

# TOXRIC: a comprehensive database of toxicological data and benchmarks

Lanlian Wu<sup>1,2,†</sup>, Bowei Yan<sup>1,3,†</sup>, Junshan Han<sup>1</sup>, Ruijiang Li<sup>1</sup>, Jian Xiao<sup>4,5</sup>, Song He<sup>1,\*</sup> and Xiaochen Bo<sup>1,2,\*</sup>

<sup>1</sup>Department of Bioinformatics, Institute of Health Service and Transfusion Medicine, Beijing, China.

<sup>2</sup>Academy of Medical Engineering and Translational Medicine, Tianjin University, Tianjin, China.

<sup>3</sup>State Key Laboratory of Genetic Engineering, Institutes of Biomedical Sciences, Fudan University, Shanghai, China.

<sup>4</sup>Department of Pharmacy, Xiangya Hospital, Central South University, Changsha, China.

<sup>5</sup>Institute for Rational and Safe Medication Practices, National Clinical Research Center for Geriatric Disorders, Xiangya Hospital, Central South University, Changsha, China.

---

\*To whom correspondence should be addressed (Xiaochen Bo). Tel: +86 010 66931207; Email: boxc@bmi.ac.cn,boxiaoc@163.com.

\*Correspondence may also be addressed to Song He. Tel: +8601066931450; Email: hes1224@163.com

†These authors contributed equally to this work.

© The Author(s)

This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (<http://creativecommons.org/licenses/by-nc/2.0/uk/>) which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

## Supplementary Table

Supplementary Table 1 The units in endpoints of typical species in TOXRIC.

Species	Endpoint	Administration	Unit
human	TDLo、LDLo	oral, skin, intravenous	mg/kg, gm/kg, ug/kg, ng/kg, uL/kg, mL/kg
rat	LD50、LDLo	oral	mg/kg, gm/kg, ug/kg, uL/kg, mL/kg
mouse	LD50、LDLo	oral	mg/kg, gm/kg, ng/kg, uL/kg, mL/kg
rabbit	LD50、LDLo	oral	mg/kg, gm/kg, ug/kg, ng/kg, mL/kg
dog	LD50、LDLo	oral	mg/kg, gm/kg, ug/kg, uL/kg, mL/kg
cat	LD50、LDLo	oral	mg/kg, gm/kg, ug/kg
guinea pig	LD50、LDLo	oral	mg/kg, gm/kg, ug/kg, ng/kg

## Supplementary Data

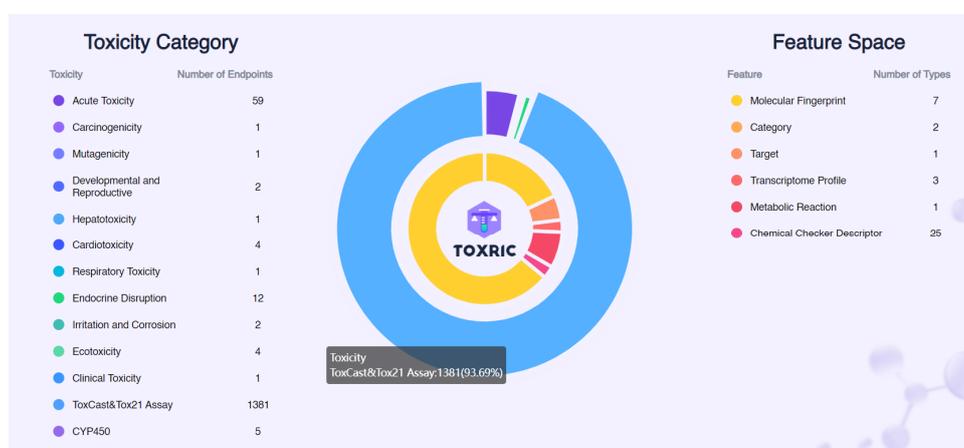
### STEP-BY-STEP TUTORIAL OF TOXRIC WEBSITE (<https://toxric.bioinforai.tech/>)

#### Data browsing.

##### 1. Browse dataset composition on Home page

TOXRIC provides 13 toxicity datasets and 6 feature datasets. Each dataset includes multiple sub-datasets of toxicity endpoints and feature types.

On the Home page, the number of sub-datasets of both toxicity and feature datasets is displayed in the form of two-layer concentric circles. The outer layer and inner layer represent toxicity and feature datasets respectively. When clicking on the dataset field, users will be linked to the corresponding dataset description on the Data Collection page.



##### 2. Browse information of toxicity datasets on Data Collection page

(1) Enter the Toxicity Dataset page on Data Collection page, or click the Toxicity Dataset field in the page, the descriptions of all toxicity datasets will be displayed, including dataset description, number of compounds, and sources. Click the Details button on the far right of each row to view the detailed information.

**TOXRIC**  
Toxicity Research & Benchmarking Consortium

Home Data Collection Search Benchmark & Representation Statistics Download Contribute Contact & About

**Toxicity Category**

- Toxic Effect
  - Acute Toxicity
  - Carcinogenicity
  - Mutagenicity
  - Developmental and Reproductive Toxicity
- Target Organ Toxicology
  - Hepatotoxicity
  - Cardiotoxicity
  - Respiratory Toxicity
  - Endocrine Disruption
  - Irritation and Corrosion
- Applied Toxicology
  - Ecotoxicity
  - Clinical Toxicity
- Other Toxicology Datasets
  - ToxCast&Tox21 Assay
  - CYP450

**Toxicity Category**

Toxicity Category	Description	Number of compounds	Sources	Details
Acute Toxicity	This dataset lists the acute systemic toxicity outcome records (e.g., lethal dose, 50% or LD50) in different species and multiple routes of administration. The data is collected from the study by Jain et al., which included data obtained from the ChemIDplus database.	59	Scientific Literature, ChemIDplus database	Details
Carcinogenicity	The data is collected and revised from the Carcinogenic Potency Database (CPDB) summary tables (CPDBAS, version 5d), which is a unique and standardized resource of long-term animal carcinogenesis study results on more than 1500 chemical substances.	1	Scientific Literature	Details
Mutagenicity	Mutagenicity refers to the induction of permanent transmissible changes in the amount or structure of the genetic material of cells or organisms. These changes may involve a single gene or gene segment, a block of genes or chromosomes. The data contains 7485 compounds associated with known mutagenicity.	1	Scientific Literature	Details

(2) The interactive filter located on the left side of the Toxicity Dataset page allows users to explore the endpoint sub-datasets. The information of sub-datasets includes dataset description, number of compounds and sources. Click the Details button on the far right of each row to view the detailed information and all the compounds included in the sub-dataset.

