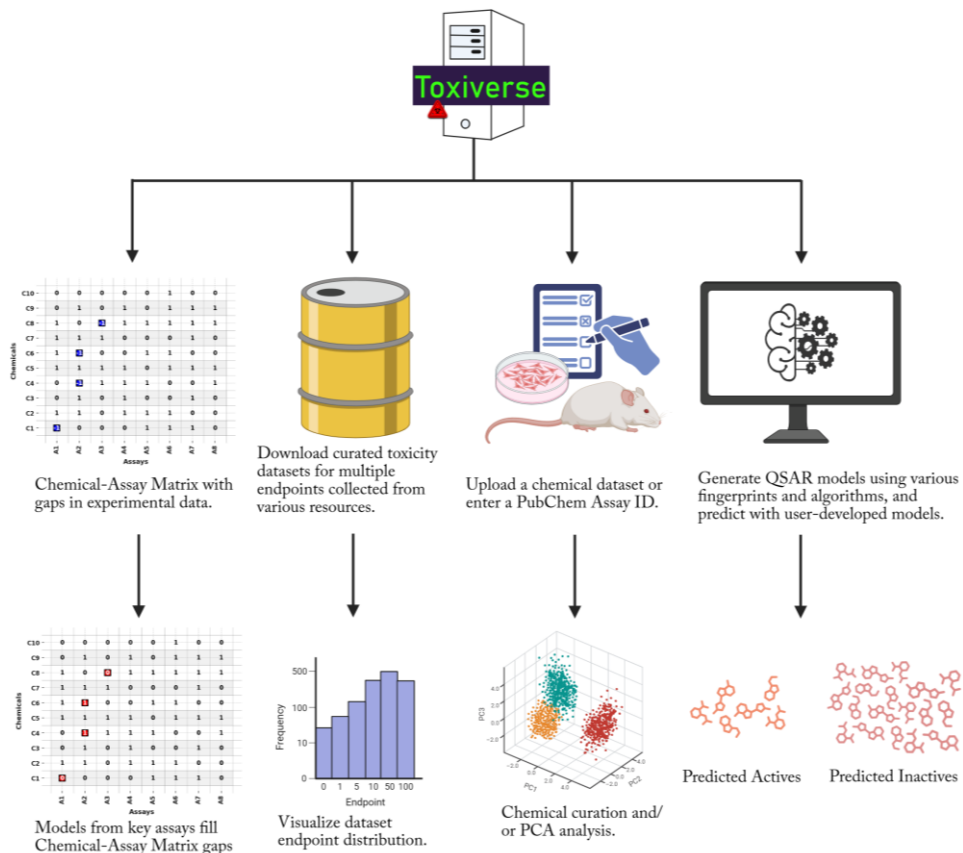




# ToxiVerse

## Tutorial

[www.toxiverse.com](http://www.toxiverse.com)



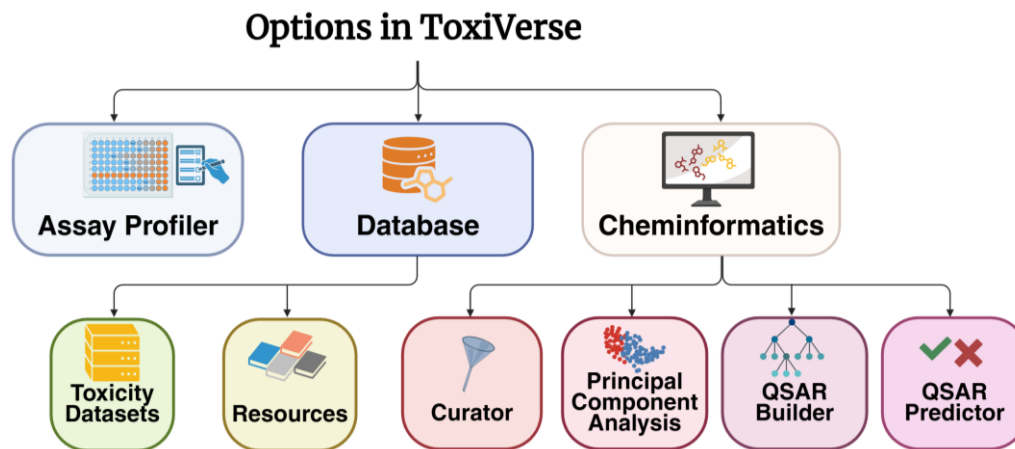
# Contents

1. Introduction.....	3
2. Assay Profiler .....	5
3. Database .....	10
4. Resources.....	13
5. Upload or Retrieve datasets.....	15
6. Curator.....	21
7. Principal Component Analysis.....	24
8. QSAR Builder.....	26
9. QSAR Predict.....	29
10. Contact us.....	31
11. About us.....	31

## 1. Introduction

Computational toxicology plays a significant role in identifying hazardous compounds to protect human health and the environment in a cost-effective manner. A major challenge in this field is the lack of publicly available and user-friendly computational tools that can be used for chemical risk assessment, especially by users with limited computational expertise.

To address this need, we developed Toxicology Universe (ToxiVerse), a web portal designed to assist toxicologists, pharmaceutical researchers, and chemists in assessing chemical safety. Please check Figure below for all the available options.



The options in ToxiVerse. They are explained in detail along with step-by-step tutorial in the following pages.

ToxiVerse offers the following functions:

- I. **Profile bioassay results** from PubChem for chemicals of interest and fill experimental data gaps using QSAR models built from key assays.
- II. **Download and visualize** curated toxicological datasets using Principal Component Analysis (PCA) and endpoint distribution plots. The integrated database includes over 40,000 chemicals across 50 toxicity endpoints, compiled from various sources.
- III. **Create QSAR models** using either user-uploaded datasets or datasets retrieved from PubChem by providing an Assay ID. A variety of molecular descriptors and machine learning algorithms are supported.
- IV. **Curate** and analyze datasets using **Principal Component Analysis** (PCA) before model development.
- V. **Predict toxicity** for new chemicals using Quantitative Structure-Activity Relationship (QSAR) models developed within the platform.

