XXXI Symposium on Bioinformatics and Computer-Aided Drug Discovery (BCADD-2025)

TOXAI ASSISTANT - AN IN SILICO ALTERNATIVE TO RATS TESTING FOR ACUTE TOXICITY

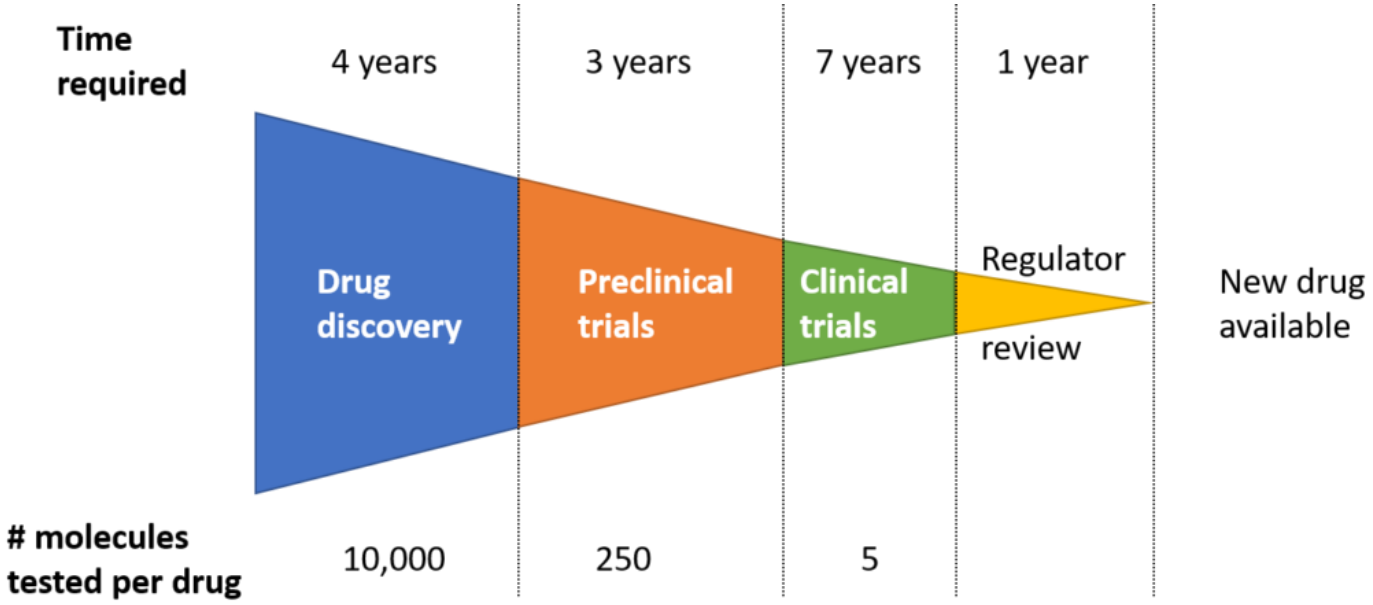
O.V. Tinkov[1], V.Y. Grigorev[2]

[1] Pridnestrovian State University, Tiraspol, Moldova;
[2] Institute of Physiologically Active Compounds at Federal Research Center of Problems of Chemical Physics and Medicinal Chemistry, Russian Academy of Sciences, Chernogolovka, Russia

Moscow, 2025

Introduction:

- According to experts [1], the development of a new drug on the global market takes an average of 10–15 years and requires investments ranging from 1 billion dollars to colossal 11.8 billion dollars;
- A key aspect of this process is predicting and understanding the toxicity profile of a drug, since approximately 40% of the developed drugs fail to reach the market due to toxicity problems.



1. Aghila Rani KG, Hamad MA, Zaher M, et al. Drug development post COVID-19 pandemic: toward a better system to meet current and future global health challenges. Expert Opin Drug Discov. 2021;16:365–371. doi: 10.1080/17460441.2021.1854221

Introduction:

Why is QSAR modeling necessary when assessing the toxicity of organic compounds that are potential drugs?

