BS6207 Project Report

Problem Definition

Structure based ligand discovery is one of the most successful approaches for augmenting the drug discovery process. The object of this project is to predict the protein – ligand structure based on the given exiting structure.

- Protein – ligand interaction

Understanding protein–ligand interactions (PLIs) and is at the core of molecular recognition and has a fundamental role in many scientific areas. PLIs have a broad area of practical applications in drug discovery such as molecular docking, structure-based design, and other type of compounds, clustering of complexes.

In recent years, deep learning has achieved remarkable success in various artificial intelligence research areas. Taking the advantages of deep learning, we could deal with this problem by predicting PLIs interactions and may also address diverse problems in drug discovery.

- Training Dataset

In our dataset, there are 3000 protein-ligand complexes that were determined experimentally with 3D structures available (For now, we are providing 100 of them). Each protein and its ligand are of one-to-one correspondence.

As for the features of these complexes, we are solely interested in coordinates and types of the atoms. Protein and ligand data files contain (x, y, z) coordinates and the type of atoms of each atom of the proteins and ligands. Specifically, hydrophobic ('h') for 'C - Carbon' and polar ('p') for others such as 'O – Oxygen', 'N – Nitrogen', etc.

- Project Task

Training a neural network that takes in as input, the (x, y, z) coordinates of atoms and atom type of a pair of protein and the ligand is required. The trained neural network is to predict if protein and the ligand complexes bind at the output of the network.

For each protein in the test data set, the identified 10 ligands that are predicted to bind the protein are required. Besides, the names of test dataset are randomly assigned. The .csv file specifying which protein random index files match to which of ligand random index files is also needed. Thus the main object.

Highlights

- Biological & chemical background

Before initialling the project, some research in terms of biological and chemical background knowledge are needed. To be more specific, whether the features chosen is biologically contribute to the prediction? Whether this project is biologically reasonable?

a. Biological support for position feature
   Steric complementarity is a great support for the position feature in dataset. Proteins usually bind their biological ligands at a single site on a structured domain. Where a biological function requires a protein to bind several distinct ligands[1]. Based on my understanding, the location of ligands is decided by the biological function. Similarly, the PLIs structure also driven by its function. Thus, the positions of protein and its ligand do relate to structure prediction.

b. Biological support for atom type feature
Complementarity of the chemical nature of the groups on the ligand and protein surfaces provide the support for the atom type feature. It would help to predict which ligands will bind with high affinity to a given protein. A polar group tend to be brought together with polar groups. hydrogen bond donors match acceptors and charge groups on a ligand are frequently neutralized by a nearby, oppositely charged protein sidechain [3].

c. Biological support for project
Besides the biological support mentioned previously, the exiting of binding affinity provides the backbone for the whole project. Binding affinity is a key to the intermolecular interactions driving biological processes, structural biology, and structure-function relationships. Binding affinity is the strength of the binding interaction between a single biomolecule (protein in out project) to its ligand partner [2]. Since such strength existing in every complex, the confidence of reasonable prediction is supported.

- Medical meaning
The project should also have its medical meaning. In reality, almost all pharmaceuticals act by binding to particular proteins, their "targets", interfering with the protein's biological function. With other types of protein target, many drugs are agonists; mimics of the natural biological ligand that binds and activates the protein [1]. Thus, the accurate prediction of PLIs do contribute to the discovery of new drugs, which lays the meaning of this project.

- Data Structure selection

For treatment of features, there are two features in the dataset — protein/ligand and hydrophobic/polar atom type. One way is to put them within one channel with different encodes. Another way is to split them into two channels, one feature in one channel.

Protein	Ligand	Hydrophobic	Polar	One channel	Two channels	
1	0	1	0	1	1	1
1	0	0	1	2	1	-1
0	1	1	0	3	-1	1
0	1	0	1	4	-1	-1

For methods of reshaping cube size to smaller, there are also two methods. One way is simply down-sampling the dataset-discard all atoms outside the standard cube, which leads to many discarded atoms. Another way is to rescale the original cube, which leads to many overlap atoms. E.g. If the sacle ratio is 4, the atom at (80,80,80) in original position will be located at (20,20,20) in the new cube.

	Discard atoms (total/mean)	Overlap atoms (total/mean)	Atoms in cubes (total/mean)
Down sampling	3383684 /1127.8	2772 / 0.92	513000/ 171
Rescale	638597 / 21.28	1000969 / 333.65	2207640/753.88

The result of comparation is shown below. It turns out two channels with down-sample gives the best performance. Thus the data structure here is down-sampled 3000 cubes with two channels.

Feature structure	Reshape method	Validation acc
One channel	Down sampling	47.78
One channel	Rescale	51.45
Two channels	Down sampling	68.75
Two channels	Rescale	51.67

Dataset Pre-processing Description

Since the structure of samples in the dataset is 3 dimensions, the 3D cubes are needed to capture the location information (x, y, z coordinates) in the dataset. Before building the cubes, the standard cube structure should be properly designed based on the properties of this dataset.

