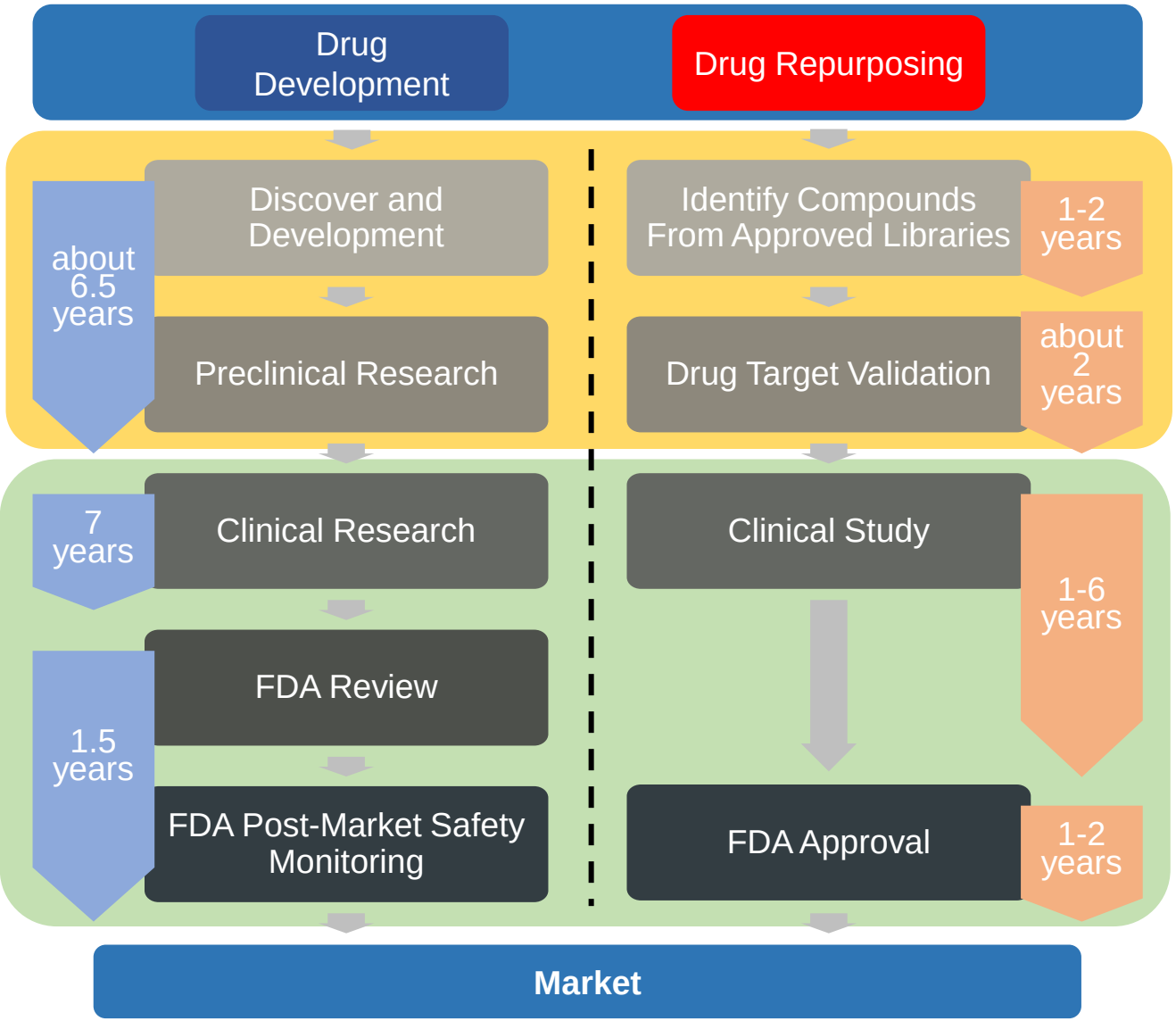
Drug repurposing significantly reduces drug discovery time.



- Development cycle
  - Traditional Drug Development: ~ 13 years[1].
  - Drug repurposing: ~6 years.

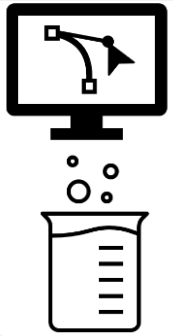
- Orange – physical or *in-silico* experiments.
- Green – animal and human experiments.

Data from Ref [1], illustration by the presenter

[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**

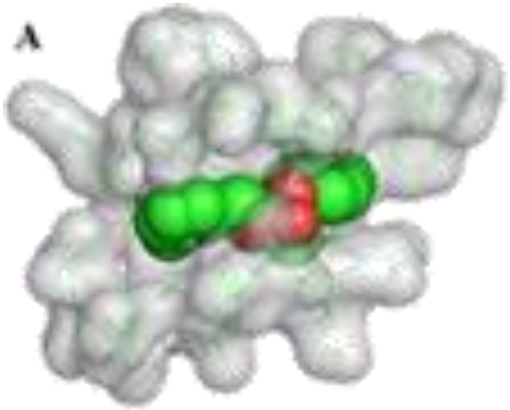


Image from https://doi.org/10.1186/s12860-020-00294-x

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 Drug/Protein Spatial information

- **Fast but inaccurate**

Traditional Machine Learning Methods: Using human selecting features to do the prediction, the precision of prediction are not sufficient for finding potential drugs.