- ✓ **Reduces the cost of medicines by reducing financial resources, allowing you to choose the optimal development strategy;**
- ✓ **Maintains the life and health of laboratory animals;**
- ✓ **Allows you to discover hidden patterns.**

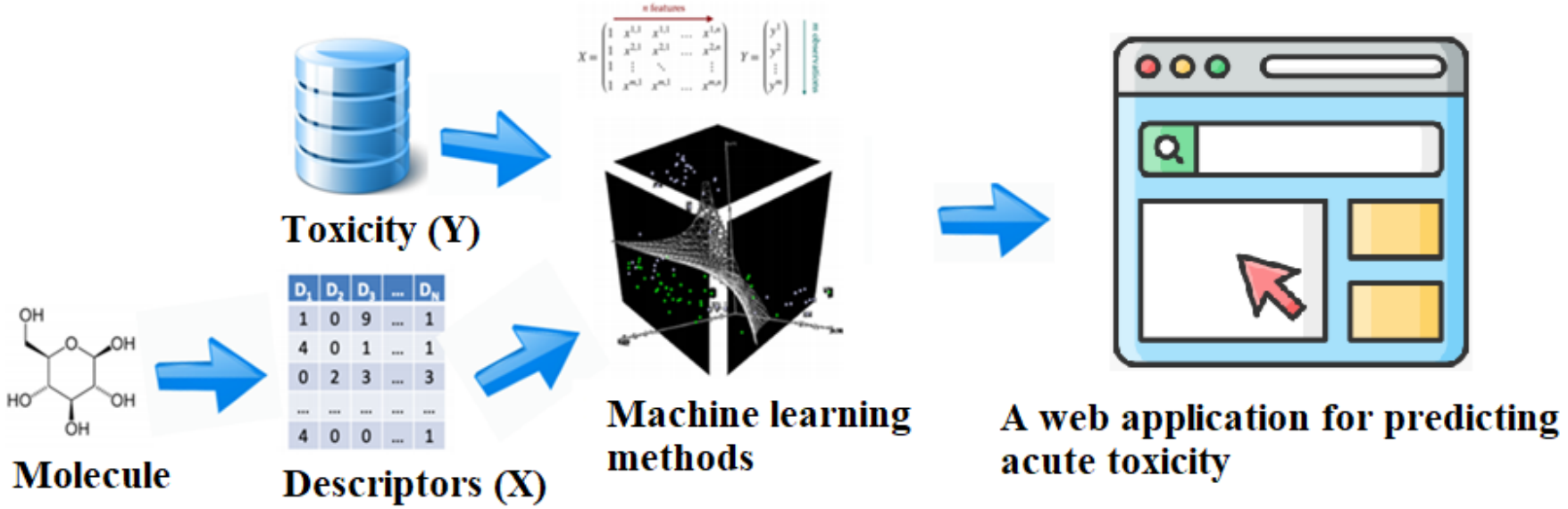
Introduction:

- When deciding whether to further investigate a chemical compound as a potential drug, it is crucial to evaluate its acute toxicity level or class;
- Currently, the World Health Organization (WHO) provides a widely accepted classification system for chemical substances based on acute toxicity levels;
- One of the most important criteria for acute toxicity classification of chemicals is LD_{50} value for oral administration in rats

