Machine Learning for Drug Development

Marinka Zitnik

Department of Biomedical Informatics
Broad Institute of Harvard and MIT
Harvard Data Science Initiative

marinka@hms.harvard.edu
https://zitniklab.hms.harvard.edu

HARVARD MEDICAL SCHOOL

HDSI | Harvard Data Science Initiative

BROAD INSTITUTE

Outline

- ✓ Overview and introduction
- ✓ Part 1: Virtual drug screening and drug repurposing
- ✓ Part 2: Adverse drug effects, drug-drug interactions
- ✓ Part 3: Clinical trial site identification, patient recruitment
- ✓ Part 4: Molecule optimization, molecular graph generation, multimodal graph-to-graph translation
- ✓ Part 5: Molecular property prediction and transformers
- Demos, resources, wrap-up & future directions ☜

Datasets to facilitate algorithmic innovation

Therapeutics are one of most exciting areas for computational scientists. However,

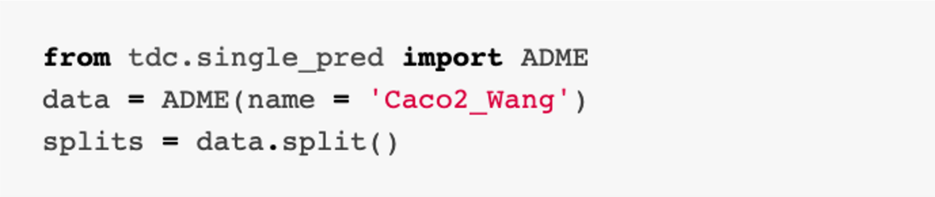
Retrieving, curating, and processing datasets is time-consuming and requires extensive domain expertise

Datasets are scattered around the bio repositories and there is no centralized data repository for a variety of therapeutics

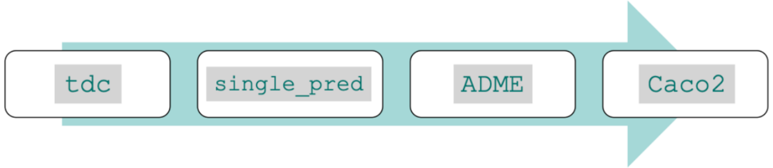
Many tasks are under-explored in AI/ML community because of the lack of data access

THERAPEUTICS DATA COMMONS

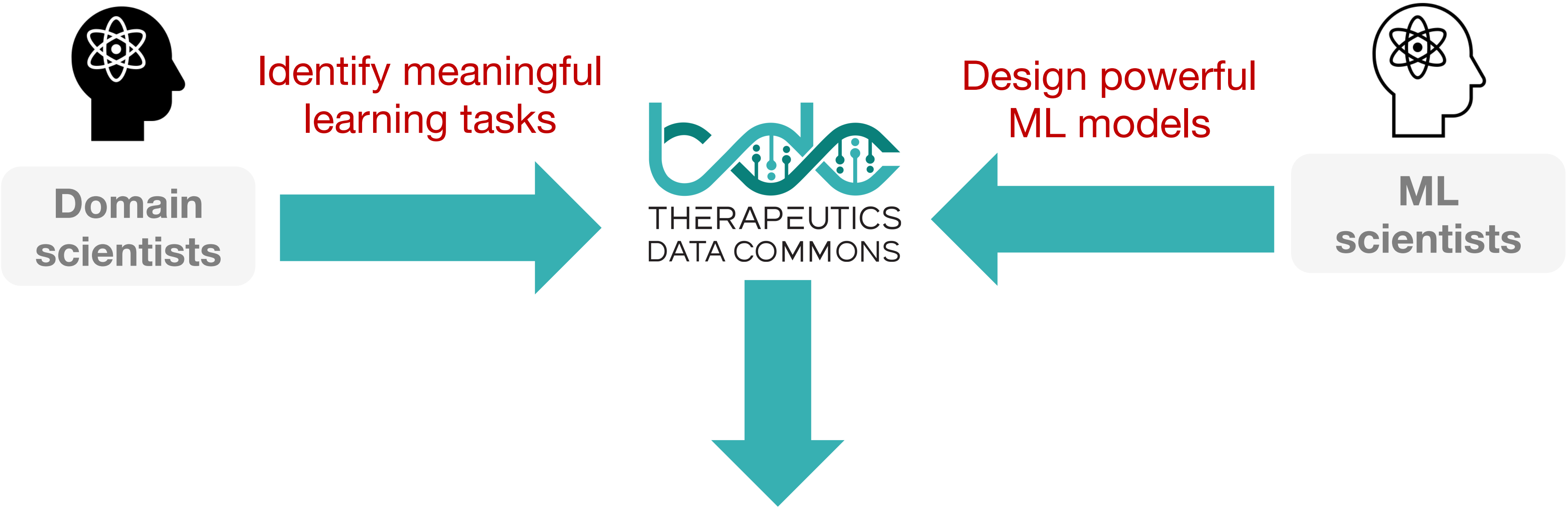
- **Open-Source ML Datasets for Therapeutics:**
  - Wide range of tasks: target discovery, activity screening, efficacy, safety, manufacturing
  - Wide range of products: small molecules, antibodies, vaccine, miRNA
- **Numerous Data Functions:**
  - Extensive data functions
  - Model evaluation, data processing and splits, molecule generation oracles, and much more
- **3 Lines of Code:**
  - Minimum package dependency, lightweight loaders



```
from tdc.single_pred import ADME
data = ADME(name = 'Caco2_Wang')
splits = data.split()
```



