Predicting RMSD or Affinity of Protein-Ligand Complexes Using a Graph Transformer

Haiping Zhang

(SIAT), Shenzhen,China

Hp.zhang@siat.ac.cn

Content

Introduction

Model developing Process

Result

Three models were developed

- **SGraph_RMSD**

for predicting the RMSD of docked protein-ligand complexes

- **Graph_RG**

for predicting affinity when no complex is available, using separate graphs for the pocket and ligand

- **SGraph_affinity**

for predicting affinity based on the given protein-ligand complex interface

Content

Introduction

Model developing Process

Result

SGraph_RMSD building process

Training data

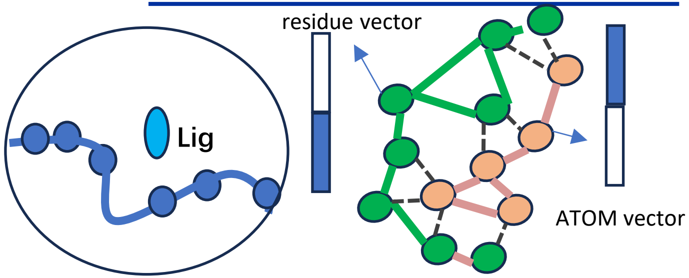
Data processing

PDBbind → Unidock → dataset

docking failures or with RMSD values exceeding 5 Å were excluded

990,000 for training

14,896 for testing.

