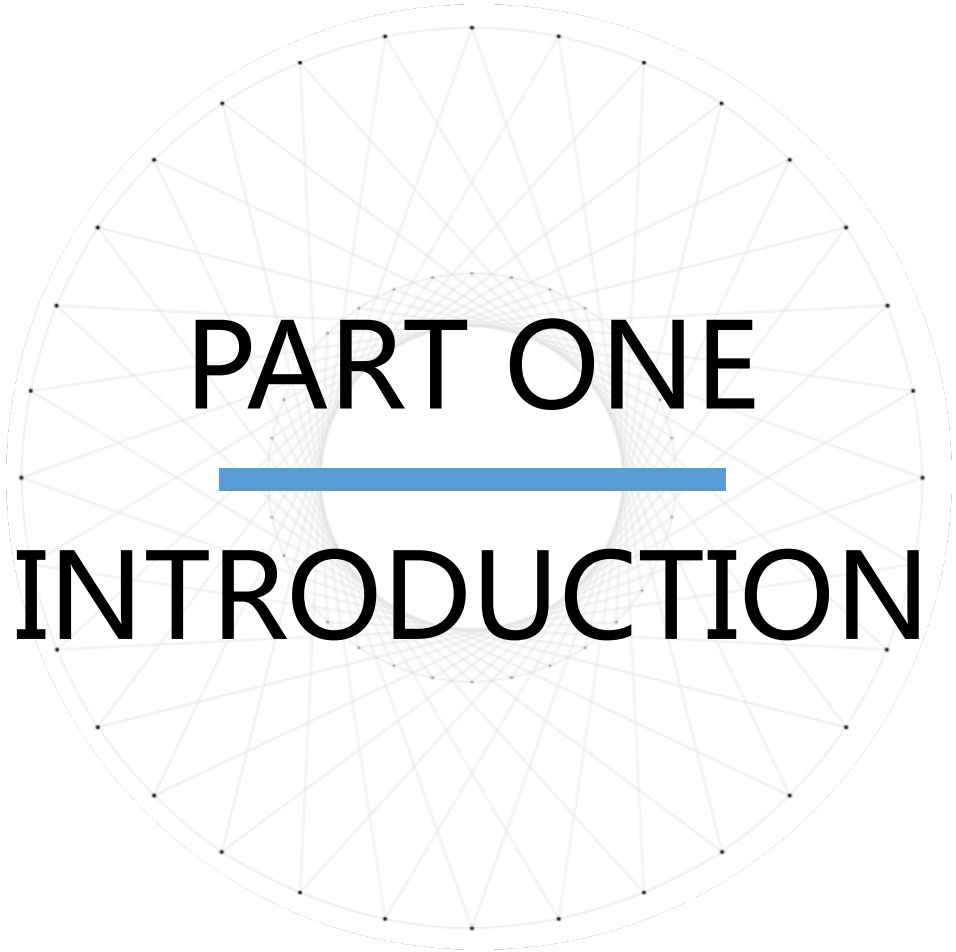
NATIONAL UNIVERSITY OF SINGAPORE

CS5242: Protein-ligand Pairs Prediction With Deep Learning

Lian Yiming
A0174446Y E0210479

Yan Maitong
A0174365Y E0210398

PRESENTED BY TEAM 23



PART ONE

INTRODUCTION

What we have in this project:

1

Mission

Our mission is clear: achieve as high accuracy as we can with reasonable and interpretable models

2

Hypotheses

With no prior knowledge in this project, we proposed two hypotheses to better model the problem.

3

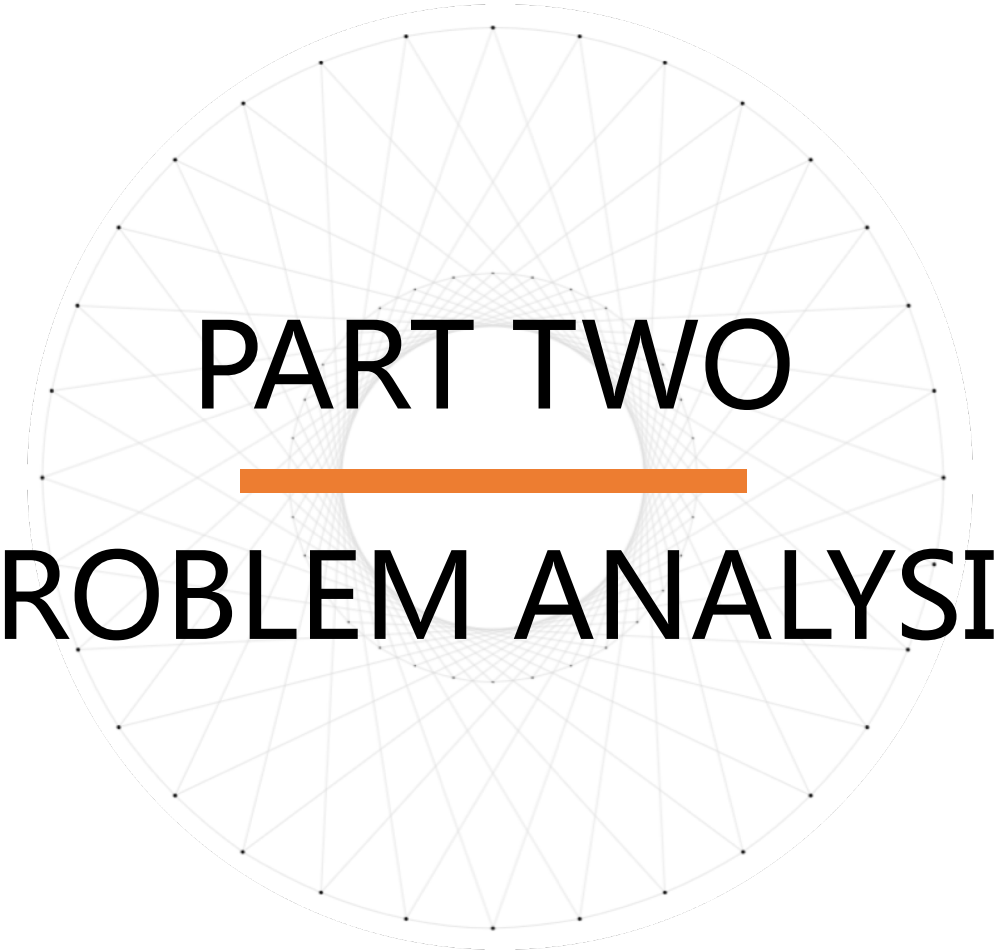
Forms of Representations

Based on previous hypotheses, we design 3 different forms of data representations: KDtree, Compressive Matrix and Octree.

4

Models

Accordingly, we developed four models, including MLP, LSTM, 3D-CNN and O-CNN.



