Protein-Ligand Binding Prediction

BS6207 Project

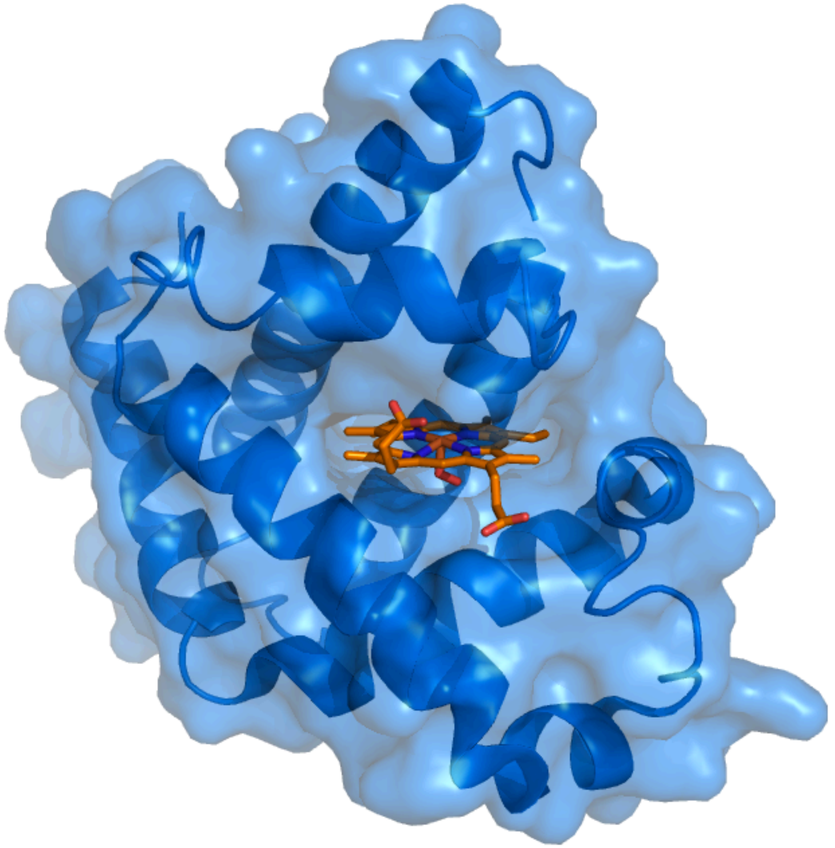
Zhang Xinxin
24-May-2021

Contents

Backgrounds

Protein-ligand complex

- Protein-ligand binding plays an important role in biological processes.
- The conformational change of the protein after protein-ligand binding enables various functionalities, such as catalysis of chemical reactions (enzyme-substrate complex), signal transduction (receptor-ligand complex), etc.
- Unlike the bond between atoms belonging to the same molecule, protein-ligand binding is usually reversible and non-covalent, such as hydrogen bonds, hydrophobic forces, van der Waals forces, pi interactions.
    - Formation of the protein-ligand complex depends on the molecular structure as well as atom properties



Problem Statement

Given the protein/ligand information in pdb format, we are asked to train a neural network model to predict protein-ligand binding pairs.

- Training/validation set: 3000 of binding pairs
- Testing set: 824 protein and 824 ligands without binding labels
- Features: X/Y/Z coordinates, atom type

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
10	ATOM	38	CB	ILE A	4	11.944	-23.314	-2.084	1.00	20.76	C
11	ATOM	43	CA	VAL A	5	7.516	-22.209	-1.354	1.00	21.45	C
12	ATOM	46	CB	VAL A	5	6.127	-22.882	-1.483	1.00	20.70	C
13	ATOM	50	CA	VAL A	6	7.351	-19.607	1.413	1.00	23.08	C
14	ATOM	53	CB	VAL A	6	8.432	-18.519	1.571	1.00	23.11	C
15	ATOM	57	CA	VAL A	7	4.032	-18.435	2.799	1.00	22.22	C
16	ATOM	60	CB	VAL A	7	2.994	-19.564	3.031	1.00	22.77	C
17	ATOM	64	CA	HIS A	8	2.839	-15.691	5.155	1.00	22.40	C
18	ATOM	67	CB	HIS A	8	3.714	-14.408	5.123	1.00	20.60	C
19	ATOM	75	CA	AGLY A	9	-0.090	-14.108	6.817	0.50	29.08	C
20	ATOM	76	CA	BGLY A	9	-0.201	-14.020	6.722	0.50	27.76	C