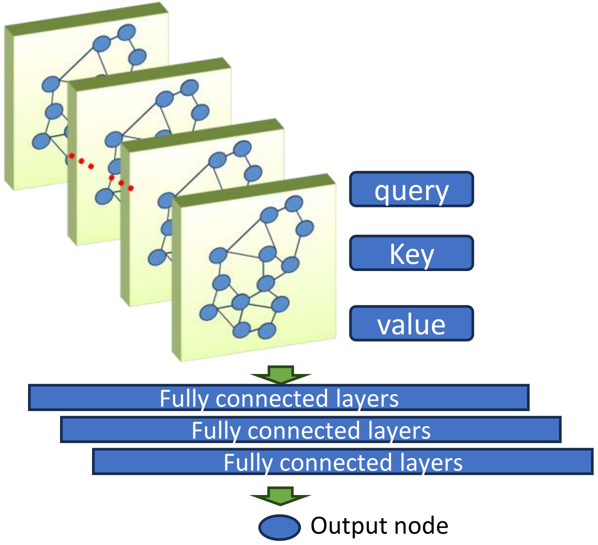


Extract pocket ⇨ Convert into graph ⇨ X

RMSD value ⇨ Y

Model construction

Graph Transformer



Predicting RMSD

Usage

1 Alphafold built model

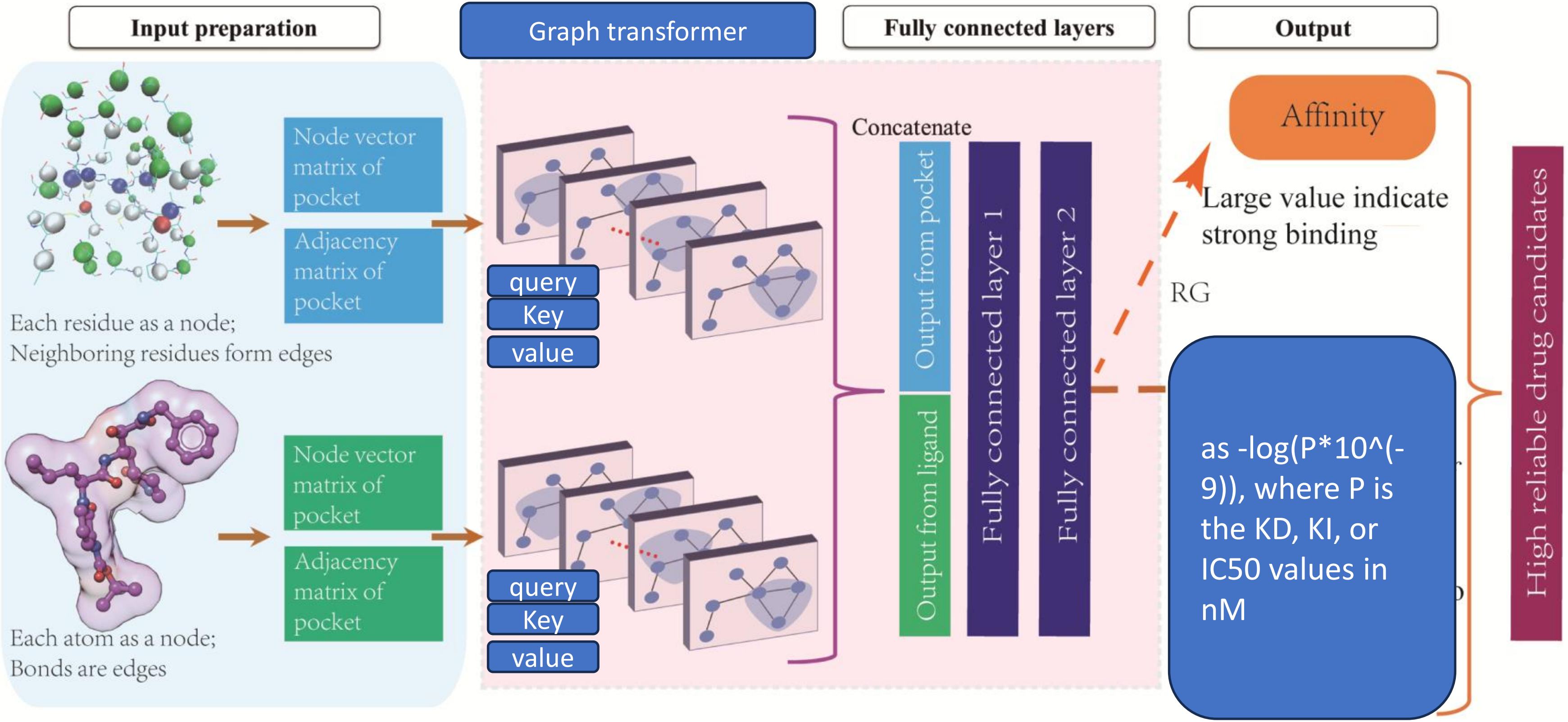
2 Cofactor find pocket

3 “Unidock generate conformations”

4 **SGraph_RMSD select conformation candidates**

5 Complex structure use for other applications

Graph_RG model architecture



SGraph_affinity building process

PDBbind V2020 database	Leave the core set 2016 as test set
Prepare pocket-ligand graph	Same as Sgraph_RMSD
Model training	Model same as Sgraph_RMSD Label as Graph_RG

Content

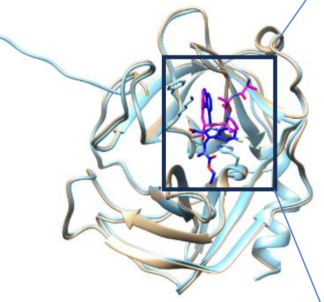
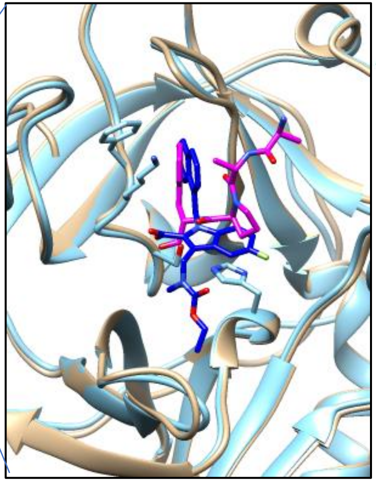
Introduction

Model developing Process

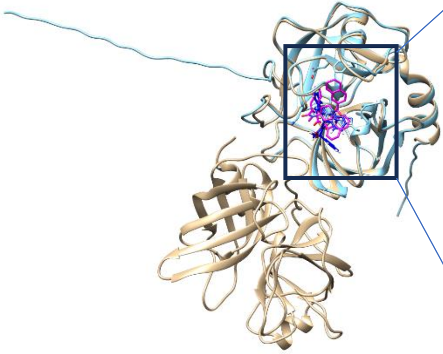
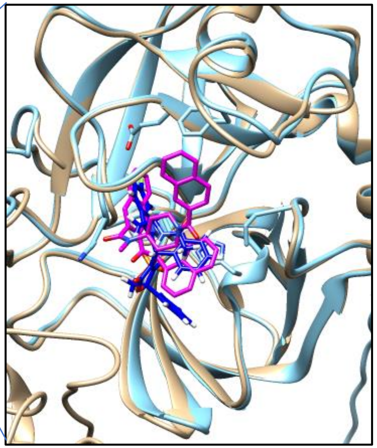
Result