PART TWO

PROBLEM ANALYSIS

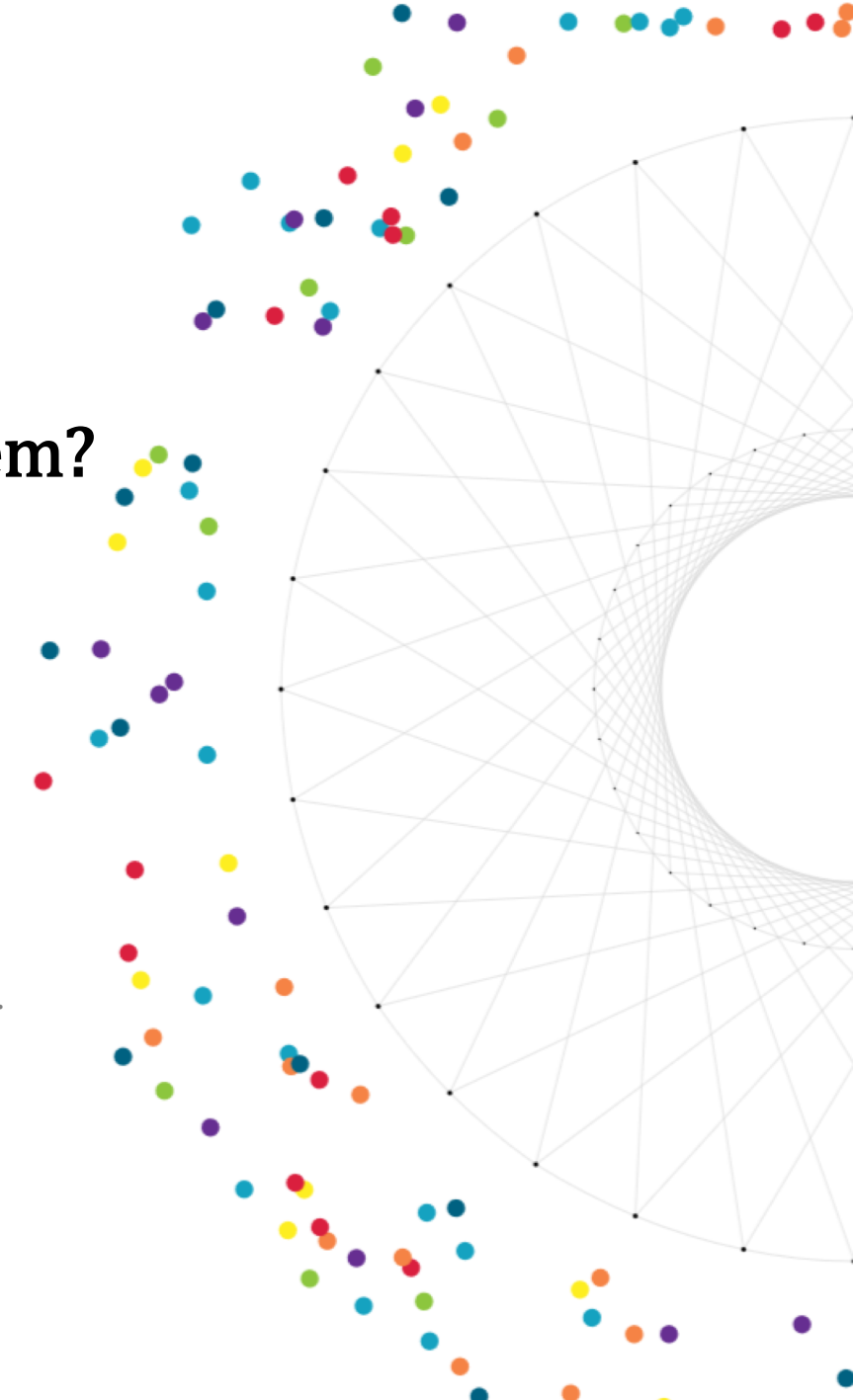
Before modeling ...

Do we have any prior knowledge to this problem?

E.G. does the distance of a protein and a ligand decide whether they can match?

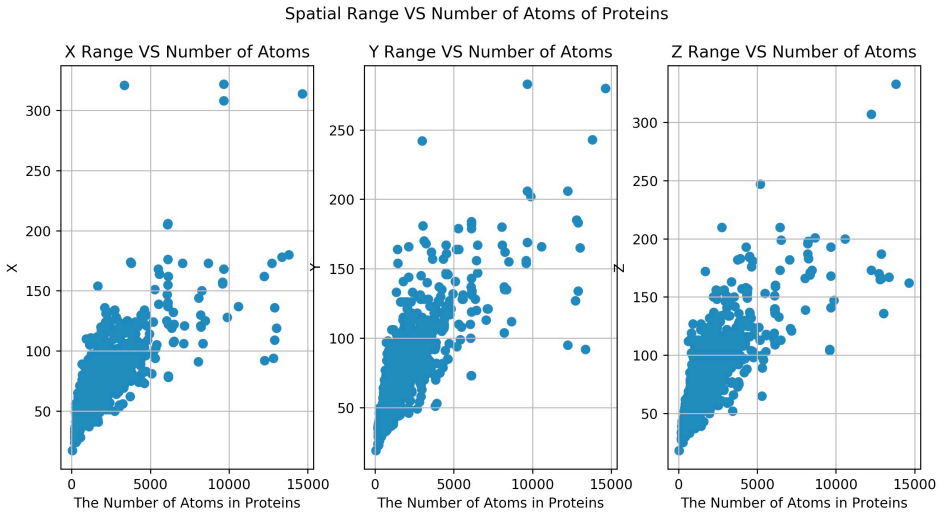
Some one says: proteins and ligands with small distances may match with each other; and for those with large distances, they are less likely to match.

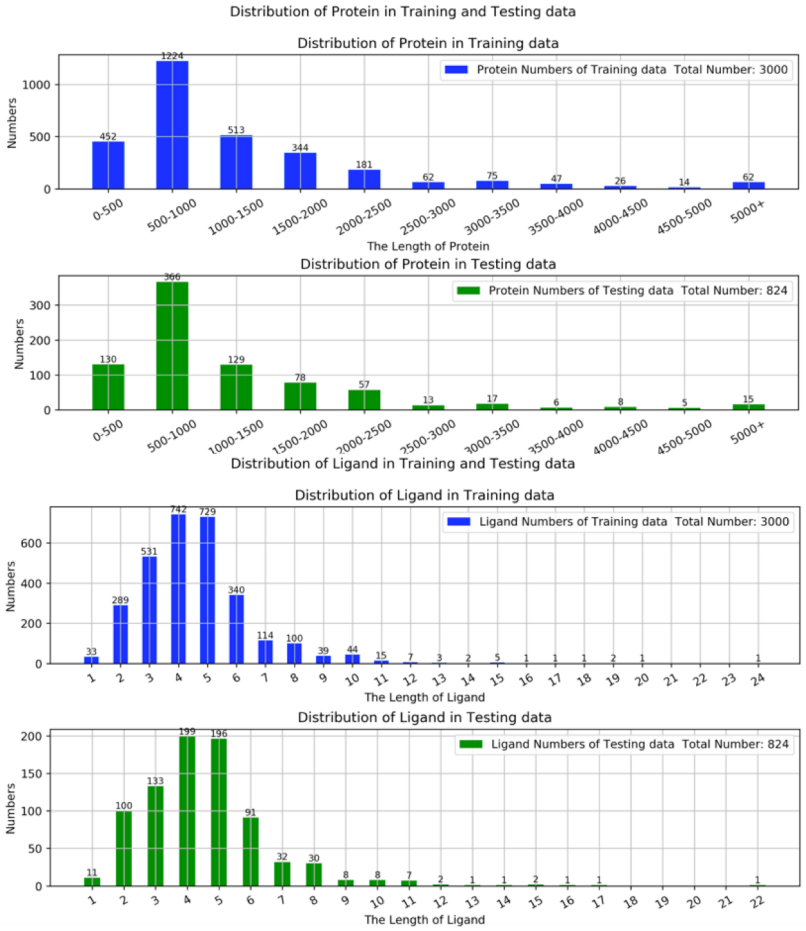
Actually, there's no such prior knowledge, and it's pretty common in the real world.