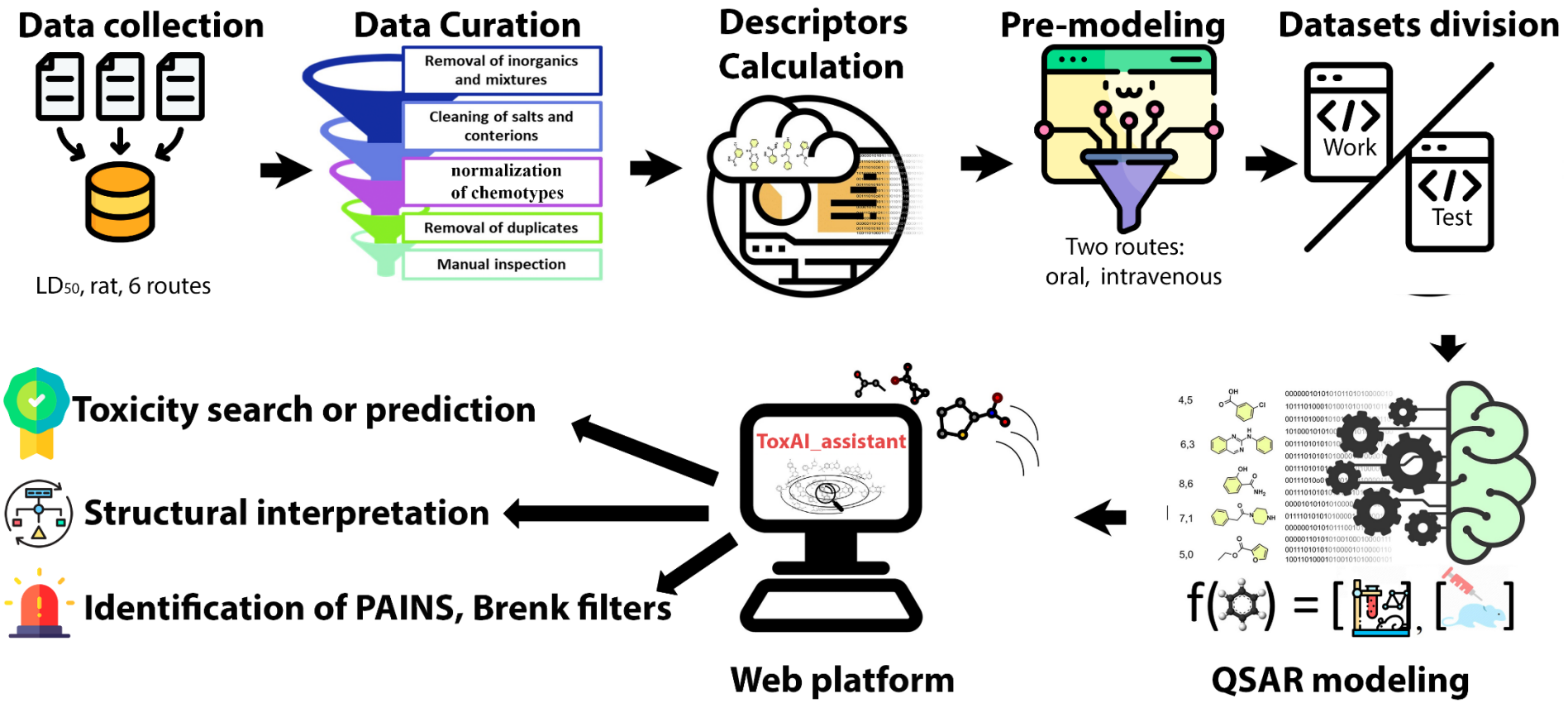
Goals:

The purposes of this work were as follows:

1. to develop regression acute toxicity QSAR models for various methods of administration of toxicants to rats using maximally representative training sets;
2. to integrate the developed QSAR models corresponding to OECD modelling principles into a web application, which is capable of classifying chemical compounds according to the WHO approach based on predicted or experimentally determined LD_{50} values



The main stages of this work:



Materials and methods:

Six data sets containing experimental LD_{50} values for rats with various routes of administration of toxicants were exported from the TOXRIC database (https://toxric.bioinforai.tech/home).

Development tools

Python:

- **RDKit -** library for solving various problems in chemoinformatics https://github.com/rdkit/rdkit-tutorials;
- **PaDELPy –** library for calculating molecular descriptorshttps://github.com/ecrl/padelpy;
- **Scikit-lear –** a library that implements various machine learning methods https://github.com/scikit-learn/scikit-learn;
- **CatBoost** – a library for building models using the gradient boosting method https://catboost.ai/

