**Multi-task Cascade Forest Framework for Predicting Acute Toxicity Across Species**

**Supplementary Materials**

**Without author info**

#### Section A Information of multi-species acute toxicity dataset

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Table A1 Information on the multi-species acute toxicity dataset** | | | | |
| ID | Species | Administration | Toxicity Type | Size |
| 01 | bird-wild | oral | LD50 | 338 |
| 02 | chicken | oral | LD50 | 353 |
| 03 | duck | oral | LD50 | 192 |
| 04 | quail | oral | LD50 | 352 |
| 05 | frog | subcutaneous | LDLo | 112 |
| 06 | mammal (species unspecified) | intraperitoneal | LD50 | 545 |
| 07 | mammal (species unspecified) | oral | LD50 | 674 |
| 08 | mammal (species unspecified) | subcutaneous | LD50 | 125 |
| 09 | mammal (species unspecified) | unreported | LD50 | 1129 |
| 10 | guinea pig | intraperitoneal | LD50 | 248 |
| 11 | guinea pig | intravenous | LD50 | 153 |
| 12 | guinea pig | intravenous | LDLo | 121 |
| 13 | guinea pig | oral | LD50 | 793 |
| 14 | guinea pig | skin | LD50 | 176 |
| 15 | guinea pig | subcutaneous | LD50 | 169 |
| 16 | guinea pig | subcutaneous | LDLo | 179 |
| 17 | mouse | intramuscular | LD50 | 571 |
| 18 | mouse | intraperitoneal | LD50 | 36295 |
| 19 | mouse | intraperitoneal | LDLo | 266 |
| 20 | mouse | intravenous | LD50 | 16978 |
| 21 | mouse | intravenous | LDLo | 102 |
| 22 | mouse | oral | LD50 | 23373 |
| 23 | mouse | oral | LDLo | 264 |
| 24 | mouse | parenteral | LD50 | 302 |
| 25 | mouse | skin | LD50 | 214 |
| 26 | mouse | subcutaneous | LD50 | 6769 |
| 27 | mouse | subcutaneous | LDLo | 252 |
| 28 | mouse | unreported | LD50 | 1739 |
| 29 | rat | intramuscular | LD50 | 300 |
| 30 | rat | intraperitoneal | LD50 | 5021 |
| 31 | rat | intraperitoneal | LDLo | 318 |
| 32 | rat | intravenous | LD50 | 2472 |
| 33 | rat | intravenous | LDLo | 135 |
| 34 | rat | oral | LD50 | 10190 |
| 35 | rat | oral | LDLo | 322 |
| 36 | rat | skin | LD50 | 835 |
| 37 | rat | subcutaneous | LD50 | 1896 |
| 38 | rat | subcutaneous | LDLo | 174 |
| 39 | rat | unreported | LD50 | 806 |
| 40 | cat | intravenous | LD50 | 261 |
| 41 | cat | intravenous | LDLo | 159 |
| 42 | cat | oral | LD50 | 171 |
| 43 | cat | oral | LDLo | 142 |
| 44 | rabbit | intraperitoneal | LD50 | 131 |
| 45 | rabbit | intravenous | LD50 | 792 |
| 46 | rabbit | intravenous | LDLo | 346 |
| 47 | rabbit | oral | LD50 | 894 |
| 48 | rabbit | oral | LDLo | 249 |
| 49 | rabbit | skin | LD50 | 1495 |
| 50 | rabbit | skin | LDLo | 181 |
| 51 | rabbit | subcutaneous | LD50 | 156 |
| 52 | rabbit | subcutaneous | LDLo | 241 |
| 53 | dog | intravenous | LD50 | 468 |
| 54 | dog | intravenous | LDLo | 360 |
| 55 | dog | oral | LD50 | 649 |
| 56 | dog | oral | LDLo | 187 |
| 57 | human | oral | TDLo | 140 |
| 58 | woman | oral | TDLo | 156 |
| 59 | man | oral | TDLo | 163 |
| **Total sample size** | | | | **122594** |

**Note:** Lethal Dose Low (LDLo), Lethal Dose 50 (LD50), Toxic Does Low (TDLo).

1. **Data Preprocessing**

The multi-species acute toxicity data involved in this study were derived from the published literature [1]. Data preprocessing followed the approach described by Jain et al.[1], which involved data extraction, unit standardization, and compound screening.

Firstly, Jain et al. extracted the acute toxicity measurements from the publicly available ChemIDplus database [2], which is a key chemical database under the Toxicology Data Network (TOXNET [3]) of the United States National Library of Medicine. ChemIDplus consolidates acute toxicity data and related compound information from multiple sources, including median lethal dose (LD50), lethal dose low (LDLo), and toxic dose low (TDLo).

Secondly, various types of measurements were standardized to units of mg/kg, ug/kg, or ng/kg, and all values were ultimately converted to mg/kg. To mitigate the magnitude differences, the toxicity values were further converted into dimensionless values of −log(mol/kg). This logarithmic transformation not only compresses the dynamic range of the data but also effectively reduces the impact of outliers on statistical analysis through molar concentration normalization.

Thirdly, compounds in the dataset were filtered and standardized. This process involved removing salts, solvents, counterions, large organic compounds (molecular weight ≥ 2000 Da), mixtures, and inorganic compounds, as well as standardizing specific chemical structures (such as aromatic rings, nitro groups, etc.).

To ensure data quality, toxicity endpoints with fewer than 100 measurements were excluded, resulting in a final dataset containing 80,081 unique compounds and 122,594 measurements across 59 toxicity endpoints. Table 1 of supplementary materials provides an overview of the data, detailing the species, administration modes, toxicity types, and sample sizes.

1. **Analysis of Missing Values and Outliers**

The feature space of the dataset expressed the chemical structure of compounds, and the label space expressed the toxicity values of these compounds at different toxicity endpoints. The dataset containing 80,081 unique compounds and 122,594 measurements across 59 toxicity endpoints, where one measurement corresponds to one sample.

The feature representation of the compounds is calculated using specific methods without missing values, which is described at lines 243-263 of the manuscript. However, not all compounds can be measured at 59 toxicity endpoints. In other words, if complete measurements are conducted, the total number of samples for 59 toxicity endpoints should be 80,081\*59. But for some toxicity endpoints, it is impossible to measure the toxicity of a large number of compounds, such as human oral toxicity (endpoints 57-59 of Table 1). Therefore, the toxicity values (label) in the dataset exhibit a high degree of sparsity with a substantial number of missing entries. More specifically, there are significant differences in the number of measurements for each toxicity endpoint. To handle the aforementioned situation, we employ multi-task learning to model each endpoint individually, which helps to alleviate the adverse effects caused by label sparsity during the modeling process.