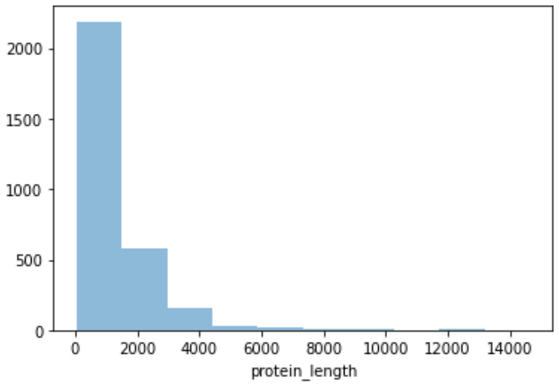
Data Processing

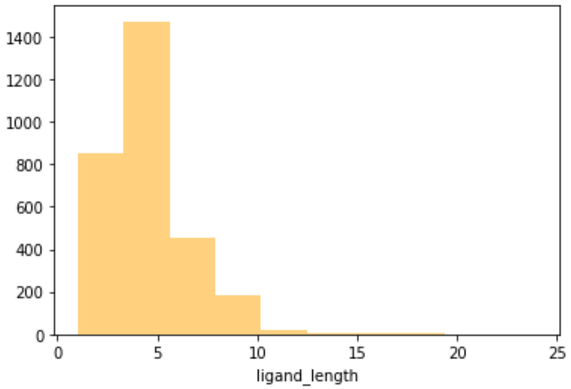
Challenges in Data Processing

Two main challenges when transforming the data from pdb files to a usable format for neural network input:

Varying sizes of protein and ligand

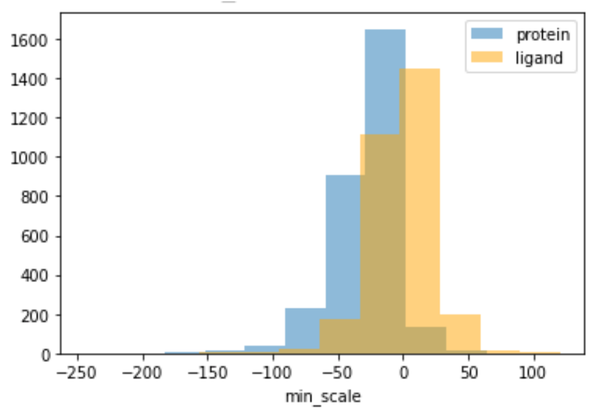
- The length of the protein/ligand varies
- Ligands are much shorter than proteins

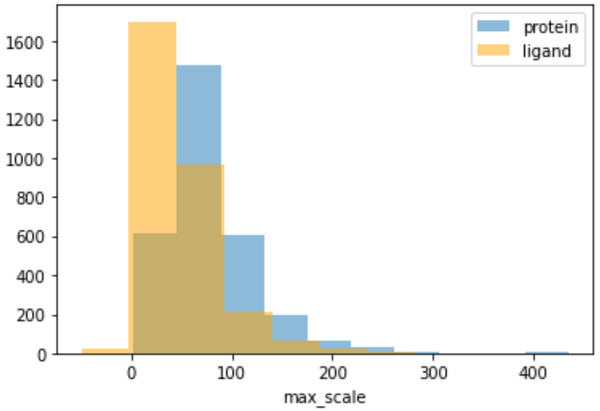




Varying scale of X/Y/Z coordinates

The scale of X/Y/Z coordinates vary by molecules





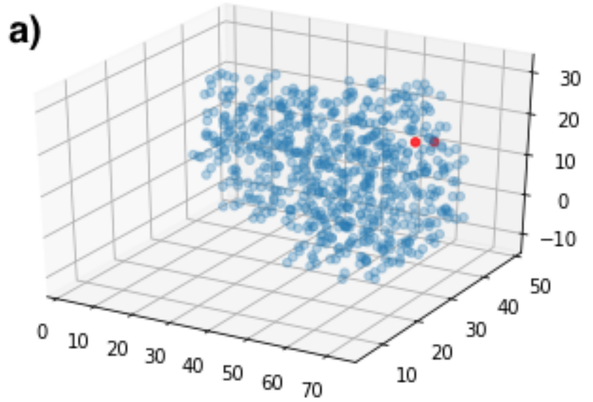
Voxelization

Data Processing (cont'd)

