DeepLPI: a novel deep learning-based model for protein-ligand interaction prediction for drug repurposing

David Wei

Princeton International School of Math and Science

Yau Science Awards 2022, North American Division, Computer Science

Project Highlights

Task

- Industry-ready, simplistic, accurate AI model
- For effective drug discovery

Work

- Started in early 2021
- Worked 1000+ hours, 9 versions, 2 million data

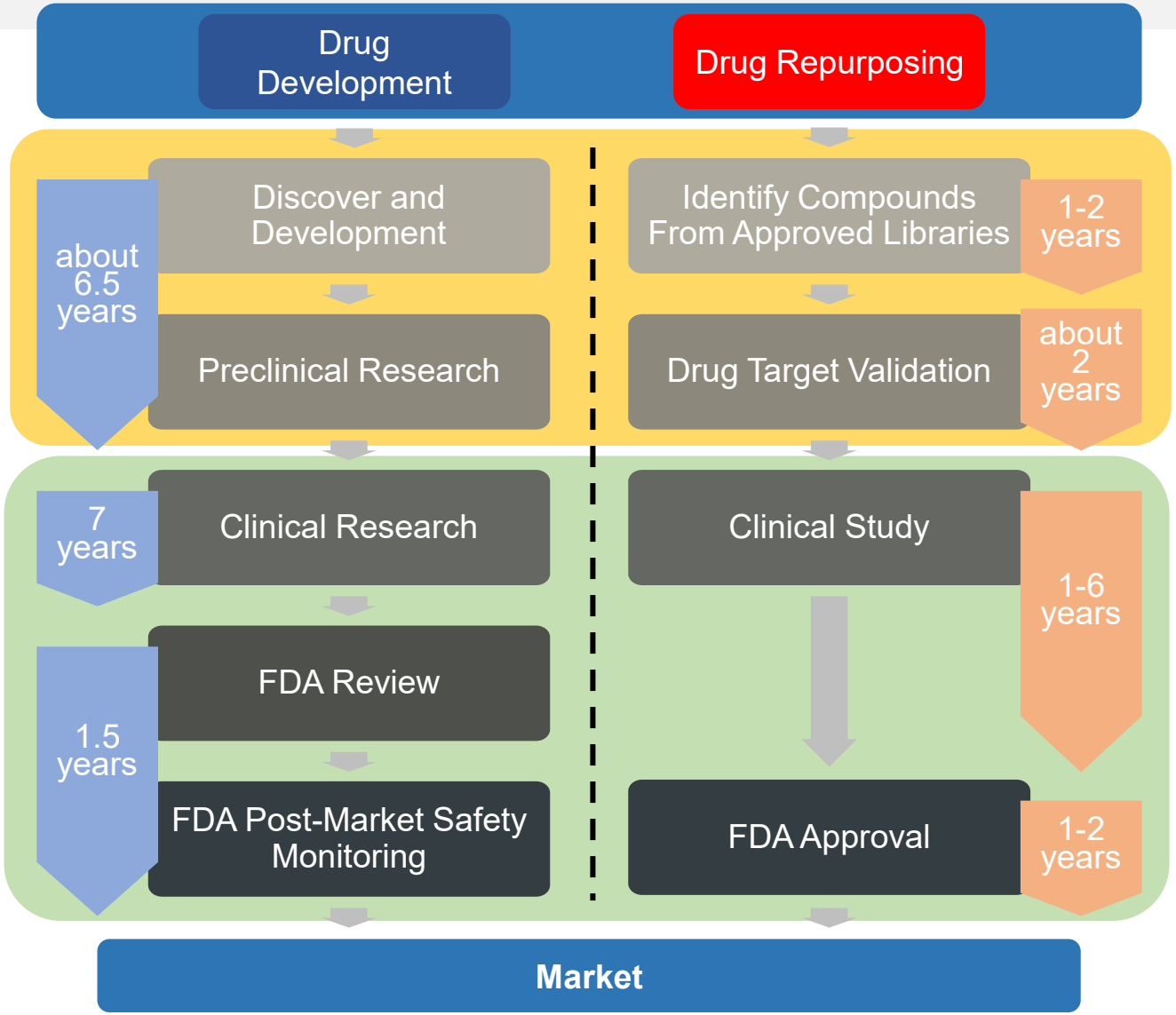
Project Highlights

Achievement

- Continued revision after submission to Yau Science Award
- Obtained significant improvement
- Presenting the most recent results today

- Poster presentation at ISMB
- Poster presentation at IDweek
- Manuscript published on Scientific Reports
  Scientific Reports vol. 12: 18200 (2022)
  https://www.nature.com/articles/s41598-022-23014-1

Drug repurposing significantly reduces drug discovery time.

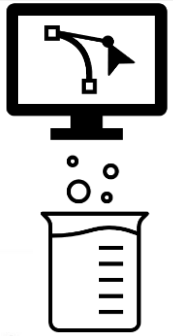


- Orange – physical or *in-silico* experiments.
- Green – animal and human experiments.

[1] S. Pushpakom *et al*, *Nat. Rev. Drug Dis.*, vol. 18, no. 1, Jan. 2019

DTI (Drug-Target Interaction)

DTI Prediction – (*in silico*-base approach)



Computer-based (*in silico*) Prediction: **Fast**

Lab Experiment Test: **Slow**

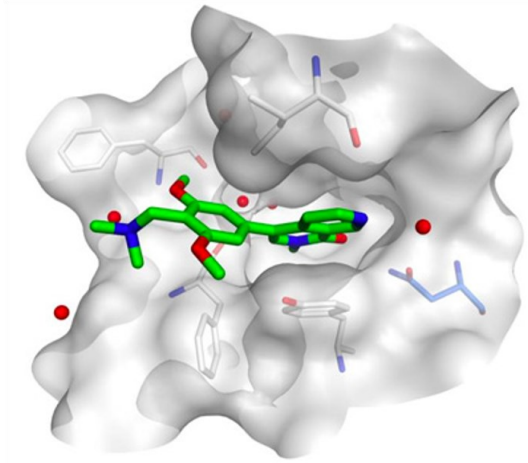


Image from DeepPurpose Documentation: https://github.com/kexinhuang12345/DeepPurpose

Binding

Strong Binding Affinity

Good Drug Candidate

Non-binding

Weak Binding Affinity

Poor Drug Candidate

Illustrations by the presenter

Input

- Drug - Molecule
- Protein – Sequence

Output

- Interaction or not
- Binding/Non-binding

Traditional Models Rely on Complex and Rare Protein Spatial Information

Fast but inaccurate

- Human selecting features

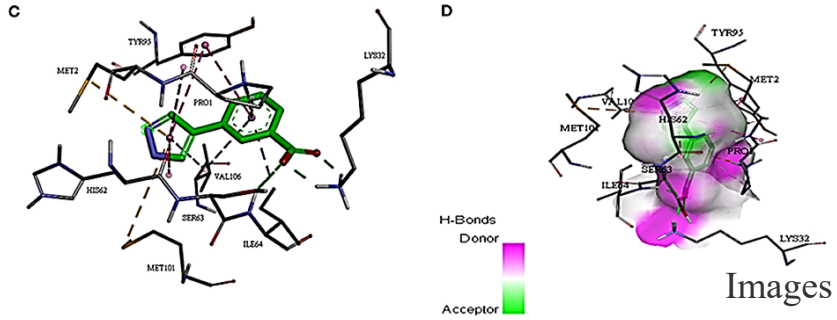
	FNN	SVM	RF	KNN
StaticF	0.687 ± 0.131	0.668 ± 0.128	0.665 ± 0.125	0.624 ± 0.120
SemiF	0.743 ± 0.124	0.704 ± 0.128	0.701 ± 0.119	0.660 ± 0.119
ECFP6	0.724 ± 0.125	0.715 ± 0.127	0.679 ± 0.128	0.669 ± 0.121
DFS8	0.707 ± 0.129	0.693 ± 0.128	0.689 ± 0.120	0.648 ± 0.120
ECFP6 + ToxF	0.731 ± 0.126	0.722 ± 0.126	0.711 ± 0.131	0.675 ± 0.122

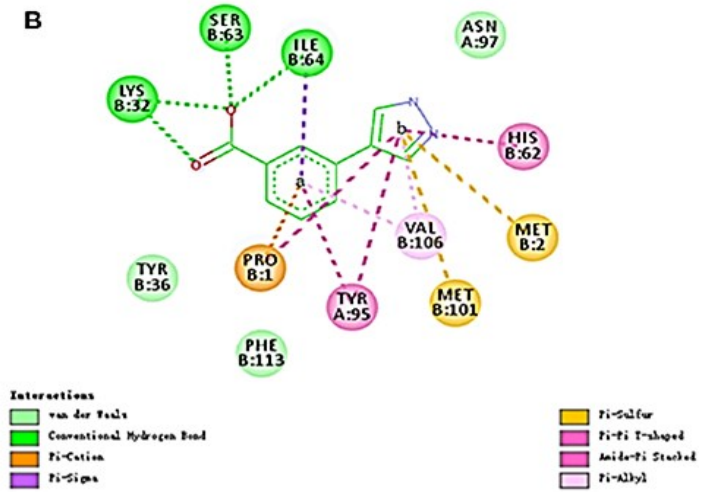
Traditional ML model prediction precision on binding affinity MSE