1. These are the main modules.

ToxiVerse

Home

Assay Profiler

Database

Cheminformatics

Tutorial

Tasks

About

Welcome to

Database

Toxicity Datasets

Resources

2. Available options under the Database module.



ToxiVerse

Cheminformatics

Upload or Retrieve Datasets

Curator

Principal Component Analysis

QSAR Builder

QSAR Predictor

3. Available options under the Cheminformatics module.

ToxiVerse is an online chemical data analysis portal that allows users to profile chemicals based on bioassay responses, download curated toxicological datasets, and perform common cheminformatics analyses. These include Quantitative Structure–Activity Relationship (QSAR) modeling, chemical space visualization, and more.

There are three main modules accessible from ToxiVerse. The **Assay Profiler** module enables users to profile chemicals based on their bioassay responses. The **Database** module provides access to curated toxicological datasets, along with detailed information about their data sources. The **Cheminformatics** module offers tools for visualizing chemical space and performing various cheminformatics analyses.



# Assay Profiler

## 2. Assay Profiler

Users can upload up to 500 chemicals to profile them using PubChem bioassays. It fills experimental data gaps using QSAR models built from key assays.

# 1. Upload a file with PubChem Compound IDs eg: sample\_dataset.txt

A sample file for Assay Profiler looks like below.

```
95170
104926
153909
155166
441071
533728
616765
643833
2723650
3032581
3032791
```

## Assay Profiler

Users can upload up to 500 chemicals to profile them using PubChem bioassays.

\*Accepts up to 500 chemicals.

Upload a file containing the CIDs of compounds you would like to profile on assay outcomes. This should be a single-column txt or csv file containing the PubChem CID's of the compounds you wish to obtain HTS data. Check the sample files provided below. If you have compounds as another identifier (e.g., SMILES) PubChem offers a helpful batch-conversion tool: <https://pubchem.ncbi.nlm.nih.gov/idxexchange/idxexchange.cgi>

Please use the Sample buttons to download them.

Download Sample CSV

Download Sample TXT

Upload Dataset File (Formats: .txt, .csv):

Choose File No file chosen

Build Bioprofile

2. Click here

Results:

Download Bioprofile

Download Heatmap

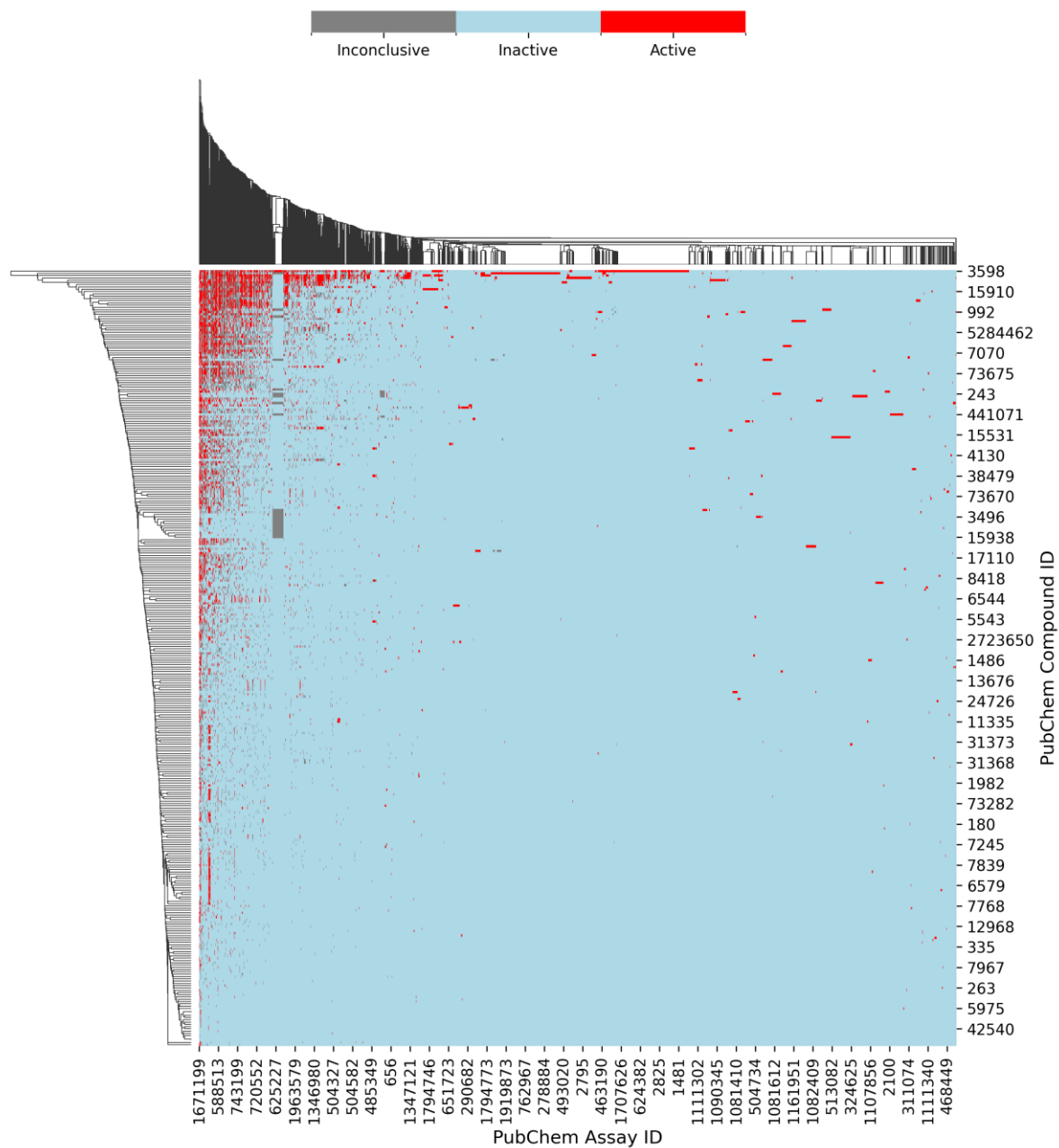
Download Model Metrics

Download Model Metrics Plot

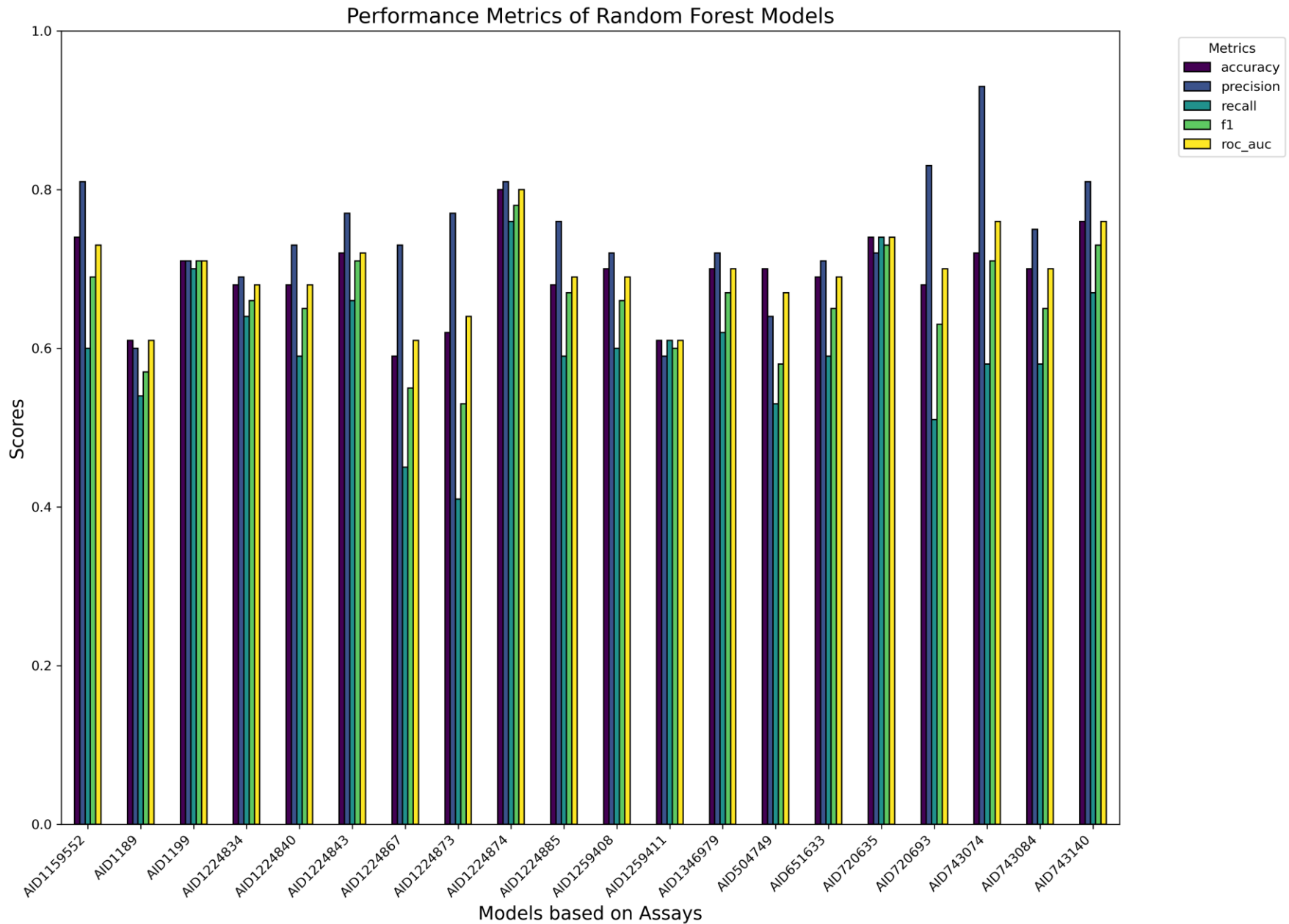
Download Filled Data Gaps Matrix

3. Click on options one by one to download the results.

## Download heatmap result looks like this



# Download Model Metrics Plot result looks like this





## Download Bioprofile result looks like this

CID	1	7	9	19	25	29	67
4	0	0	0	0	0	0	0
6	0	0	0	0	0	0	0
11	-1	0	0	0	0	0	0
13	0	0	0	0	0	0	0
15	0	0	0	0	0	0	0
33	0	-1	0	0	-1	0	0
34	0	0	0	0	0	0	0
51	0	0	0	0	0	0	0
66	0	0	-1	0	0	0	0
72	0	0	0	0	0	0	0
76	0	0	0	0	0	0	0
79	0	0	0	0	0	0	0

## Download Model Metrics Plot looks like this

Model	accuracy	precision	recall	f1	roc_auc
AID1159515_rf_model	0.67	0.81	0.47	0.6	0.67
AID1159555_rf_model	0.74	0.78	0.65	0.71	0.74
AID1224841_rf_model	0.72	0.79	0.55	0.65	0.71
AID1224847_rf_model	0.72	0.75	0.62	0.68	0.72
AID1224871_rf_model	0.7	0.75	0.57	0.65	0.7

## Download Filled Data Gaps Matrix result looks like this

CID	1	7	9	19	25	29	67
4	0	0	0	0	0	0	0
6	0	0	0	0	0	0	0
11	1	0	0	0	0	0	0
13	0	0	0	0	0	0	0
15	0	0	0	0	0	0	0
33	0	0	0	0	1	0	0
34	0	0	0	0	0	0	0
51	0	0	0	0	0	0	0
66	0	0	0	0	0	0	0
72	0	0	0	0	0	0	0
76	0	0	0	0	0	0	0
79	0	0	0	0	0	0	0

Please compare the top and bottom Tables. The CIDs (chemicals) and columns (AIDs) are shown. The above circled -1 values (blue circled) are replaced with 1s or 0s below (red circled) based on prediction scores.

