THERAPEUTICS DATA COMMONS

Artificial Intelligence Foundation for Therapeutic Science

Kexin Huang*, Tianfan Fu*, Wenhao Gao*, Yue Zhao, Yusuf Roohani, Jure Leskovec, Connor W. Coley, Cao Xiao, Jimeng Sun, Marinka Zitnik

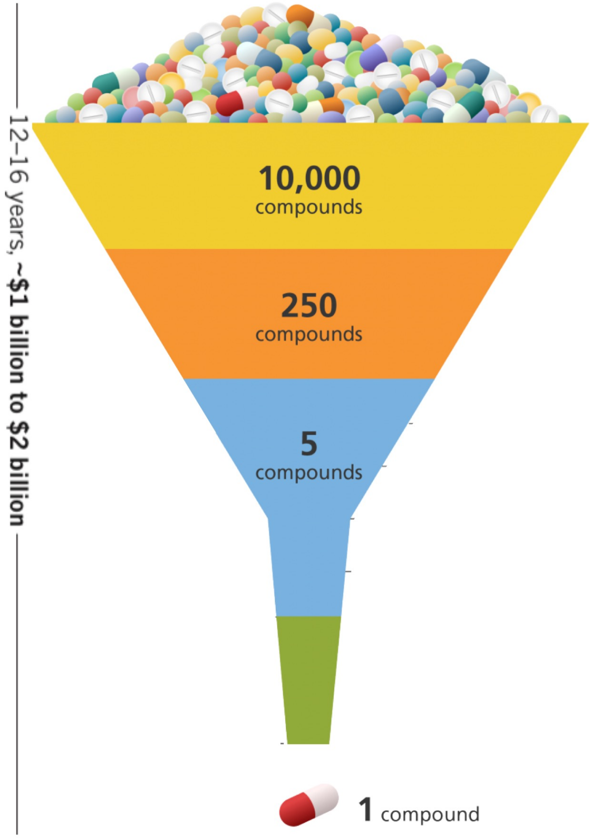
ACS Fall 2022, Division of Computers in Chemistry

8/22/2022

Huang, K., Fu, T., Gao, W., ... & Zitnik, M. (2022). Artificial intelligence foundation for therapeutic science. *Nature chemical biology* (accepted)

Challenges in Drug Discovery & Development

High cost, long time

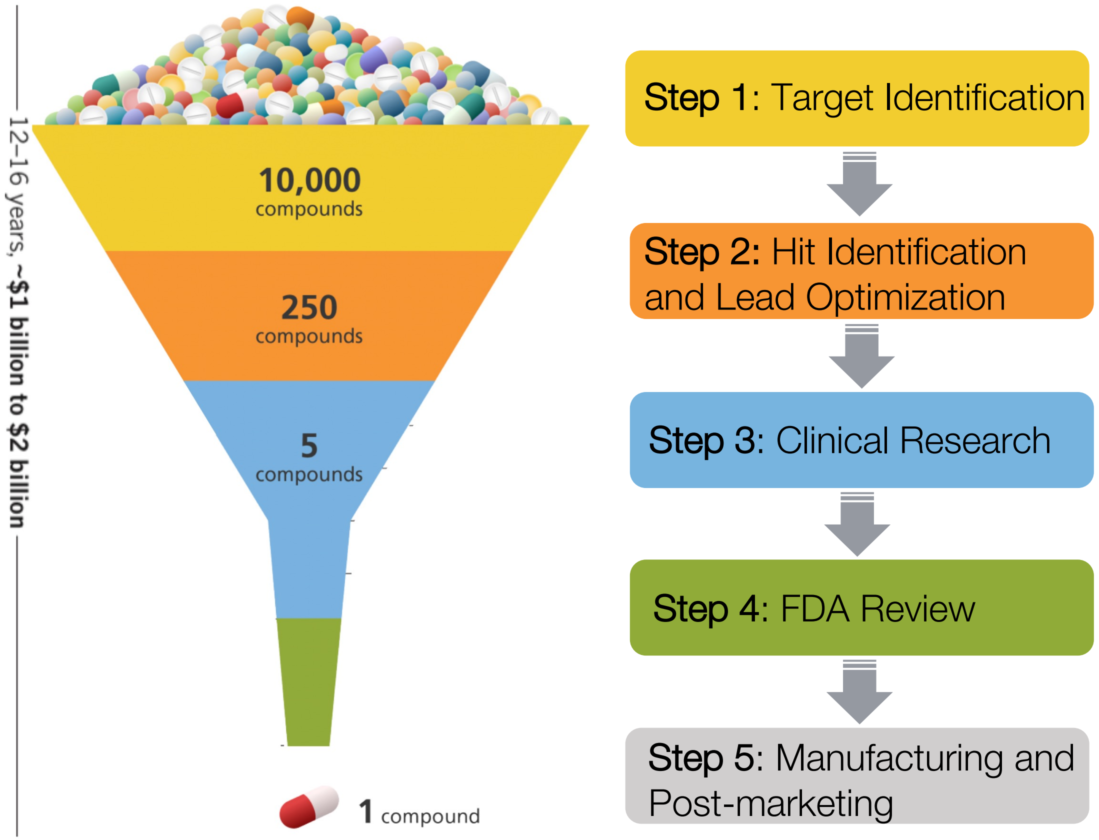


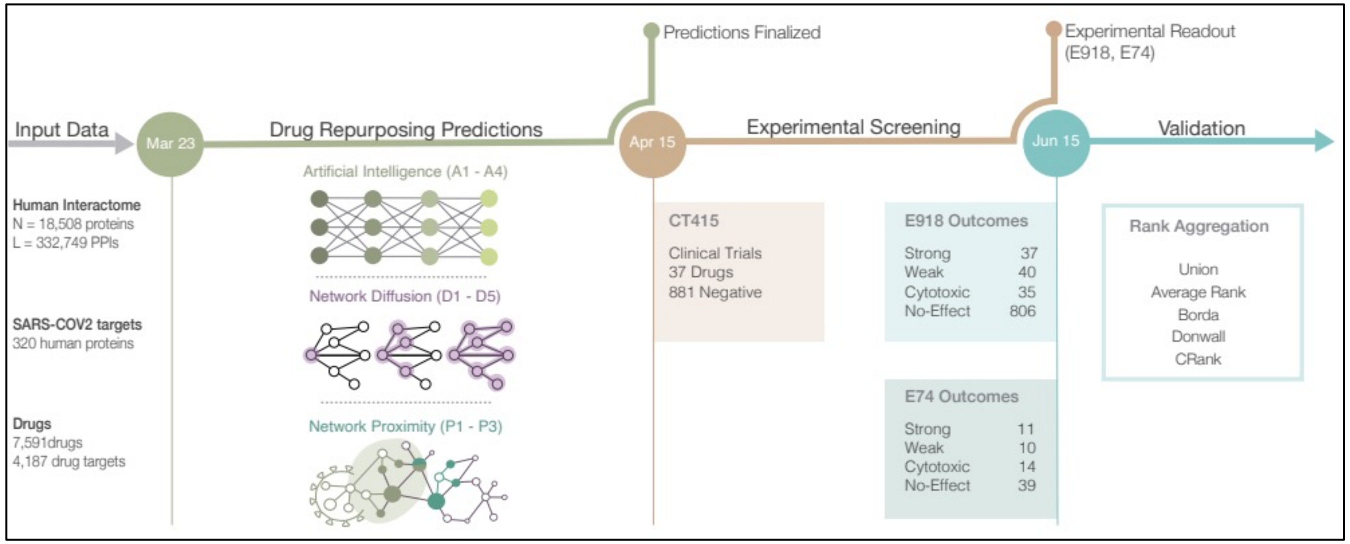
- **Step 1:** Target Identification
- **Step 2:** Hit Identification and Lead Optimization
- **Step 3:** Clinical Research
- **Step 4:** FDA Review
- **Step 5:** Manufacturing and Post-marketing