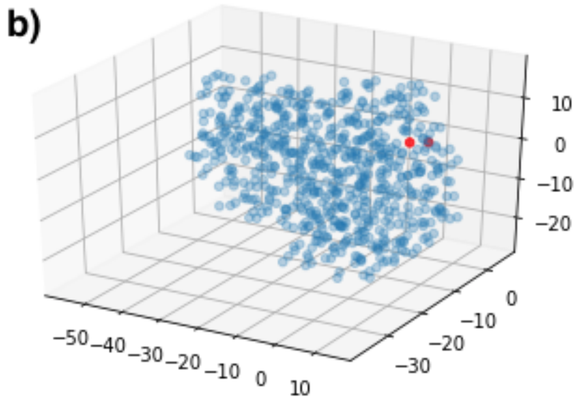
Voxelization

Based on the X/Y/Z coordinates, the protein and ligand can be voxelized in to a 3D matrix with 1 channel, value of which represents the molecule type and atom type.

(Hydrophobic protein - 1, polar protein - 0, hydrophobic ligand - 200, polar ligand - 100)

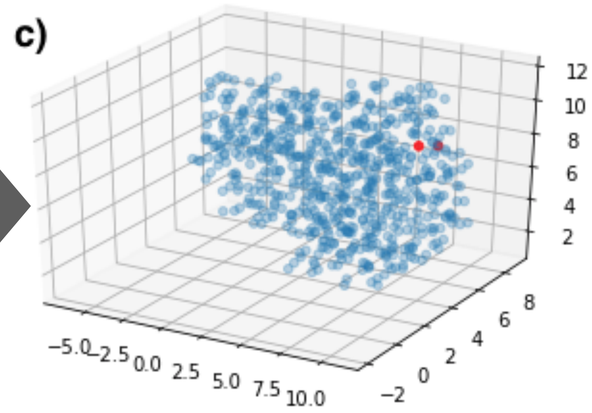


Original



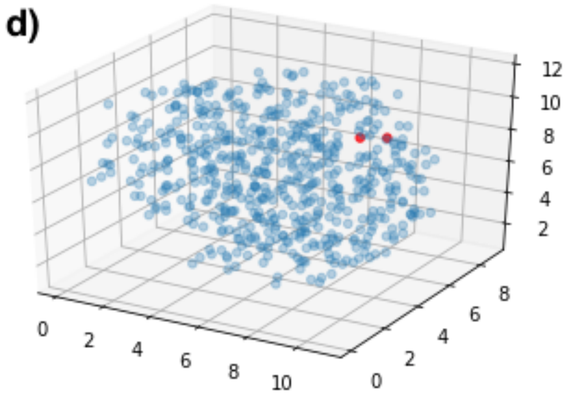
Centralized

- Use the centroid of the ligand coordinates to centralise the protein and ligand



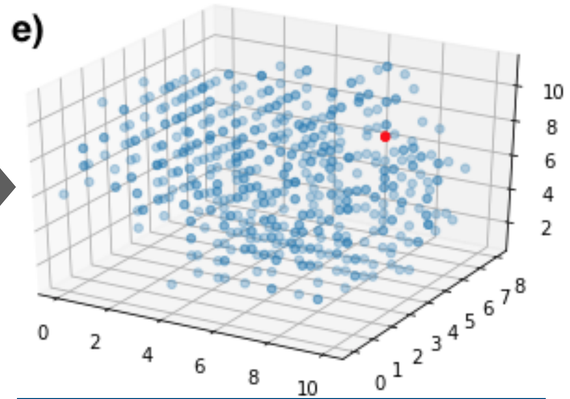
Shifted & Scaled

- Shift the protein and ligand to the positive space
- Scale the atoms to make it more condensed, so that more atoms can be retained from trimming



Trimmed

- Assume atoms at the outskirt does NOT impact the protein-ligand binding
- Trim the 3D matrix using a fixed box size to exclude atoms at the outskirt of the structure