**TOXRIC** Home Data Collection Search Benchmark & Representation Statistics Download Contribute Contact & About

**Toxicity Category**

- Toxic Effect
  - Acute Toxicity**
  - Carcinogenicity
  - Mutagenicity
  - Developmental and Reproductive Toxicity
- Target Organ Toxicology
  - Hepatotoxicity
  - Cardiotoxicity
  - Respiratory Toxicity
  - Endocrine Disruption
  - Irritation and Corrosion

**Toxicity Category** Search

Endpoint	Description	Number of compounds	Sources	Details
mouse_intraperitoneal_LD50	This file contains the Lethal Dose Fifty (LD50) values for mice via the intraperitoneal route.	35299	Scientific Literature	
mammal (species unspecified)_intraperitoneal_LD50	This file contains the Lethal Dose Fifty (LD50) values for mammals (species unspecified) via the intraperitoneal route.	537	Scientific Literature	
guinea pig_intraperitoneal_LD50	This file contains the Lethal Dose Fifty (LD50) values for guinea pigs via the intraperitoneal route.	237	Scientific Literature	
rat_intraperitoneal_LD50	This file contains the Lethal Dose Fifty (LD50) values for rats via the intraperitoneal route.	4798	Scientific Literature	

On the detail information page, all compounds contained in an endpoint sub-dataset are listed in the form of a molecular graph. Clicking on a specific compound will open the compound information page that displays affluent chemical, toxicological, and feature data of the compound.

**TOXRIC** Home Data Collection Search Benchmark & Representation Statistics Download Contribute Contact & About

**Endpoint**

Endpoint: **mouse\_intraperitoneal\_LD50**

Toxicity category: **Acute Toxicity**

Description: This file contains the Lethal Dose Fifty (LD50) values for mice via the intraperitoneal route.

Number of endpoints: **35299**

Sources: 1. Jain S, Siramshetty VB, Alves VM, Muratov EN, Kleinreuter N, Tropsha A, Nicklaus MC, Simeonov A, Zakharov AV. Large-Scale Modeling of Multispecies Acute Toxicity End Points Using Consensus of Multitask Deep Learning Methods. *J Chem Inf Model.* 2021;61(2):653-663. 2. Liwanag PM, Hudson VW, Hazard GF Jr. ChemIDplus: A Web-Based Chemical Search System. *NLM Tech Bull.* 2000; (313):e3.

Benchmark task: **Regression**

 <b>TOX-1273</b> Name: <b>Indane</b>	 <b>TOX-1279</b> Name: <b>metformin</b>	 <b>TOX-1282</b> Name: <b>CARVACROL</b>	 <b>TOX-1289</b> Name: <b>PAEONOL</b>	 <b>TOX-1292</b> Name: <b>Glucoxy</b>	 <b>TOX-1293</b> Name: <b>Esculetin</b>
 <b>TOX-1298</b> Name: <b>(E,Z)-ferulic acid</b>	 <b>TOX-1307</b> Name: <b>No data</b>	 <b>TOX-1311</b> Name: <b>Melatonin</b>	 <b>TOX-1318</b> Name: <b>daidzein</b>	 <b>TOX-145</b> Name: <b>hydroquinone</b>	 <b>TOX-245</b> Name: <b>thiotepa</b>

ALL 35299 items

1 2 3 4 5 6 ... 2942 12/page Go to 1

### 3. Browse information of feature datasets on Data Collection page

The interactive filter located on the left side of the Feature Dataset page allows users to explore the feature type sub-datasets. The information of sub-datasets includes dataset description, feature dimension, number of compounds and sources. Click the Details button on the far right of each row to view the detailed information and all the compounds included in the sub-dataset.

Type	Description	Feature dimension	Number of compounds	Sources	Details
ECFP2	ECFP2 is a 2048-length bits vector, which represent the neighbor hood environment of each atom using the extended connectivity fingerprint encoding a circular substructure of diameter 2 bonds. An individual bit has no definite meaning.	2048	110000	Rdkit	
ECFP4	ECFP4 is a 2048-length bits vector, which represent the neighbor hood environment of each atom using the extended connectivity fingerprint encoding a circular substructure of diameter 4 bonds. An individual bit has no definite meaning.	2048	110000	Rdkit	
ECFP6	ECFP6 is a 2048-length bits vector, which represent the neighbor hood environment of each atom using the extended connectivity fingerprint encoding a circular substructure of diameter 6 bonds. An individual bit has no definite meaning.	2048	110001	Rdkit	

On the detail information page, all compounds contained in a feature type sub-dataset are listed in the form of a molecular graph. Clicking on a specific compound will open the compound information page that displays affluent chemical, toxicological, and feature data of the compound.

**Feature Space for compounds**

Type: ECFP2  
 Feature: Molecular Fingerprints  
 Description: ECFP2 is a 2048-length bits vector, which represent the neighbor hood environment of each atom using the extended connectivity fingerprint encoding a circular substructure of diameter 2 bonds. An individual bit has no definite meaning.  
 Feature dimension: 2048  
 Number of compounds: 110000  
 Sources: Rdkit

 TOX-11 Name: No data	 TOX-12 Name: L-ascorbate	 TOX-2 Name: nitrate	 TOX-1 Name: glucarate	 TOX-10 Name: No data	 TOX-3 Name: Nitroxyl
 TOX-4 Name: Fluorophosphate	 TOX-5 Name: No data	 TOX-6 Name: Thiosulphate	 TOX-7 Name: Bortezomib	 TOX-8 Name: Ixazomib	 TOX-9 Name: ethinamate

ALL 110000 items

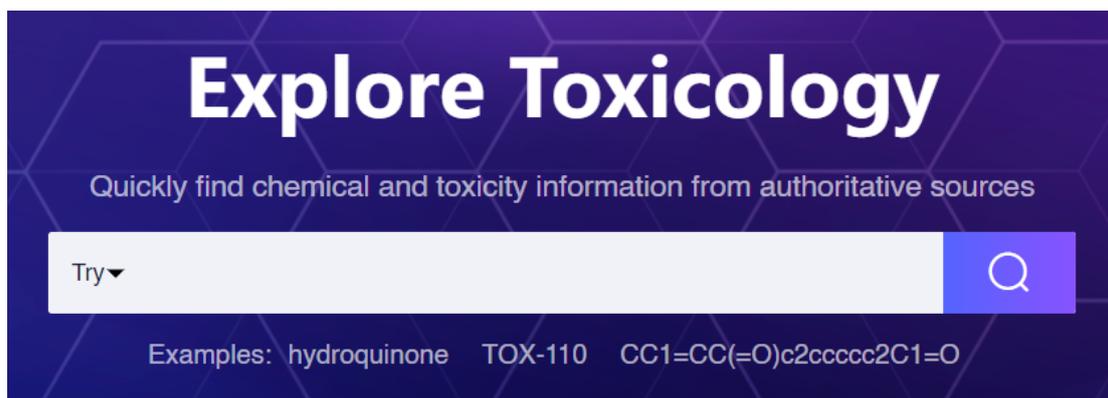
1 2 3 4 5 6 ... 9167 > 12/page Go to 1

## Data retrieval.

### 1. Keyword search for a compound

(1) Enter a keyword in the search box of the Home page. The search box accepts both complete or partial keywords of TAID, name, IUPAC name, PubChem CID, SMILES, InChIKey, and InChI identifiers. Fuzzy search is allowed.

There are three examples below the search box, users can click the example keywords to view the search results.



(2) After clicking the search button or pressing the Enter key, a list of compound entities will be provided. The query keyword is highlighted in red.