Challenges in Drug Discovery & Development

High cost, long time

Various and emerging diseases



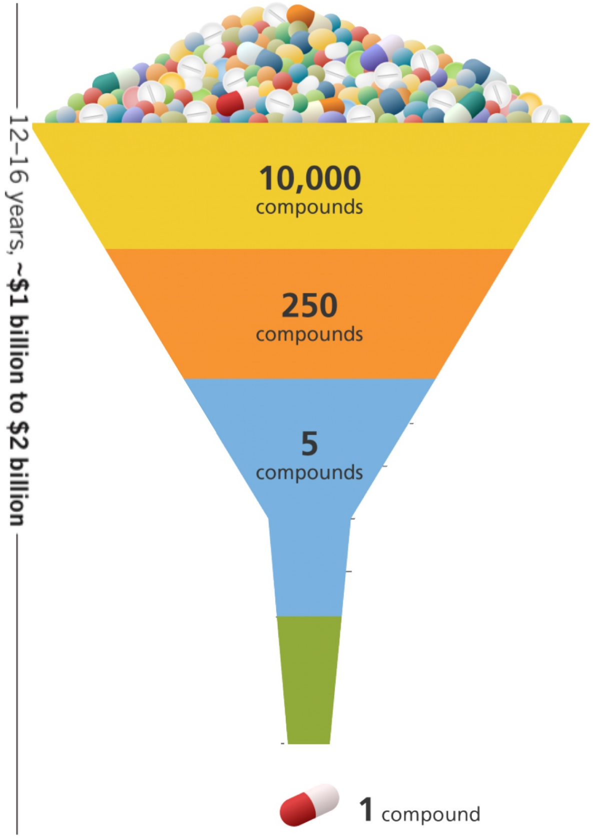


Challenges in Drug Discovery & Development

High cost, long time

Various and emerging diseases

Abundant molecular modality



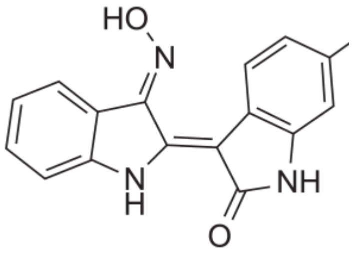
Step 1: Target Identification

Step 2: Hit Identification and Lead Optimization

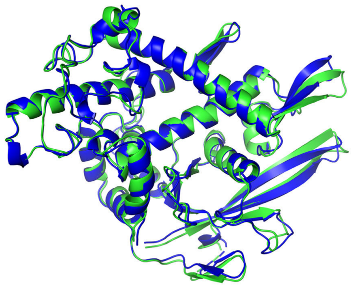
Step 3: Clinical Research

Step 4: FDA Review

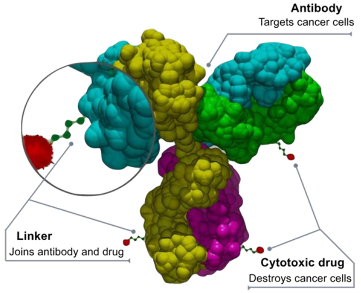
Step 5: Manufacturing and Post-marketing



Small molecules

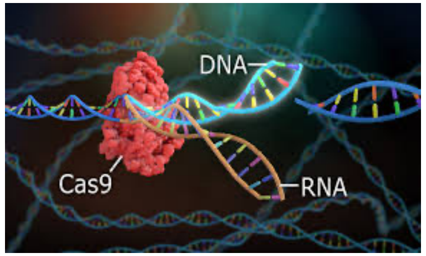


Proteins

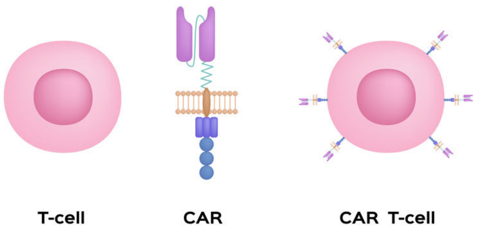


Anti- Nano-bodies

Vaccines



Gene-editing



CAR-T

Data-driven Methods Demonstrated a Clear Impact

MAIN ABOUT PHARMA.AI PIPELINE NEWS & MEDIA CAREERS ENG 简

From Start to Phase 1 in 30 Months: AI-discovered and AI-designed Anti-fibrotic Drug Enters Phase I Clinical Trial

2022.02.24

Exscientia ABOUT PATIENT-FIRST AI PIPELINE INVESTORS & MEDIA JOIN US CONTACT

OVERVIEW PRESS RELEASES EVENTS & PRESENTATIONS STOCK INFORMATION CORPORATE GOVERNANCE FINANCIALS & FILINGS INVESTOR RESOURCES

AI-Designed Immuno-Oncology Drug to Enter Clinical Trials

ertise now responsible for world's first and second AI-designed drugs to enter Phase I testing

Recursion Investors Governance News SEC Filings Stock Resources

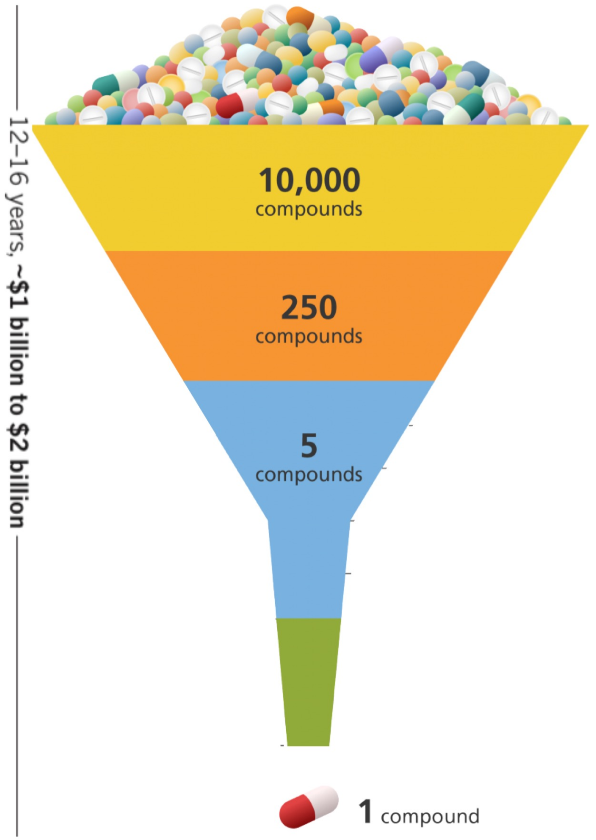
News

Recursion is Granted FDA Fast Track Designation for REC-2282 for the Potential Treatment of NF2-Mutated Meningiomas

SALT LAKE CITY, Oct. 7, 2021 /PRNewswire/ -- Recursion (NASDAQ: RXRX), a clinical-stage biotechnology company decoding biology by

investors.exscientia.ai/press-releases/ https://insilico.com/phase1 ir.recursion.com

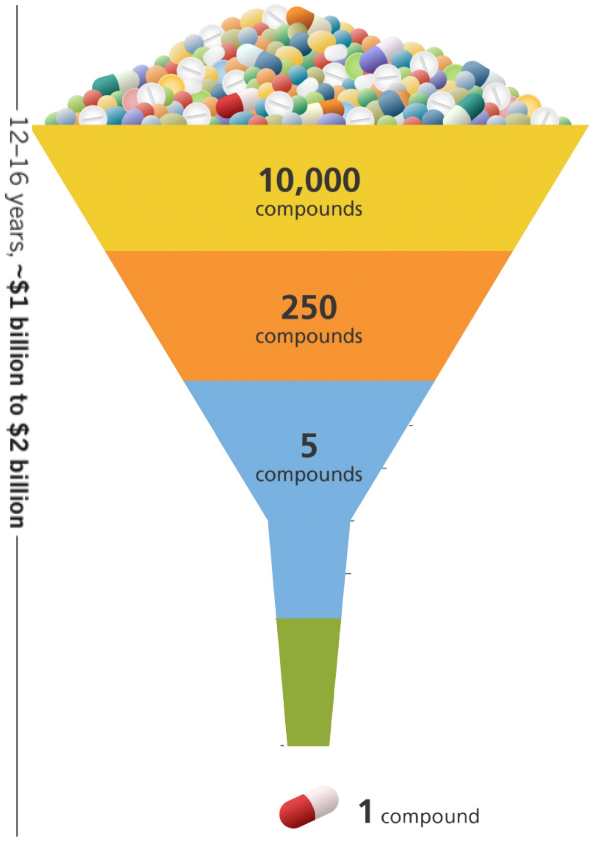
Challenges of AI in Drug Discovery & Development



- **Step 1**: Target Identification
- **Step 2:** Hit Identification and Lead Optimization
- **Step 3**: Clinical Research
- **Step 4**: FDA Review
- **Step 5**: Manufacturing and Post-marketing

Challenges of AI in Drug Discovery & Development

- Formulate various tasks as machine-learning-solvable tasks.



Step 1: Target Identification — Gene-disease/drug interaction prediction; PPI prediction …

Step 2: Hit Identification and Lead Optimization — Molecular optimization; Protein structure prediction; Molecule property prediction …

Step 3: Clinical Research — Clinical trial recruitment; Drug repurposing …

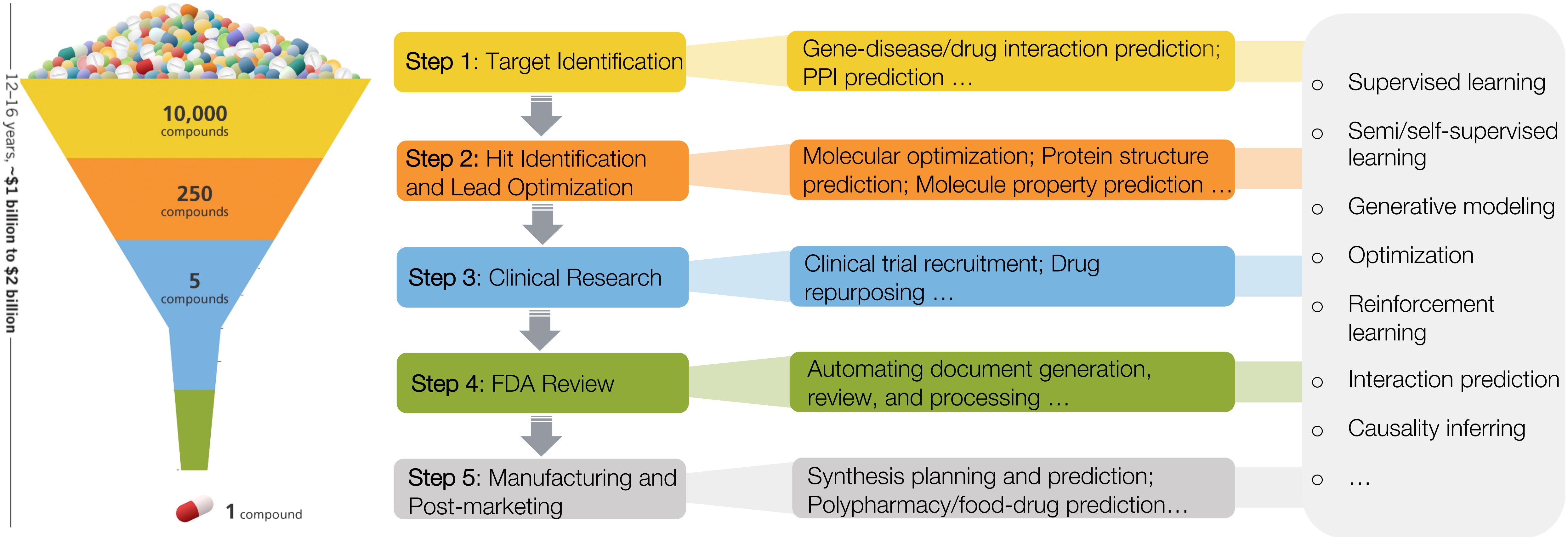
Step 4: FDA Review — Automating document generation, review, and processing …

