig2_id	lig3_id	lig4_id	lig5_id	lig6_id	lig7_id	lig8_id	lig9_id	lig10_id
1977	1977	2440	1457	1761	2411	1485	0562	2951	2159	1224
2827	2827	1236	1680	1594	1186	0820	1947	1869	1232	0477
1963	1963	2638	1075	1485	2512	0503	1769	1796	1756	0920
0750	0459	0750	1288	2473	0021	0064	1169	1148	2011	2498
1957	1957	0993	1654	2199	1669	1180	2706	0209	1073	0971
2683	2126	2683	2423	1072	1085	1765	0691	2262	0240	2548
2745	2745	0534	1947	0820	1784	2220	2021	1886	1903	1896
2403	2403	2093	0230	2267	1264	2129	1045	1110	0416	2276
2155	2290	0495	2663	2621	1392	1238	2429	2155	0434	2039
2739	2739	0608	0227	1644	1174	2656	1705	0604	1828	0084
1778	1778	1758	1305	1872	0946	2776	2165	2489	2487	2764
2359	2359	1556	0850	1974	1011	2445	1477	0676	0085	2655
0840	0738	0840	1368	2082	2675	2222	1127	1412	2157	0159
2709	0273	2709	1798	1475	1360	2134	1480	1885	2959	1141
2859	1391	1860	0082	1387	1954	2859	2869	0541	0601	2667
0527	2829	0527	0403	0518	0582	0928	0609	1012	1102	0087
2906	1504	1360	2642	2589	1993	0760	2819	1981	2906	0545
2898	0288	2966	0962	2725	2898	1259	2612	2563	2129	1077
0179	2284	0424	0179	2138	1656	2508	1068	1346	1828	0282
0859	0504	0859	0427	1799	1825	1601	1441	2727	0912	1687
0897	0897	1645	2104	2153	1857	2909	0524	0599	0834	1152
0027	1205	1563	0449	0027	0149	1523	2559	0054	0674	0982
1432	1463	2645	1508	1432	0563	1475	2945	1844	1384	0245

Success rate=95.3%, (571 out of 600)
nDCG₁₀=0.776 (between 1st and 2nd)

Average 1st NDCG₁₀=1
 Average 2nd NDCG₁₀=0.631

Contents



Introduction



Preprocessing



Hyperparameter tuning



Results



Conclusion

Conclusion

1

- Dropout probability 0.8, lr 0.001, max distance 10 and resolution 1 could make the model achieve highest acc 98.25%.

2

- The prediction performance of test is expected to be 95.3% SR and 0.776 nDCG, the average ranking of binding ligand is expected to be between 1st and 2nd.

Limitation:

- Did not follow the 'train as test' principal, train as a classification problem, but the test is the ranking problem for top10.
- The structure of model in max distance and resolution tuning changed.

Thank you for your listening!

BS 6207
ADVANCED ARTIFICIAL INTELLIGENCE
FOR BIOMEDICAL DATA SCIENCE

Ni Yuxin