Our Vision for TDC



Domain scientists

Identify meaningful learning tasks

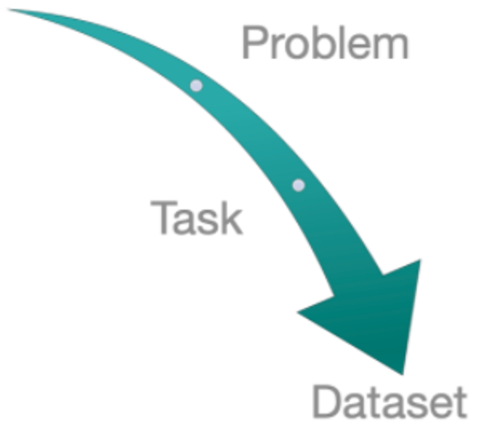
THERAPEUTICS DATA COMMONS

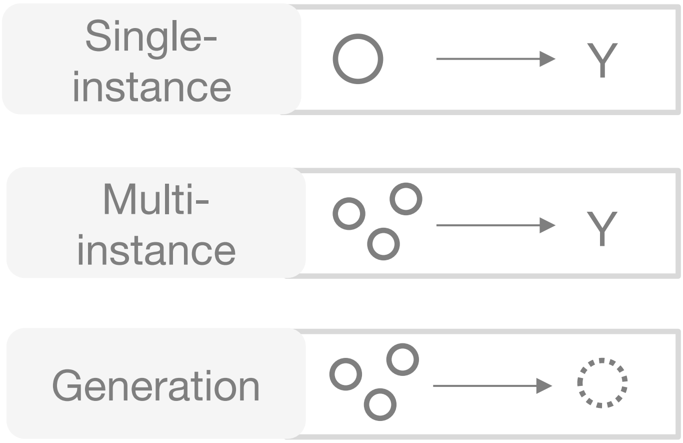
Design powerful ML models

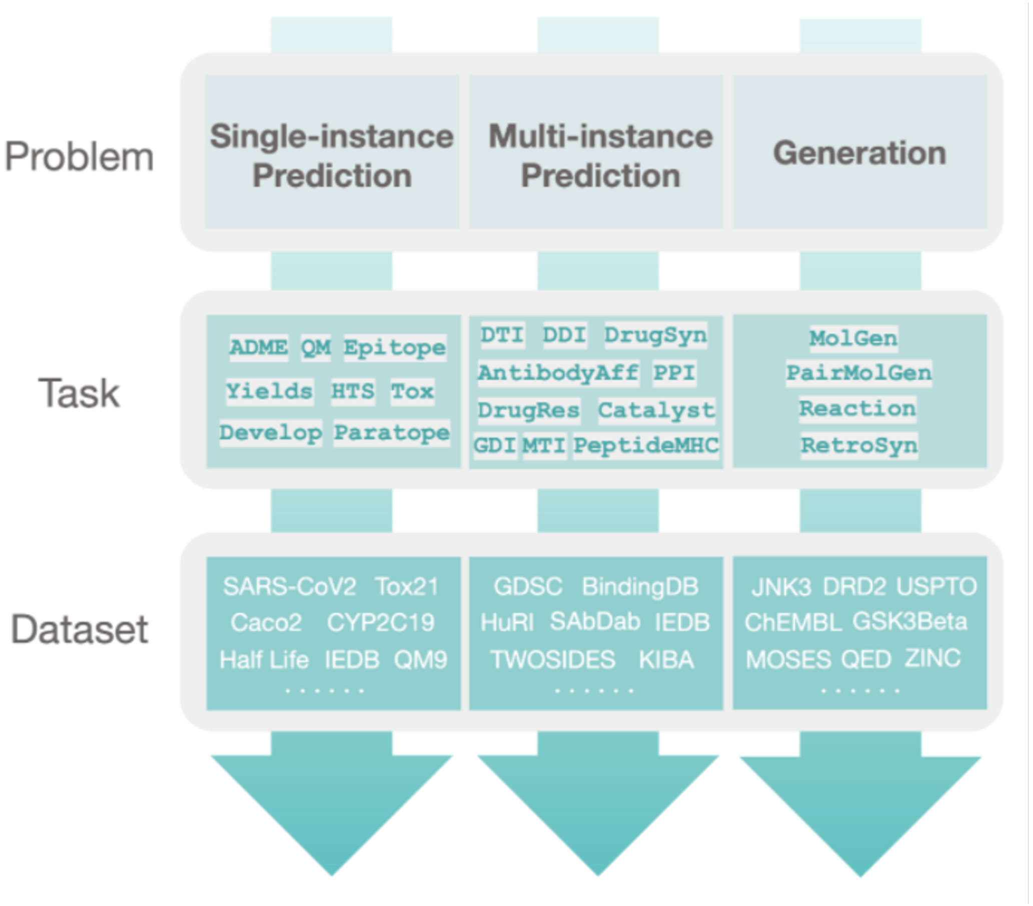
ML scientists

Advancing algorithms for key therapeutics problems

Modular Structure of TDC

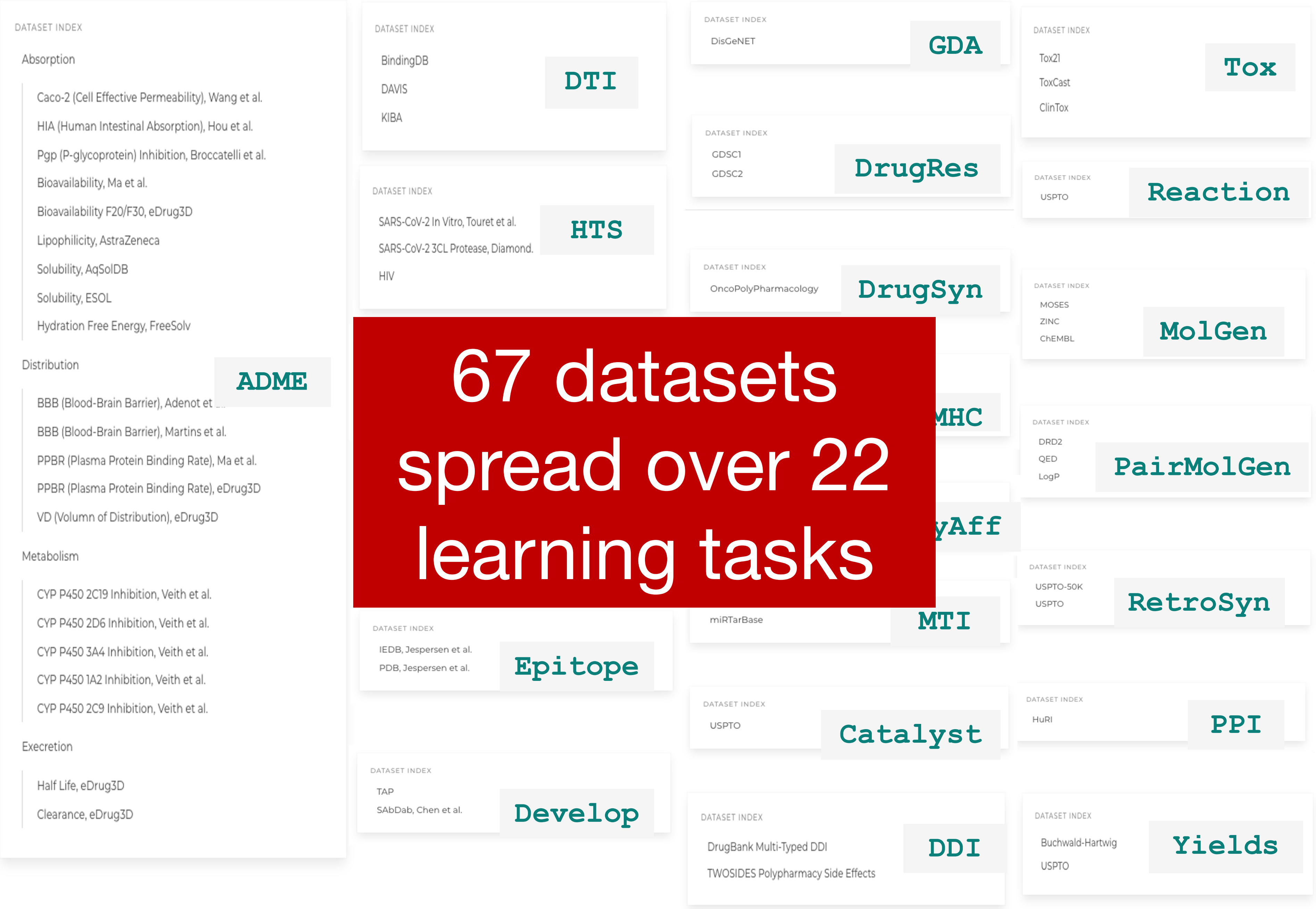






Problem	Single-instance Prediction	Multi-instance Prediction	Generation
Task	ADME QM Epitope Yields HTS Tox Develop Paratope	DTI DDI DrugSyn AntibodyAff PPI DrugRes Catalyst GDI MTI PeptideMHC	MolGen PairMolGen Reaction RetroSyn
Dataset	SARS-CoV2 Tox21 Caco2 CYP2C19 Half Life IEDB QM9	GDSC BindingDB HuRI SAbDab IEDB TWOSIDES KIBA	JNK3 DRD2 USPTO ChEMBL GSK3Beta MOSES QED ZINC . . . . . .

67 datasets spread over 22 learning tasks



DATASET INDEX

Absorption

- Caco-2 (Cell Effective Permeability), Wang et al.
- HIA (Human Intestinal Absorption), Hou et al.
- Pgp (P-glycoprotein) Inhibition, Broccatelli et al.
- Bioavailability, Ma et al.
- Bioavailability F20/F30, eDrug3D
- Lipophilicity, AstraZeneca
- Solubility, AqSolDB
- Solubility, ESOL
- Hydration Free Energy, FreeSolv

Distribution

- BBB (Blood-Brain Barrier), Adenot et ...
- BBB (Blood-Brain Barrier), Martins et al.
- PPBR (Plasma Protein Binding Rate), Ma et al.
- PPBR (Plasma Protein Binding Rate), eDrug3D
- VD (Volumn of Distribution), eDrug3D

Metabolism

- CYP P450 2C19 Inhibition, Veith et al.
- CYP P450 2D6 Inhibition, Veith et al.
- CYP P450 3A4 Inhibition, Veith et al.
- CYP P450 1A2 Inhibition, Veith et al.
- CYP P450 2C9 Inhibition, Veith et al.

Execretion

- Half Life, eDrug3D
- Clearance, eDrug3D

ADME

DATASET INDEX

BindingDB

DAVIS

KIBA

DTI

DATASET INDEX

SARS-CoV-2 In Vitro, Touret et al.

SARS-CoV-2 3CL Protease, Diamond.

HIV

HTS

DATASET INDEX

IEDB, Jespersen et al.

PDB, Jespersen et al.

Epitope

DATASET INDEX

TAP

SAbDab, Chen et al.

Develop

DATASET INDEX

DisGeNET

GDA

DATASET INDEX

GDSC1

GDSC2

DrugRes

DATASET INDEX

OncoPolyPharmacology

DrugSyn

MHC

yAff

miRTarBase

MTI

DATASET INDEX

USPTO

Catalyst

DATASET INDEX

DrugBank Multi-Typed DDI

TWOSIDES Polypharmacy Side Effects

DDI

DATASET INDEX

Tox21

ToxCast

ClinTox

Tox

DATASET INDEX

USPTO

Reaction

DATASET INDEX

MOSES

ZINC

ChEMBL

MolGen

DATASET INDEX

DRD2

QED

LogP

PairMolGen

DATASET INDEX

USPTO-50K

USPTO

RetroSyn

DATASET INDEX

HuRI

PPI

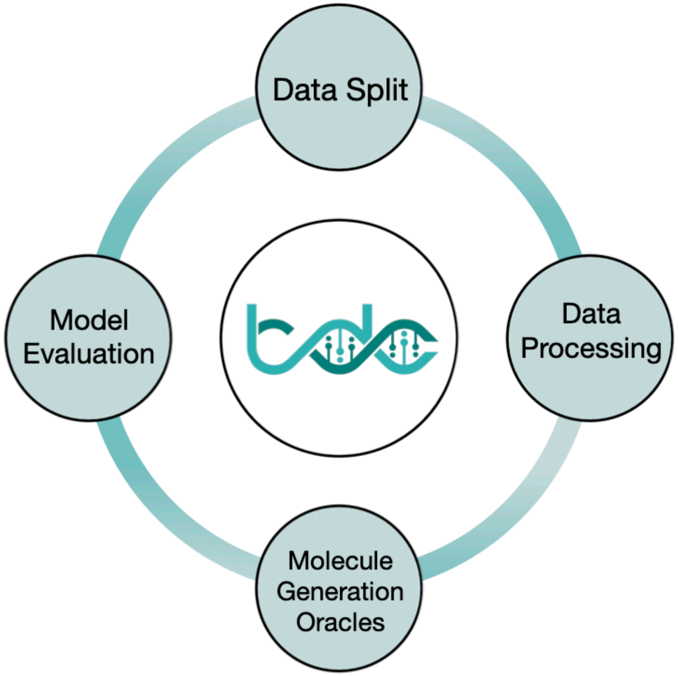
DATASET INDEX

Buchwald-Hartwig

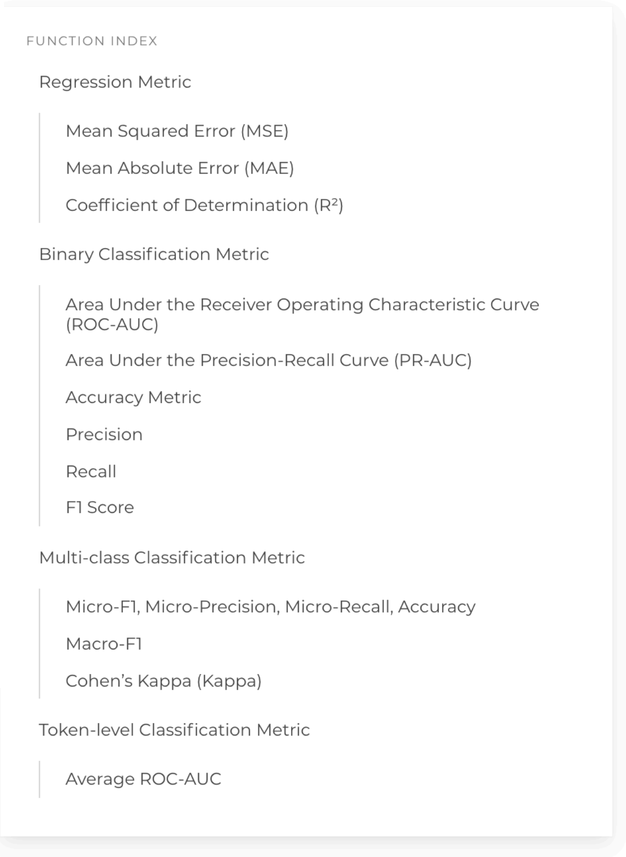
USPTO

Yields

Data Functions to Support Your Research



Model performance evaluators


FUNCTION INDEX

Regression Metric

- Mean Squared Error (MSE)
- Mean Absolute Error (MAE)
- Coefficient of Determination (R^2)

Binary Classification Metric

- Area Under the Receiver Operating Characteristic Curve (ROC-AUC)
- Area Under the Precision-Recall Curve (PR-AUC)
- Accuracy Metric
- Precision
- Recall
- F1 Score

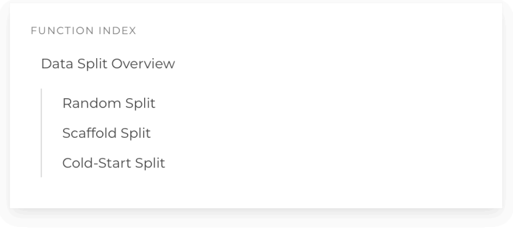
Multi-class Classification Metric

- Micro-F1, Micro-Precision, Micro-Recall, Accuracy
- Macro-F1
- Cohen's Kappa (Kappa)