Step 5: Manufacturing and Post-marketing — Synthesis planning and prediction; Polypharmacy/food-drug prediction…

- Supervised learning
- Semi/self-supervised learning
- Generative modeling
- Optimization
- Reinforcement learning
- Interaction prediction
- Causality inferring
- …

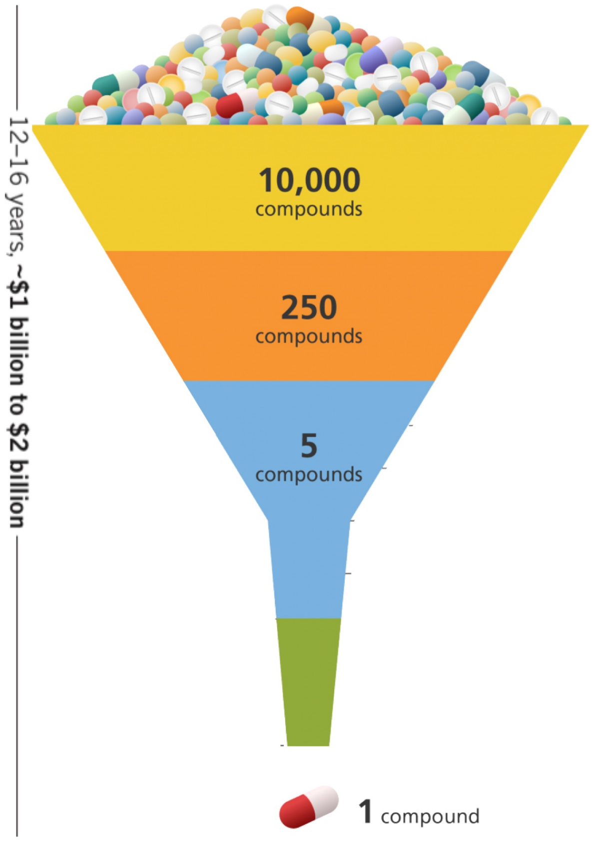
Challenges of AI in Drug Discovery & Development

- Formulate various tasks as machine-learning-solvable tasks.
- Identify, retrieve, and process datasets of many different types scattered around.



Challenges of AI in Drug Discovery & Development

- Formulate various tasks as machine-learning-solvable tasks.
- Identify, retrieve, and process datasets of many different types scattered around.
- Assess algorithmic advances to align with real-world and clinical deployment.

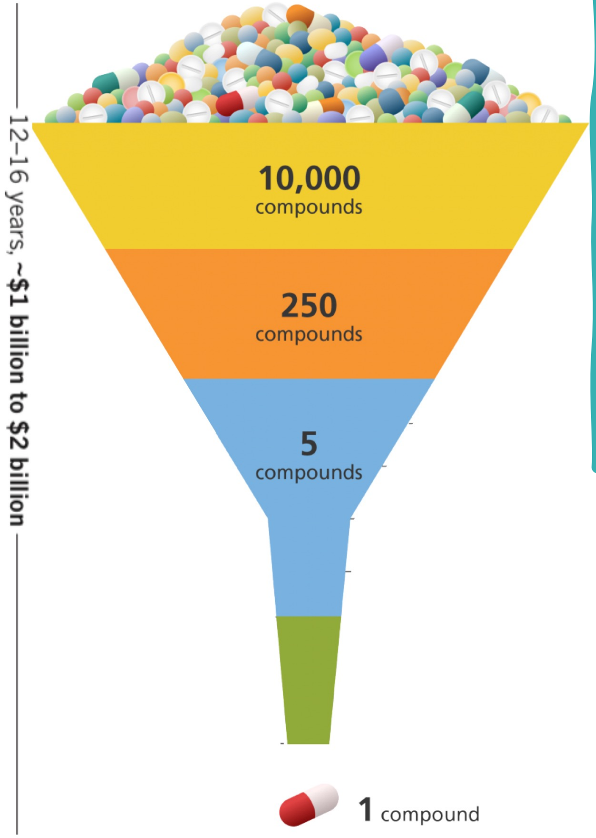


Step 1: Target Identification	Gene-disease/drug interaction prediction; PPI prediction …
Step 2: Hit Identification and Lead Optimization	Molecular optimization; Protein structure prediction; Molecule property prediction …
Step 3: Clinical Research	Clinical trial recruitment; Drug repurposing …
Step 4: FDA Review	Automating document generation, review, and processing …
Step 5: Manufacturing and Post-marketing	Synthesis planning and prediction; Polypharmacy/food-drug prediction…

- Supervised learning
- Semi/self-supervised learning
- Generative modeling
- Optimization
- Reinforcement learning
- Interaction prediction
- Causality inferring
- …

Challenges of AI in Drug Discovery & Development

- Formulate various tasks as machine-learning-solvable tasks.
- Identify, retrieve, and process datasets of many different types scattered around.
- Assess algorithmic advances to align with real-world and clinical deployment.



How to Advance Data-driven Therapeutics Design?

n;

…

Step 4: FDA Review — Automating document generation, review, and processing …

Step 5: Manufacturing and Post-marketing — Synthesis planning and prediction; Polypharmacy/food-drug prediction…

- Supervised learning
- Semi/self-supervised learning
- Generative modeling
- Optimization
- Reinforcement learning
- Interaction prediction
- Causality inferring
- …

Therapeutics Data Commons

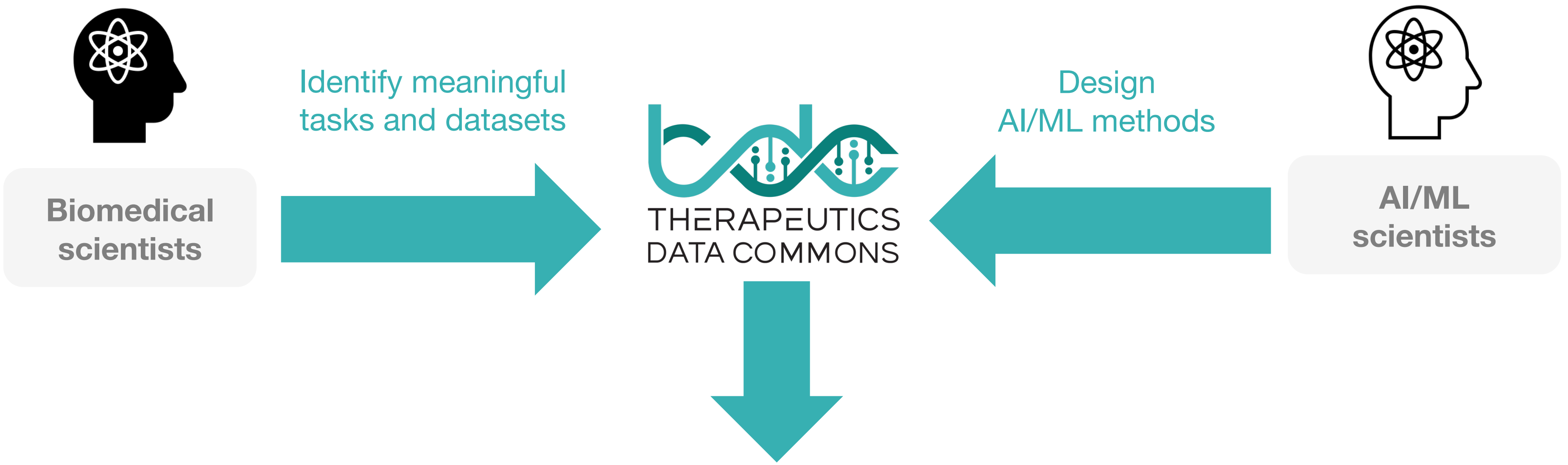
THERAPEUTICS DATA COMMONS

Website: https://tdcommons.ai (or QR code)