We visualize the distribution of the labels. The left histogram in Figure A1 shows that the trend of label values follows a normal distribution. The right box plot in Figure 1 demonstrates that there are a few outliers beyond the upper and lower bounds of the interquartile range. However, we believe that a small number of outliers caused by non-measurement errors in large datasets should be retained to restore the authenticity of the data.

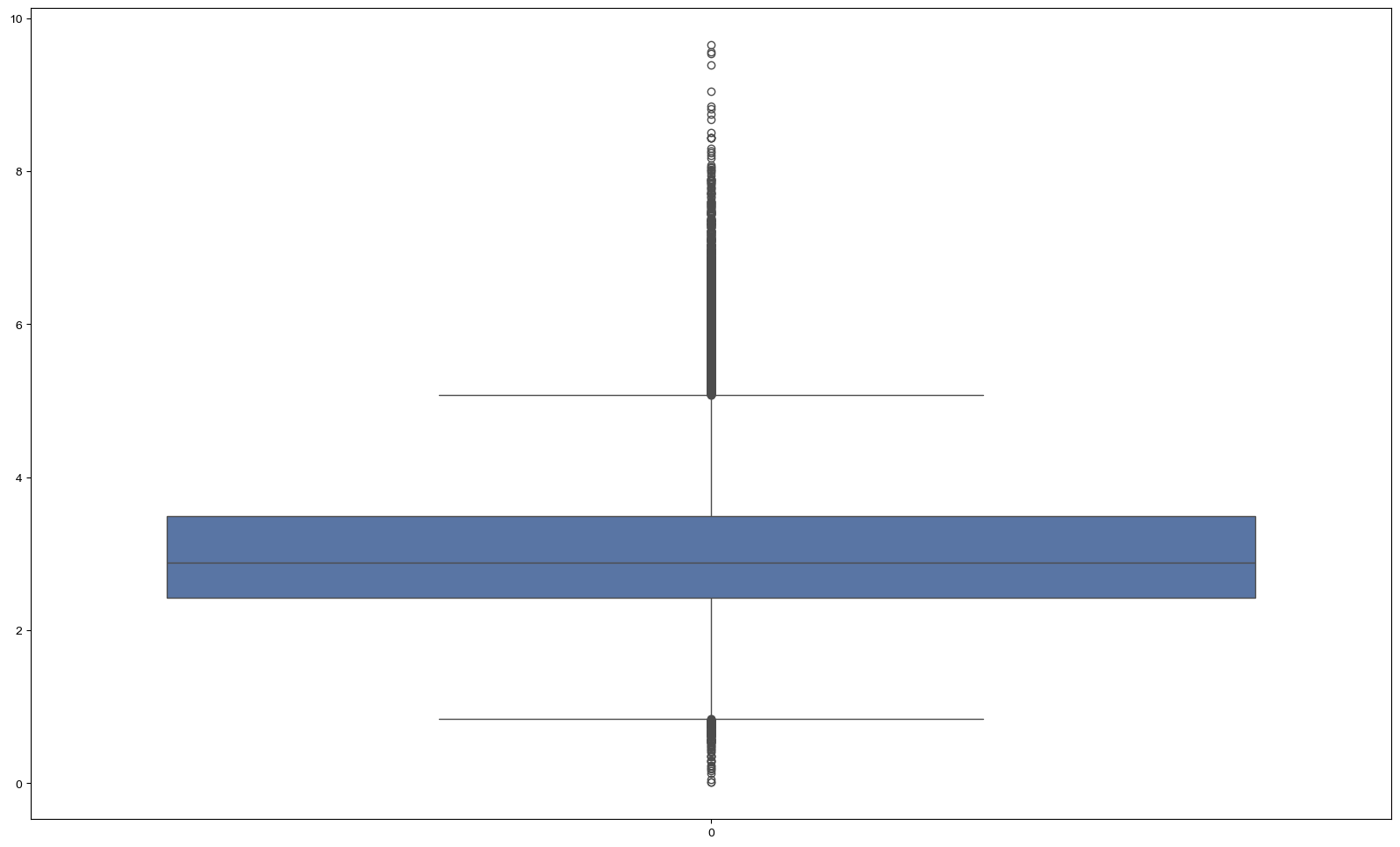
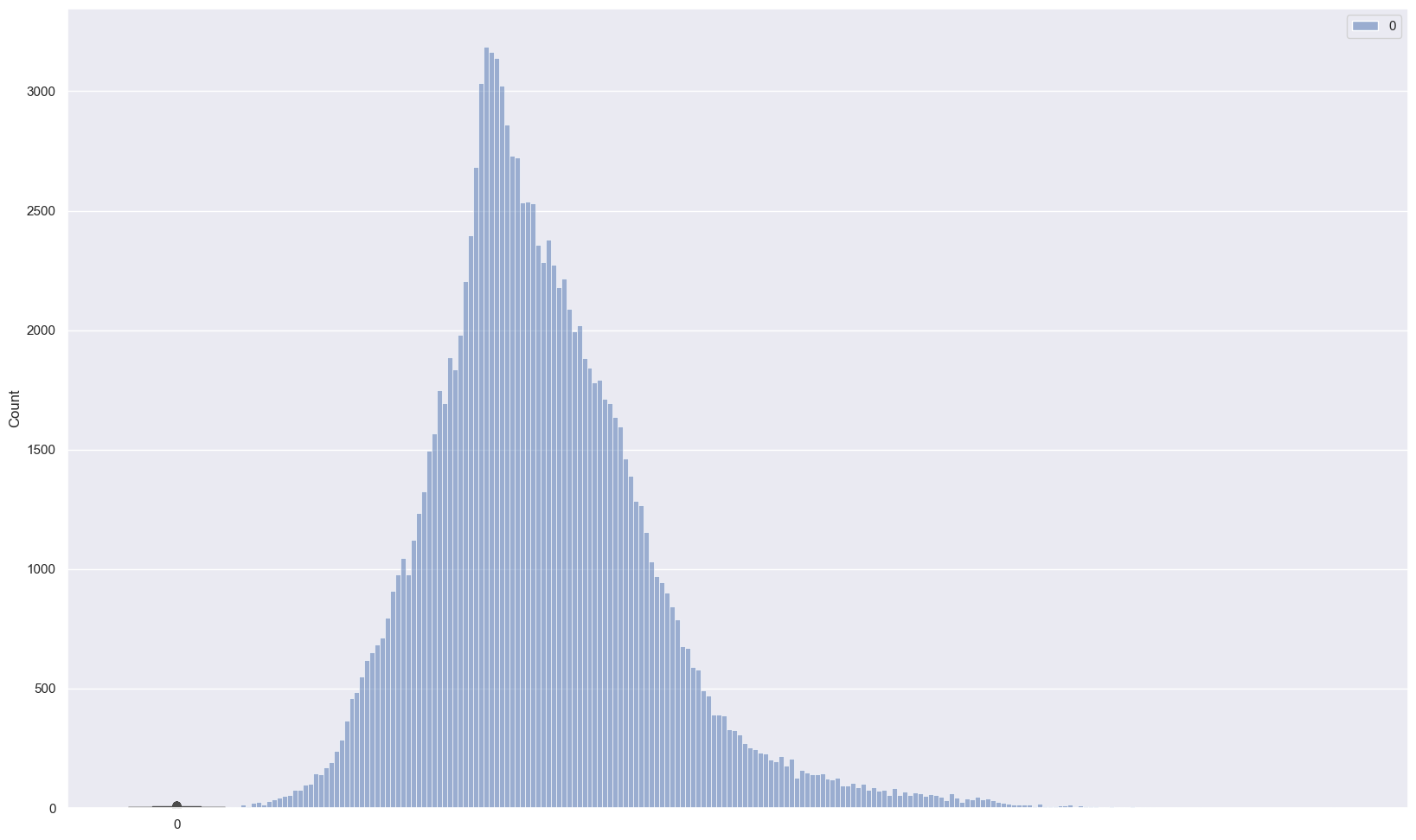


