# THERAPEUTICS DATA COMMONS


**Kexin Huang**
Harvard
kexinhuang@hsph.harvard.edu

**Tianfan Fu**
Georgia Tech
tfu42@gatech.edu

**Wenhao Gao**
MIT
whgao@mit.edu

**Yue Zhao**
CMU
zhaoy@cmu.edu

**Marinka Zitnik**
Harvard
marinka@hms.harvard.edu





# Therapeutics are one of most exciting areas for machine learning

## However…

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

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

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



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

THERAPEUTICS DATA COMMONS

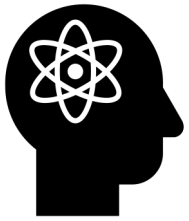
# Machine Learning Datasets for Therapeutics

- **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 and model evaluators
  - Data processing and splits, molecule generation oracles, and much more
- **3 Lines of Code:**
  - Minimum package dependency, lightweight loaders



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

# Our Vision for TDC

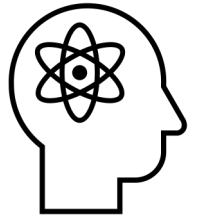


Domain scientists

Identify meaningful therapeutics tasks

THERAPEUTICS DATA COMMONS

Design powerful ML models



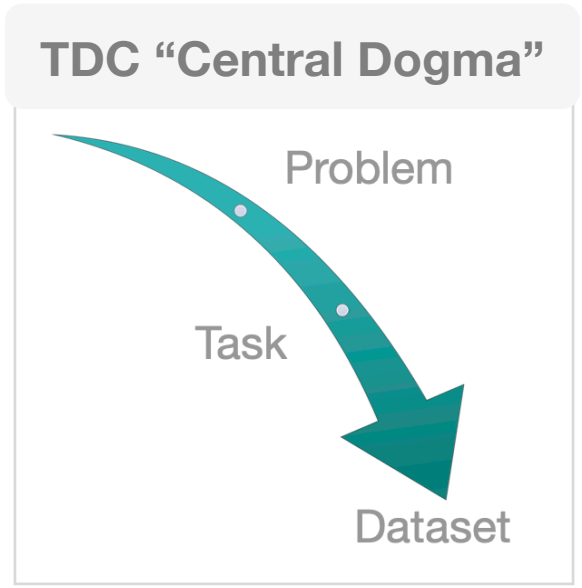
ML scientists

**Advancing algorithms for key therapeutics problems**

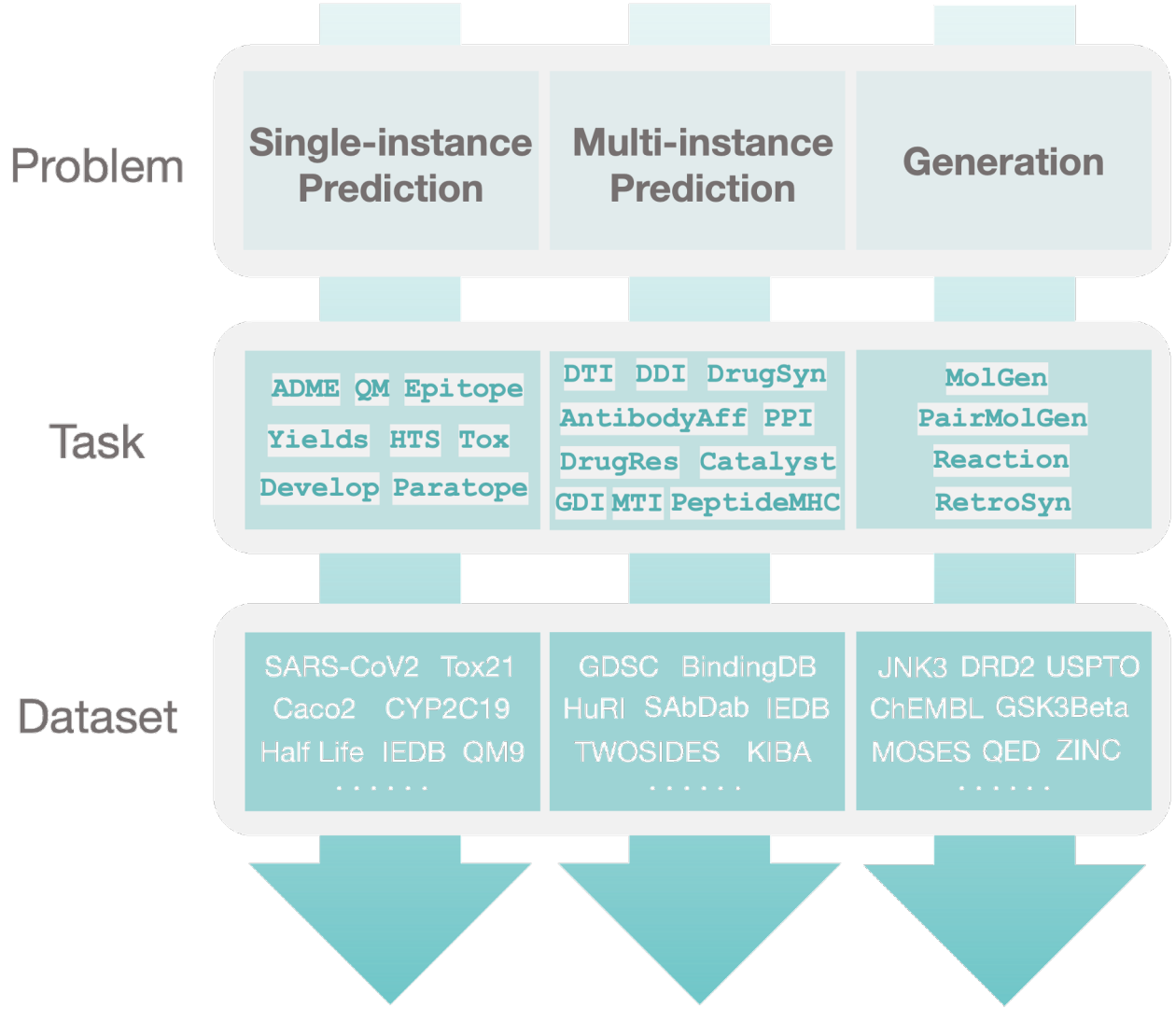


https://github.com/mims-harvard/TDC

# Modular Structure of TDC








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

# Diverse Coverage of Tasks

## Single-instance Prediction

| Products / Areas | Target Discovery | Activity | Efficacy and Safety | Manufacturing |
|---|---|---|---|---|
| Small Molecule | | HTS QM | ADME Tox | Yields |
| Biologics | | Paratope Epitope | Develop | |

## Multi-instance Prediction

| Products / Areas | Target Discovery | Activity | Efficacy and Safety | Manufacturing |
|---|---|---|---|---|
| Small Molecule | DTI GDA | DTI PPI DrugRes DrugSyn | DDI | Catalyst |
| Biologics | MTI | PPI PeptideMHC AntibodyAff | | |

## Generation

| Products / Areas | Target Discovery | Activity | Efficacy and Safety | Manufacturing |
|---|---|---|---|---|
| Small Molecule | | MolGen PairMolGen | MolGen PairMolGen | RetroSyn Reaction |
| Biologics | | | | |



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

# Initial release:
# 62 datasets

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 al.
- 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

- QM7
- QM8
- QM9

**QM**

DATASET INDEX

- SAbDab, Liberis et al.

**Paratope**

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**

DATASET INDEX

- MHC Class II, IEDB, Jensen et al.