Paper: https://arxiv.org/abs/2102.09548

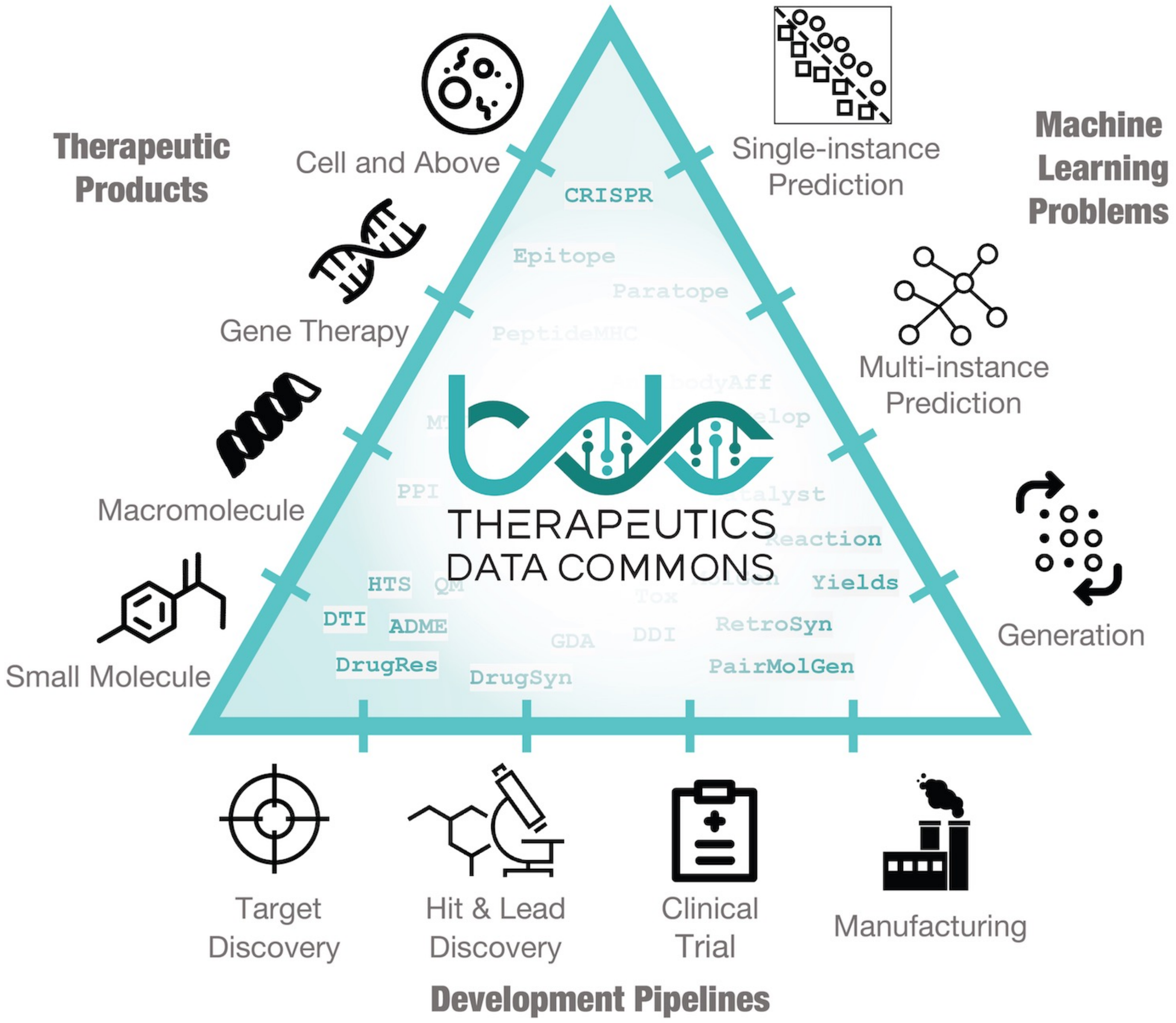
Github: https://github.com/mims-harvard/TDC