- Scale the dataset

  The location of centre of these structures in 3D space is different from complexes to complexes. Thus, the scaling procussure is needed here to unify the position of each cube. The ligand centre is regarded as the centre of whole structures, which should be in (0,0,0). The reason is that the cubes need to be cropped in the further procussure and we don't want to loss ligand information. After scaling, all the complexes are located at the centre of the cubes. (Fig 1)

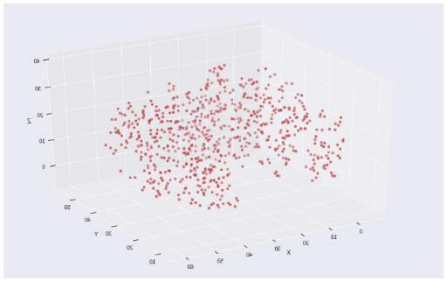
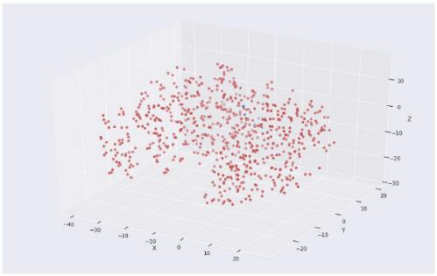


Fig1 a. 3D Structure before scaling



b. 3D Structure after scaling

	x_max	y_max	z_max
count	3000.000000	3000.000000	3000.000000
mean	39.992333	40.369753	42.849592
std	18.789355	19.764223	21.467783
min	8.642429	10.115857	10.379714
25%	27.067038	27.489500	28.384525
50%	34.487250	34.577167	36.072083
75%	47.492777	47.796625	50.304625
max	282.239667	261.038600	271.938000

C. Size of scaled complexes

- Design of the cube size

  After scaling, the complexes size is checked here, the number represent the radius of each complex. (Fig 1c) From the description table, the 75% of these complexes structure size is within 100. Thus, the proper size of cubes should be 100*100*100. However, since the size is too large, we should down sample the dataset. After adjusting cubes size, the proper size should be 25 * 25* 25, which has best space and time efficiency.

Training and testing procedure

- Training procedure

The architecture of model here is a deep convolutional neural network with a single output neuron for prediction. The model consists of two parts: the convolutional and dense parts, with different types of connections between layers (Fig 2).

First, the input is processed by a block of 3D convolutional layers combined with a max pooling layer. The CNN uses 3 convolutional layers with 64, 128 and 256 filters. Each layer has batch normalization and is followed by a max pooling layer. The result of the last convolutional layer is flattened and used as input for a block of dense (fully-connected) layers. The 3 dense layers with 1024, 512 and 256neurons. In order to avoid overfitting, dropout with drop probability of 0.4 was used for all dense layers. There is a SoftMax layer in the last layer to convert the logic value into probability for backward process.

The model tune procedure will be illustrated in the experimental study part.

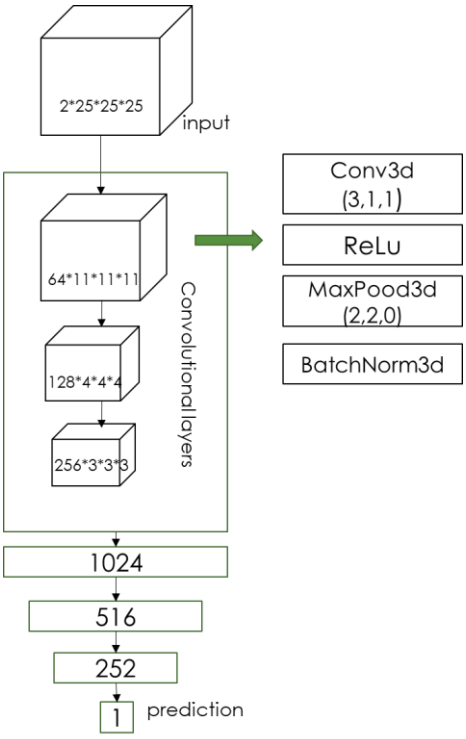


Fig 2. Architecture of model

- Testing procedure

The final result of this project is to recommend 10 candidates of ligands for each protein. To accomplish this, for each protein, the 824 matched cubes are built with 824 ligands. Then the matched cubes are fed in the trained model as batches (batch size = 64, 13 batches in total). After collecting prediction, all the probability are collected. Among them, the top 10 probability with its index are selected with 'torch.topk' method.

Experimental Study

- Cube size and Learning rate

  For model tuning, the first consideration is cube size. Since the size larger than 30 is too time consuming, here I only tried size smaller than 30. It turns out the best size is 25. Besides, different learning rate is tried here, the best one is 0.001.