Identify correct pocket by exsiting methods

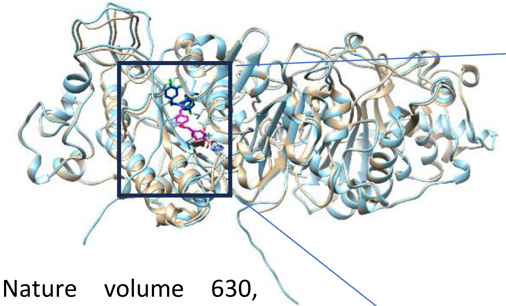
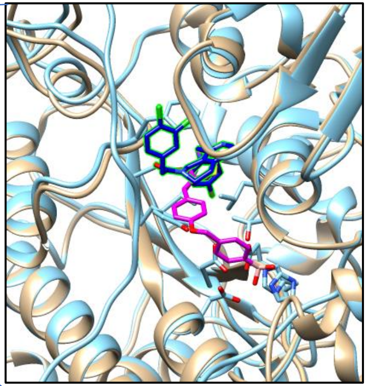
1. Alphafold3 online version (1)
2. Cofactor (2)

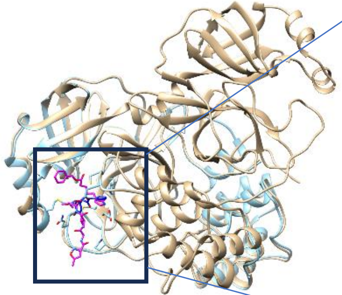
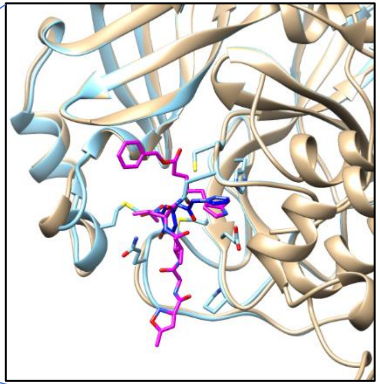
L1001

L2001

L2001

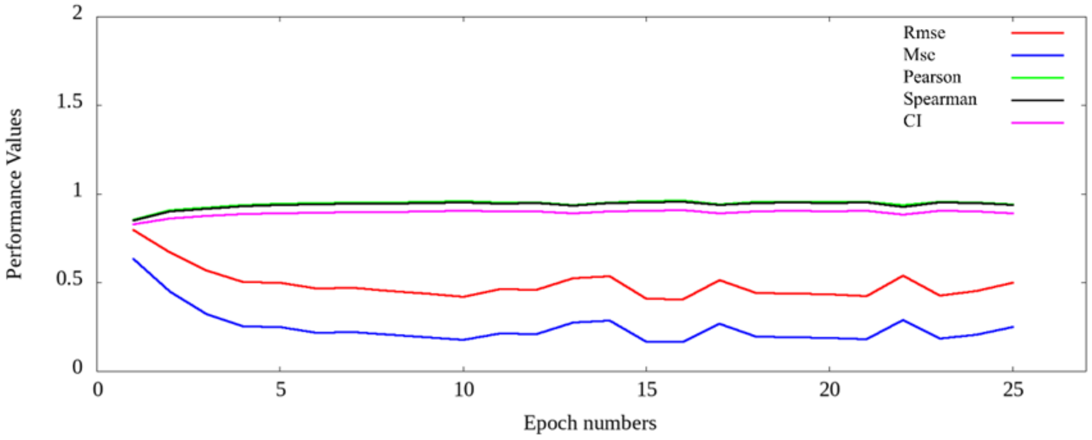



L4001

(1) Nature volume 630, pages493–500 (2024)
(2) Nucleic Acids Research, Volume 45, Issue W1, 3 , 2017