THERAPEUTICS DATA COMMONS

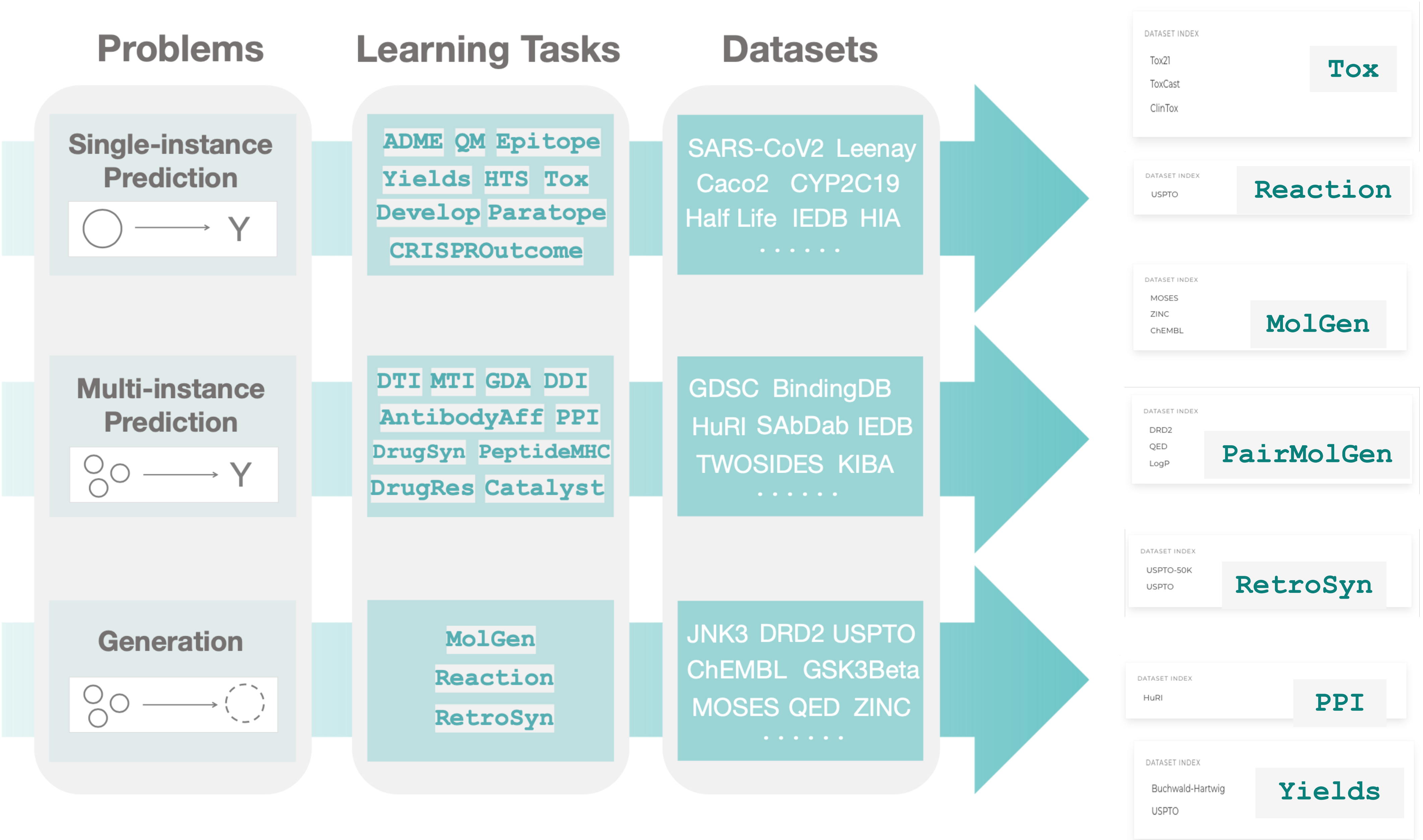


Facilitate algorithmic and scientific advance in therapeutics

Wide Range of Therapeutic Modalities and Pipeline



Three-Tier Design

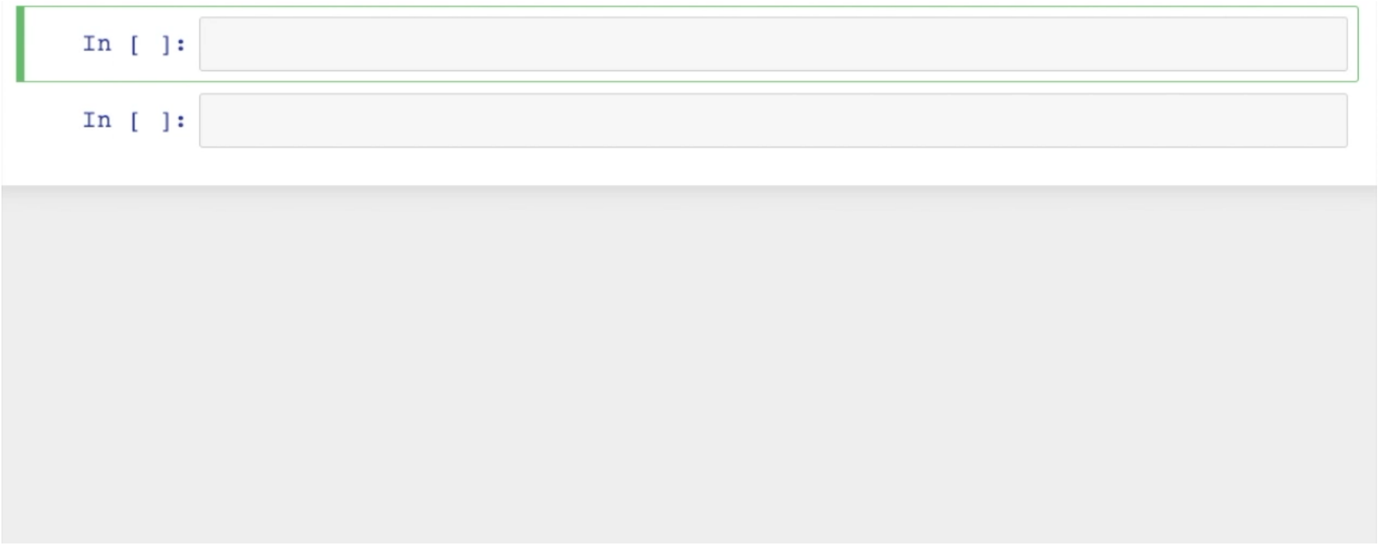


Unified, Light-weighted and User-friendly

```
pip install PyTDC
```

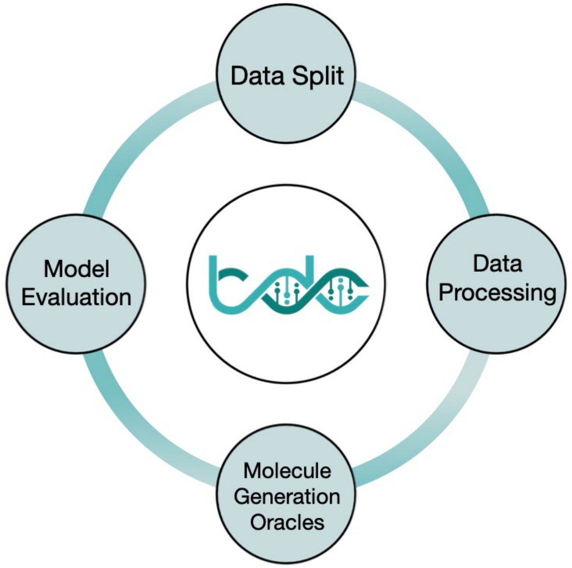
3 Lines of Code

The core TDC library uses minimum packages thus is installed hassle-free. Data loaders are simplified so that you can get access to ML-ready datasets within only 3 lines of code.



TDC has more than Datasets

- Ecosystem of tools, leaderboards, and community resources, including data functions, strategies for systematic model evaluation, meaningful data splits, data processors, and molecule generation oracles



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

Meaningful data splits

FUNCTION INDEX

Data Split Overview

- Random Split
- Scaffold Split
- Cold-Start Split

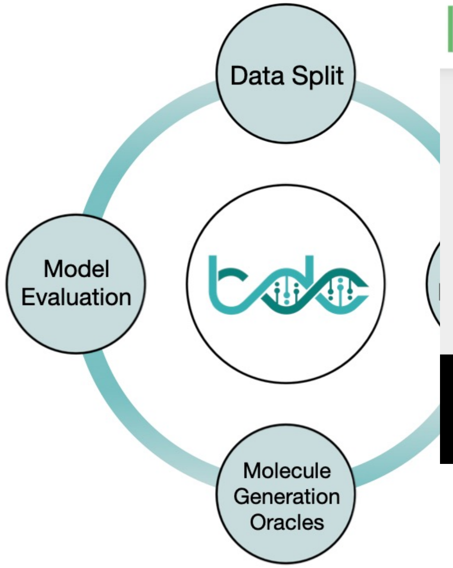
Data processing helpers

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

TDC has more than Datasets

- Ecosystem of tools, leaderboards, and community resources, including data functions, strategies for systematic model evaluation, meaningful data splits, data processors, and molecule generation oracles



[redacted]

In [ ]:

[redacted]

[redacted]ingful data splits

erview
split
plit
Split

Multi-class Classification Metric

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

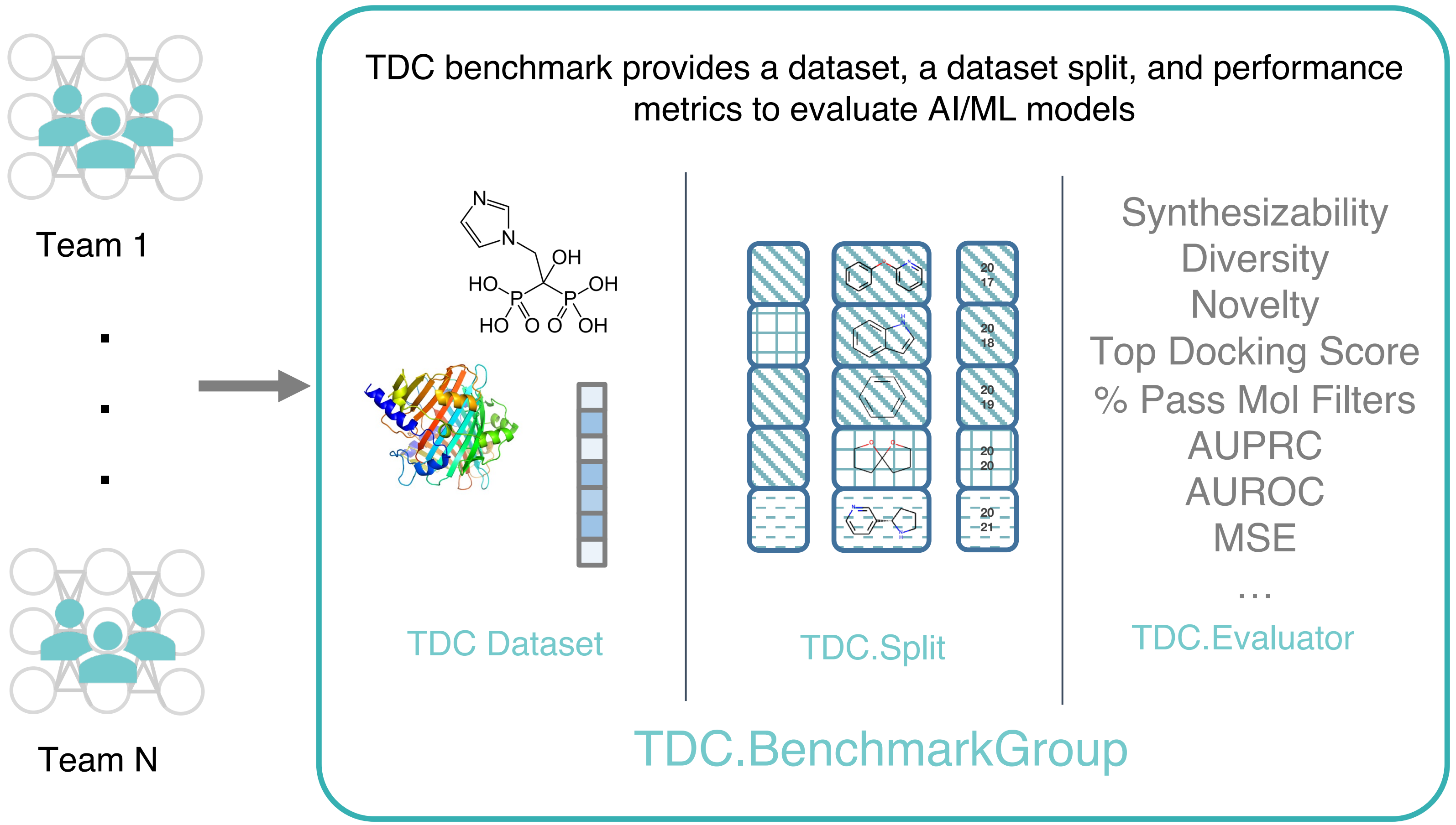
Token-level Classification Metric

- Average ROC-AUC

[redacted] processing helpers

tion Visualization
ation
onversion
g
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

Benchmarks and Leaderboards



Benchmarks and Leaderboards

Loading TDC benchmark class

Get ML-ready train/val/test data

Train your model

Return ready to submit metrics

```
from tdc.benchmark_group import admet_group
group = admet_group(path = 'data/')
benchmark = group.get('Caco2_Wang')

predictions = {}
name = benchmark['name']
train_val, test = benchmark['train_val'], benchmark['test']

## --- train your model --- ##

predictions[name] = y_pred
group.evaluate(predictions)
# {'caco2_wang': {'mae': 0.234}}
```