# Database



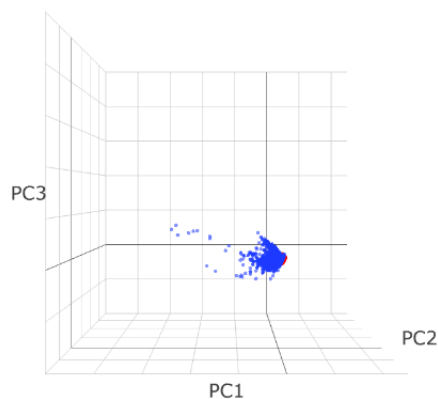
# Toxicity Datasets

### **3. Toxicity Datasets**

Allows users to download and visualize curated toxicological datasets, including endpoint distributions, and provides relevant bioassays for the selected endpoints. The dataset contains over 40,000 records for 26,000 chemicals across 43 endpoints, collected from various sources.

# Select an endpoint in a toxicity dataset

## Principal Component Analysis



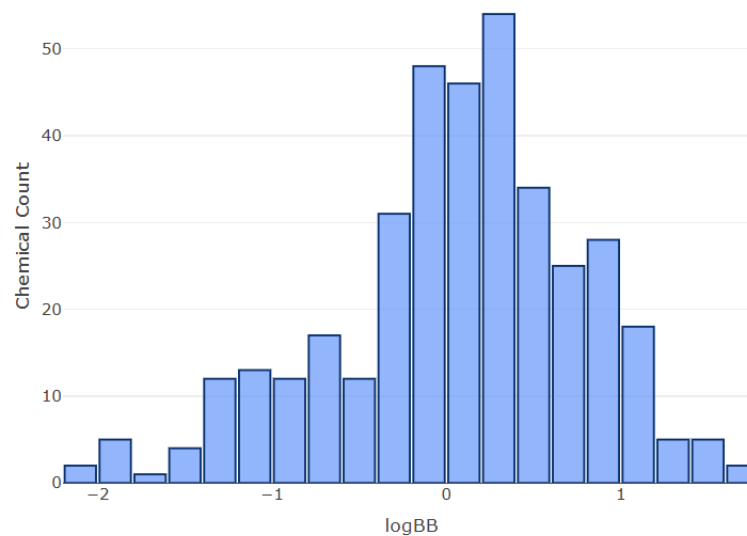
- Compounds with endpoint data in this dataset
- Compounds without endpoint data in this dataset
- Compounds from other datasets

## Endpoint Distribution

Select endpoint:

logBB (BBB\_curated)

Distribution of logBB



Relevant bioassays -----> This is relevant Bioassays to the selected endpoint.

This is the table of bioassays ranked by their active rates (active compounds numbers in all bioassays).

Show 

10

 entries

Search:

AID	Bioassay name	Inactive	Inconclusive	Active	Active rate
625256	DRUGMATRIX: Dopamine Transporter radioligand binding (ligand: [125I] RTI-55)	0	644	49	1
678713	Inhibition of human CYP2C9 assessed as ratio of IC50 in absence of NADPH to IC50 for presence of NADPH using 7-methoxy-4-trifluoromethylcoumarin-3-acetic acid as substrate after 30 mins	0	142	206	1
625204	DRUGMATRIX: Adrenergic beta1 radioligand binding (ligand: [125I] Cyanopindolol)	0	640	53	1
625205	DRUGMATRIX: Adrenergic beta2 radioligand binding (ligand: [3H] CGP-12177)	0	637	56	1
625207	DRUGMATRIX: Norepinephrine Transporter radioligand binding (ligand: [125I] RTI-55)	0	576	117	1
625215	DRUGMATRIX: Calcium Channel Type L, Benzothiazepine radioligand binding (ligand: [3H] Diltiazem)	0	603	90	1
625217	DRUGMATRIX: Serotonin (5-Hydroxytryptamine) 5-HT2B radioligand binding (ligand: [3H] Lysergic acid diethylamide)	0	537	156	1
625218	DRUGMATRIX: Serotonin (5-Hydroxytryptamine) 5-HT2C radioligand binding (ligand: [3H] Mesulergine)	0	550	143	1
625219	DRUGMATRIX: Serotonin (5-Hydroxytryptamine) 5-HT3 radioligand binding (ligand: [3H] GR-65630)	0	660	33	1
625221	DRUGMATRIX: Serotonin (5-Hydroxytryptamine) 5-HT6 radioligand binding (ligand: [3H] Lysergic acid diethylamide)	0	598	95	1

Select database to download

Select database:

BBB\_curated

Download database

-----> Select a database and click to download it as csv file. 12

# Database



# Resources

#### **4. Resources**

The details of the downloadable datasets available including references.