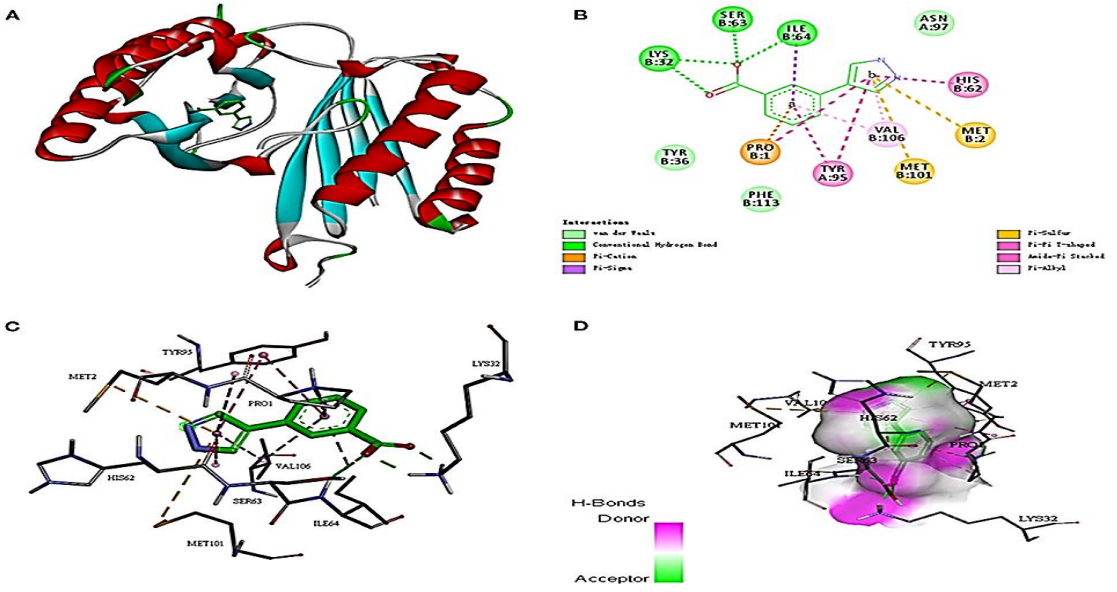
- **Accurate but limited/rare**

3D-Structure-Based Models:
High accuracy but the model may cannot be deployed into real-life situation.

Research Question: Can we use simpler input (drug/protein) information to make the model both fast and accurate?

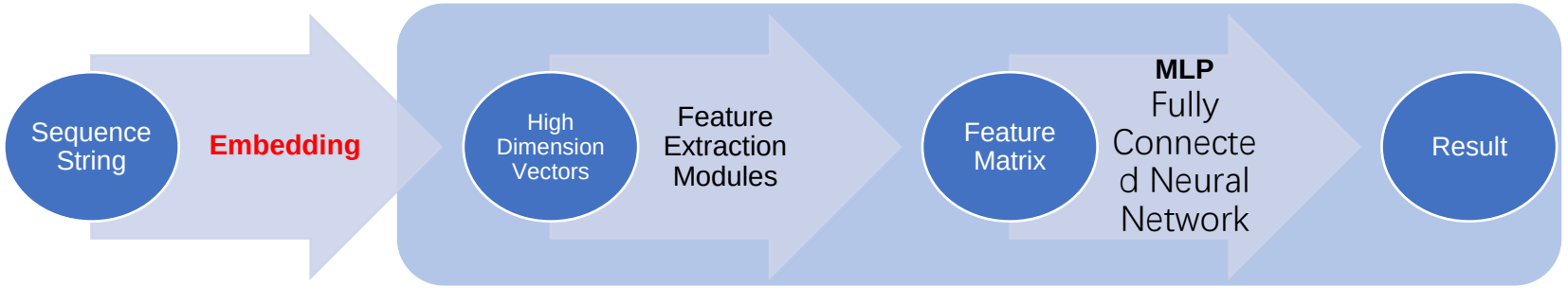
	FNN	SVM	RF	KNN
StaticF	0.687 ± 0.131	0.668 ± 0.128	0.665 ± 0.125	0.624 ± 0.120
SemiF	0.743 ± 0.124	0.704 ± 0.128	0.701 ± 0.119	0.660 ± 0.119
ECFP6	0.724 ± 0.125	0.715 ± 0.127	0.679 ± 0.128	0.669 ± 0.121
DFS8	0.707 ± 0.129	0.693 ± 0.128	0.689 ± 0.120	0.648 ± 0.120
ECFP6 + ToxF	0.731 ± 0.126	0.722 ± 0.126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

Tranditional 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



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

Our proposed model DeepLPI (Ligand-Protein Interaction)



Illustrations by the presenter

Our Model: Treat drug and protein as language and adapt NLP techniques for DTI.