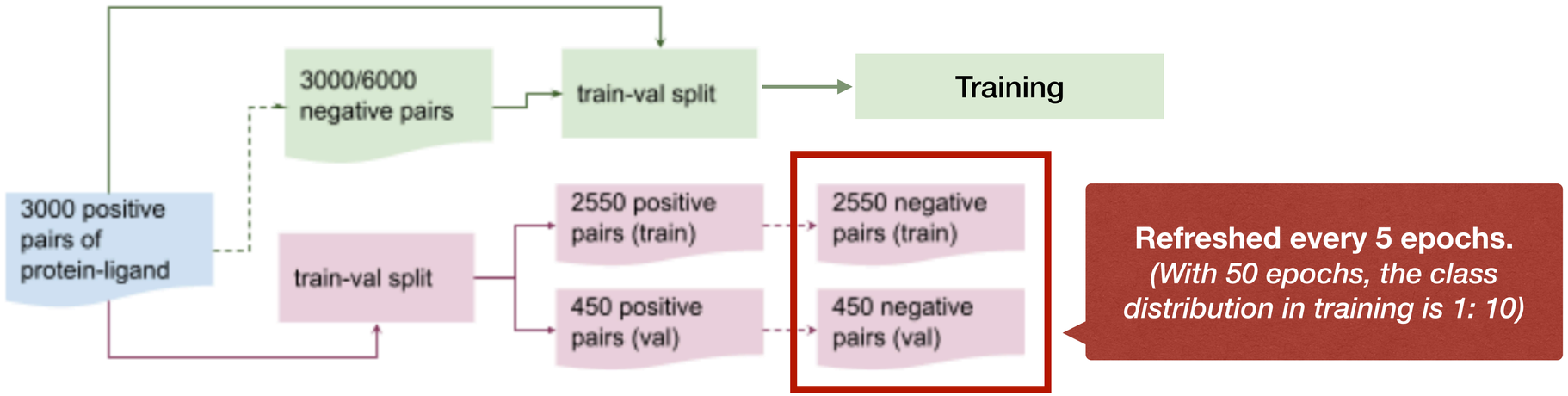
Rounded

- Round the coordinates of each atom to the nearest integer to form a sparse 3D matrix

Data Processing (cont'd)

Lack of negative cases & Test set class imbalance issue

- **The training set provided does NOT include any non-binding pairs.**
  - Non-binding pairs were generated by random shuffling of the protein/ligand sequences.
- **The prediction test is to be done on an extremely imbalanced dataset.**
  - In test set, binding : non-binding = 1: 823
  - Training the model on a balanced dataset may not yield good prediction performance on the imbalanced test set
  - Hence, more non-binding pairs were generated either statically (option 1) or dynamically (option 2)
  - Model performance was measured in an imbalanced validation set of 1:10 (pos: neg) using confusion matrix



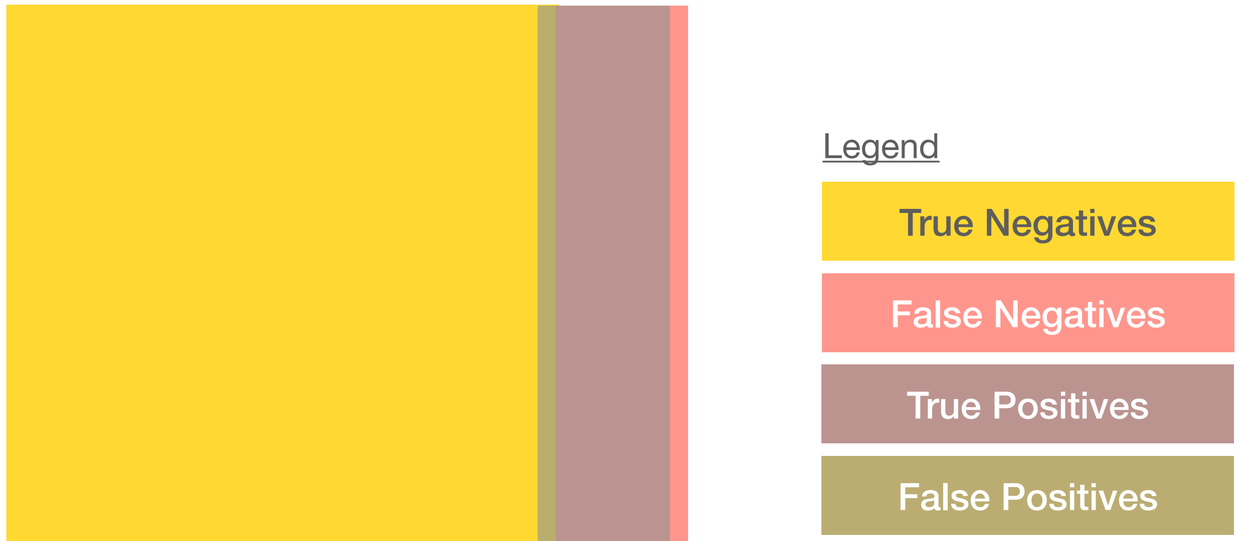
Data Imbalance Issue

Low FPR and High TPR is wanted

To predict the binding pairs in an imbalanced set with much more non-binding pairs, it is essential to increase the model TPR and reduce the model FPR

- High TPR ensures that all positive cases can be predicted as positive correctly
- Low FPR reduces the number of non-binding pairs predicted to be binding

This is to be handled by adjusting the training class ratio in the training set as well as tuning the loss function weights.