Exploration of Raw Data

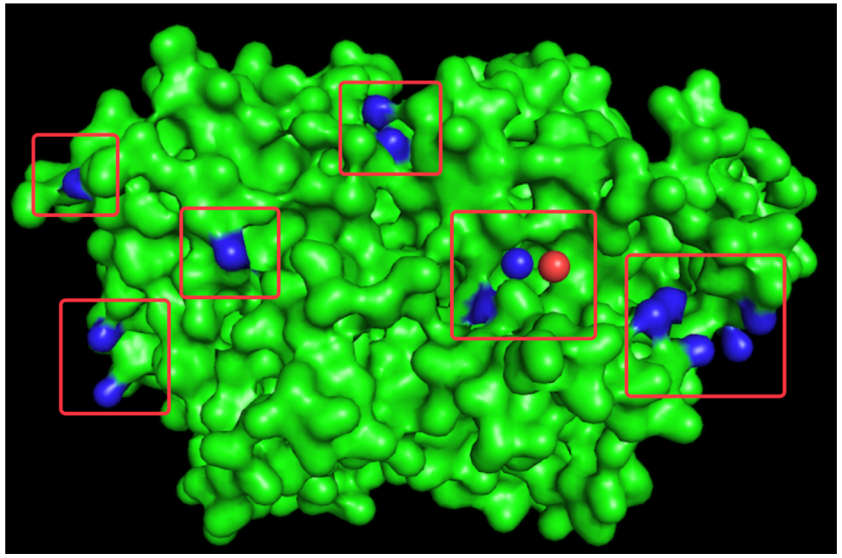
- Most of the spatial ranges (X, Y, Z) of a protein is less than **150 units with 2 digits**;
- Most of the number of atoms within a ligand is less than **10 units**;
- Most of the proteins consist of more than **500** atoms.



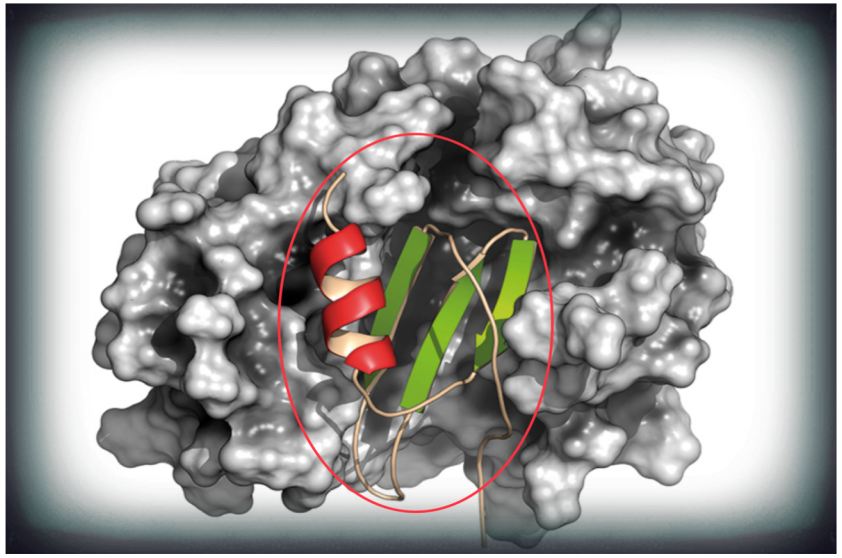


We proposed two hypotheses

- The binding relation is determined by the **local features** (i.e. distance) of atoms.
- The binding relation is determined by the **global features** (i.e. shape) of ligands.



Local perspective: for each P-L pair, we only capture the neighbor protein atoms around a ligand atom.



Global perspective: for each P-L pair, we reserve as more atoms data as we can to capture the global info

PART THREE

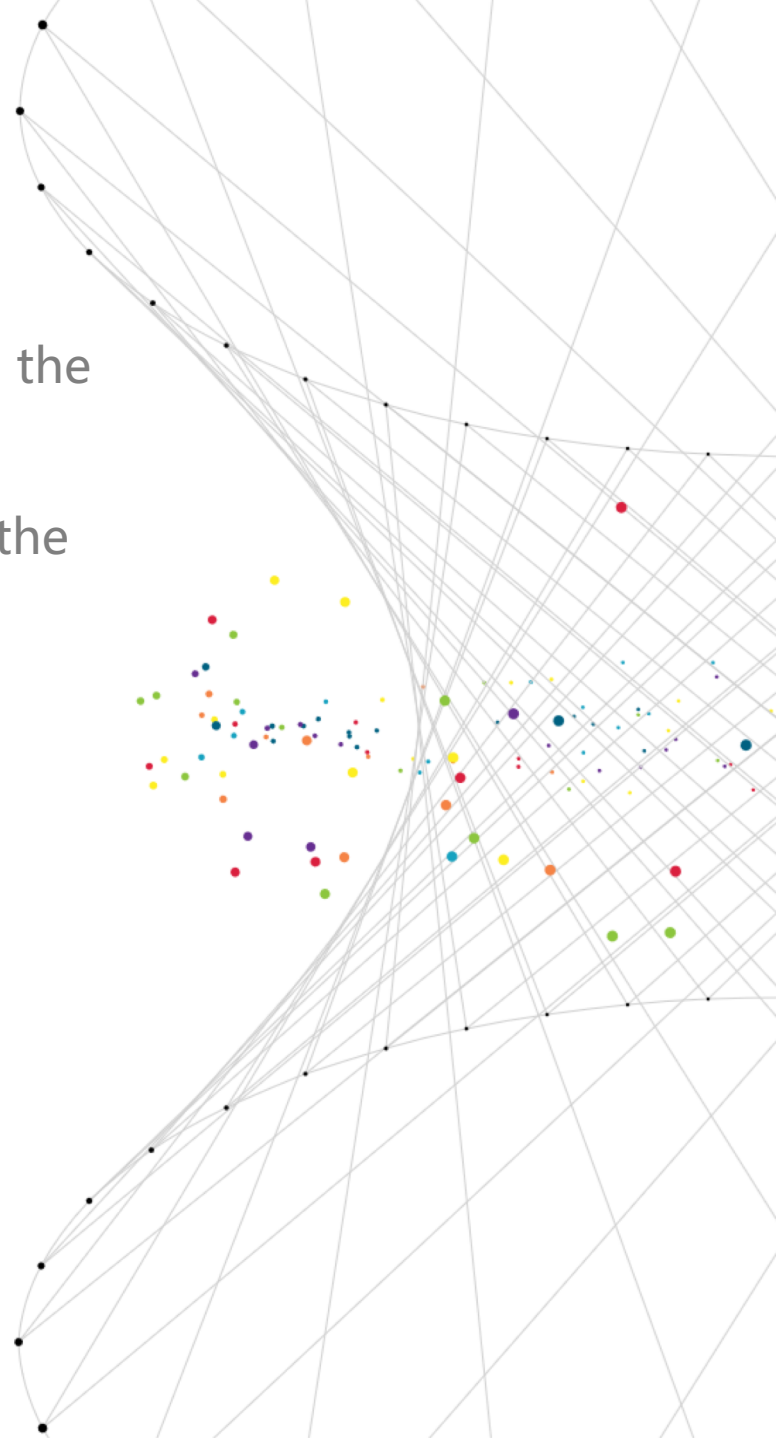
DATA PREPROCESSING

Step 1: Centroid Normalization

- We cannot assume all pairs of proteins and ligands are measured in the same coordinate system.
- Normalize the proteins first, and use proteins' centroids to adjust the coordinates of ligands.

