DeepPurpose

a Deep Learning Library for Drug and Target Molecular Modeling

Applications to DTI, DDI, PPI, Compound Property and Protein Function Prediction

Kexin Huang, Tianfan Fu, Lucas Glass, Marinka Zitnik, Cao Xiao, Jimeng Sun

HARVARD UNIVERSITY

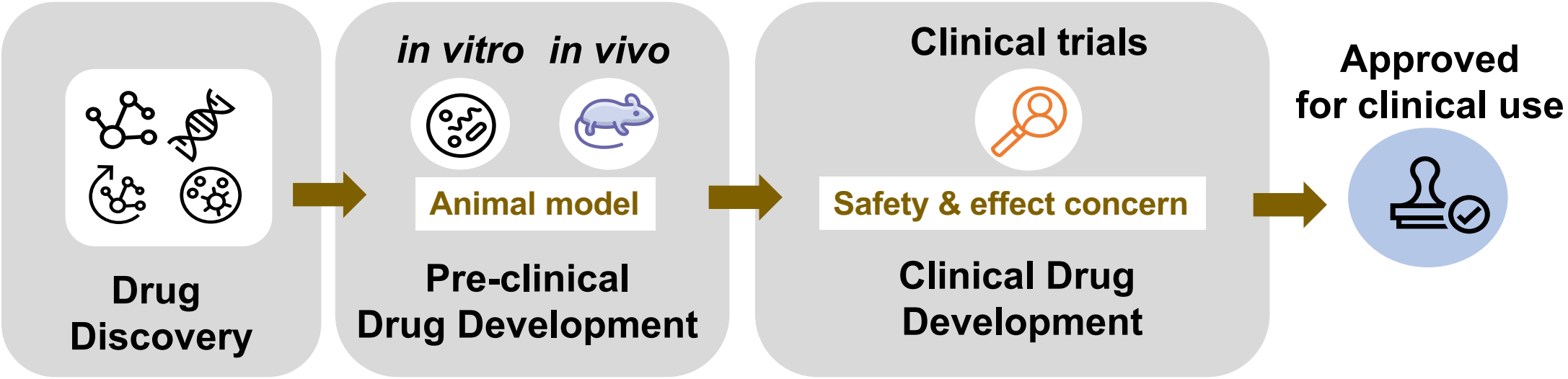
Georgia Tech

IQVIA™

ILLINOIS
UNIVERSITY OF ILLINOIS AT URBANA-CHAMPAIGN

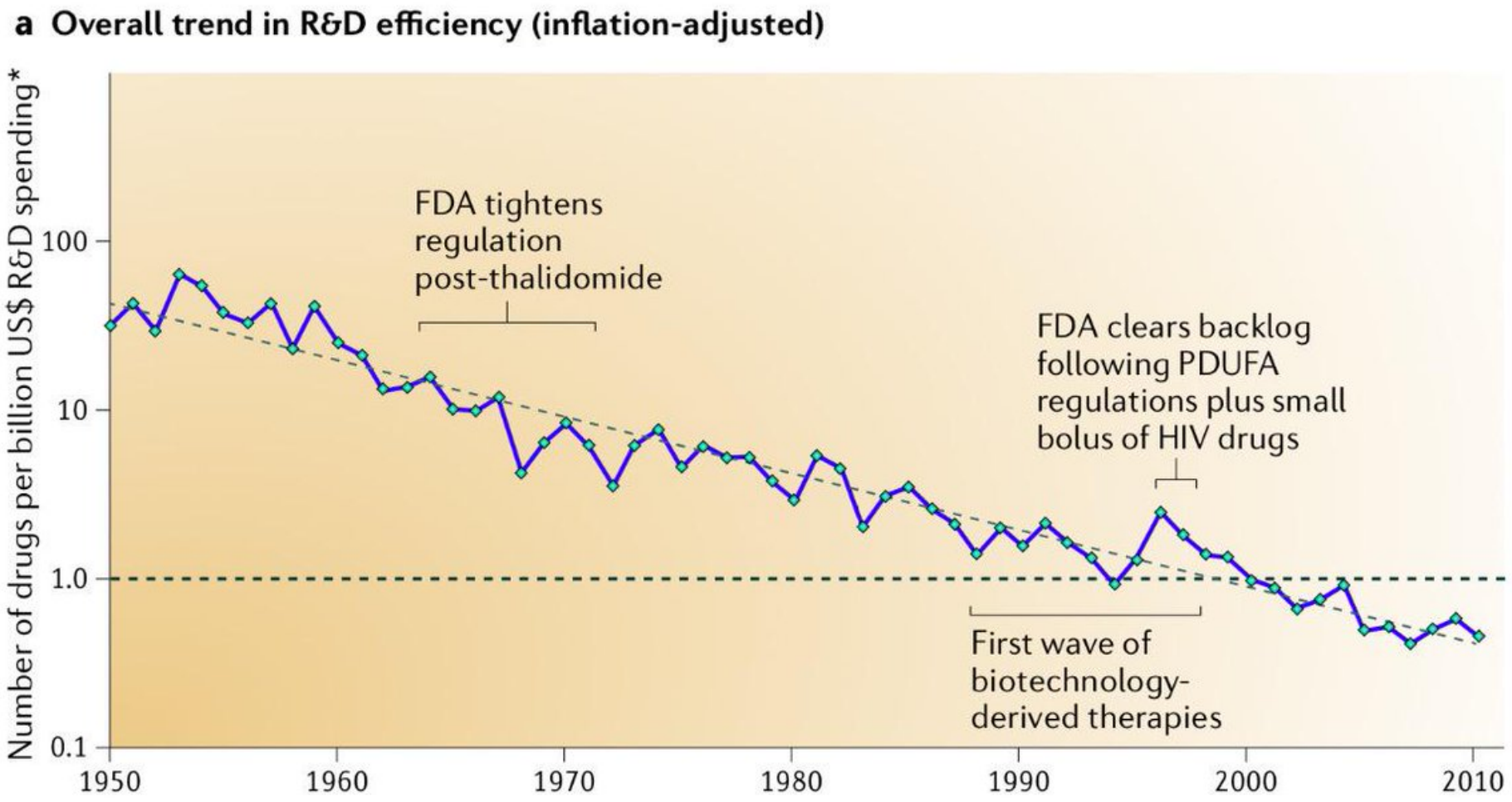
in Bioinformatics

Traditional Drug Discovery & Development Process



	Drug discovery	Pre-clinical	Phase 1	Phase 2	Phase 3
Time spent	4-5 years	1-2 years	1-2 years	1-2 years	2-3 years
$ spent	$550M	$125M	$225M	$250M	$250M
Output	5,000 - 10,000 compounds	10-20 candidates	5-10 candidates	2-5 candidates	1-2 candidates

Eroom's Law


a Overall trend in R&D efficiency (inflation-adjusted)
Number of drugs per billion US$ R&D spending*
100
10
1.0
0.1
1950
1960
1970
1980
1990
2000
2010
FDA tightens regulation post-thalidomide
FDA clears backlog following PDUFA regulations plus small bolus of HIV drugs
First wave of biotechnology-derived therapies

NIH Clinical Trial Shows Remdesivir Accelerates Recovery from Advanced COVID-19

April 29, 2020

Hospitalized patients with advanced COVID-19 and lung involvement who received remdesivir recovered faster than similar patients who received placebo, according to a preliminary data analysis from a randomized, controlled trial involving 1063 patients, which began on February 21. The trial (known as the Adaptive COVID-19 Treatment Trial, or ACTT), sponsored by the National Institute of Allergy and Infectious Diseases (NIAID), part of the National Institutes of Health, is the first clinical trial launched in the United States to evaluate an experimental treatment for COVID-19.

An independent data and safety monitoring board (DSMB) overseeing the trial met on April 27 to review data and shared their interim analysis with the study team. Based upon their review of the data, they noted that remdesivir was better than placebo from the perspective of the primary endpoint, time to recovery, a metric often used in influenza trials. Recovery in this study was defined as being well enough for hospital discharge or returning to normal activity level.

New Drug Discovery

10+ Years! $2.6 billion!

Patients Cannot Wait!

↓

Drug Repurposing

New uses for existing drugs

Aspirin, Sildenafil,...

Remdesivir: < 4 Months!

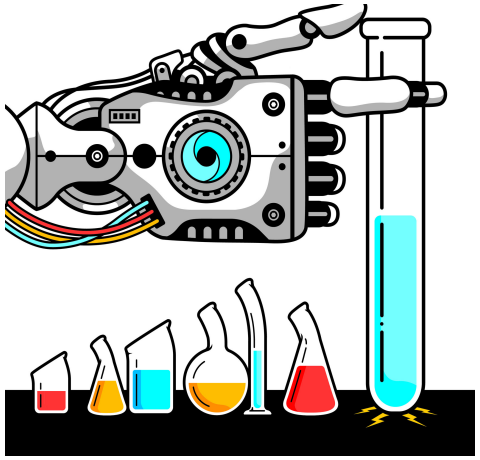
ML Accelerates Drug Discovery

MERCK
Be well

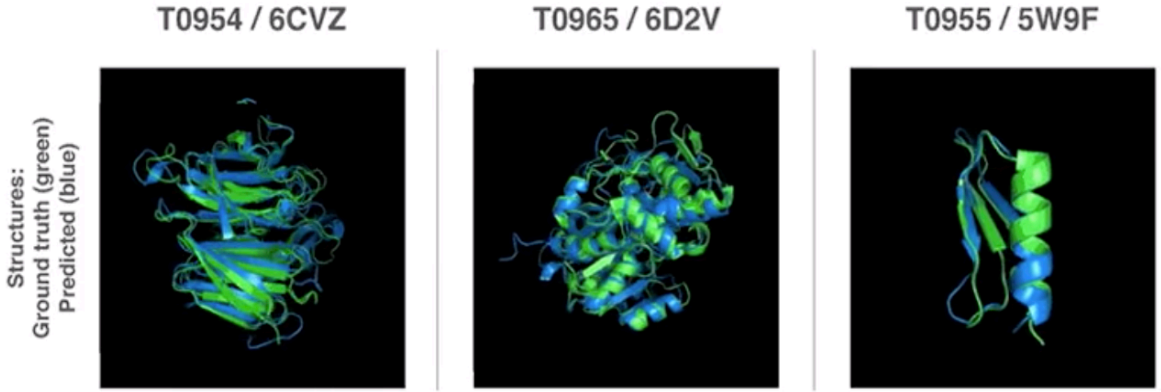
Merck Molecular Activity Challenge

Help develop safe and effective medicines by predicting molecular activity.

$40,000 · 236 teams · 7 years ago



DeepMind's AI will accelerate drug discovery by predicting how proteins fold



Leading Companies - Advanced AI in Healthcare and Drug Discovery / 2019 Q1

AI-Companies

Insilico Medicine 英科智能
Atomwise
GNS HEALTHCARE
CYCLICA
BenevolentAI
insitro
XtalPi
DeepMind Health
WuXiNextCODE
e-therapeutics
deep genomics
Numerate
RECURSION pharmaceuticals
Biovista
Exscientia
SCHRÖDINGER.
iCarbonX
CytoReason
NURITAS
PathAI
Lantern Pharma
twoXAR
ardigen

Pharma Corporations

AMGEN
astellas
MERCK
SANOFI
BAYER
Roche
gsk
AstraZeneca
NOVARTIS
Bristol-Myers Squibb
illumina
Pfizer
Johnson & Johnson
Boehringer Ingelheim
evotec
药明康德 WuXi AppTec
abbvie
novo nordisk
Celgene
IPSEN
Santen
DAEWOONG
Takeda
Sumitomo Dainippon Pharma
TEVA

