Supplementary information

Nucleic Acids Research

TOXRIC: a comprehensive database of toxicological data and benchmarks

Lanlian Wu [1,2,†], **Bowei Yan** [1,3,†], **Junshan Han** [1], **Ruijiang Li** [1], **Jian Xiao** [4,5], **Song He** [1,*] **and Xiaochen Bo** [1,2,*]

[1]Department of Bioinformatics, Institute of Health Service and Transfusion Medicine, Beijing, China.
[2]Academy of Medical Engineering and Translational Medicine, Tianjin University, Tianjin, China.
[3]State Key Laboratory of Genetic Engineering, Institutes of Biomedical Sciences, Fudan University, Shanghai, China.
[4]Department of Pharmacy, Xiangya Hospital, Central South University, Changsha, China.
[5]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.

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

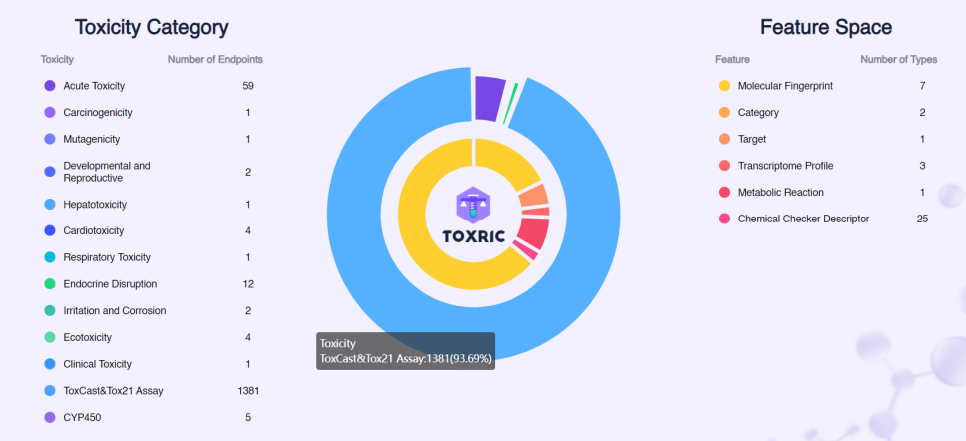
STEP-BY-STEP TUTORIAL OF TOXRIC WEBSITE (https://toxric.bioinforai.tech/)

Data browsing.

1. Browse dataset composition on Home page

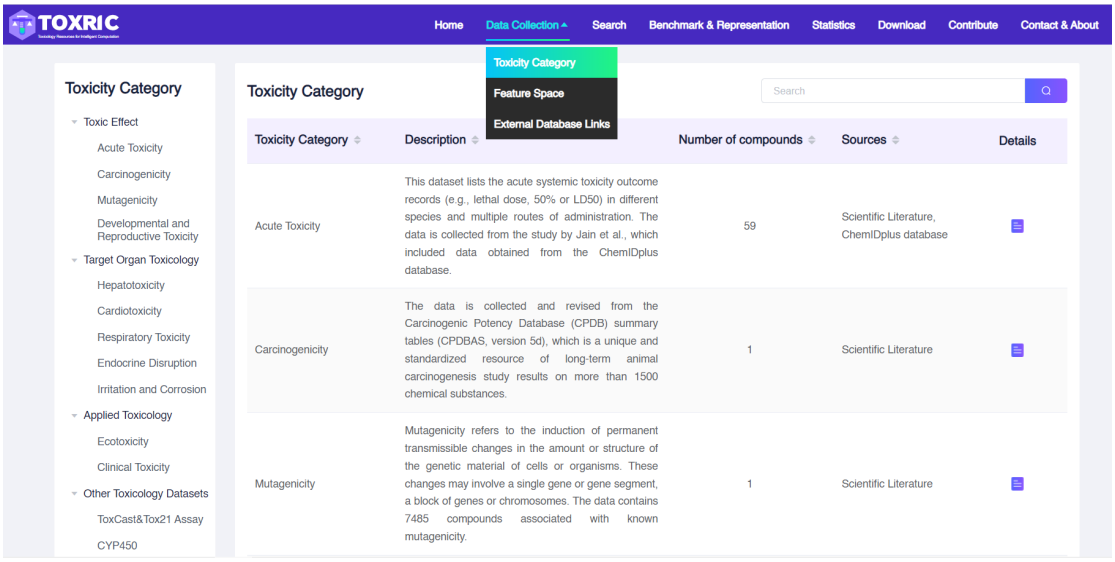
TOXRIC provides 13 toxicity datasets and 6 feature datasets. Each dataset includes multiple sub-datasets of toxicity endpoints and feature types.

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.

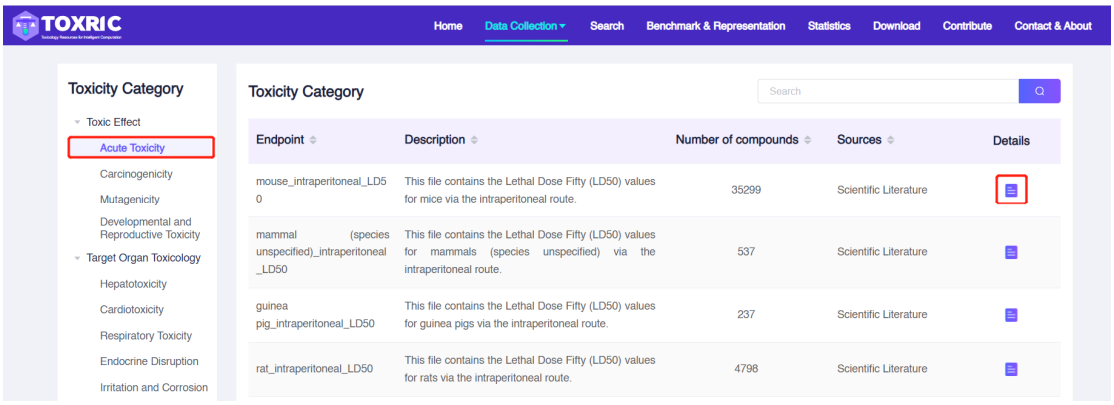


2. Browse information of toxicity datasets on Data Collection page

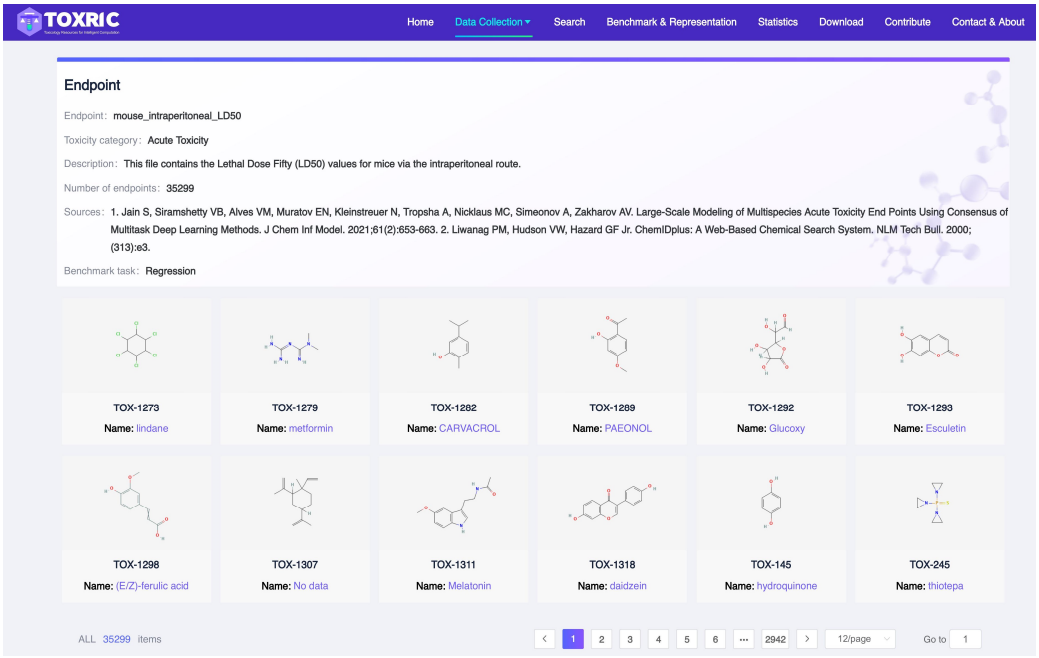
(1) Enter the Toxicity Dataset page on Data Collection page, or click the Toxicity Dataset field in the page, the descriptions of all toxicity datasets will be displayed, including dataset description, number of compounds, and sources. Click the Details button on the far right of each row to view the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



(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.