Data from references [1] https://doi.org/10.1186/s13321-017-0209-z [2] https://doi.org/10.1039/C8SC00148K

Accurate but limited/rare

- 3D-Structure-Based Models
- Limited data and variation



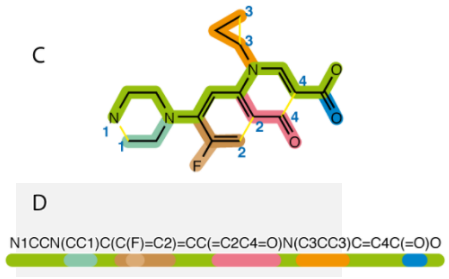


Images taken from references [3] https://doi.org/10.3389/fgene.2020.607824 [4] https://arxiv.org/abs/1510.02855

DeepLPI model overview (best after 12 versions)

Input

Drug Molecule -- SMILES format



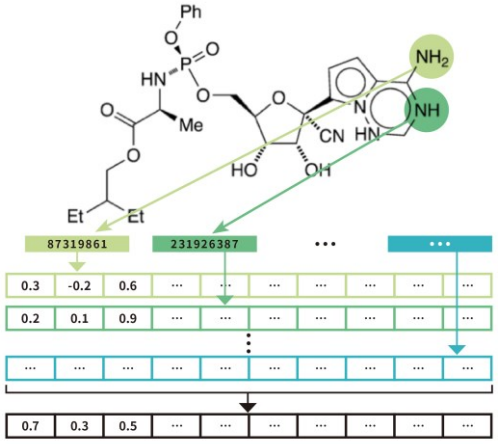
Target Protein -- FASTA format

```
;LCBO - Prolactin precursor - Bovine
; a sample sequence in FASTA format
MDSKGSSQKGSRLLLLLVVSNLLLCQGVVSTPVCPNGPGNCQVSLRDLFDRAVMVSHYIHDLSS
EMFNEFDKRYAQGKGFITMALNSCHTSSLPTPEDKEQAQQTHHEVLMSLILGLLRSWNDPLYHL
VTEVRGMKGAPDAILSRAIEIEEENKRLLEGMEMIFGQVIPGAKETEPYPVWSGLPSLQTKDED
ARYSAFYNLLHCLRRDSSKIDTYLKLLNCRIIYNNNC*
```

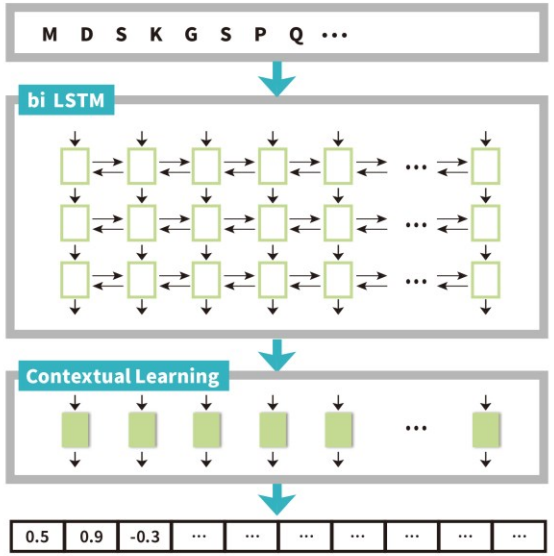
Image and Data from [1] https://en.wikipedia.org/wiki/FASTA_format
[2] https://en.wikipedia.org/wiki/Simplified_molecular-input_line-entry_system

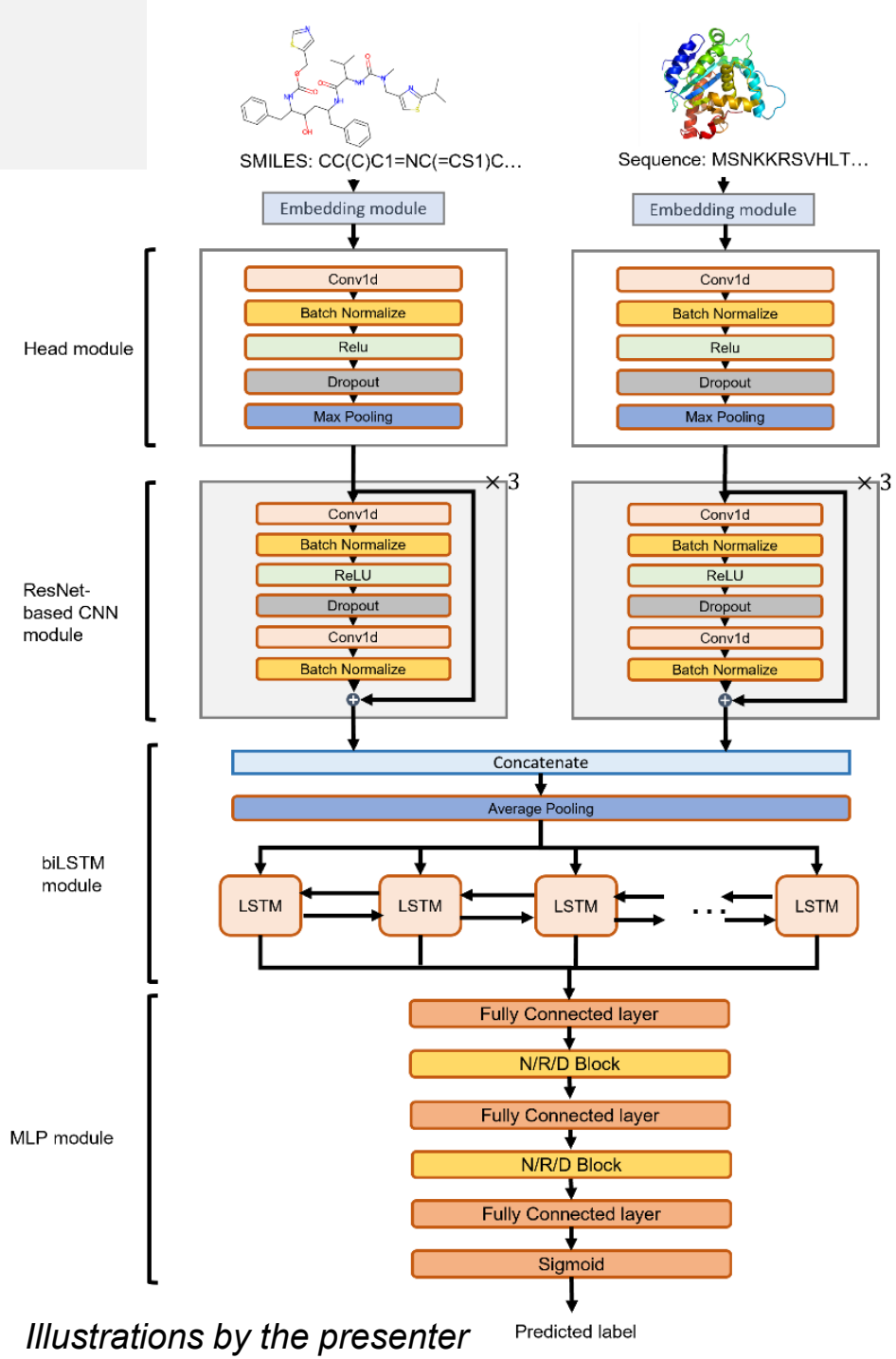
Embedding

Mol2Vec Embedding



ProSE Embedding





Illustrations by the presenter

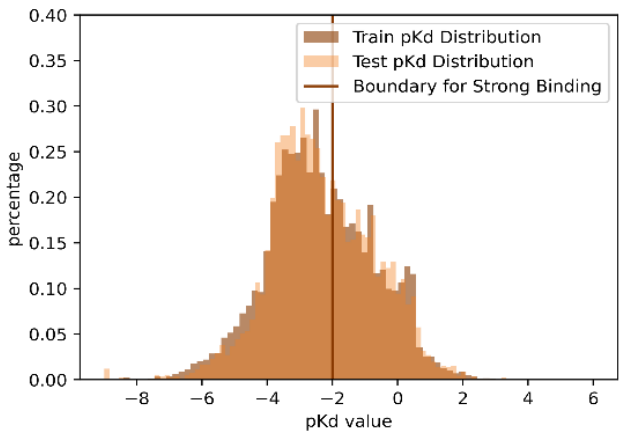
Train Data Selection

No duplicates, High Confidence Experiment, Balanced Label

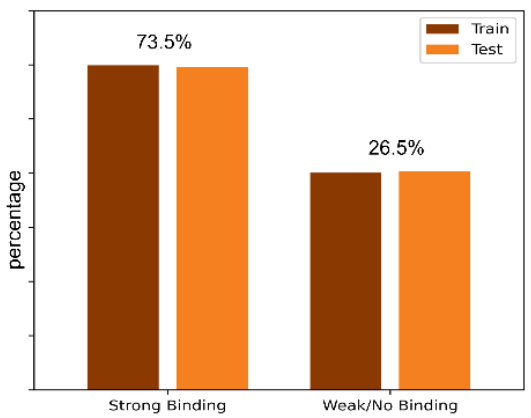
All Images on this page created by the presenter