Traditional	• use drug/protein spatial information, **complex**
Our model	• use drug formula/protein sequence, **simple**

DeepLPI model overview (best after 9 versions)

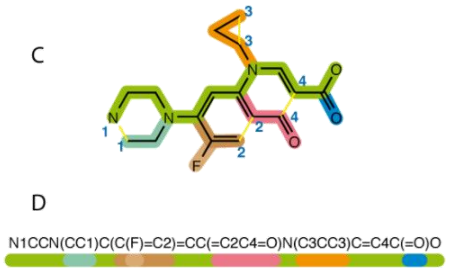
Illustrations by the presenter

Common Setup

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

Input

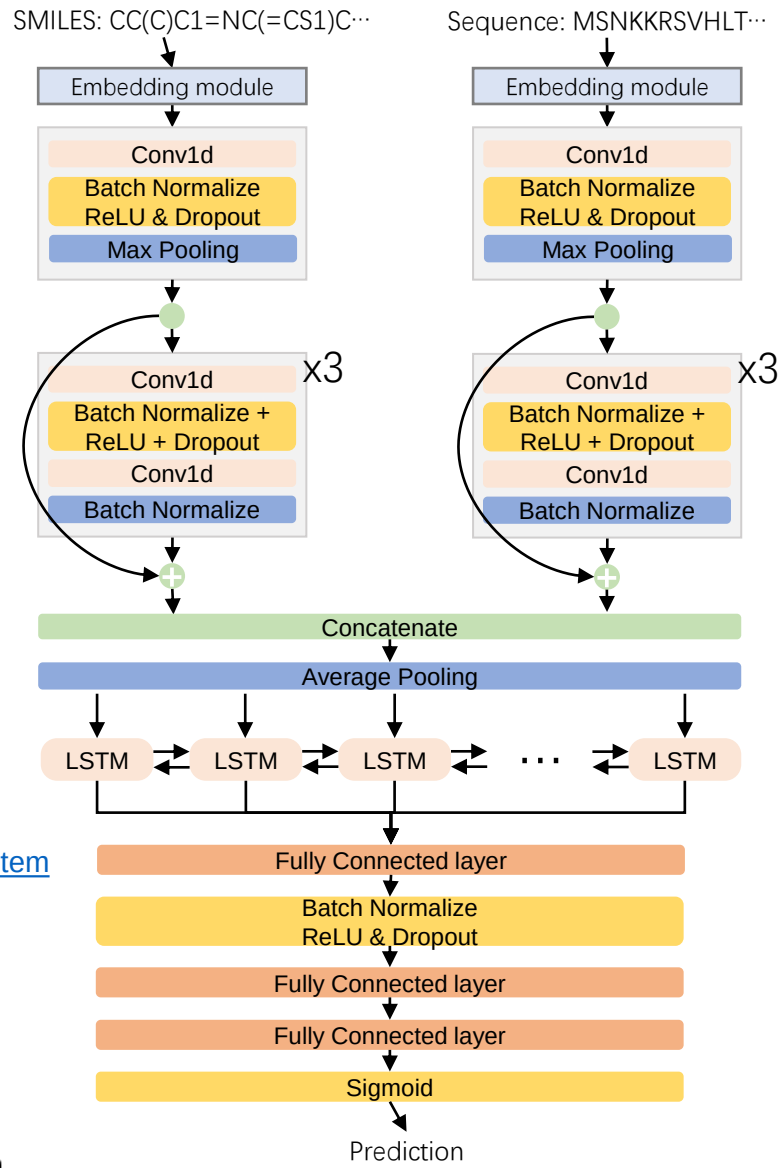
Drug Molecule -- SMILES format



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

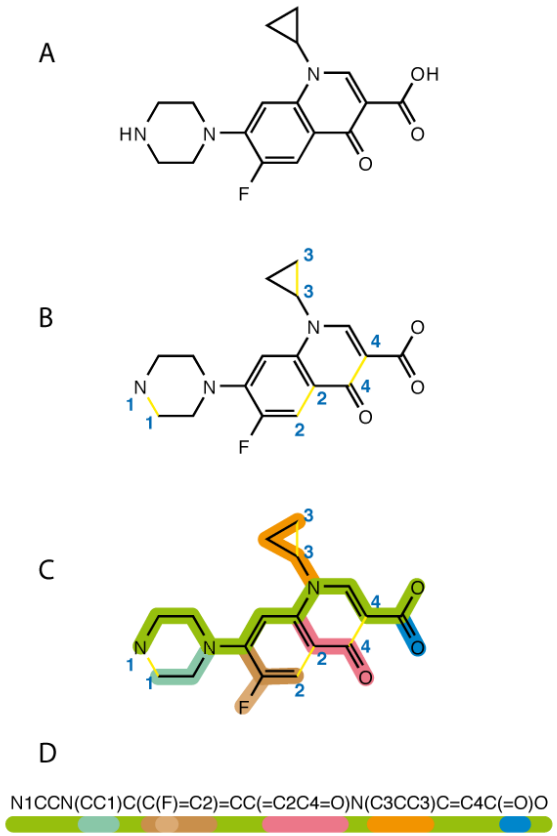


Loss Function = Binary Cross Entropy + L2 regularization

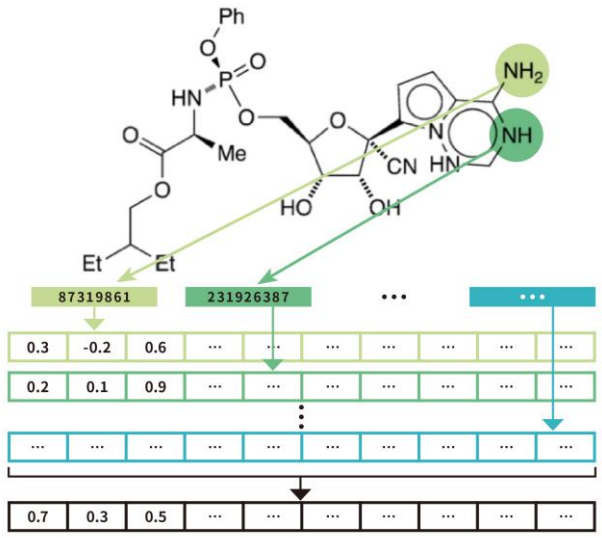
$$\text{Loss} = \underbrace{-\frac{1}{N}\sum_{i=1}^{N}[y_i \cdot \log(\hat{y}_i) + (1-y_i)\cdot\log(1-\hat{y}_i)]}_{\text{BCE loss}} + \underbrace{\alpha\|W\|_2^2}_{\text{L2-norm regularization}}$$

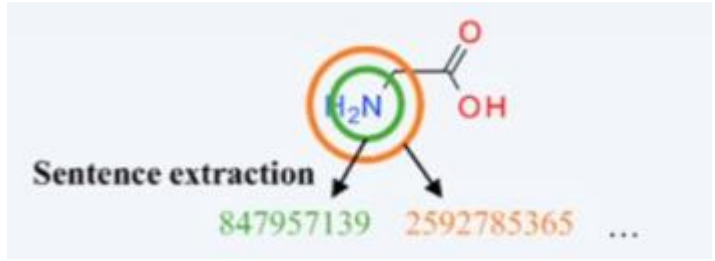
Molecule Embedding by Mol2Vec

Image from Ref. [2]



Represent molecules with a sequence: SMILES format