Token-level Classification Metric

- Average ROC-AUC

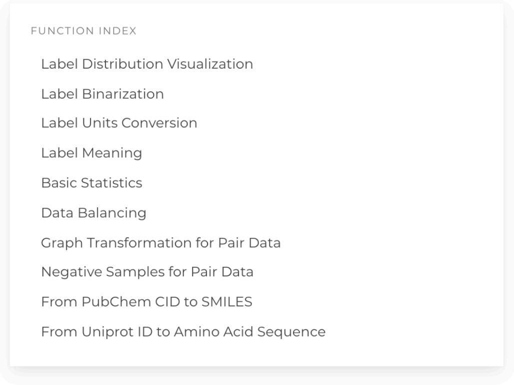
A variety of data splits


FUNCTION INDEX

Data Split Overview

- Random Split
- Scaffold Split
- Cold-Start Split

Data processing helpers


FUNCTION INDEX

Label Distribution Visualization

Label Binarization

Label Units Conversion

Label Meaning

Basic Statistics

Data Balancing

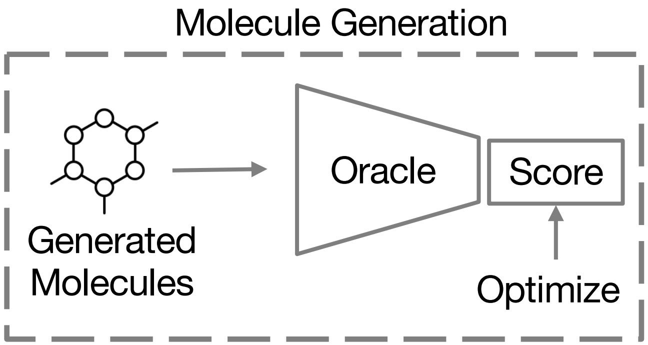
Graph Transformation for Pair Data

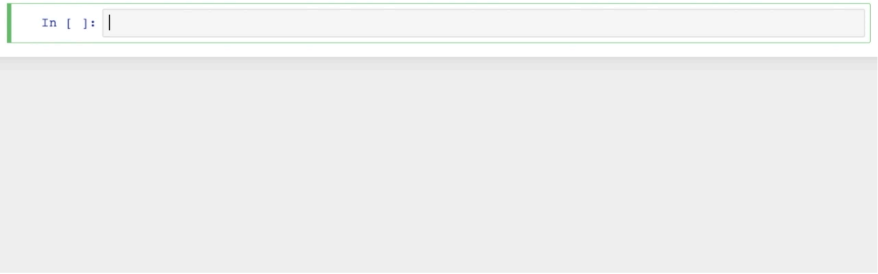
Negative Samples for Pair Data

From PubChem CID to SMILES

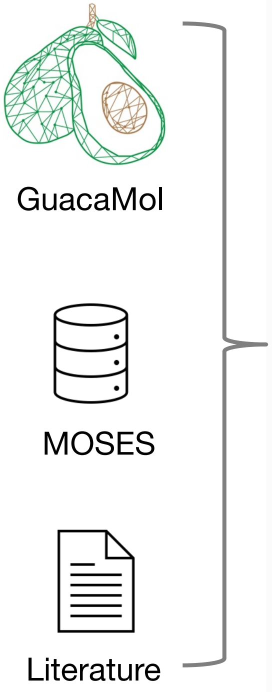
From Uniprot ID to Amino Acid Sequence

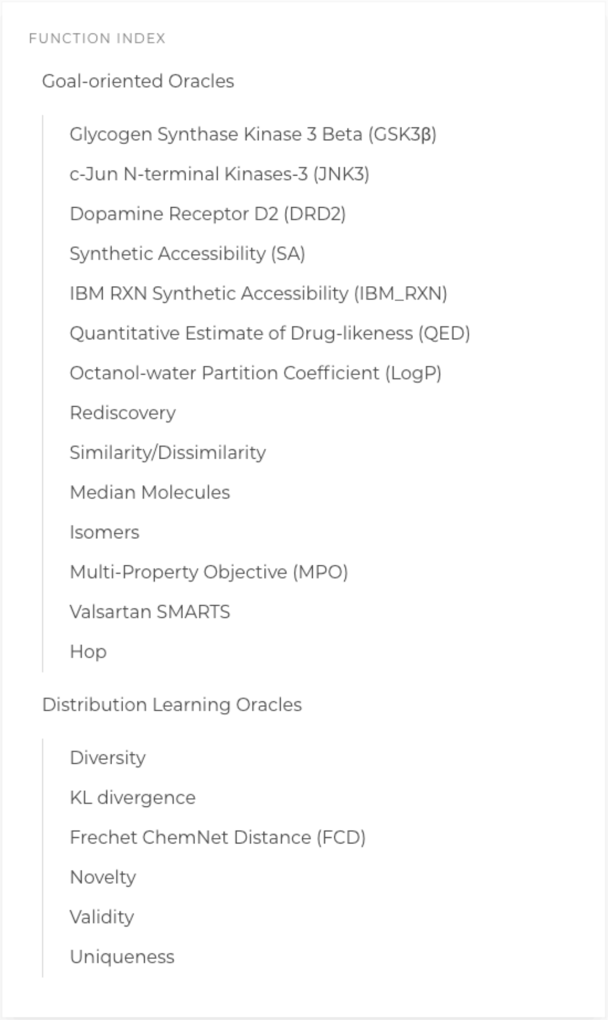
Molecule Generation Oracles





```
In [ ]:
```





FUNCTION INDEX

Goal-oriented Oracles

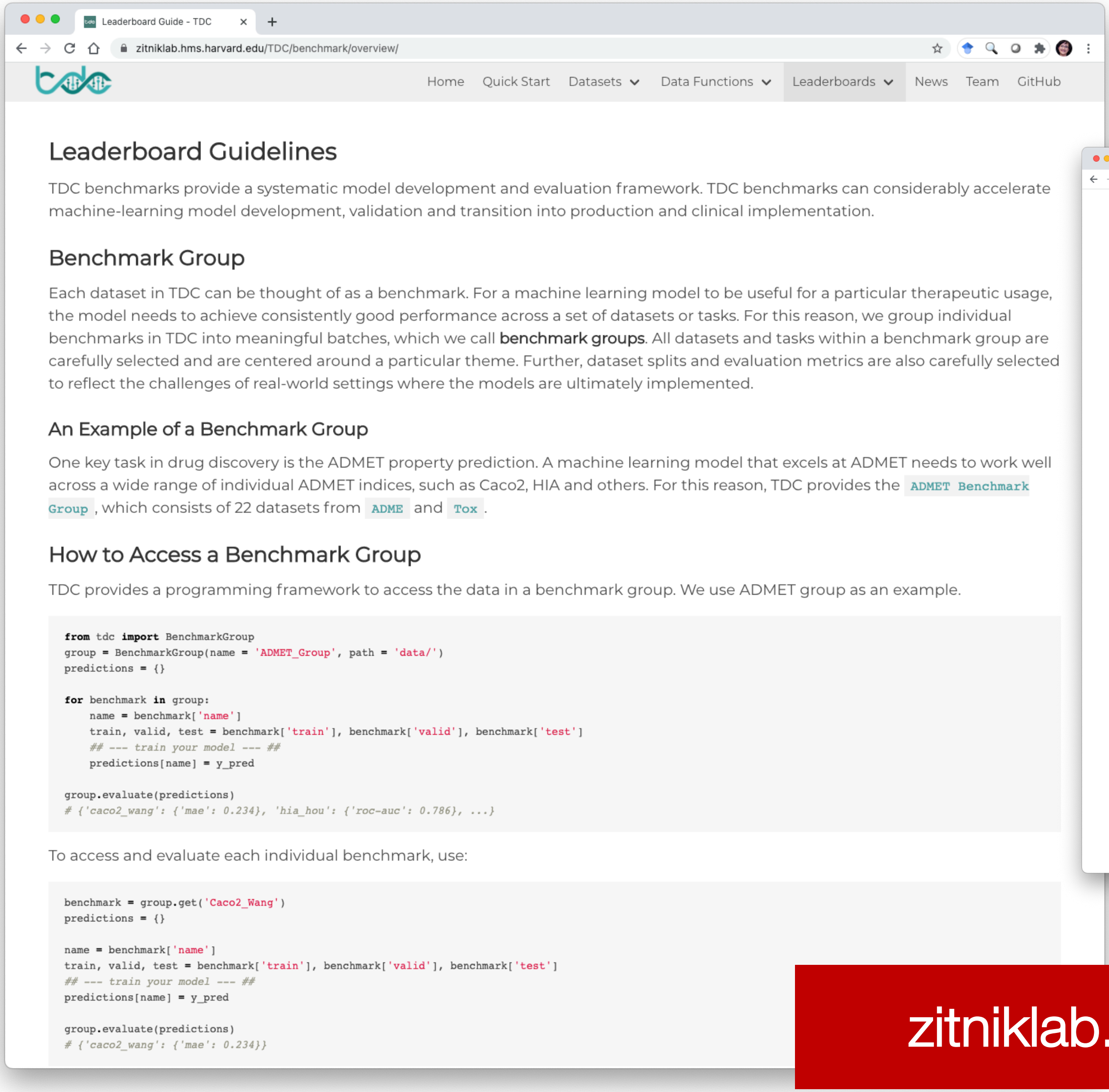
- Glycogen Synthase Kinase 3 Beta (GSK3β)
- c-Jun N-terminal Kinases-3 (JNK3)
- Dopamine Receptor D2 (DRD2)
- Synthetic Accessibility (SA)
- IBM RXN Synthetic Accessibility (IBM_RXN)
- Quantitative Estimate of Drug-likeness (QED)
- Octanol-water Partition Coefficient (LogP)
- Rediscovery
- Similarity/Dissimilarity
- Median Molecules
- Isomers
- Multi-Property Objective (MPO)
- Valsartan SMARTS
- Hop

Distribution Learning Oracles

- Diversity
- KL divergence
- Frechet ChemNet Distance (FCD)
- Novelty
- Validity
- Uniqueness

GuacaMol: Benchmarking Models for de Novo Molecular Design, J. Chem. Inf. Model., 2019
MOSES: A Benchmarking Platform for Molecular Generation Models, Frontiers in Pharmacology, 2020

Leaderboards: Submit your Models



Leaderboard Guidelines

TDC benchmarks provide a systematic model development and evaluation framework. TDC benchmarks can considerably accelerate machine-learning model development, validation and transition into production and clinical implementation.

Benchmark Group

Each dataset in TDC can be thought of as a benchmark. For a machine learning model to be useful for a particular therapeutic usage, the model needs to achieve consistently good performance across a set of datasets or tasks. For this reason, we group individual benchmarks in TDC into meaningful batches, which we call **benchmark groups**. All datasets and tasks within a benchmark group are carefully selected and are centered around a particular theme. Further, dataset splits and evaluation metrics are also carefully selected to reflect the challenges of real-world settings where the models are ultimately implemented.

An Example of a Benchmark Group

One key task in drug discovery is the ADMET property prediction. A machine learning model that excels at ADMET needs to work well across a wide range of individual ADMET indices, such as Caco2, HIA and others. For this reason, TDC provides the `ADMET Benchmark Group`, which consists of 22 datasets from `ADME` and `Tox`.