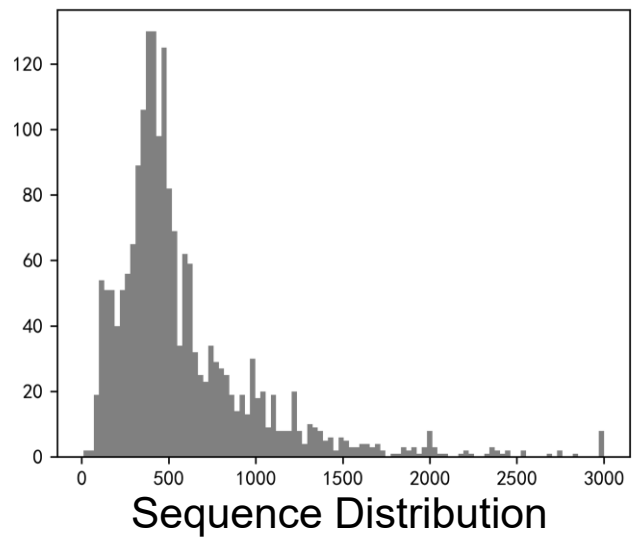
Data stats and processing for BindingDB dataset. Davis dataset follow a similar pre-processing and stats.

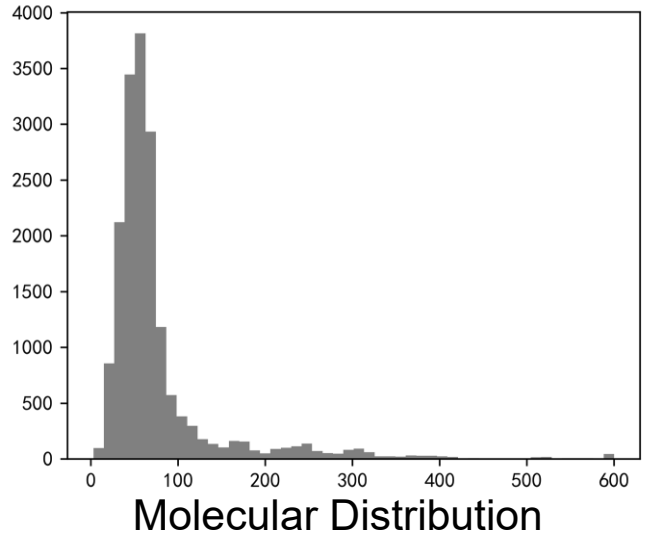
(a)



(b)







Original
Total: 2,407,235
Drug: 1,036,498 Seq: 8,130

↓

Remove Multichain
Total: 2,291,976
Drug: 992,203 Seq: 7,969

↓

Choose Only Kd Label
Total: 93,978
Drug: 21,148 Seq: 1,944

↓

Remove Invalid Value
Total: 93,977
Drug: 21,148 Seq: 1,944

↓

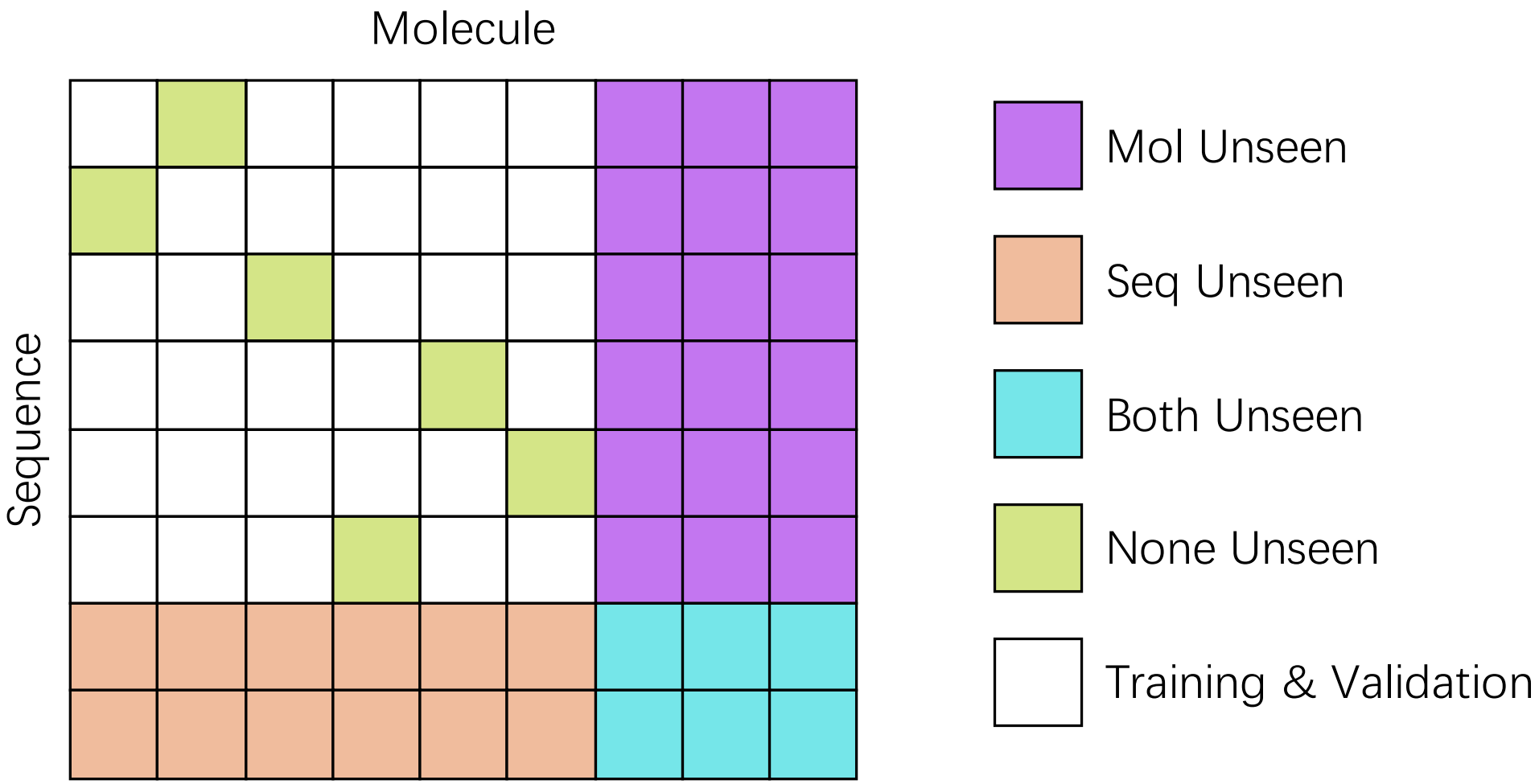
Remove Duplicates
Total: 62,825
Drug: 21,148 Seq: 1,944

Convert to 0-1 Classification

Train Data Selection

On ***BindingDB*** dataset with DeepLPI

All Images on this page created by the presenter



Experiment Settings: Traning

On ***BindingDB*** dataset with DeepLPI

Common Setup			
Dropout	0.3	Learning rate (LR)	0.001
Weight initialization	Kaiming		0.0001
Optimizer	Adam	LR decay rate	0.8
Batch size	256		

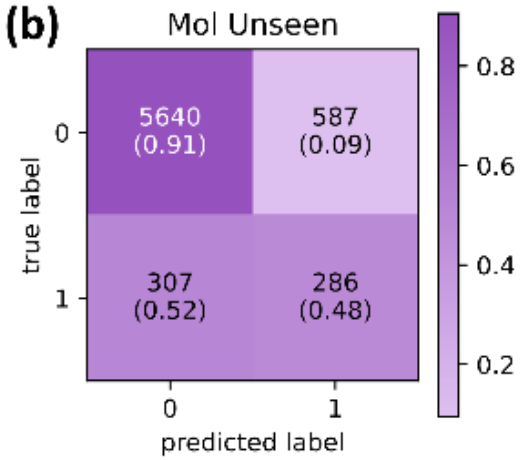
Loss Function = Binary Cross Entropy + L2 regularization

$$\text{Loss} = \underbrace{-\frac{1}{N}\sum_{i=1}^{N}[y_i \cdot \log(\hat{y}_i) + (1-y_i) \cdot \log(1-\hat{y}_i)]}_{\text{BCE loss}} + \underbrace{\alpha \|W\|_2^2}_{\text{L2-norm regularization}}$$