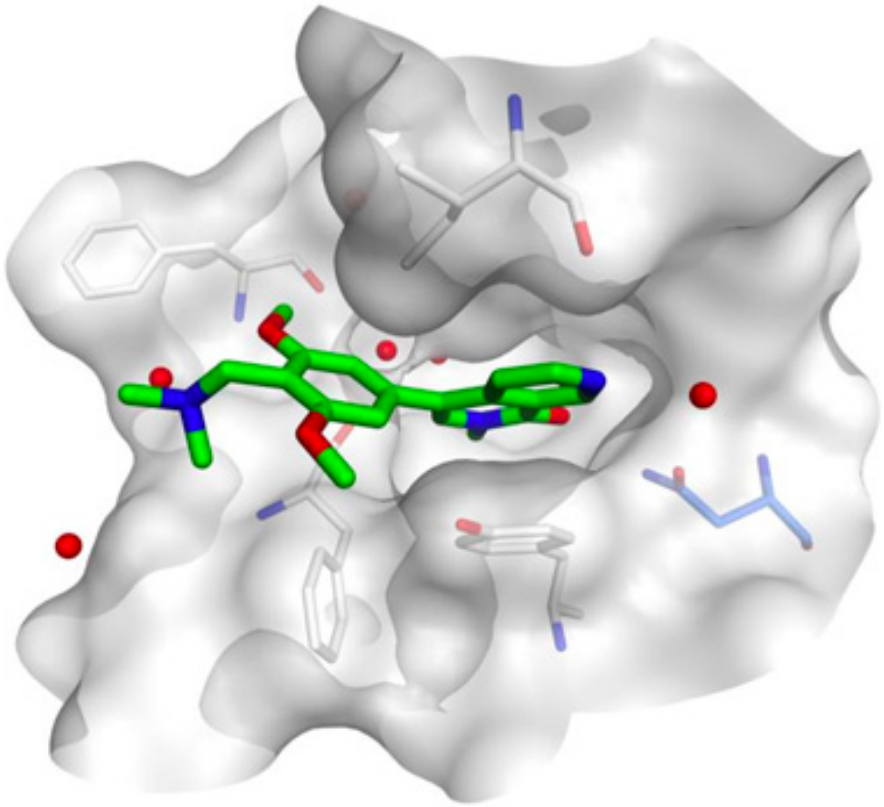
Tech Corporations

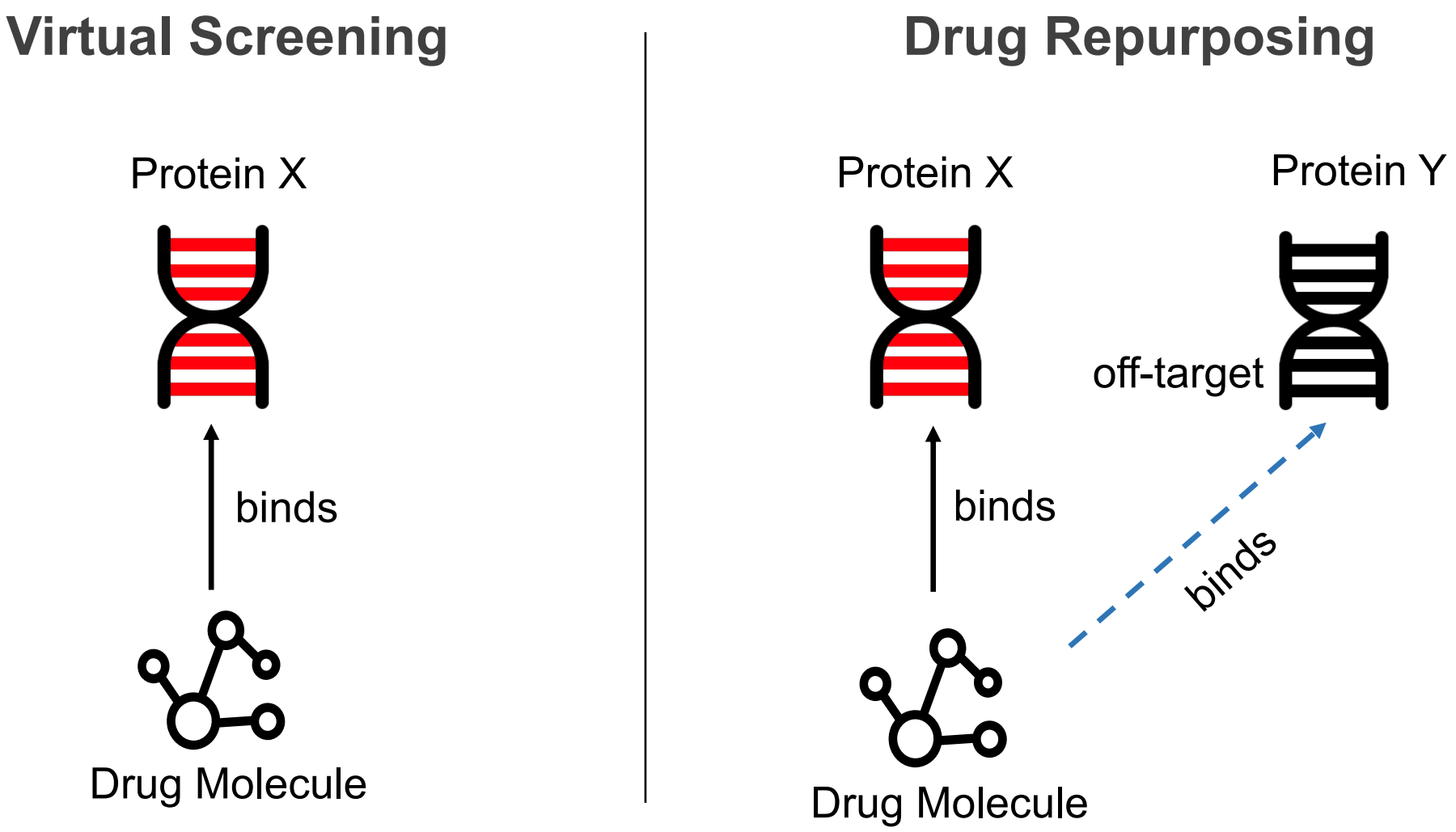
intel
Google
Canon
Alibaba Group
HITACHI
Baidu 百度
Tencent 腾讯
SAMSUNG
HUAWEI
SIEMENS
Adobe
IBM
cisco
NVIDIA
UNISYS
workday
ORACLE
FUJITSU
FOXCONN
SAP
DELL
salesforce

Investors

AME
Y
ZhenFund 真格基金
SEQUOIA
octopus ventures
AMGEN Ventures
Startx
G/
500
OS FUND
Felicis Ventures
DCVC
FORESITE CAPITAL
药明康德 WuXi AppTec
REFACTOR CAPITAL
Amadeus Capital
DRAPER ASSOCIATES
ARBORETUM VENTURES
Arch
LILLY VENTURES
Menlo
ANDREESSEN HOROWITZ
THIRD ROCK
SOSV THE ACCELERATOR VC
Baidu ventures
NLV
BILL & MELINDA GATES foundation
khosla ventures

DEEP KNOWLEDGE ANALYTICS
PHARMA DIVISION



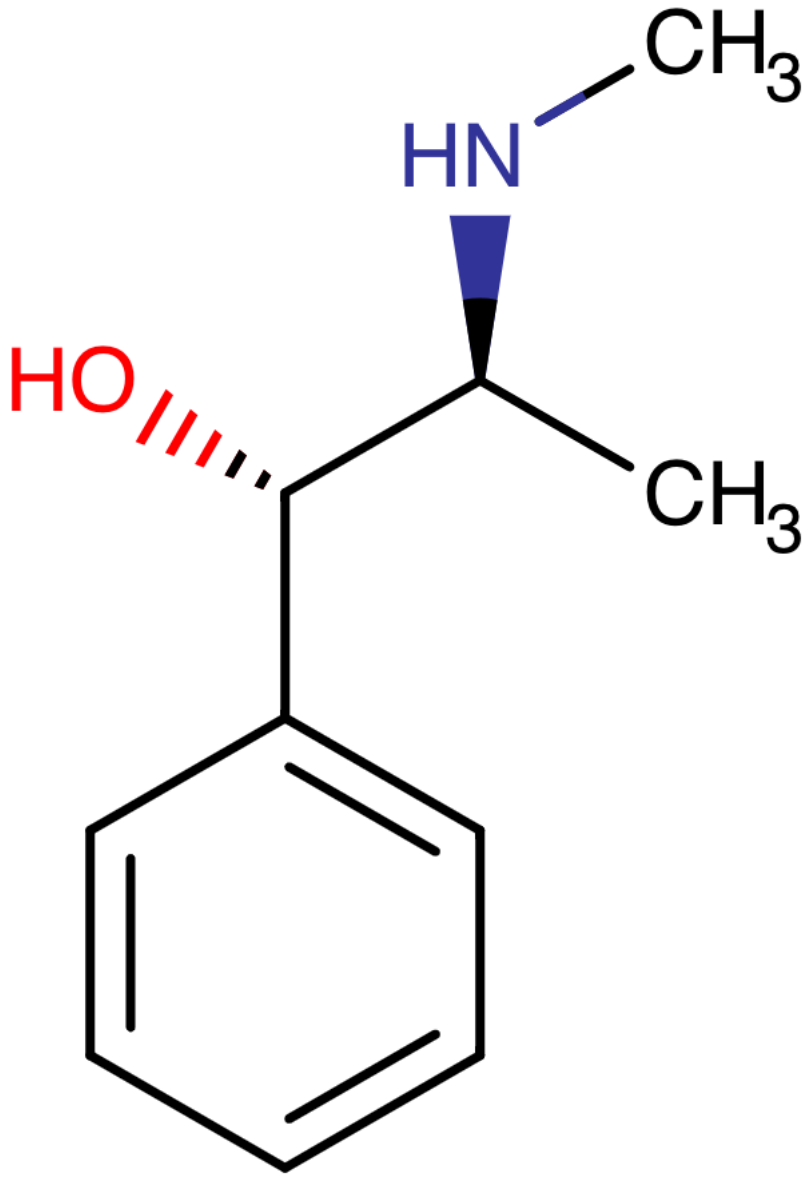

Virtual Screening
Protein X
binds
Drug Molecule
Drug Repurposing
Protein X
Protein Y
off-target
binds
binds
Drug Molecule