Fig.A1. The distribution of label values.

#### Section B Algorithm flowchart

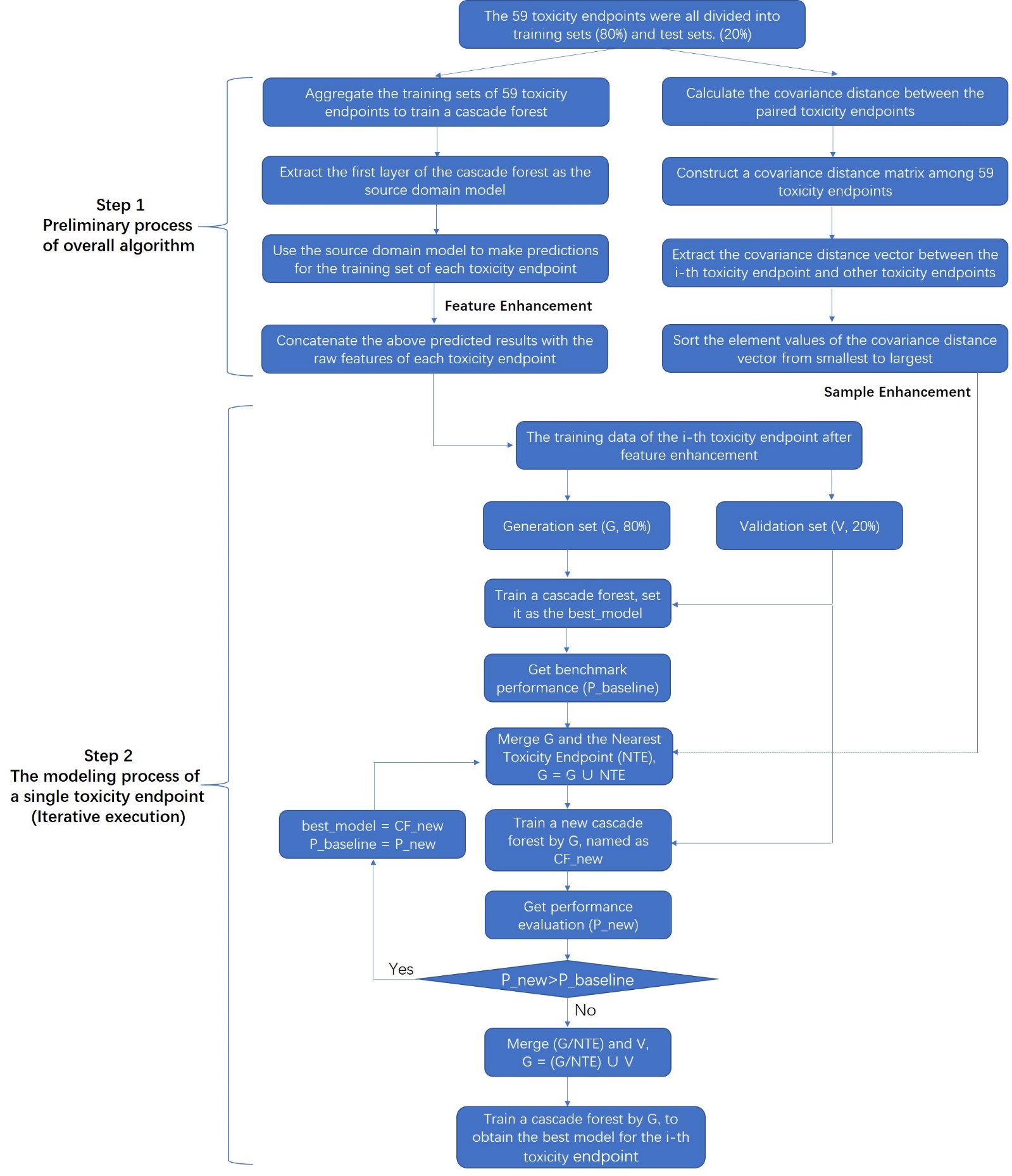


Fig.B1. The flowchart of the proposed algorithm.