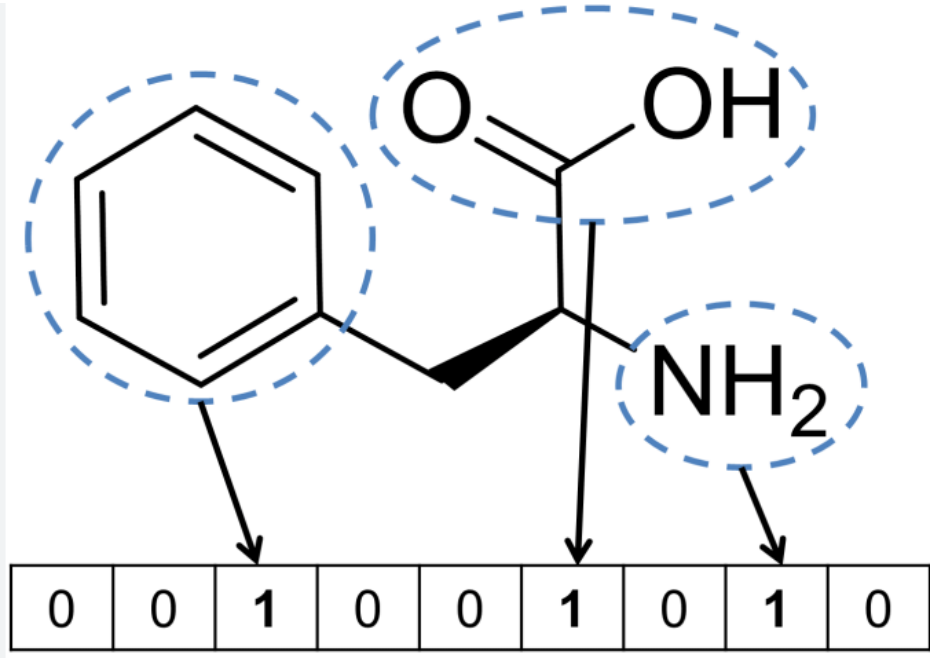
Materials and methods :

QSAR model development methods

- Gradient boosting Regressor (GBR);
- Support Vector Machine (SVM);
- K-nearest neighbors algorithm (k-NN);
- Multi-layer perceptron regressor (MLP)

Molecular descriptors

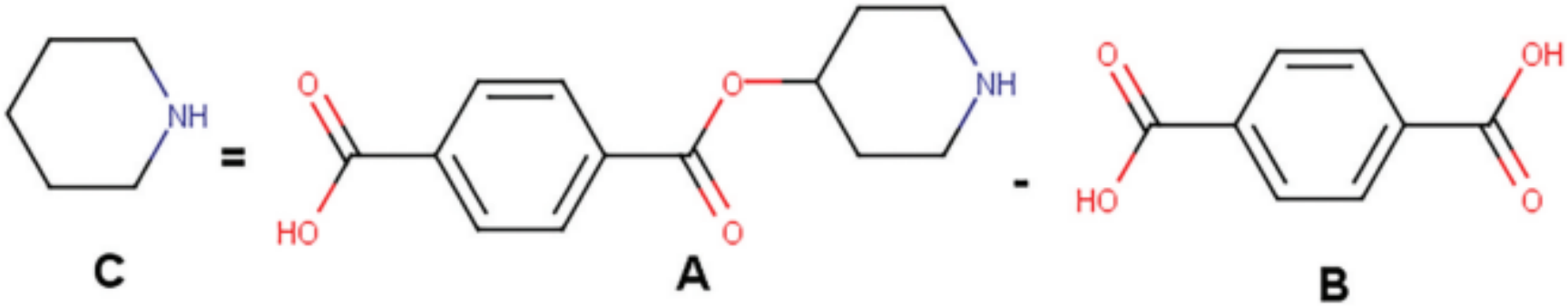
1) Morgan fingerprints (MF);
2) PubChem fingerprints;
3) KlekotaRoth fingerprints;
4) AtomPairs fingerprints
5) MACCS fingerprints;
6) RDKit descriptors



Materials and methods :

The structural interpretation

The structural interpretation was performed according to the approach described in [2], where the contribution of a molecular fragment **(C)** was calculated as the difference between the toxicity values calculated for the parent structure **(A)** and a hypothetical structure **(B)** generated by removing the target fragment **(C)** from **(A)**



$$W(C)=X(A)-X(B)$$

Schematic representation of the structural interpretation approach. W(C) is the contribution of fragment (C); X(A) is the predicted toxicity of the parent structure (A); and X(B) is the predicted toxicity of a hypothetical structure (B).

2. Polishchuk P, Tinkov O, Khristova T, et al. Structural and physico-chemical interpretation (SPCI) of QSAR models and its comparison with matched molecular pair analysis. J Chem Inf Model. 2016;56:1455–1469. doi: 10.1021/acs.jcim.6b00371

Materials and methods :

Criteria for assessing the quality of QSAR models

$$Q^2 = 1 - \frac{\sum_{i} (y_i - \hat{y}_i)^2}{\sum_{i} (y_i - y_{mean})^2}$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{m} (y_i - \hat{y}_i)^2}{m - 1}}$$

where y_i is the observed activity of the ith compound, $\hat{y}_i$ is the activity predicted for the ith compound, *mean* is the mean observed activity, and m is the number of compounds in the set.