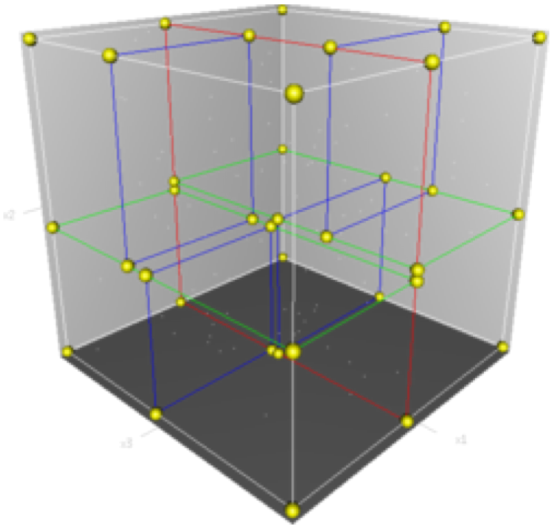
Step 2: Hypothesis-based Data Representation

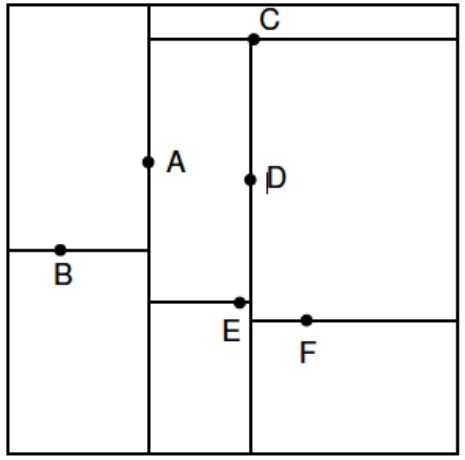
- Local features -> KDtree -> MLP, LSTM
- Global features -> Compressive Matric, Octree -> 3D-CNN, O-CNN



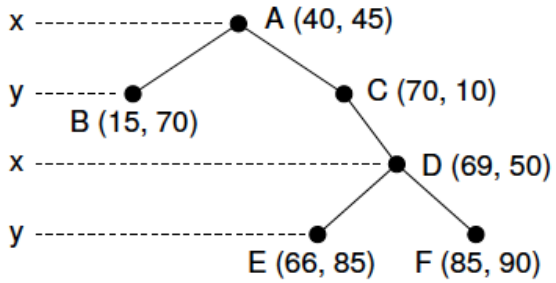
KDtree for Local Hypothesis:

- K-dimensional tree. A space-partitioning structure for point clouds.
- For one ligand, up to 10 atoms would be taken into account.
- For each atom in one ligand, find 3 nearest protein atoms when the ligand is paired to one protein. (Maximum 160 atoms for one pair)
- Reduce complexity from O(m*n) to O(n).





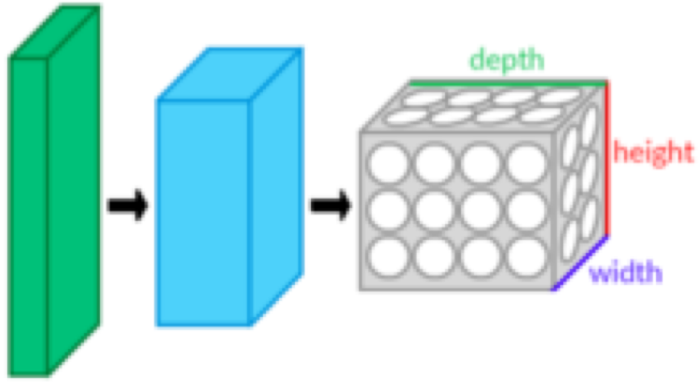
(a)



(b)

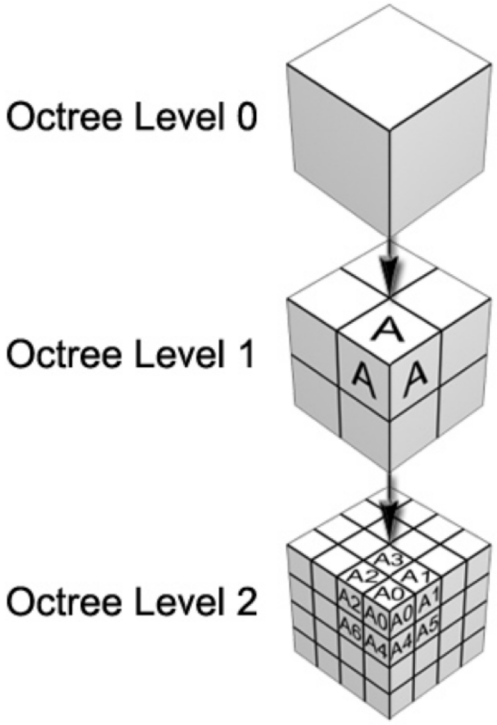
Compressive Matrix for Global Hypothesis:

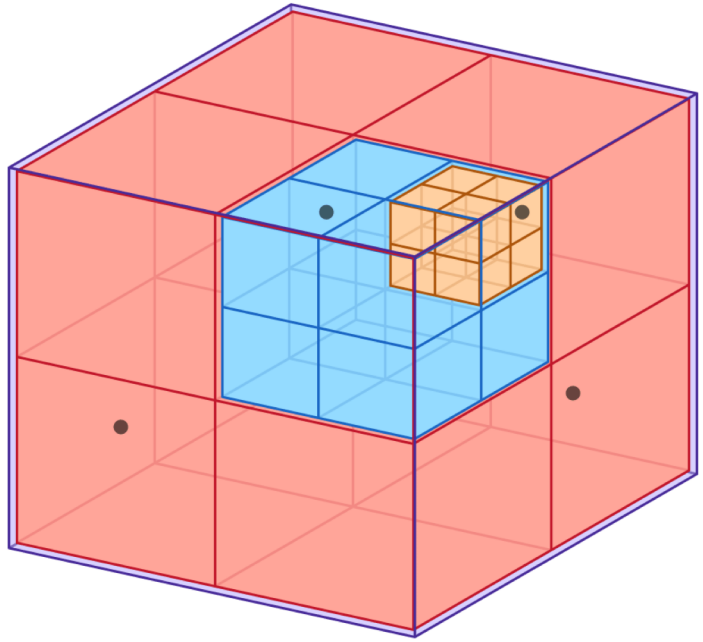
- Reserve most of the atoms of proteins and ligands.
- 27*27*27 matrix for proteins; length of receptive field is 135.
- 6*6*6 matrix for ligands; length of receptive filed is 30.
- Atoms mapped to the same cell of matrix will be added.