#### Section C Parameter configuration

|  |  |  |
| --- | --- | --- |
| **Table C1 Parameter configuration of all algorithms** | | |
| **Model architectures for MT-DNN** | **Parameter** | **Value** |
|  | number of hidden layers | 3 |
|  | number of neurons for each layer | 2000, 700, 500 |
|  | activation function | ReLu |
|  | optimizer | Adam |
|  | dense\_size\_candidates | 1500, 500, 100, 1 |
|  | learning rate (Adam) | 0.0001 |
|  | number of epochs | 50 |
|  | batch size | 128 |
| **Model architectures for DLCA** | **Parameter** | **Value** |
| Descriptor 1-3 (Fingerprints) | number of hidden layers | 4 |
|  | number of neurons for each layer | 8000, 2000, 1000, 700 |
| Descriptor 4  (RDKit physicochemical properties) | number of hidden layers | 4 |
|  | number of neurons for each layer | 8000, 2000, 1000, 700 |
|  | dropout | 0.3 |
| Descriptor 5 (SMILES) | architecture | 1D convolutional |
|  | Pooling | GlobalMaxPooling1D |
|  | number of hidden layers | 1 |
|  | number of neurons for each layer | 200 |
|  | optimizer | Adagrad |
|  | learning rate | 0.01 |
| Common configuration for above 5 networks | activation function | relu |
|  | batch size | 128 |
|  | number of epochs | 20 |
| **Model architectures for MT-GCNN** | **Parameter** | **Value** |
|  | depth | 5 |
|  | dropout | 0.3 |
|  | ffn\_num\_layers | 3 |
|  | hidden\_size | 1700 |
| **Model architectures for Consen A** | **Parameter** | **Value** |
|  | the ensemble of MT-DNN and MT-GCNN | / |
| **Model architectures for Consen B** | **Parameter** | **Value** |
|  | the ensemble of MT-DNN, MT-GCNN, and DLCA | / |
| **Model architectures for TranAda** | **Parameter** | **Value** |
|  | estimator type | Ridge regression |
|  | estimator number | 10 |
|  | learning rate | 1 |
| **Model architectures for LinInt** | **Parameter** | **Value** |
|  | estimator type | Ridge regression |
|  | proportion | 0.5 |
| **Model architectures for CFF-AT** | **Parameter** | **Value** |
|  | base framework | Cascade forest regression |
|  | estimator number in each layer | 4 |
|  | estimator type in each layer | RandomForest\*2, ExtraTrees\*2 |
|  | layer number | self-adaptive |
|  | task dianstance metric | covariance |