**PeptideMHC**

DATASET INDEX

- SAbDab

**AntibodyAff**

DATASET INDEX

- 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**



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

## 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.

```
pip install PyTDC
```

```
In [1]: from tdc.single_pred import ADME
        data = ADME(name = 'Caco2_Wang')
        split = data.get_split(seed = 'benchmark')
```

```
Downloading...
100%|██████████| 84.3k/84.3k [00:00<00:00, 970kiB/s]
Loading...
Done!
```

```
In [2]: split['test'].head(2)
```

Out[2]:

| | Drug_ID | Drug | Y |
|---|---|---|---|
| 0 | VLA-4 antagonist 3 | S1CN(S(=O)(=O)c2cn(nc2)C)[C@H](C(=O)N[C@@H](Cc... | -5.17 |
| 1 | Astilbin | O1[C@@H](C)[C@H](O)[C@@H](O)[C@@H](O)[C@@H]1O[... | -6.82 |



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

# Highlight: 24 ADMET Datasets

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
- Hydration Free Energy, FreeSolv

Distribution

- BBB (Blood-Brain Barrier), Adenot et al.
- 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.

Excretion

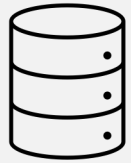
- Half Life, eDrug3D
- Clearance, eDrug3D

Toxicity

- Tox21
- ToxCast
- ClinTox

Data sources

**Paper Supplementary**



**Public Database**

**Bioassays**

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

# Highlight: Precision Polytherapy

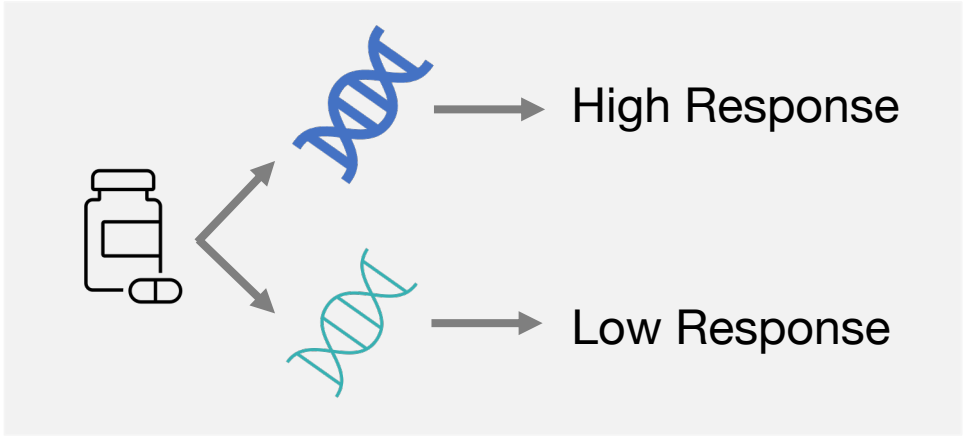
## Drug Response Prediction

DATASET INDEX

GDSC1

GDSC2

`DrugRes`

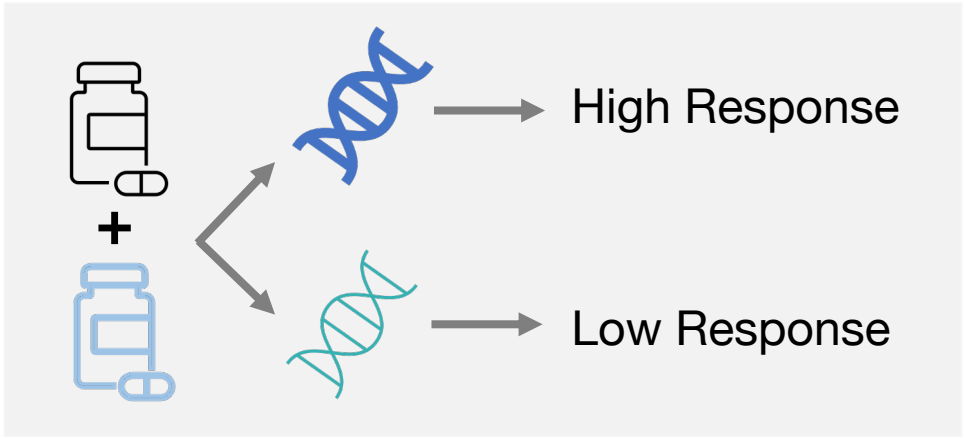



## Drug Synergy Prediction

DATASET INDEX

OncoPolyPharmacology

`DrugSyn`






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

# Highlight: 10 Biologics Datasets

## Paratope Prediction

| DATASET INDEX | |
|---|---|
| SAbDab, Liberis et al. | `Paratope` |

## Epitope Prediction

| DATASET INDEX | |
|---|---|
| IEDB, Jespersen et al.<br>PDB, Jespersen et al. | `Epitope` |