Neural Network & Experimental Study

Overview

The experimental study can be summarised as below:

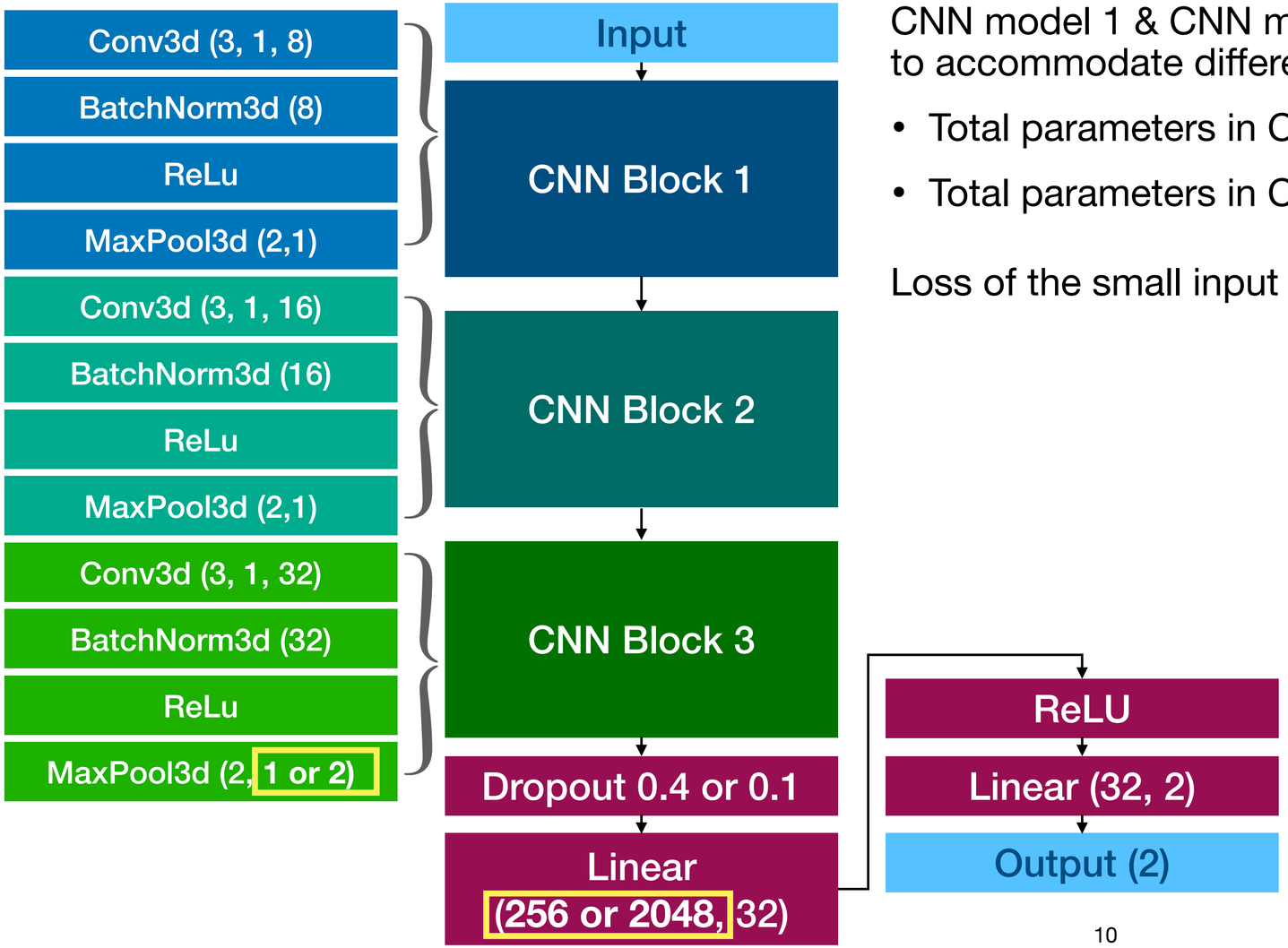
<table>
<tr><th>Input Size</th><th>Negative data generation</th><th>Neural Network Model</th><th>Loss function</th></tr>
<tr><td>Small (16, 16, 16, 1)</td><td>Static (pos: neg = 1:1 ~ 1:3)</td><td>CNN Model 1</td><td>Cross Entropy</td></tr>
<tr><td rowspan="3">Big (25, 25, 25, 1)</td><td>Static (pos: neg = 1:1 ~ 1:3)</td><td rowspan="3">CNN Model 2</td><td>Cross Entropy</td></tr>
<tr><td rowspan="2">Dynamic</td><td>Weighted Cross Entropy</td></tr>
<tr><td>Cost-sensitive loss based on cross entropy</td></tr>
</table>