#### Section D Performance comparison

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Table D1 Performance comparison of all algorithms on 59 toxicity endpoints based on R2** | | | | | | | | |
| **Simplified Task Name** | **MT-DNN** | **MT-GCNN** | **DLCA** | **ConsenA** | **ConsenB** | **LinInt** | **TranAda** | **CFF-AT** |
| 1.bird-wild\_oral\_LD50 | 0.38 (6.5) | 0.41 (5.0) | 0.46 (3.0) | 0.45 (4.0) | 0.47 (2.0) | 0.38 (6.5) | 0.17 (8.0) | 0.57 (1.0) |
| 2.chicken\_oral\_LD50 | 0.36 (5.0) | 0.31 (8.0) | 0.45 (2.0) | 0.41 (4.0) | 0.43 (3.0) | 0.33 (6.0) | 0.32 (7.0) | 0.56 (1.0) |
| 3.duck\_oral\_LD50 | 0.48 (5.0) | 0.44 (6.0) | 0.55 (2.0) | 0.52 (4.0) | 0.54 (3.0) | 0.33 (8.0) | 0.38 (7.0) | 0.68 (1.0) |
| 4.quail\_oral\_LD50 | 0.46 (5.0) | 0.44 (6.0) | 0.51 (3.0) | 0.49 (4.0) | 0.52 (2.0) | 0.35 (8.0) | 0.41(7.0) | 0.61 (1.0) |
| 5.frog\_subc\_LDLo | 0.54 (4.5) | 0.56 (3.0) | 0.54 (4.5) | 0.58 (1.5) | 0.58 (1.5) | 0.35 (7.0) | 0.32 (8.0) | 0.39 (6.0) |
| 6.mammal\_intrap\_LD50 | 0.64 (2.0) | 0.54 (6.0) | 0.59 (5.0) | 0.63 (4.0) | 0.64 (2.0) | 0.34 (7.0) | 0.32 (8.0) | 0.64 (2.0) |
| 7.mammal\_oral\_LD50 | 0.63 (2.5) | 0.54 (6.0) | 0.57 (5.0) | 0.63 (2.5) | 0.64 (1.0) | 0.32 (7.0) | 0.28 (8.0) | 0.59 (4.0) |
| 8.mammal\_subc\_LD50 | 0.20 (5.0) | 0.28 (4.0) | 0.18 (6.0) | 0.36 (2.0) | 0.32 (3.0) | 0.13 (7.0) | 0.04 (8.0) | 0.41 (1.0) |
| 9.mammal\_unrep\_LD50 | 0.50 (4.0) | 0.45 (6.0) | 0.49 (5.0) | 0.53 (3.0) | 0.54 (2.0) | 0.25 (8.0) | 0.27 (7.0) | 0.62 (1.0) |
| 10.gp\_intrap\_LD50 | 0.57 (3.0) | 0.49 (6.0) | 0.56 (4.0) | 0.55 (5.0) | 0.58 (2.0) | 0.45 (7.0) | 0.40 (8.0) | 0.72 (1.0) |
| 11.gp\_intrav\_LD50 | 0.62 (5.0) | 0.60 (6.0) | 0.66 (3.5) | 0.66 (3.5) | 0.67 (1.5) | 0.56 (7.0) | 0.21 (8.0) | 0.67 (1.5) |
| 12.gp\_intrav\_LDLo | 0.61 (6.0) | 0.57 (7.0) | 0.66 (2.5) | 0.62 (4.5) | 0.66 (2.5) | 0.62 (4.5) | 0.51 (8.0) | 0.75 (1.0) |
| 13.gp\_oral\_LD50 | 0.61 (6.0) | 0.63 (5.0) | 0.64 (4.0) | 0.67 (2.5) | 0.67 (2.5) | 0.45 (8.0) | 0.47 (7.0) | 0.76 (1.0) |
| 14.gp\_skin\_LD50 | 0.33 (5.0) | 0.32 (6.0) | 0.40 (3.0) | 0.37 (4.0) | 0.42 (2.0) | 0.21 (8.0) | 0.25 (7.0) | 0.48 (1.0) |
| 15.gp\_subc\_LD50 | 0.58 (3.5) | 0.48 (6.0) | 0.56 (5.0) | 0.58 (3.5) | 0.59 (2.0) | 0.41 (8.0) | 0.43 (7.0) | 0.67 (1.0) |
| 16.gp\_subc\_LDLo | 0.65 (5.5) | 0.68 (4.0) | 0.65 (5.5) | 0.70 (2.0) | 0.69 (3.0) | 0.56 (8.0) | 0.61 (7.0) | 0.81 (1.0) |
| 17.mouse\_intram\_LD50 | 0.58 (6.0) | 0.61 (5.0) | 0.63 (4.0) | 0.64 (3.0) | 0.66 (2.0) | 0.45 (7.0) | 0.41 (8.0) | 0.71 (1.0) |
| 18.mouse\_intrap\_LD50 | 0.59 (5.0) | 0.57 (6.0) | 0.63 (3.5) | 0.63 (3.5) | 0.65 (2.0) | 0.33 (7.0) | -0.46 (8.0) | 0.66 (1.0) |
| 19.mouse\_intrap\_LDLo | 0.46 (6.0) | 0.49 (5.0) | 0.53 (2.5) | 0.53 (2.5) | 0.55 (1.0) | 0.27 (7.0) | 0.18 (8.0) | 0.52 (4.0) |
| 20.mouse\_intrav\_LD50 | 0.57 (6.0) | 0.58 (5.0) | 0.64 (2.0) | 0.63 (3.0) | 0.65 (1.0) | 0.39 (7.0) | -0.31 (8.0) | 0.61 (4.0) |
| 21.mouse\_intrav\_LDLo | 0.53 (5.0) | 0.49 (6.0) | 0.58 (2.5) | 0.56 (4.0) | 0.58 (2.5) | 0.06 (8.0) | 0.13 (7.0) | 0.62 (1.0) |
| 22.mouse\_oral\_LD50 | 0.51 (5.0) | 0.50 (6.0) | 0.55 (4.0) | 0.56 (3.0) | 0.58 (2.0) | 0.27 (7.0) | -0.53 (8.0) | 0.59 (1.0) |
| 23.mouse\_oral\_LDLo | 0.34 (6.0) | 0.36 (4.0) | 0.35 (5.0) | 0.39 (2.0) | 0.40 (1.0) | 0.15 (7.0) | 0.11 (8.0) | 0.38 (3.0) |
| 24.mouse\_parent\_LD50 | 0.46 (4.0) | 0.44 (5.5) | 0.44 (5.5) | 0.53 (1.0) | 0.52 (2.0) | 0.33 (8.0) | 0.41 (7.0) | 0.51 (3.0) |
| 25.mouse\_skin\_LD50 | 0.36 (6.0) | 0.49 (2.0) | 0.38 (5.0) | 0.46 (4.0) | 0.47 (3.0) | 0.29 (7.0) | 0.15 (8.0) | 0.52 (1.0) |
| 26.mouse\_subc\_LD50 | 0.59 (5.0) | 0.58 (6.0) | 0.62 (4.0) | 0.63 (3.0) | 0.65 (2.0) | 0.41 (7.0) | 0.14 (8.0) | 0.71 (1.0) |
| 27.mouse\_subc\_LDLo | 0.31 (5.0) | 0.36 (4.0) | 0.30 (6.0) | 0.38 (2.5) | 0.38 (2.5) | 0.23 (7.0) | 0.20 (8.0) | 0.61 (1.0) |
| 28.mouse\_unrep\_LD50 | 0.68 (5.0) | 0.63 (6.0) | 0.70 (4.0) | 0.71 (3.0) | 0.72 (2.0) | 0.47 (7.5) | 0.47 (7.5) | 0.74 (1.0) |