Extract molecule fingerprint in all loci and turn into words

Construct a sentence

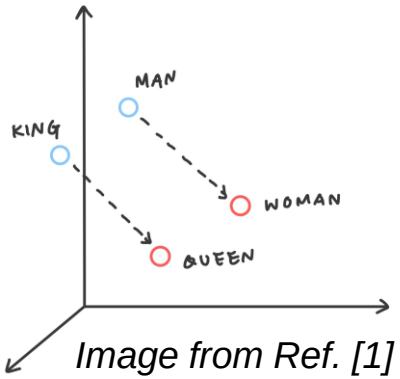


Image from Ref. [1]

Train/Apply machine learning model for high-D representation

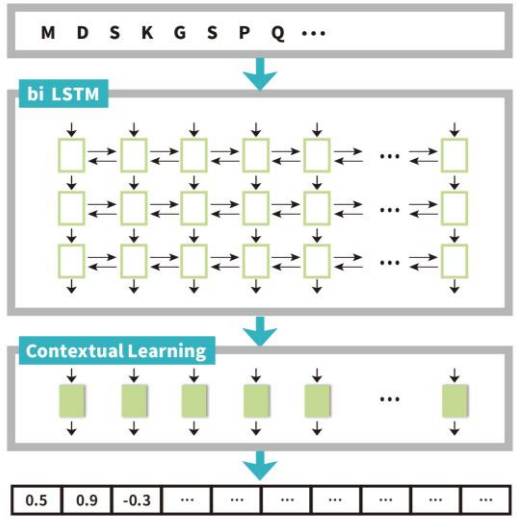
Example of drug molecule after embedding

id	Mol	0	1	2	3	4	5	6	7	8	...	290	
0	COc1ccc(Br)c2C[C@@H]3[C@H](C[C@H](CN3C)C(=O)N3...	0.352985	-0.415436	0.212324	-0.093876	-0.937583	0.117071	0.137606	0.924334	-0.643350	...	0.539983	0.32
1	CC(C)C[C@H](NC(=O)CNC(=O)CNC(=O)[C@H](Cc1ccc(C...	-2.158546	-1.040054	-1.089538	-1.080631	2.828738	-2.466355	-2.914023	-3.128932	3.317803	...	1.728877	2.30
2	OC(C(=O)O[C@H]1CN2CCC1CC2)	-0.167114	0.315020	0.241171	0.109566	-0.769069	-0.153241	0.331229	0.322230	-0.286682	...	-0.436014	-0.17

[1] S. Jaeger, S. Fulle, and S. Turk, "Mol2vec: Unsupervised Machine Learning Approach with Chemical Intuition."