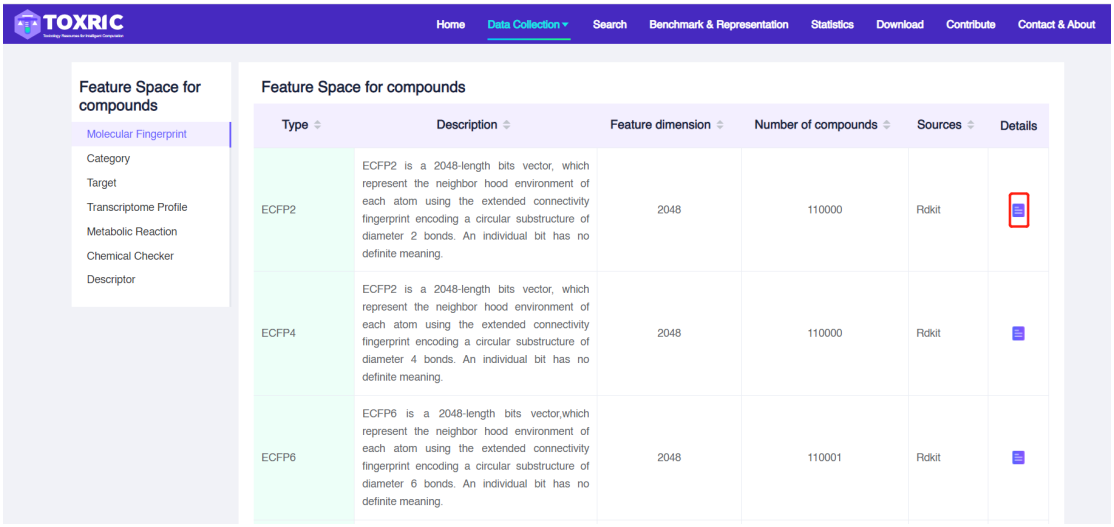


On the detail information page, all compounds contained in an endpoint sub-dataset are listed in the form of a molecular graph. Clicking on a specific compound will open the compound information page that displays affluent chemical, toxicological, and feature data of the compound.

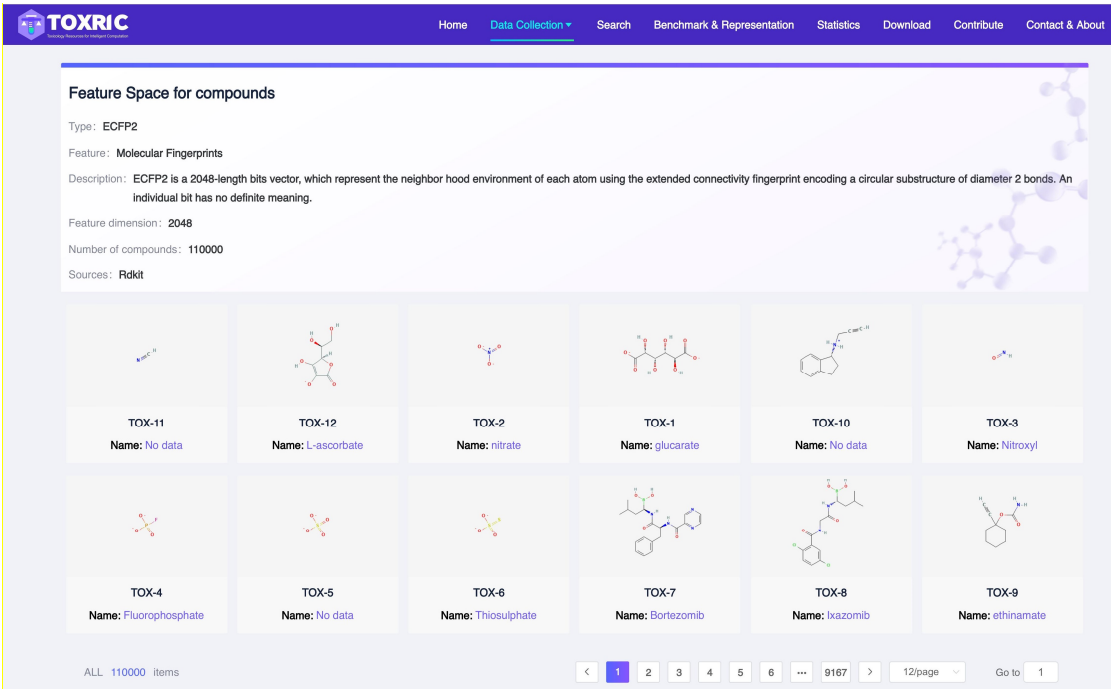


3. Browse information of feature datasets on Data Collection page

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.



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.



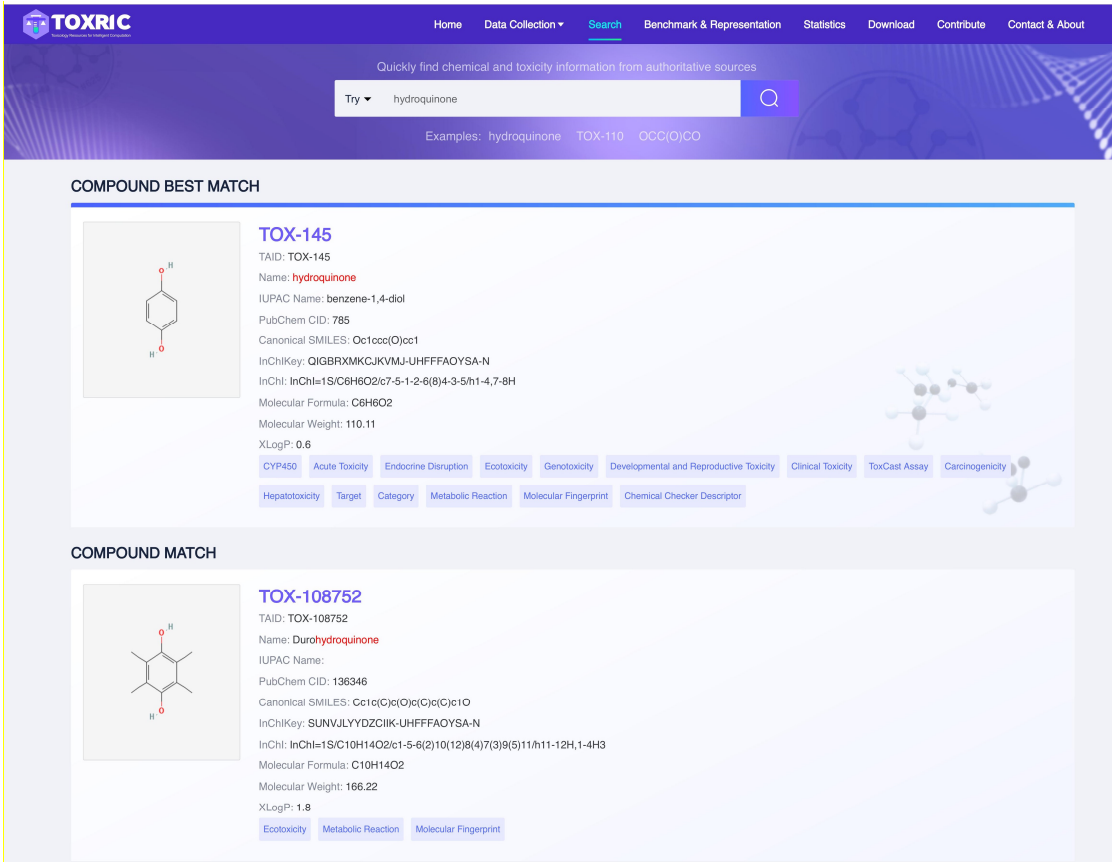
Data retrieval.