| 29.rat\_intram\_LD50 | 0.57 (6.0) | 0.61 (4.0) | 0.58 (5.0) | 0.63 (3.0) | 0.64 (2.0) | 0.37 (8.0) | 0.44 (7.0) | 0.69 (1.0) |
| 30.rat\_intrap\_LD50 | 0.56 (5.5) | 0.56 (5.5) | 0.59 (4.0) | 0.60 (3.0) | 0.62 (2.0) | 0.34 (7.0) | 0.04 (8.0) | 0.72 (1.0) |
| 31.rat\_intrap\_LDLo | 0.53 (6.0) | 0.58 (3.5) | 0.57 (5.0) | 0.58 (3.5) | 0.62 (1.0) | 0.31 (7.0) | 0.29 (8.0) | 0.59 (2.0) |
| 32.rat\_intrav\_LD50 | 0.67 (5.0) | 0.65 (6.0) | 0.69 (4.0) | 0.70 (3.0) | 0.71 (2.0) | 0.44 (7.0) | 0.25 (8.0) | 0.77 (1.0) |
| 33.rat\_intrav\_LDLo | 0.48 (4.0) | 0.37 (5.0) | 0.55 (1.0) | 0.50 (3.0) | 0.52 (2.0) | 0.18 (7.0) | 0.01 (8.0) | 0.46 (5.0) |
| 34.rat\_oral\_LD50 | 0.60 (6.0) | 0.61 (5.0) | 0.64 (4.0) | 0.65 (3.0) | 0.66 (2.0) | 0.41 (7.0) | 0.01 (8.0) | 0.69 (1.0) |
| 35.rat\_oral\_LDLo | 0.37 (5.0) | 0.33 (6.0) | 0.44 (2.5) | 0.42 (4.0) | 0.44 (2.5) | 0.26 (7.0) | 0.15 (8.0) | 0.49 (1.0) |
| 36.rat\_skin\_LD50 | 0.36 (6.0) | 0.41 (5.0) | 0.43 (4.0) | 0.44 (3.0) | 0.46 (2.0) | 0.27 (7.0) | -0.06 (8.0) | 0.51 (1.0) |
| 37.rat\_subc\_LD50 | 0.53 (5.5) | 0.53 (5.5) | 0.58 (3.5) | 0.58 (3.5) | 0.60 (2.0) | 0.36 (7.0) | 0.27 (8.0) | 0.73 (1.0) |
| 38.rat\_subc\_LDLo | 0.68 (2.5) | 0.65 (5.0) | 0.68 (2.5) | 0.67 (4.0) | 0.71 (1.0) | 0.43 (7.5) | 0.43 (7.5) | 0.63 (6.0) |
| 39.rat\_unrep\_LD50 | 0.56 (6.0) | 0.57 (5.0) | 0.60 (4.0) | 0.61 (3.0) | 0.62 (2.0) | 0.42 (7.5) | 0.42 (7.5) | 0.72 (1.0) |
| 40.cat\_intrav\_LD50 | 0.63 (6.0) | 0.65 (4.5) | 0.65 (4.5) | 0.67 (2.5) | 0.67 (2.5) | 0.61 (7.0) | 0.40 (8.0) | 0.75 (1.0) |
| 41.cat\_intrav\_LDLo | 0.71 (5.0) | 0.66 (7.0) | 0.72 (3.0) | 0.72 (3.0) | 0.73 (1.0) | 0.67 (6.0) | 0.34 (8.0) | 0.72 (3.0) |
| 42.cat\_oral\_LD50 | 0.47 (7.0) | 0.46 (8.0) | 0.56 (4.5) | 0.52 (6.0) | 0.56 (4.5) | 0.62 (2.0) | 0.59 (3.0) | 0.71 (1.0) |
| 43.cat\_oral\_LDLo | 0.41 (4.5) | 0.39 (6.0) | 0.51 (1.0) | 0.45 (3.0) | 0.50 (2.0) | 0.21 (8.0) | 0.22 (7.0) | 0.41 (4.5) |
| 44.rabbit\_intrap\_LD50 | 0.43 (6.0) | 0.51 (3.5) | 0.48 (5.0) | 0.51 (3.5) | 0.53 (2.0) | 0.28 (7.0) | 0.11 (8.0) | 0.64 (1.0) |
| 45.rabbit\_intrav\_LD50 | 0.72 (4.5) | 0.71 (6.0) | 0.72 (4.5) | 0.74 (3.0) | 0.75 (2.0) | 0.54 (7.0) | 0.45 (8.0) | 0.80 (1.0) |
| 46.rabbit\_intrav\_LDLo | 0.61 (5.0) | 0.59 (6.0) | 0.63 (3.0) | 0.64 (2.0) | 0.66 (1.0) | 0.49 (7.0) | 0.15 (8.0) | 0.62 (4.0) |
| 47.rabbit\_oral\_LD50 | 0.56 (5.5) | 0.56 (5.5) | 0.57 (4.0) | 0.59 (3.0) | 0.61 (2.0) | 0.33 (7.0) | 0.30 (8.0) | 0.69 (1.0) |
| 48.rabbit\_oral\_LDLo | 0.57 (3.0) | 0.49 (6.0) | 0.53 (5.0) | 0.55 (4.0) | 0.58 (2.0) | 0.41 (7.5) | 0.41 (7.5) | 0.60 (1.0) |
| 49.rabbit\_skin\_LD50 | 0.43 (6.0) | 0.44 (5.0) | 0.47 (4.0) | 0.48 (3.0) | 0.50 (1.5) | 0.25 (7.0) | -0.01 (8.0) | 0.50 (1.5) |
| 50.rabbit\_skin\_LDLo | 0.47 (1.0) | 0.31 (6.0) | 0.45 (4.0) | 0.46 (2.5) | 0.46 (2.5) | 0.29 (7.0) | 0.03 (8.0) | 0.41 (5.0) |
| 51.rabbit\_subc\_LD50 | 0.62 (4.0) | 0.59 (5.0) | 0.56 (6.5) | 0.66 (2.0) | 0.65 (3.0) | 0.43 (8.0) | 0.56 (6.5) | 0.72 (1.0) |
| 52.rabbit\_subc\_LDLo | 0.73 (5.0) | 0.76 (4.0) | 0.77 (3.0) | 0.78 (2.0) | 0.79 (1.0) | 0.51 (8.0) | 0.55 (7.0) | 0.71 (6.0) |
| 53.dog\_intrav\_LD50 | 0.64 (6.0) | 0.67 (4.0) | 0.66 (5.0) | 0.69 (3.0) | 0.70 (2.0) | 0.49 (7.0) | 0.20 (8.0) | 0.75 (1.0) |
| 54.dog\_intrav\_LDLo | 0.62 (5.0) | 0.61 (6.0) | 0.64 (4.0) | 0.66 (3.0) | 0.67 (2.0) | 0.48 (7.0) | 0.16 (8.0) | 0.72 (1.0) |
| 55.dog\_oral\_LD50 | 0.38 (6.0) | 0.39 (5.0) | 0.44 (3.5) | 0.44 (3.5) | 0.47 (2.0) | 0.23 (7.0) | 0.17 (8.0) | 0.54 (1.0) |
| 56.dog\_oral\_LDLo | 0.52 (3.0) | 0.46 (6.0) | 0.51 (4.5) | 0.51 (4.5) | 0.54 (2.0) | 0.31 (8.0) | 0.36 (7.0) | 0.56 (1.0) |
| 57.human\_oral\_TDLo | 0.32 (1.0) | 0.11 (7.0) | 0.27 (5.5) | 0.31 (2.0) | 0.3 (3.0) | 0.28 (4.0) | -0.34 (8.0) | 0.27 (5.5) |
| 58.woman\_oral\_TDLo | 0.20 (5.0) | 0.19 (6.0) | 0.13 (7.0) | 0.26 (2.5) | 0.26 (2.5) | 0.22 (4.0) | -0.64 (8.0) | 0.38 (1.0) |