General settings used in all trials above:

- Adam optimizer with learning rate of 0.0005
- Early stopping with patience of 13 ~ 15
- Batch size = 128
- Epochs: 50 / 80

Input Size & Neural Network

Larger input 3D matrix generates better performance

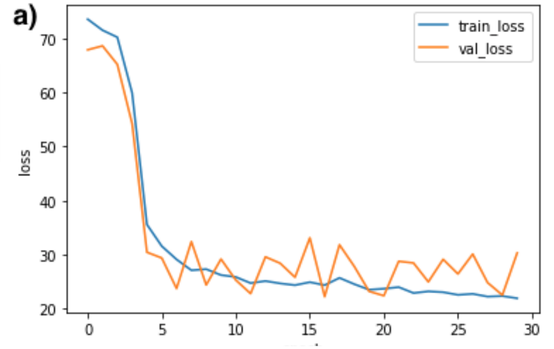


CNN model 1 & CNN model 2 have the same structure, with minor tweaks to accommodate different size of the input 3D matrix, respectively.

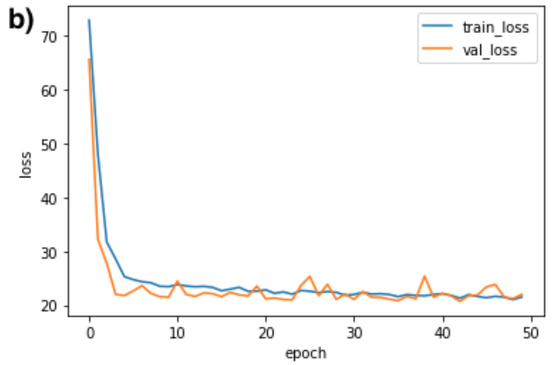
- Total parameters in CNN Model 1 with small input (16, 16, 16, 1): 25 954
- Total parameters in CNN Model 2 with large input (25, 25, 25, 1): 83 298

Loss of the small input oscillates much more than that of the large input.

Loss of small input (class ratio = 1:3)



Loss of large input (class ratio = 1:3)