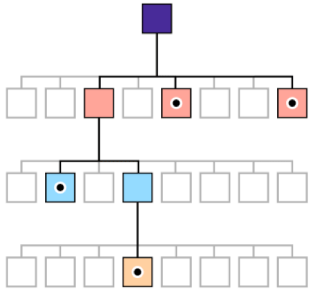


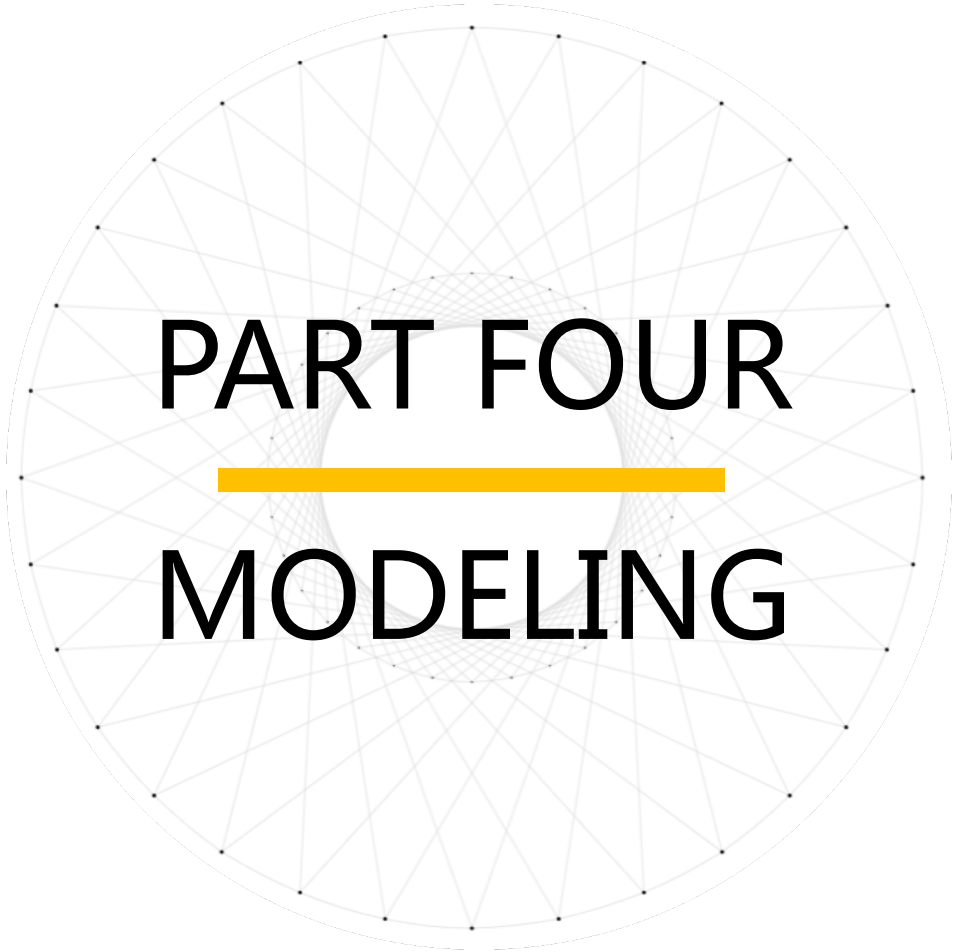
Octree for Global Hypothesis:

- Reserve all the atoms of proteins and ligands.
- Peng-Shuai Wang, Yang Liu et al. 2017. O-CNN: Octree-based Convolutional Neural Networks for 3D Shape Analysis. ACM Trans.







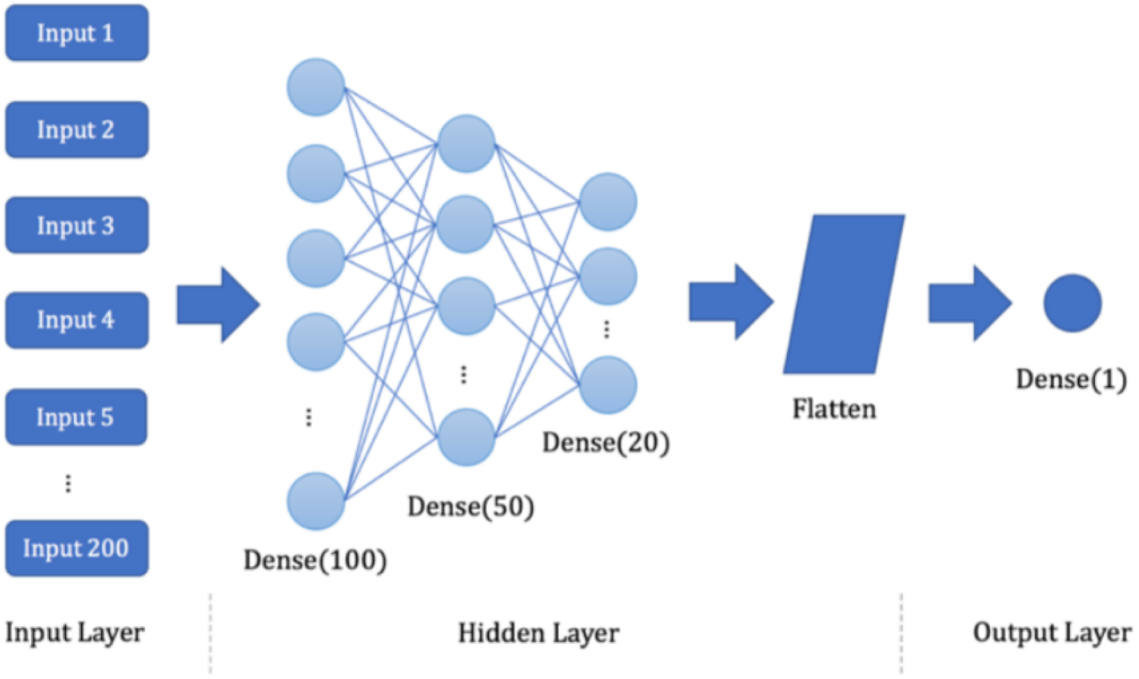


PART FOUR

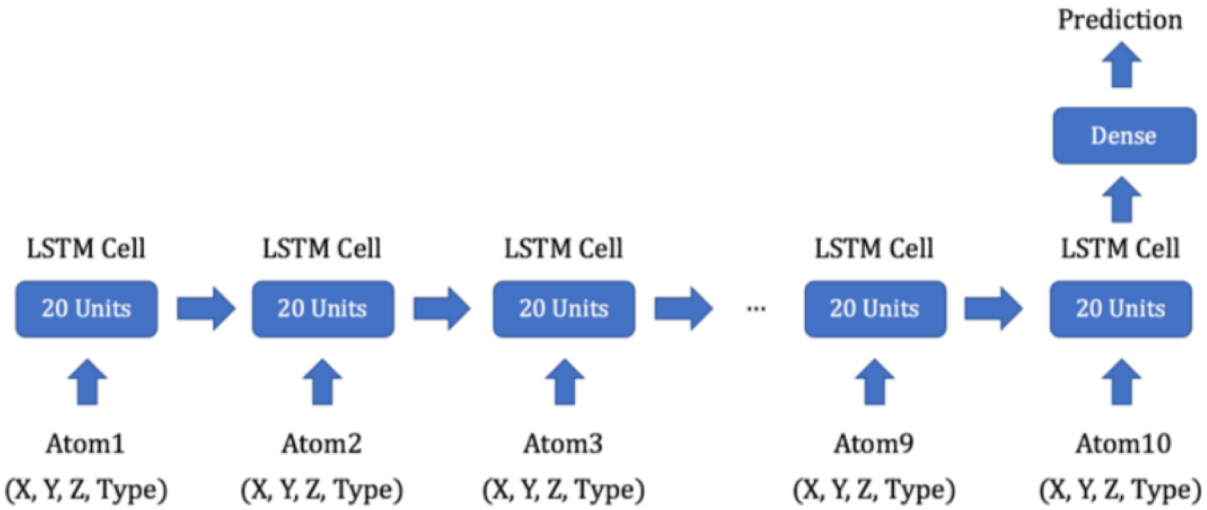
MODELING

PART FOUR: MODELING

With local hypothesis, use 10 atoms of a ligand and their neighbor in a protein to represent the whole pair in MLP and LSTM