| 59.man\_oral\_TDLo | 0.17 (5.0) | 0.15 (6.0) | 0.28 (2.5) | 0.26 (4.0) | 0.28 (2.5) | 0.11 (7.0) | -0.51 (8.0) | 0.29 (1.0) |
| **Average R2 (Rank)** | **0.51 (4.9)** | **0.50 (5.4)** | **0.54 (4.0)** | **0.55 (3.2)** | **0.57 (2.1)** | **0.36 (7.0)** | **0.20 (7.6)** | **0.61 (1.9)** |
|  |  |  |  |  |  |  |  |  |
| **Table D2 Performance comparison of all algorithms on 59 toxicity endpoints based on RMSE** | | | | | | | | |
| **Simplified Task Name** | **MT-DNN** | **MT-GCNN** | **DLCA** | **ConsenA** | **ConsenB** | **LinInt** | **TranAda** | **CFF-AT** |
| 1.bird-wild\_oral\_LD50 | 0.72 (6.0) | 0.71 (5.0) | 0.67 (2.5) | 0.68 (4.0) | 0.67 (2.5) | 0.76 (7.0) | 0.88 (8.0) | 0.64 (1.0) |
| 2.chicken\_oral\_LD50 | 0.84 (5.0) | 0.87 (7.5) | 0.77 (2.0) | 0.81 (4.0) | 0.79 (3.0) | 0.86 (6.0) | 0.87 (7.5) | 0.69 (1.0) |
| 3.duck\_oral\_LD50 | 0.92 (5.0) | 0.94 (6.0) | 0.85 (2.0) | 0.88 (4.0) | 0.86 (3.0) | 1.03 (8.0) | 1.00 (7.0) | 0.71 (1.0) |
| 4.quail\_oral\_LD50 | 0.74 (6.0) | 0.75 (7.0) | 0.71 (3.0) | 0.72 (4.5) | 0.70 (2.0) | 0.76 (8.0) | 0.72 (4.5) | 0.59 (1.0) |
| 5.frog\_subc\_LDLo | 0.84 (5.5) | 0.82 (3.0) | 0.84 (5.5) | 0.80 (1.5) | 0.80 (1.7) | 0.86 (7.0) | 0.90 (8.0) | 0.83 (4.0) |
| 6.mammal\_intrap\_LD50 | 0.40 (3.0) | 0.45 (6.0) | 0.43 (5.0) | 0.40 (3.0) | 0.40 (3.0) | 0.53 (7.0) | 0.54 (8.0) | 0.39 (1.0) |
| 7.mammal\_oral\_LD50 | 0.49 (3.0) | 0.54 (6.0) | 0.53 (5.0) | 0.48 (1.5) | 0.48 (1.5) | 0.66 (7.0) | 0.68 (8.0) | 0.51 (4.0) |
| 8.mammal\_subc\_LD50 | 0.47 (5.5) | 0.45 (4.0) | 0.48 (7.0) | 0.42 (2.0) | 0.44 (3.0) | 0.47 (5.5) | 0.50 (8.0) | 0.40 (1.0) |
| 9.mammal\_unrep\_LD50 | 0.44 (4.5) | 0.46 (6.0) | 0.44 (4.5) | 0.42 (2.5) | 0.42 (2.5) | 0.53 (8.0) | 0.52 (7.0) | 0.37 (1.0) |
| 10.gp\_intrap\_LD50 | 0.61 (3.0) | 0.66 (6.0) | 0.62 (4.0) | 0.63 (5.0) | 0.60 (2.0) | 0.71 (7.0) | 0.74 (8.0) | 0.51 (1.0) |
| 11.gp\_intrav\_LD50 | 0.75 (5.0) | 0.77 (6.0) | 0.71 (3.5) | 0.71 (3.5) | 0.70 (2.0) | 0.80 (7.0) | 1.07 (8.0) | 0.68 (1.0) |
| 12.gp\_intrav\_LDLo | 0.65 (5.0) | 0.67 (6.0) | 0.61 (2.5) | 0.64 (4.0) | 0.61 (2.5) | 0.70 (7.0) | 0.79 (8.0) | 0.53 (1.0) |
| 13.gp\_oral\_LD50 | 0.72 (6.0) | 0.70 (5.0) | 0.69 (4.0) | 0.66 (2.5) | 0.66 (2.5) | 0.88 (8.0) | 0.87 (7.0) | 0.57 (1.0) |
| 14.gp\_skin\_LD50 | 0.77 (5.0) | 0.78 (6.0) | 0.74 (3.0) | 0.75 (4.0) | 0.72 (2.0) | 0.84 (8.0) | 0.83 (7.0) | 0.68 (1.0) |
| 15.gp\_subc\_LD50 | 0.81 (4.0) | 0.88 (6.5) | 0.83 (5.0) | 0.80 (3.0) | 0.79 (2.0) | 0.91 (8.0) | 0.88 (6.5) | 0.61 (1.0) |
| 16.gp\_subc\_LDLo | 0.75 (5.5) | 0.72 (4.0) | 0.75 (5.5) | 0.70 (2.5) | 0.70 (2.5) | 0.82 (8.0) | 0.77 (7.0) | 0.54 (1.0) |
| 17.mouse\_intram\_LD50 | 0.58 (6.0) | 0.57 (5.0) | 0.55 (4.0) | 0.54 (3.0) | 0.53 (2.0) | 0.67 (7.0) | 0.70 (8.0) | 0.48 (1.0) |
| 18.mouse\_intrap\_LD50 | 0.45 (5.5) | 0.45 (5.5) | 0.42 (3.5) | 0.42 (3.5) | 0.41 (1.5) | 0.57 (7.0) | 0.84 (8.0) | 0.41 (1.5) |
| 19.mouse\_intrap\_LDLo | 0.67 (6.0) | 0.65 (5.0) | 0.62 (3.5) | 0.62 (3.5) | 0.61 (2.0) | 0.74 (7.0) | 0.77 (8.0) | 0.58 (1.0) |
| 20.mouse\_intrav\_LD50 | 0.48 (6.0) | 0.47 (5.0) | 0.44 (2.5) | 0.44 (2.5) | 0.43 (1.0) | 0.57 (7.0) | 0.84 (8.0) | 0.46 (4.0) |
| 21.mouse\_intrav\_LDLo | 0.83 (5.0) | 0.86 (6.0) | 0.79 (3.0) | 0.80 (4.0) | 0.78 (2.0) | 0.97 (7.0) | 1.01 (8.0) | 0.66 (1.0) |
| 22.mouse\_oral\_LD50 | 0.42 (5.0) | 0.43 (6.0) | 0.40 (3.5) | 0.40 (3.5) | 0.39 (1.5) | 0.52 (7.0) | 0.75 (8.0) | 0.39 (1.5) |
| 23.mouse\_oral\_LDLo | 0.68 (5.0) | 0.68 (5.0) | 0.68 (5.0) | 0.66 (3.0) | 0.65 (1.5) | 0.77 (7.0) | 0.78 (8.0) | 0.65 (1.5) |
| 24.mouse\_parent\_LD50 | 0.71 (5.5) | 0.72 (7.5) | 0.72 (7.5) | 0.66 (2.0) | 0.67 (3.5) | 0.71 (5.5) | 0.67 (3.5) | 0.61 (1.0) |
| 25.mouse\_skin\_LD50 | 0.79 (5.5) | 0.71 (2.0) | 0.79 (5.5) | 0.73 (3.5) | 0.73 (3.5) | 0.81 (7.0) | 0.88 (8.0) | 0.66 (1.0) |
| 26.mouse\_subc\_LD50 | 0.55 (5.0) | 0.56 (6.0) | 0.53 (4.0) | 0.52 (3.0) | 0.51 (2.0) | 0.67 (7.0) | 0.80 (8.0) | 0.47 (1.0) |
| 27.mouse\_subc\_LDLo | 0.76 (5.0) | 0.73 (3.5) | 0.77 (6.0) | 0.73 (3.5) | 0.72 (2.0) | 0.83 (7.0) | 0.85 (8.0) | 0.59 (1.0) |
