Final Project

MolDesigner UI and Model Implementation

Kotini, Jessica

CSCI E-104 Advanced Deep Learning, 2024

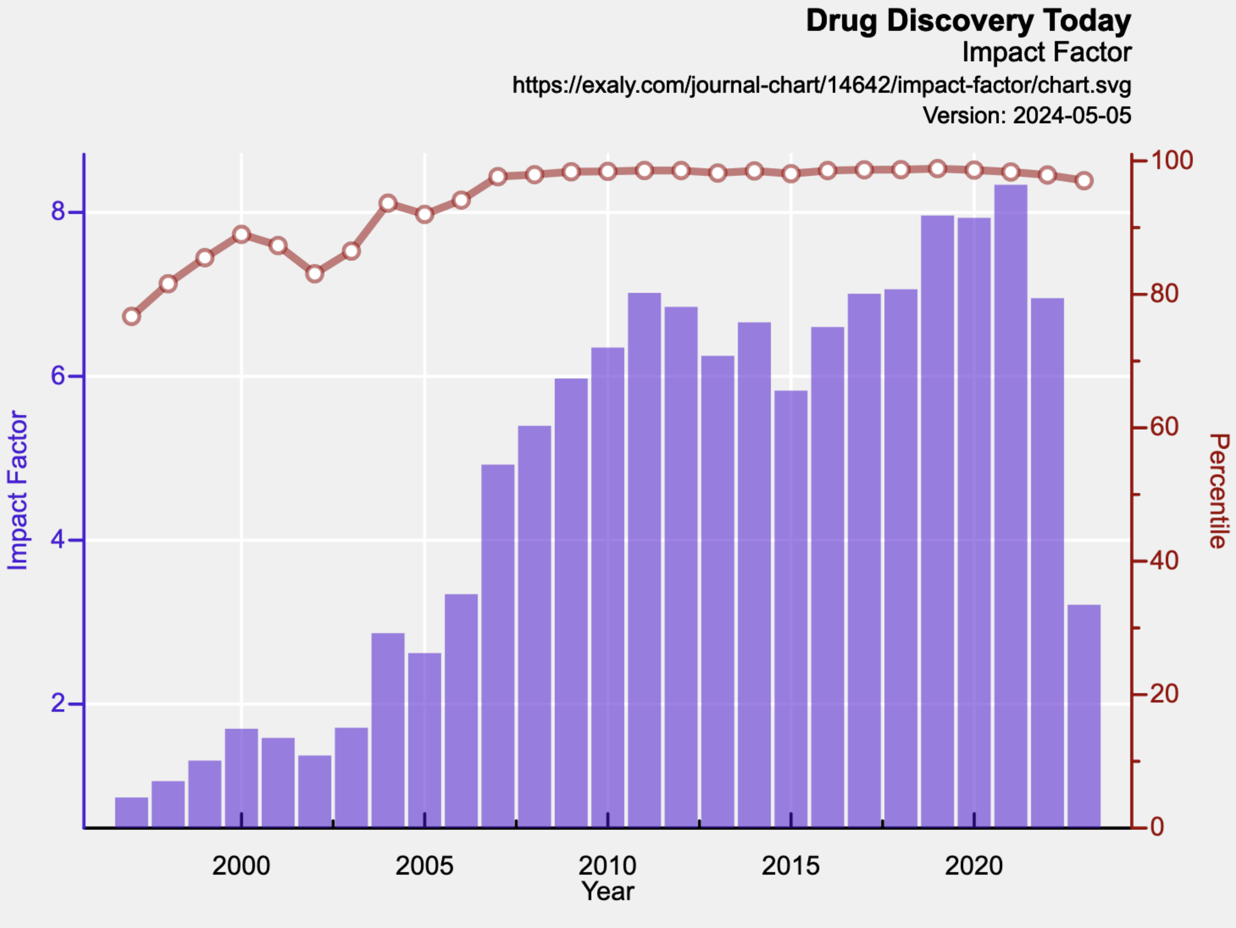
Harvard University Extension School

Prof. Zoran B. Djordjević

Introduction

- Field of drug-discovery is rapidly evolving

- Need tools for fast visualization of drug-target interaction data

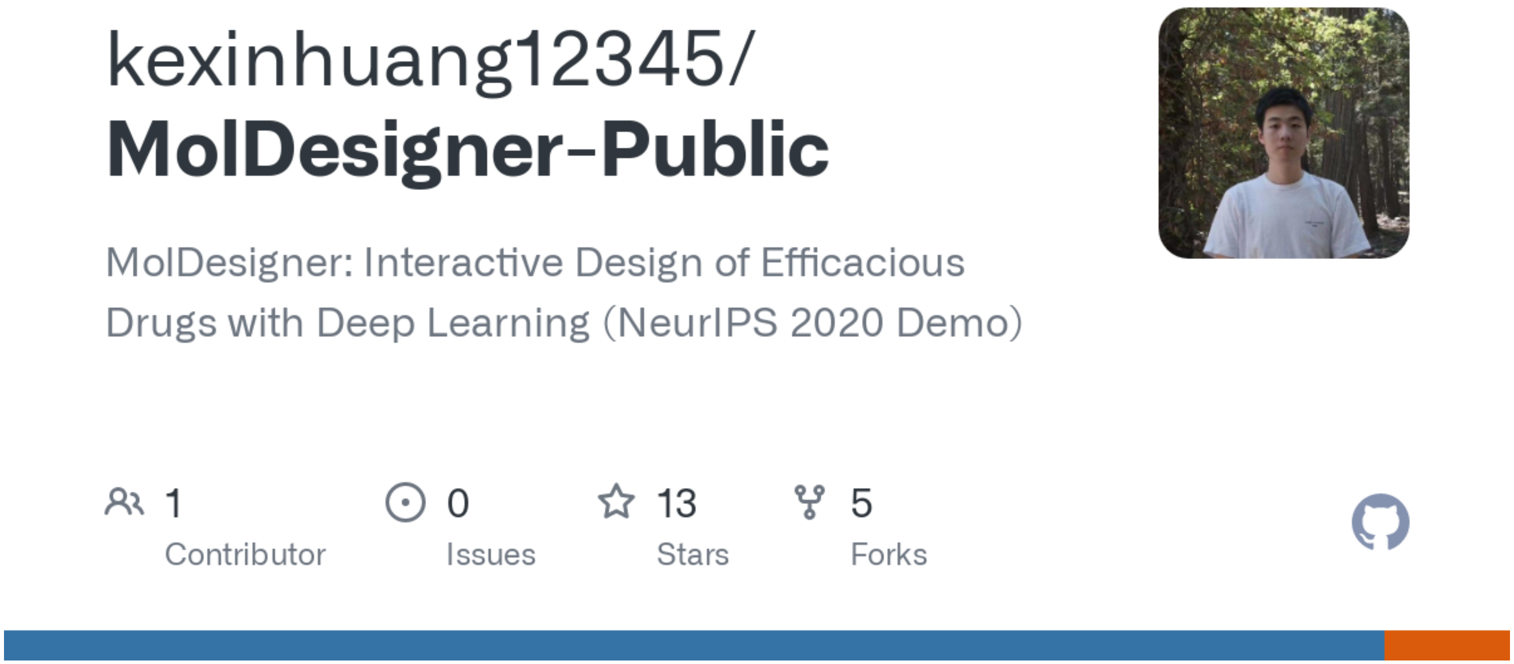


The graph shows the changes in the impact factor of **Drug Discovery Today** and its corresponding percentile for the sake of comparison with the entire literature. Impact Factor is the most common scientometric index, which is defined by the number of citations of papers in two preceding years divided by the number of papers published in those years.