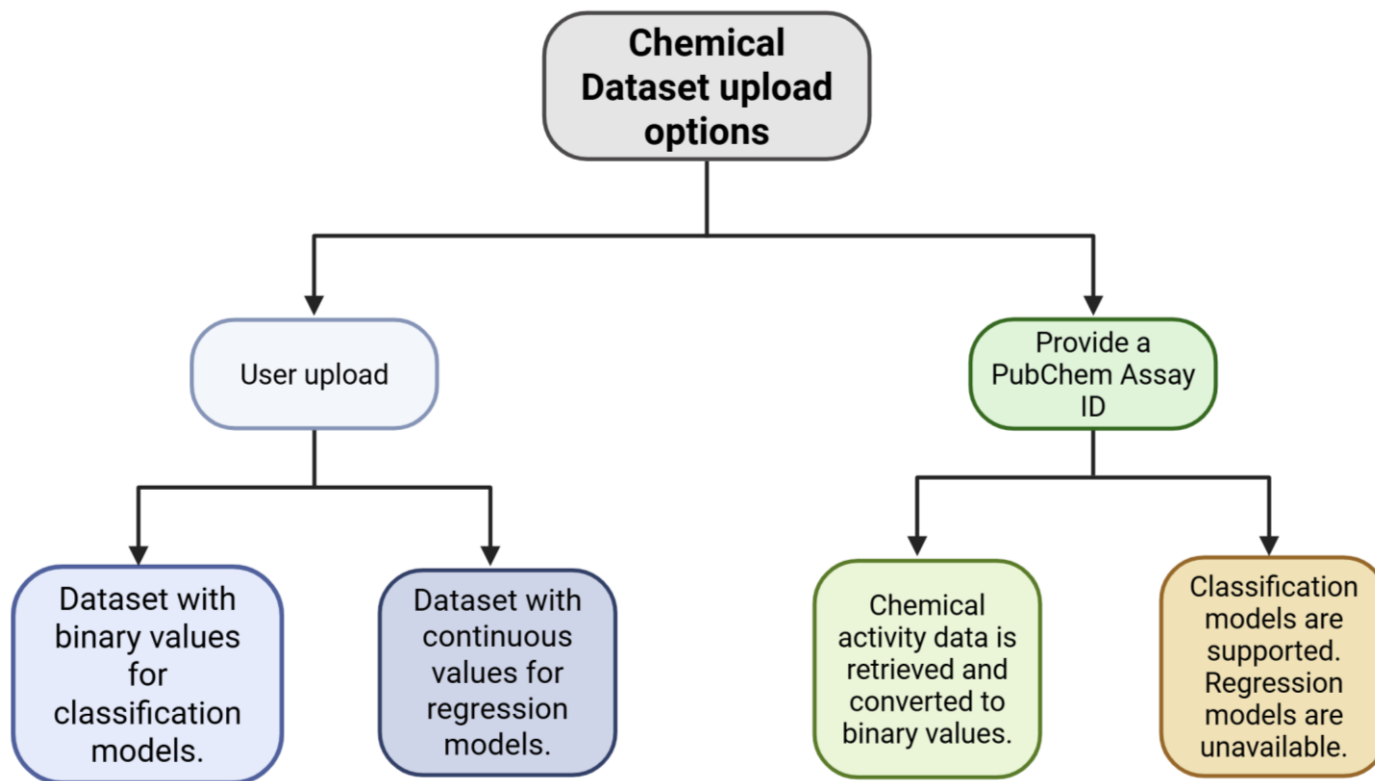
## First few rows of the Resources Table

### Curated Datasets

Dataset Name	Total # of Compounds	Data Type	Dataset Source	Dataset Description
BBB (Blood Brain Barrier)	438	logBB	Wang et al.	Compounds with experimental logBB values was compiled and curated using ChemAxon and CASE Ultra tools.
BCRP (Breast Cancer Resistance Protein)	395	$\mu$ M (evidence of inhibition at 10 $\mu$ M)	Sedykh et al. Zhao et al.	The BCRP dataset was curated for experimental consistency and structural quality, and filtered to include only reliable binary classification labels for substrates and inhibitors.
Bioavailability	1159	oral bioavailability (%F)	Kim et al.	Compiled across public and literature sources. Chemical structures were standardized, and %F values were harmonized to resolve discrepancies.
BSEP (Bile Salt Export Pump)	725	$\mu$ M (evidence of inhibition at 100 $\mu$ M)	Zhao et al.	Collected from publicly available experimental data. Structures were curated and standardized to ensure consistency and dataset includes binary labels.
Cancer (Human Oral Carcinogenicity)	342	Binary, 0=Non-Carcinogen; 1=Carcinogen	Chung et al.	342 unique organic compounds from the EPA's IRIS database, labeled as carcinogenic or noncarcinogenic based on oral slope factor (OSF), a quantitative measure for oral cancer risk.
Cosmetics	4129	---	Chung et al.	Cosmetic dataset collected from COSMOS Cosmetics Inventory knowledge base.
DART (Developmental and Reproductive Toxicity)	1452	Oral Developmental, Inhalation Maternal, ToxRefDB Maternal	Ciallella et al.	Collected from U.S. EPA's in vivo prenatal developmental toxicity studies in rats and rabbits based on oral or inhalation studies.
Drugbank	8055	---	Chung et al.	Collected from DrugBank database.

# Cheminformatics

## Upload or Retrieve datasets



### 5. Upload or Retrieve datasets

Data can be in the Comma-Separated Values (CSV) or Structure Data Format (SDF) format to upload. Sample files provided. You can upload or retrieve up to 1000 chemicals. Instead of uploading a dataset, you may also import structure-activity information from PubChem by entering the PubChem Assay Identifier (AID).

1. Upload a dataset (eg: sample\_dataset.csv).

2. eg: cid.

3. eg: Activity\_Binary.

4. eg: SMILES.

5. Choose dataset type and click on Upload dataset.

### Upload a dataset.

Please choose a file type to upload dataset. Dataset should be in an CSV file. For CSV file, it must contain a column named "SMILES" contains SMILES in for each record.