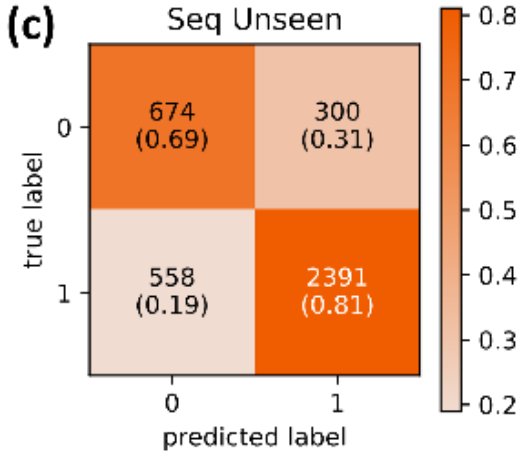
Independent Testing Results

On ***BindingDB*** dataset with DeepLPI

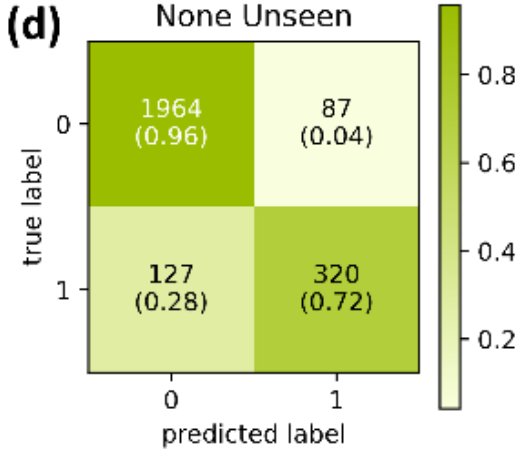
All Images on this page created by the presenter



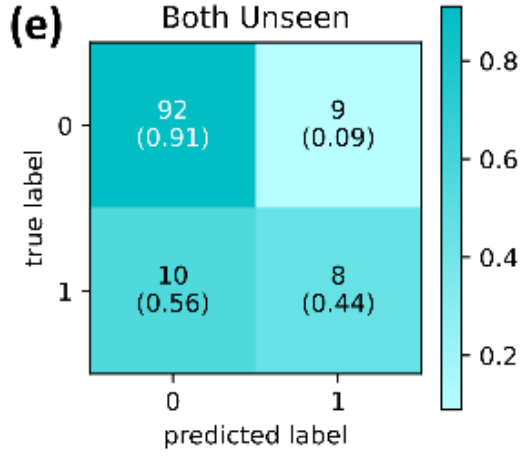
AUROC: 0.78
ACC: 0.87



AUROC: 0.86
ACC: 0.78



AUROC: 0.94
ACC: 0.91



AUROC: 0.78
ACC: 0.84

AUROC: Area under the receiver operating characteristic

ACC: Accuracy, percent of value predict correctly

Independent Testing Results

On ***BindingDB*** dataset with DeepLPI-6165

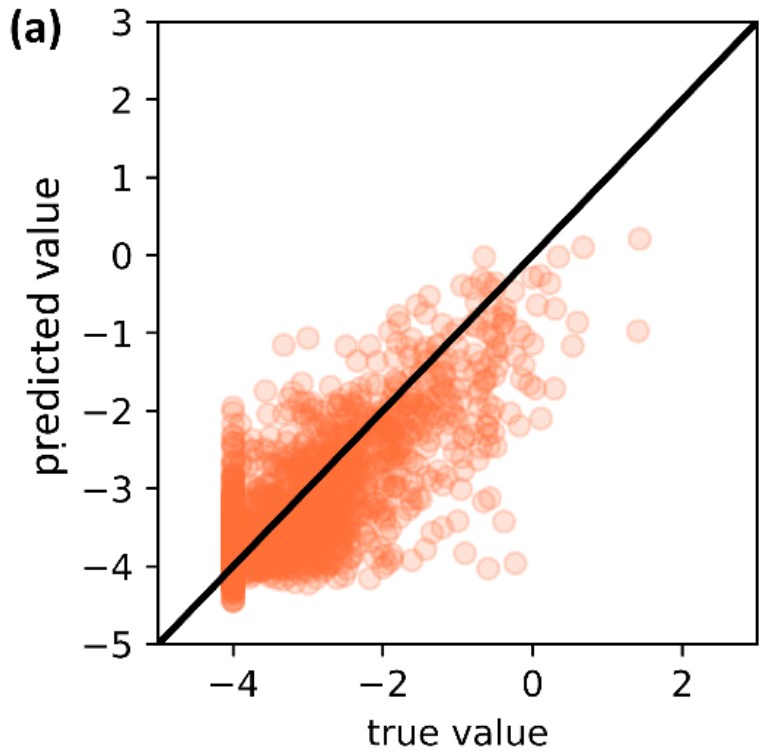
All Images on this page created by the presenter

AUROC	Mol Unseen	Protein Unseen	Both Unseen	None Unseen
Control	**0.782**	**0.862**	**0.781**	**0.942**
Mol Masked	0.625 (0.003)	0.726 (0.004)	0.737 (0.037)	0.726 (0.003)
Seq Masked	0.407 (0.008)	0.727 (0.005)	0.403 (0.046)	0.713 (0.007)

Independent Testing Results

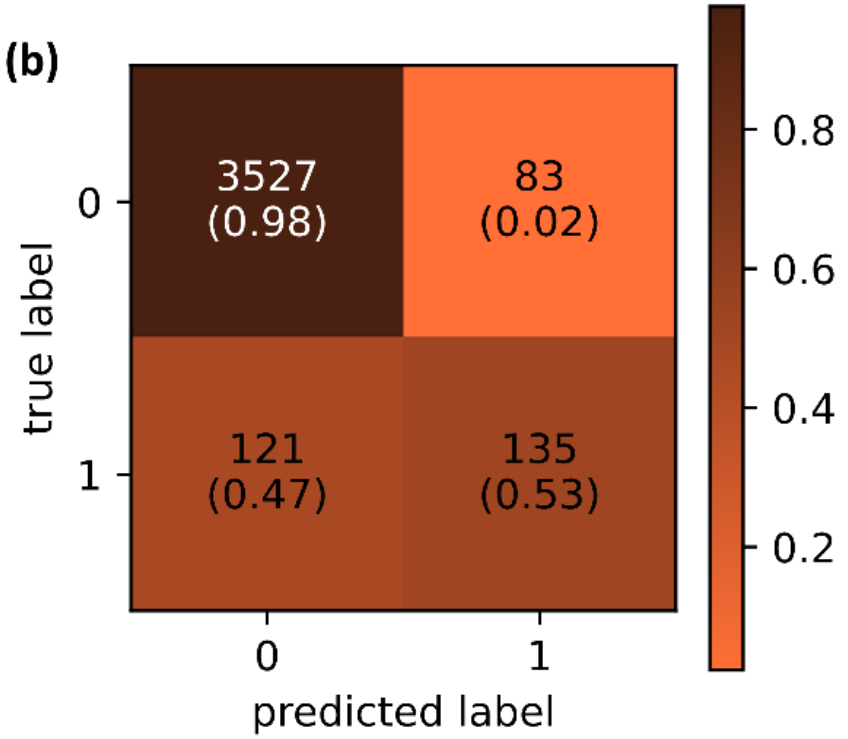
On ***Davis*** dataset with DeepLPI

All Images on this page created by the presenter



R2: 0.69

MSE: 0.196



AUROC: 0.90

ACC: 0.95

Performance Comparison

Model	AUROC	Sensitivity	Specificity
DeepLPI	0.94	0.72	0.96
DeepCDA	0.88	0.79	0.80
DeepDTA	0.89	0.77	0.86
Unseen Testsets Combined			
This work	0.79	0.68	0.77
DeepCDA	0.45	0.00	1.00

Comparing model performance on BindingDB

Model	AUROC	Sensitivity	Specificity
DeepLPI	0.923	0.930	0.730
DeepCDA	0.912	0.766	0.896
DeepDTA	0.909	0.865	0.795
DTITR	0.932	---	---

Comparing model performance on Davis

	R^2	MSE
DeepLPI	0.70	0.196
DeepCDA	0.74	0.208
DeepDTA	0.75	0.215
DTITR	0.77	0.192

Comparing model performance on Davis, Regression

	AUROC	Sensitivity	Specificity
DeepLPI	0.610	0.538	0.576
DeepCDA	0.400	0.000	1.000

Comparing model performance on Covid-19 data

Limitations and Future Work

Advanced Testing Methods

- Significantly different training and testing sets, distribution-wise
- Quantifying the difference
- Molecular Scaffold
- Protein Similarity
- Other metrics

More on repurposing drugs

- Potential adverse effects
- Due to new interactions between the drug and the proposed disease target, or a new group of population.
- Interactions with traditional drugs on the new disease to give adverse effects

Acknowledgements

Prof. Yue Zhang, School of Medicine, University of Utah

Dr. Xiang Gong, PRISMS

Prof. Shuyun Dong, School of Pharmacy, University of Utah

Key References

[1] S. Pushpakom et al., “Drug repurposing: progress, challenges and recommendations,” Nature Reviews Drug Discovery, vol. 18, no. 1, Jan. 2019, doi: 10.1038/nrd.2018.168.