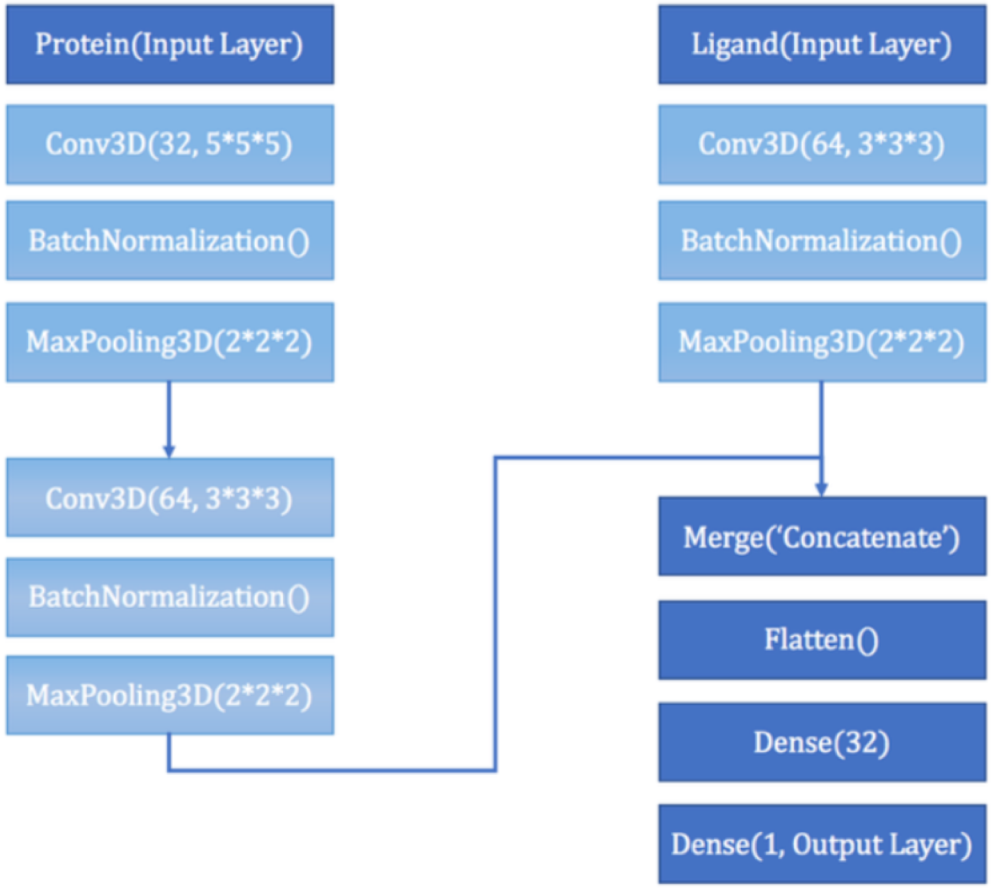
MLP(Multilayer Perceptron)



LSTM(Long Short Term Memory)

PART FOUR: MODELING

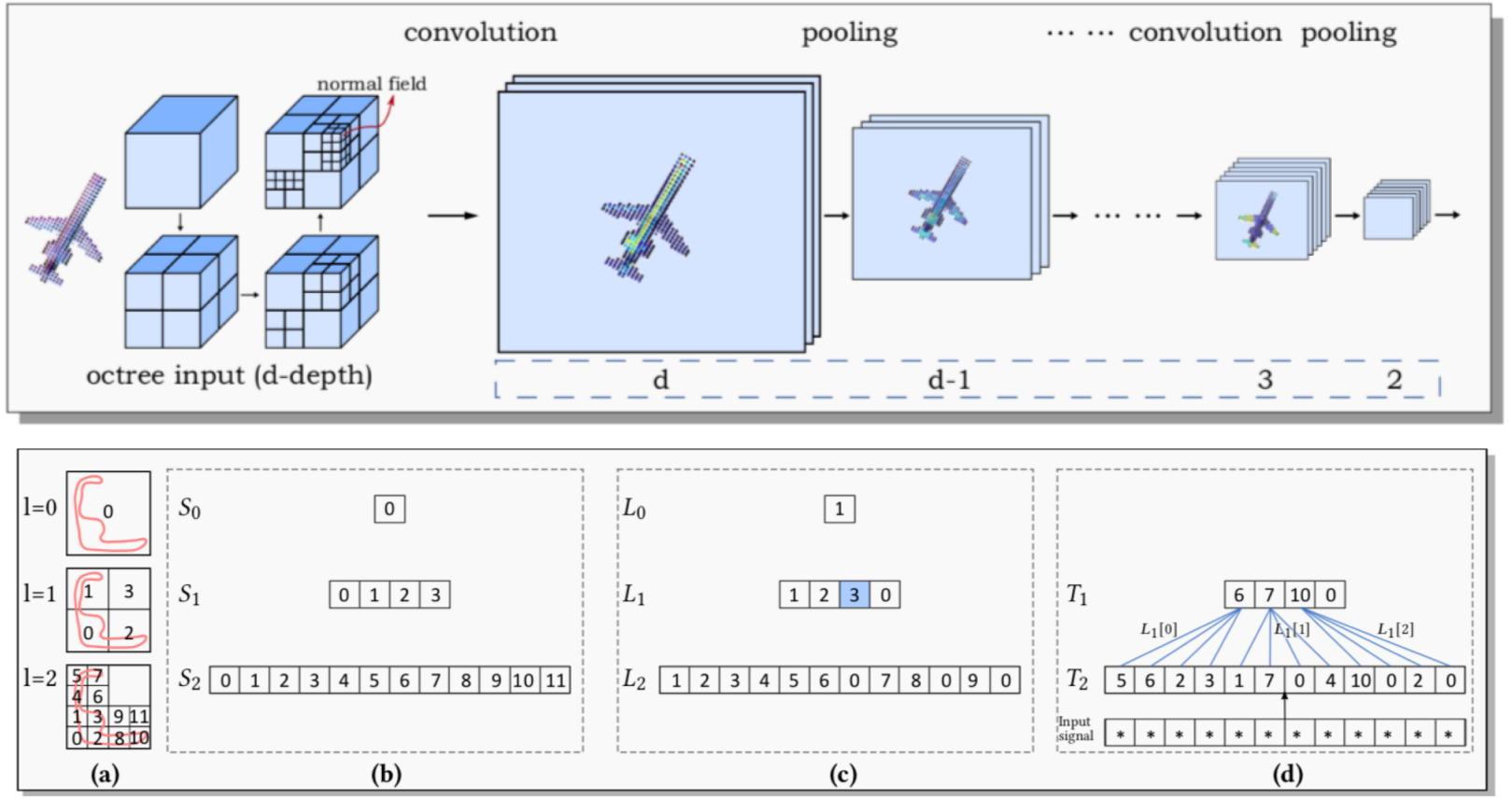
With global hypothesis, use 3D-CNN and reserve most of the atoms .