## Antibody-Antigen Affinity Prediction

| DATASET INDEX | |
|---|---|
| SAbDab | `AntibodyAff` |

## Antibody Developability Prediction

| DATASET INDEX | |
|---|---|
| TAP<br>SAbDab, Chen et al. | `Develop` |

## Peptide-MHC Binding Prediction

| DATASET INDEX | |
|---|---|
| MHC Class I, IEDB-IMGT, Nielsen et al.<br>MHC Class II, IEDB, Jensen et al. | `Peptide MHC` |

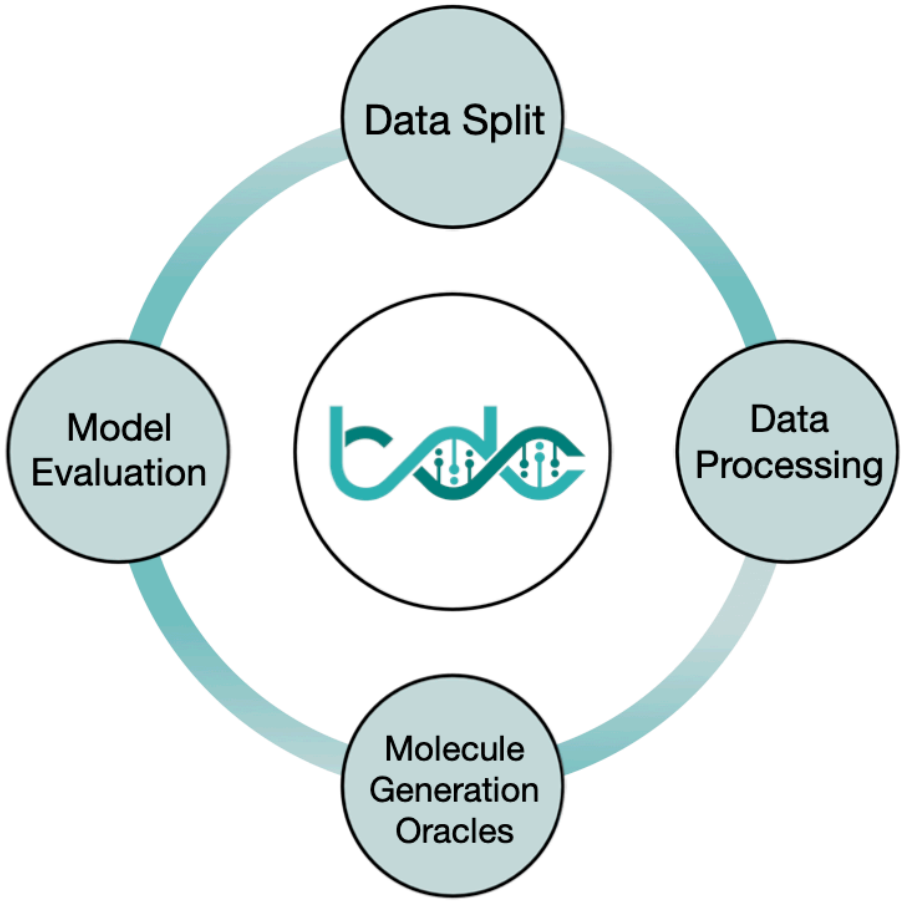
## miRNA-Target Interaction Prediction

| DATASET INDEX | |
|---|---|
| miRTarBase | `MTI` |



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

# 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

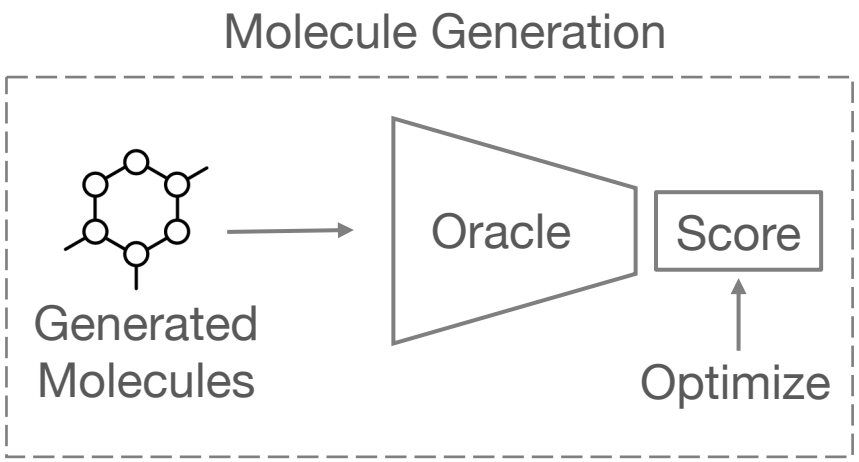
## Data processing helpers

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

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

# Molecule Generation Oracles




3 Lines of Code

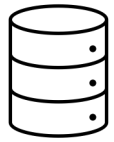
```
In [1]: from tdc import Oracle
        oracle = Oracle(name = 'GSK3B')
        oracle('CC1=CN=C(N1)C2=CN=C(N=C2C3=C(C=C(C=C3)Cl)Cl)NCCNC4=NC=C(C=C4)C#N.Cl')

        Downloading Oracle...
        100%|██████████| 27.8M/27.8M [00:01<00:00, 16.5MiB/s]
        Done!

Out[1]: 0.68

In [ ]:
```