[2] O. J. Wouters, M. McKee, and J. Luyten, “Estimated Research and Development Investment Needed to Bring a New Medicine to Market, 2009-2018,” JAMA, vol. 323, no. 9, Mar. 2020, doi: 10.1001/jama.2020.1166.

[10] T. He, M. Heidemeyer, F. Ban, A. Cherkasov, and M. Ester, “SimBoost: a read-across approach for predicting drug–target binding affinities using gradient boosting machines,” Journal of Cheminformatics, vol. 9, no. 1, Dec. 2017, doi: 10.1186/s13321-017-0209-z.

[11] R. Özçelik, H. Öztürk, A. Özgür, and E. Ozkirimli, “ChemBoost: A Chemical Language Based Approach for Protein – Ligand Binding Affinity Prediction,” Molecular Informatics, vol. 40, no. 5, May 2021, doi: 10.1002/minf.202000212.

[12] P.-W. Hu, K. C. C. Chan, and Z.-H. You, “Large-scale prediction of drug-target interactions from deep representations,” Jul. 2016. doi: 10.1109/IJCNN.2016.7727339.

[13] H. Öztürk, A. Özgür, and E. Ozkirimli, “DeepDTA: Deep drug-target binding affinity prediction,” in Bioinformatics, Sep. 2018, vol. 34, no. 17, pp. i821–i829. doi: 10.1093/bioinformatics/bty593.

[14] I. Wallach, M. Dzamba, and A. Heifets, “AtomNet: A Deep Convolutional Neural Network for Bioactivity Prediction in Structure-based Drug Discovery,” Oct. 2015.

[15] S. Wang et al., “SE-OnionNet: A Convolution Neural Network for Protein–Ligand Binding Affinity Prediction,” Frontiers in Genetics, vol. 11, Feb. 2021, doi: 10.3389/fgene.2020.607824.

[20] S. Jaeger, S. Fulle, and S. Turk, “Mol2vec: Unsupervised Machine Learning Approach with Chemical Intuition.”

[21] T. Bepler and B. Berger, “Learning the protein language: Evolution, structure, and function,” Cell Systems, vol. 12, no. 6, Jun. 2021, doi: 10.1016/j.cels.2021.05.017.

[22] T. Liu, Y. Lin, X. Wen, R. N. Jorissen, and M. K. Gilson, “BindingDB: a web-accessible database of experimentally determined protein-ligand binding affinities,” Nucleic Acids Research, vol. 35, no. Database, Jan. 2007, doi: 10.1093/nar/gkl999.

[23] D. Rogers and M. Hahn, “Extended-Connectivity Fingerprints,” Journal of Chemical Information and Modeling, vol. 50, no. 5, May 2010, doi: 10.1021/ci100050t.

[24] K. He, X. Zhang, S. Ren, and J. Sun, “Deep Residual Learning for Image Recognition,” Dec. 2015.

[25] K. He, X. Zhang, S. Ren, and J. Sun, “Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification,” Feb. 2015.

[26] D. P. Kingma and J. Ba, “Adam: A Method for Stochastic Optimization,” Dec. 2014.

[27] M. I. Davis et al., “Comprehensive analysis of kinase inhibitor selectivity,” Nature Biotechnology, vol. 29, no. 11, Nov. 2011, doi: 10.1038/nbt.1990.

[28] J. Tang et al., “Making Sense of Large-Scale Kinase Inhibitor Bioactivity Data Sets: A Comparative and Integrative Analysis,” Journal of Chemical Information and Modeling, vol. 54, no. 3, Mar. 2014, doi: 10.1021/ci400709d.

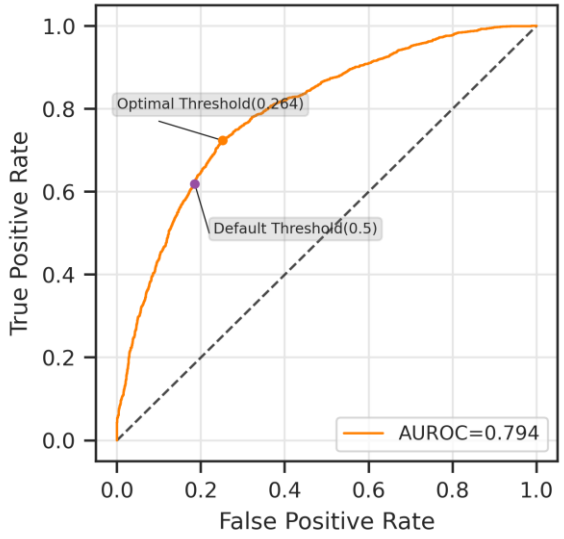
[29] K. Abbasi, P. Razzaghi, A. Poso, M. Amanlou, J. B. Ghasemi, and A. Masoudi-Nejad, “DeepCDA: deep cross-domain compound–protein affinity prediction through LSTM and convolutional neural networks,” Bioinformatics, vol. 36, no. 17, Nov. 2020, doi: 10.1093/bioinformatics/btaa544.

Thanks for listening!

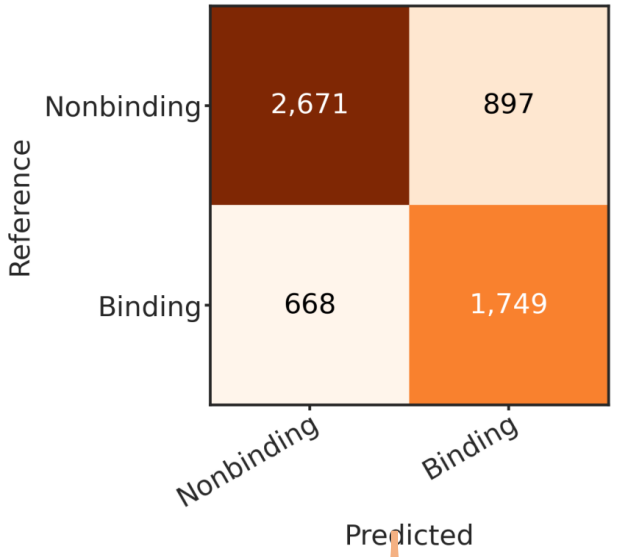
Independent Testing Results

On ***BindingDB*** dataset with DeepLPI-6165

All Images on this page created by the presenter



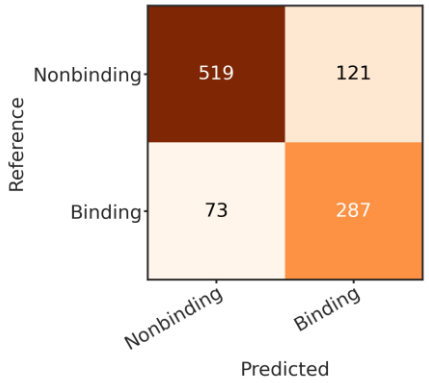
From the AUROC curve, we determine an optimal threshold for dividing the predicted Y values into binary (0/1) binding/non-binding values.



Overall confusion matrix

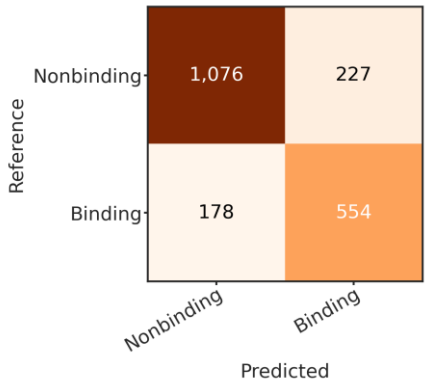
AUROC = 0.794
Sensitivity:0.724
specificity:0.749
PPV:0.661
NPV:0.800
(based on optimal threshold)

Both seen, AUROC = 0.877



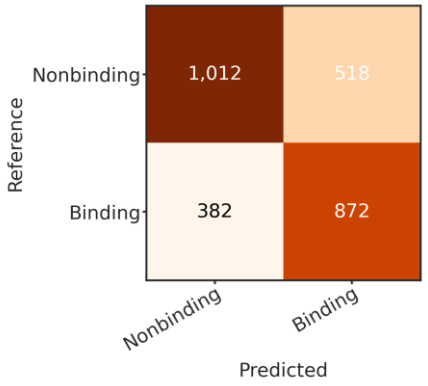
Sensitivity:0.797; PPV:0.703
specificity:0.811; NPV:0.877

Molecule unseen, AUROC = 0.857



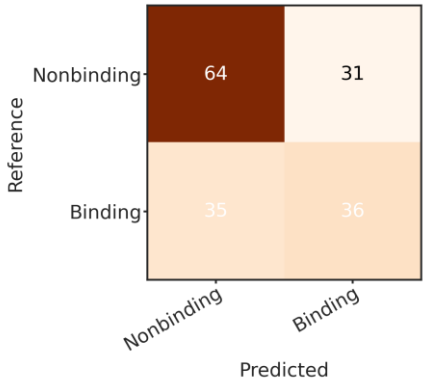
Sensitivity:0.757; PPV:0.709
specificity:0.826; NPV:0.858