Choose file format:

☒ CSV [Download Sample CSV](#)

☐ SDF [Download Sample SDF](#)

[Choose File](#) No file chosen

Compound id (has to be a property in the SDF file or a column in CSV file):

Activity name (has to be a property in the SDF or a column in CSV file):

SMILES column name (Only for CSV file):

Select Dataset Type:

☒ Binary  
☐ Continuous

[Upload dataset](#)

Please use the Sample buttons to download them.

\*Accepts up to 1000 chemicals.



A sample file for Upload Dataset looks like this.

A	B	C	D
cid	SMILES	Activity_Binary	Activity_Reg
11	<chem>C(CCl)Cl</chem>	1	0.86
13	<chem>C1=CC(=C(C=C1Cl)Cl)Cl</chem>	0	0.31
174	<chem>C(CO)O</chem>	0	0.5
180	<chem>CC(=O)C</chem>	0	-0.02
240	<chem>C1=CC=C(C=C1)C=O</chem>	0	-1.11
241	<chem>C1=CC=CC=C1</chem>	1	-1.11
243	<chem>C1=CC=C(C=C1)C(=O)O</chem>	0	0.69
263	<chem>CCCCO</chem>	0	-0.22
299	<chem>C1(=O)C2(C3(C4(C1(C5(C2(C3(C(C45Cl)(Cl)Cl)Cl)Cl)Cl)Cl)Cl)Cl)Cl</chem>	0	0.32
335	<chem>CC1=CC=CC=C1O</chem>	0	1.4
342	<chem>CC1=CC(=CC=C1)O</chem>	0	-0.16
712	<chem>C=O</chem>	0	0.02
727	<chem>C1(C(C(C(C1Cl)Cl)Cl)Cl)Cl</chem>	1	0.38
887	<chem>CO</chem>	0	0.03
931	<chem>C1=CC=C2C=CC=CC2=C1</chem>	0	-0.15
949	<chem>CN(C)C1=CC=CC=C1</chem>	0	0.36
992	<chem>C1(=C(C(=C(C=C1Cl)Cl)Cl)Cl)O</chem>	1	-0.14
996	<chem>C1=CC=C(C=C1)O</chem>	0	0.95
1049	<chem>C1=CC=NC=C1</chem>	0	0.48
1140	<chem>CC1=CC=CC=C1</chem>	0	-0.05
1480	<chem>C1=C(C(=CC(=C1Cl)Cl)Cl)OCC(=O)O</chem>	0	-0.05
1486	<chem>C1=CC(=C(C=C1Cl)Cl)OCC(=O)O</chem>	0	-0.12
1489	<chem>C1=CC(=C(C=C1Cl)Cl)OCCCC(=O)O</chem>	0	-0.15
1493	<chem>C1=CC(=C(C=C1[N+](=O)[O-])[N+](=O)[O-])O</chem>	0	-0.37
1982	<chem>CC(=O)NP(=O)(OC)SC</chem>	0	-0.84

## Import a PubChem Bioassay.

Instead of uploading a dataset, you may also import structure-activity information from PubChem by entering the PubChem Assay Identifier (AID) below.

Enter PubChem AID (eg: 1259248):

Import dataset

1. Enter an AID and click to retrieve the dataset

\*Retrieves up to 1000 random chemicals. 500 actives and 500 inactives.

Please check in Tasks option if the job is Complete.

## Tasks

Description	Status
Importing structure-activity data for AID 1000: Bioassay 1000	<a href="#">Complete</a>

# Selected dataset display once uploaded.

## Upload a dataset.

Please choose a file type to upload a new dataset. Dataset should be in an CSV or SDF file. For CSV file, it must contain a column named "SMILES" contains SMILES information for each record.

Choose file format:

☒ CSV [Download Sample CSV](#)

☐ SDF [Download Sample SDF](#)

[Choose File](#) No file chosen

Compound ID (has to be a property in the SDF file or a column in CSV file):

Activity name (has to be a property in the SDF or a column in CSV file):

SMILES column name (Only for CSV file):

Select  
Dataset  
Type:

☒ Binary  
☐ Continuous

[Upload dataset](#)

Select dataset:

sample\_dataset\_Binary

Show 10 entries

Search:

Chemical	Activity	Structure
TOX-1823	1	
TOX-1678	1	
TOX-1695	1	
TOX-1694	1	
TOX-1693	1	
TOX-1692	1	
TOX-1691	1	

Showing 1 to 10 of 1,000 entries

[Previous](#)

1

2

3

4

5

...

100

[Next](#)

[Remove dataset](#)

[Download dataset as CSV file](#)

20

Download or remove  
the selected dataset

# Cheminformatics



Curator

## 6. Curator

Cleans the chemical structures and prepares them for next steps such as model generation by the following steps:

- i. Check and clean chemical structures.
- ii. Standardize chemical structure representation (e.g, updating valencies, removing charges, etc.).
- iii. Strip salts and remove mixtures by keeping the largest organic component.
- iv. Merge or remove duplicated structures.

Please check in Tasks if the job is Finished.

## Chemical curator

Proper chemical curation is a crucial step in Quantitative Structure-Activity Relationship development. This module cleans chemical structures and prepares them by using the following steps:

1. Check and clean chemical structures.
2. Standardize chemical structure representation (e.g, updating valencies, removing charges, etc.).
3. Strip salts and remove mixtures by keeping the largest organic component.
4. Merge or remove duplicated structures.

### 1. Select your dataset.

Select Dataset:

sample\_dataset\_Binary

Select how to handle duplicates:

- ☒ Merge and take **highest** activity  
☐ Merge and take **lowest** activity  
☐ Remove all

Replace dataset or create new:

- ☒ Replace dataset  
☐ Create new dataset

Curate

### 2. Click to run the curation.

Select Dataset:

sample\_dataset\_Binary

sample\_dataset\_Binary

sample\_dataset\_Binary\_curated

Select how to handle duplicates:

- ☒ Merge and take **highest** activity  
☐ Merge and take **lowest** activity  
☐ Remove all

Replace dataset or create new:

- ☒ Replace dataset  
☐ Create new dataset

Curate

A file named sample\_dataset\_Binary\_curated is generated if 'create new dataset' option was chosen.