Test Performance during SGraph_RMSD training

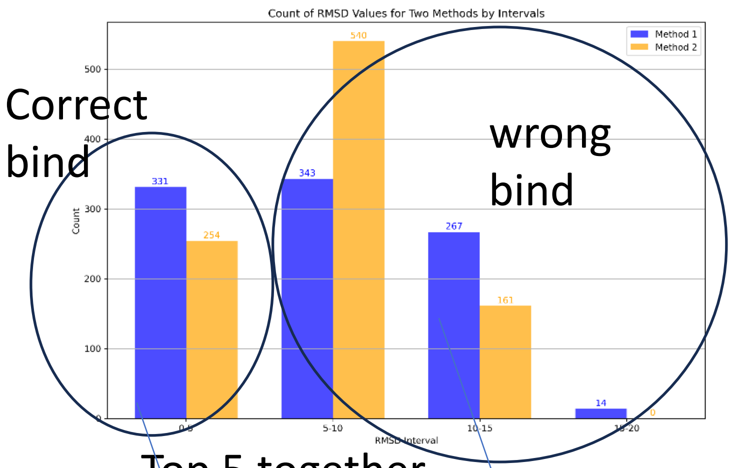
Epoch	Rmse	Mse	Pearson	Spearman	CI
1	0.797	0.636	0.855	0.851	0.828
2	0.671	0.451	0.909	0.903	0.862
3	0.569	0.324	0.923	0.917	0.875
4	0.505	0.255	0.940	0.931	0.886
5	0.499	0.249	0.945	0.937	0.891
6	0.466	0.217	0.950	0.940	0.895
7	0.470	0.221	0.952	0.945	0.899
8	0.455	0.207	0.953	0.946	0.899
9	0.439	0.193	0.956	0.950	0.904
10	0.421	0.177	0.960	0.953	0.906
11	0.462	0.213	0.952	0.946	0.901
12	0.459	0.211	0.954	0.948	0.901
13	0.525	0.276	0.940	0.935	0.890
14	0.537	0.288	0.955	0.948	0.901
15	0.409	0.167	0.961	0.954	0.907
16	0.407	0.165	0.962	0.956	0.910
17	0.516	0.266	0.943	0.938	0.892
18	0.442	0.195	0.956	0.950	0.903
19	0.438	0.192	0.957	0.952	0.906
20	0.436	0.190	0.957	0.950	0.903
21	0.424	0.180	0.958	0.952	0.904
22	0.538	0.290	0.938	0.929	0.883
23	0.429	0.184	0.957	0.952	0.905
24	0.453	0.205	0.954	0.948	0.901
25	0.501	0.251	0.941	0.939	0.892



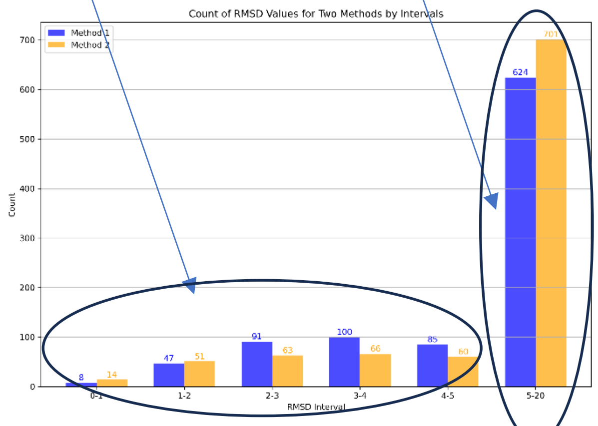
The performance of the **SGraph_RMSD with** different training epochs over the testing set.