https://exaly.com/journal/14642/drug-discovery-today/

Goal

1. Demonstrate utility of MolDesigner
2. Integrate my own models into MolDesigner

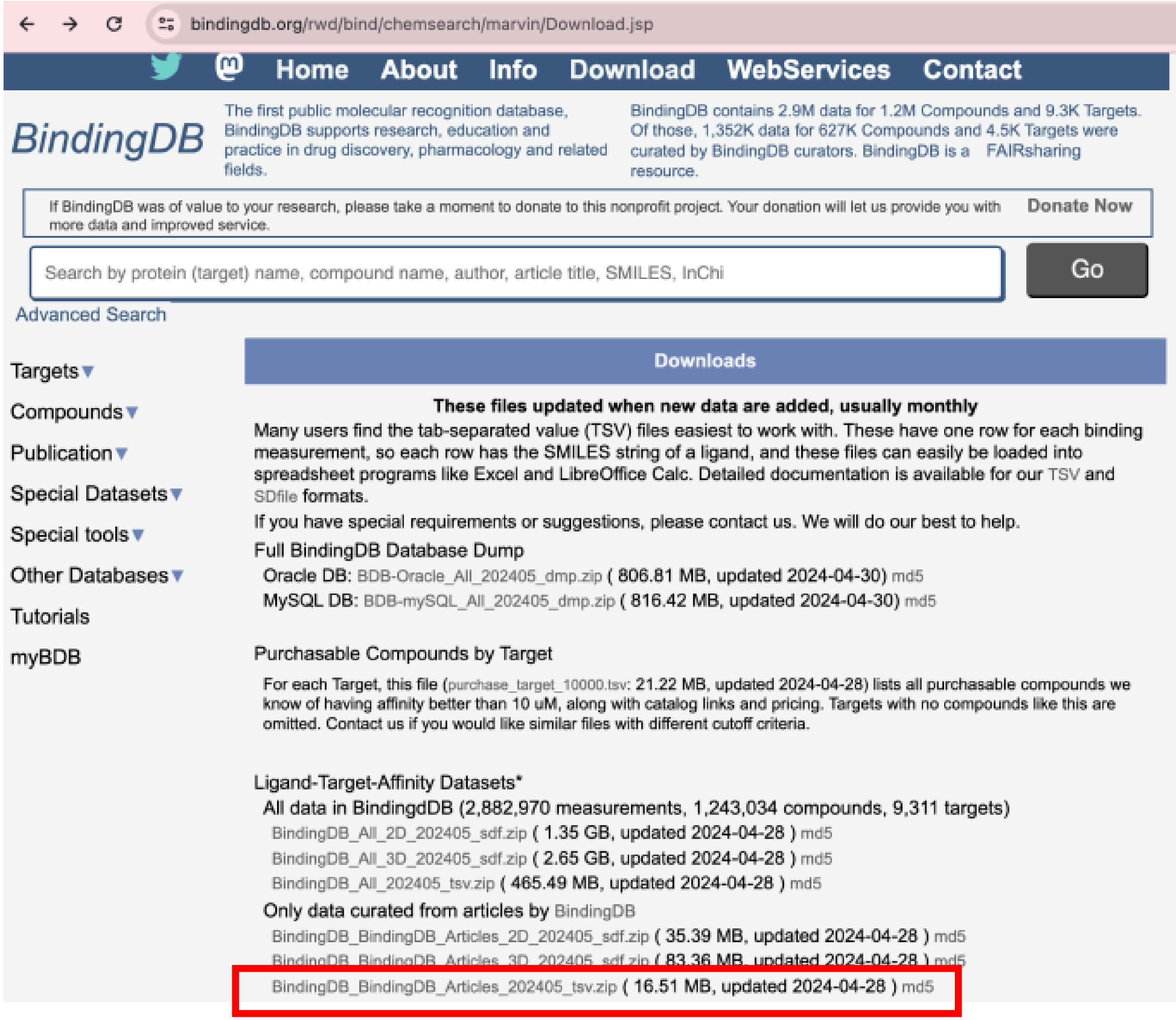


kexinhuang12345/
MolDesigner-Public

MolDesigner: Interactive Design of Efficacious Drugs with Deep Learning (NeurIPS 2020 Demo)

1	0	13	5
Contributor	Issues	Stars	Forks

Dataset



The first public molecular recognition database, BindingDB supports research, education and practice in drug discovery, pharmacology and related fields.

BindingDB contains 2.9M data for 1.2M Compounds and 9.3K Targets. Of those, 1,352K data for 627K Compounds and 4.5K Targets were curated by BindingDB curators. BindingDB is a FAIRsharing resource.

If BindingDB was of value to your research, please take a moment to donate to this nonprofit project. Your donation will let us provide you with more data and improved service.

These files updated when new data are added, usually monthly

Many users find the tab-separated value (TSV) files easiest to work with. These have one row for each binding measurement, so each row has the SMILES string of a ligand, and these files can easily be loaded into spreadsheet programs like Excel and LibreOffice Calc. Detailed documentation is available for our TSV and SDfile formats.

If you have special requirements or suggestions, please contact us. We will do our best to help.

Full BindingDB Database Dump

Oracle DB: BDB-Oracle_All_202405_dmp.zip (806.81 MB, updated 2024-04-30) md5
MySQL DB: BDB-mySQL_All_202405_dmp.zip (816.42 MB, updated 2024-04-30) md5

Purchasable Compounds by Target

For each Target, this file (purchase_target_10000.tsv: 21.22 MB, updated 2024-04-28) lists all purchasable compounds we know of having affinity better than 10 uM, along with catalog links and pricing. Targets with no compounds like this are omitted. Contact us if you would like similar files with different cutoff criteria.