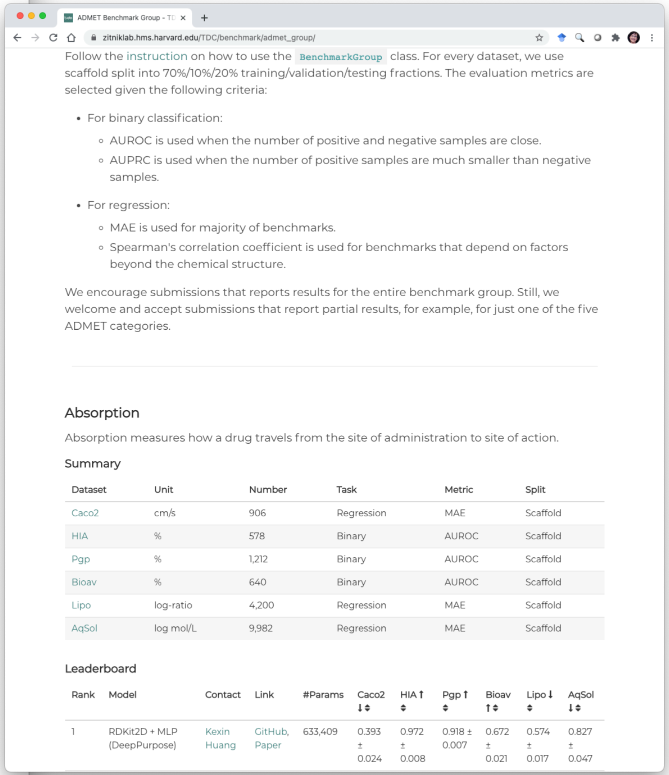
How to Access a Benchmark Group

TDC provides a programming framework to access the data in a benchmark group. We use ADMET group as an example.

```
from tdc import BenchmarkGroup
group = BenchmarkGroup(name = 'ADMET_Group', path = 'data/')
predictions = {}

for benchmark in group:
    name = benchmark['name']
    train, valid, test = benchmark['train'], benchmark['valid'], benchmark['test']
    ## --- train your model --- ##
    predictions[name] = y_pred

group.evaluate(predictions)
# {'caco2_wang': {'mae': 0.234}, 'hia_hou': {'roc-auc': 0.786}, ...}
```

To access and evaluate each individual benchmark, use:

```
benchmark = group.get('Caco2_Wang')
predictions = {}

name = benchmark['name']
train, valid, test = benchmark['train'], benchmark['valid'], benchmark['test']
## --- train your model --- ##
predictions[name] = y_pred

group.evaluate(predictions)
# {'caco2_wang': {'mae': 0.234}}
```


ADMET Benchmark Group - TD
zitniklab.hms.harvard.edu/TDC/benchmark/admet_group/

Follow the instruction on how to use the `BenchmarkGroup` class. For every dataset, we use scaffold split into 70%/10%/20% training/validation/testing fractions. The evaluation metrics are selected given the following criteria:

- For binary classification:
  - AUROC is used when the number of positive and negative samples are close.
  - AUPRC is used when the number of positive samples are much smaller than negative samples.
- For regression:
  - MAE is used for majority of benchmarks.
  - Spearman's correlation coefficient is used for benchmarks that depend on factors beyond the chemical structure.

We encourage submissions that reports results for the entire benchmark group. Still, we welcome and accept submissions that report partial results, for example, for just one of the five ADMET categories.

Absorption

Absorption measures how a drug travels from the site of administration to site of action.

Summary

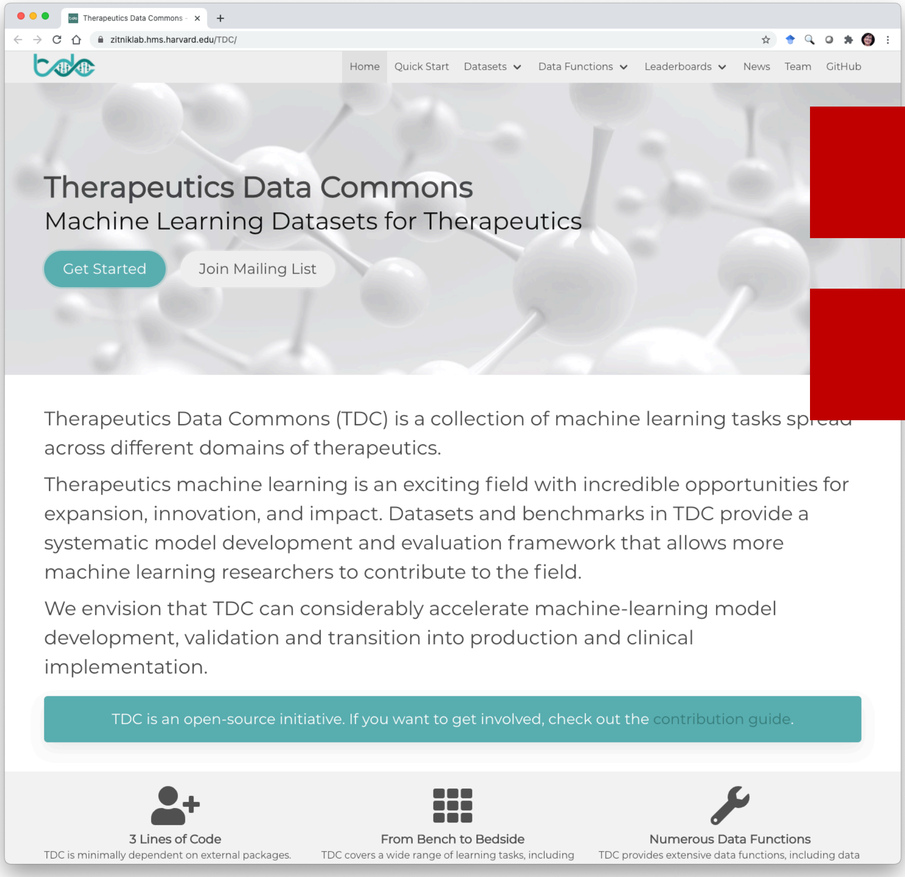
Dataset	Unit	Number	Task	Metric	Split
Caco2	cm/s	906	Regression	MAE	Scaffold
HIA	%	578	Binary	AUROC	Scaffold
Pgp	%	1,212	Binary	AUROC	Scaffold
Bioav	%	640	Binary	AUROC	Scaffold
Lipo	log-ratio	4,200	Regression	MAE	Scaffold
AqSol	log mol/L	9,982	Regression	MAE	Scaffold

Leaderboard

Rank	Model	Contact	Link	#Params	Caco2 ↓	HIA ↑	Pgp ↑	Bioav ↑	Lipo ↓	AqSol ↓
1	RDKit2D + MLP (DeepPurpose)	Kexin Huang	GitHub, Paper	633,409	0.393 ± 0.024	0.972 ± 0.008	0.918 ± 0.007	0.672 ± 0.021	0.574 ± 0.017	0.827 ± 0.047

zitniklab.hms.harvard.edu/TDC

You Are Invited to Join TDC! TDC is an Open-Source, Community Effort



Therapeutics Data Commons

Machine Learning Datasets for Therapeutics

Therapeutics Data Commons (TDC) is a collection of machine learning tasks spread across different domains of therapeutics.

Therapeutics machine learning is an exciting field with incredible opportunities for expansion, innovation, and impact. Datasets and benchmarks in TDC provide a systematic model development and evaluation framework that allows more machine learning researchers to contribute to the field.

We envision that TDC can considerably accelerate machine-learning model development, validation and transition into production and clinical implementation.

TDC is an open-source initiative. If you want to get involved, check out the contribution guide.

3 Lines of Code

TDC is minimally dependent on external packages.

From Bench to Bedside

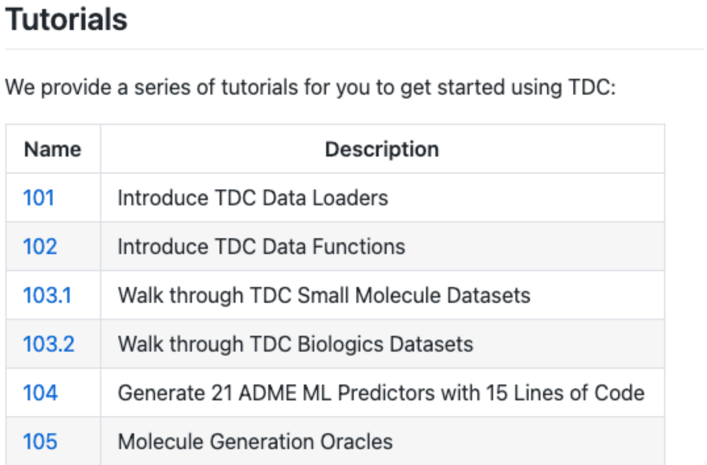
TDC covers a wide range of learning tasks, including

Numerous Data Functions

TDC provides extensive data functions, including data

zitniklab.hms.harvard.edu/TDC

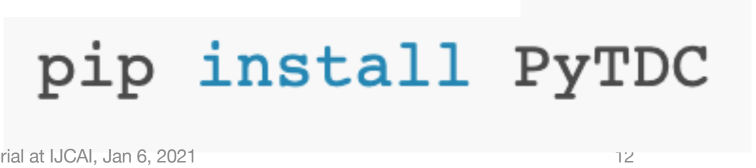
github.com/mims-harvard/TDC



Tutorials

We provide a series of tutorials for you to get started using TDC:

Name	Description
101	Introduce TDC Data Loaders
102	Introduce TDC Data Functions
103.1	Walk through TDC Small Molecule Datasets
103.2	Walk through TDC Biologics Datasets
104	Generate 21 ADME ML Predictors with 15 Lines of Code
105	Molecule Generation Oracles

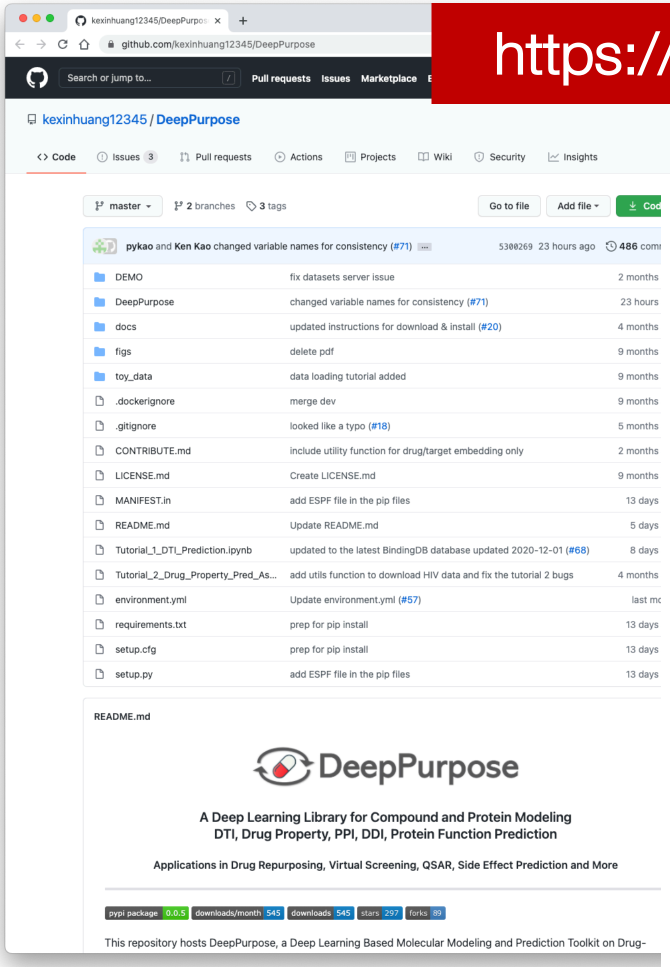


```
pip install PyTDC
```

Demos, tools, and implementations

DeepPurpose: Deep Learning Library for Compound and Protein Modeling DTI, Drug Property, PPI, DDI, Protein Function Prediction

https://github.com/kexinhuang12345/DeepPurpose



kexinhuang12345 / DeepPurpose

2 branches 3 tags

pykao and Ken Kao changed variable names for consistency (#71) … 5300269 23 hours ago 486 comm