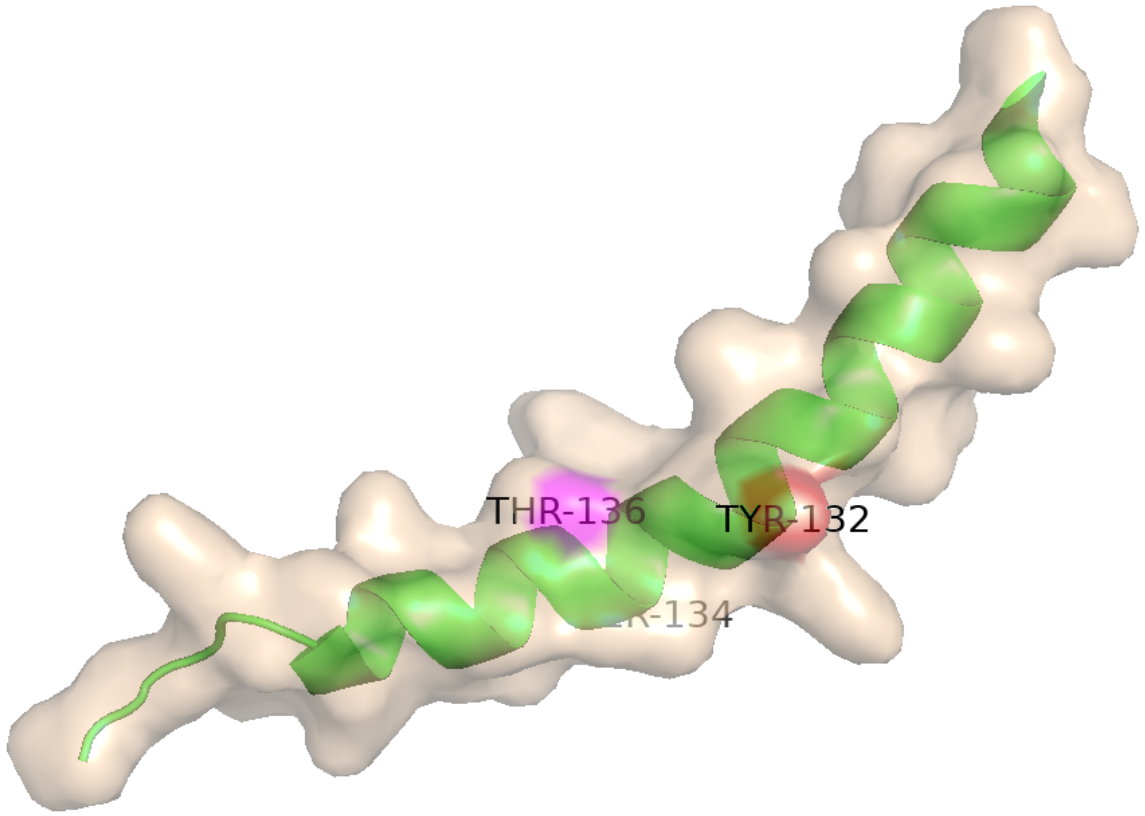
Pseudoephedrine

SMILES

Simplified molecular-input line-entry system

CC(C(C1=CC=CC=C1)O)NC

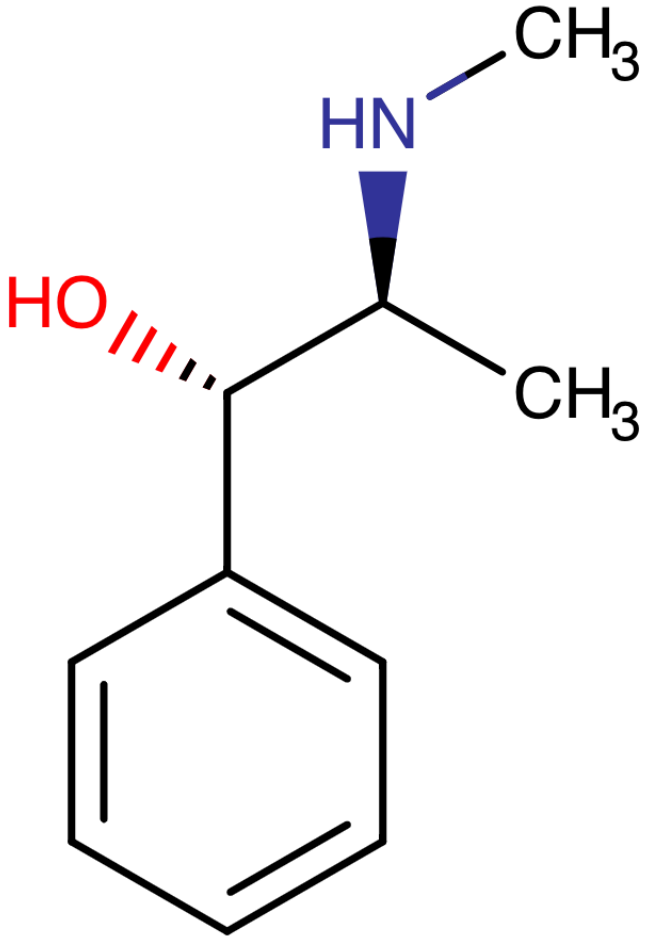


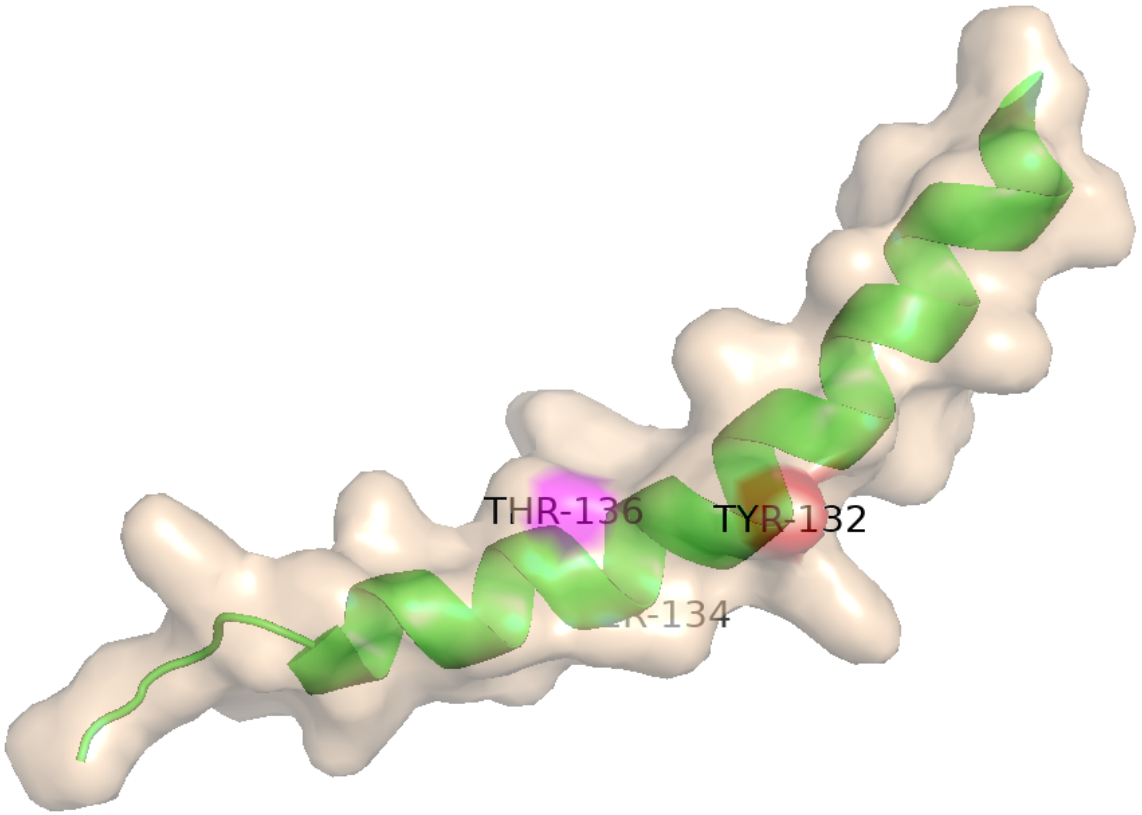


Alpha-2A receptor

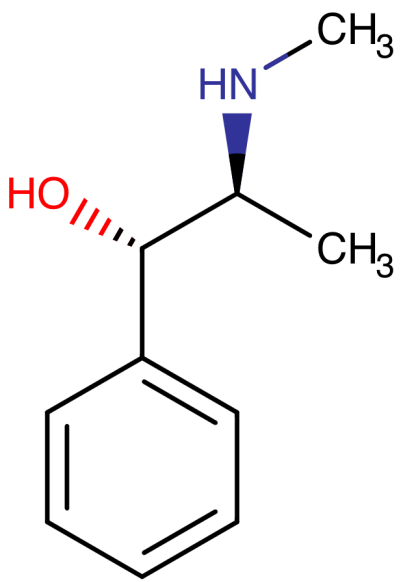
Amino Acid Sequence

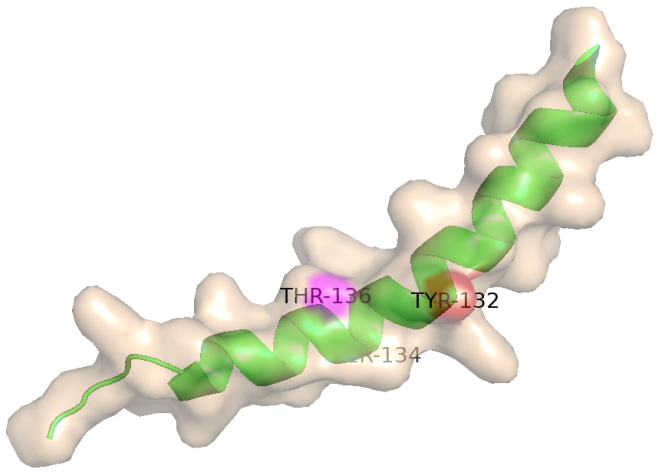
MFRQEQPLAEGSFAPMGSLQPDAGNASWNGTEA
PGGGARATPYSLQVTLTLVCLAGLLMLLTVFGNVLV
IIAVFTSRALKAPQNLFLVSLASADILVATLVIPFSLAN
EVMGYWYFGKAWCEIYLALDVLFCTSSIVHLCAISL
DRYWSITQAIEYNLKRTPRRIKAIIITVWVISAVISFPP
LISIEKKGGGGGPQPAEPRCEINDQKWYVISSCIGS
FFAPCLIMILVYVRIYQIAKRRTRVPPSRRGPDAVAA
PPGGTERRPNGLGPERSAGPGGAEAEPLPTQLNG
APGEPAPAGPRDTDALDLEESSSSDHAERPPGPRR
PERGPRGKGKARASQVKPGDSLPRRGPGATGIGT
PAAGPGEERVGAAKASRWRGRQNREKRFTFVLAV
VIGVFVVCWFPFFFTYTLTAVGCSVPRTLFKFFFWF
GYCNSSLNPVIYTIFNHDFRRAFKKILCRGDRKRIV





Q: Will they bind?





A machine learning question:

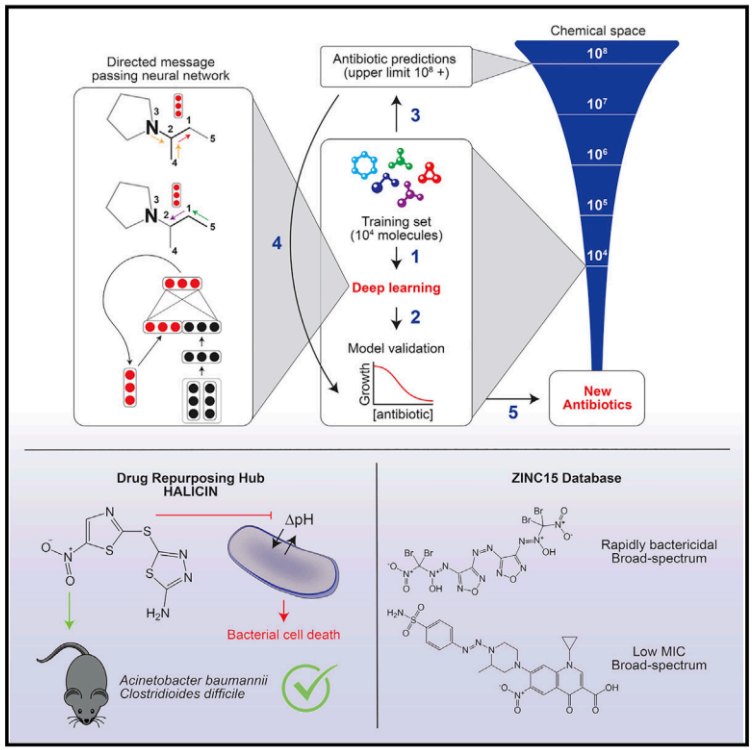
Given drug SMILES, target amino acid sequence, what is their predicted binding affinity score?

Article

Cell

A Deep Learning Approach to Antibiotic Discovery

Graphical Abstract



Highlights

- A deep learning model is trained to predict antibiotics based on structure
- Halicin is predicted as an antibacterial molecule from the Drug Repurposing Hub
- Halicin shows broad-spectrum antibiotic activities in mice
- More antibiotics with distinct structures are predicted from the ZINC15 database

Authors

Jonathan M. Stokes, Kevin Yang, Kyle Swanson, ..., Tommi S. Jaakkola, Regina Barzilay, James J. Collins

Correspondence

regina@csail.mit.edu (R.B.), jimjc@mit.edu (J.J.C.)

In Brief

A trained deep neural network predicts antibiotic activity in molecules that are structurally different from known antibiotics, among which Halicin exhibits efficacy against broad-spectrum bacterial infections in mice.

Deep learning shows great promise!

A scikit-learn style framework is missing!

↓

DeepPurpose

DeepPurpose

For Biomedical Scientist

ONE line of code to do:

drug repurposing
virtual screening
property prediction

results aggregated from 5 pretrained SOTA deep learning models

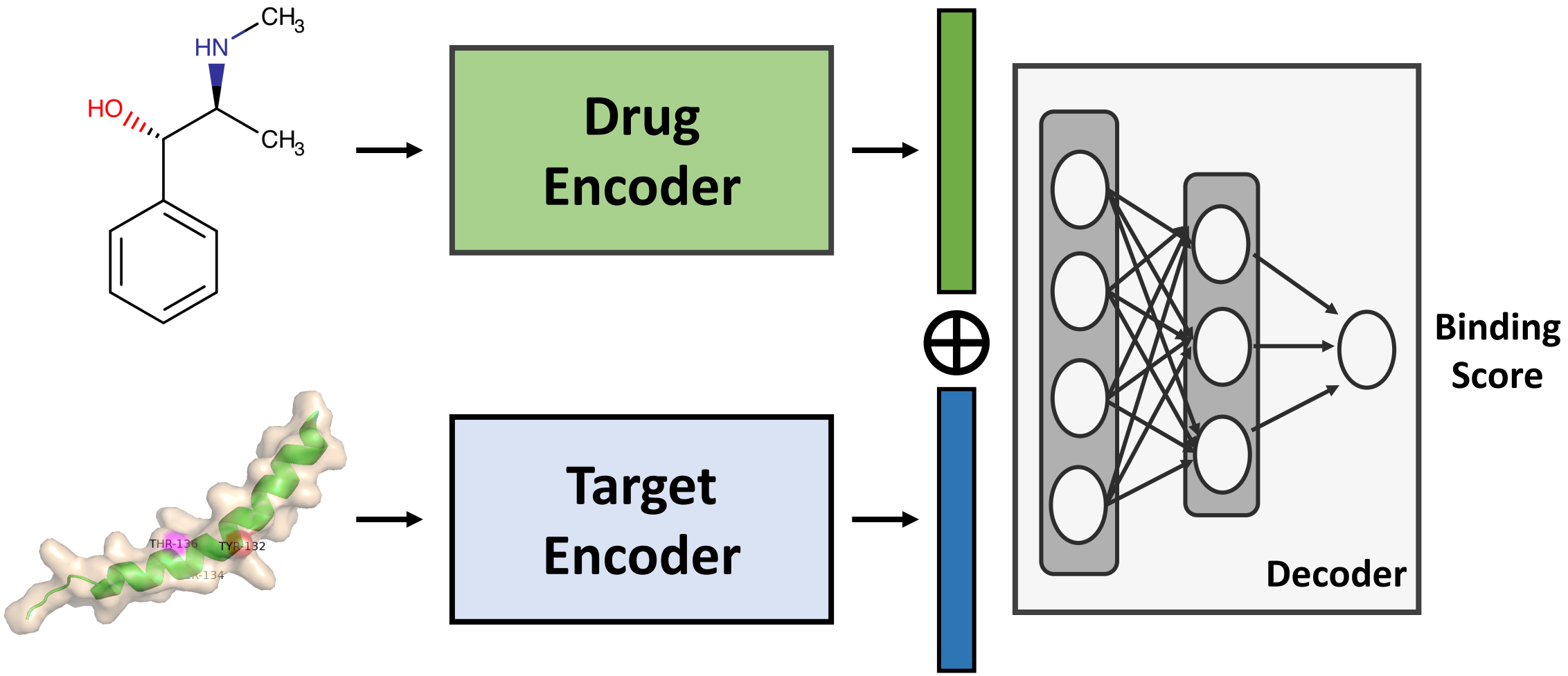
accept customized training dataset

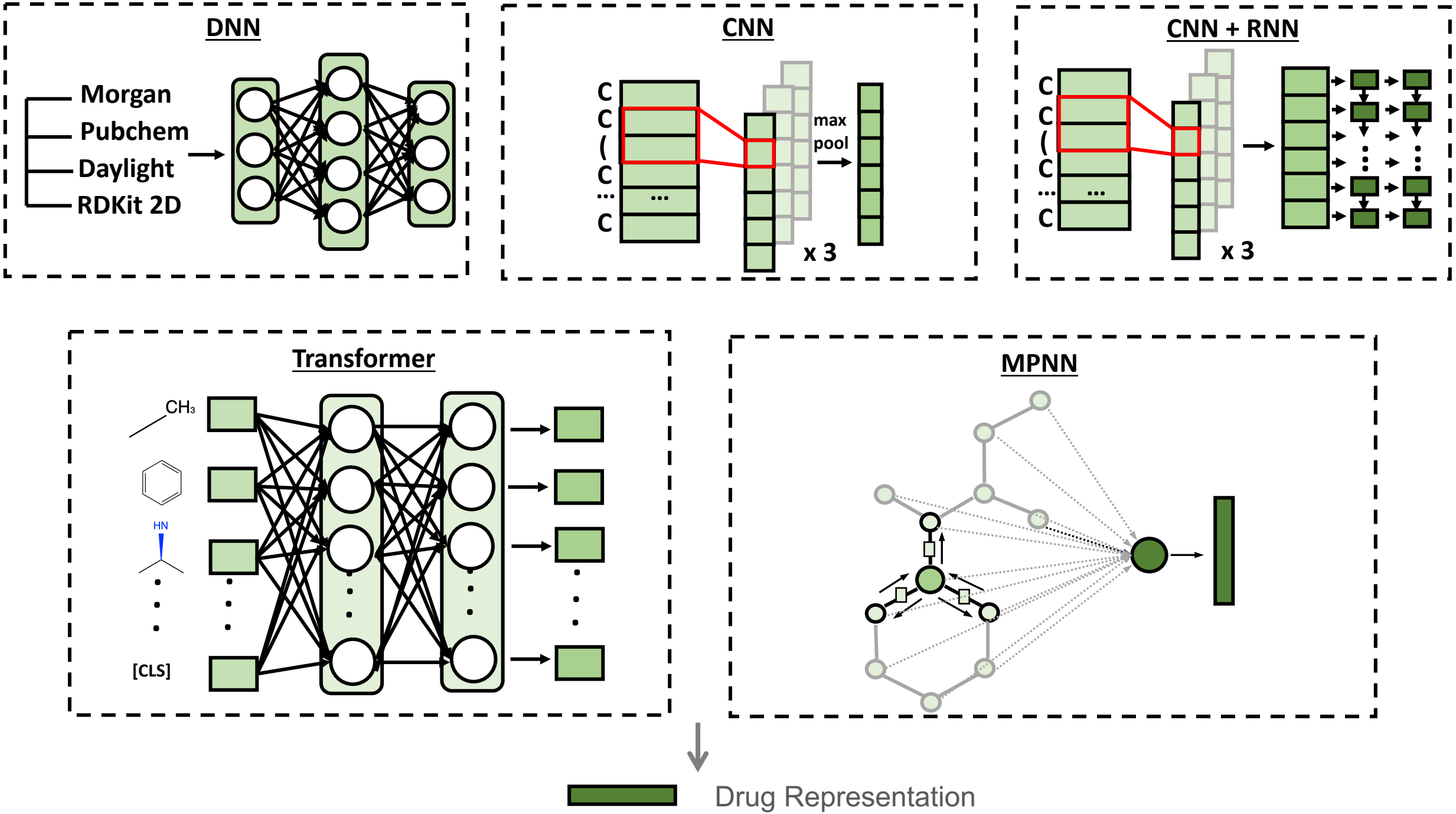
For ML Researcher

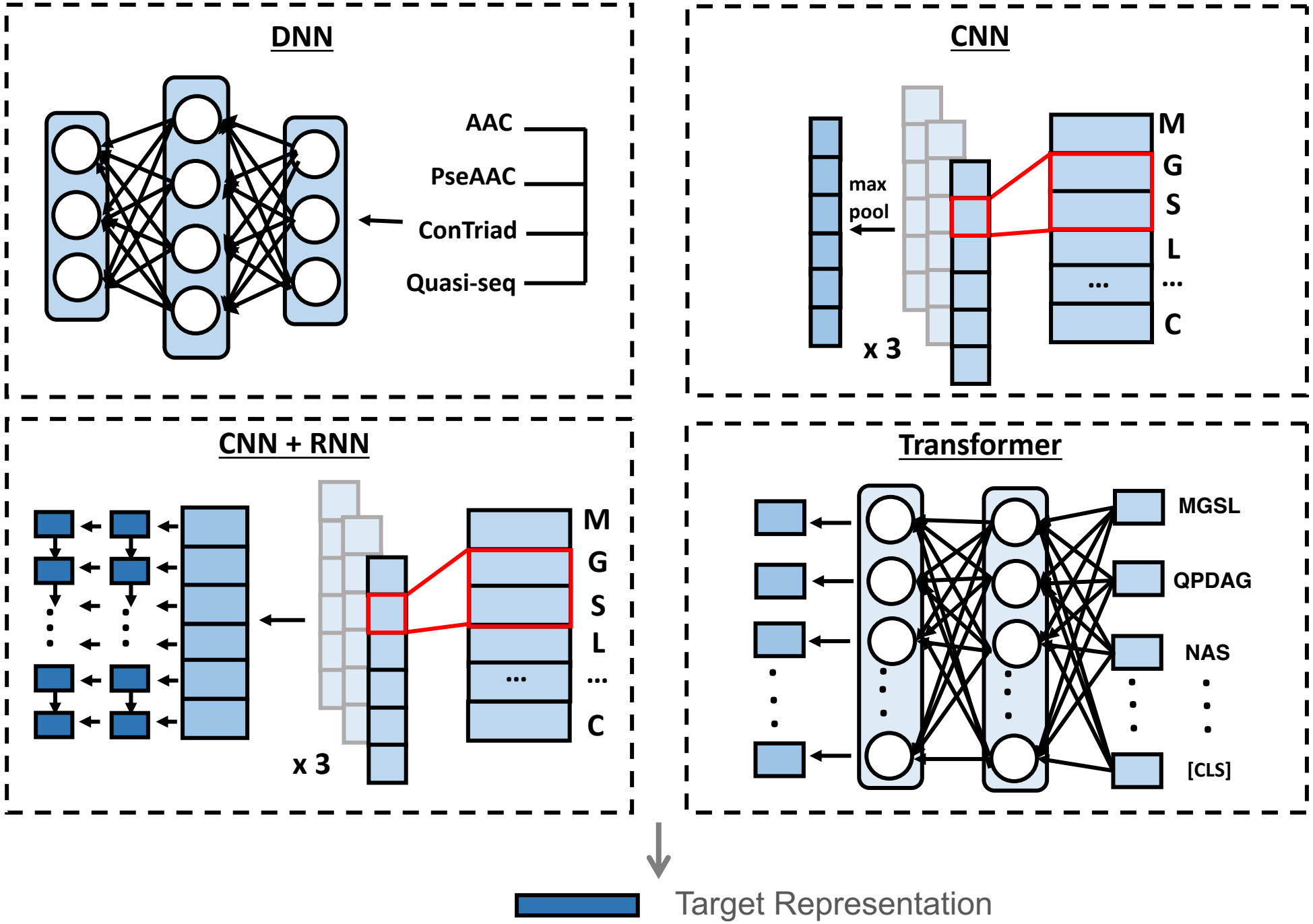
10 lines framework to unlock:

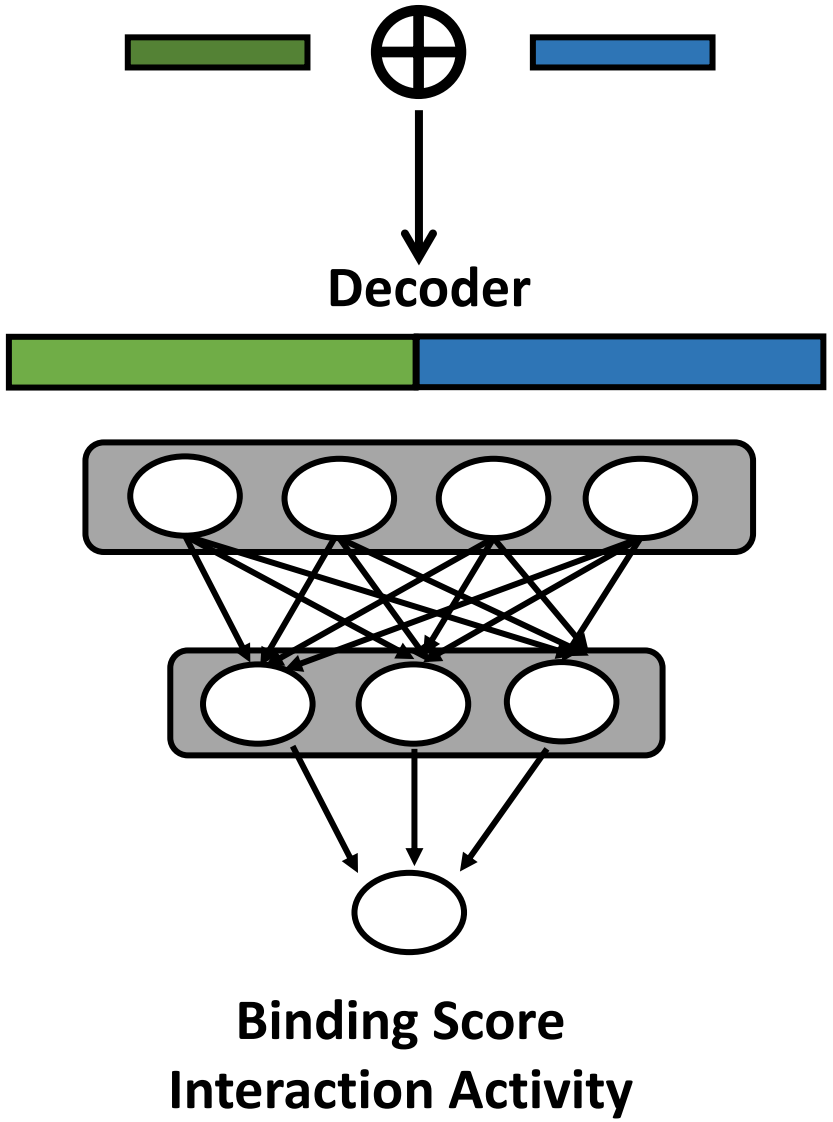
50+ novel models
15+ novel encoders
10+ pretrained models
5+ benchmark datasets

automatic result figures generation
training monitoring
robustness evaluation
result ensembles