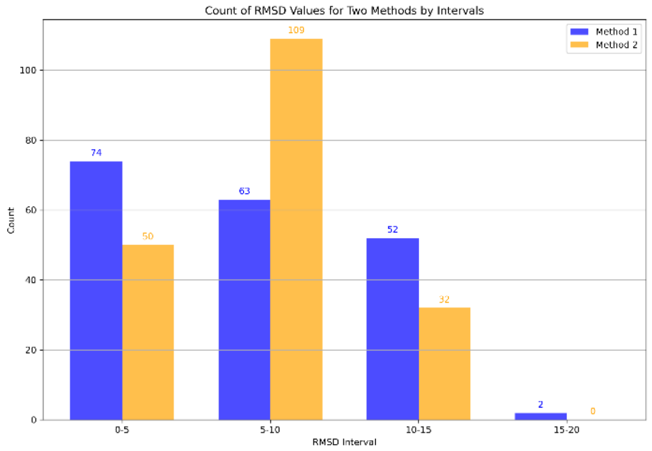
Compare with top predict of UniDock

Stage 1 task



Top 5 together





Top 1

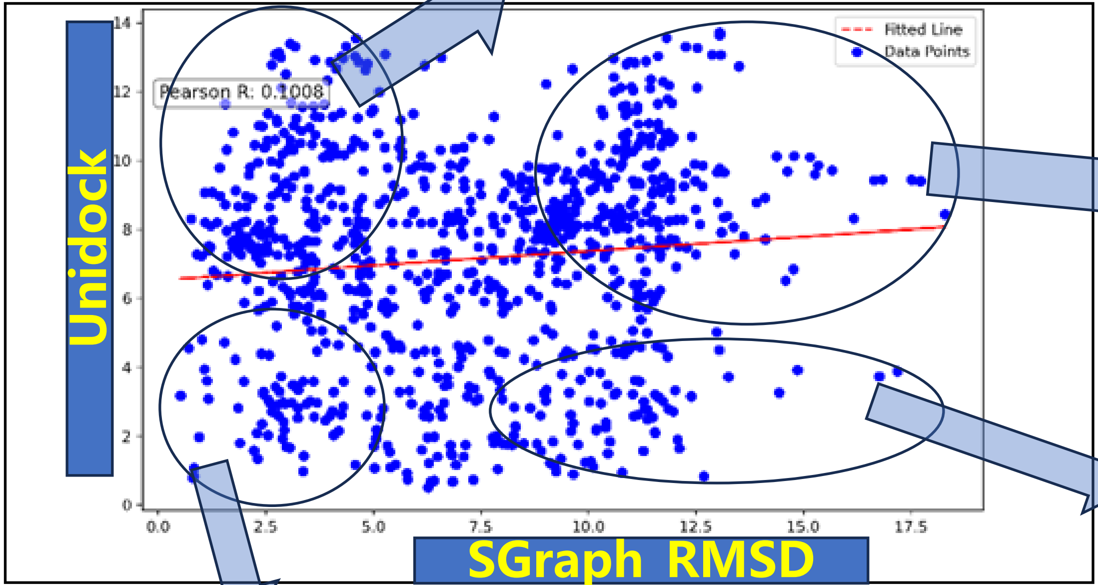
Data for fast evaluate

All data exclude:

1. those native structure will bind to A and B
2. L4 data

Can & Cannot (SGraph_RMSD)

SGraph_RMSD good, Unidock not good



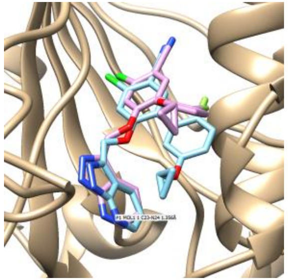
Both methods perform bad

Unidock good, SGraph_RMSD not good

Both methods perform good

Can & Cannot (SGraph_RMSD)

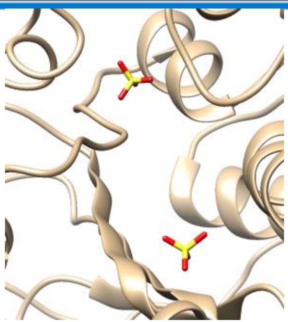
SGraph_RMSD top	original rank	SGraph_RMSD	Vina top	Vina
1	60	1.954	1	8.19
2	134	2.64	2	8.275
3	76	12.589	3	8.147
4	45	7.264	4	8.528
5	42	2.052	5	8.656



L3006_0

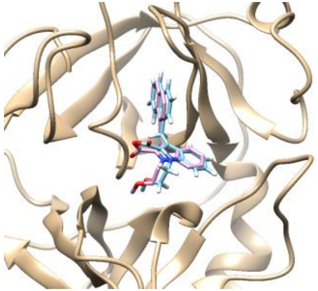
SGraph_RMSD good, Unidock perform not good

SGraph_RMSD top	original rank	SGraph_RMSD	Vina top	Vina
1	168	17.726	1	9.41
2	125	8.446	2	9.529
3	172	16.648	3	9.422
4	117	16.832	4	9.449
5	119	9.652	5	9.383



L3006_1

Both methods perform bad

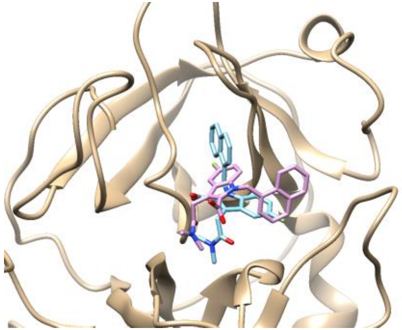


SGraph_RMSD top	original rank	SGraph_RMSD	Vina top	Vina
1	1	0.835	1	0.835

L1013_0

Both methods perform good

SGraph_RMSD top	original rank	SGraph_RMSD	Vina top	Vina
1	153	5.787	1	0.93
2	120	7.893	2	1.046
3	283	7.696	3	3.895
4	74	8.513	4	4.237
5	77	8.224	5	1.751