Please check in Tasks option if the job is Complete.

## Tasks

Description	Status
Curating sample_dataset_Binary chemicals	<a href="#">Complete</a>

# Cheminformatics



# Principal Component Analysis

## **7. Principal Component Analysis**

Principal Component Analysis (PCA) is a dimension reduction technique helps visualizing chemical space.



# Principal Component Analysis

Principal Component Analysis (PCA) is a dimension reduction technique useful for visualizing chemical space. Select a dataset and click "Perform PCA" to visualize its chemical space. Chemicals will be colored according to their assigned activity (active, 1: red; inactive, 0: blue). If the Activity values are continuous, a continuous color scale is used.

Select dataset:

sample\_dataset\_Binary

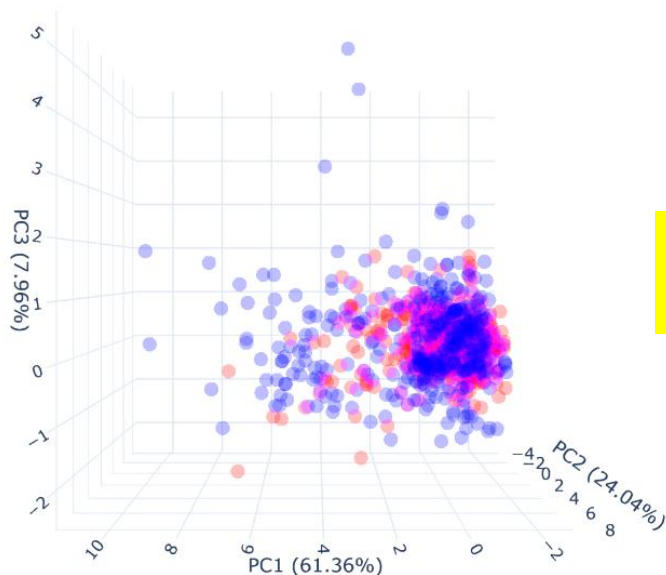
1. Select your dataset.

Perform PCA

2. Click to generate PCA plot.



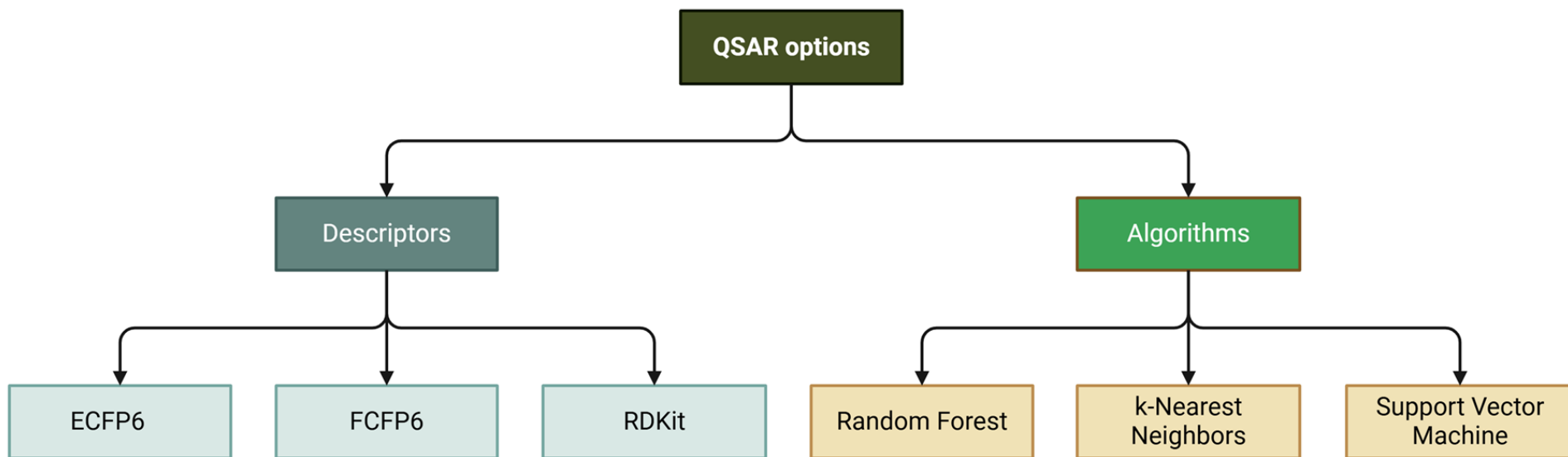
Activity  
● 1.0  
● 0.0



PCA figure is generated like shown with many options including save and zoom.

# Cheminformatics

## ↳ QSAR Builder



### 8. QSAR Builder

Create QSAR models using a variety of descriptors and algorithms with either a user-uploaded dataset or a dataset obtained by providing a PubChem Assay ID.

# QSAR Builder

Quantitative Structure-Activity Relationship (QSAR) models are statistical models relating chemical structures to observed biological activities. A model is defined in this tab as a pair of features (e.g., ECFP6 fingerprints) and a machine learning algorithm (e.g., Random Forest). A QSAR model can be built by selecting a dataset, feature set, and machine learning algorithm and pressing the "Build QSAR" button. Depending on numerous factors, modeling training can take a while.