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.

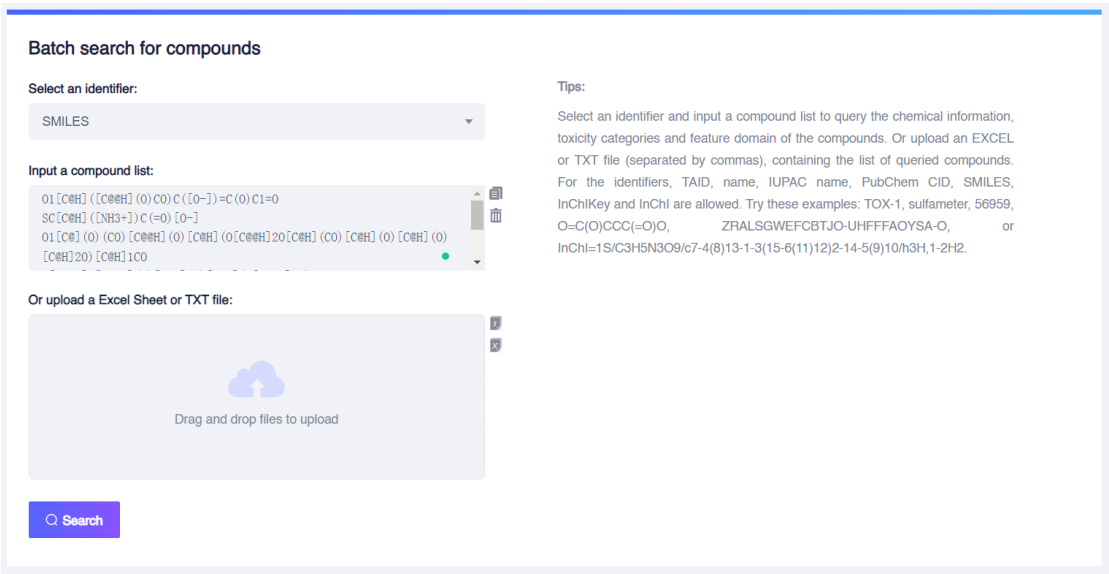
(2) After clicking the search button or pressing the Enter key, a list of compound entities will be provided. The query keyword is highlighted in red.



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.



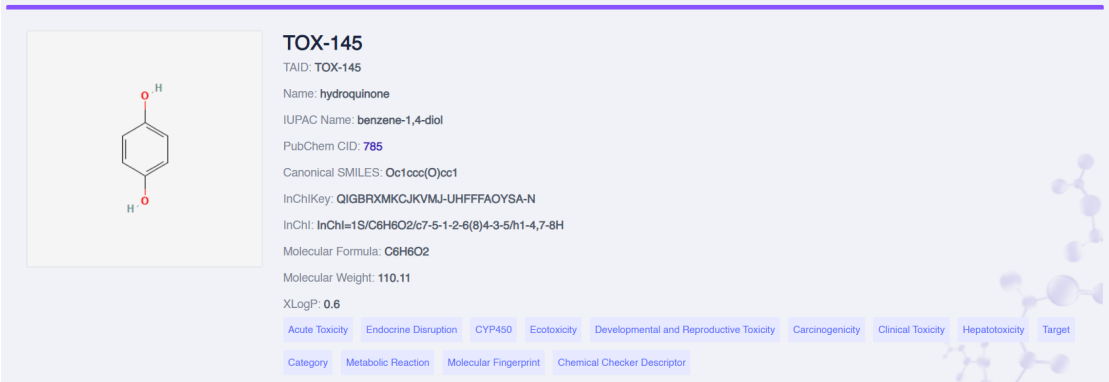
(3) Or you can upload an EXCEL or TXT file to query the information of the compounds. The template file can be downloaded by clicking the button to the right of the upload box.

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

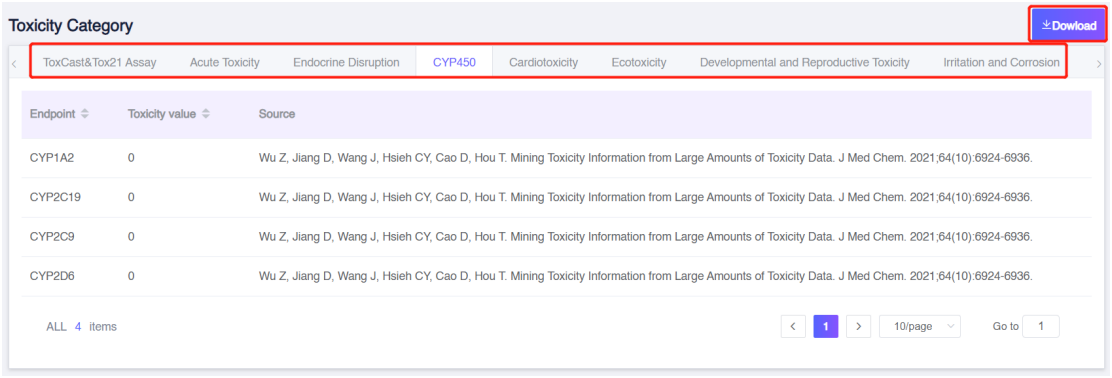
The chemical information section provides seven commonly used identifier types and physicochemical properties of compounds. Click on the PubChem CID to link to the Compound page of PubChem website.



(2) Toxicity category

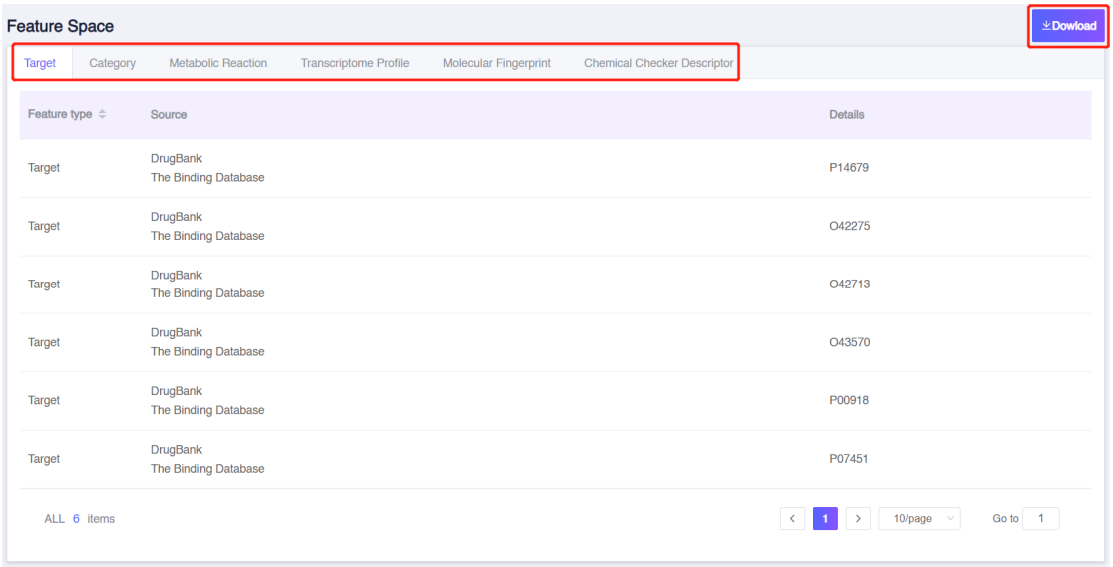
Click on the toxicity categories in the title bar to view the toxicity values for each endpoint. If no toxicity value is currently collected for the compound, it will be shown as “Data is not available now”.