DEMO	fix datasets server issue	2 months
DeepPurpose	changed variable names for consistency (#71)	23 hours
docs	updated instructions for download & install (#20)	4 months
figs	delete pdf	9 months
toy_data	data loading tutorial added	9 months
.dockerignore	merge dev	9 months
.gitignore	looked like a typo (#18)	5 months
CONTRIBUTE.md	include utility function for drug/target embedding only	2 months
LICENSE.md	Create LICENSE.md	9 months
MANIFEST.in	add ESPF file in the pip files	13 days
README.md	Update README.md	5 days
Tutorial_1_DTI_Prediction.ipynb	updated to the latest BindingDB database updated 2020-12-01 (#68)	8 days
Tutorial_2_Drug_Property_Pred_As…	add utils function to download HIV data and fix the tutorial 2 bugs	4 months
environment.yml	Update environment.yml (#57)	last mo
requirements.txt	prep for pip install	13 days
setup.cfg	prep for pip install	13 days
setup.py	add ESPF file in the pip files	13 days

README.md

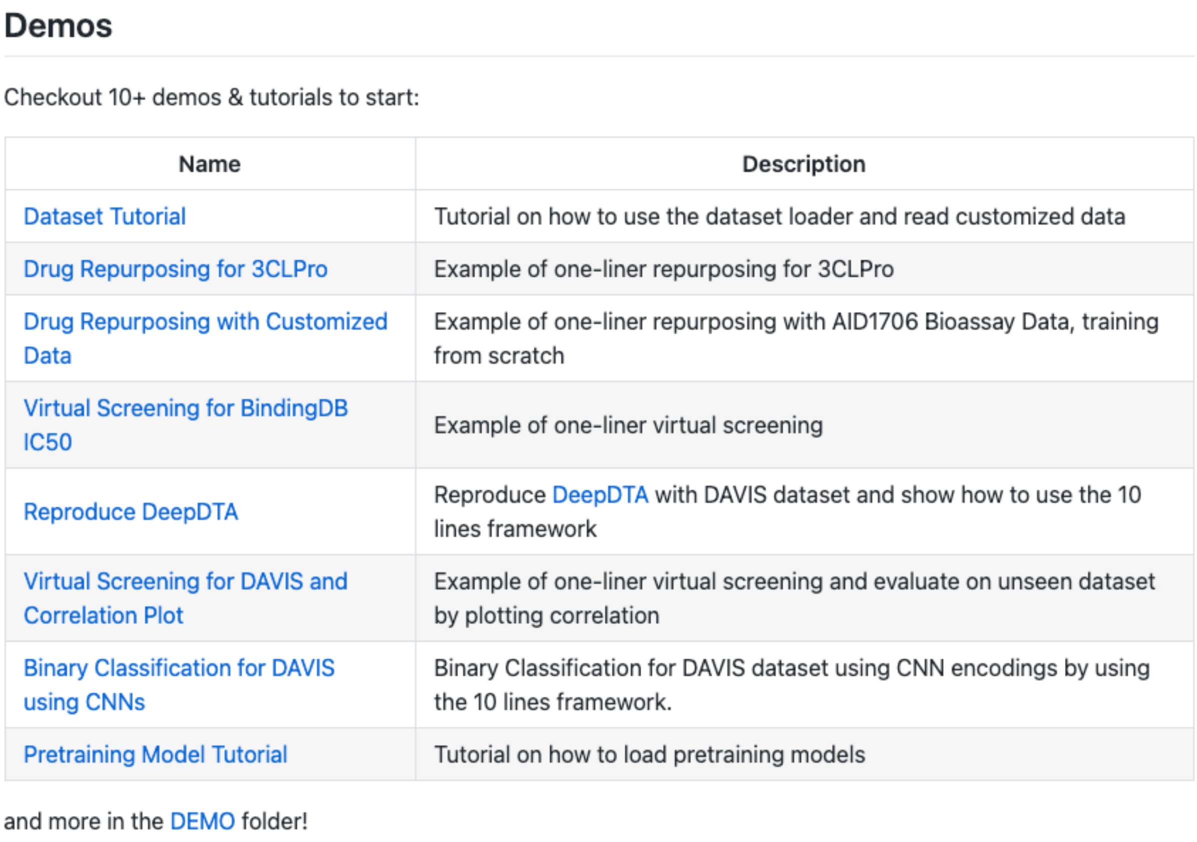
DeepPurpose

A Deep Learning Library for Compound and Protein Modeling
DTI, Drug Property, PPI, DDI, Protein Function Prediction

Applications in Drug Repurposing, Virtual Screening, QSAR, Side Effect Prediction and More

pypi package 0.0.5 | downloads/month 545 | downloads 545 | stars 297 | forks 89

This repository hosts DeepPurpose, a Deep Learning Based Molecular Modeling and Prediction Toolkit on Drug-



Demos

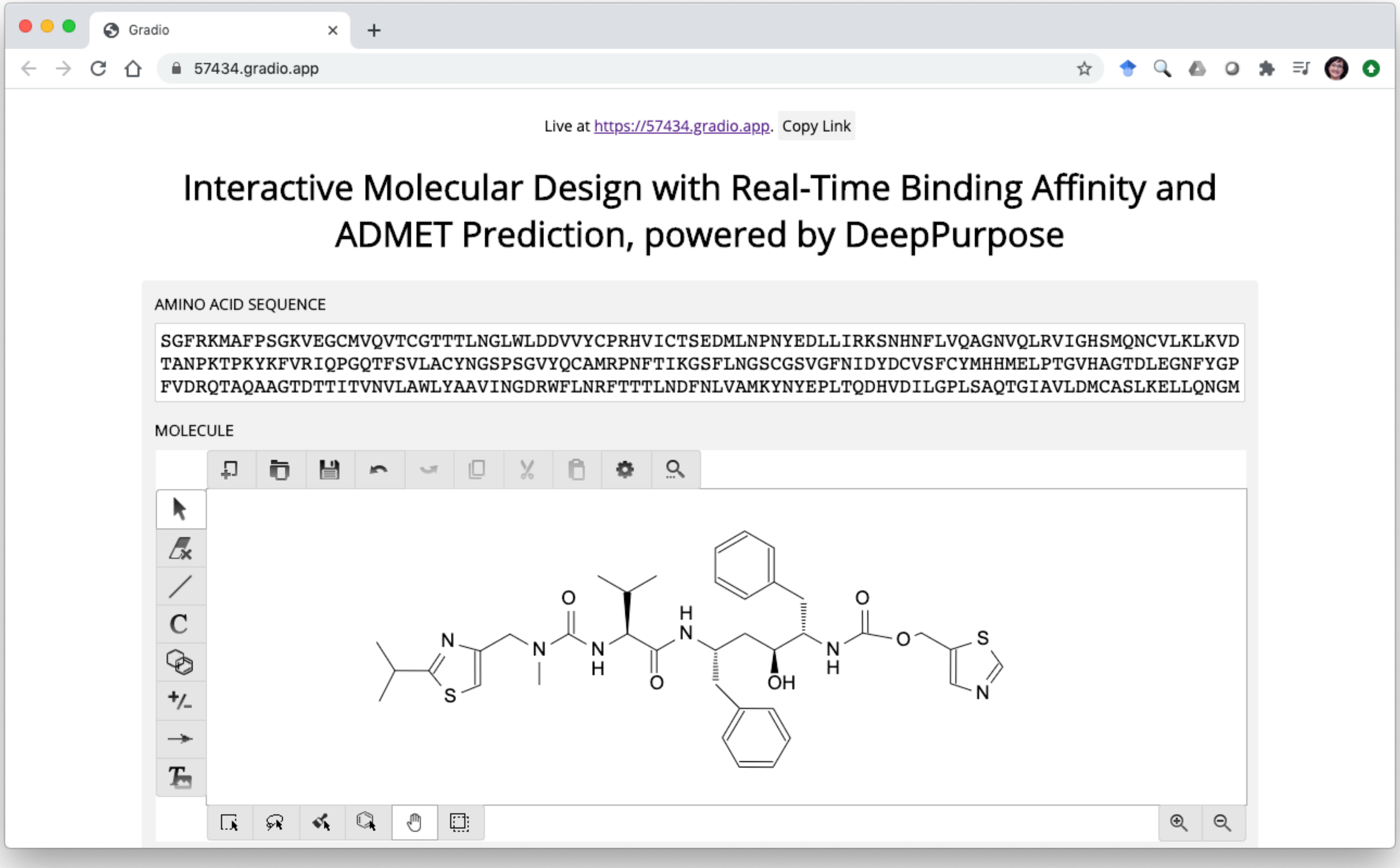
Checkout 10+ demos & tutorials to start:

Name	Description
Dataset Tutorial	Tutorial on how to use the dataset loader and read customized data
Drug Repurposing for 3CLPro	Example of one-liner repurposing for 3CLPro
Drug Repurposing with Customized Data	Example of one-liner repurposing with AID1706 Bioassay Data, training from scratch
Virtual Screening for BindingDB IC50	Example of one-liner virtual screening
Reproduce DeepDTA	Reproduce DeepDTA with DAVIS dataset and show how to use the 10 lines framework
Virtual Screening for DAVIS and Correlation Plot	Example of one-liner virtual screening and evaluate on unseen dataset by plotting correlation
Binary Classification for DAVIS using CNNs	Binary Classification for DAVIS dataset using CNN encodings by using the 10 lines framework.
Pretraining Model Tutorial	Tutorial on how to load pretraining models

and more in the DEMO folder!

DeepPurpose: a Deep Learning Library for Drug-Target Interaction Prediction, *Bioinformatics* 2020

How can domain scientists interact with AI systems?



Live at https://57434.gradio.app. Copy Link

Interactive Molecular Design with Real-Time Binding Affinity and ADMET Prediction, powered by DeepPurpose