Protein unseen, AUROC = 0.718



Sensitivity:0.695; PPV:0.627
specificity:0.661; NPV:0.726

None seen, AUROC = 0.655

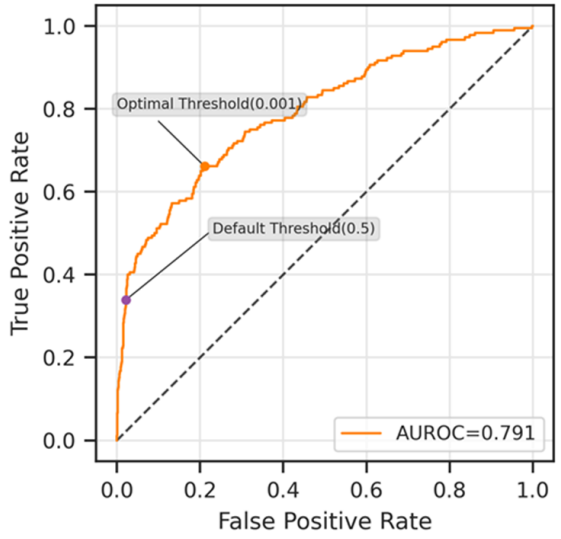


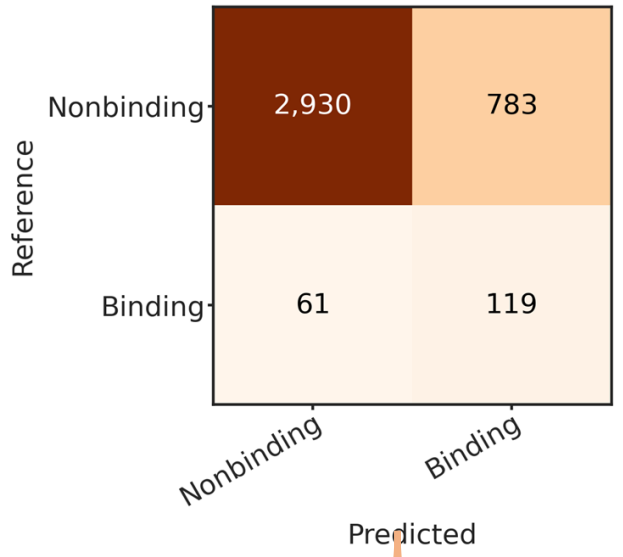
Sensitivity:0.507; PPV:0.537
specificity:0.674; NPV:0.646

Independent Testing Results

On ***Davis*** dataset with DeepLPI-6165

All Images on this page created by the presenter

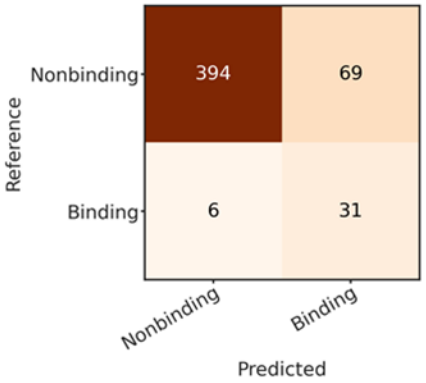




Overall
confusion matrix

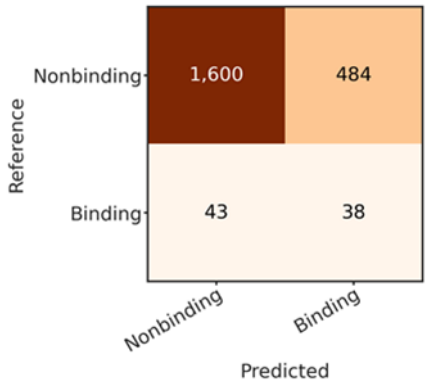
AUROC = 0.791
Sensitivity:0.661
specificity:0.789
PPV:0.132
NPV:0.980
(based on optimal threshold)

Both seen, AUROC = 0.844



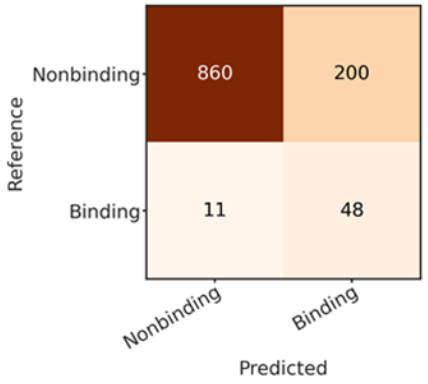
Sensitivity:0.838; PPV:0.310
specificity:0.851; NPV:0.985

Molecule unseen, AUROC = 0. 618



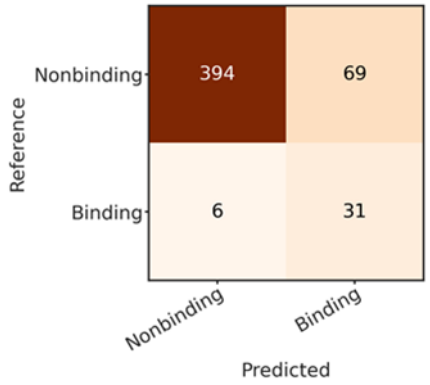
Sensitivity:0.469; PPV:0.073
specificity:0.768; NPV:0.974