Benchmarks and Leaderboards

Loading TDC benchmark class ⟹

Get ML-ready train/val/test data ⟹

Train your model ⟹

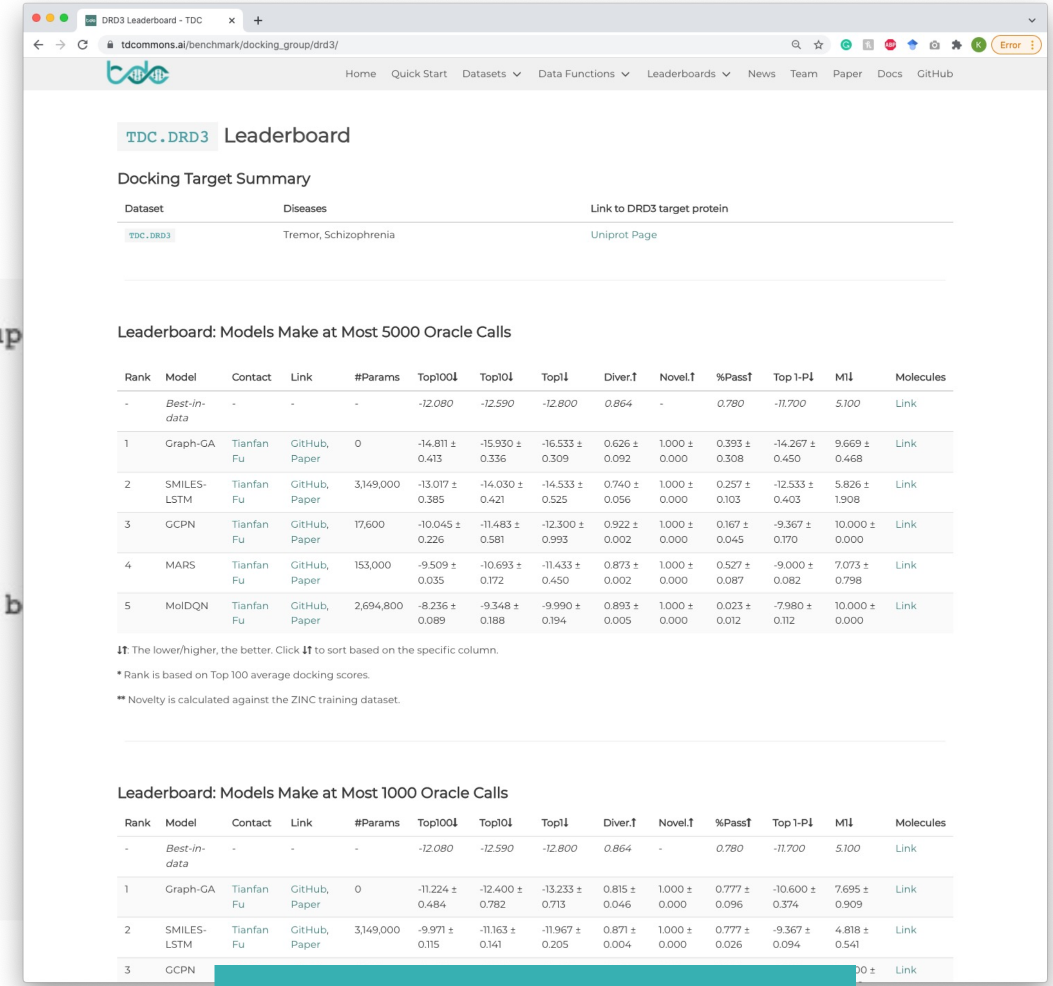
Return ready to submit metrics ⟹

```python
from tdc.benchmark_group import admet_group
group = admet_group(path = 'data/')
benchmark = group.get('Caco2_Wang')

predictions = {}
name = benchmark['name']
train_val, test = benchmark['train_val'], b

## --- train your model --- ##

predictions[name] = y_pred
group.evaluate(predictions)
# {'caco2_wang': {'mae': 0.234}}
```



DRD3 Leaderboard - TDC

tdcommons.ai/benchmark/docking_group/drd3/

Home Quick Start Datasets Data Functions Leaderboards News Team Paper Docs GitHub

TDC.DRD3 Leaderboard

Docking Target Summary

Dataset	Diseases	Link to DRD3 target protein
TDC.DRD3	Tremor, Schizophrenia	Uniprot Page

Leaderboard: Models Make at Most 5000 Oracle Calls

Rank	Model	Contact	Link	#Params	Top100↓	Top10↓	Top1↓	Diver.↑	Novel.↑	%Pass↑	Top 1-P↓	M1↓	Molecules
-	*Best-in-data*	-	-	-	*-12.080*	*-12.590*	*-12.800*	*0.864*	-	*0.780*	*-11.700*	*5.100*	Link
1	Graph-GA	Tianfan Fu	GitHub, Paper	0	-14.811 ± 0.413	-15.930 ± 0.336	-16.533 ± 0.309	0.626 ± 0.092	1.000 ± 0.000	0.393 ± 0.308	-14.267 ± 0.450	9.669 ± 0.468	Link
2	SMILES-LSTM	Tianfan Fu	GitHub, Paper	3,149,000	-13.017 ± 0.385	-14.030 ± 0.421	-14.533 ± 0.525	0.740 ± 0.056	1.000 ± 0.000	0.257 ± 0.103	-12.533 ± 0.403	5.826 ± 1.908	Link
3	GCPN	Tianfan Fu	GitHub, Paper	17,600	-10.045 ± 0.226	-11.483 ± 0.581	-12.300 ± 0.993	0.922 ± 0.002	1.000 ± 0.000	0.167 ± 0.045	-9.367 ± 0.170	10.000 ± 0.000	Link
4	MARS	Tianfan Fu	GitHub, Paper	153,000	-9.509 ± 0.035	-10.693 ± 0.172	-11.433 ± 0.450	0.873 ± 0.002	1.000 ± 0.000	0.527 ± 0.087	-9.000 ± 0.082	7.073 ± 0.798	Link
5	MolDQN	Tianfan Fu	GitHub, Paper	2,694,800	-8.236 ± 0.089	-9.348 ± 0.188	-9.990 ± 0.194	0.893 ± 0.005	1.000 ± 0.000	0.023 ± 0.012	-7.980 ± 0.112	10.000 ± 0.000	Link

↓↑: The lower/higher, the better. Click ↓↑ to sort based on the specific column.

* Rank is based on Top 100 average docking scores.

** Novelty is calculated against the ZINC training dataset.

Leaderboard: Models Make at Most 1000 Oracle Calls

Rank	Model	Contact	Link	#Params	Top100↓	Top10↓	Top1↓	Diver.↑	Novel.↑	%Pass↑	Top 1-P↓	M1↓	Molecules
-	*Best-in-data*	-	-	-	*-12.080*	*-12.590*	*-12.800*	*0.864*	-	*0.780*	*-11.700*	*5.100*	Link
1	Graph-GA	Tianfan Fu	GitHub, Paper	0	-11.224 ± 0.484	-12.400 ± 0.782	-13.233 ± 0.713	0.815 ± 0.046	1.000 ± 0.000	0.777 ± 0.096	-10.600 ± 0.374	7.695 ± 0.909	Link
2	SMILES-LSTM	Tianfan Fu	GitHub, Paper	3,149,000	-9.971 ± 0.115	-11.163 ± 0.141	-11.967 ± 0.205	0.871 ± 0.004	1.000 ± 0.000	0.777 ± 0.026	-9.367 ± 0.094	4.818 ± 0.541	Link
3	GCPN	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	Link

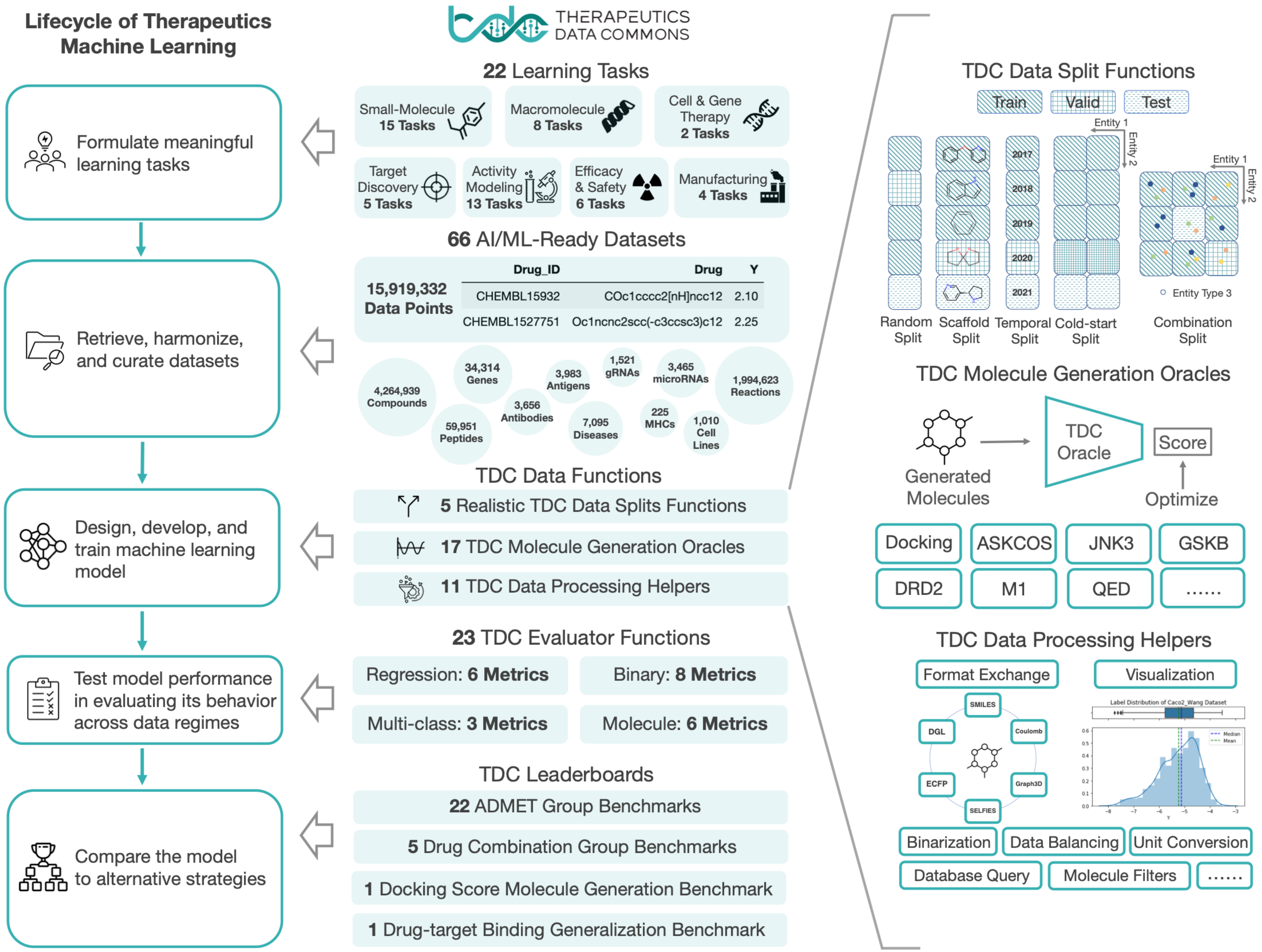
https://tdcommons.ai/benchmark/overview/

Benchmarks and Leaderboards

- Scaffold split
- No single method has the best performance across the board!