CH3
HN
HO
CH3
Drug Encoder
Target Encoder
Decoder
Binding Score

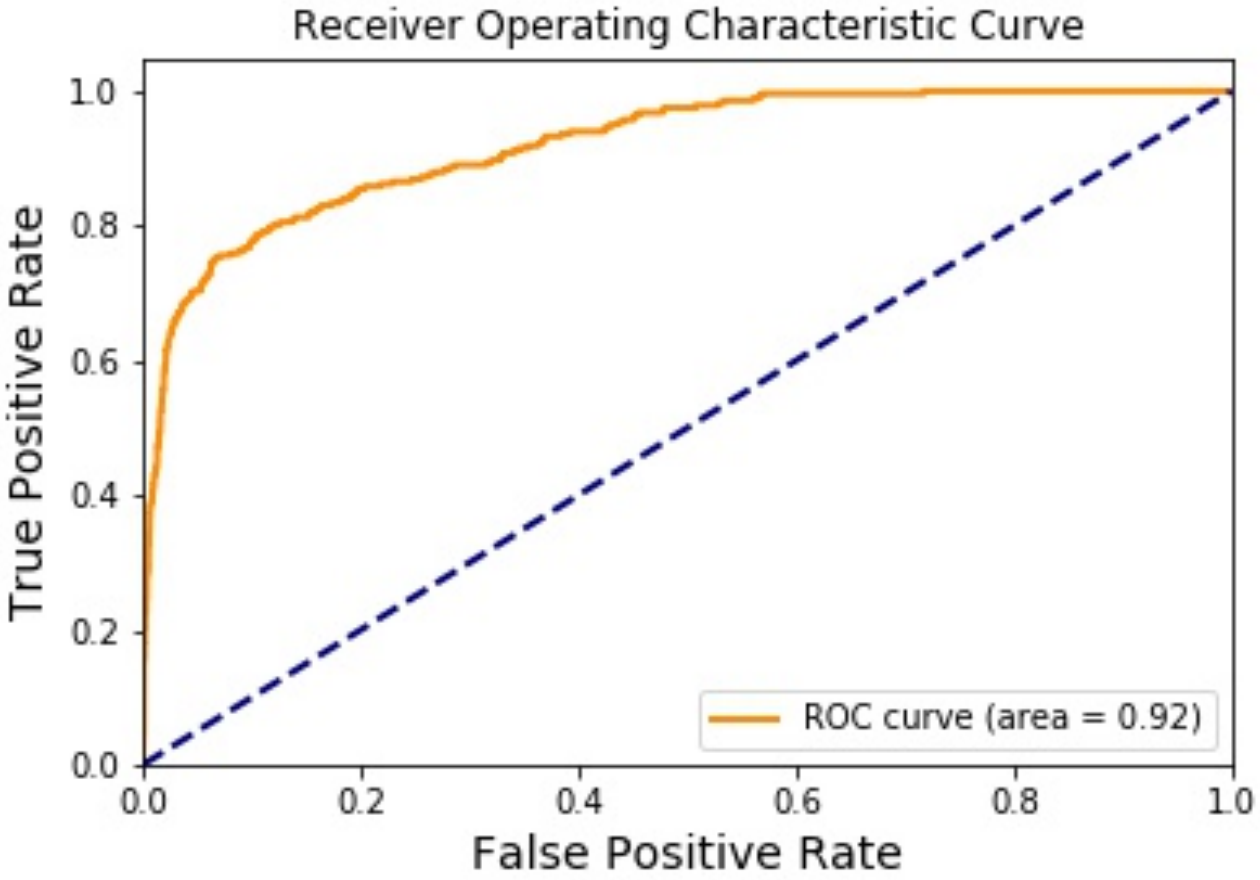

DNN
Morgan
Pubchem
Daylight
RDKit 2D
CNN
C
C
(
C
...
C
...
max
pool
x 3
CNN + RNN
C
C
(
C
...
C
...
x 3
Transformer
CH3
HN
[CLS]
MPNN
Drug Representation

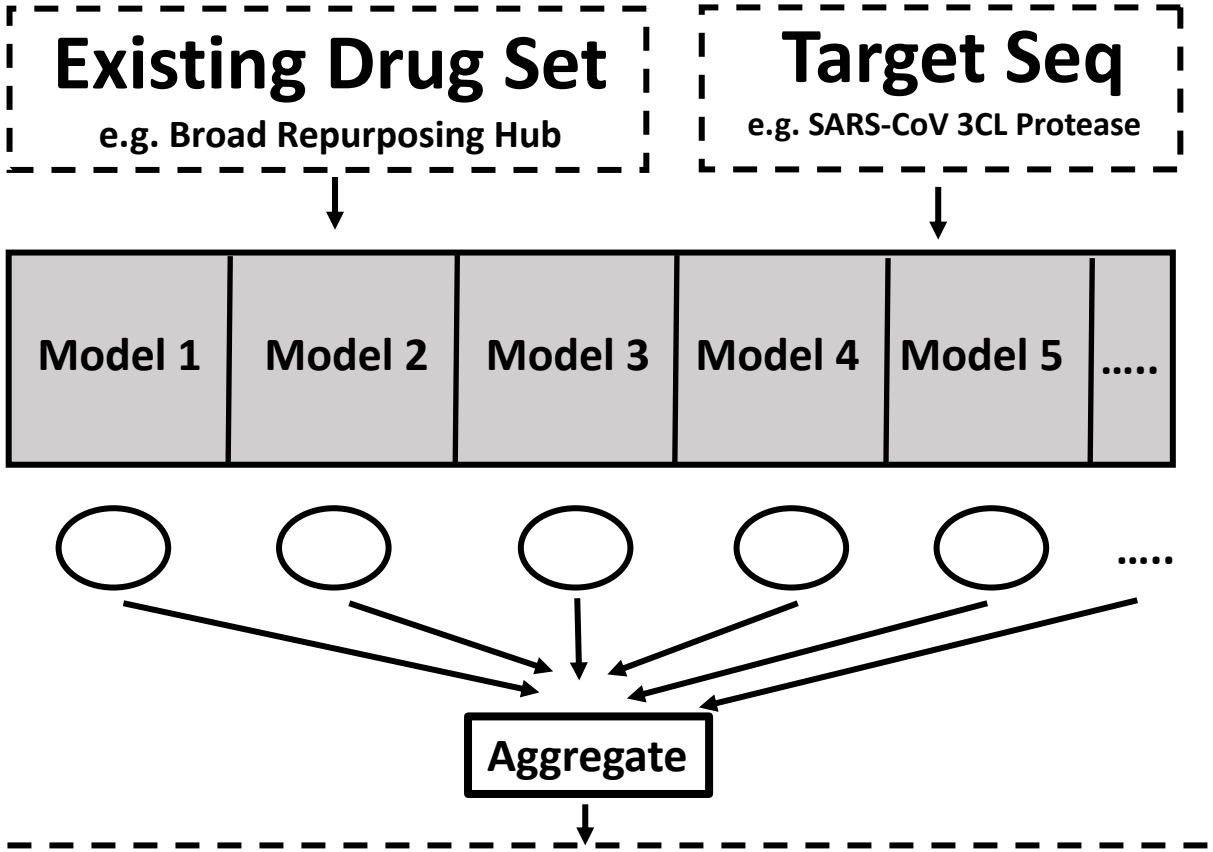

DNN
AAC
PseAAC
ConTriad
Quasi-seq
CNN
max
pool
x 3
M
G
S
L
...
C
CNN + RNN
x 3
M
G
S
L
...
C
Transformer
MGSL
QPDAG
NAS
[CLS]
Target Representation



Auto-Generated Test Set Performance Table & Figure:

MSE	Pearson Correlation	with p-value	Concordance Index
0.2795	0.8317	0.0000	0.8838





Binding Ranked Drug Candidate List:

Drug Repurposing Result for SARS-CoV2 3CL Protease

Rank	Drug Name	Target Name	Binding Score
1	Sofosbuvir	SARS-CoV2 3CL Protease	360.22
2	Daclatasvir	SARS-CoV2 3CL Protease	424.06
3	Vicriviroc	SARS-CoV2 3CL Protease	623.78
4	Efavirenz	SARS-CoV2 3CL Protease	768.33

Case Study I: Drug Repurposing for 3CLPro

```
>>> from DeepPurpose import oneliner
>>> from DeepPurpose.dataset import *
>>>
>>> oneliner.repurpose(*read_file_target_sequence('target.txt'), \
                       *read_file_repurposing_library('repurpose.txt'))
```

Rank		Drug Name	Target Name	Binding Score
1	*	Sofosbuvir	SARS-CoV2 3CL Protease	190.25
2		Daclatasvir	SARS-CoV2 3CL Protease	214.58
3		Vicriviroc	SARS-CoV2 3CL Protease	315.70
4	*	Simeprevir	SARS-CoV2 3CL Protease	396.53
5		Etravirine	SARS-CoV2 3CL Protease	409.34
6	*	Amantadine	SARS-CoV2 3CL Protease	419.76
7		Letermovir	SARS-CoV2 3CL Protease	460.28
8		Rilpivirine	SARS-CoV2 3CL Protease	470.79
9	+	Darunavir	SARS-CoV2 3CL Protease	472.24
10	+	Lopinavir	SARS-CoV2 3CL Protease	473.01
11		Maraviroc	SARS-CoV2 3CL Protease	474.86
12		Fosamprenavir	SARS-CoV2 3CL Protease	487.45
13	+	Ritonavir	SARS-CoV2 3CL Protease	492.19

* Supported by other Literature Evidence

+ Undergo Clinical Trial for COVID-19

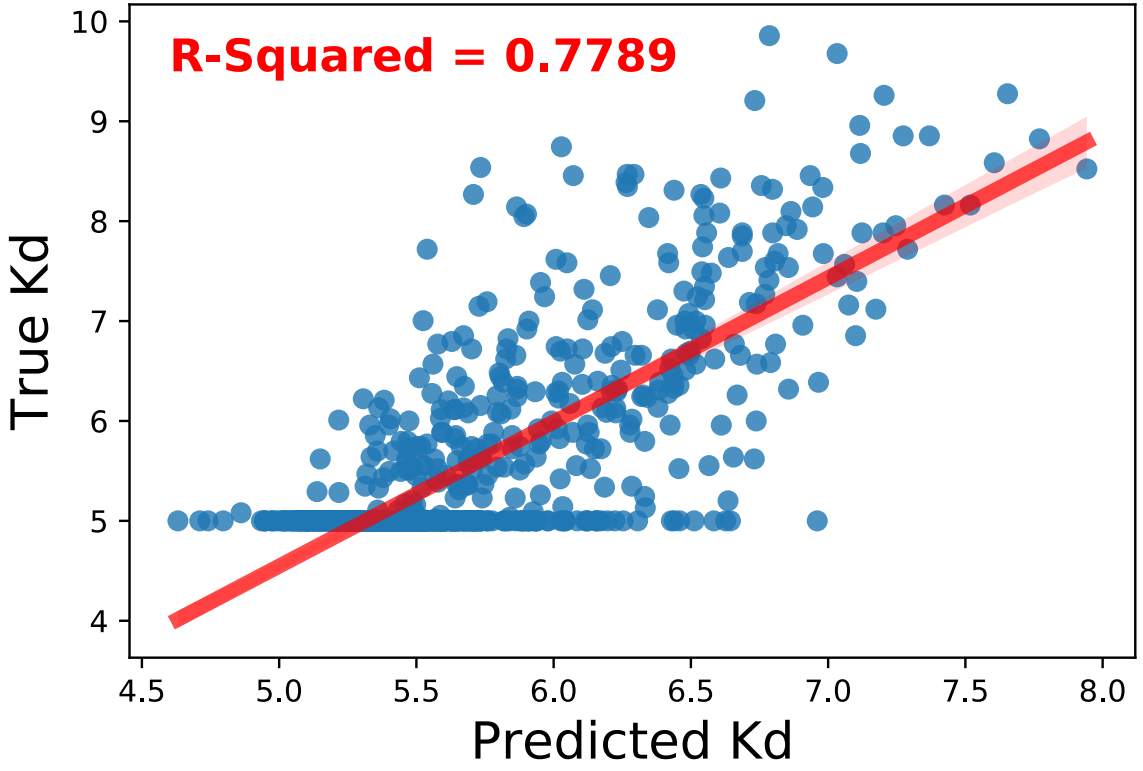
Case Study II: Virtual Screening using One Line and Binding Predictive Performance

```
>>> from DeepPurpose import oneliner
>>> from DeepPurpose.dataset import *
>>>
>>> oneliner.virtual_screening(['MKK...LIDL', ...], ['CC1=C...C4)N', ...])
```

Test set performance on pretraining BindingDB dataset:

DeepPurpose Model	MSE	Concordance Index
MPNN+CNN	0.635 (0.014)	0.841 (0.004)
CNN+CNN	0.600 (0.007)	0.857 (0.003)
Morgan+CNN	0.631 (0.002)	0.846 (0.005)
Morgan+AAC	0.629 (0.034)	0.848 (0.005)
Daylight+AAC	0.649 (0.014)	0.841 (0.004)

Performance on UNSEEN DAVIS dataset:



Case Study III: Drug Repurposing with Customized Data

target.txt

```
SARS-CoV2_3CL_Protease SGF...TFQ
```

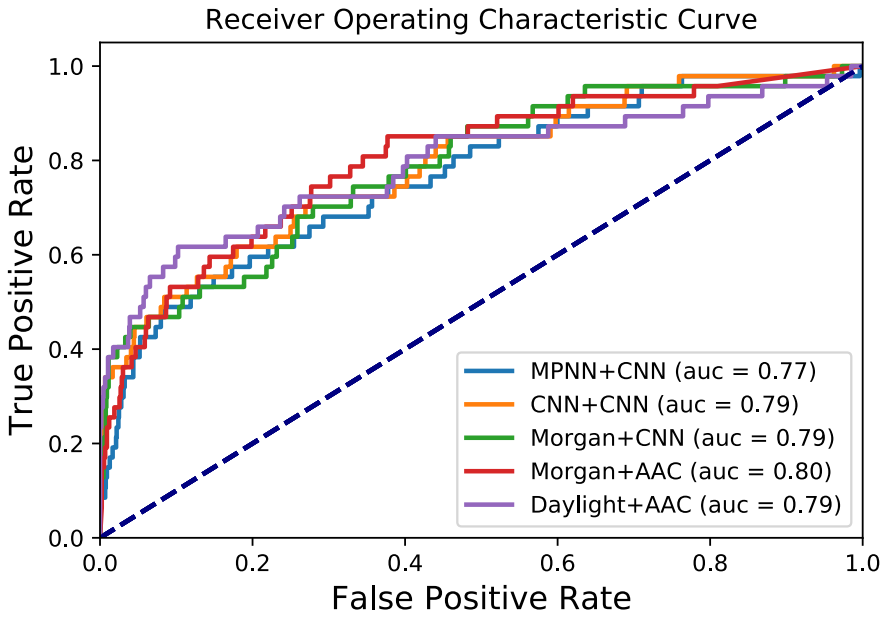
repurpose.txt

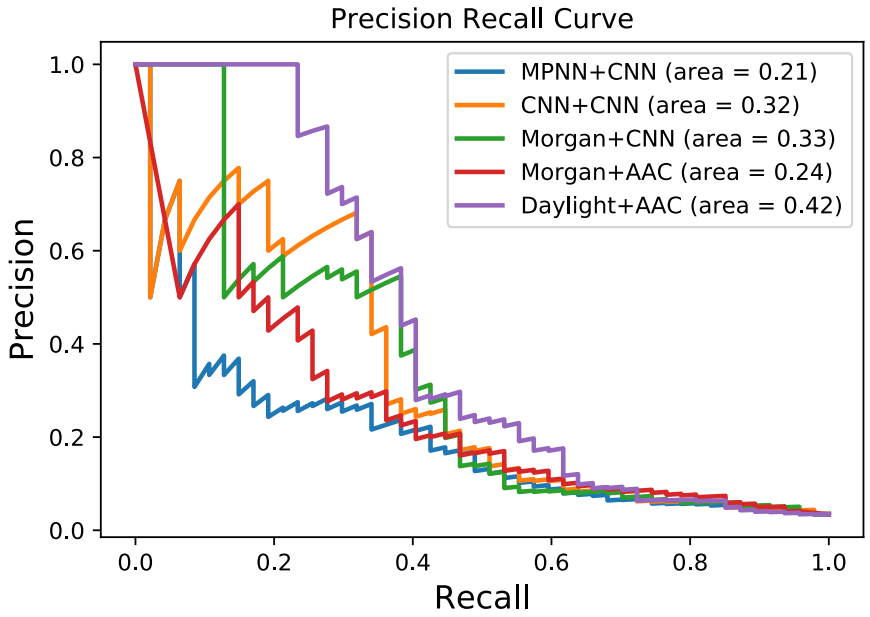
```
Rufloxacin     CN1...2=O
Sparfloxacin   C[C...CCl
...
```

train.txt

```
SGFKK...VGGVRLQ
CCOC1...C=CC=N4  0
CC1=C...C=CC=C2  1
...
```

```
>>> from DeepPurpose import oneliner
>>> from DeepPurpose.dataset import *
>>>
>>> oneliner.repurpose(*read_file_target_sequence('target.txt'), \
                       *read_file_repurposing_library('repurpose.txt'), \
                       *read_file_training_dataset_bioassay('train.txt'),\
                       split='HTS', convert_y = False, \
                       frac=[0.8,0.1,0.1], finetune_LR = 1e-3,\
                       pretrained = False, agg = 'max_effect')
```





A DTI Prediction Framework

```
>>> from DeepPurpose import models
>>> from DeepPurpose.utils import *
>>> from DeepPurpose.dataset import *
>>>
>>> X_drug, X_target, y = load_process_DAVIS(SAVE_PATH, binary=False)
>>>
>>> drug_encoding, target_encoding = 'CNN', 'CNN'
>>> train, val, test = data_process(X_drug, X_target, y, drug_encoding, \
                                    target_encoding, split_method='random', \
                                    frac=[0.7,0.1,0.2], random_seed = 1)
>>>
>>> config = generate_config(drug_encoding, target_encoding, \
                             cls_hidden_dims = [1024,1024,512], \
                             train_epoch = 100, LR = 0.001, batch_size = 256, \
                             cnn_drug_filters = [32,64,96], \
                             cnn_drug_kernels = [4,8,12], \
                             cnn_target_filters = [32,64,96], \
                             cnn_target_kernels = [4,8,12])
>>>
>>> model = models.model_initialize(**config)
>>> model.train(train, val, test)
```

Dataset 1: DAVIS			
	Model	MSE	Concordance Index
Baselines	KronRLS	0.329 (0.019)	0.847 (0.006)
	GraphDTA	0.263 (0.015)	0.864 (0.007)
	DeepDTA	0.262 (0.022)	0.870 (0.003)
DeepPurpose	CNN+CNN	0.254 (0.018)	0.879 (0.011)
	MPNN+CNN	0.271 (0.012)	0.858 (0.007)
	MPNN+AAC	0.242 (0.009)	0.881 (0.005)
	CNN+Trans	0.282 (0.009)	0.852 (0.006)
	Morgan+CNN	0.271 (0.012)	0.858 (0.007)
	Morgan+AAC	0.258 (0.012)	0.861 (0.008)
	Daylight+AAC	0.277 (0.014)	0.861 (0.008)
Dataset 2: KIBA			
	Model	MSE	Concordance Index
Baselines	KronRLS	0.852 (0.014)	0.688 (0.003)
	GraphDTA	0.183 (0.003)	0.862 (0.005)
	DeepDTA	0.196 (0.008)	0.864 (0.002)
DeepPurpose	CNN+CNN	0.196 (0.005)	0.856 (0.004)
	MPNN+CNN	0.222 (0.006)	0.825 (0.003)
	MPNN+AAC	0.178 (0.002)	0.872 (0.001)
	CNN+Trans	0.240 (0.013)	0.818 (0.004)
	Morgan+CNN	0.229 (0.008)	0.825 (0.004)
	Morgan+AAC	0.233 (0.009)	0.823 (0.004)
	Daylight+AAC	0.252 (0.014)	0.808 (0.008)

THE LANCET

ARTICLES | VOLUME 395, ISSUE 10229, P1054-1062, MARCH 28, 2020

Clinical course and risk factors for mortality of adult inpatients with COVID-19 in Wuhan, China: a retrospective cohort study

Fei Zhou, MD † • Ting Yu, MD † • Ronghui Du, MD † • Guohui Fan, MS † • Ying Liu, MD † • Zhibo Liu, MD † • et al.

Show all authors • Show footnotes

Published: March 11, 2020 • DOI: https://doi.org/10.1016/S0140-6736(20)30566-3 • Check for updates

Summary

Background

Since December, 2019, Wuhan, China, has experienced an outbreak of coronavirus disease 2019 (COVID-19), caused by the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2). Epidemiological and clinical characteristics of patients with COVID-19 have been reported but risk factors for mortality and a detailed clinical course of illness, including viral shedding, have not been well described.

Of the 54 non-survivors observed, 27 of these patients had a secondary infection (**50**%).

Contrarily, only **one** of the 137 surviving patients tracked in this paper had a secondary infection (**~0.7**%).

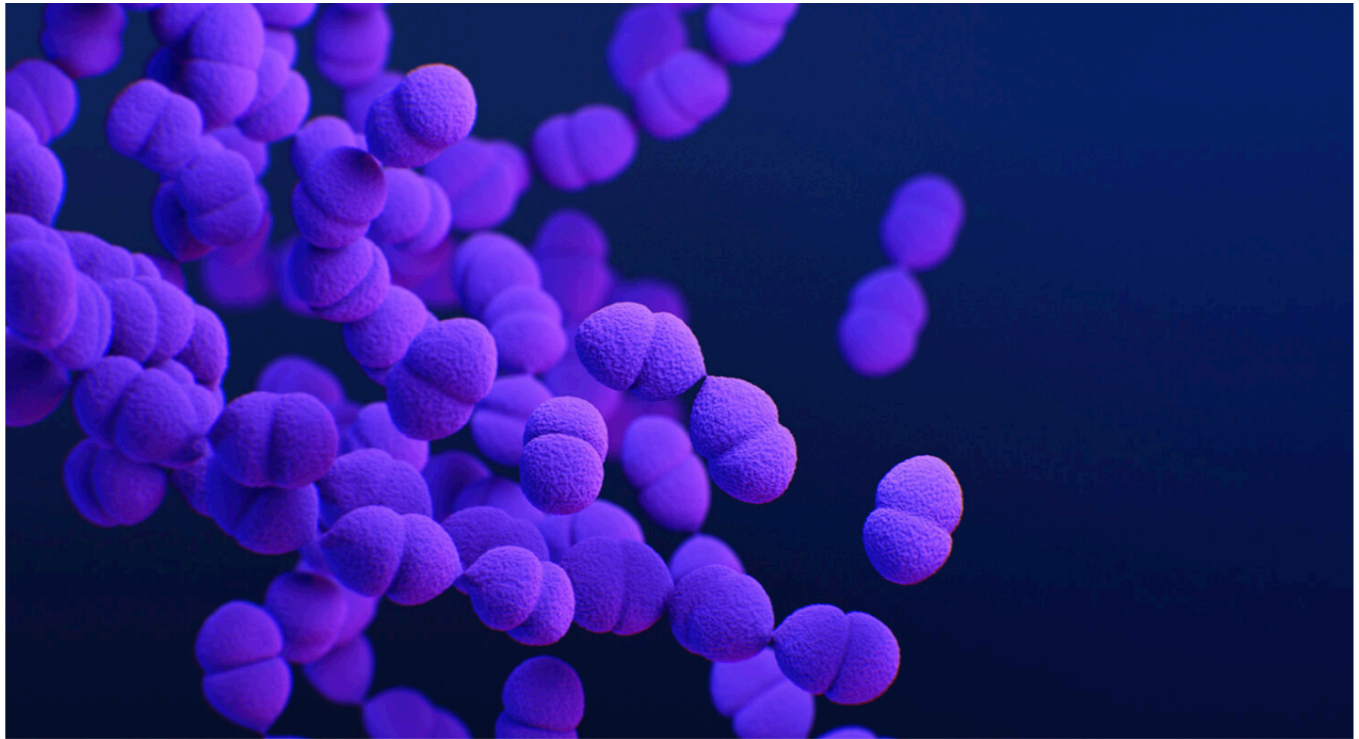
Lots of the infection is due to multi-drug resistant bacteria.

The Question:

Can we identify existing drugs to cure secondary infection for COVID-19?

Mar 25 · 2 min read

First Open Task: Fighting Secondary Effects of COVID-19



Drug repurposing for pseudomonas aeruginosa (**PA01**):

A high-throughput screening data of pseudomonas aeruginosa's activity

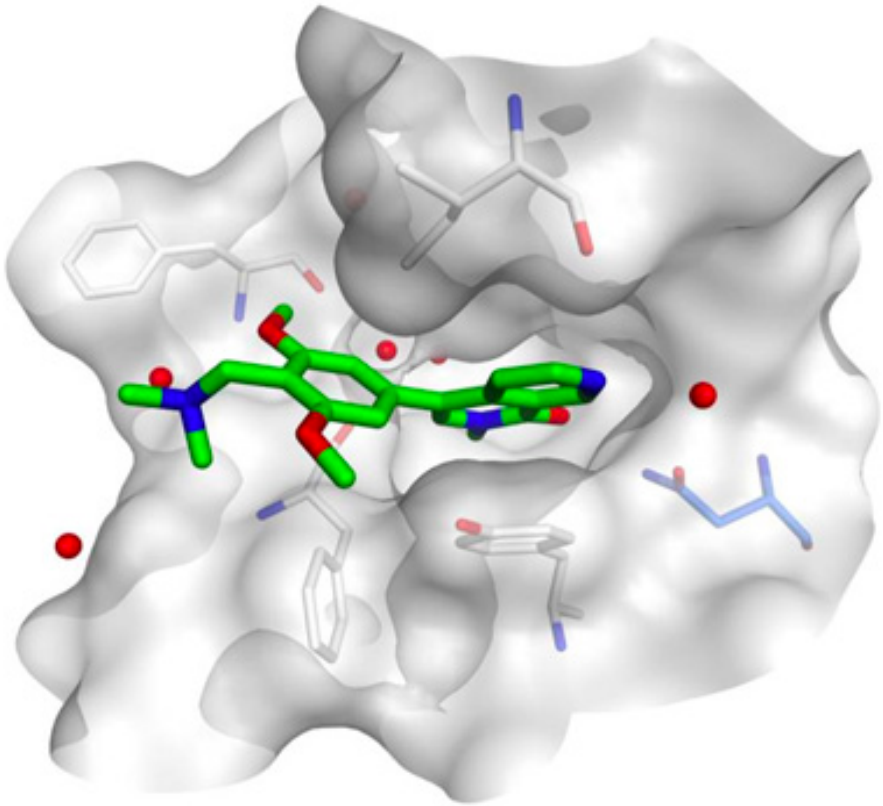
2,335 molecules

Molecules that inhibited growth >80% were labelled as active.