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.

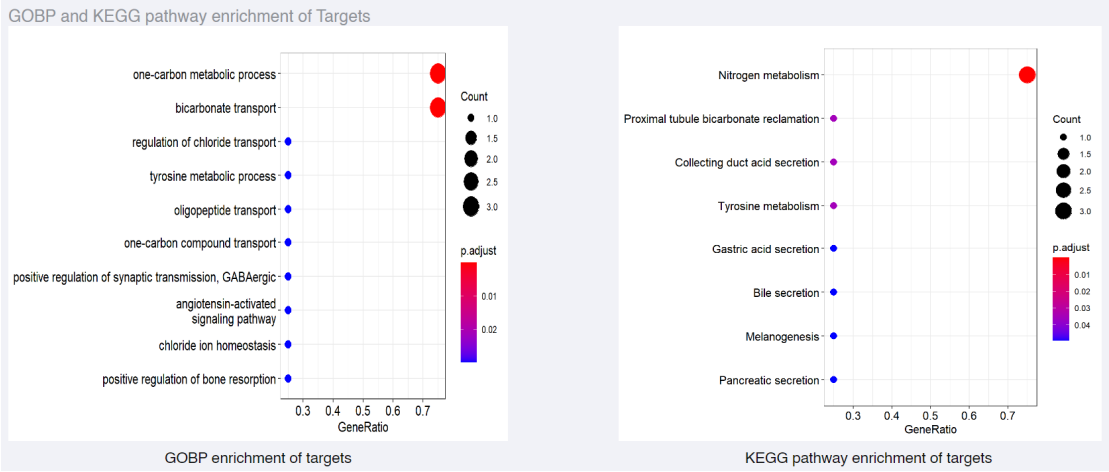


(3) Feature space

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.



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.



GOBP enrichment of targets

KEGG pathway enrichment of targets

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.

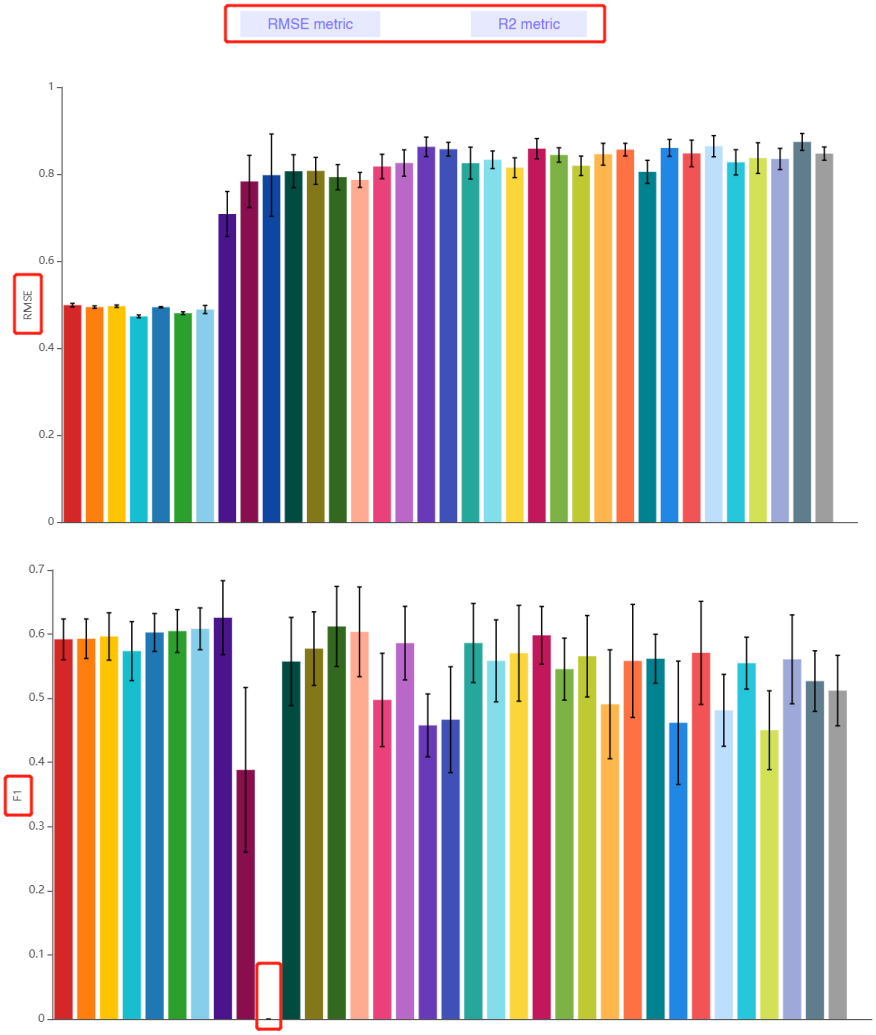
In the legend, the parentheses after each feature type indicate the feature space it belongs to. MF represents Molecular Fingerprint, TP represents Transcriptome Profile, and CC represents Chemical Checker Descriptor.

ECFP2 Fingerprint (MF)	ECFP4 Fingerprint (MF)	ECFP6 Fingerprint (MF)	MACCS Fingerprint (MF)
Morgan Fingerprint (MF)	Pubchem Fingerprint (MF)	Rdkit2D descriptor (MF)	Target
LINCS L1000 (TP)	Open TG-GATEs (TP)	DrugMatrix (TP)	2D fingerprints (CC)
3D fingerprints (CC)	Scaffolds (CC)	Structural keys (CC)	Physicochemical parameters (CC)
Mechanisms of action (CC)	Metabolic genes (CC)	Crystals (CC)	Binding (CC)
HTS bioassays (CC)	Small-molecule roles (CC)	Small molecule pathways (CC)	Signaling pathways (CC)
Biological processes (CC)	Gene expression (CC)	Interactome (CC)	Cancer cell lines (CC)
Chemical genetics (CC)	Morphology (CC)	Cell bioassays (CC)	Therapeutic areas (CC)
Indications (CC)	Side effects (CC)	Diseases (CC)	Drug-drug interactions (CC)

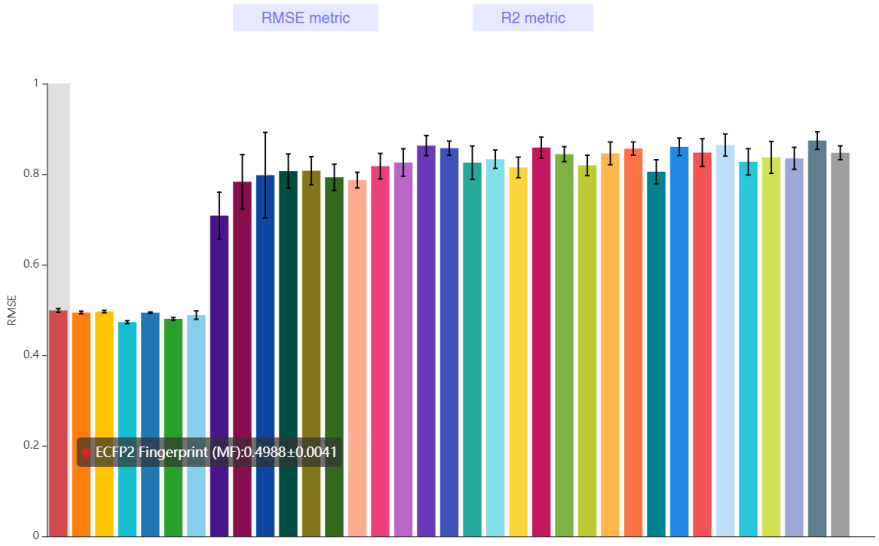
Note: MF: Molecular Fingerprint; TP: Transcriptome Profile; CC: Chemical Checker Descriptor

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.