**COMPOUND BEST MATCH**

**TOX-145**  
 TAID: TOX-145  
 Name: **hydroquinone**  
 IUPAC Name: benzene-1,4-diol  
 PubChem CID: 785  
 Canonical SMILES: Oc1ccc(O)cc1  
 InChIKey: QIGBRXMKCJJKVMJ-UHFFFAOYSA-N  
 InChI: InChI=1S/C6H6O2/c7-5-1-2-6(8)/4-3-5/h1-4,7-8H  
 Molecular Formula: C6H6O2  
 Molecular Weight: 110.11  
 XLogP: 0.6

**COMPOUND MATCH**

**TOX-108752**  
 TAID: TOX-108752  
 Name: **Durohydroquinone**  
 IUPAC Name:  
 PubChem CID: 136346  
 Canonical SMILES: Cc1c(C)c(O)c(C)c(O)c1O  
 InChIKey: SUNVJLYDZCIK-UHFFFAOYSA-N  
 InChI: InChI=1S/C10H14O2/c1-5-6(2)10(12)(8)(4)7(3)(9)(5)11/h11-12H,1-4H3  
 Molecular Formula: C10H14O2  
 Molecular Weight: 166.22  
 XLogP: 1.8

## 2. Batch search for a list of compounds

(1) Select the type of identifier to search for a list of compounds on the Search page. For the identifiers, IUPAC name, PubChem CID, SMILES, InChIKey and InChI are allowed.

(2) You can enter a compound list to query the information of the compounds.

**Batch search for compounds**

Select an identifier:  
 SMILES

Input a compound list:

```

O1 [C@H] ([C@H] (O) CO) C ([O-])=C (O) C1=O
SC [C@H] ([NH3+]) C (=O) [O-]
O1 [C@] (O) (CO) [C@H] (O) [C@H] (O) [C@H] 2O [C@H] (CO) [C@H] (O) [C@H] (O)
[C@H] 2O [C@H] 1CO
  
```

Or upload a Excel Sheet or TXT file:

Drag and drop files to upload

Search

**Tips:**  
 Select an identifier and input a compound list to query the chemical information, toxicity categories and feature domain of the compounds. Or upload an EXCEL or TXT file (separated by commas), containing the list of queried compounds. For the identifiers, TAID, name, IUPAC name, PubChem CID, SMILES, InChIKey and InChI are allowed. Try these examples: TOX-1, sulfamer, 56959, O=C(O)CCC(=O)O, ZRALSGWFCBTJQ-UHFFFAOYSA-O, or InChI=1S/C3H5N3O9/c7-4(8)13-1-3(15-6(11)12)2-14-5(9)10/h3H,1-2H2.

(3) Or you can upload an EXCEL or TXT file to query the information of the compounds. The template file can be downloaded by clicking the button to the right of the upload box.

### 3. Browse compound information on Compound page

The Compound page consists of three sections, i.e., chemical information, toxicity category, and feature space.

#### (1) Chemical information

The chemical information section provides seven commonly used identifier types and physicochemical properties of compounds. Click on the PubChem CID to link to the Compound page of PubChem website.

**TOX-145**  
 TAID: TOX-145  
 Name: hydroquinone  
 IUPAC Name: benzene-1,4-diol  
 PubChem CID: 785  
 Canonical SMILES: Oc1ccc(O)cc1  
 InChIKey: QIGBRXMKCJVKMJ-UHFFFAOYSA-N  
 InChI: InChI=1S/C6H6O2/c7-5-1-2-6(8)4-3-5/h1-4,7-8H  
 Molecular Formula: C6H6O2  
 Molecular Weight: 110.11  
 XLogP: 0.6

Acute Toxicity | Endocrine Disruption | CYP450 | Ecotoxicity | Developmental and Reproductive Toxicity | Carcinogenicity | Clinical Toxicity | Hepatotoxicity | Target

Category | Metabolic Reaction | Molecular Fingerprint | Chemical Checker Descriptor

#### (2) Toxicity category

Click on the toxicity categories in the title bar to view the toxicity values for each endpoint. If no toxicity value is currently collected for the compound, it will be shown as “Data is not available now”.

Clicking the download button in the upper-right corner will download the toxicity values for all endpoints under the selected toxicity category in .csv format.

**Toxicity Category** Download

[ToxCast&Tox21 Assay](#)
[Acute Toxicity](#)
[Endocrine Disruption](#)
[CYP450](#)
[Cardiotoxicity](#)
[Ecotoxicity](#)
[Developmental and Reproductive Toxicity](#)
[Irritation and Corrosion](#)

Endpoint	Toxicity value	Source
CYP1A2	0	Wu Z, Jiang D, Wang J, Hsieh CY, Cao D, Hou T. Mining Toxicity Information from Large Amounts of Toxicity Data. J Med Chem. 2021;64(10):6924-6936.
CYP2C19	0	Wu Z, Jiang D, Wang J, Hsieh CY, Cao D, Hou T. Mining Toxicity Information from Large Amounts of Toxicity Data. J Med Chem. 2021;64(10):6924-6936.
CYP2C9	0	Wu Z, Jiang D, Wang J, Hsieh CY, Cao D, Hou T. Mining Toxicity Information from Large Amounts of Toxicity Data. J Med Chem. 2021;64(10):6924-6936.
CYP2D6	0	Wu Z, Jiang D, Wang J, Hsieh CY, Cao D, Hou T. Mining Toxicity Information from Large Amounts of Toxicity Data. J Med Chem. 2021;64(10):6924-6936.

ALL 4 items 10/page Go to 1

### (3) Feature space

Click on the feature spaces in the title bar to view the features for each feature type. The targets, categories, and metabolic reactions of compounds are listed in text format to be queried, while the feature vectors of transcriptome profiles, molecular fingerprints, and CC descriptors should be downloaded to view because the length of the vectors is too long to display. Clicking the download button in the upper-right corner to download.

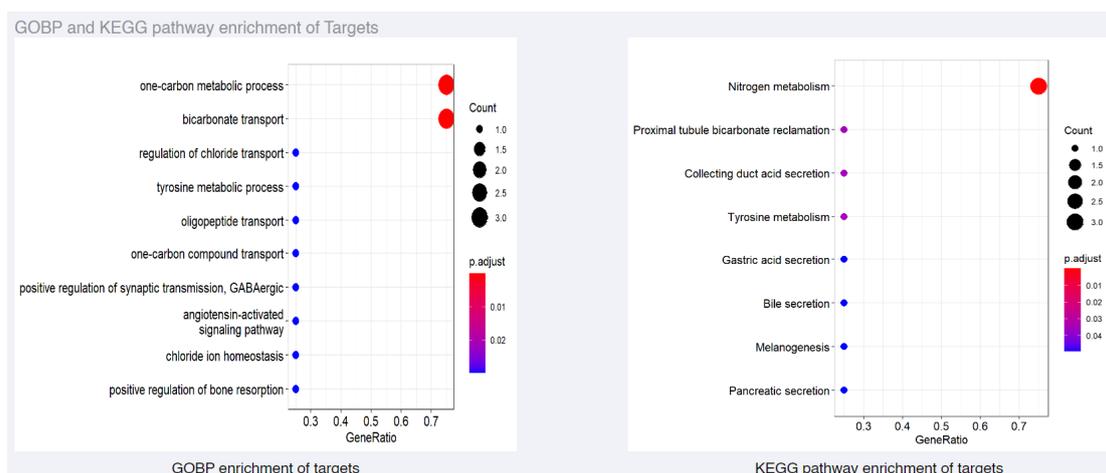
**Feature Space** Download

[Target](#)
[Category](#)
[Metabolic Reaction](#)
[Transcriptome Profile](#)
[Molecular Fingerprint](#)
[Chemical Checker Descriptor](#)

Feature type	Source	Details
Target	DrugBank The Binding Database	P14679
Target	DrugBank The Binding Database	O42275
Target	DrugBank The Binding Database	O42713
Target	DrugBank The Binding Database	O43570
Target	DrugBank The Binding Database	P00918
Target	DrugBank The Binding Database	P07451

ALL 6 items 10/page Go to 1

Below the feature list, the top 10 GOBP (Gene Ontology Biological Process) and KEGG pathway enrichment results of compounds' target proteins are displayed in a bubble plot.