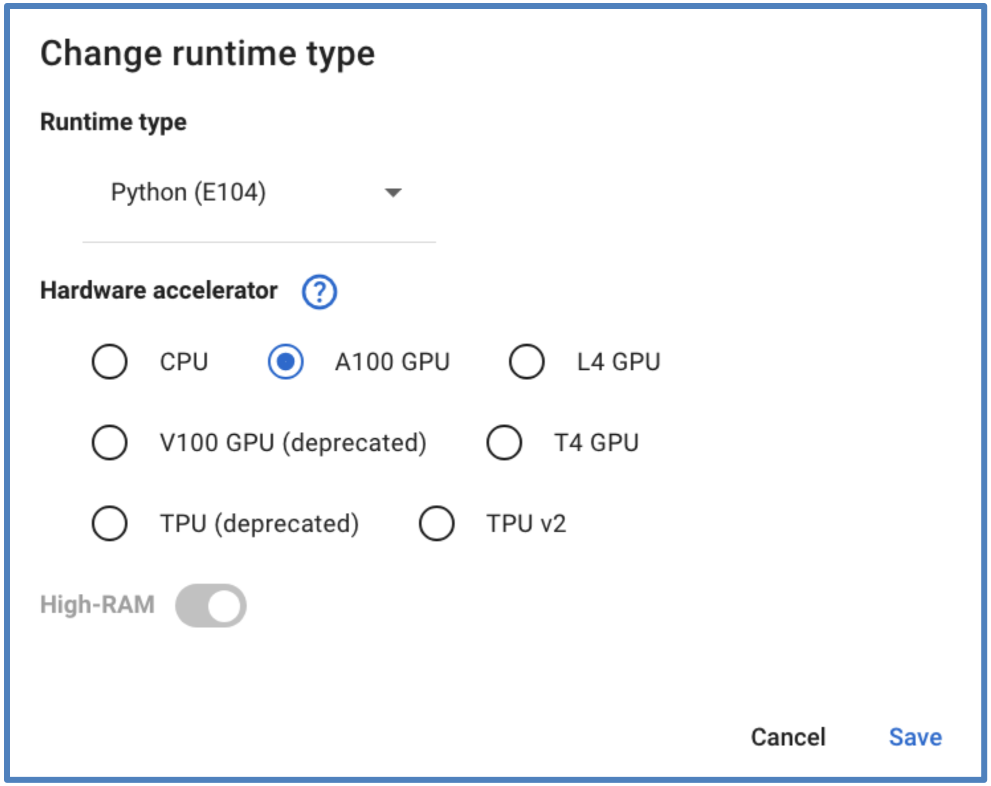
Ligand-Target-Affinity Datasets*

All data in BindingdDB (2,882,970 measurements, 1,243,034 compounds, 9,311 targets)
BindingDB_All_2D_202405_sdf.zip (1.35 GB, updated 2024-04-28) md5
BindingDB_All_3D_202405_sdf.zip (2.65 GB, updated 2024-04-28) md5
BindingDB_All_202405_tsv.zip (465.49 MB, updated 2024-04-28) md5

Only data curated from articles by BindingDB
BindingDB_BindingDB_Articles_2D_202405_sdf.zip (35.39 MB, updated 2024-04-28) md5
BindingDB_BindingDB_Articles_3D_202405_sdf.zip (83.36 MB, updated 2024-04-28) md5
BindingDB_BindingDB_Articles_202405_tsv.zip (16.51 MB, updated 2024-04-28) md5

Demo – Installation & Configuration

- Google Colab Pro+ highly recommended with A100 hardware accelerator



Change runtime type

Runtime type

Python (E104)

Hardware accelerator

○ CPU ● A100 GPU ○ L4 GPU

○ V100 GPU (deprecated) ○ T4 GPU

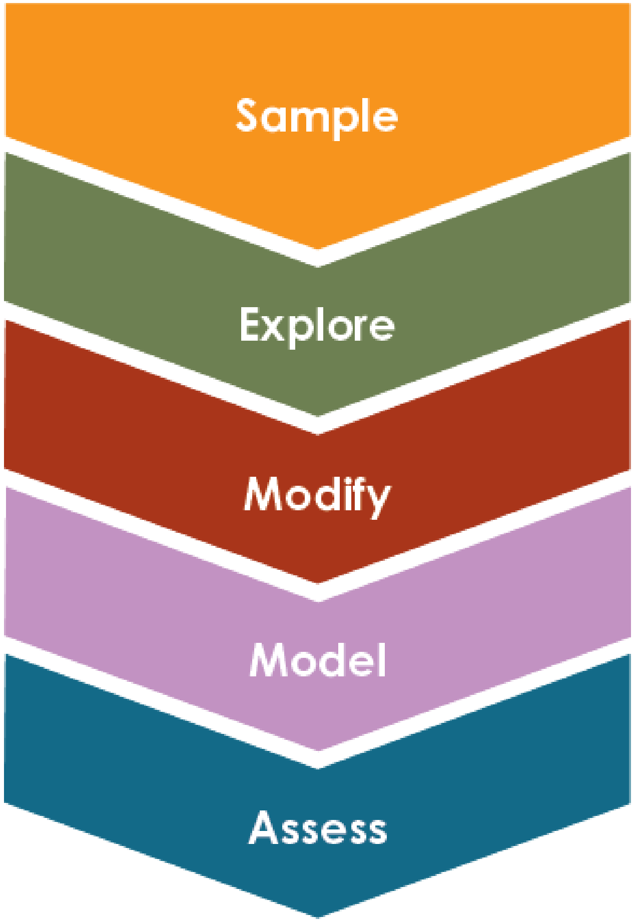
○ TPU (deprecated) ○ TPU v2

High-RAM

Demo - Methodology

- Followed a SEMMA style framework
  - Loaded, processed, sampled data
  - Explored and viewed data
  - Modelled data
  - Viewed loss curves for each iteration within and across epochs
  - Displayed predictions with Imatinib and ABL1 receptor (common drug-target pair)