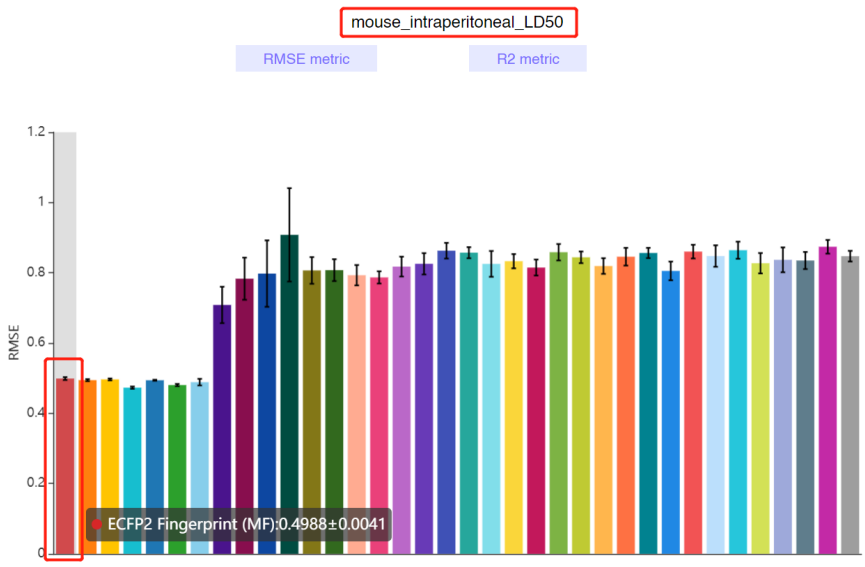
For RMSE metric, lower value represents the higher prediction performance. While for R2 and F1 metric, the higher the value, the higher the prediction performance. It should be noted that if the value of metric is 0, it represents the number of samples with the feature type at the endpoint sub-dataset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



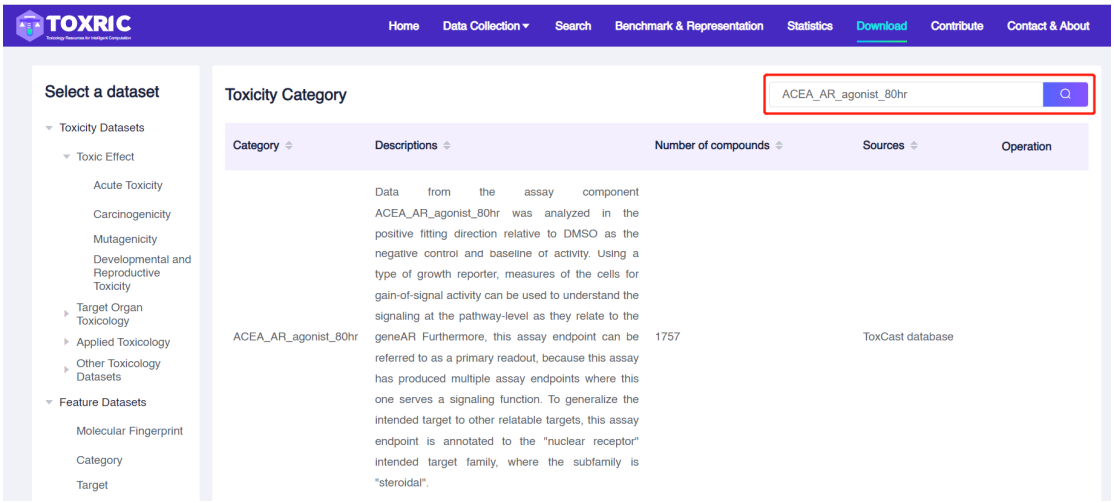
When the mouse is suspended on the bar, the mean and standard deviation are showed



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.



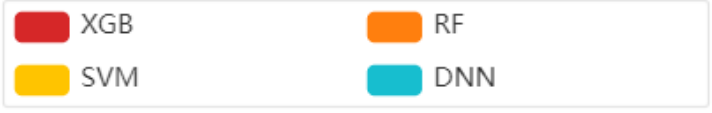
In addition, users can enter the keywords of required endpoint and feature to search.



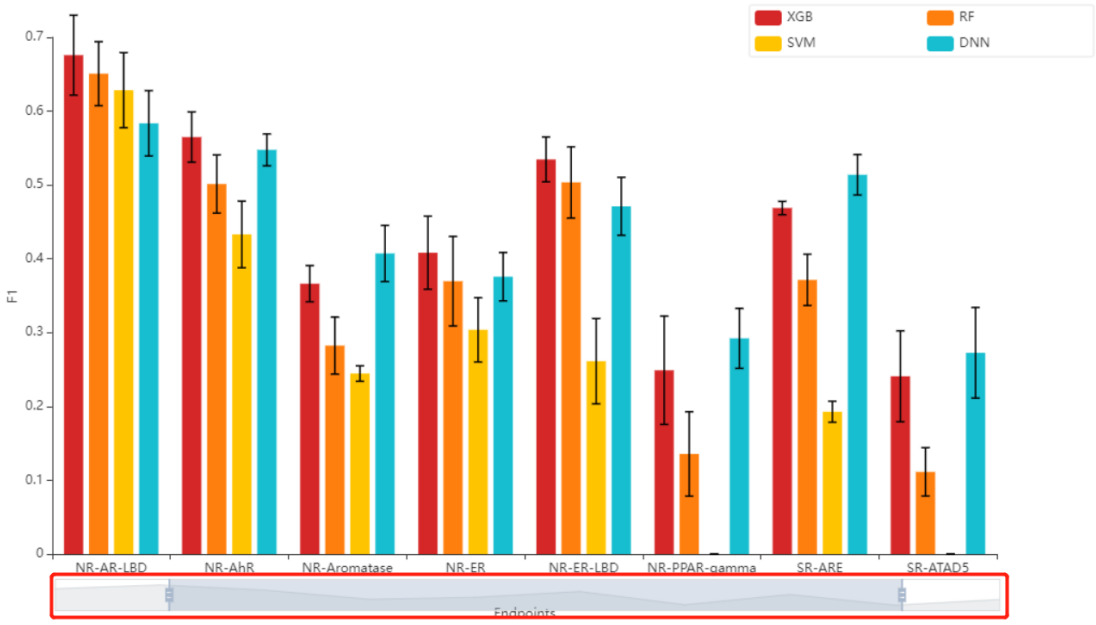
2. View benchmarks for algorithms

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.