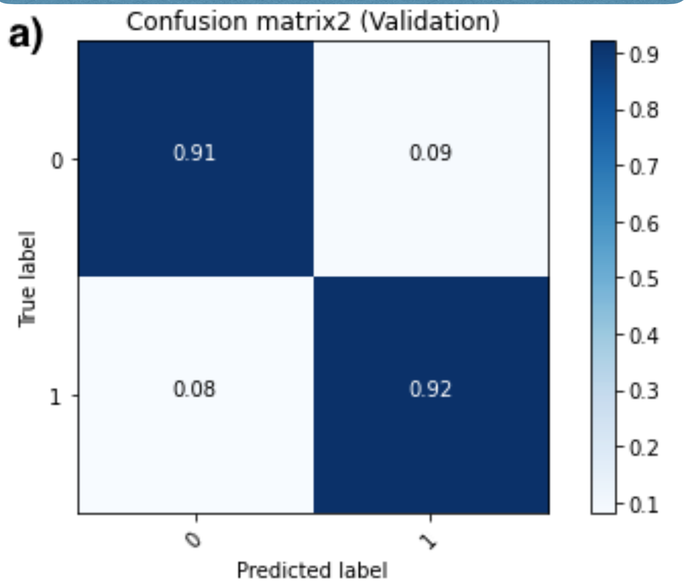


Static and Dynamic negative case generation

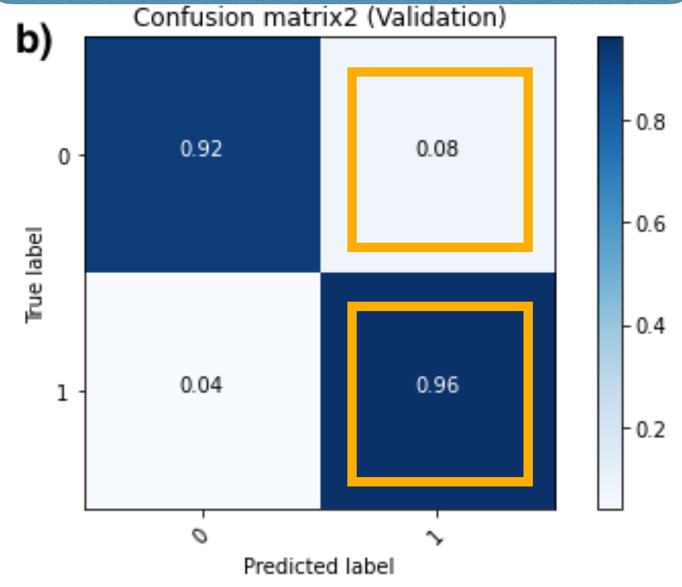
Dynamically refreshed negative cases yield highest recall

Model performance was evaluated on an imbalanced validation set (pos: neg = 1:10) to compare the different method of generation negative cases.

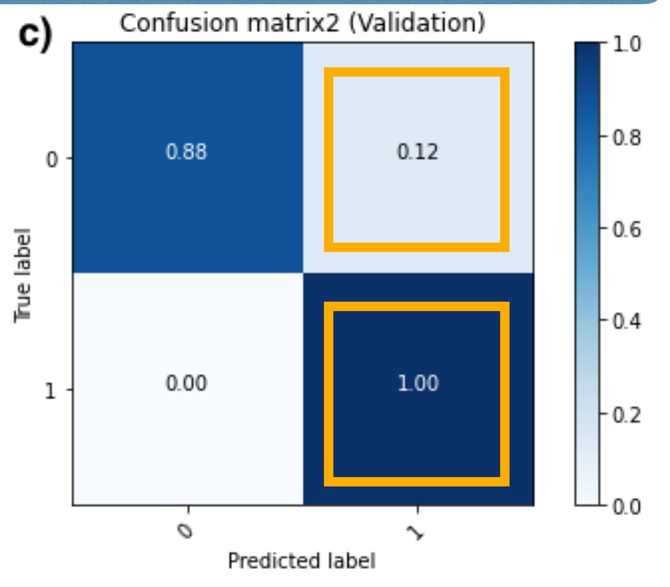
Generated and fixed before training starts (pos: neg = 1:1)



Generated and fixed before training starts (pos: neg = 1:2)



Dynamically refreshed every 5 epochs (pos: neg = 1:1 for each refresh)



Weighted Loss Function

To increase precision with imbalanced class

The extreme class imbalance in the test set may lead to high False Positive Rate. To account for the this issue, class weightage was added to the loss function to boost the cost of false predictions.

- **Weighted cross entropy** - boost the misclassification errors for the less-frequent class (i.e. negative class)
- **Cost-sensitive loss function based on cross entropy**

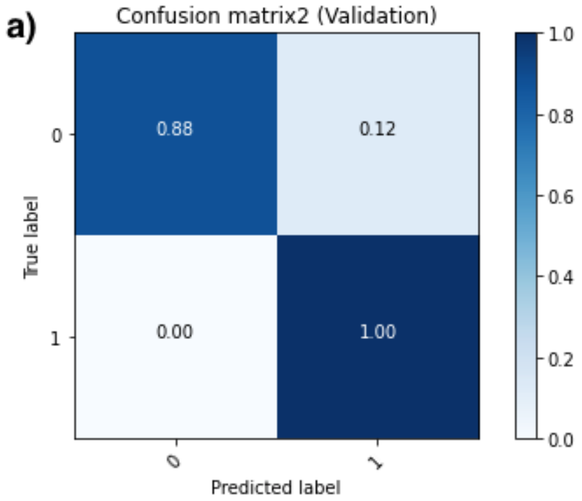
**Model performance was evaluated on an imbalanced validation set (pos: neg = 1:10)*