Drug-Target Pairs	Unique Drugs	Unique Target Sequence
2385	1300	438



Sample

Explore

Modify

Model

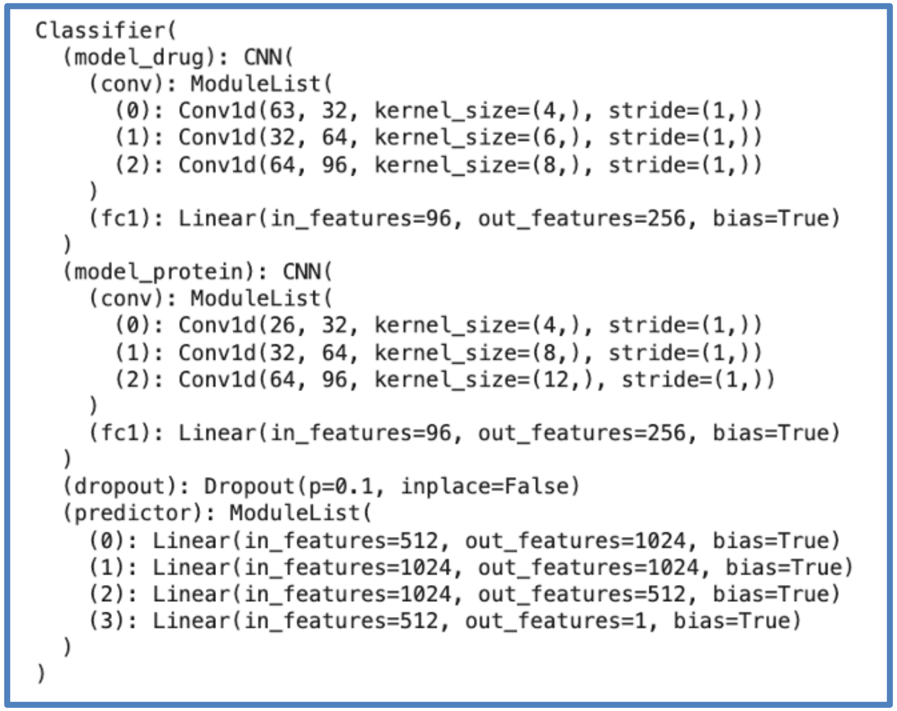
Assess

Demo Methodology

- Separate convolutional neural network (CNN) architectures for processing drugs and proteins