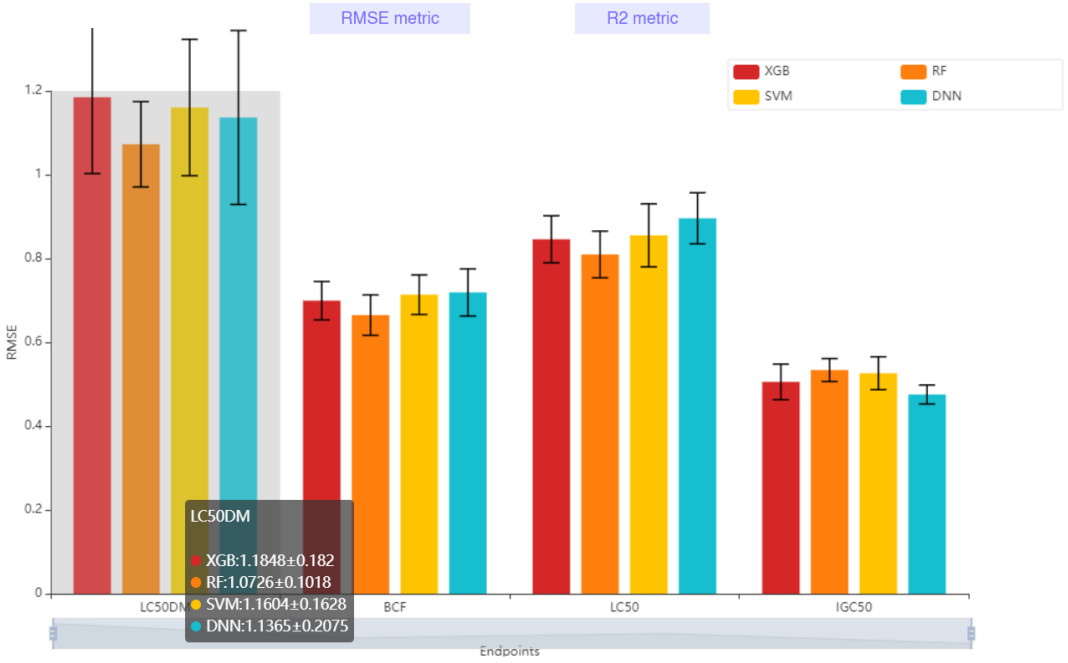
Each picture shows 10 endpoints, slide the mouse on the bar chart and drag the scroll bar below the chart to view the results of 10 endpoints.



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.



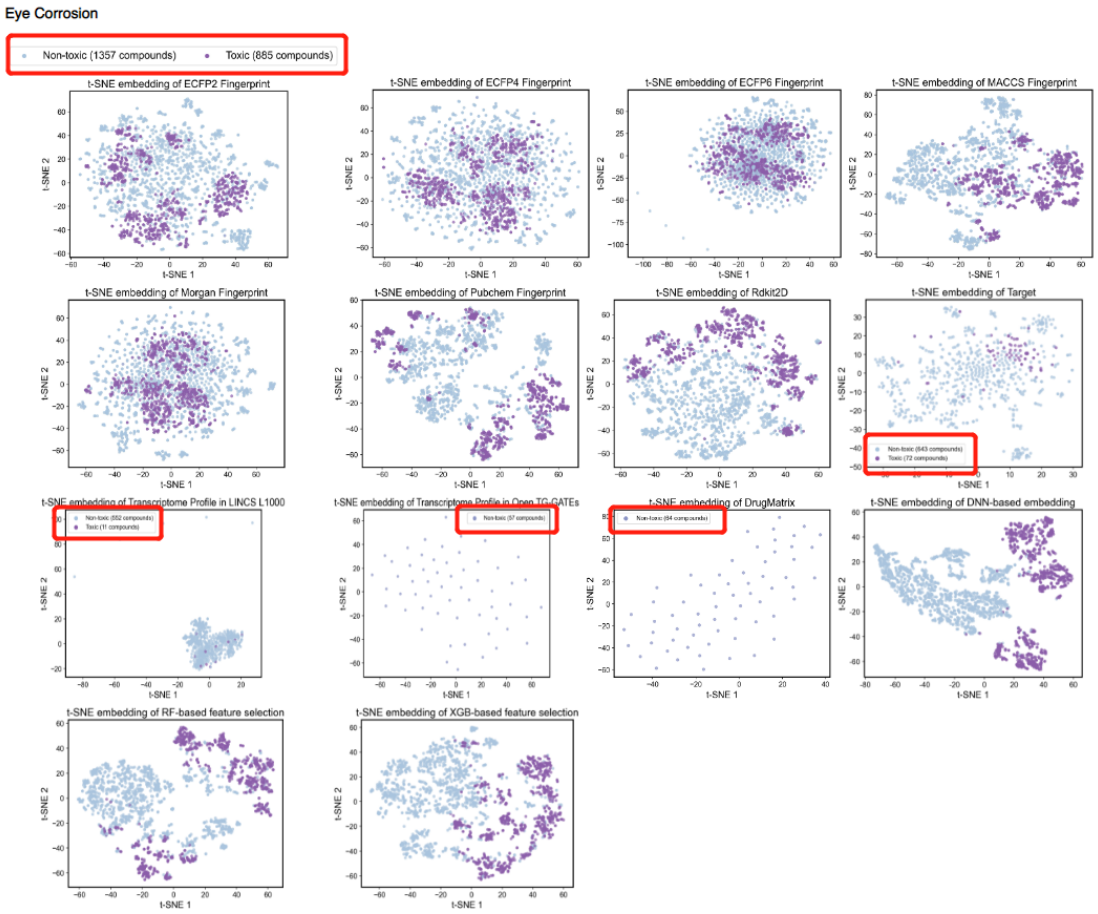
3.View t-SNE embedding of molecular representations

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

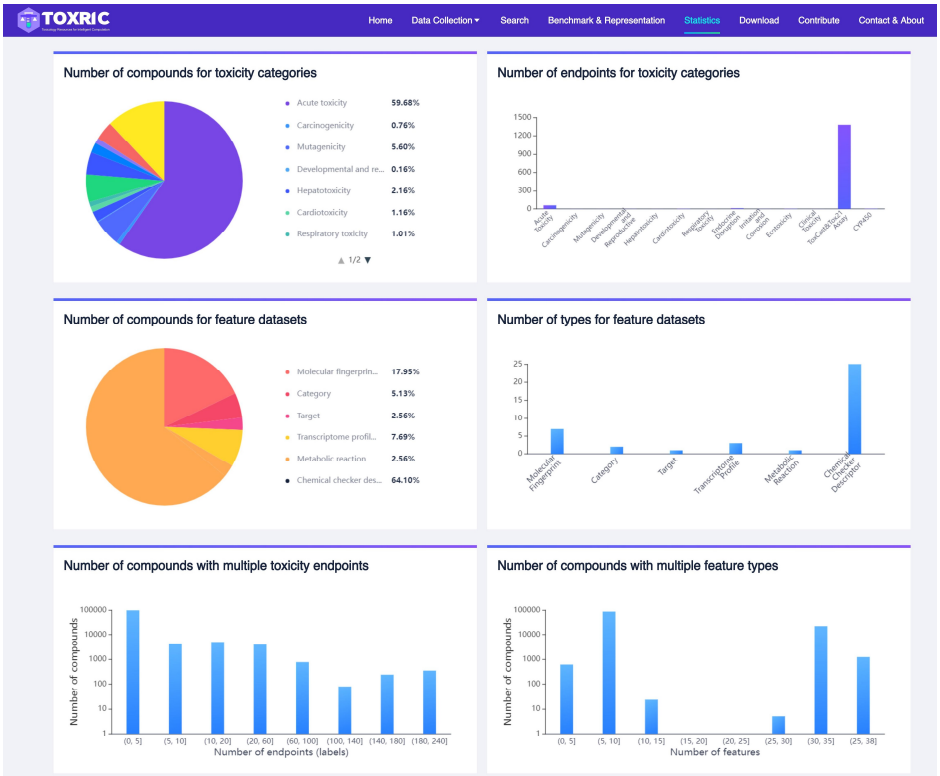
representations on the classification endpoint sub-datasets. The representations include 10 original features and three ML-based representations.

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.