Materials and methods :

- Model predictive ability was evaluated and compared with the reference statistical QSAR model indices developed in [3] using 5-fold cross-validation (CV).
- The inclusion of the test set compounds in AD was determined using the similarity distance. A test-set compound is considered to belong to the AD of a QSAR model if its similarity distance does not exceed the threshold value Dc calculated by formula (1)

$$D_c = Z\sigma + \bar{y} \quad (1)$$

where $\bar{y}$ and σ are the mean and the standard deviation, respectively, for the Euclidean distances in the chemical space of descriptors between all objects from the training set and their nearest neighbours in it and Z is a constant usually taken to be 0.5.

Data coverage (Cov) in the AD was calculated as the ratio of the number of compounds from a test set included in the AD to the total number of compounds in the test set,

$$Cov = num\_AD/num\_All \quad (2)$$

where *num_AD* is the number of compounds from the test set included in the AD and *num_All* is the total number of compounds in the test set.

3. Wu L, Yan B, Han J, et al. TOXRIC: a comprehensive database of toxicological data and benchmarks. Nucleic Acids Res. 2023;51:D1432–D1445. doi: 10.1093/nar/gkac1074

Results and discussion

At the first stage, we performed preliminary QSAR modelling. Adequate QSAR models (Q^2_{cv} >= 0.60) were created for only two sets, respectively, with oral and intravenous methods of administration of toxicants.
These sets were then divided into training and test sets for the development, validation, and physicochemical interpretation of QSAR models.

Table 1. Statistical characteristics of the preliminary QSAR models

Administration	The best algorithms and statistical parameters of this study				Reference statistical parameters [3]	
	Descriptors	Method	Q^2_{cv}	RMSE	Q^2_{cv}	RMSE
oral	MACCS	CatBoost	**0.60**	0.57	0.59	0.59
intraperitoneal	RDKit	CatBoost	0.51	0.60	0.53	0.62
intravenous	Morgan fingerprints	CatBoost	**0.61**	0.62	0.64	0.64
skin	RDKit	CatBoost	0.37	0.95	0.33	0.98
intramuscular	RDKit	CatBoost	0.34	0.89	0.43	0.90
subcutaneous	RDKit	CatBoost	0.48	0.73	0.48	0.73

3. Wu L, Yan B, Han J, et al. TOXRIC: a comprehensive database of toxicological data and benchmarks. Nucleic Acids Res. 2023;51:D1432–D1445. doi: 10.1093/nar/gkac1074