[2] https://en.wikipedia.org/wiki/Simplified_molecular-input_line-entry_system

Protein Embedding by ProSE



We tested a few embedding methods, including AllenNLP, SeqVec, and etc, and finally decided to use **ProSE** to embed the protein sequence in FASTA format and obtained both a 6165-dimension embedded vector, and a 100-dimension embedded vector.

We test both embedding for our model because the longer vecter might contain more useful information but may potentially lead to overfitting, while the shorter vector might lose some information but could suppress overfitting and run at faster speed.

Therefore, we have two models: **DeepLPI-6165** and **DeepLPI-100**.

ProSE maximizes global similary and residue contacts between proteins

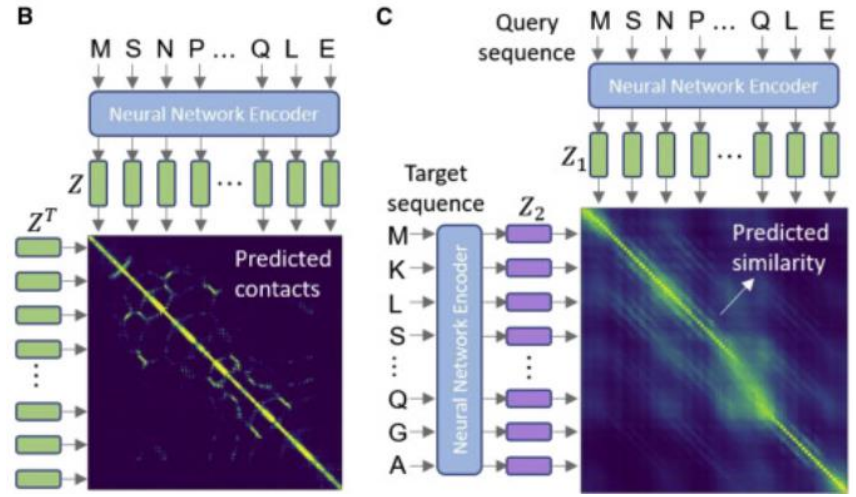


Image from Ref. [1]

Example of Protein after embedding

id	Seq	0	1	2	3	4	5	6	7
1	MKKFFDSRREQGGSGLGSGSSGGGGSTSGLGSGYIGRVFGIGRQQV...	0.066187	0.105748	-0.170256	0.108832	0.120057	-0.012809	0.099857	0.040916
2	PFWKILNPLLERGTYYYFMGQQPGKVLGDQRRPSLPALHFIKGAGK...	0.053404	0.007162	-0.109380	0.046826	0.004983	-0.014588	0.033338	-0.006568
3	PFWKILNPLLERGTYYYFMGQQPGKVLGDQRRPSLPALHFIKGAGK...	0.053404	0.007162	-0.109380	0.046826	0.004983	-0.014588	0.033338	-0.006568
4	PFWKILNPLLERGTYYYFMGQQPGKVLGDQRRPSLPALHFIKGAGK...	0.053404	0.007162	-0.109380	0.046826	0.004983	-0.014588	0.033338	-0.006568
5	PFWKILNPLLERGTYYYFMGQQPGKVLGDQRRPSLPALHFIKGAGK...	0.053404	0.007162	-0.109380	0.046826	0.004983	-0.014588	0.033338	-0.006568

Image and Data created by the presenter

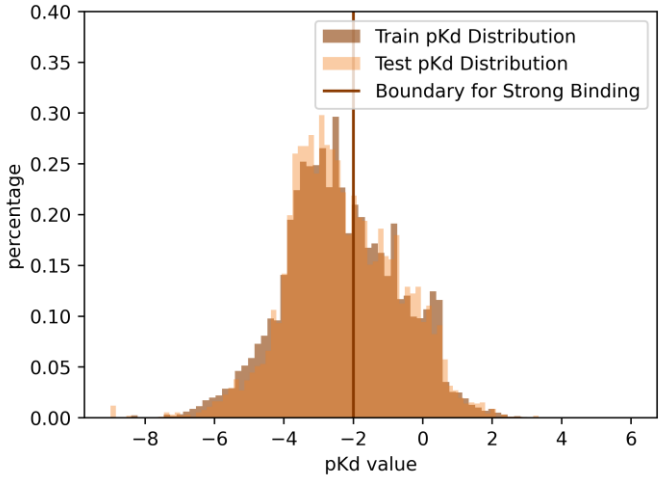
[1] T. Bepler and B. Berger, "Learning the protein language: Evolution, structure, and function," *Cell Systems*, vol. 12, no. 6, (2021)