GuacaMol



MOSES

Literature

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

**20 Oracles**

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

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

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

## Contribute

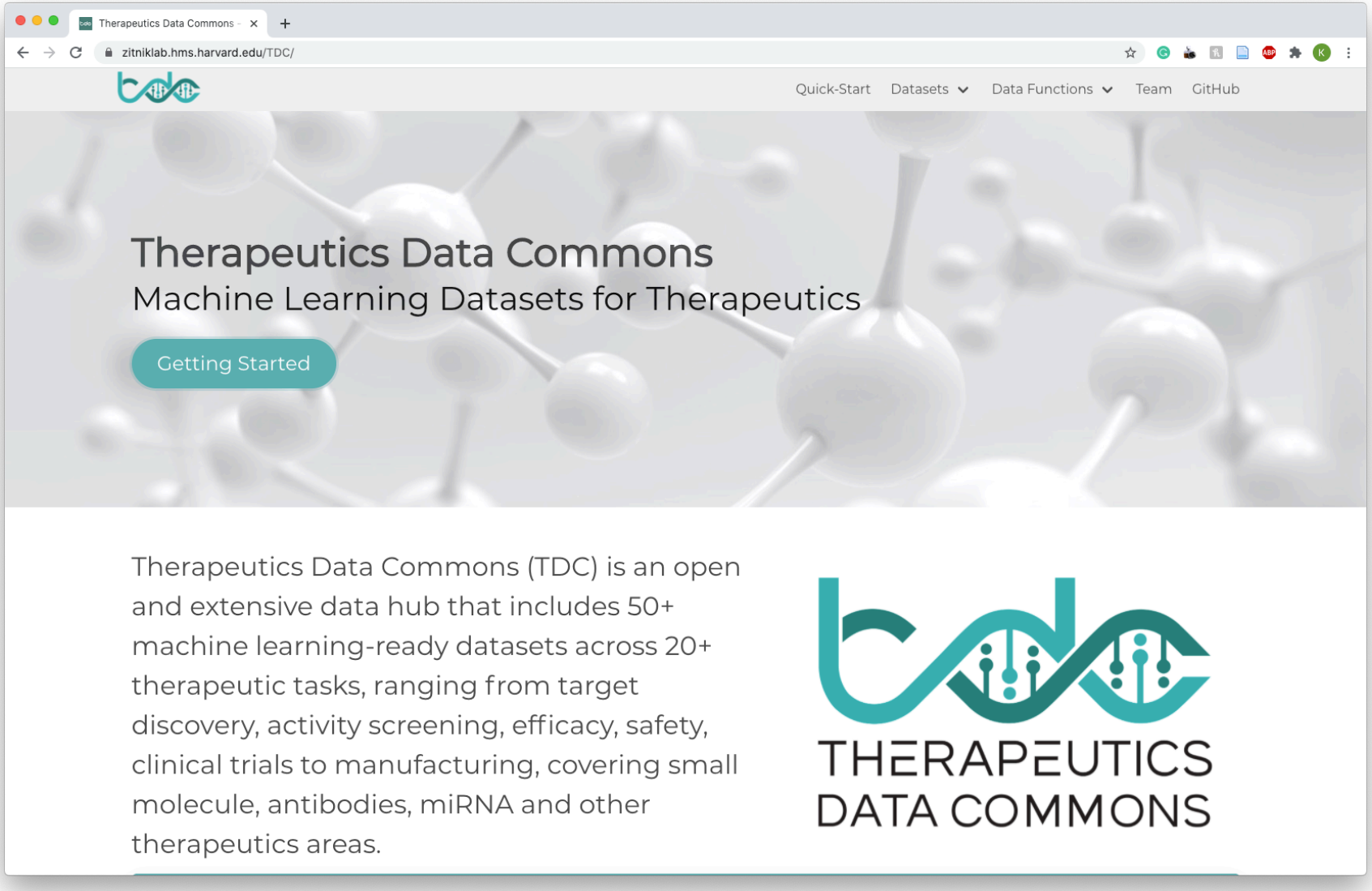
| Tasks | Datasets | Data Functions |
| --- | --- | --- |
| Clinical Trials,<br>CRISPR,<br>Phenotypic Screening,<br>Protein Contact,<br>Crystal Structure<br>..... | HTS,<br>ADME,<br>Drug Response,<br>Drug Synergy,<br>Reactions,<br>Antibody affinity,<br>..... | Data Wrangling,<br>Data Visualization,<br>Realistic Splits,<br>Molecule Generation Oracles,<br>..... |