TN (10)	**FP (Highest loss, 400)**
FN (2nd highest loss, 200)	**TP (Reward, 0)**

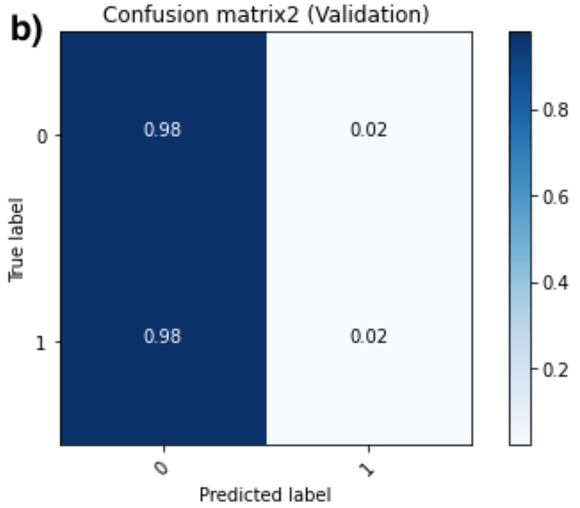
Unwanted

Wanted

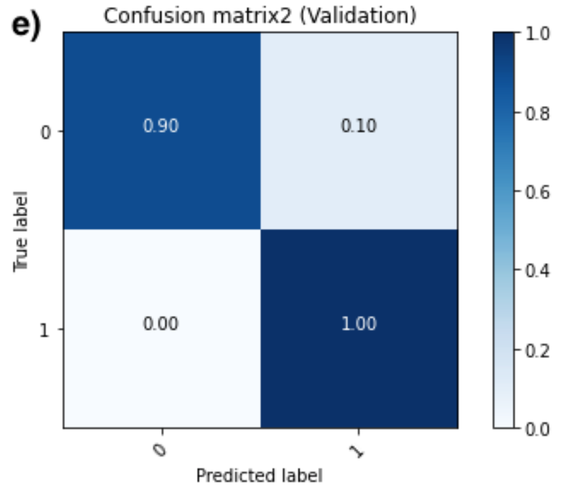
Normal cross entropy (baseline)



Weighted cross entropy (neg: pos = 200 : 1)



Cost-sensitive loss function (TP:TN:FP:FN = 0: 10: 400: 200)



Lesson learnt from experimental study

Some issues encountered during experimental study:

1. **Training/validation accuracy during training stuck at 50% (with pos: neg = 1:1)**
   - Reason: training data was not shuffled
   - Fix: shuffle the training data before constructing the dataloader
2. **Validation confusion matrix and accuracy were extremely high**
   - Reason: validation set overlaps with training set, due to the dynamically refreshed training data
   - Fix: split the positive pairs into training/validation first, and randomly shuffle the sequences within the two separate pool to generate negative pairs

Test Prediction

Final model = larger input matrix + dynamic training data refresh + cost-sensitive loss function

- The model was run to predict all possible protein-ligand pairs in the test set (824 x 824).
- The class with higher predicted logit score is the predicted class of the protein-ligand pair.
- For each protein, rank all pairs by the logit score to get the top 10 predictions.

pro_id	lig1_id	lig2_id	lig3_id	lig4_id	lig5_id	lig6_id	lig7_id	lig8_id	lig9_id	lig10_id
1	766	78	104	751	582	489	534	156	664	388
2	167	403	1	729	285	148	593	334	587	765
3	534	314	496	259	582	716	455	593	638	449
4	612	512	449	674	439	649	378	518	760	710
5	466	91	14	222	90	650	493	315	624	385
6	714	407	137	79	661	606	576	404	125	19
7	411	335	768	454	527	177	146	426	461	698
8	356	778	798	172	58	529	243	246	577	455
9	411	101	146	237	335	768	675	527	620	357
10	360	120	562	666	72	803	439	350	234	742
11	727	63	614	390	95	622	751	263	239	47
12	552	401	771	669	335	345	490	668	411	486
13	600	452	9	58	410	356	520	229	687	566
14	439	248	562	336	803	79	358	275	321	16
15	527	595	461	607	411	485	426	129	454	698
16	424	328	708	788	659	791	293	419	627	619
17	579	84	678	672	295	770	717	21	255	301
18	119	442	32	221	82	252	279	364	118	811
19	252	733	100	805	364	119	489	603	113	247
20	664	824	522	95	377	314	766	210	383	159
21	633	97	397	226	163	330	632	592	394	383
22	147	92	168	465	818	737	392	597	308	130
23	163	397	633	592	97	226	767	632	722	711
24	177	491	630	502	335	345	353	609	390	348
25	484	232	362	403	591	593	1	729	167	688
26	795	552	486	384	300	348	551	345	177	401
27	481	442	444	345	103	758	719	145	486	221
28	584	491	454	680	557	18	696	768	524	154
29	544	205	578	500	550	191	223	495	186	689
30	533	284	266	381	499	102	3	713	508	819

Conclusion and Future Improvement

- The final model has achieved relatively good prediction performance.
- Future improvements:
  - Use a larger input matrix with less information loss
  - Further fine tune of the class weights in the loss function
  - Train with more epochs

Q & A