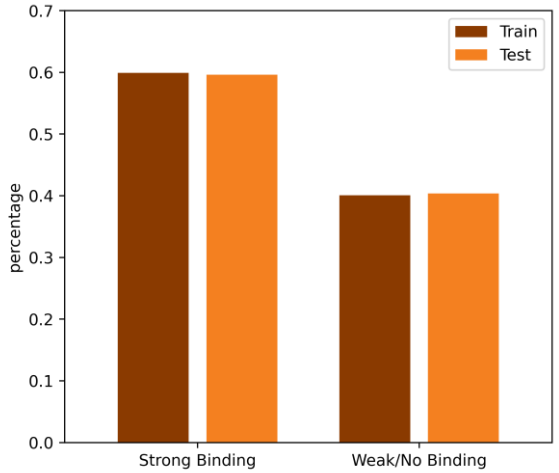
Train Data Selection

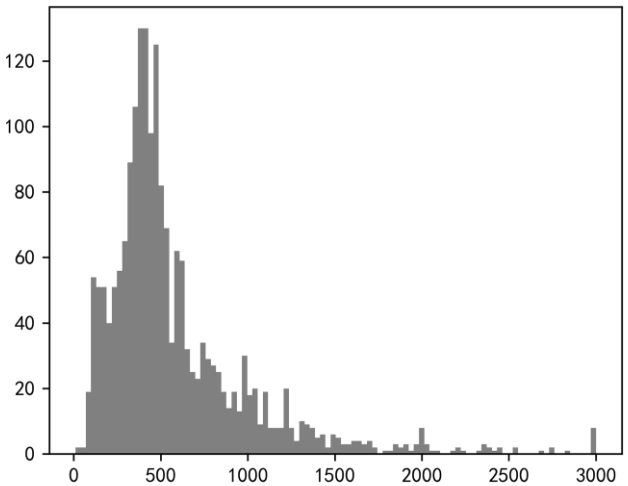
No duplicates, High Confidence Experiment, Balanced Label

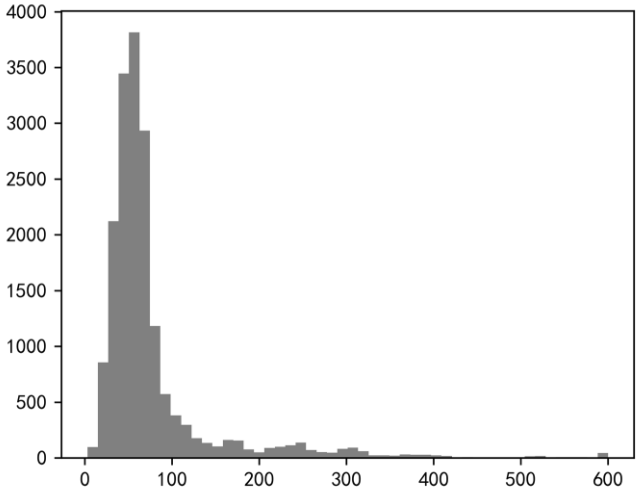
All Images on this page created by the presenter

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









Original
Total: 2,333,459
Drug:1,007,991 Seq:8091

↓

Remove Multichain
Total: 2,221,864
Drug: 964,811 Seq: 7933

↓

Choose Only Kd Label
Total: 91,214
Drug:19,974 Seq:1939

↓

Remove Invalid Value
Total: 52,390
Drug:17,773 Seq:1915

↓

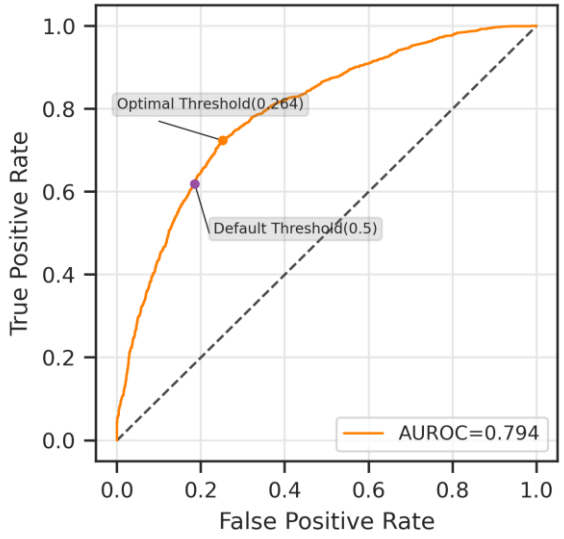
Remove Duplicates
Total: 36,111
Drug:17,773 Seq:1915