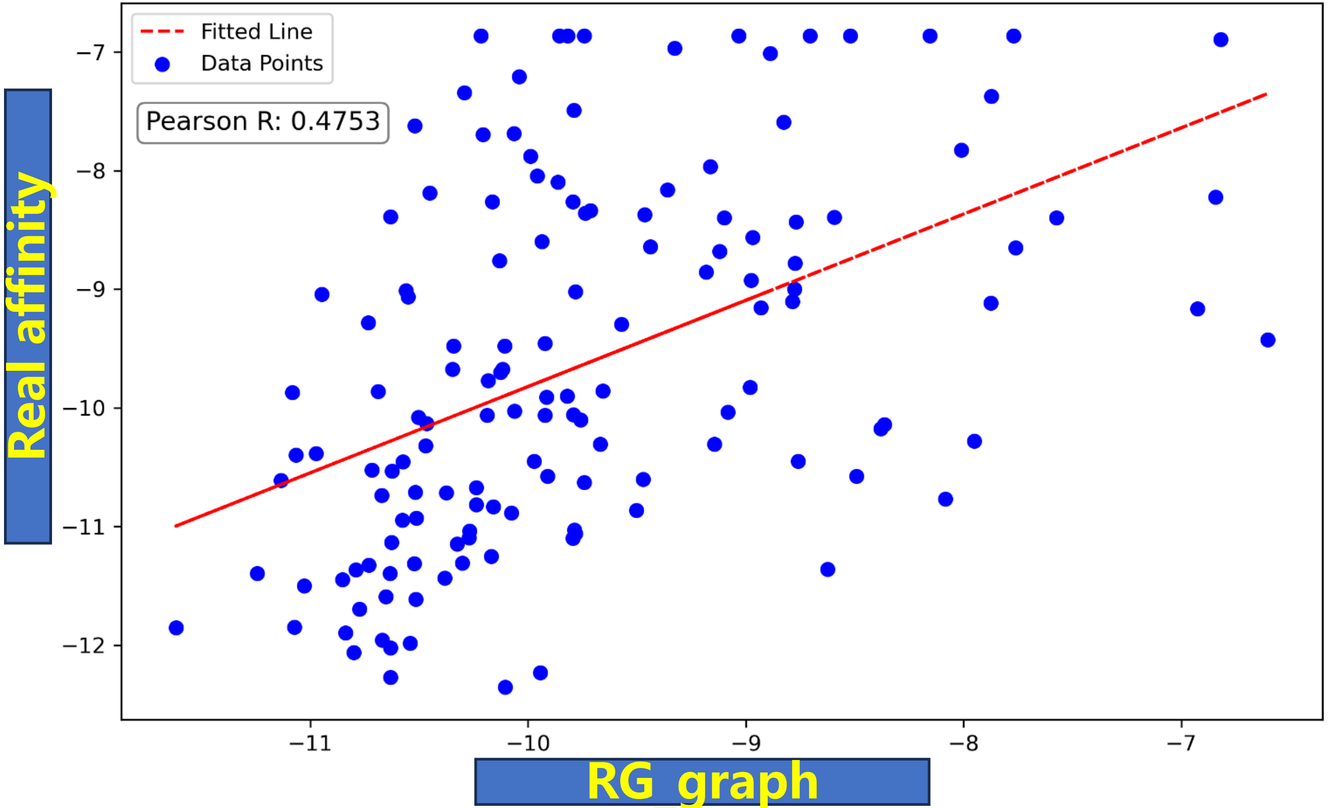


L1011_0

Unidock good, SGraph_RMSD perform not good

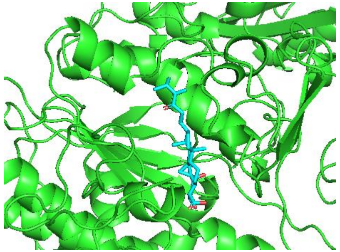
RG_graph prediction result

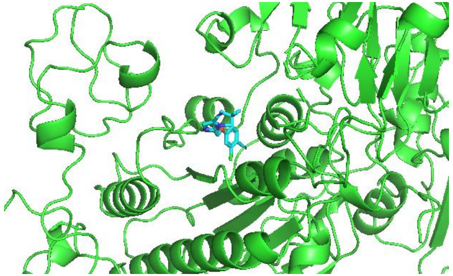
Stage 1 task



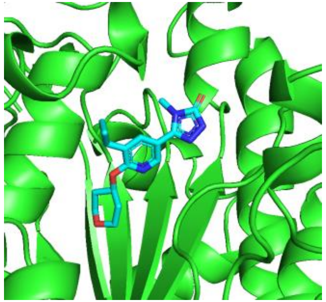
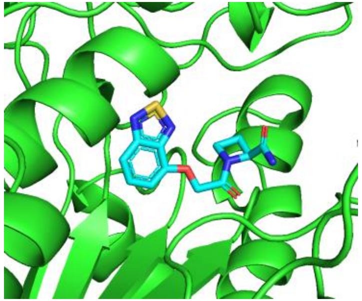
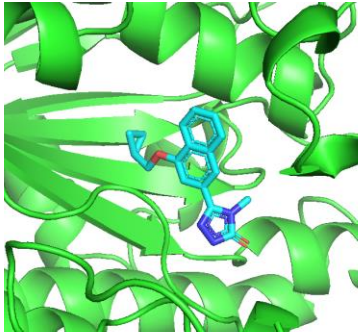
Can & Cannot (Graph_RG)

Differences are smallest in the following three lines:
L3109LG016_1: Your Value = -9.92, Experimental Value = -9.91, Difference = 0.01