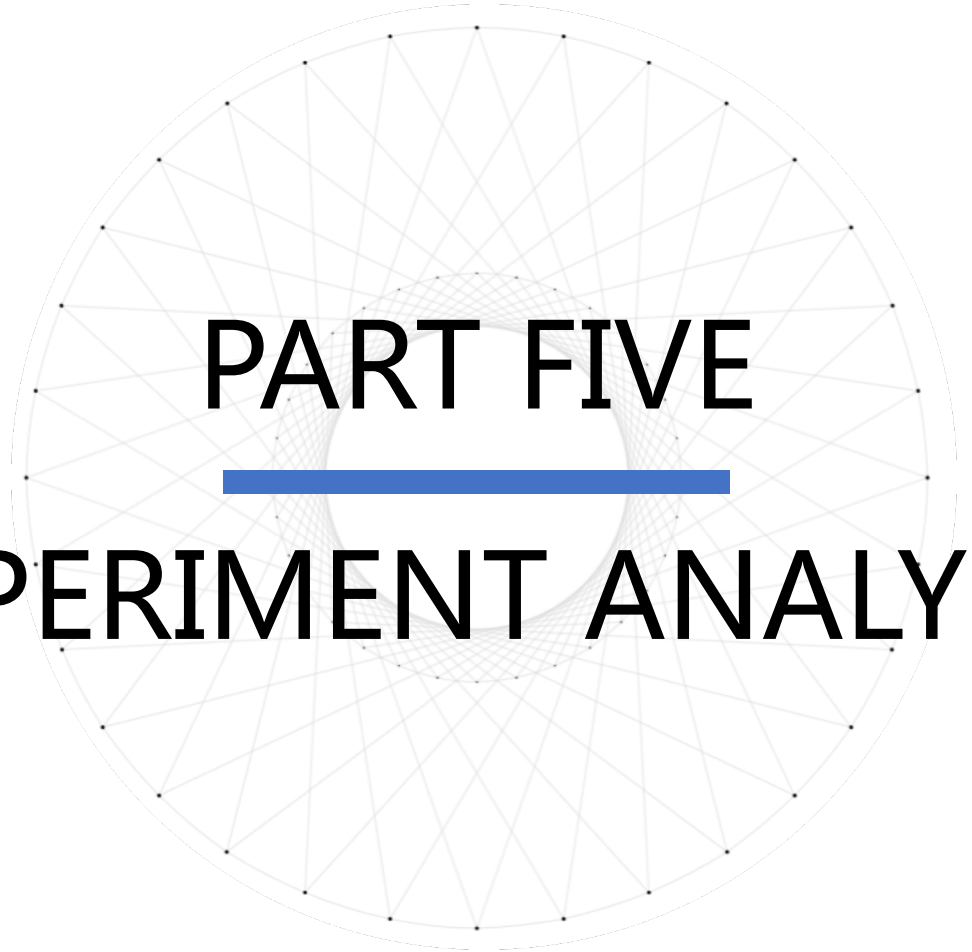
3D-CNN

PART FOUR: MODELING

With global hypothesis, we propose Octree-based-CNN is a promising choice that reserve all the atoms.



O-CNN



PART FIVE

EXPERIMENT ANALYSIS

PART FIVE: EXPERIMENT ANALYSIS

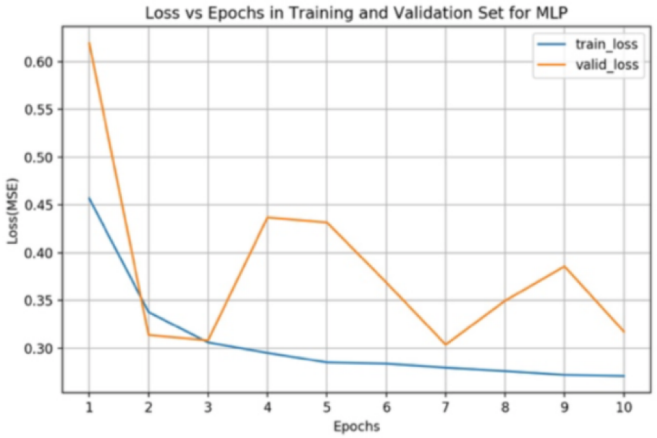
1. Comparison of MLP's Configuration

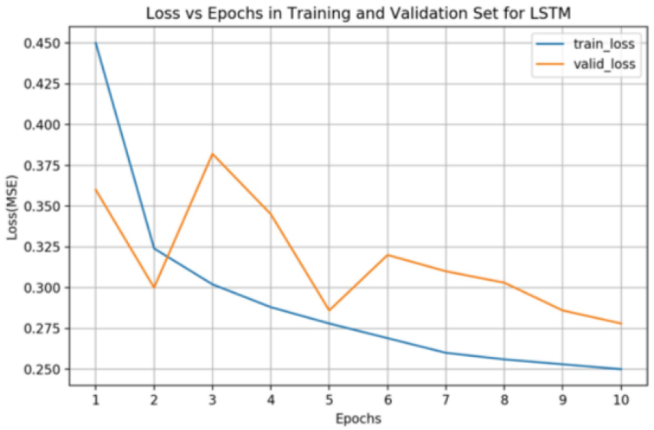
Dense_1	200	100	100	50	50	100	100
Dense_2	20	50	20	20	20	20	50
Dense_3	nil	nil	nil	nil	10	10	20
Accuracy (%)	52.3	56	58.3	57.3	57.3	58.7	59.7

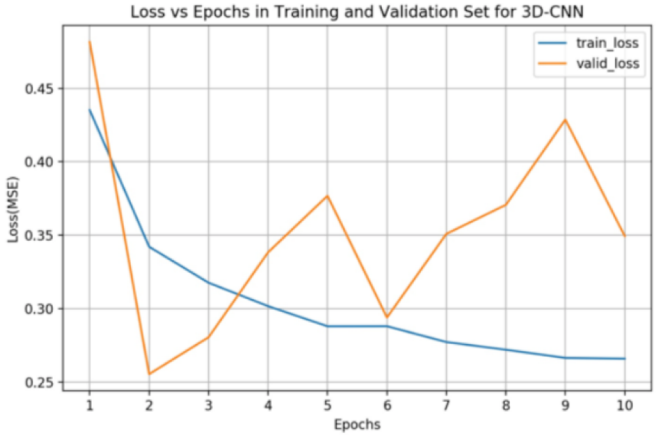
2. Comparison of Models' Performance

Models	Epoch(s)	Optimizer	Accuracy on Validation set (%)
MLP	10	Adam (lr=0.001)	59.7
LSTM	10	Adam (lr=0.001)	61.3
3D-CNN	10	Adam (lr=0.001)	43.5

Outlier Analysis - Why Validation loss fluctuates?