Convert to 0-1 Classification

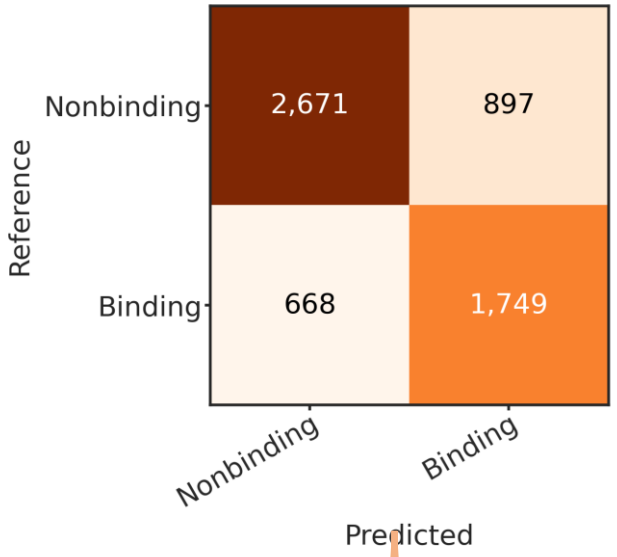
Independent Testing Results

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

All Images on this page created by the presenter



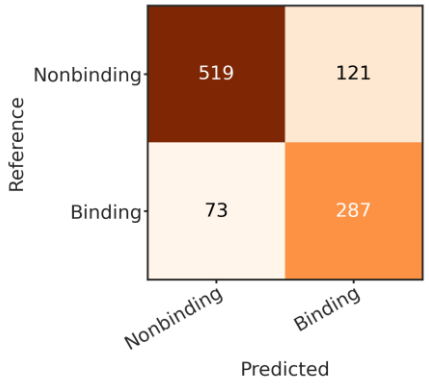
From the AUROC curve, we determine an optimal threshold for dividing the predicted Y values into binary (0/1) binding/non-binding values.



Overall confusion matrix

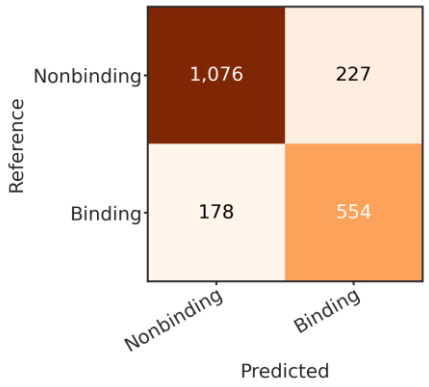
AUROC = 0.794
Sensitivity:0.724
specificity:0.749
PPV:0.661
NPV:0.800
(based on optimal threshold)

Both seen, AUROC = 0.877



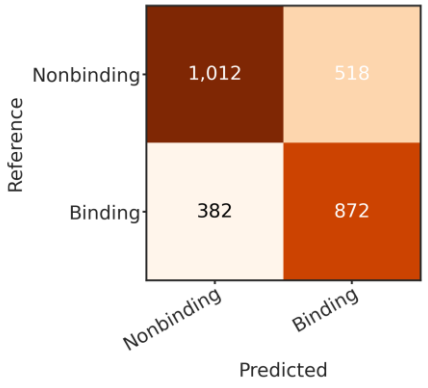
Sensitivity:0.797; PPV:0.703
specificity:0.811; NPV:0.877

Molecule unseen, AUROC = 0.857



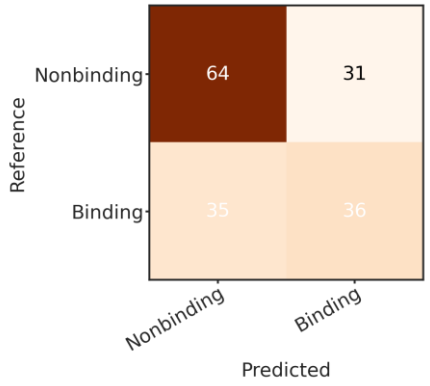
Sensitivity:0.757; PPV:0.709
specificity:0.826; NPV:0.858

Protein unseen, AUROC = 0.718



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

None seen, AUROC = 0.655

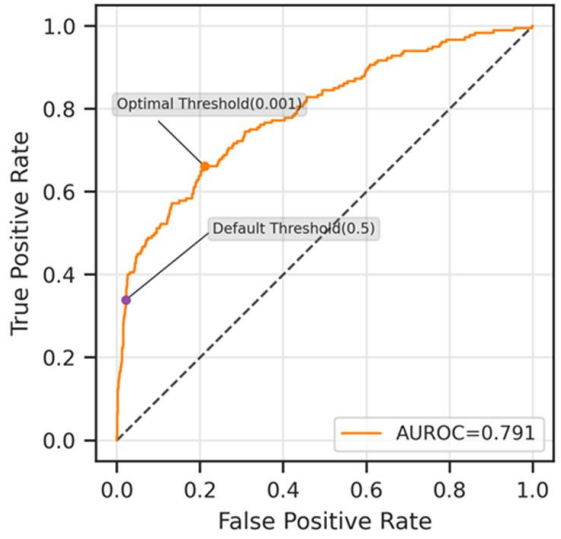


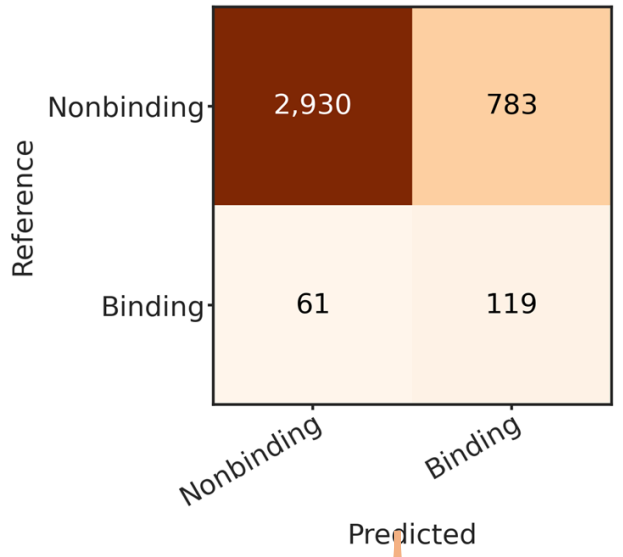
Sensitivity:0.507; PPV:0.537
specificity:0.674; NPV:0.646