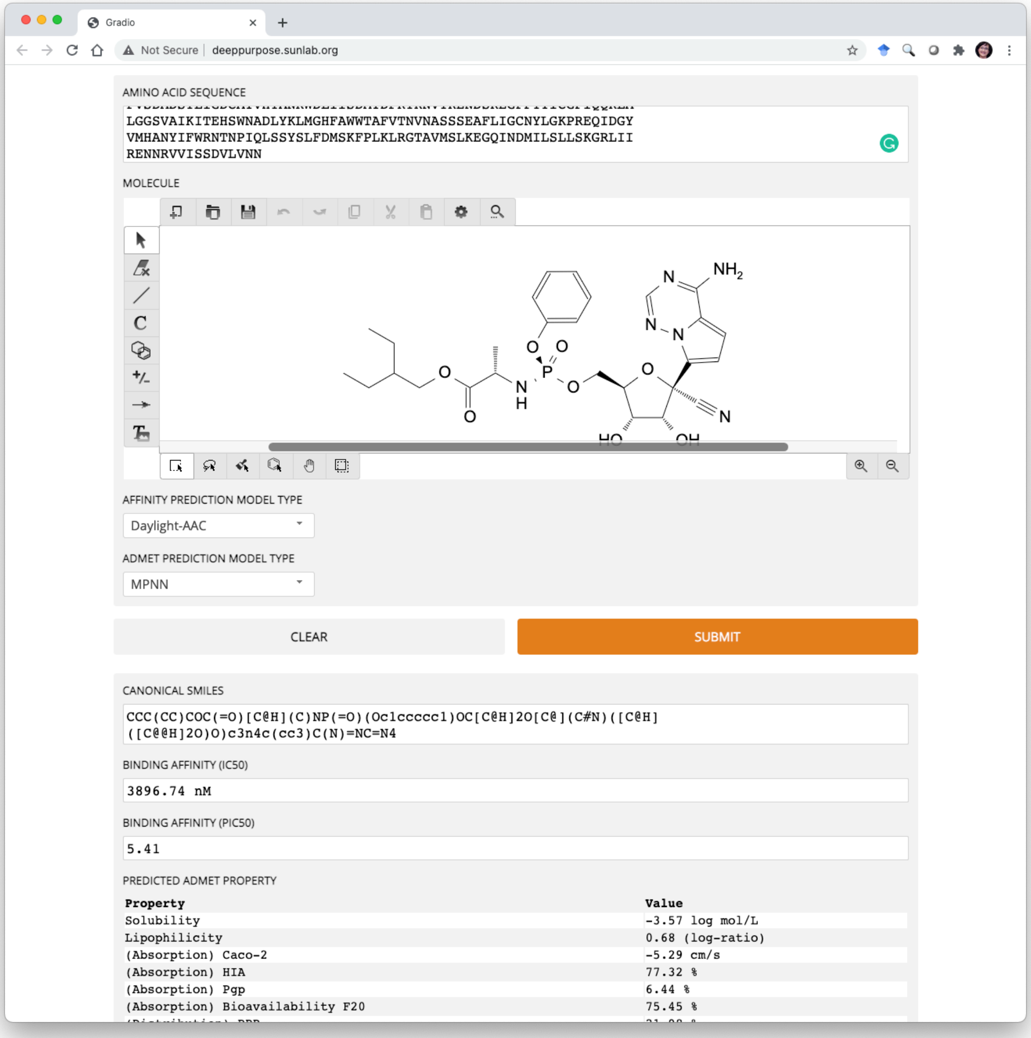
AMINO ACID SEQUENCE

```
SGFRKMAFPSGKVEGCMVQVTCGTTTLNGLWLDDVVYCPRHVICTSEDMLNPNYEDLLIRKSNHNFLVQAGNVQLRVIGHSMQNCVLKLKVD
TANPKTPKYKFVRIQPGQTFSVLACYNGSPSGVYQCAMRPNFTIKGSFLNGSCGSVGFNIDYDCVSFCYMHHMELPTGVHAGTDLEGNFYGP
FVDRQTAQAAGTDTTITVNVLAWLYAAVINGDRWFLNRFTTTLNDFNLVAMKYNYEPLTQDHVDILGPLSAQTGIAVLDMCASLKELLQNGM
```

MOLECULE

DeepPurpose: a Deep Learning Library for Drug-Target Interaction Prediction, *Bioinformatics* 2020
MolDesigner: Interactive Design of Efficacious Drugs with Deep Learning, *NeurIPS* 2020

MolDesigner: Interactive Design of Drugs with Deep Learning



AMINO ACID SEQUENCE

```
LGGSVAIKITEHSWNADLYKLMGHFAWWTAFVTNVNASSSEAFLIGCNYLGKPREQIDGY
VMHANYIFWRNTNPIQLSSYSLFDMSKFPLKLRGTAVMSLKEGQINDMILSLLSKGRLII
RENNRVVISSDVLVNN
```

MOLECULE

AFFINITY PREDICTION MODEL TYPE

Daylight-AAC

ADMET PREDICTION MODEL TYPE

MPNN

CANONICAL SMILES

```
CCC(CC)COC(=O)[C@H](C)NP(=O)(Oc1ccccc1)OC[C@H]2O[C@](C#N)([C@H]
([C@@H]2O)O)c3n4c(cc3)C(N)=NC=N4
```

BINDING AFFINITY (IC50)

3896.74 nM

BINDING AFFINITY (PIC50)

5.41

PREDICTED ADMET PROPERTY

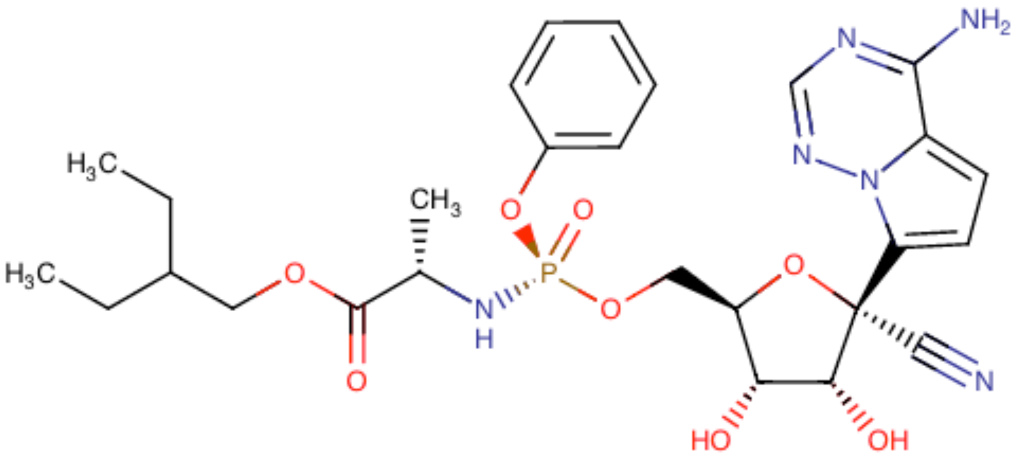
Property	**Value**
Solubility	-3.57 log mol/L
Lipophilicity	0.68 (log-ratio)
(Absorption) Caco-2	-5.29 cm/s
(Absorption) HIA	77.32 %
(Absorption) Pgp	6.44 %
(Absorption) Bioavailability F20	75.45 %
(Distribution) BBB	[illegible]

http://deeppurpose.sunlab.org

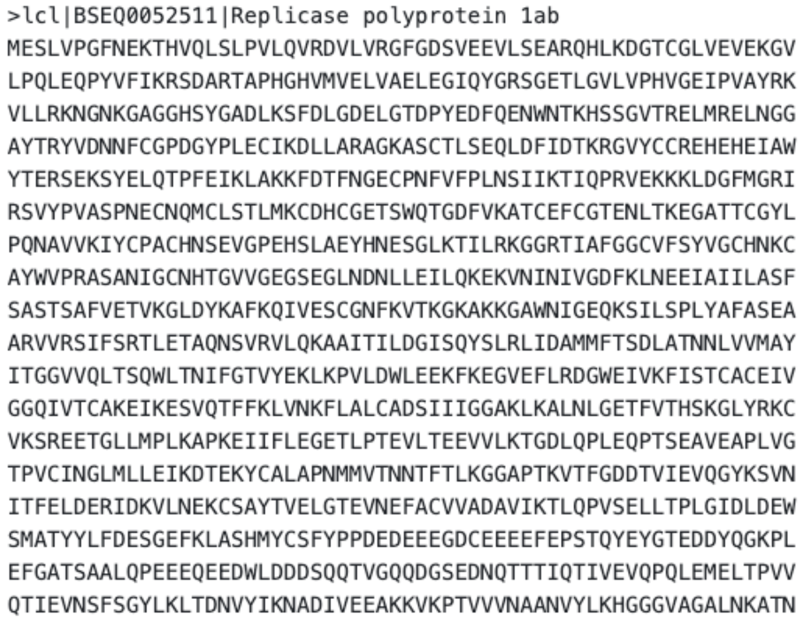
DEMO: DRUG-TARGET INTERACTION PREDICTION

Drug: Remdesivir Remdesivir is indicated for the treatment of adult and pediatric patients aged 12 years and over weighing at least 40 kg for coronavirus disease 2019 (COVID-19) infection requiring hospitalization.

Target protein: Replicase polyprotein 1ab. Multifunctional protein involved in the transcription and replication of viral RNAs



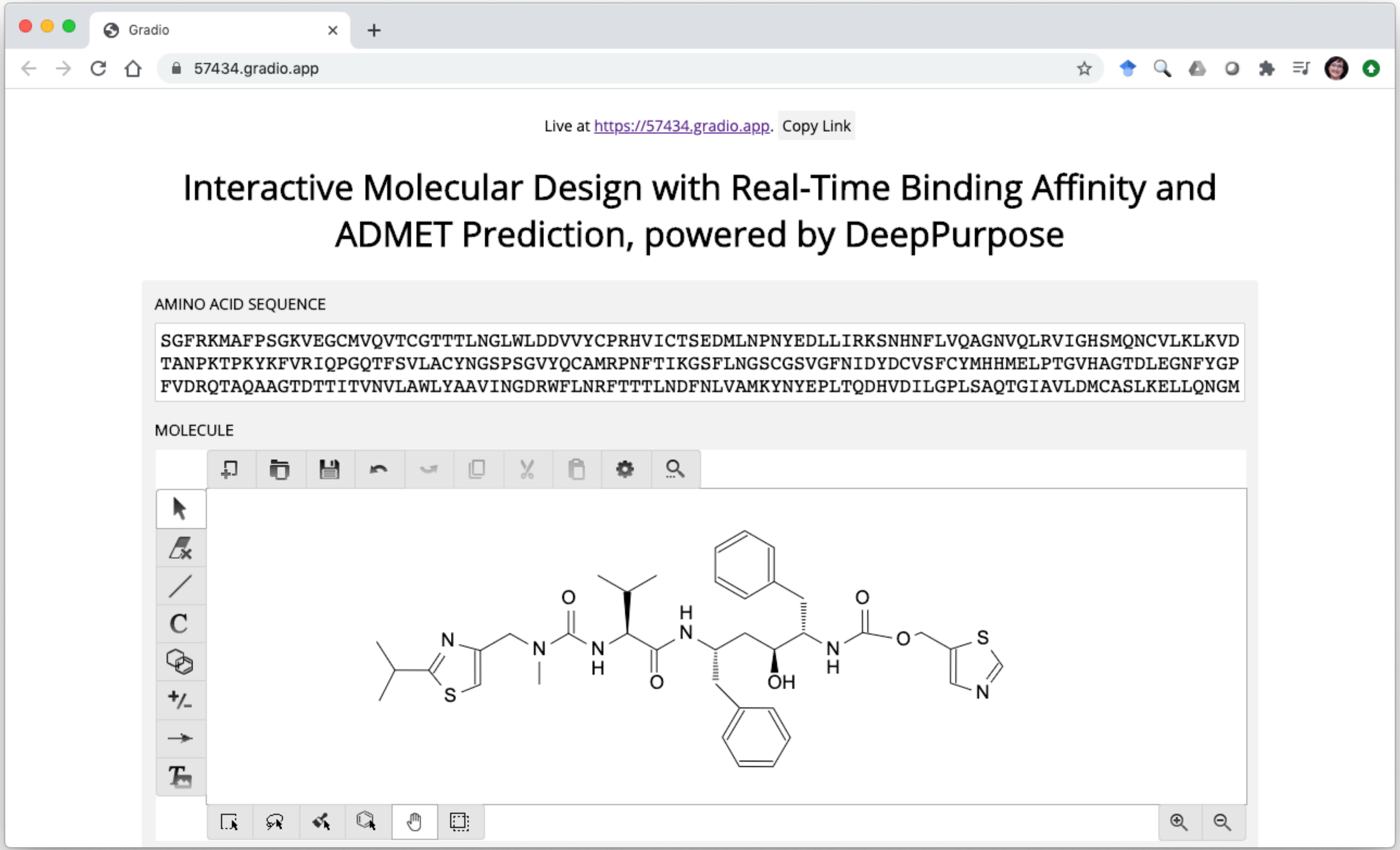
Molecular structure of Remdesivir



```
>lcl|BSEQ0052511|Replicase polyprotein 1ab
MESLVPGFNEKTHVQLSLPVLQVRDVLVRGFGDSVEEVLSEARQHLKDGTCGLVEVEKGV
LPQLEQPYVFIKRSDARTAPHGHVMVELVAELEGIQYGRSGETLGVLVPHVGEIPVAYRK
VLLRKNGNKGAGGHSYGADLKSFDLGDELGTDPYEDFQENWNTKHSSGVTRELMRELNGG
AYTRYVDNNFCGPDGYPLECIKDLLARAGKASCTLSEQLDFIDTKRGVYCCREHEHEIAW
YTERSEKSYELQTPFEIKLAKKFDTFNGECPNFVFPLNSIIKTIQPRVEKKKLDGFMGRI
RSVYPVASPNECNQMCLSTLMKCDHCGETSWQTGDFVKATCEFCGTENLTKEGATTCGYL
PQNAVVKIYCPACHNSEVGPEHSLAEYHNESGLKTILRKGGRTIAFGGCVFSYVGCHNKC
AYWVPRASANIGCNHTGVVGEGSEGLNDNLLEILQKEKVNINIVGDFKLNEEIAIILASF
SASTSAFVETVKGLDYKAFKQIVESCGNFKVTKGKAKKGAWNIGEQKSILSPLYAFASEA
ARVVRSIFSRTLETAQNSVRVLQKAAITILDGISQYSLRLIDAMMFTSDLATNNLVVMAY
ITGGVVQLTSQWLTNIFGTVYEKLKPVLDWLEEKFKEGVEFLRDGWEIVKFISTCACEIV
GGQIVTCAKEIKESVQTFFKLVNKFLALCADSIIIGGAKLKALNLGETFVTHSKGLYRKC
VKSREETGLLMPLKAPKEIIFLEGETLPTEVLTEEVVLKTGDLQPLEQPTSEAVEAPLVG
TPVCINGLMLLEIKDTEKYCALAPNMMVTNNTFTLKGGAPTKVTFGDDTVIEVQGYKSVN
ITFELDERIDKVLNEKCSAYTVELGTEVNEFACVVADAVIKTLQPVSELLTPLGIDLDEW
SMATYYLFDESGEFKLASHMYCSFYPPDEDEEEGDCEEEEFEPSTQYEYGTEDDYQGKPL
EFGATSAALQPEEEQEEDWLDDDSQQTVGQQDGSEDNQTTTIQTIVEVQPQLEMELTPVV
QTIEVNSFSGYLKLTDNVYIKNADIVEEAKKVKPTVVVNAANVYLKHGGGVAGALNKATN
```