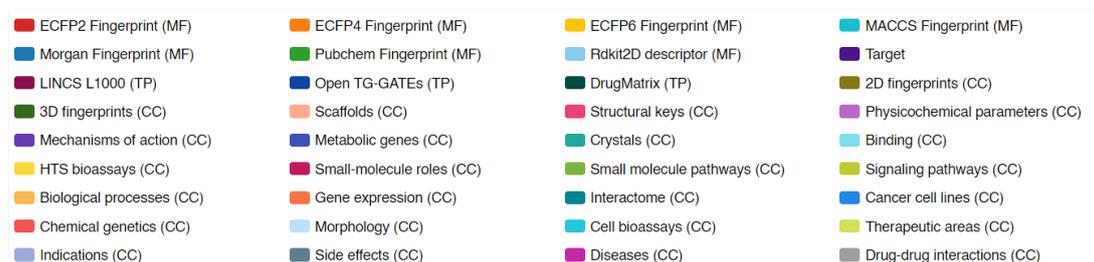


## View Benchmark and representation distribution

### 1. View benchmarks for feature types

On Benchmarks for Feature Types page, the bar charts show the predictive effect of 36 feature types on all toxicity endpoints.

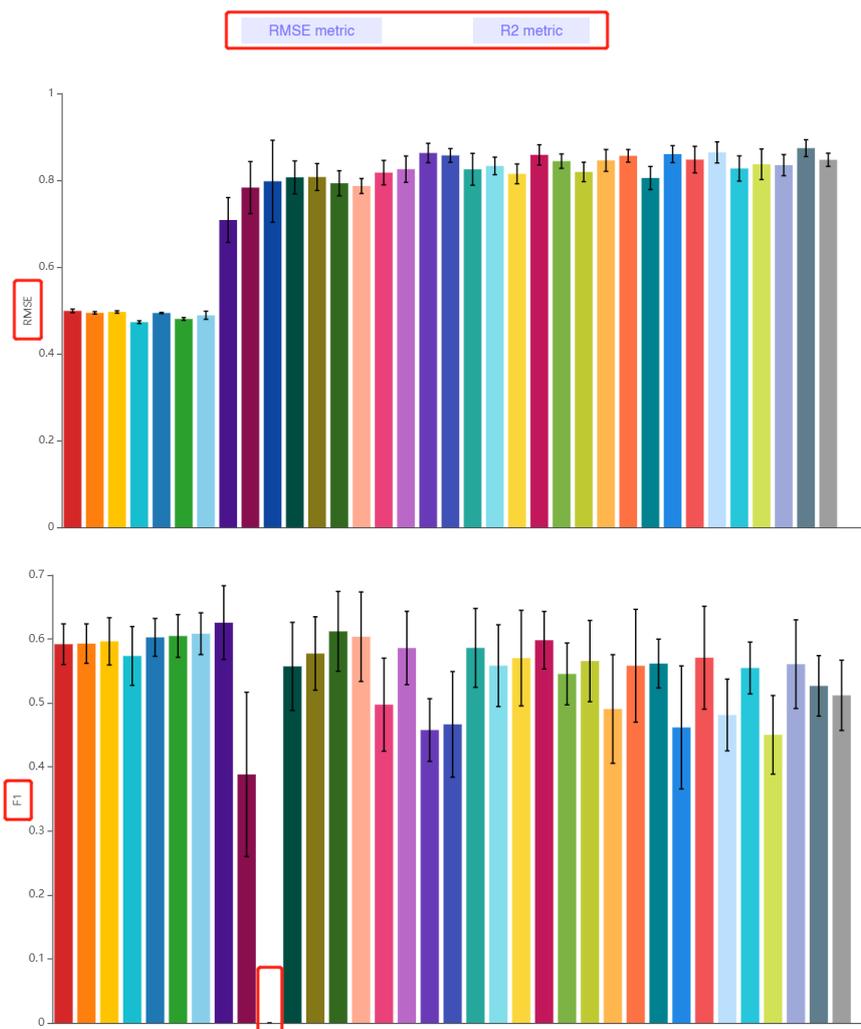
In the legend, the parentheses after each feature type indicate the feature space it belongs to. MF represents Molecular Fingerprint, TP represents Transcriptome Profile, and CC represents Chemical Checker Descriptor.



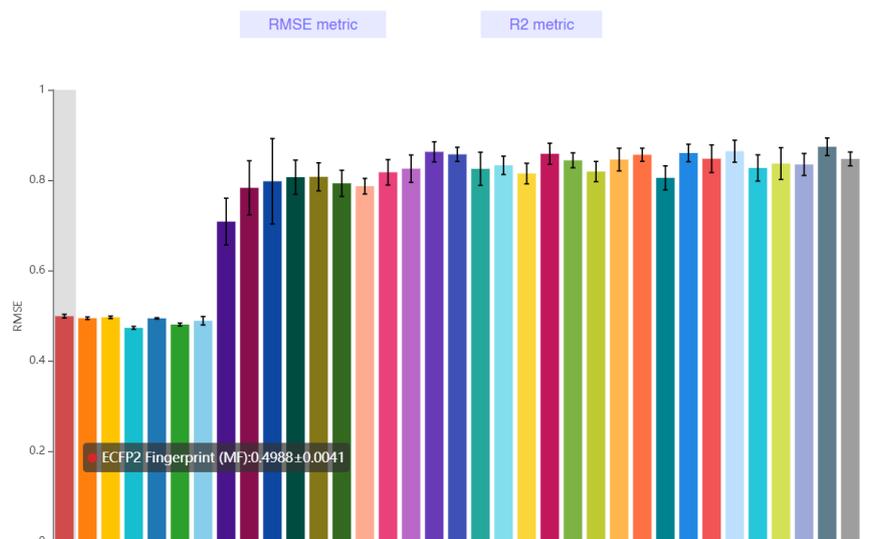
Note: MF: Molecular Fingerprint; TP: Transcriptome Profile; CC: Chemical Checker Descriptor

Three metrics are used to evaluate the performance. The classification datasets use F1 metric, while the regression datasets (Acute Toxicity, Ecotoxicity) use RMSE and R2 metrics. For the regression datasets, click on RMSE or R2 button to view performance results.