Reasoning: The imbalance of positive and negative instances in the validation set

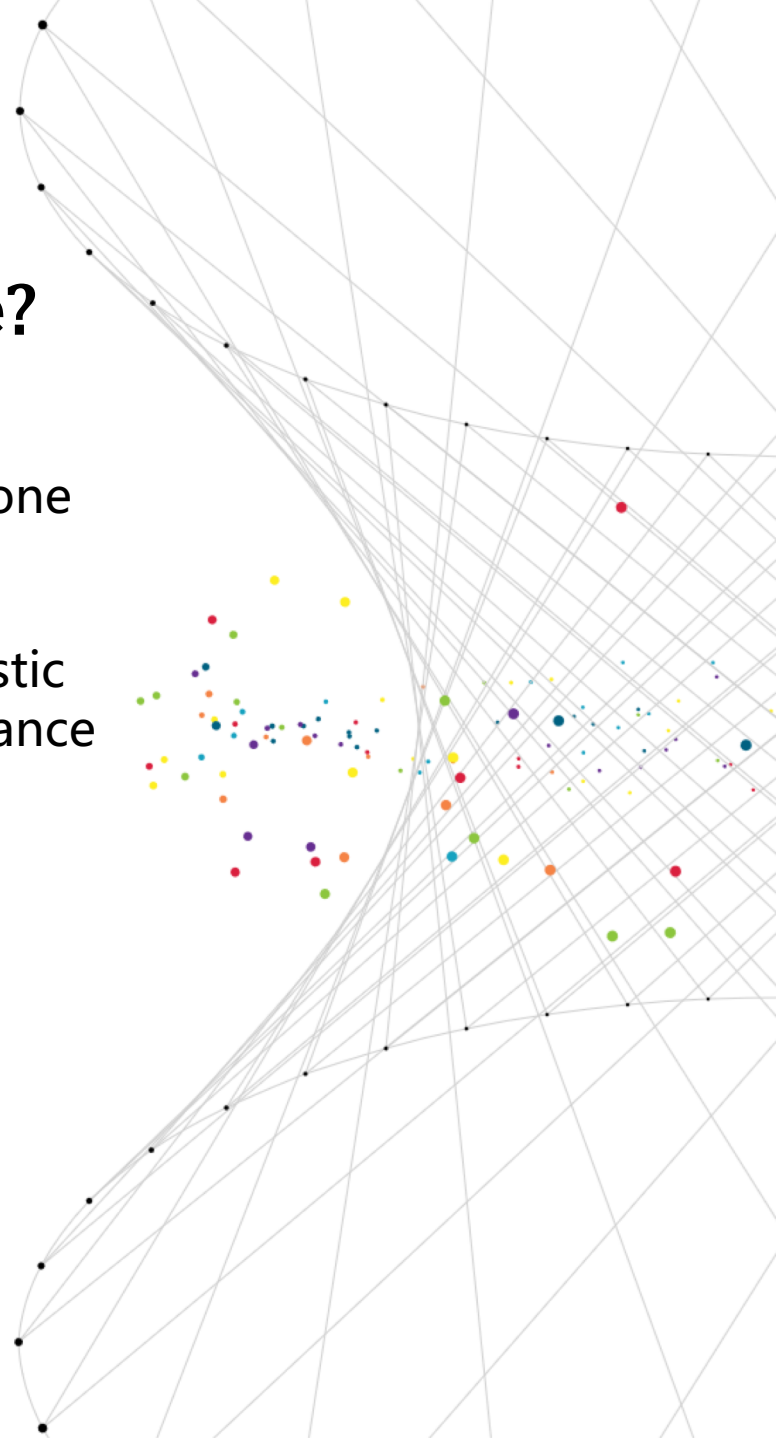
Outlier Analysis - Why does 3D-CNN not converge?

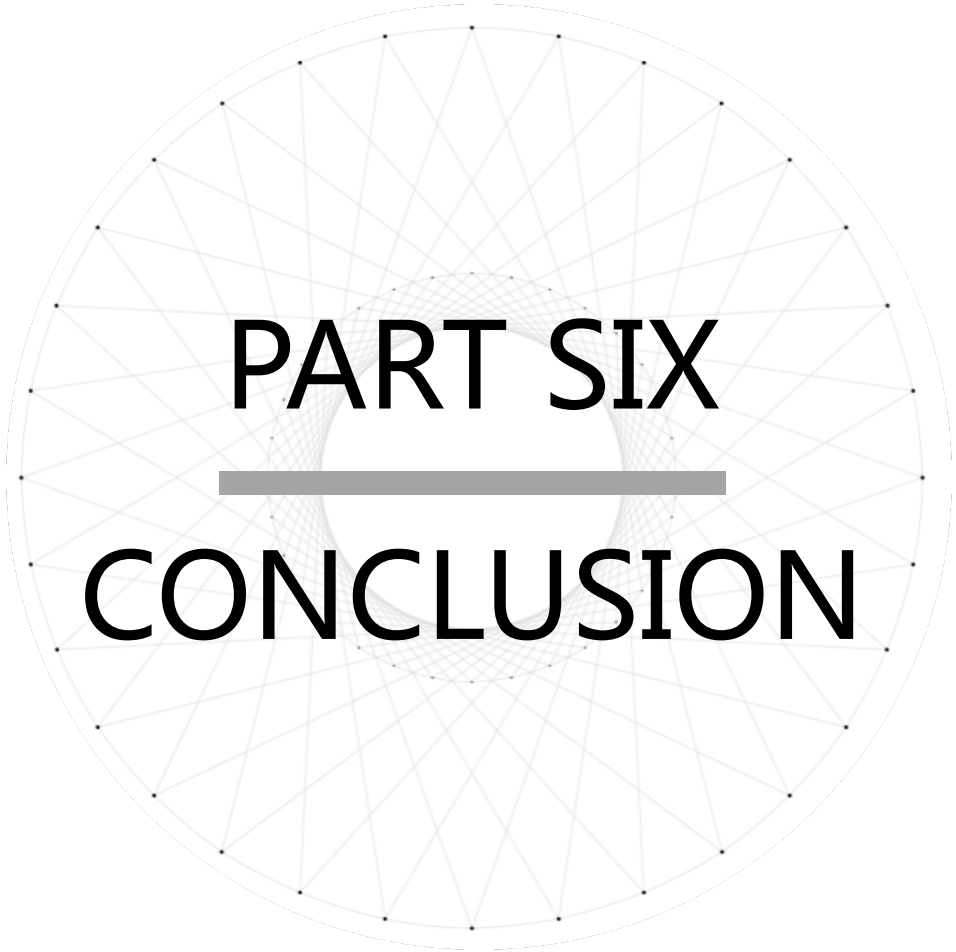
Reasoning 1: In many instances, a lot of atoms were squeezed into one cell, hence we lost lots of information.

Reasoning 2: The order that we fed data is fixed, rather than stochastic (i.e. we use a generator when training the model, and 1 positive instance is always followed by 2 negative instances).

Reasoning 3: Inappropriate activation functions.

Reasoning 4: A relatively high learning rate.





PART SIX

CONCLUSION

PART SIX: CONCLUSION

1. Prior exploration is important to know well about data.
2. Represent data with reasonable hypotheses to improve algorithm efficiency.
3. Combine multiple models to get better result.
4. Never forget to analyze the outliers to optimize your model.

NATIONAL UNIVERSITY OF SINGAPORE

THANK YOU!

Lian Yiming
A0174446Y E0210479

Yan Maitong
A0174365Y E0210398

PRESENTED BY TEAM 23