L1008LG016_1: Your Value = -8.77, Experimental Value = -8.78, Difference = 0.01
L3038LG016_1: Your Value = -10.06, Experimental Value = -10.03, Difference = 0.04





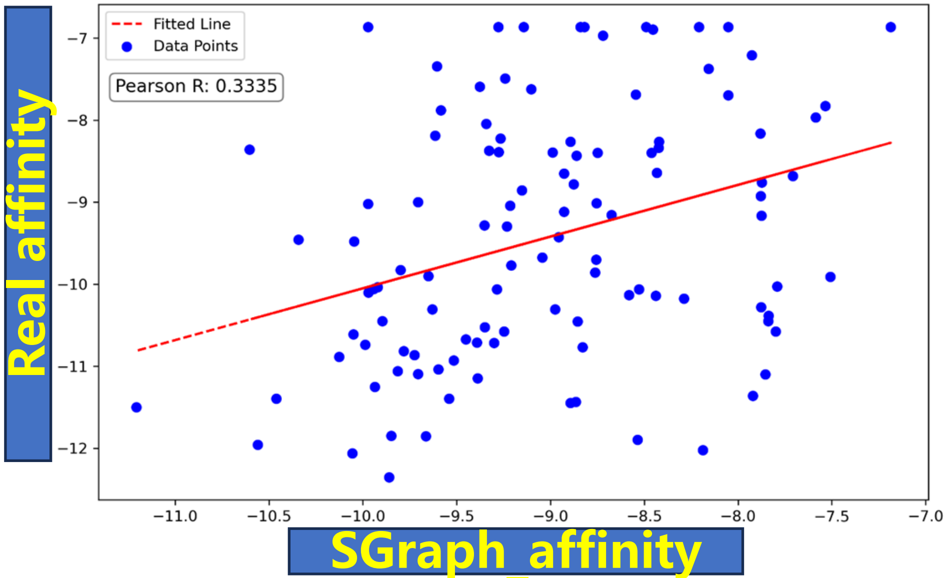
Differences are largest in the following three lines:
L3154LG016_1: Your Value = -9.82, Experimental Value = -6.86, Difference = 2.96
L3056LG016_1: Your Value = -9.86, Experimental Value = -6.86, Difference = 2.99
L3066LG016_1: Your Value = -10.22, Experimental Value = -6.86, Difference = 3.35