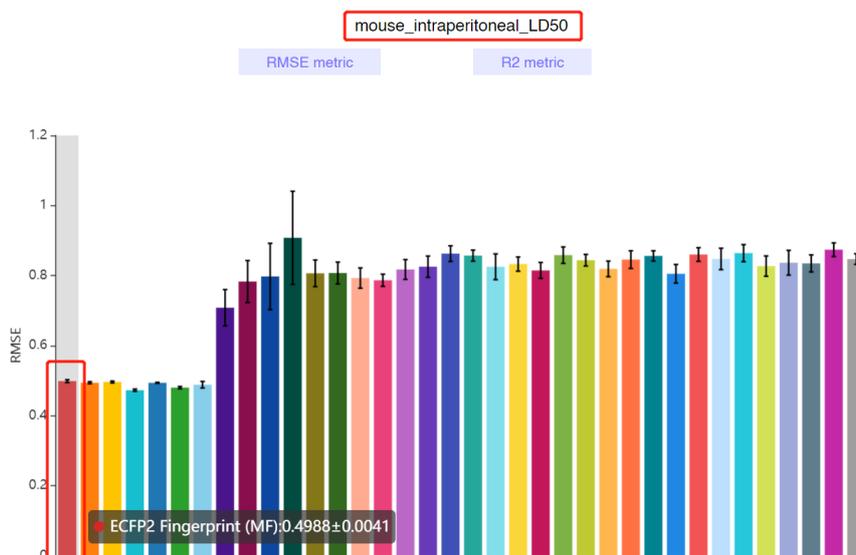
For RMSE metric, lower value represents the higher prediction performance. While for R2 and F1 metric, the higher the value, the higher the prediction performance. It should be noted that if the value of metric is 0, it represents the number of samples with the feature type at the endpoint sub-dataset is less than 10 and no benchmark experiment is performed.



When the mouse is suspended on the bar, the mean and standard deviation are showed



Click on a bar or the title of an endpoint, the corresponding feature or endpoint dataset on the Download page will open in a new tab.



In addition, users can enter the keywords of required endpoint and feature to search.

TOXRIC

Home Data Collection Search Benchmark & Representation Statistics Download Contribute Contact & About

Select a dataset

- Toxicity Datasets
  - Toxic Effect
    - Acute Toxicity
    - Carcinogenicity
    - Mutagenicity
    - Developmental and Reproductive Toxicity
    - Target Organ Toxicology
    - Applied Toxicology
    - Other Toxicology Datasets
  - Feature Datasets
    - Molecular Fingerprint
    - Category
    - Target

Toxicity Category

ACEA\_AR\_agonist\_80hr

Category	Descriptions	Number of compounds	Sources	Operation
ACEA_AR_agonist_80hr	Data from the assay component ACEA_AR_agonist_80hr was analyzed in the positive fitting direction relative to DMSO as the negative control and baseline of activity. Using a type of growth reporter, measures of the cells for gain-of-signal activity can be used to understand the signaling at the pathway-level as they relate to the geneAR. Furthermore, this assay endpoint can be referred to as a primary readout, because this assay has produced multiple assay endpoints where this one serves a signaling function. To generalize the intended target to other reliable targets, this assay endpoint is annotated to the "nuclear receptor" intended target family, where the subfamily is "steroidal".	1757	ToxCast database	

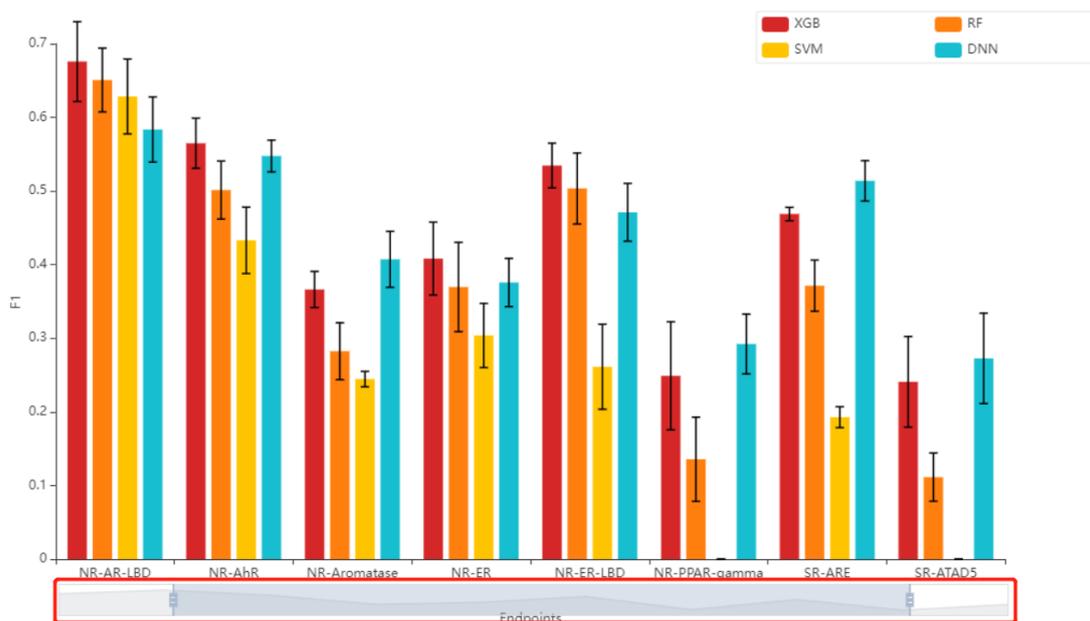
## 2. View benchmarks for algorithms

On Benchmarks for Algorithms page, the bar charts show the predictive effect of 4 algorithms on all toxicity endpoints.

XGB represents eXtreme Gradient Boosting. RF represents Random Forest. SVM represents Support Vector Machine. DNN represents Deep Neural Network.



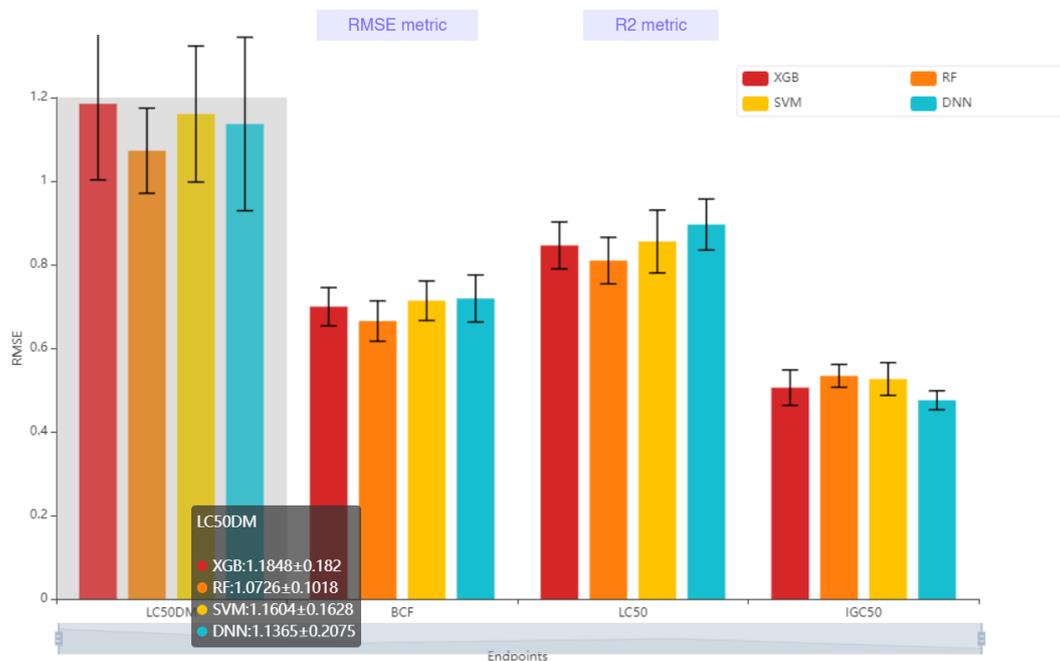
Each picture shows 10 endpoints, slide the mouse on the bar chart and drag the scroll bar below the chart to view the results of 10 endpoints.