Raw Feature Type		Expert-Curated Methods		SMILES	Molecular Graph-Based Methods (state-of-the-Art in ML)				
Dataset	Metric	Morgan [31]	RDKit2D [24]	CNN [18]	NeuralFP [7]	GCN [23]	AttentiveFP [43]	AttrMasking [16]	ContextPred [16]
	# Params.	1477K	633K	227K	480K	192K	301K	2067K	2067K
TDC.Caco2 (↓)	MAE	0.908±0.060	**0.393±0.024**	0.446±0.036	0.530±0.102	0.599±0.104	0.401±0.032	0.546±0.052	0.502±0.036
TDC.HIA (↑)	AUROC	0.807±0.072	0.972±0.008	0.869±0.026	0.943±0.014	0.936±0.024	0.974±0.007	**0.978±0.006**	0.975±0.004
TDC.Pgp (↑)	AUROC	0.880±0.006	0.918±0.007	0.908±0.012	0.902±0.020	0.895±0.021	0.892±0.012	**0.929±0.006**	0.923±0.005
TDC.Bioav (↑)	AUROC	0.581±0.086	**0.672±0.021**	0.613±0.013	0.632±0.036	0.566±0.115	0.632±0.039	0.577±0.087	0.671±0.026
TDC.Lipo (↓)	MAE	0.701±0.009	0.574±0.017	0.743±0.020	0.563±0.023	0.541±0.011	0.572±0.007	0.547±0.024	**0.535±0.012**
TDC.AqSol (↓)	MAE	1.203±0.019	0.827±0.047	1.023±0.023	0.947±0.016	0.907±0.020	**0.776±0.008**	1.026±0.020	1.040±0.045
TDC.BBB (↑)	AUROC	0.823±0.015	0.889±0.016	0.781±0.030	0.836±0.009	0.842±0.016	0.855±0.011	0.892±0.012	**0.897±0.004**
TDC.PPBR (↓)	MAE	12.848±0.362	9.994±0.319	11.106±0.358	**9.292±0.384**	10.194±0.373	9.373±0.335	10.075±0.202	9.445±0.224
TDC.VD (↑)	Spearman	0.493±0.011	**0.561±0.025**	0.226±0.114	0.258±0.162	0.457±0.050	0.241±0.145	0.559±0.019	0.485±0.092
TDC.CYP2D6-I (↑)	AUPRC	0.587±0.011	0.616±0.007	0.544±0.053	0.627±0.009	0.616±0.020	0.646±0.014	0.721±0.009	**0.739±0.005**
TDC.CYP3A4-I (↑)	AUPRC	0.827±0.009	0.829±0.007	0.821±0.003	0.849±0.004	0.840±0.010	0.851±0.006	0.902±0.002	**0.904±0.002**
TDC.CYP2C9-I (↑)	AUPRC	0.715±0.004	0.742±0.006	0.713±0.006	0.739±0.010	0.735±0.004	0.749±0.004	0.829±0.003	**0.839±0.003**
TDC.CYP2D6-S (↑)	AUPRC	0.671±0.066	0.677±0.047	0.485±0.037	0.572±0.062	0.617±0.039	0.574±0.030	0.704±0.028	**0.736±0.024**
TDC.CYP3A4-S (↑)	AUROC	0.633±0.013	0.639±0.012	**0.662±0.031**	0.578±0.020	0.590±0.023	0.576±0.025	0.582±0.021	0.609±0.025
TDC.CYP2C9-S (↑)	AUPRC	0.380±0.015	0.360±0.040	0.367±0.059	0.359±0.059	0.344±0.051	0.375±0.032	0.381±0.045	**0.392±0.026**
TDC.Half_Life (↑)	Spearman	**0.329±0.083**	0.184±0.111	0.038±0.138	0.177±0.165	0.239±0.100	0.085±0.068	0.151±0.068	0.129±0.114
TDC.CL-Micro (↑)	Spearman	0.492±0.020	**0.586±0.014**	0.252±0.116	0.529±0.015	0.532±0.033	0.365±0.055	0.585±0.034	0.578±0.007
TDC.CL-Hepa (↑)	Spearman	0.272±0.068	0.382±0.007	0.235±0.021	0.401±0.037	0.366±0.063	0.289±0.022	0.413±0.028	**0.439±0.026**
TDC.hERG (↑)	AUROC	0.736±0.023	**0.841±0.020**	0.754±0.037	0.722±0.034	0.738±0.038	0.825±0.007	0.778±0.046	0.756±0.023
TDC.AMES (↑)	AUROC	0.794±0.008	0.823±0.011	0.776±0.015	0.823±0.006	0.818±0.010	0.814±0.008	**0.842±0.008**	0.837±0.009
TDC.DILI (↑)	AUROC	0.832±0.021	0.875±0.019	0.792±0.016	0.851±0.026	0.859±0.033	0.886±0.015	**0.919±0.008**	0.861±0.018
TDC.LD50 (↓)	MAE	0.649±0.019	0.678±0.003	0.675±0.011	0.667±0.020	0.649±0.026	0.678±0.012	**0.685±0.025**	0.669±0.030

TDC Serves the Whole Lifecycle



And Innovation!

Lifecycle of Therapeutics Machine Learning

Formulate meaningful learning tasks

Retrieve, harmonize, and curate datasets

Design, develop, and train machine learning model

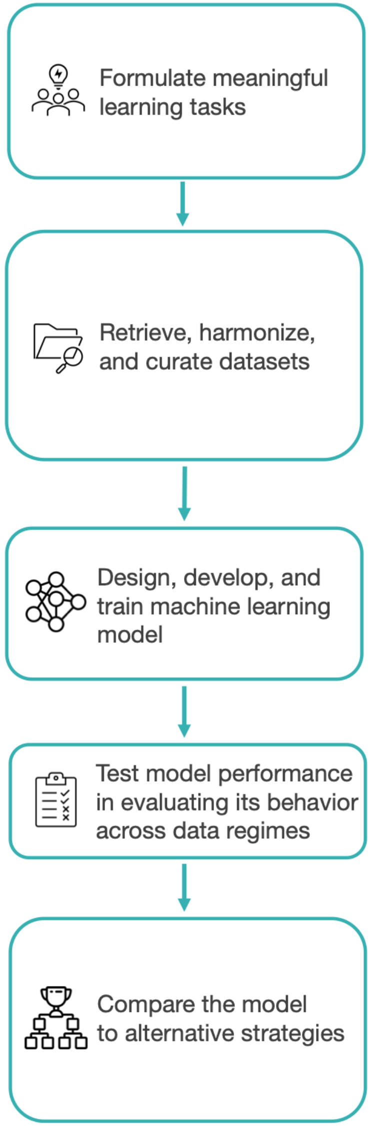
. . .

Test model performance in evaluating its behavior across data regimes

Compare the model to alternative strategies

And Innovation!

Lifecycle of Therapeutics Machine Learning



Not just a low-code implementation of existing models!

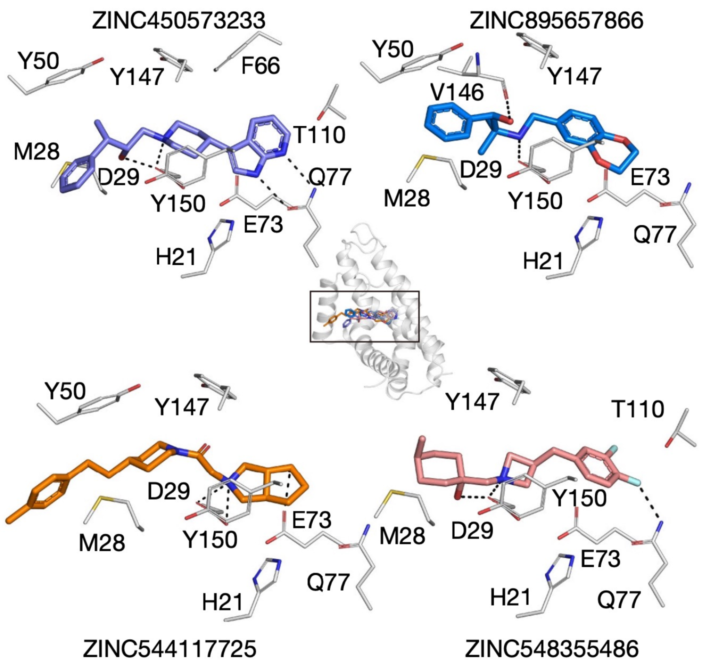
```python
from torchdrug import core, models, tasks, utils

model = models.GIN(input_dim=dataset.node_feature_dim,
                   hidden_dims=[256, 256, 256, 256],
                   short_cut=True, batch_norm=True, concat_hidden=True)
task = tasks.PropertyPrediction(model, task=dataset.tasks,
                                criterion="bce", metric=("auprc", "auroc"))
```