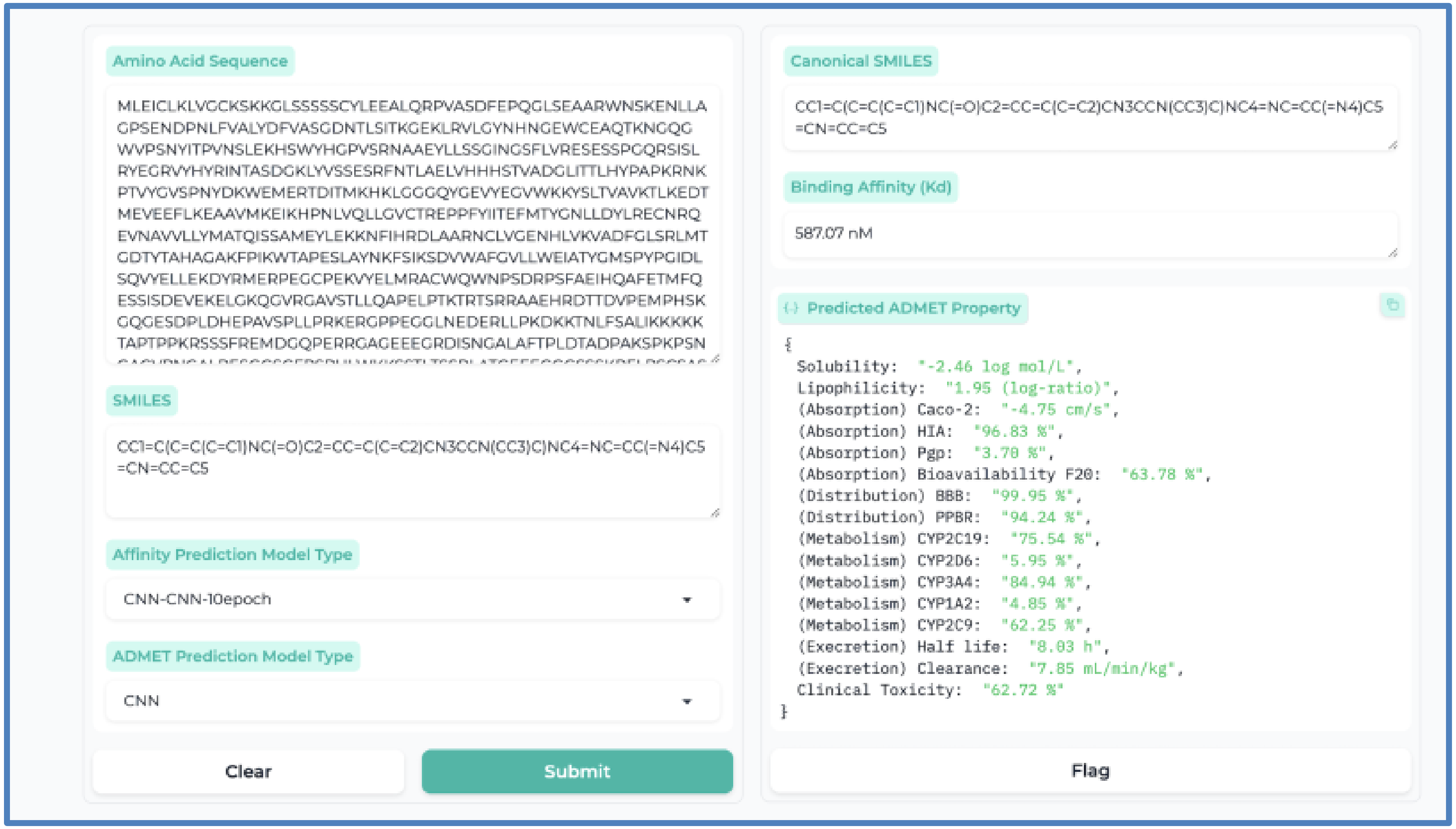
- Each segment (drug and protein) features 3 convolutional layers
  - Increasing kernel sizes, filter counts
  - Linear layer, with dropout layer before final classification layers

CNN_CNN_BindingDB



```
Classifier(
  (model_drug): CNN(
    (conv): ModuleList(
      (0): Conv1d(63, 32, kernel_size=(4,), stride=(1,))
      (1): Conv1d(32, 64, kernel_size=(6,), stride=(1,))
      (2): Conv1d(64, 96, kernel_size=(8,), stride=(1,))
    )
    (fc1): Linear(in_features=96, out_features=256, bias=True)
  )
  (model_protein): CNN(
    (conv): ModuleList(
      (0): Conv1d(26, 32, kernel_size=(4,), stride=(1,))
      (1): Conv1d(32, 64, kernel_size=(8,), stride=(1,))
      (2): Conv1d(64, 96, kernel_size=(12,), stride=(1,))
    )
    (fc1): Linear(in_features=96, out_features=256, bias=True)
  )
  (dropout): Dropout(p=0.1, inplace=False)
  (predictor): ModuleList(
    (0): Linear(in_features=512, out_features=1024, bias=True)
    (1): Linear(in_features=1024, out_features=1024, bias=True)
    (2): Linear(in_features=1024, out_features=512, bias=True)
    (3): Linear(in_features=512, out_features=1, bias=True)
  )
)
```

Results



Amino Acid Sequence

MLEICLKLVGCKSKKGLSSSSSCYLEEALQRPVASDFEPQGLSEAARWNSKENLLA
GPSENDPNLFVALYDFVASGDNTLSITKGEKLRVLGYNHNGEWCEAQTKNGQG
WVPSNYITPVNSLEKHSWYHGPVSRNAAEYLLSSGINGSFLVRESESSPGQRSISL
RYEGRVYHYRINTASDGKLYVSSESRFNTLAELVHHHSTVADGLITTLHYPAPKRNK
PTVYGVSPNYDKWEMERTDITMKHKLGGGQYGEVYEGVWKKYSLTVAVKTLKEDT
MEVEEFLKEAAVMKEIKHPNLVQLLGVCTREPPFYIITEFMTYGNLLDYLRECNRQ
EVNAVVLLYMATQISSAMEYLEKKNFIHRDLAARNCLVGENHLVKVADFGLSRLMT
GDTYTAHAGAKFPIKWTAPESLAYNKFSIKSDVWAFGVLLWEIATYGMSPYPGIDL
SQVYELLEKDYRMERPEGCPEKVYELMRACWQWNPSDRPSFAEIHQAFETMFQ
ESSISDEVEKELGKQGVRGAVSTLLQAPELPTKTRTSRRAAEHRDTTDVPEMPHSK
GQGESDPLDHEPAVSPLLPRKERGPPEGGLNEDERLLPKDKKTNLFSALIKKKKK
TAPTPPKRSSSFREMDGQPERRGAGEEEGRDISNGALAFTPLDTADPAKSPKPSN