1. Choose your dataset

Select Dataset:

sample\_dataset\_Binary

2. Choose Features, Algorithms and Activity type to build QSAR models. Multiple options can be chosen

Select Features:

- ☒ ECFP6
- ☐ FCFP6
- ☐ RDKit Descriptors

Select Algorithm:

- ☒ Random Forest (RF)
- ☐ k-Nearest Neighbors (kNN)
- ☐ Support Vector Machine (SVM)

Select Activity:

- ☒ Classification
- ☐ Regression

Build QSAR

## Trained Models

3. Check the evaluation metrics of your model.

Model	Type	Classification-Accuracy	Classification-F1 Score	Classification-Precision	Classification-Recall	Classification-Specificity	Regression-R2 Score	Regression-Max Error	Regression-Mean Squared Error
sample_dataset_Binary-ECFP6-RF-Classification	Classification	0.72	0.75	0.71	0.81	0.63	None	None	None

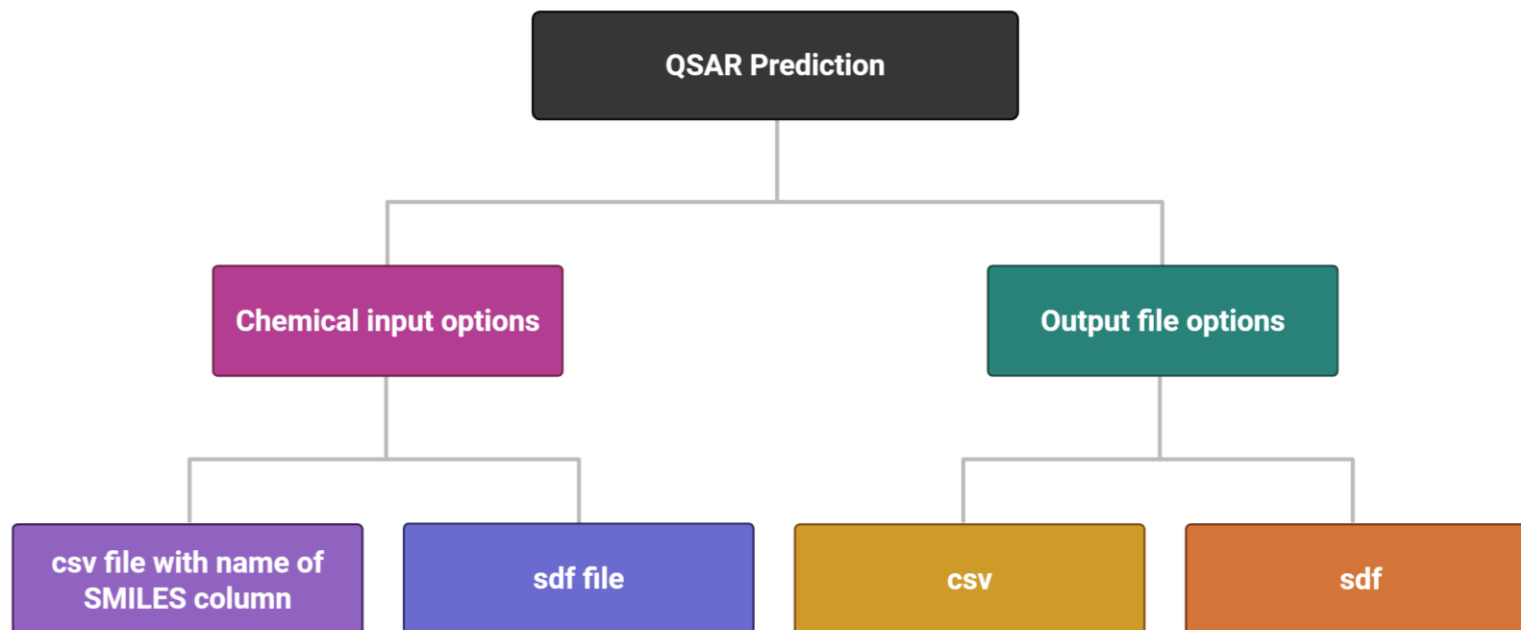
Please check in Tasks option if the job is Complete.

## Tasks

Description	Status
Building QSAR Model on sample_dataset_Binary&ECFP6&RF&Classification	<a href="#">Complete</a>
Curating sample_dataset_Binary chemicals	<a href="#">Complete</a>

# Cheminformatics

## ↳ QSAR Predictor



### 9. QSAR Predict

Predict toxicity for new chemicals using user-developed models.

# QSAR Predictor

In this tab, you can select one or more previously built QSAR models to predict the activity of a set of 100 chemicals. After prediction, a modified CSV or SDF file containing the results will be automatically downloaded to your computer.

\*Accepts up to 100 chemicals.

Select QSAR model:

1. Select QSAR your QSAR model.

Select Output File Format:

- ☒ CSV  
☐ SDF

\*Multiple QSAR models can be selected.

2. Select an output format.

Select CSV or SDF File:

No file chosen

SMILES column name in the CSV file:

3. Choose a file eg: Prediction\_dataset.csv and enter eg: SMILES, then click on Predict.

An output file eg: Prediction\_dataset\_predicted.csv is downloaded with a new column eg: sample\_dataset\_Binary-ECFP6-RF-Classification\_Prediction contains predicted scores for the chemicals.

## 10. Contact us

Rowan University: 201 Mullica Hill Rd, Robinson Hall, Glassboro, NJ 08028

Tulane University: Hutchinson Memorial Building (School of Medicine), 1415 Tulane Ave, New Orleans, LA 70112

Questions, comments, and general inquiries can be emailed to [toxiverse.help@gmail.com](mailto:toxiverse.help@gmail.com).

## 11. About us

The Zhu Lab uses cheminformatics algorithms, workflows, and other computational tools to model chemical toxicity, ADME (Absorption, Distribution, Metabolism, and Excretion), and other biological activities. These models support regulatory chemical toxicity assessments and the computer-aided drug discovery (CADD) process.



TULANE UNIVERSITY  
SCHOOL *of* MEDICINE