Protein unseen, AUROC = 0.812



Sensitivity:0.814; PPV:0.194
specificity:0.811; NPV:0.987

None seen, AUROC = 0.692

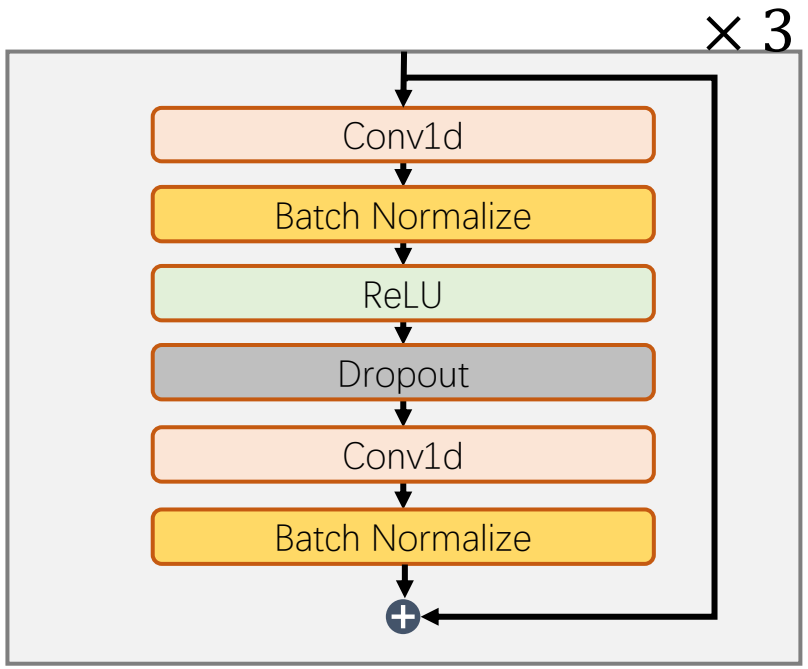


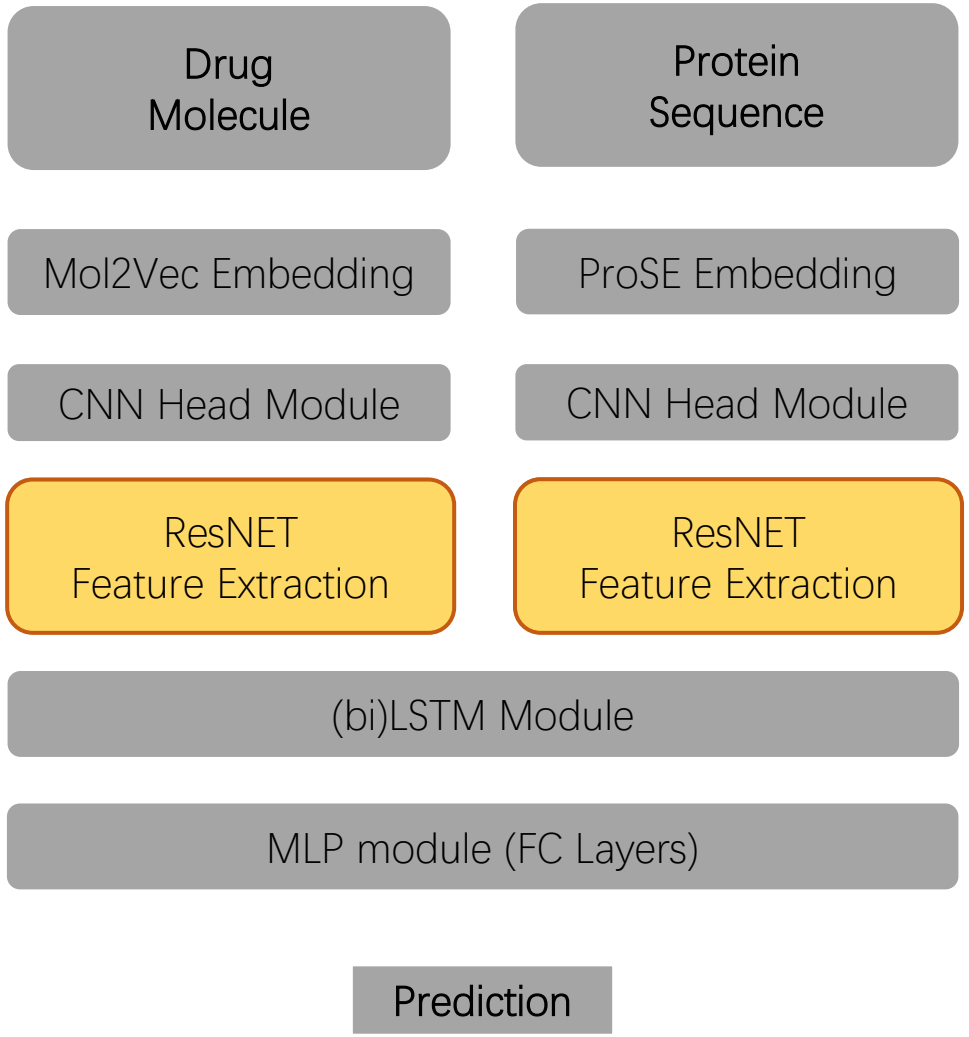
Sensitivity:0.667; PPV:0.062
specificity:0.717; NPV:0.987

Model Method

ResNet Based CNN Module

-- Main Information layer, to extract information from vector

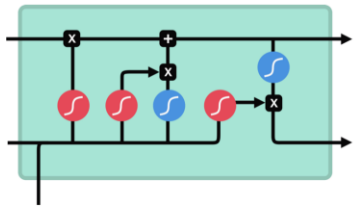


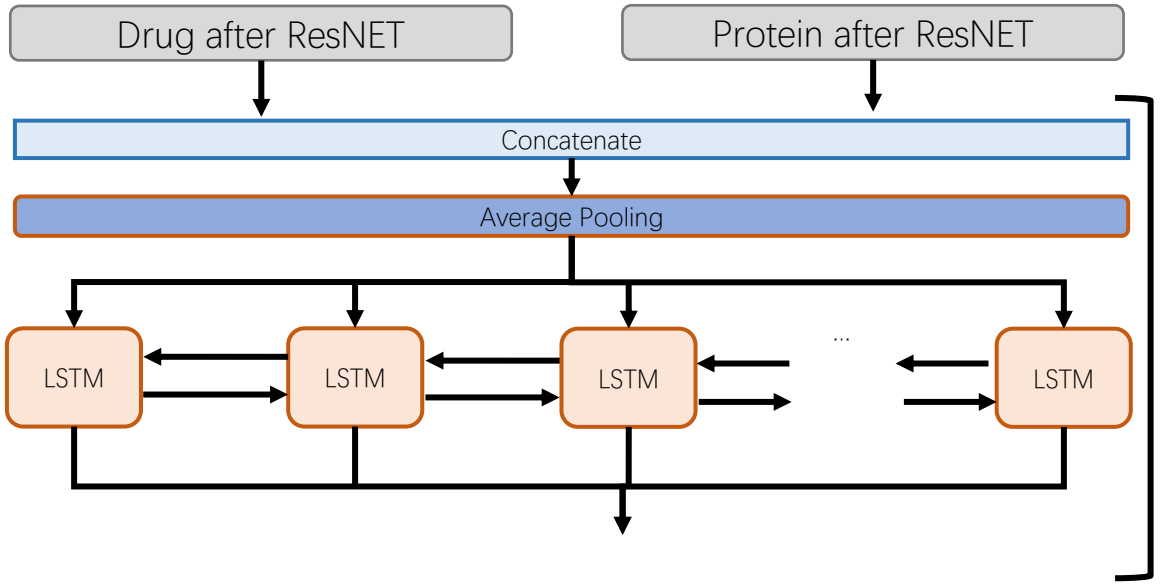


Model Method

biLSTM

-- combination and encoding module





Drug Molecule

Protein Sequence

Mol2Vec Embedding

ProSE Embedding

CNN Head Module

CNN Head Module

ResNET Feature Extraction

ResNET Feature Extraction

(bi)LSTM Module

MLP module (FC Layers)