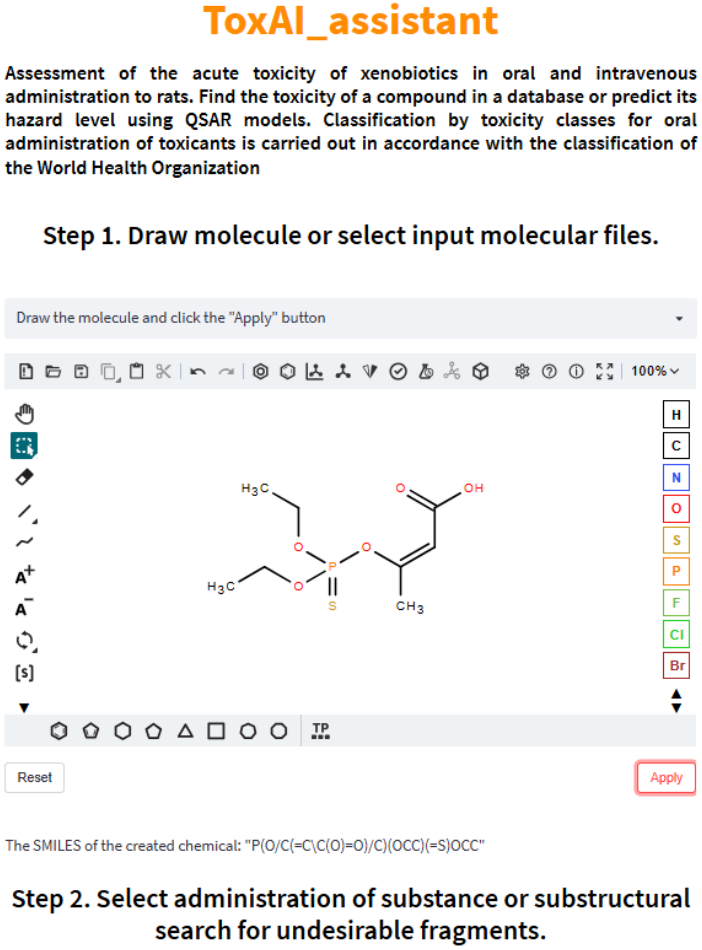
Results and discussion

For both administration routes, the models developed using the RDKit descriptors and the CatBoost method exhibit the best combination of descriptive and predictive abilities and the largest data coverage in the AD.

Table 2. Statistical characteristics of the developed QSAR models for oral and intravenous administration in rats

Administration	Training set, 5-fold CV		Test set				
			All compounds		Cov	Compounds in AD	
	Q^2_{cv}	RMSE	Q^2_{ts}	RMSE		Q^2_{ts}	RMSE
oral	0.58	0.58	0.60	0.57	0.78	0.66	0.53
intravenous	0.62	0.61	0.54	0.67	0.79	0.66	0.56

ToxAI_assistant application



https://tox-ai-assistant.streamlit.app/.

- created using the Streamlit framework;
- Information about the chemical structure of the studied compounds can be entered using a chemical editor, linear SMILES notation, or CSV or SDF format files;
- automatic verification and standardization of investigated chemical structures using the MolVS library.

Streamlit

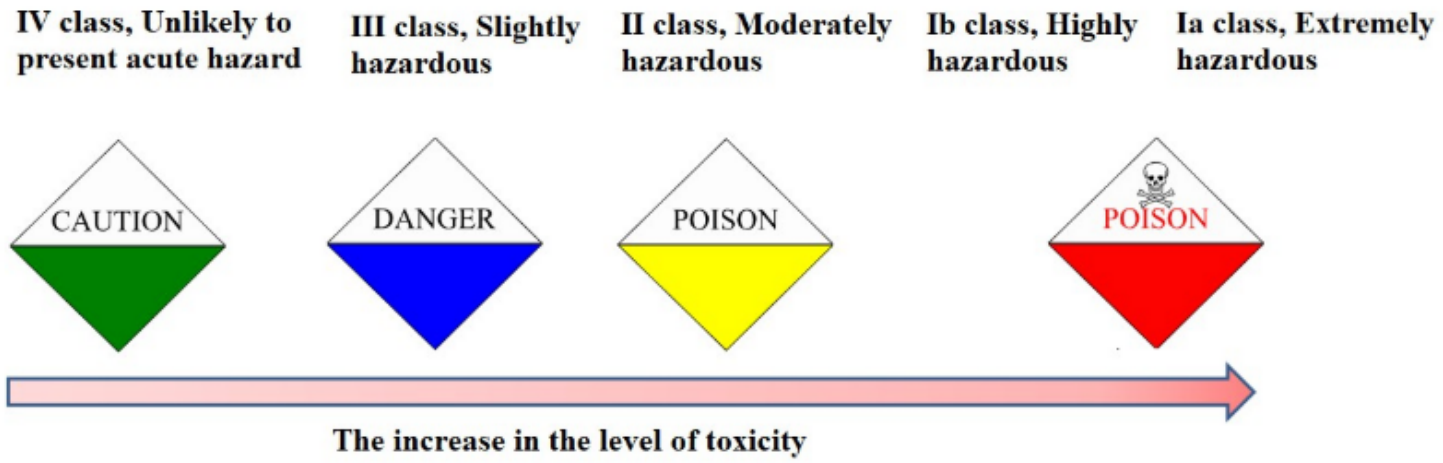
Results of the " ToxAI_assistant" web application

For chemical structures that have successfully passed verification and standardization, the presence of experimental LD_{50} values is checked for the selected method of administration. If the test compound has experimental data, the corresponding LD_{50} value is displayed in the web application, as well as the CAS identification number of the substance. In this case, no prediction of toxicity is carried out.