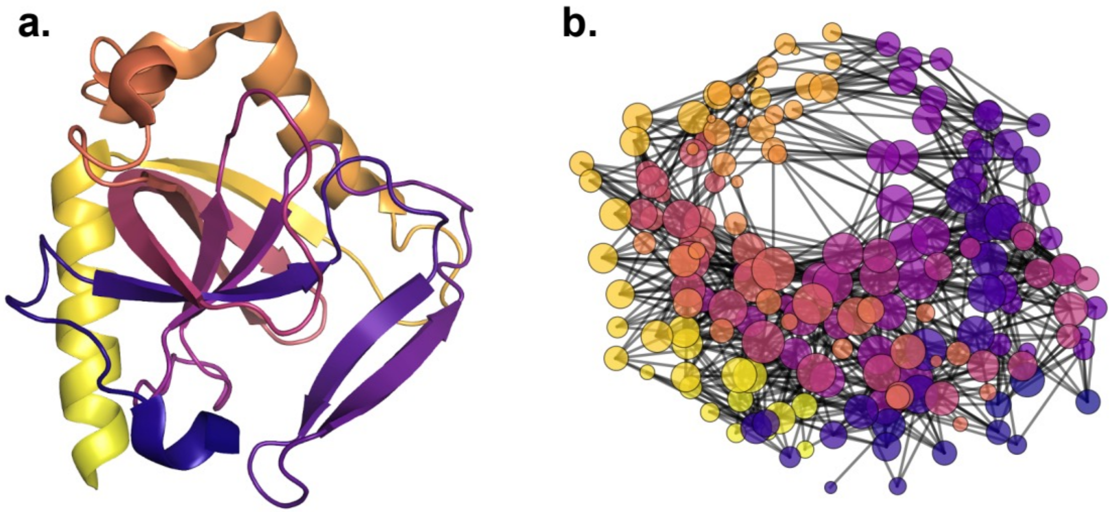
```python
import deepchem as dc
from deepchem.models.graph_models import GraphConvModel
```

More Coming Soon

- 3D pocket molecule generation datasets.
- 3D representation convertor
- Protein representation with Graphein
- ……

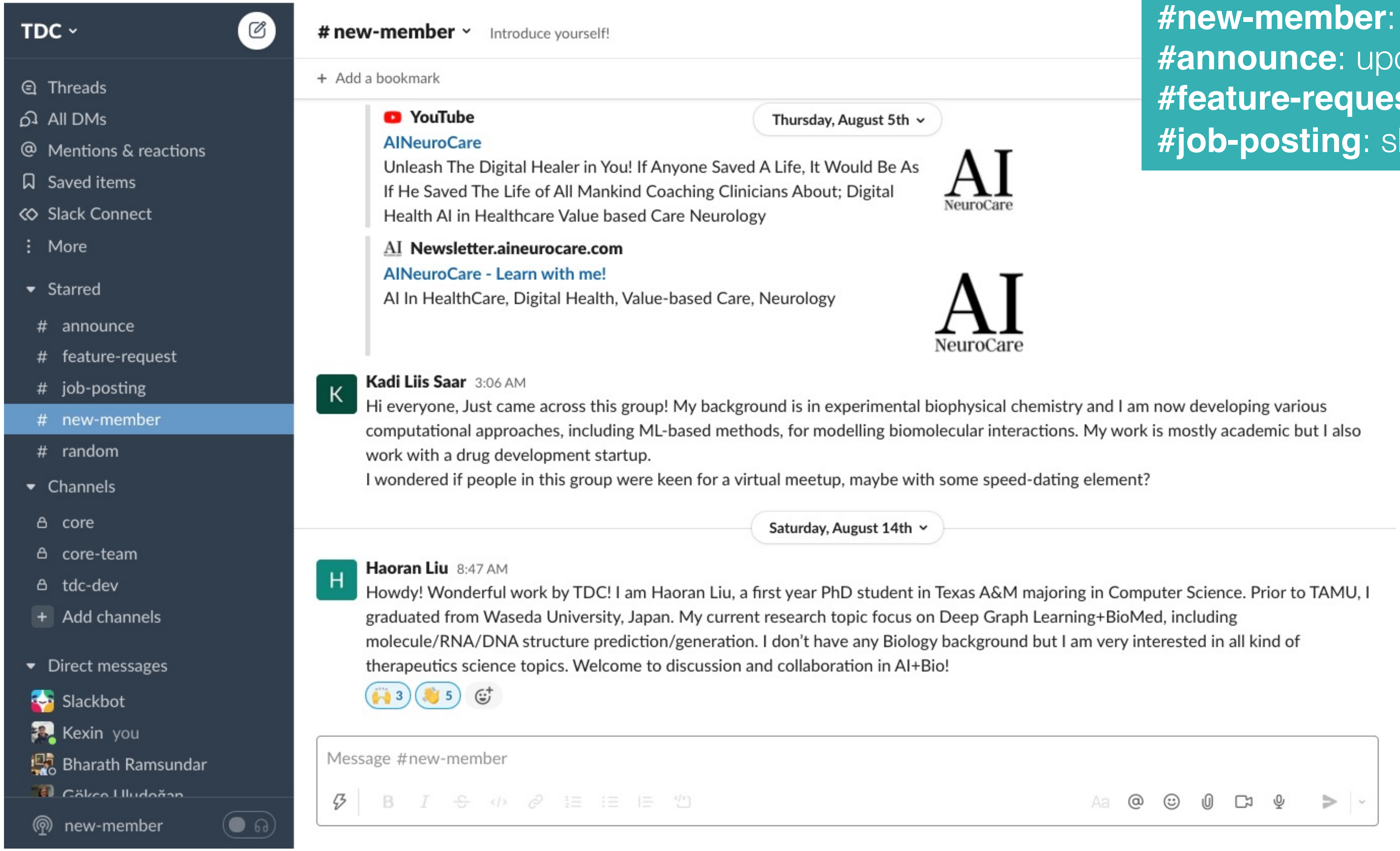


Pocket-based 3D molecular generation



Protein graphs representation

Slack Channel for Users



#new-member: introduce and connect
#announce: update from TDC team
#feature-request: discuss new features
#job-posting: sharing relevant positions

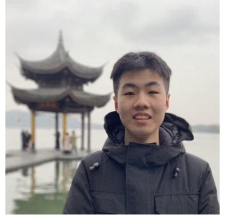
tinyurl.com/tdc-slack

TDC is an Open Science Initiative!

https://tdcommons.ai

THERAPEUTICS DATA COMMONS

- We welcome contributions from valuable therapeutic problems posing to data deposit and adding functions.

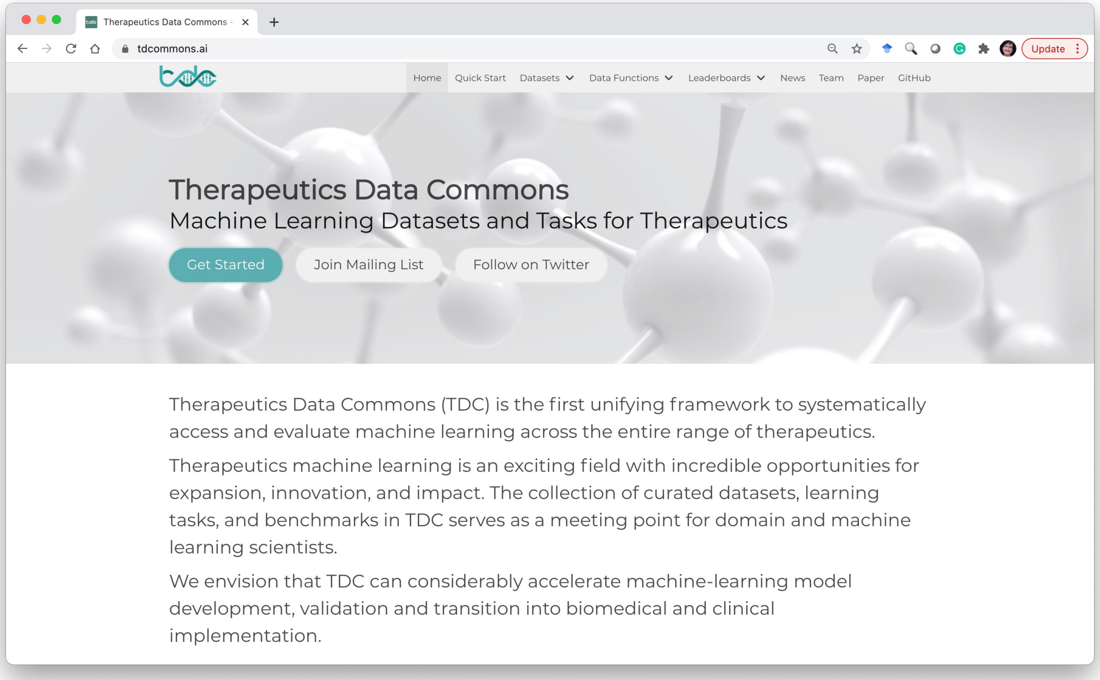


Kexin Huang

Tianfan Fu

Wenhao Gao

Marinka Zitnik



HARVARD UNIVERSITY

Georgia Tech

Massachusetts Institute of Technology

Carnegie Mellon University

Stanford University

IQVIA

ILLINOIS