SMILES

CC1=C(C=C(C=C1)NC(=O)C2=CC=C(C=C2)CN3CCN(CC3)C)NC4=NC=CC(=N4)C5=CN=CC=C5

Affinity Prediction Model Type

CNN-CNN-10epoch

ADMET Prediction Model Type

CNN

Canonical SMILES

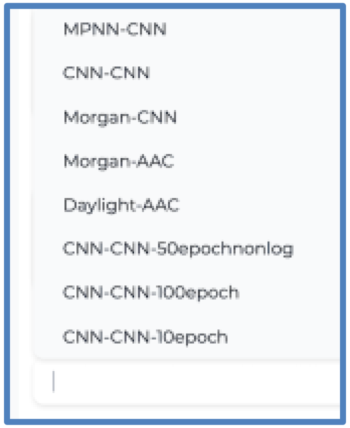
CC1=C(C=C(C=C1)NC(=O)C2=CC=C(C=C2)CN3CCN(CC3)C)NC4=NC=CC(=N4)C5=CN=CC=C5

Binding Affinity (Kd)

587.07 nM

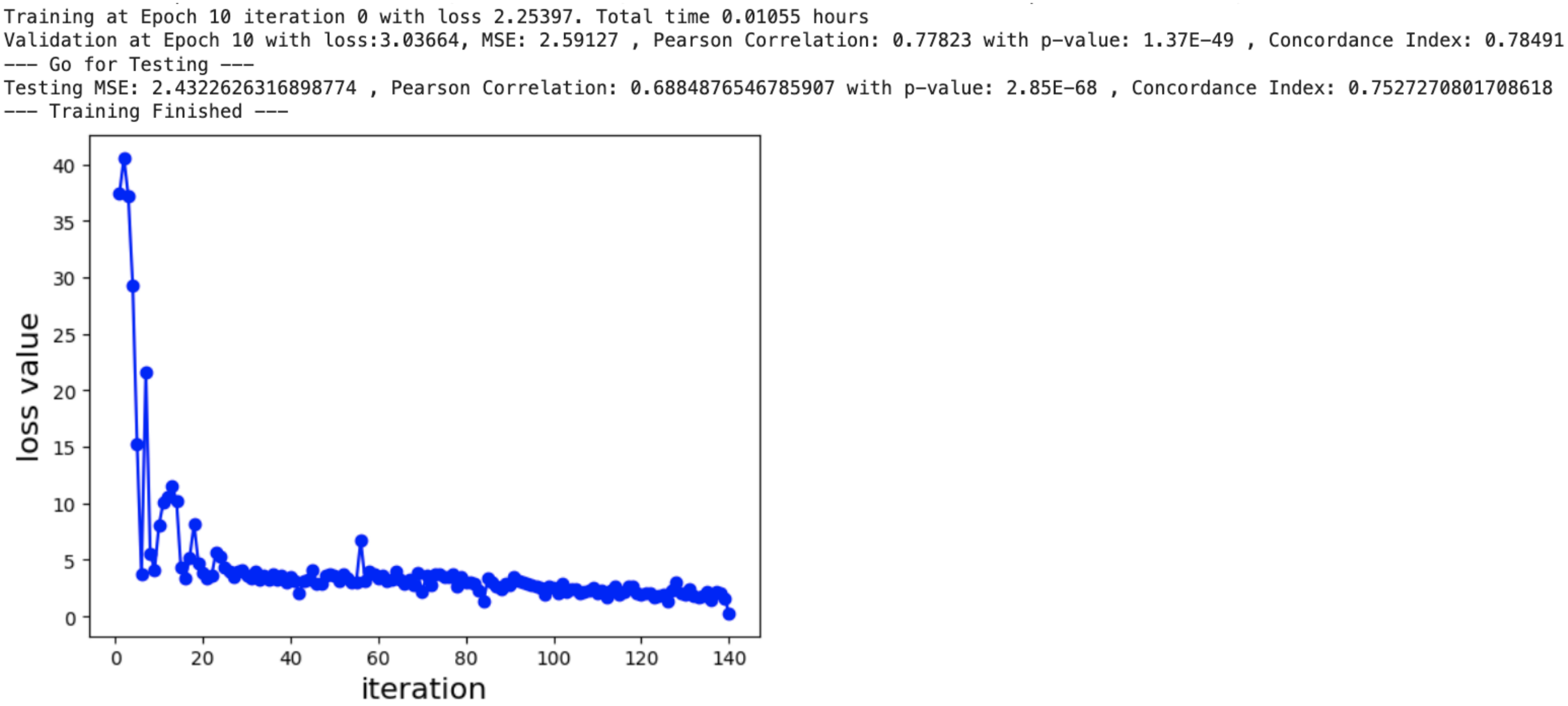
Predicted ADMET Property

```
{
  Solubility:  "-2.46 log mol/L",
  Lipophilicity:  "1.95 (log-ratio)",
  (Absorption) Caco-2:  "-4.75 cm/s",
  (Absorption) HIA:  "96.83 %",
  (Absorption) Pgp:  "3.70 %",
  (Absorption) Bioavailability F20:  "63.78 %",
  (Distribution) BBB:  "99.95 %",
  (Distribution) PPBR:  "94.24 %",
  (Metabolism) CYP2C19:  "75.54 %",
  (Metabolism) CYP2D6:  "5.95 %",
  (Metabolism) CYP3A4:  "84.94 %",
  (Metabolism) CYP1A2:  "4.85 %",
  (Metabolism) CYP2C9:  "62.25 %",
  (Execretion) Half life:  "8.03 h",
  (Execretion) Clearance:  "7.85 mL/min/kg",
  Clinical Toxicity:  "62.72 %"
}
```