Three metrics are used to evaluate the performance. The classification datasets use F1 metric, while the regression datasets (Acute Toxicity, Ecotoxicity) use RMSE and R2 metrics. For the regression datasets, click on RMSE or R2 button to view performance results.

For RMSE metric, lower value represents the higher prediction performance. While for R2 and F1 metric, the higher the value, the higher the prediction performance.

When the mouse is suspended on the bar, the mean and standard deviation are showed.



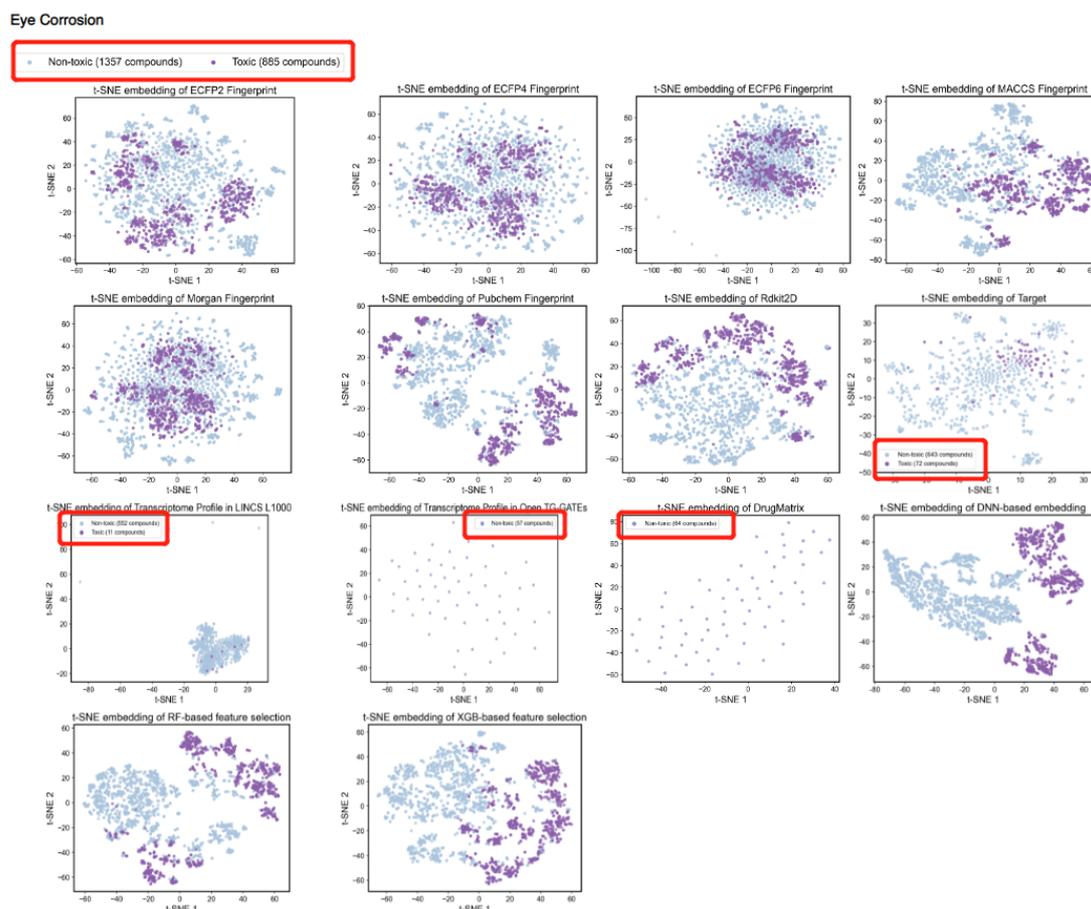
### 3.View t-SNE embedding of molecular representations

The T-SNE Embedding of Molecular Representations page shows the clustering effects of multiple

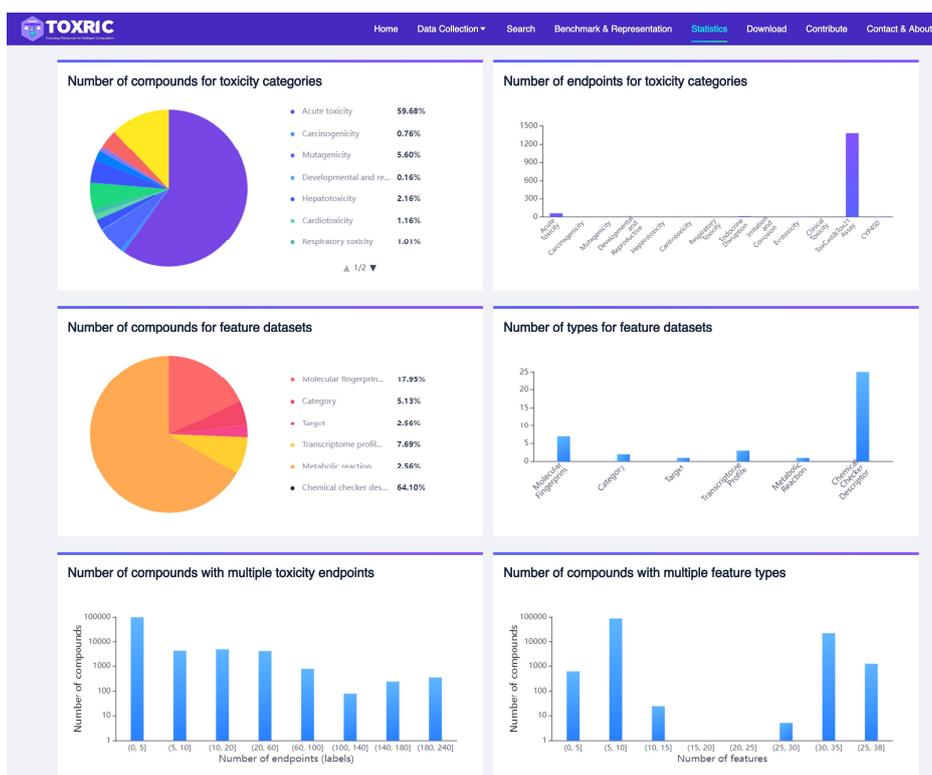
representations on the classification endpoint sub-datasets. The representations include 10 original features and three ML-based representations.

Above all scatter plots, the number of samples for both classes (Toxic, Non-toxic) in most images is shown. The number of samples in four images is different and marked separately. That is because there are compounds missing these four feature types.

Click to view or save each image.



[View statistical information of datasets](#)



### 1. Number of compounds for toxicity categories

The number of compounds under all toxicity categories is displayed in the form of a pie chart. Hover the mouse to view the number and proportion of compounds under this category.

### 2. Number of endpoints for toxicity categories

The number of endpoints under each toxicity category is displayed in the form of a bar chart. Hover the mouse to view the number of endpoints under this category.