Amino acid sequence of Replicase polyprotein 1ab

How can domain scientists interact with AI systems?



Gradio

57434.gradio.app

Live at https://57434.gradio.app. Copy Link

Interactive Molecular Design with Real-Time Binding Affinity and ADMET Prediction, powered by DeepPurpose

AMINO ACID SEQUENCE

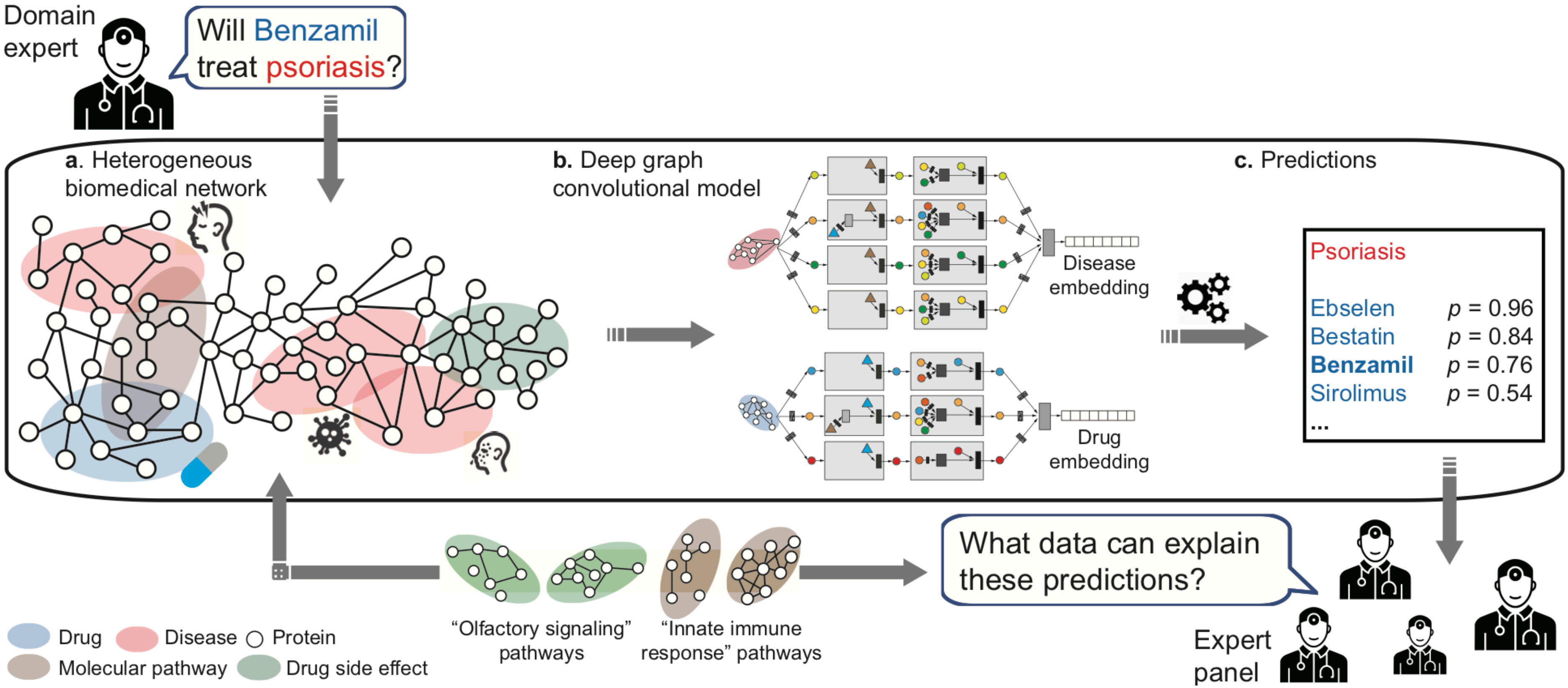
```
SGFRKMAFPSGKVEGCMVQVTCGTTTLNGLWLDDVVYCPRHVICTSEDMLNPNYEDLLIRKSNHNFLVQAGNVQLRVIGHSMQNCVLKLKVD
TANPKTPKYKFVRIQPGQTFSVLACYNGSPSGVYQCAMRPNFTIKGSFLNGSCGSVGFNIDYDCVSFCYMHHMELPTGVHAGTDLEGNFYGP
FVDRQTAQAAGTDTTITVNVLAWLYAAVINGDRWFLNRFTTTLNDFNLVAMKYNYEPLTQDHVDILGPLSAQTGIAVLDMCASLKELLQNGM
```

MOLECULE

DeepPurpose: a Deep Learning Library for Drug-Target Interaction Prediction, *Bioinformatics* 2020

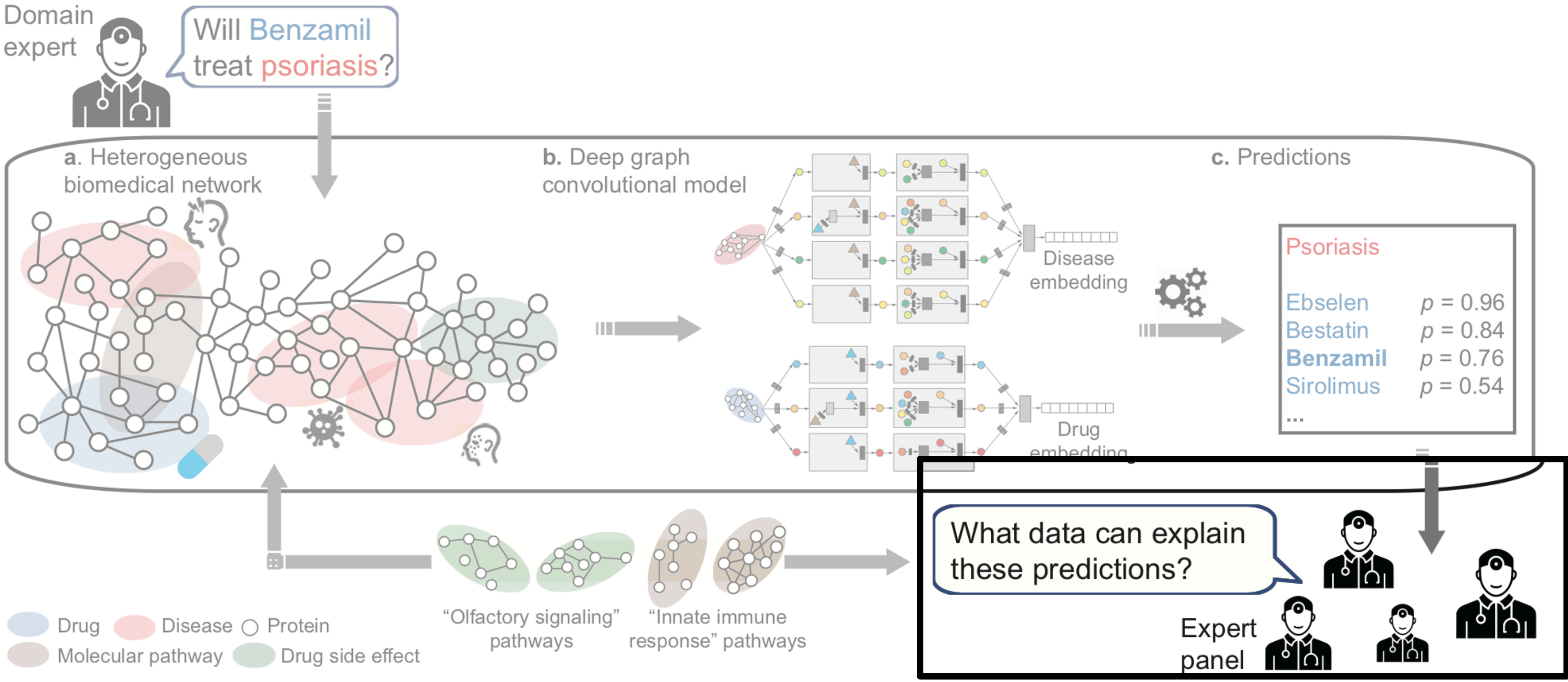
MolDesigner: Interactive Design of Efficacious Drugs with Deep Learning, *NeurIPS* 2020

Automating Science


Domain expert
Will Benzamil treat psoriasis?
a. Heterogeneous biomedical network
b. Deep graph convolutional model
Disease embedding
Drug embedding
c. Predictions
Drug
Disease
○ Protein
Molecular pathway
Drug side effect
"Olfactory signaling" pathways
"Innate immune response" pathways
What data can explain these predictions?
Expert panel

Psoriasis	
Ebselen	p = 0.96
Bestatin	p = 0.84
Benzamil	p = 0.76
Sirolimus	p = 0.54
...	