- Rotaion and early stopping

  Random rotation in the transform compose is used here to improve generalization.The preformance of model is much better after the rotation processure. The early stop method is also applied here. Form the result, the model without early stopping could overfit and thus the accuracy will drop(Fig 3).

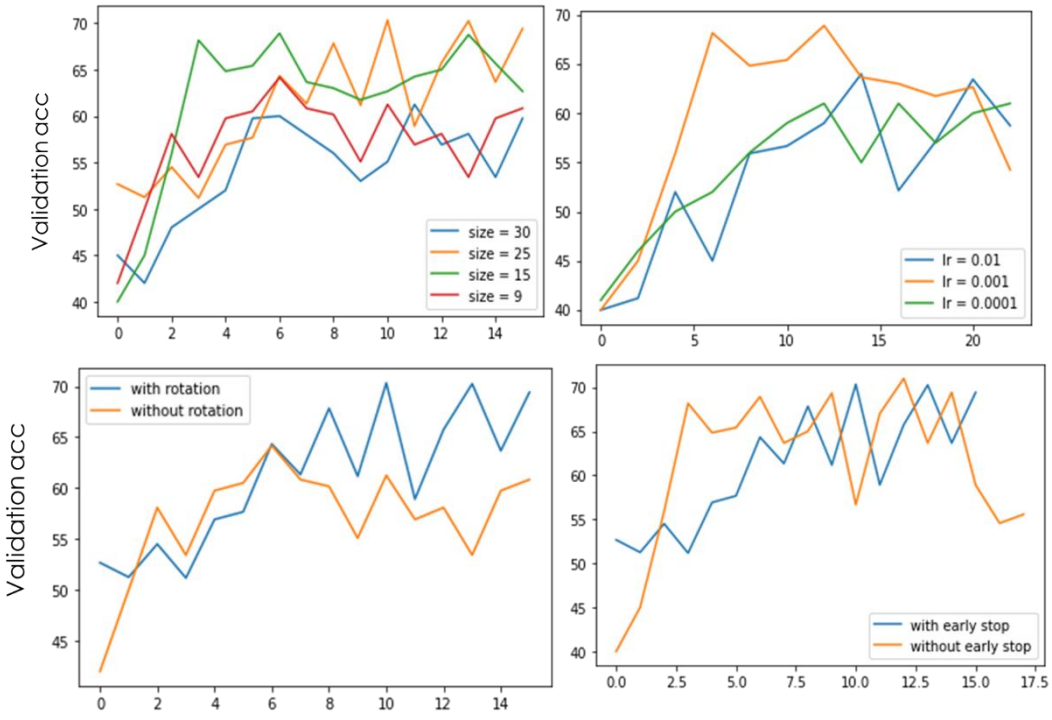


Fig 3 Result of parameter tuning

- Analysis of preformance

		Prediction	
		Negative-Pairing	Positive-Pairing
True label	Negative-Pairing	557	43
	Positive-Pairing	489	111

From the confusion matrix of the model, the prefermance of the model is not promissing. There are false negatives, which cause the lost of real pairing. The true positive is alow, which means the model could not capture the structure of positive pairing.

The reason of this poor perfermance largerly becase I made a mistask of buliding the data structure, which costed too much time. In the end, I have no time fine tune my model and explore other methods. I will continue to improve my model and hope to show a better one in the presentation.

I also find other reason from the biological side. There are lots of other factors may have impact on the PLIs structure. For example, the Ligand-binding sites are often modification by mutation may change the binding site and affect the affinity or specificity for ligands with little impact on the structure of the domain as a whole.

Besides, the precise thermodynamics and kinetics of association are determined not only by the changes in protein–ligand interactions as the protein and ligand come together but also by the changes in protein–water and ligand–water interactions. Thus, the interactions and not complementarity alone which determines ligand binding[3].

Reference

[1] Marta M Stepniewska-Dziubinska, Piotr Zielenkiewicz, Pawel Siedlecki, Development and evaluation of a deep learning model for protein–ligand binding affinity prediction, Bioinformatics, Volume 34, Issue 21, 01 November 2018, Pages 3666–3674, https://doi.org/10.1093/bioinformatics/bty374

[2] KDEEP: Protein–Ligand Absolute Binding Affinity Prediction via 3D-Convolutional Neural Networks José Jiménez, Miha Škalič, Gerard Martínez-Rosell, and Gianni De Fabritiis Journal of Chemical Information and Modeling 2018 58 (2), 287-296 DOI: 10.1021/acs.jcim.7b00650

[3] Mark A. Williams and Tina Daviter (eds.), Protein-Ligand Interactions: Methods and Applications, Methods in Molecular Biology, vol. 1008, DOI 10.1007/978-1-62703-398-5_1, # Springer Science+Business Media New York 2013