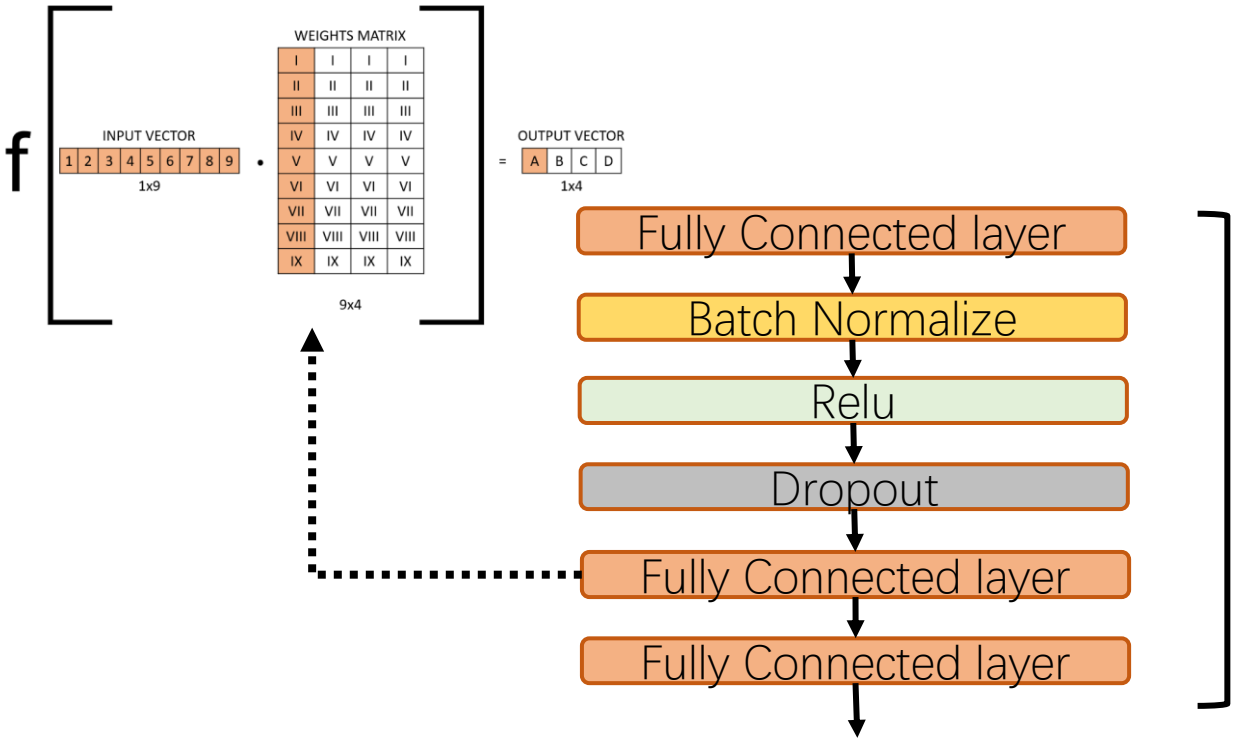
Prediction

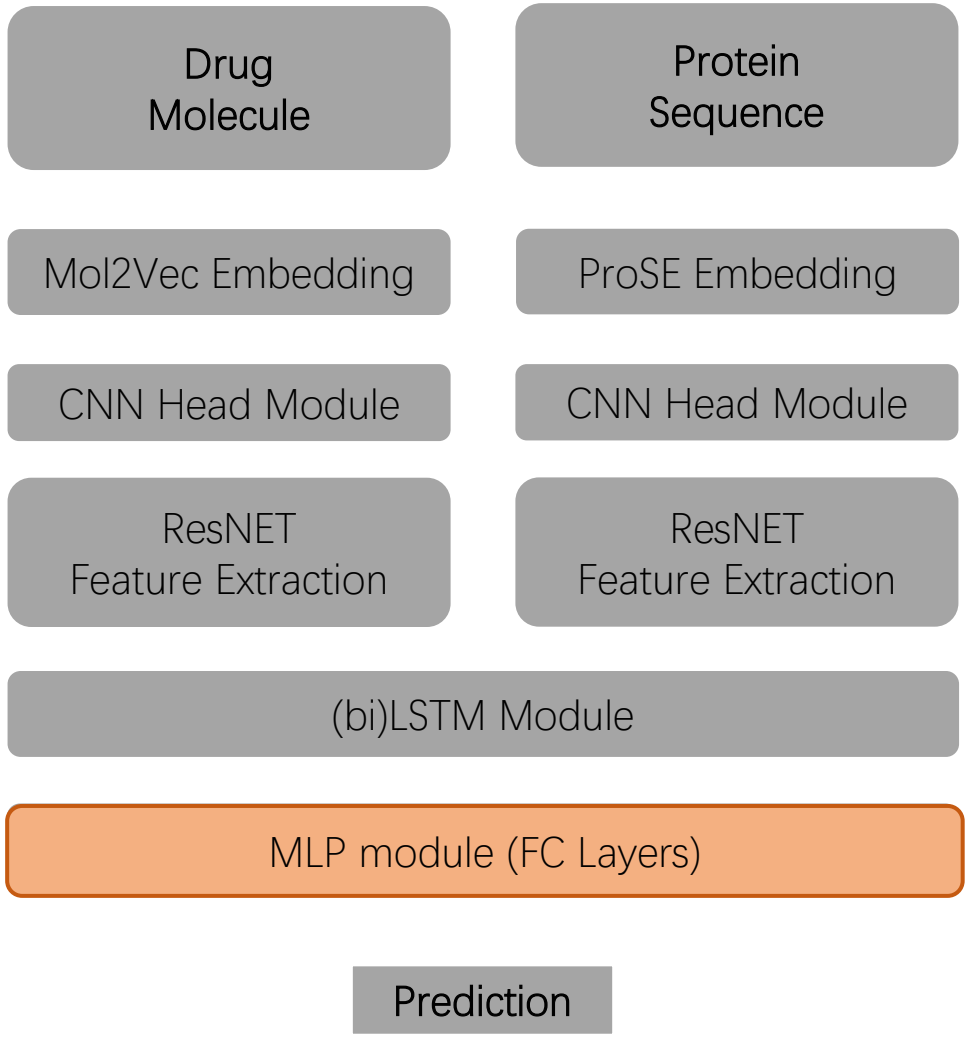
https://towardsdatascience.com/illustrated-guide-to-lstms-and-gru-s-a-step-by-step-explanation-44e9eb85bf21

Model Method

FC layers

- output the prediction value





https://towardsdatascience.com/convolutional-layers-vs-fully-connected-layers-364f05ab460b

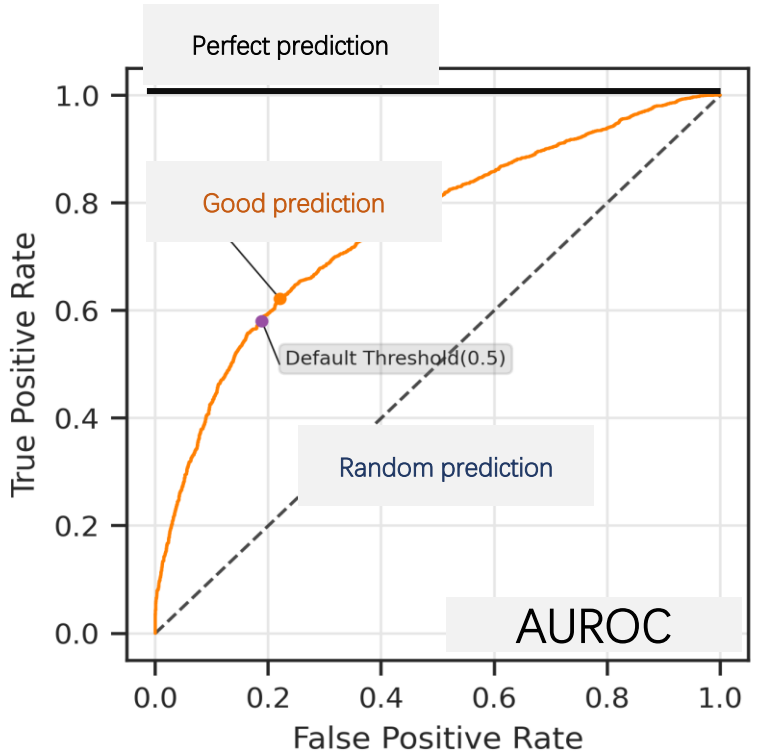
Evaluation Metrics for Classification (binding/non-binding)

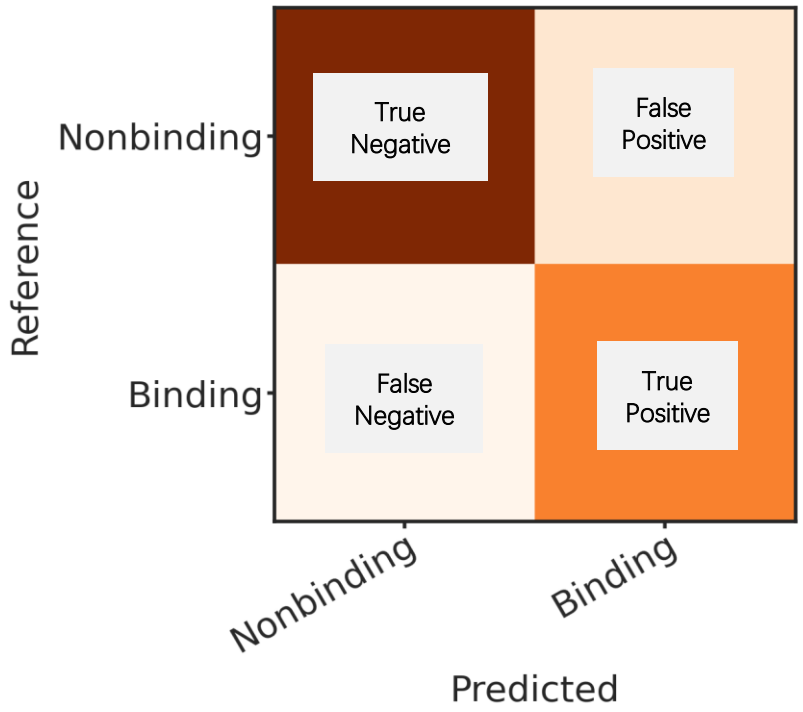
AUROC

- Area Under the Receiver Operating Characteristic Curve
- Measures the probability that a randomly selected “1” will have a higher predicted probability of being a “1”

Confusion Matrix

- Sensitivity (Recall) = TP / (TP + FN)
- Specificity = TN / (TN + FP)
- PPV (Precision) = TP / (TP + FP)
- NPV = TN / (FN + TN)





Result – Datasets Overview

BindingDB

- 2.3 million experiment data, general purpose
- 36,111 data w/ required label after quality screening
- Kd <= 100 nM : Strong Binding; Kd > 100 nM, Non-Binding
- Kd < 10000 nM: Weak Binding; Kd > 10000 nM, Non-Binding

Davis

- 30,056 experiment data, general purpose
- 24,548 non-duplicated data
- Kd <= 100 nM : Strong Binding; Kd > 100 nM, Non-Binding
- Kd < 10000 nM: Weak Binding; Kd > 10000 nM, Non-Binding

COVID-19

- 850 experiment data entry
- Binary: active or inactive

Model Performance Test

2 Public Training Datasets

BindingDB | Davis

3 Models' Comparison

DeepPLA -6165 | DeepPLA -100 | DeepCDA

Best Literature Model

Transfer-ability Test

BindingDB → COVID Data

BindingDB → Davis

Performance Metric (Classification)

AUROC | Specificity, Sensitivity, PPV, NPV

Results:

External Test Dataset of COVID-19 Stats

Description:
A large XChem crystallographic fragment screen against SARS-CoV-2 main protease at high resolution. From MIT AiCures.

Entry:
879 Drug entries

Target:
SARS-CoV-2 3CL Protease

Data sources:

Diamond Light Source:
https://www.diamond.ac.uk/covid-19/for-scientists/Main-protease-structure-and-XChem.html

MIT AI Cure:
https://www.aicures.mit.edu/data

Therapeutic Data Commons:
https://tdcommons.ai/single_pred_tasks/hts/#sars-cov-2-3cl-protease-diamond

	Drug_ID	Drug	Y
0	0	Oc1ccccc1CNc1nc2ccccc2[nH]1	1
1	1	CC(=O)NCCc1c[nH]c2ccc(F)cc12	1
2	2	NC(=O)[C@H]1CCC[C@H]1c1ccsc1	1
3	3	CN1CCCc2ccc(S(N)(=O)=O)cc21	1
4	4	CC(=O)Nc1ccc(Oc2ncccn2)cc1	1
...	...	...	...
875	875	CC(C)c1ccc(NC(=O)N2CCOCC2)cc1	0
876	876	CN(CC(=O)O)C(=O)c1ccccn1	0
877	877	CN1CCN(C(=O)c2ccc(F)c(F)c2)CC1	0
878	878	O=C(c1c(F)cccc1F)N1CCCCCC1	0
879	879	Fc1cccnc1NCC1CCOCC1	0