Automating Science



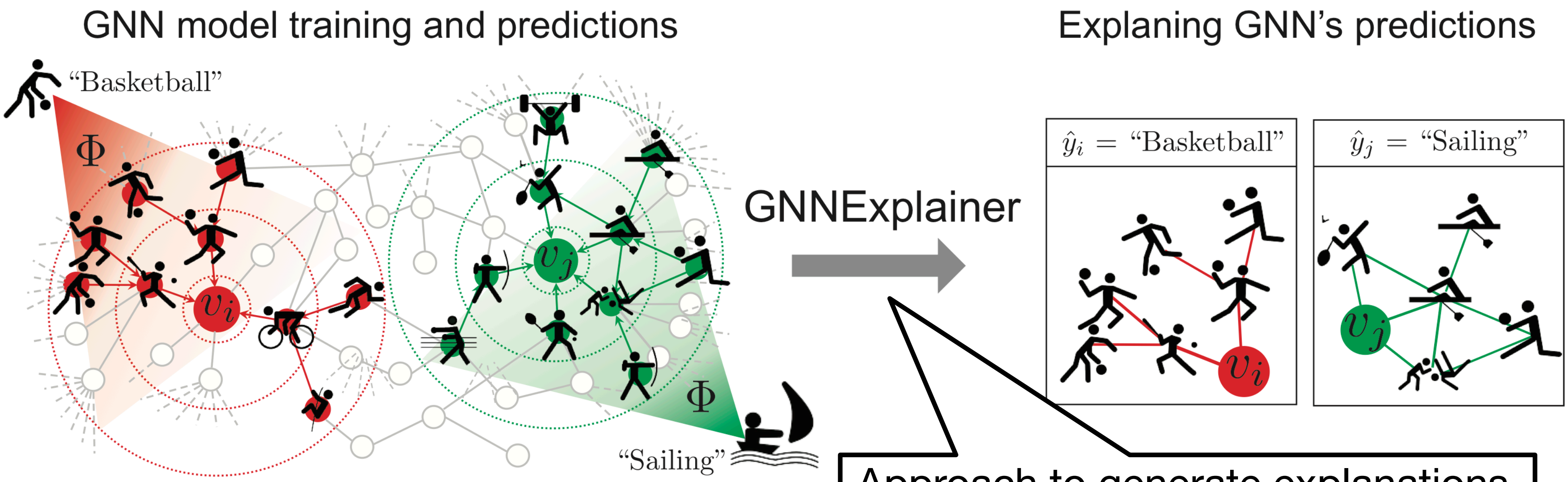
Will Benzamil treat psoriasis?

What data can explain these predictions?

How to explain predictions?

Key idea:

- Summarize where in the data the model “looks” for evidence for its prediction
- Find a small subgraph most influential for the prediction

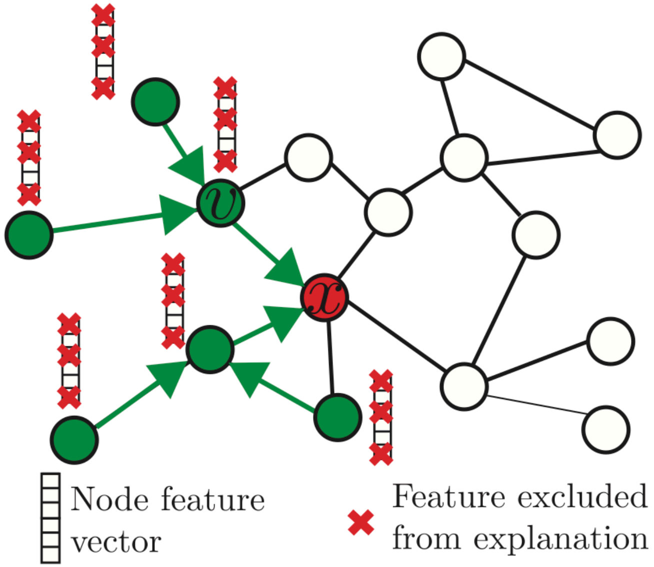


Approach to generate explanations for graph neural networks based on counterfactual reasoning

GNNExplainer: Generating Explanations for Graph Neural Networks, *NeurIPS* 2019

GNNExplainer: Key Idea

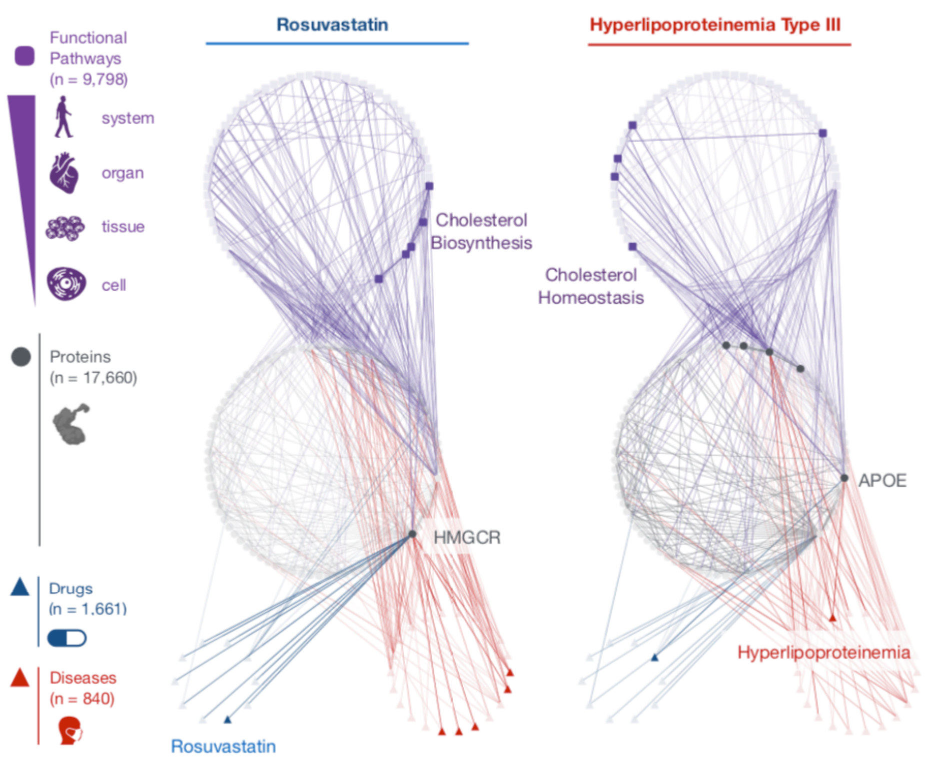
- Input: Given prediction $f(x)$ for node/link x
- Output: Explanation, a small subgraph M_x together with a small subset of node features:
  - M_x is most influential for prediction $f(x)$
- Approach: Learn M_x via counterfactual reasoning
  - Intuition: If removing v from the graph strongly decreases the probability of prediction ⇒ v is a good counterfactual explanation for the prediction

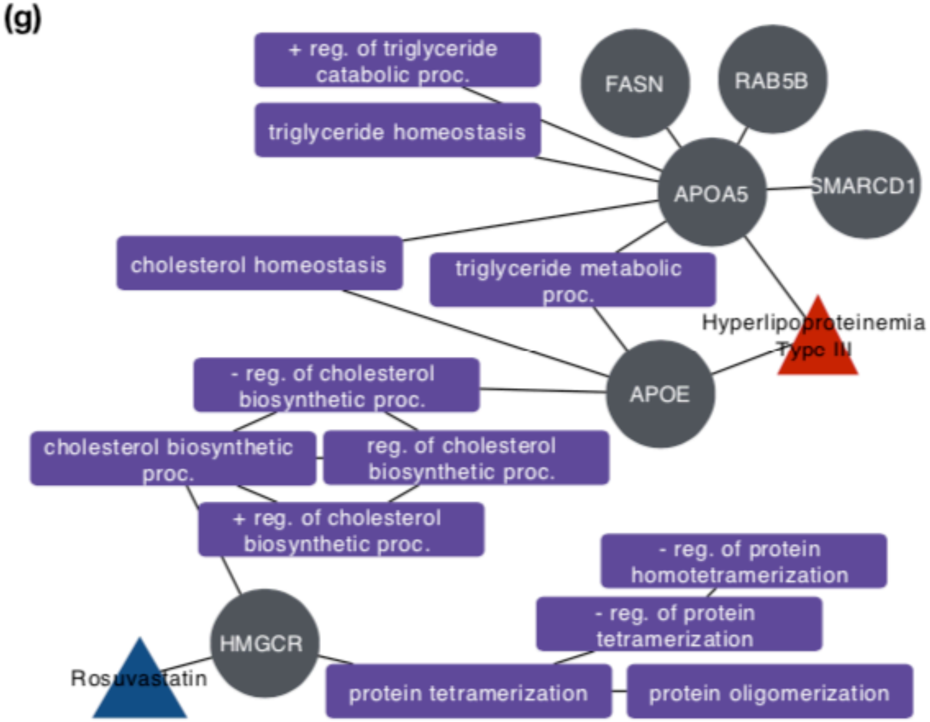


GNN Explainer: Generating Explanations for Graph Neural Networks, *NeurIPS* 2019

Examples of Explanations

”Will rosuvastatin treat hyperlipidemia? What is the disease treatment mechanism?”




(g)
+ reg. of triglyceride catabolic proc.
triglyceride homeostasis
FASN
RAB5B
APOA5
SMARCD1
cholesterol homeostasis
triglyceride metabolic proc.
Hyperlipoproteinemia Type III
- reg. of cholesterol biosynthetic proc.
APOE
cholesterol biosynthetic proc.
reg. of cholesterol biosynthetic proc.
+ reg. of cholesterol biosynthetic proc.
- reg. of protein homotetramerization
- reg. of protein tetramerization
HMGCR
Rosuvastatin
protein tetramerization
protein oligomerization

New Algorithms: GNNExplainer: Generating Explanations for Graph Neural Networks, *NeurIPS* 2019

New Insights: Discovery of Disease Treatment Mechanisms through the Multiscale Interactome, *Nature Communications 2021* (in press)

Open challenges and future directions

Learn about Therapeutics ML!

National Symposium on Drug Repurposing for Future Pandemics

17-18th November 2020 | Virtual Symposium

Registration Closed

https://www.drugsymposium.org

Videos from the presentations are now publicly available to everyone through the Symposium Video Channel

Open Challenges

- Disconnected, uncoupled biomedical knowledge:
  - Challenge: Need to combine data in their broadest sense to close the gap between research and patient data
- Diverse mechanisms of drug action:
  - Challenge: Need to consider diverse mechanisms through which a drug can treat a disease
- Novel drugs in development, emerging diseases:
  - Challenge: Need to learn and reason about never-before-seen phenomena
- Datasets for a variety of therapeutics tasks:
  - Challenge: Need datasets and benchmarks to accelerate ML model development, validation and transition into production and clinical implementation

Outline

- ✓ Overview and introduction
- ✓ Part 1: Virtual drug screening and drug repurposing
- ✓ Part 2: Adverse drug effects, drug-drug interactions
- ✓ Part 3: Clinical trial site identification, patient recruitment
- ✓ Part 4: Molecule optimization, molecular graph generation, multimodal graph-to-graph translation
- ✓ Part 5: Molecular property prediction and transformers
- ✓ Demos, resources, wrap-up & future directions

https://zitniklab.hms.harvard.edu/drugml