Prediction results:

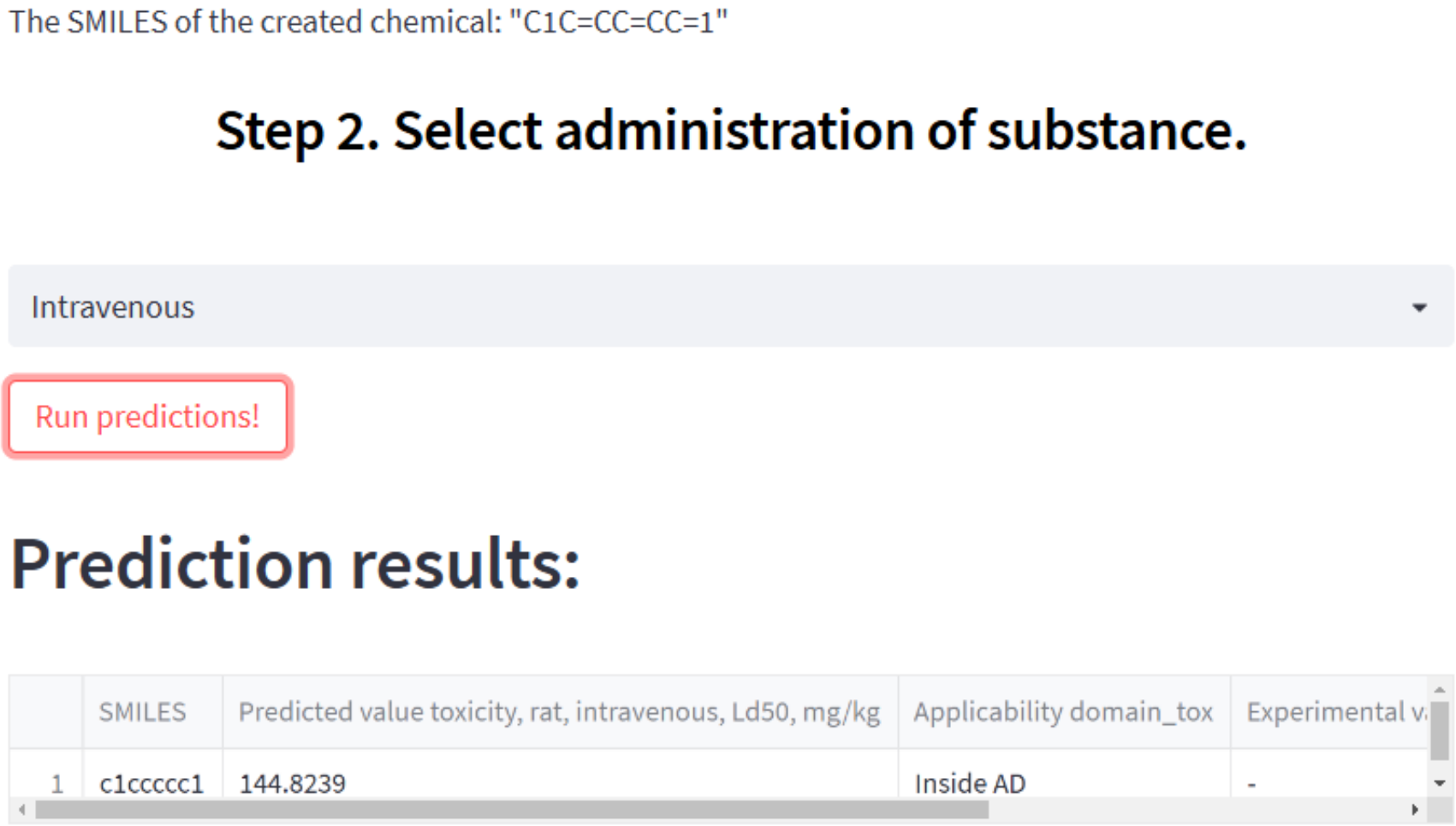
	SMILES	Predicted value toxicity, rat, oral, Ld50, mg/kg	Applicability domain_tox	Experimental value tox
1	c1ccccc1	see experimental value	-	929.2017