SGraph_affinity prediction result

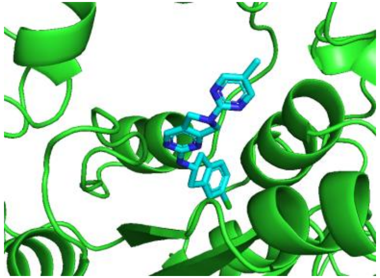
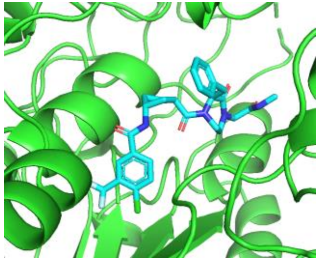
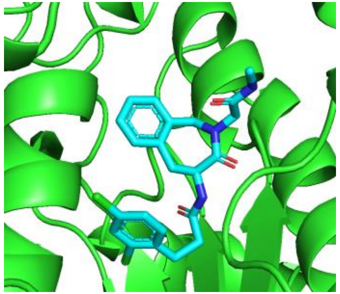
Stage 2 task

Epoch	Rmse	Mse	Pearson	Spearman	CI
11	1.616	2.611	0.710	0.710	0.759
12	1.311	1.720	0.716	0.717	0.763
13	1.396	1.949	0.722	0.720	0.764
14	1.317	1.735	0.723	0.718	0.764
15	1.339	1.793	0.730	0.730	0.769
16	1.547	2.394	0.713	0.714	0.762
17	1.472	2.167	0.735	0.736	0.771
18	1.431	2.049	0.731	0.733	0.770
19	1.275	1.625	0.735	0.734	0.771
20	1.273	1.620	0.742	0.743	0.775
21	1.384	1.914	0.738	0.736	0.773
22	1.392	1.937	0.738	0.738	0.773
23	1.377	1.895	0.740	0.741	0.774
24	1.388	1.926	0.745	0.744	0.776
25	1.268	1.608	0.738	0.737	0.773
26	1.350	1.822	0.737	0.740	0.774
27	1.430	2.045	0.732	0.734	0.771
28	1.503	2.260	0.740	0.739	0.774
29	1.483	2.201	0.738	0.740	0.774
30	1.309	1.714	0.739	0.740	0.774

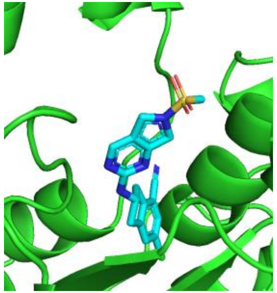
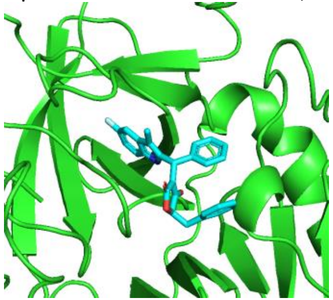
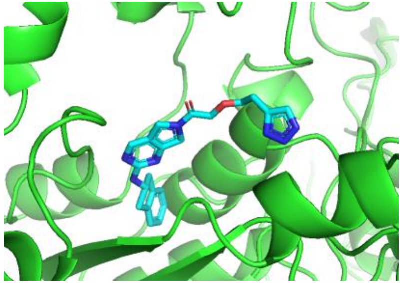


Can & Cannot (SGraph_affinity)

Differences are smallest in the following three lines:
L3130: Your Value = -9.80, Experimental Value = -9.83, Difference = 0.03
L3120: Your Value = -9.23, Experimental Value = -9.29, Difference = 0.06
L3028: Your Value = -8.46, Experimental Value = -8.40, Difference = 0.06

Differences are largest in the following three lines:
L3131: Your Value = -8.54, Experimental Value = -11.90, Difference = 3.36
L1009: Your Value = -7.92, Experimental Value = -11.36, Difference = 3.44
L3047: Your Value = -8.19, Experimental Value = -12.02, Difference = 3.84

Conclusion

1. SGraph_RMSD

1) Deep learning can help to identify more accuracy binding pose compare to tradition method.

2) To highly accurate predict binding pose is still challenge.

2. Graph_RG

1)Only pocket information and ligand information without interface residue-atom pairs information can effectively estimate affinity.

2)It is not perfect but still a current valuable choice in drug screening task.

3. SGraph_affinity

1) Single conformation may not enough to accurately estimate free energy

2) Small data set with complicated input representation and model architecture may lead to overfitting

Acknowledge

Thanks CASP16 organizers provide us such opportunity to check our Methods

John Moult, Gilson Michael, Andriy Kryshtafovych, … …

Thanks SIAT Colleagues

John Z.H. Zhang

Hei Wun Kan

Konda Mani Saravanan

Rongfeng Zou

Many thanks to the developers of the softwares(such as alphafold， Vina, UniDock, Cofactor, etc.) that we applied in this work