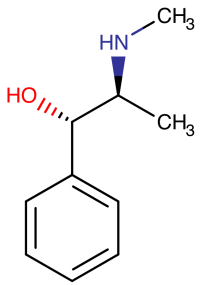
Goal:
Train a model that can predict the HTS data accurately and then use the model to screen a large set of repurposing library.

Expediate the Process!

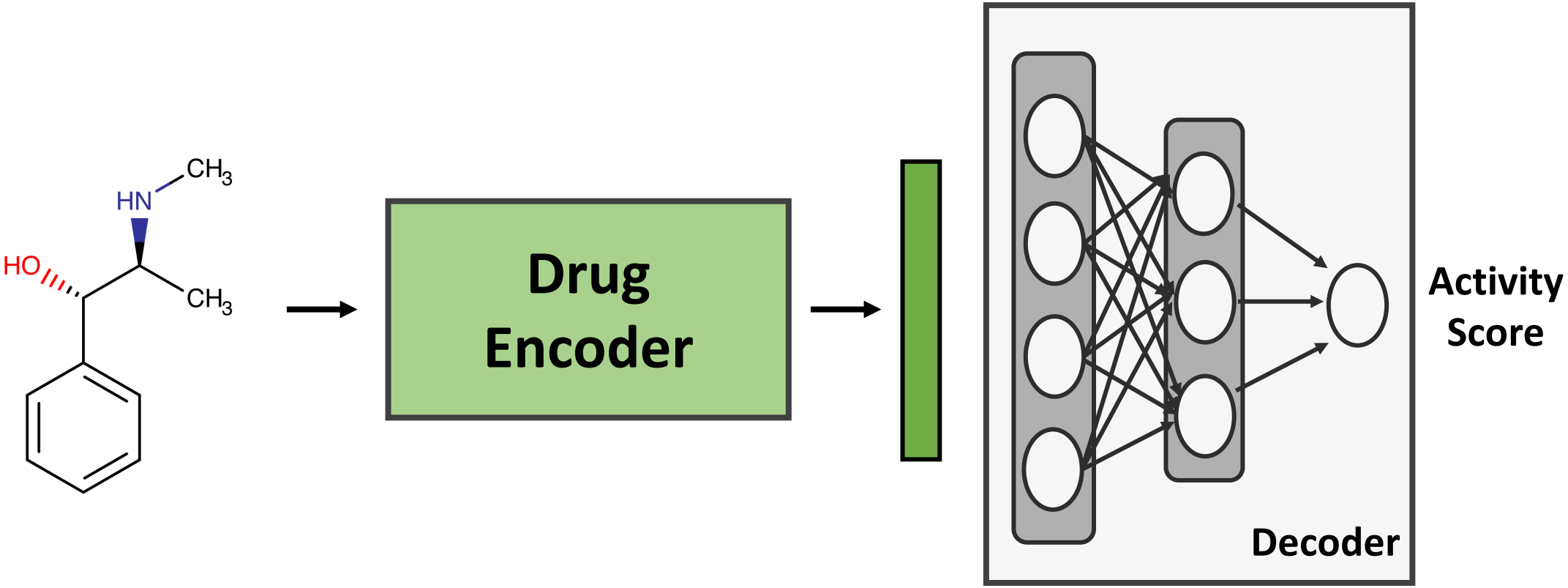
DTI prediction requires
BOTH drug and target information

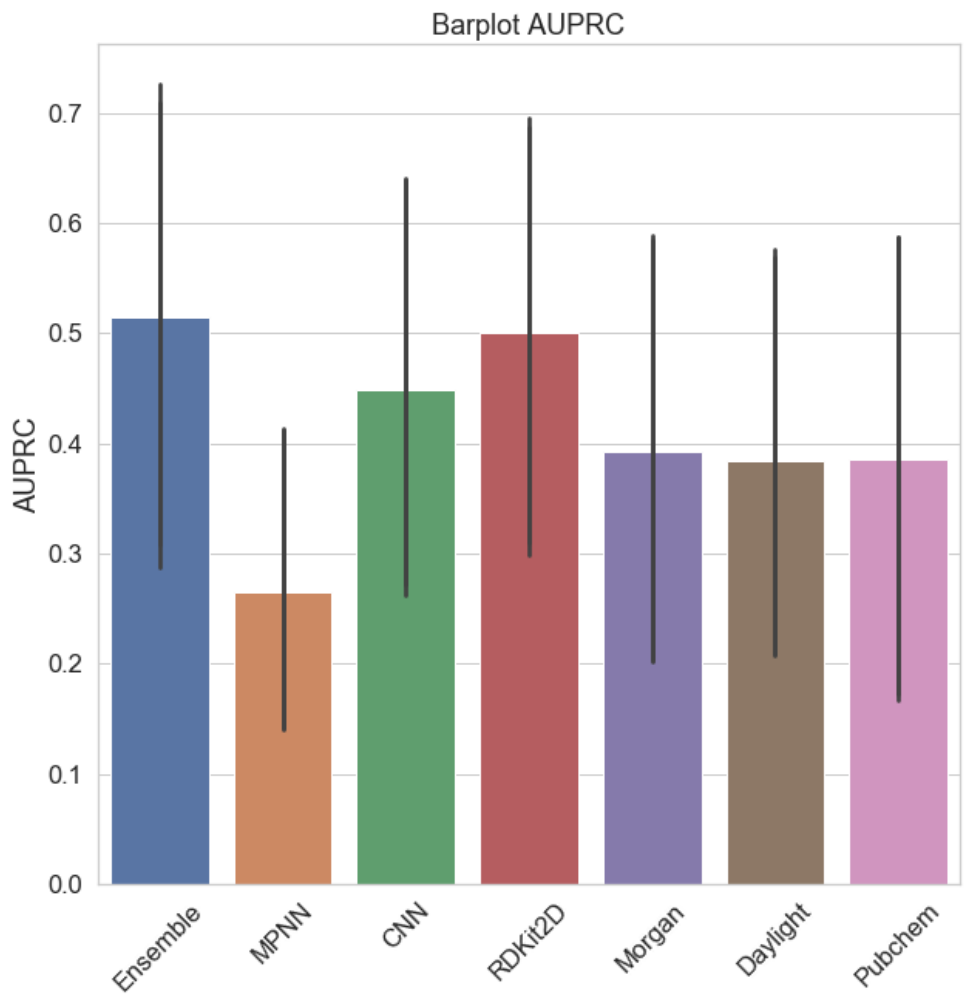


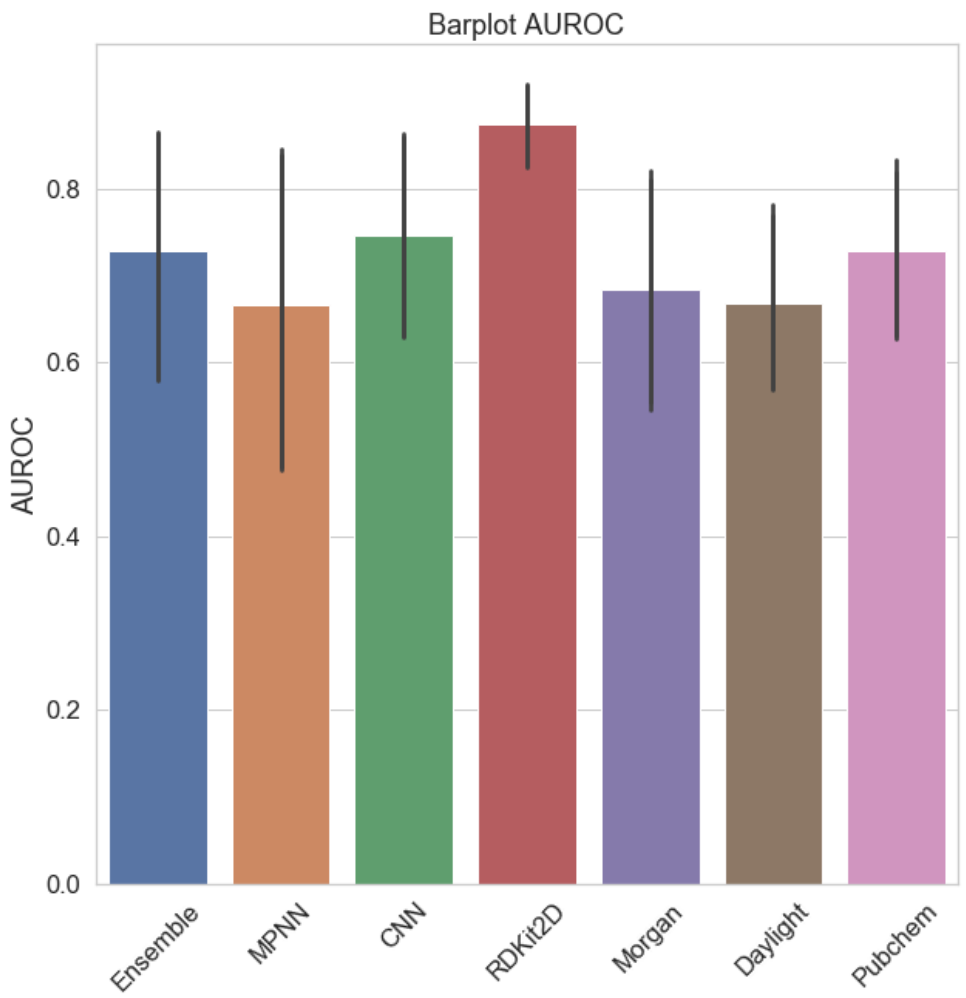
However, now, we only have the activity score for drugs, there is NO protein target for bacteria.



?


HN
CH3
HO
CH3
Drug
Encoder
Decoder
Activity
Score


Barplot AUPRC
AUPRC
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
Ensemble
MPNN
CNN
RDKit2D
Morgan
Daylight
Pubchem


Barplot AUROC
AUROC
0.0
0.2
0.4
0.6
0.8
Ensemble
MPNN
CNN
RDKit2D
Morgan
Daylight
Pubchem

```
>>> import DeepPurpose.property_pred as models
>>> from DeepPurpose.utils import *
>>> from DeepPurpose.dataset import *
>>>
>>> X_drug, drug_names = read_file_repurposing_library(PATH)
>>>
>>> model = models.model_pretrained(MODEL_PATH)
>>> models.repurpose(X_drug, model, drug_names)
```

Rank	Drug Name	Interaction	Probability
1	Elvitegravir	YES	0.92
2	Letermovir	NO	0.44
3	Bictegravir	NO	0.39
4	Dolutegravir	NO	0.26
5	Ibacitabine	NO	0.13
6	Cidofovir	NO	0.00
7	Emtricitabine	NO	0.00
8	Zanamivir	NO	0.00
9	Docosanol	NO	0.00
10	Vidarabine	NO	0.00

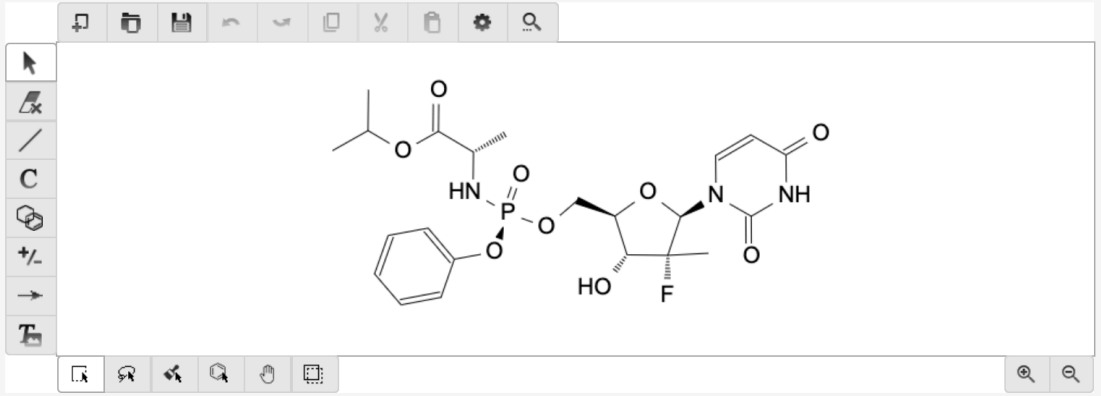
“Elvitegravir has a quinolone moiety and was confirmed to have antibacterial activity in the reverse mutation assay (23.4 µg/plate).”