Independent Testing Results

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

All Images on this page created by the presenter

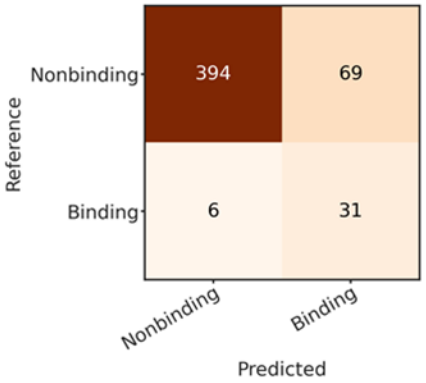




Overall
confusion matrix

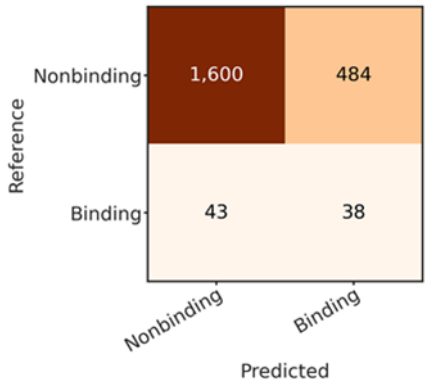
AUROC = 0.791
Sensitivity:0.661
specificity:0.789
PPV:0.132
NPV:0.980
(based on optimal threshold)

Both seen, AUROC = 0.844



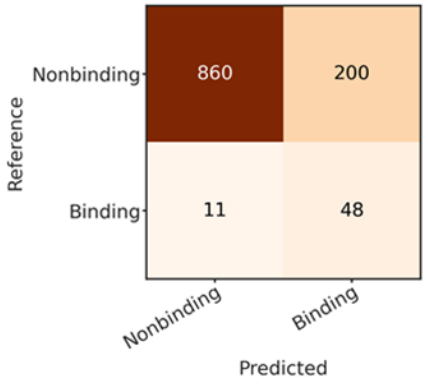
Sensitivity:0.838; PPV:0.310
specificity:0.851; NPV:0.985

Molecule unseen, AUROC = 0. 618



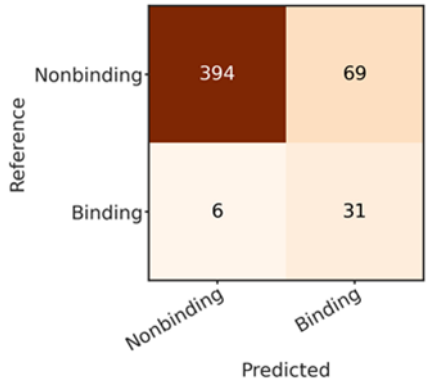
Sensitivity:0.469; PPV:0.073
specificity:0.768; NPV:0.974

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

Performance Comparison

BindingDB	AUROC	Sensitivity	Specificity	PPV	NPV	Remark
Our 6165	**0.790**	**0.684**	0.773	**0.671**	**0.783**	
Our 100	0.751	0.541	0.818	0.668	0.725	
DeepCDA	0.448	0.000	**1.000**	Nan	0.596	All nonbinding
Transfer to COVID Data						
Our 6165	**0.610**	0.538	0.576	0.110	0.928	
Our 100	0.475	0.692	0.332	0.092	0.912	
DeepCDA	0.400	0.000	1.0	nan	0.911	All nonbinding

Davis	AUROC	Sensitivity	Specificity	PPV	NPV	Remark
Our 6165	**0.791**	0.661	0.789	0.132	0.980	
Our 100	0.673	0.395	0.820	0.439	0.791	
DeepCDA	0.741	0.511	0.813	0.495	0.823	
Transfer to COVID Data						
Our 6165	**0.534**	0.000	1.000	nan	0.911	All nonbinding
Our 100	0.482	0.040	1.000	1	0.914	
DeepCDA	0.413	0.000	1.000	nan	0.911	All nonbinding

Result Summary

Use 1-dimension drug SMILES and protein sequence as input

Use NLP technique to treat drug and protein

Model DeepLPI-6165 performance in classification on BindingDB dataset is **76% better** than the state-of-the-art DeepCDA model

Model DeepLPI-6165 performance in classification on transferability is 25% (Davis to Covid) and 50% (BindingDB to Covid) better than DeepCDA