### 3. Number of compounds for feature spaces

The number of compounds under all feature spaces is displayed in the form of a pie chart. Hover the mouse to view the number and proportion of compounds under this space.

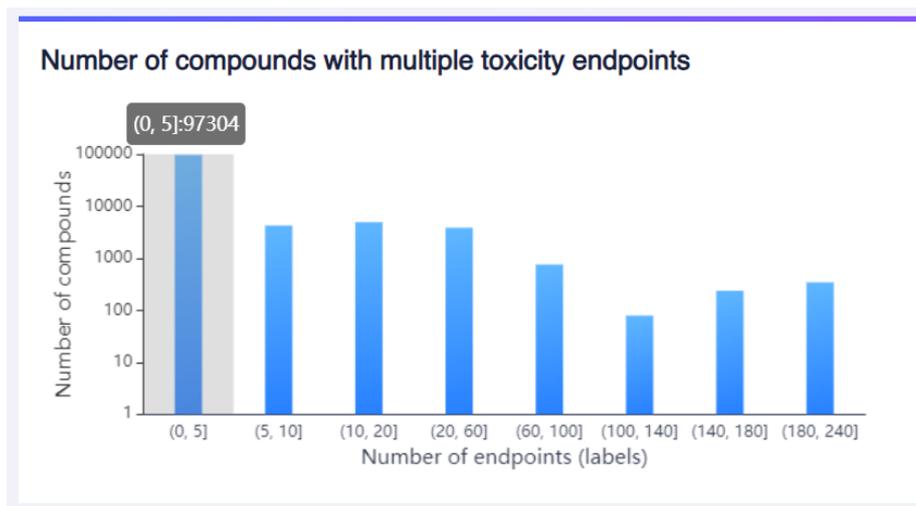
### 4. Number of types for feature spaces

The number of feature types under each feature space is displayed in the form of a bar chart. Hover the mouse to view the number of types under this feature space.

### 5. Number of compounds with multiple toxicity endpoints

For a compound, TOXRIC provides toxicity values across multiple endpoints. The number of compounds with multiple endpoint values is presented as a bar chart. Hover the mouse to view the number of compounds.

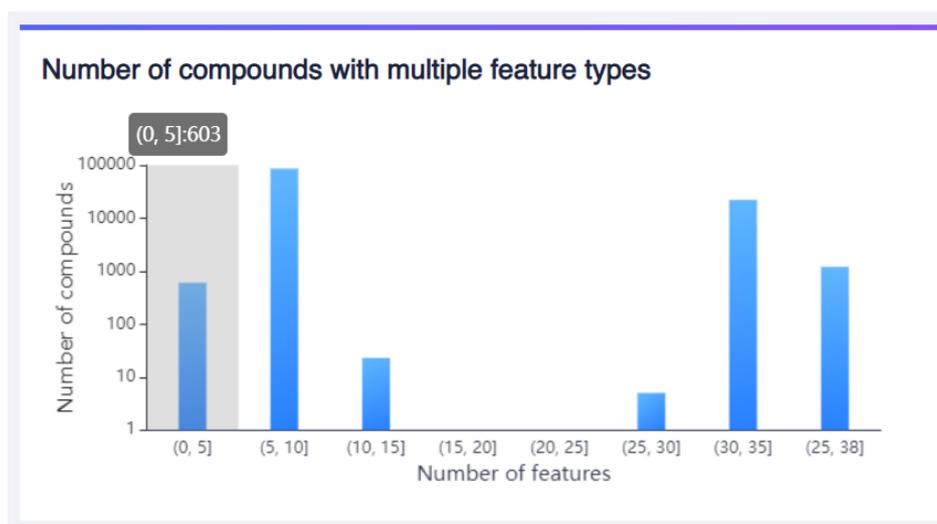
For example, the first bar indicates that 97,304 compounds have 0-5 endpoints values.



#### 6. Number of compounds with multiple features

For a compound, TOXRIC provides multiple features types. The number of compounds with multiple feature types is presented as a bar chart. Hover the mouse to view the number of compounds.

For example, the first bar indicates that 603 compounds have 0-5 feature types.



#### Download required dataset.

##### 1. Download the toxicity endpoint sub-dataset

Select a toxicity category and an endpoint dataset of interest. Click the Detail button to view the detailed information of this dataset. Click the Download button to download the dataset in .csv format.

**Select a dataset**

- Toxicity Datasets
  - Toxic Effect
    - Acute Toxicity
    - Carcinogenicity**
    - Mutagenicity
    - Developmental and Reproductive Toxicity
  - Target Organ Toxicology

**Toxicity Category**

Category	Descriptions	Number of compounds	Sources	Operation
Carcinogenicity	This file contain the data from the Carcinogenic Potency Database (CPDB) summary tables (CPDBAS, version 5d), which is a unique and standardized resource of long-term animal carcinogenesis study results on more than 1500 chemical substances.	1021	Scientific Literature	

It should be noted that in Acute Toxicity and Ecotoxicity datasets, two types of endpoint data, the values with (mg/kg) or (mg/L) units and the dimensionless values, are provided to download.

**Select a dataset**

- Toxicity Datasets
  - Toxic Effect
    - Acute Toxicity**
    - Carcinogenicity
    - Mutagenicity
    - Developmental and Reproductive Toxicity
  - Target Organ Toxicology
  - Applied Toxicology

**Toxicity Category**

Category	Descriptions	Number of compounds	Sources	Operation
mouse_intraperitoneal_LD50	This file contains the Lethal Dose Fifty (LD50) values for mice via the intraperitoneal route.	35299	Scientific Literature	
mammal (species unspecified)_intraperitoneal_LD50	This file contains the Lethal Dose Fifty (LD50) values for mammals (species unspecified) via the intraperitoneal route.	537	Scientific Literature	

## 2. Download the feature type sub-dataset

Select a feature space and a feature type dataset of interest. Click the Detail button to view the detailed information of this dataset. Click the Download button to download the dataset in .csv format.

**Select a dataset**

- Toxicity Datasets
  - Toxic Effect
    - Acute Toxicity
    - Carcinogenicity
    - Mutagenicity
    - Developmental and Reproductive Toxicity
  - Target Organ Toxicology
  - Applied Toxicology
  - Other Toxicology Datasets
  - Feature Datasets
    - Molecular Fingerprint**
    - Category
    - Target
    - Transcriptome Profile
    - Metabolic Reaction

**Feature Datasets**

Category	Descriptions	Number of compounds	Sources	Operation
ECFP2	ECFP2 is a 2048-length bits vector, which represent the neighbor hood environment of each atom using the extended connectivity fingerprint encoding a circular substructure of diameter 2 bonds. An individual bit has no definite meaning.	110000	Rdkit	
ECFP4	ECFP4 is a 2048-length bits vector, which represent the neighbor hood environment of each atom using the extended connectivity fingerprint encoding a circular substructure of diameter 4 bonds. An individual bit has no definite meaning.	110000	Rdkit	
ECFP6	ECFP6 is a 2048-length bits vector, which represent the neighbor hood environment of each atom using the extended connectivity fingerprint encoding a circular substructure of diameter 6 bonds. An individual bit has no definite meaning.	110001	Rdkit	

## Example application for toxicity prediction

This section describes how to use TOXRIC for toxicity prediction using the mouse\_intraperitoneal\_LD50 sub-dataset (Acute Toxicity) as an example.