- FDA NDA of Elvitegravir

AMINO ACID SEQUENCE

```
SGFRKMAFPSGKVEGCMVQVTCGTTTLNGLWLDDVVYCPRHVICTSEDMLNPNYEDLLIRKSNHNFLVQAGNVQLRVIGHSMQNCVLKLKV
DTANPKTPKYKFVRIQPGQTFSVLACYNGSPSGVYQCAMRPNFTIKGSFLNGSCGSVGFNIDYDCVSFCYMHHMELPTGVHAGTDLEGNFY
GPFVDRQTAQAAGTDTTITVNVLAWLYAAVINGDRWFLNRFTTTLNDFNLVAMKYNYEPLTQDHVDILGPLSAQTGIAVLDMCASLKELLQ
```

MOLECULE



AFFINITY PREDICTION MODEL TYPE

MPNN-CNN

ADMET PREDICTION MODEL TYPE

MPNN

CLEAR | SUBMIT

Input Interface

CANONICAL SMILES

```
O=c1[nH]c(=O)n(cc1)[C@@H]2O[C@@H]([C@@H]([C@@]2(F)C)O)COP(=O)(N[C@@H]
(C(OC(C)C)=O)C)Oc3ccccc3
```

BINDING AFFINITY (KD)

749.91 nM

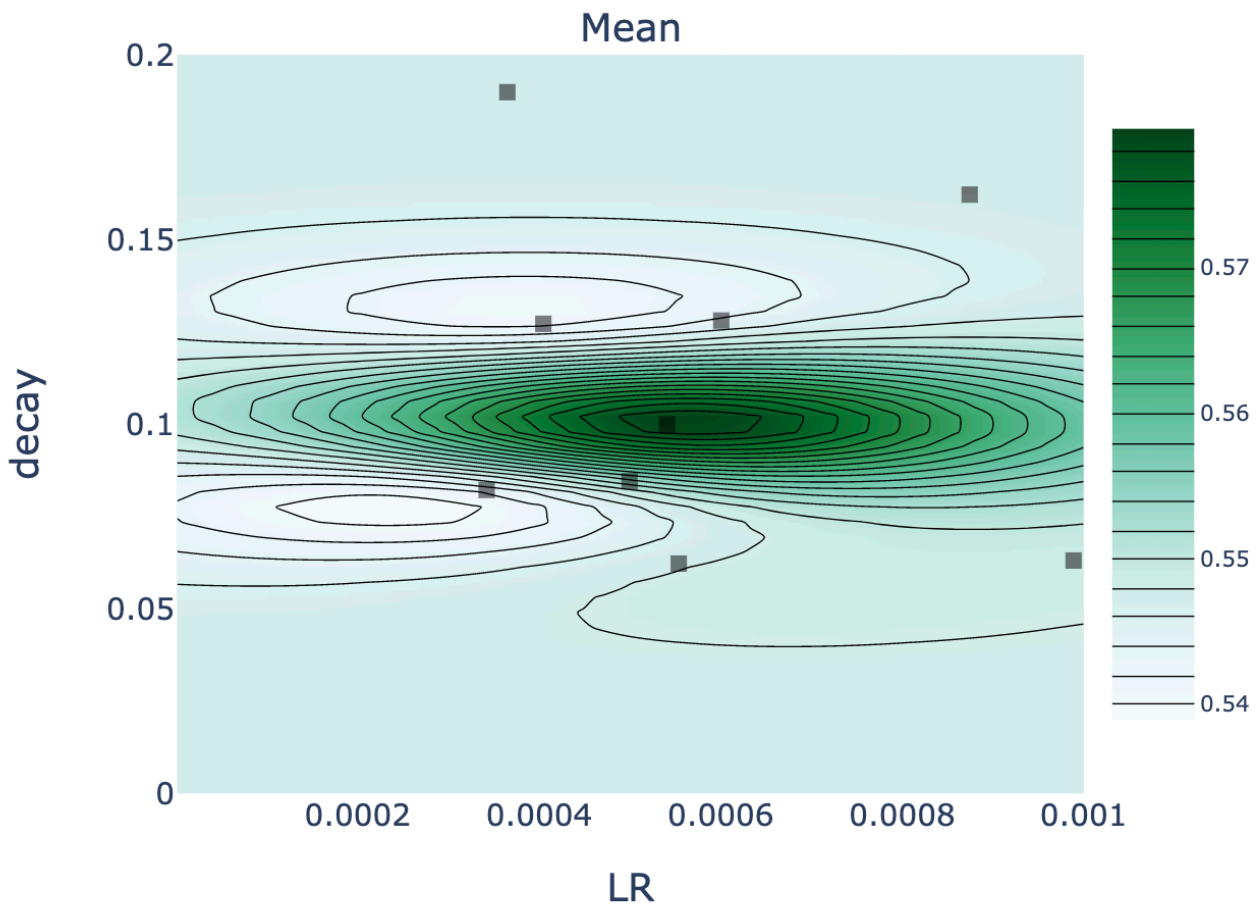
PREDICTED ADMET PROPERTY

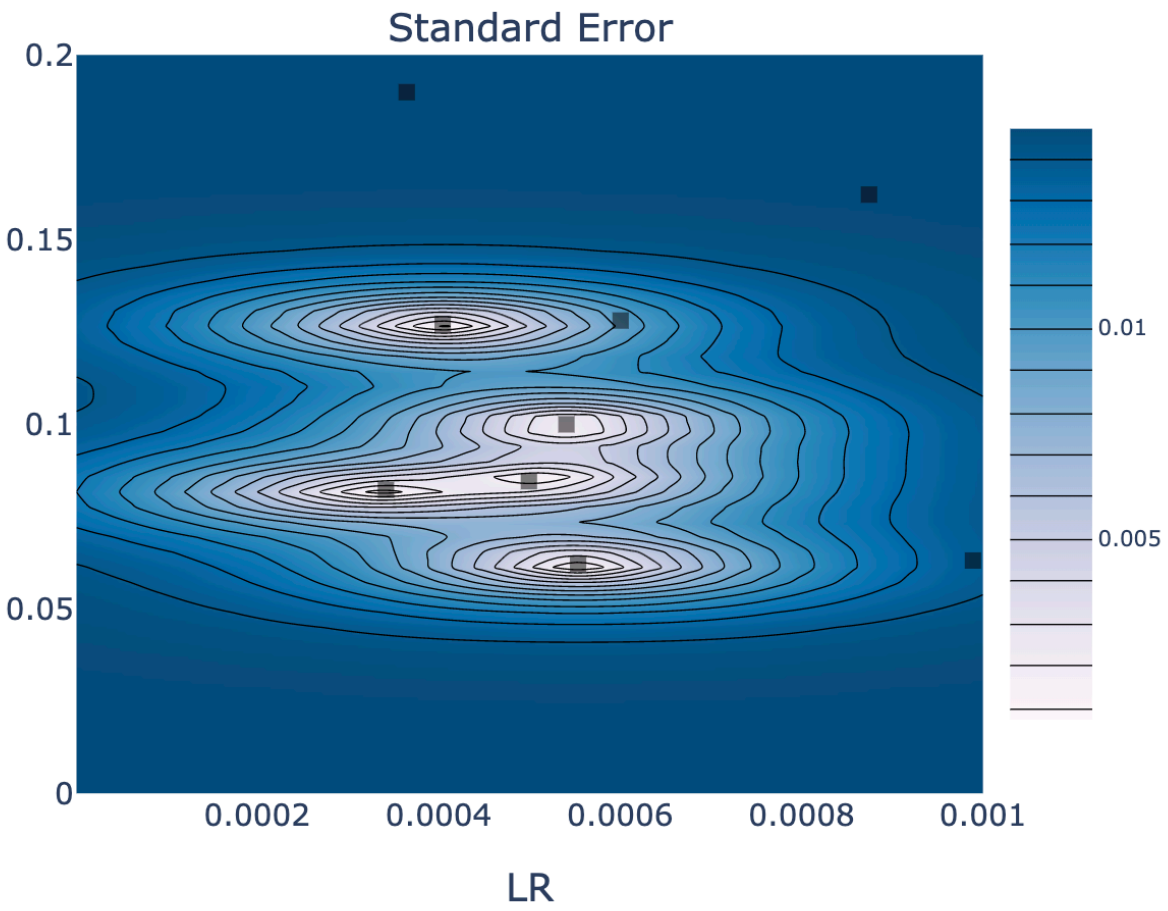
Property	Value
Solubility	-2.88 log mol/L
Lipophilicity	1.21 (log-ratio)
(Absorption) Caco-2	-5.39 cm/s
(Absorption) HIA	67.58 %
(Absorption) Pgp	2.71 %
(Absorption) Bioavailability F20	74.56 %
(Distribution) BBB	57.17 %
(Distribution) PPBR	26.57 %
(Metabolism) CYP2C19	9.52 %
(Metabolism) CYP2D6	1.15 %
(Metabolism) CYP3A4	10.25 %
(Metabolism) CYP1A2	1.63 %
(Metabolism) CYP2C9	1.56 %
(Execretion) Half life	8.28 h
(Execretion) Clearance	8.08 mL/min/kg
Clinical Toxicity	28.47 %

Latency: 0.90s

SCREENSHOT | FLAG

Output Interface





Support Hyperparameter Tuning using Bayesian Optimization!

launch binder

Summary

- **Single line of code to apply state-of-the-art deep learning to do drug repurposing for biomedical scientist.**
- **Flexible framework with 15+ encoders and 50+ models to experiment on drug repurposing, drug target interaction prediction for machine learning researcher.**
- **User-friendly interface with numerous features support.**
- **Enable deep learning accessibility for drug discovery and improve patient care in the end.**

DeepPurpose Deep Dive

Tutorial 1: Training a Drug-Target Interaction Model from Scratch

@KexinHuang5

In this tutorial, we take a deep dive into DeepPurpose and show how it builds a drug-target interac scratch.

Agenda:

- Part I: Overview of DeepPurpose and Data
- Part II: Drug Target Interaction Prediction
  - DeepPurpose Framework
  - Applications to Drug Repurposing and Virtual Screening
  - Pretrained Models
  - Hyperparameter Tuning
  - Model Robustness Evaluation

Let's start!

DeepPurpose Deep Dive

Tutorial 2: Training a Drug Property Prediction Model fror for Assay Data

@KexinHuang5

In this tutorial, we further extends the use cases of DeepPurpose to assay data where information and its affinity score to the protein in the assay.

Agenda:

- Part I: Introduction to Assay Data
- Part II: Drug Property Prediction

Let's start!

master DeepPurpose / DEMO / Go to file Add file

kexinhuang12345 Add files via upload ... 10 days ago History

..

CNN-Binary-Example-DAVIS.ipynb	history clean accident	4 months ago
CNN_CNN-Binary-SARS-CoV-3C...	history clean accident	4 months ago
CNN_Transformer_Davis.ipynb	history clean accident	4 months ago
CNN_Transformer_KIBA-gpu.ipynb	history clean accident	4 months ago
CNN_Transformer_Kiba.ipynb	history clean accident	4 months ago
DeepDTA_Reproduce_KIBA.ipynb	history clean accident	4 months ago
Drug_Property_Pred-Ax-Hyperpar...	Ax hyperparam tuning, verbose option.	3 months ago
Drug_Property_Pred-Ax-Hyperpar...	aicures data perform	3 months ago
Drug_Property_Prediction_Bacteri...	aicures data perform	3 months ago
Drug_Property_Prediction_Bacteri...	aicures data perform	3 months ago
MPNN_AAC_Davis.ipynb	history clean accident	4 months ago
MPNN_AAC_Kiba.ipynb	history clean accident	4 months ago
MPNN_CNN-Binary-SARS-CoV-3...	history clean accident	4 months ago
MPNN_CNN_Davis.ipynb	history clean accident	4 months ago
MPNN_CNN_Kiba.ipynb	history clean accident	4 months ago
Make-DAVIS-Correlation-Figure.i...	New demo on draw correlation plot for DAVIS	4 months ago
Morgan_CNN_Morgan_AAC_Dayli...	history clean accident	4 months ago
Morgan_CNN_Morgan_AAC_Dayli...	history clean accident	4 months ago
Transformer+CNN_BindingDB.ipy...	history clean accident	4 months ago
case-study-I-Drug-Repurposing-...	aicures data perform	3 months ago

Thank you!

Paper

Code

DeepPurpose

https://github.com/kexinhuang12345/DeepPurpose

Star, Share, and Contribute!