CNN_CNN_BindingDB
599.65 nM
CNN-CNN-10epoch
587.07 nM

Results



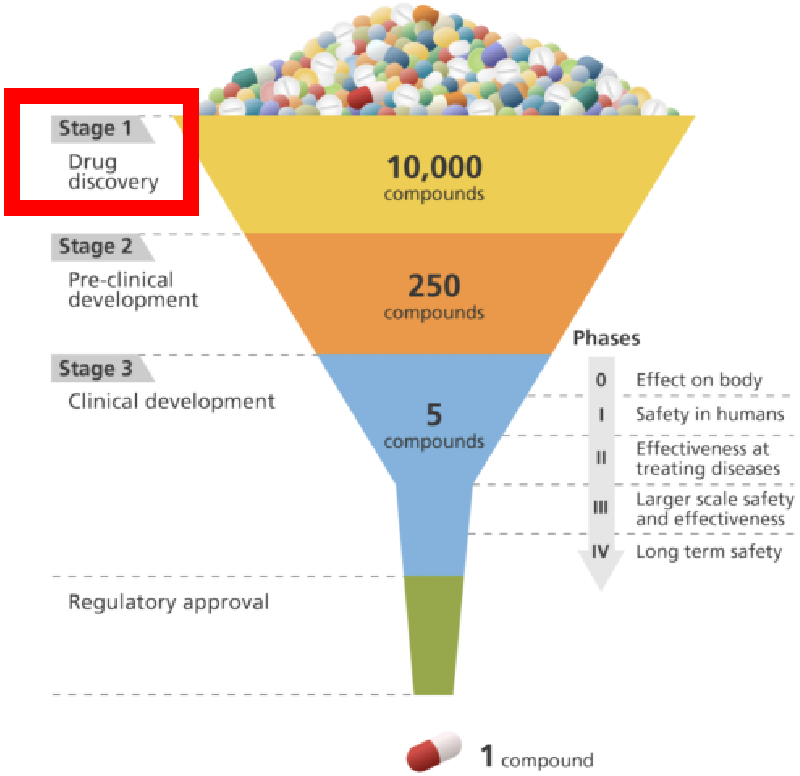
```
Training at Epoch 10 iteration 0 with loss 2.25397. Total time 0.01055 hours
Validation at Epoch 10 with loss:3.03664, MSE: 2.59127 , Pearson Correlation: 0.77823 with p-value: 1.37E-49 , Concordance Index: 0.78491
--- Go for Testing ---
Testing MSE: 2.4322626316898774 , Pearson Correlation: 0.6884876546785907 with p-value: 2.85E-68 , Concordance Index: 0.7527270801708618
--- Training Finished ---
```

Learnings/Future Work

- Learnings
  - ML moves quickly, expect deprecation/updates
  - Design flexible code for handling updated data structures
- Future Work
  - Model optimization, better results visualizations
    - train vs validation loss profiles
    - train vs validation MSE profiles
  - Model training with KIBA

Conclusion

- Fun project, learned a lot about DeepPurpose API
  - Successfully trained models using DeepPurpose
  - Successfully integrated models into MolDesigner UI
- Need for fast responsive and user-friendly DTI tools is growing fast
  - Easy to use for users who are not programmers
  - Facilitate faster drug development cycles by aiding drug discovery



YouTube URLs

- 2 minutes video URL (short): https://www.youtube.com/watch?v=BhsMjA6aVAs
- 15 minutes video URL(long): https://www.youtube.com/watch?v=87XkMBEzmhs