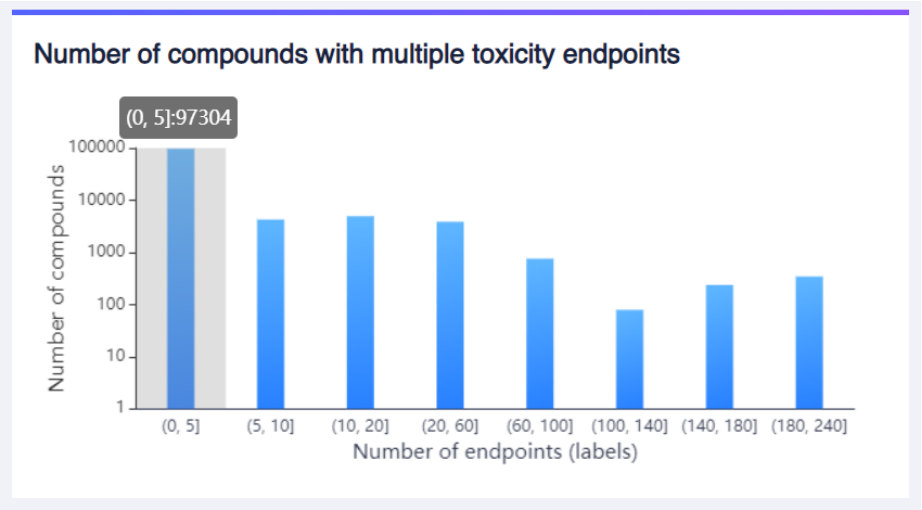
4. Number of types for feature spaces

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.

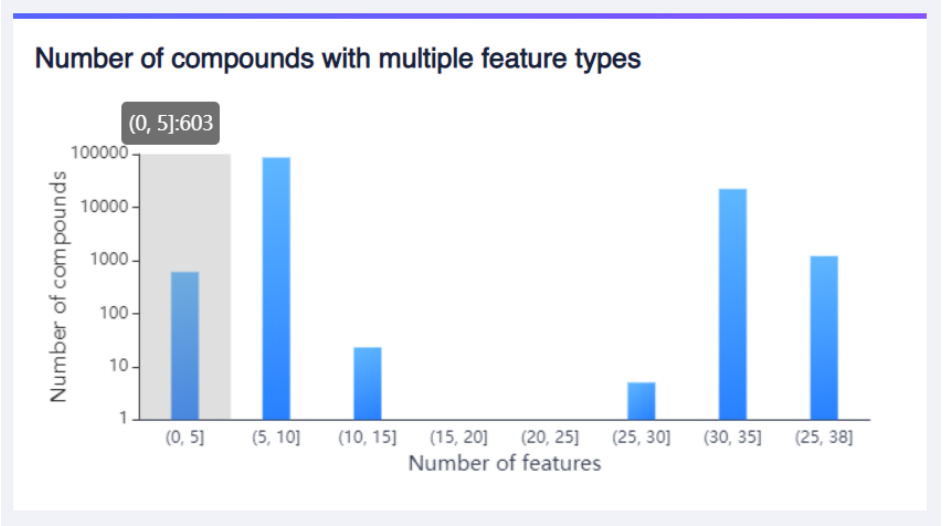
For example, the first bar indicates that 97,304 compounds have 0-5 endpoints values.



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.

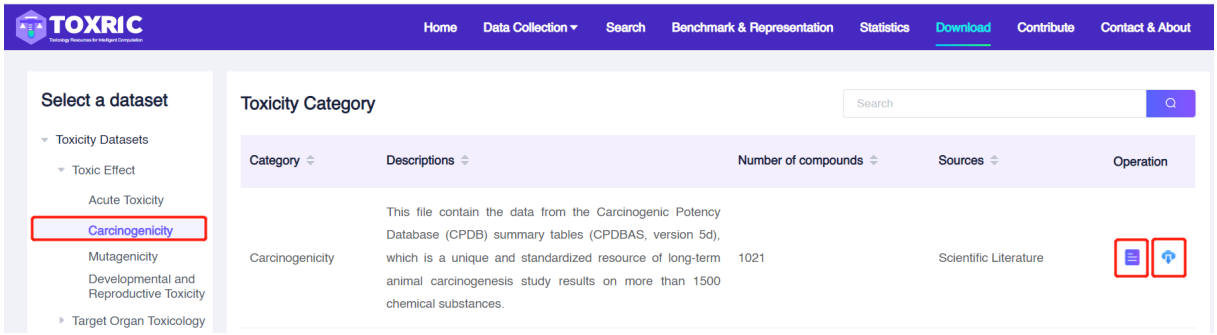
For example, the first bar indicates that 603 compounds have 0-5 feature types.



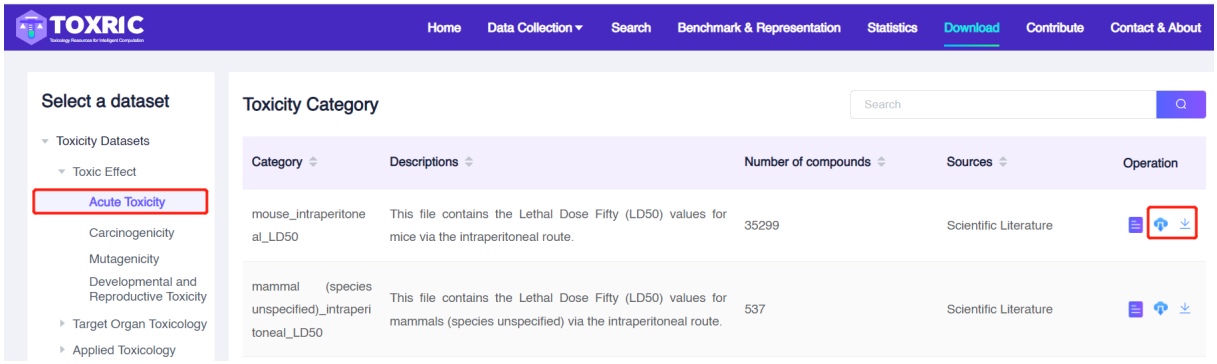
Download required dataset.

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.

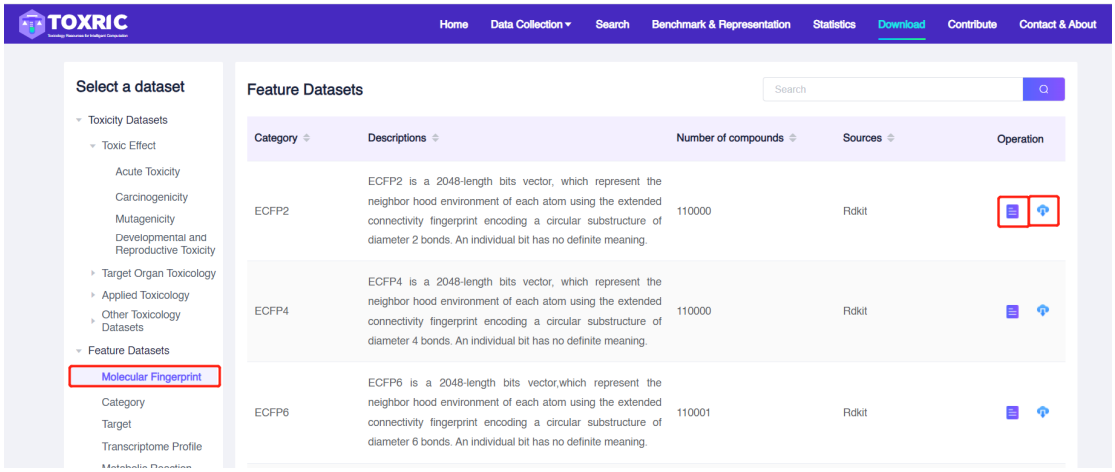


It should be noted that in Acute Toxicity and Ecotoxicity datasets, two types of endpoint data, the values with (mg/kg) or (mg/L) units and the dimensionless values, are provided to download.