1. Browse the Data Collection->Toxicity Category page. Select a toxicity category (Acute Toxicity). Select an endpoint dataset as the sample set.

TOXRIC  
Toxicity Research by Open Access

Home Data Collection Search Benchmark & Representation Statistics Download Contribute Contact & About

**Toxicity Category**

- Toxic Effect
  - Acute Toxicity**
  - Carcinogenicity
  - Mutagenicity
  - Developmental and Reproductive Toxicity
- Target Organ Toxicology
  - Hepatotoxicity
  - Cardiotoxicity
  - Respiratory Toxicity
  - Endocrine Disruption
  - Irritation and Corrosion
- Applied Toxicology
  - Ecotoxicity

**Toxicity Category**

Endpoint	Description	Number of compounds	Sources	Details
mouse_intraperitoneal_LD50	This file contains the Lethal Dose Fifty (LD50) values for mice via the intraperitoneal route.	35299	Scientific Literature	
mammal (species unspecified)_intraperitoneal_LD50	This file contains the Lethal Dose Fifty (LD50) values for mammals (species unspecified) via the intraperitoneal route.	537	Scientific Literature	
guinea pig_intraperitoneal_LD50	This file contains the Lethal Dose Fifty (LD50) values for guinea pigs via the intraperitoneal route.	237	Scientific Literature	
rat_intraperitoneal_LD50	This file contains the Lethal Dose Fifty (LD50) values for rats via the intraperitoneal route.	4798	Scientific Literature	
rabbit_intraperitoneal_LD50	This file contains the Lethal Dose Fifty (LD50) values for rabbits via the intraperitoneal route.	113	Scientific Literature	

2. On Download page, download this endpoint sub-dataset as the label data.

TOXRIC  
Toxicity Research by Open Access

Home Data Collection Search Benchmark & Representation Statistics Download Contribute Contact & About

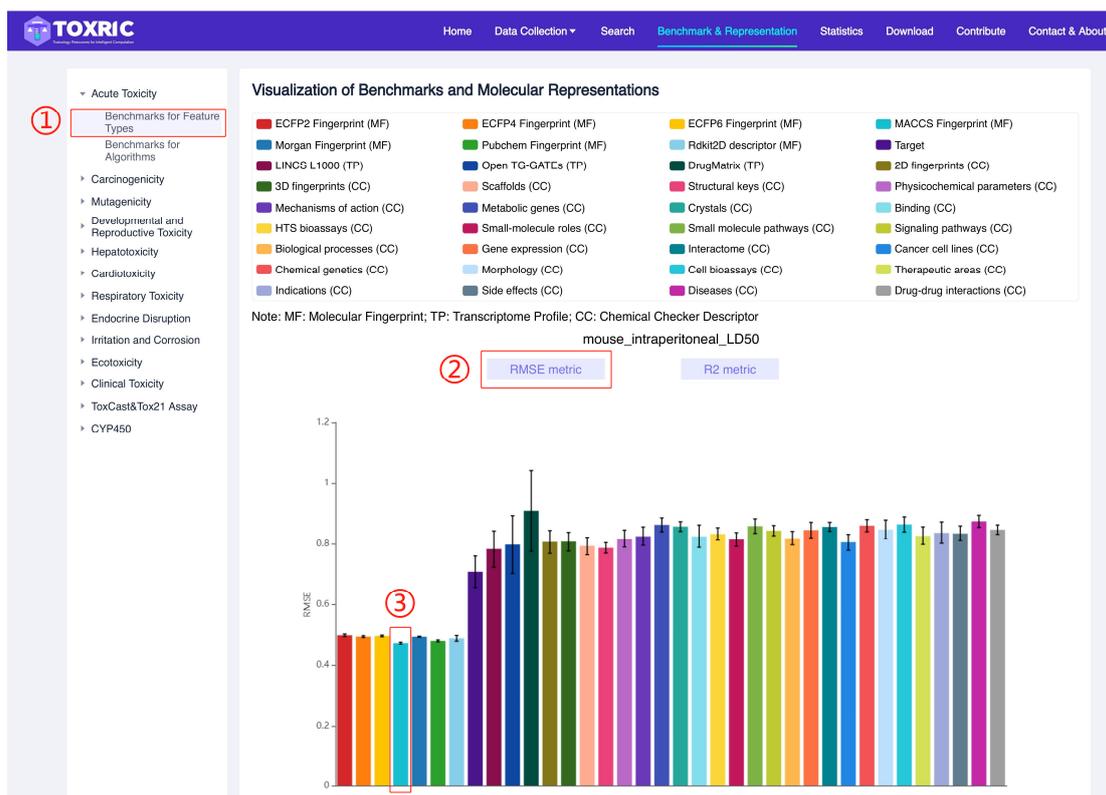
**Select a dataset**

- Toxicity Datasets
  - Toxic Effect
    - Acute Toxicity**
    - Carcinogenicity
    - Mutagenicity
    - Developmental and Reproductive Toxicity
  - Target Organ Toxicology
  - Applied Toxicology
  - Other Toxicology Datasets
- Feature Datasets
  - Molecular Fingerprint
  - Category
  - Target
  - Transcriptome Profile

**Toxicity Category**

Category	Descriptions	Number of compounds	Sources	Operation
mouse_intraperitoneal_LD50	This file contains the Lethal Dose Fifty (LD50) values for mice via the intraperitoneal route.	35299	Scientific Literature	
mammal (species unspecified)_intraperitoneal_LD50	This file contains the Lethal Dose Fifty (LD50) values for mammals (species unspecified) via the intraperitoneal route.	537	Scientific Literature	
guinea pig_intraperitoneal_LD50	This file contains the Lethal Dose Fifty (LD50) values for guinea pigs via the intraperitoneal route.	237	Scientific Literature	
rat_intraperitoneal_LD50	This file contains the Lethal Dose Fifty (LD50) values for rats via the intraperitoneal route.	4798	Scientific Literature	

3. On the Benchmark&Representation page, view the benchmarks of feature types on this endpoint. It is found that the MACCS molecular fingerprint achieved the best performance.



4. Click the bar of MACCS molecular fingerprint and enter the Download page. Then, the MACCS fingerprint sub-dataset can be downloaded as the input feature.

**Select a dataset**

- Toxicity Datasets
  - Toxic Effect
    - Acute Toxicity
    - Carcinogenicity
    - Mutagenicity
    - Developmental and Reproductive Toxicity
  - Target Organ Toxicology
  - Applied Toxicology
  - Other Toxicology Datasets

**Feature Datasets**

MACCS Fingerprint

false

Category	Descriptions	Number of compounds	Sources	Operation
MACCS	MACCS is a 168-length bits vector, which is substructure key-based fingerprints representing the presence of certain substructures or fragments from a given list of structural keys in the compound.	110000	Rdkit	

ALL 1 items

1 / 10 page

Now, The input and output datasets are ready for toxicity prediction.

5. On the Benchmark&Representation page, view the benchmarks of algorithms on this endpoint.



It is found that the RF algorithm achieved the best performance.

Therefore, in this dataset, RF can be considered as the baseline for the development of new ML algorithms.