Fill in this form: **rb.gy/ytbyfl**



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

# zitniklab.hms.harvard.edu/TDC






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

# Star, Share, and Contribute to TDC

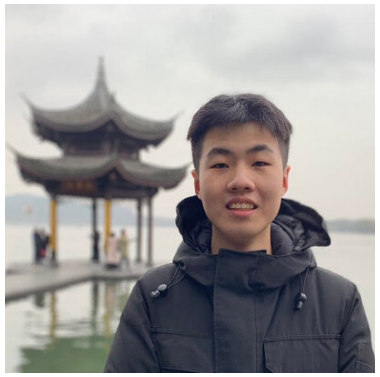
GitHub

zitniklab.hms.harvard.edu/TDC

github.com/mims-harvard/TDC

groups.io/g/tdc

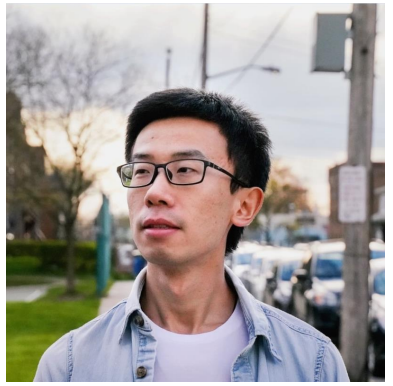
Website

Kexin Huang
@KexinHuang5
Harvard
kexinhuang@hsph.harvard.edu

Tianfan Fu
@TianfanFu
Georgia Tech
tfu42@gatech.edu

Wenhao Gao
@WenhaoGao1
MIT
whgao@mit.edu



Yue Zhao
@yzhao062
CMU
zhaoy@cmu.edu

Marinka Zitnik
@marinkazitnik
Harvard
marinka@hms.harvard.edu