2.Download the feature type sub-dataset

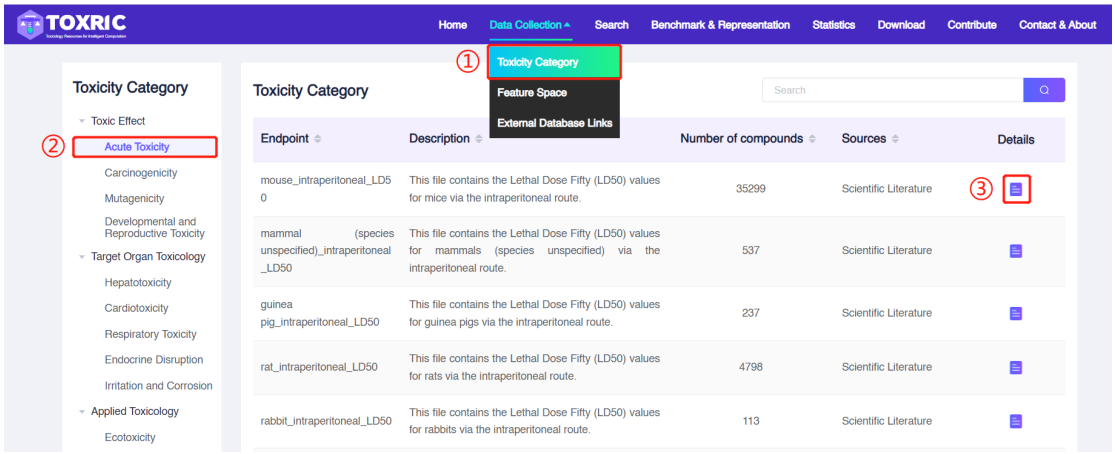
Select a feature space and a feature type dataset of interest. Click the Detail button to view the detailed information of this dataset. Click the Download button to download the dataset in .csv format.



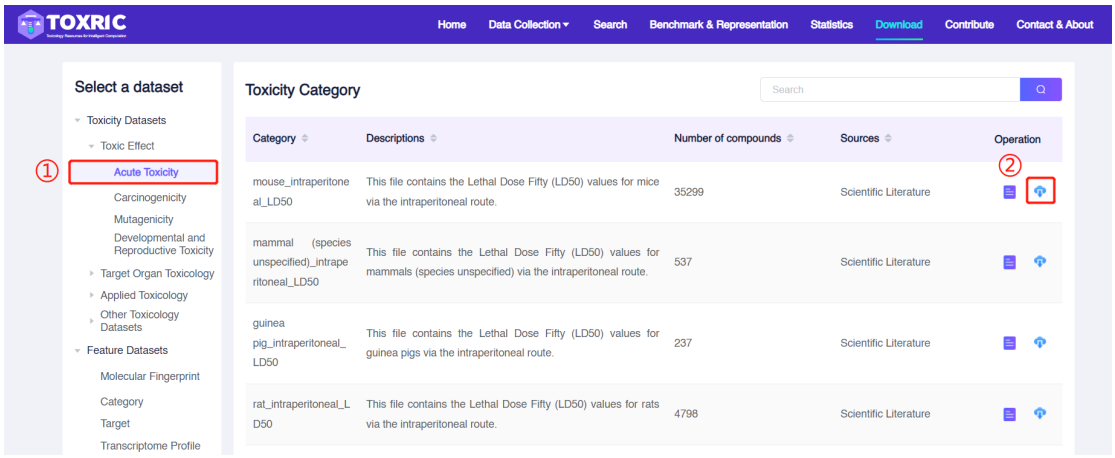
Example application for toxicity prediction

This section describes how to use TOXRIC for toxicity prediction using the mouse_intraperitoneal_LD50 sub-dataset (Acute Toxicity) as an example.

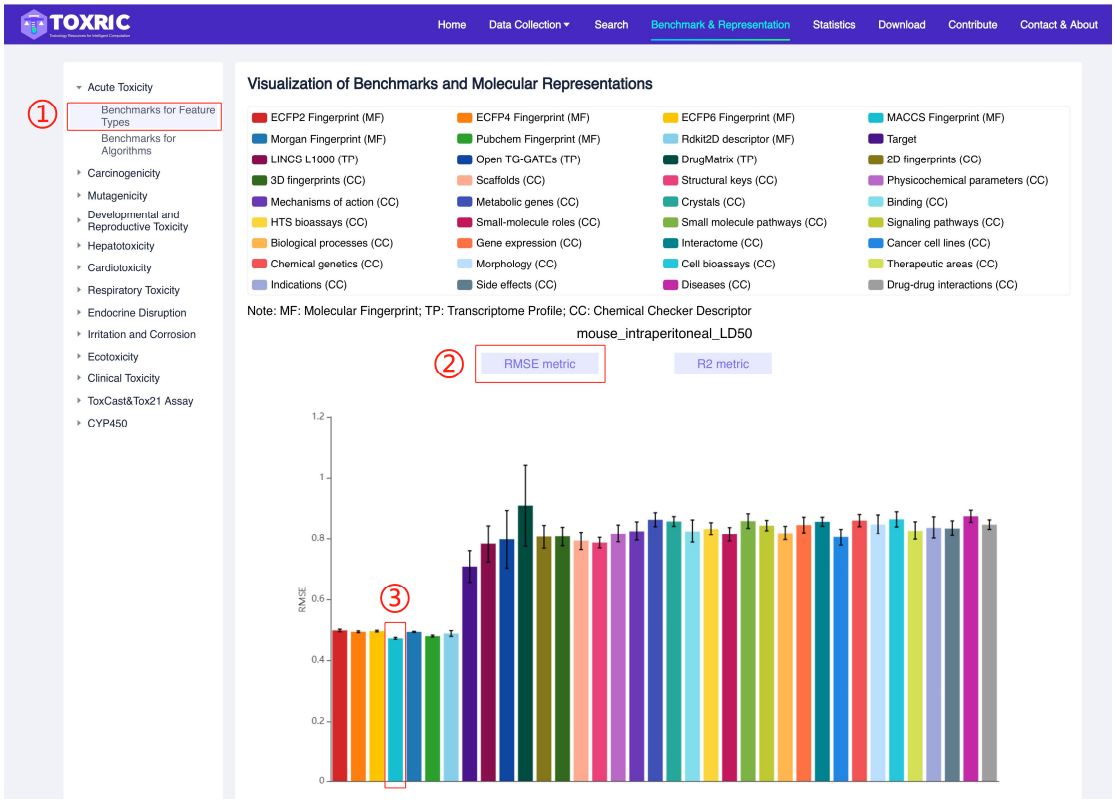
1. Browse the Data Collection->Toxicity Category page. Select a toxicity category (Acute Toxicity). Select an endpoint dataset as the sample set.



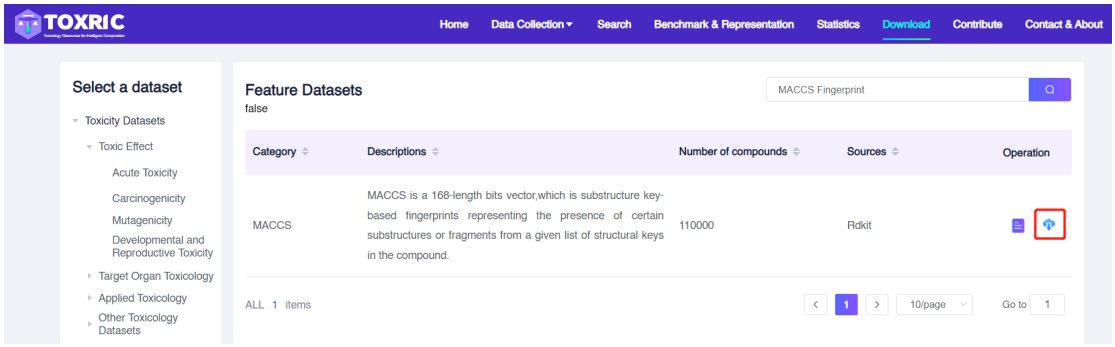
2. On Download page, download this endpoint sub-dataset as the label data.



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.



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.