Classification into toxicity classes (see column "Hazard_Categories") is carried out in accordance with the classification of the World Health Organization (https://www.who.int/publications/i/item/9789240005662)



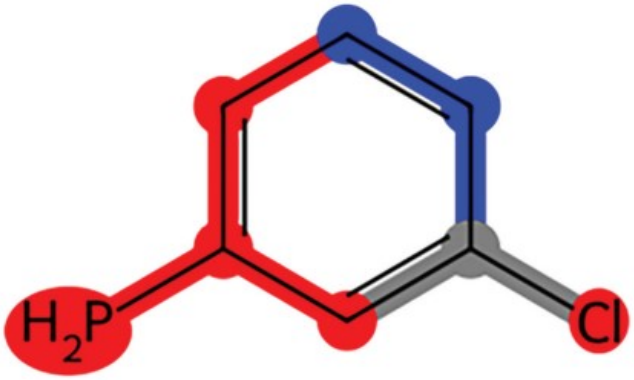
Results of the " ToxAI_assistant" web application

In the absence of experimental data for the studied compound, the LD_{50} value is predicted using the above-mentioned QSAR models for oral and intravenous administration of toxicants.



Results of the " ToxAI_assistant" web application

In addition, when predicting the level of toxicity for individual compounds for both oral and intravenous administration of toxicants, the ToxAI_assistant web application determines the contributions of structural fragments to the acute toxicity; that is, it is visually possible to assess which atoms increase or decrease toxicity.



Structural interpretation: fragments increasing toxicity are red, fragments decreasing toxicity are blue, and neutral fragments are grey.

The found tox_alerts_subst: phosphorothionate

Tanimoto coefficient: 0.14

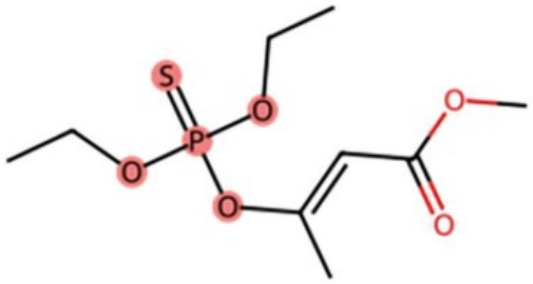


Table 3 Comparison of the main functions of the ToxAI_assistant web application with other freely available programs used to predict acute toxicity following oral administration in rats

Comparison parameter	TEST	ADMETlab 3.0	STopTox	ToxAI_assistant
Implementation form	Desktop application	Web application	Web application	Web application
Type of implemented QSAR models	Regression	Binary classification	Binary classification	Regression
Total number of compounds for QSAR modelling	7413	7327	8495	9843
Type of molecular descriptor	Chemistry Development Kit (CDK)	RDKit 2D	Morgan, MACCS, and Mordred	RDKit 2D
Machine learning methods	Consensus (Hierarchical clustering+ Nearest neighbor)	Directed Message Passing Neural Network (MPNN)	Random forest	Gradient Boosting (CatBoost)
Structural interpretation	no	no	yes	yes
Detection of common medicinal chemistry filters (PAINS, Brenk filters, structural alerts)	no	yes	no	yes
Preliminary check for experimental toxicity	no	no	no	yes

The developed QSAR models, their construction algorithms, statistical characteristics in the form of Jupyter Notebook program files and the program code of the ToxAI_assistant web application are freely available at https://github.com/ovttiras/ToxAI_assistant and can be used for virtual screening

Conclusions:

- ToxAI_assistant significantly advances the state of acute toxicity modelling;
- By combining a very large LD_{50} dataset with regression-based QSAR modelling, WHO-derived hazard categories, and substructural interpretation, we have filled a major gap in this field;
- All models satisfy the OECD validation principles;

Co-author

Veniamin Yurievich Grigorev,
Doctor of Chemical Sciences

Institute of Physiologically Active Compounds at Federal Research Center of Problems of Chemical Physics and Medicinal Chemistry, Russian Academy of Sciences, Chernogolovka, Russia

Funding

Part of this work was supported by the budget of the Institute of Physiologically Active Compounds of the Russian Academy of Sciences (IPAC RAS) State Targets – 2024 [topic No. FFSG-2024–0019].

THANKS FOR THE ATTENTION!