| 28.mouse\_unrep\_LD50 | 0.50 (5.0) | 0.54 (6.0) | 0.48 (3.5) | 0.48 (3.5) | 0.47 (2.0) | 0.66 (7.5) | 0.66 (7.5) | 0.46 (1.0) |
| 29.rat\_intram\_LD50 | 0.70 (5.5) | 0.67 (4.0) | 0.70 (5.5) | 0.65 (2.5) | 0.65 (2.5) | 0.84 (8.0) | 0.79 (7.0) | 0.59 (1.0) |
| 30.rat\_intrap\_LD50 | 0.56 (5.5) | 0.56 (5.5) | 0.55 (4.0) | 0.54 (3.0) | 0.52 (2.0) | 0.69 (7.0) | 0.83 (8.0) | 0.45 (1.0) |
| 31.rat\_intrap\_LDLo | 0.58 (6.0) | 0.55 (4.0) | 0.56 (5.0) | 0.54 (2.5) | 0.52 (1.0) | 0.70 (7.0) | 0.71 (8.0) | 0.54 (2.5) |
| 32.rat\_intrav\_LD50 | 0.56 (5.0) | 0.57 (6.0) | 0.54 (3.5) | 0.54 (3.5) | 0.52 (2.0) | 0.74 (7.0) | 0.85 (8.0) | 0.48 (1.0) |
| 33.rat\_intrav\_LDLo | 0.80 (5.0) | 0.88 (6.0) | 0.76 (1.0) | 0.79 (4.0) | 0.78 (2.5) | 0.98 (7.0) | 1.09 (8.0) | 0.78 (2.5) |
| 34.rat\_oral\_LD50 | 0.57 (6.0) | 0.56 (5.0) | 0.53 (3.5) | 0.53 (3.5) | 0.52 (2.0) | 0.69 (7.0) | 0.89 (8.0) | 0.49 (1.0) |
| 35.rat\_oral\_LDLo | 0.68 (5.0) | 0.69 (6.0) | 0.64 (3.5) | 0.64 (3.5) | 0.63 (2.0) | 0.75 (7.0) | 0.80 (8.0) | 0.62 (1.0) |
| 36.rat\_skin\_LD50 | 0.66 (6.0) | 0.63 (5.0) | 0.62 (3.5) | 0.62 (3.5) | 0.61 (2.0) | 0.67 (7.0) | 0.80 (8.0) | 0.56 (1.0) |
| 37.rat\_subc\_LD50 | 0.68 (5.5) | 0.68 (5.5) | 0.65 (4.0) | 0.64 (3.0) | 0.63 (2.0) | 0.81 (7.0) | 0.86 (8.0) | 0.52 (1.0) |
| 38.rat\_subc\_LDLo | 0.66 (3.0) | 0.69 (6.0) | 0.66 (3.0) | 0.67 (5.0) | 0.63 (1.0) | 0.82 (7.0) | 0.83 (8.0) | 0.66 (3.0) |
| 39.rat\_unrep\_LD50 | 0.64 (6.0) | 0.62 (5.0) | 0.60 (4.0) | 0.59 (3.0) | 0.58 (2.0) | 0.74 (7.5) | 0.74 (7.5) | 0.51 (1.0) |
| 40.cat\_intrav\_LD50 | 0.80 (6.0) | 0.78 (4.5) | 0.78 (4.5) | 0.76 (3.0) | 0.75 (2.0) | 0.85 (7.0) | 1.04 (8.0) | 0.68 (1.0) |
| 41.cat\_intrav\_LDLo | 0.68 (4.0) | 0.74 (6.0) | 0.66 (2.0) | 0.67 (3.0) | 0.65 (1.0) | 0.76 (7.0) | 1.08 (8.0) | 0.70 (5.0) |
| 42.cat\_oral\_LD50 | 0.83 (7.0) | 0.84 (8.0) | 0.76 (2.5) | 0.79 (5.0) | 0.76 (2.5) | 0.77 (4.0) | 0.80 (6.0) | 0.68 (1.0) |
| 43.cat\_oral\_LDLo | 0.69 (5.0) | 0.70 (6.0) | 0.64 (1.5) | 0.67 (4.0) | 0.64 (1.5) | 0.75 (7.0) | 0.77 (8.0) | 0.66 (3.0) |
| 44.rabbit\_intrap\_LD50 | 0.66 (6.0) | 0.62 (4.0) | 0.63 (5.0) | 0.61 (3.0) | 0.60 (2.0) | 0.75 (7.0) | 0.83 (8.0) | 0.51 (1.0) |
| 45.rabbit\_intrav\_LD50 | 0.63 (5.0) | 0.64 (6.0) | 0.62 (4.0) | 0.61 (3.0) | 0.59 (2.0) | 0.81 (7.0) | 0.89 (8.0) | 0.53 (1.0) |
| 46.rabbit\_intrav\_LDLo | 0.74 (4.5) | 0.75 (6.0) | 0.71 (2.5) | 0.71 (2.5) | 0.69 (1.0) | 0.83 (7.0) | 1.09 (8.0) | 0.74 (4.5) |
| 47.rabbit\_oral\_LD50 | 0.62 (6.0) | 0.61 (5.0) | 0.60 (3.5) | 0.60 (3.5) | 0.58 (2.0) | 0.74 (7.0) | 0.76 (8.0) | 0.50 (1.0) |
| 48.rabbit\_oral\_LDLo | 0.58 (2.5) | 0.63 (6.0) | 0.61 (5.0) | 0.59 (4.0) | 0.58 (2.5) | 0.68 (7.5) | 0.68 (7.5) | 0.55 (1.0) |
| 49.rabbit\_skin\_LD50 | 0.56 (6.0) | 0.55 (5.0) | 0.54 (4.0) | 0.53 (2.5) | 0.53 (2.5) | 0.64 (7.0) | 0.73 (8.0) | 0.52 (1.0) |
| 50.rabbit\_skin\_LDLo | 0.67 (1.0) | 0.77 (7.0) | 0.70 (5.0) | 0.68 (3.0) | 0.68 (3.0) | 0.74 (6.0) | 0.86 (8.0) | 0.68 (3.0) |
| 51.rabbit\_subc\_LD50 | 0.81 (4.0) | 0.83 (5.0) | 0.86 (6.0) | 0.76 (2.0) | 0.78 (3.0) | 1.00 (8.0) | 0.88 (7.0) | 0.69 (1.0) |
| 52.rabbit\_subc\_LDLo | 0.66 (6.0) | 0.61 (4.5) | 0.61 (4.5) | 0.59 (2.0) | 0.57 (1.0) | 0.80 (8.0) | 0.77 (7.0) | 0.60 (3.0) |
| 53.dog\_intrav\_LD50 | 0.66 (6.0) | 0.64 (4.5) | 0.64 (4.5) | 0.62 (2.0) | 0.60 (3.0) | 0.79 (7.0) | 0.98 (8.0) | 0.54 (1.0) |
| 54.dog\_intrav\_LDLo | 0.71 (5.0) | 0.72 (6.0) | 0.70 (4.0) | 0.68 (3.0) | 0.67 (2.0) | 0.86 (7.0) | 1.10 (8.0) | 0.63 (1.0) |
| 55.dog\_oral\_LD50 | 0.77 (6.0) | 0.76 (5.0) | 0.73 (3.5) | 0.73 (3.5) | 0.71 (2.0) | 0.86 (7.0) | 0.89 (8.0) | 0.66 (1.0) |
| 56.dog\_oral\_LDLo | 0.83 (3.0) | 0.88 (6.0) | 0.84 (4.5) | 0.84 (4.5) | 0.82 (2.0) | 0.95 (8.0) | 0.91 (7.0) | 0.76 (1.0) |
| 57.human\_oral\_TDLo | 1.27 (1.0) | 1.44 (7.0) | 1.32 (4.0) | 1.29 (2.5) | 1.29 (2.5) | 1.33 (5.0) | 1.81 (8.0) | 1.34 (6.0) |
| 58.woman\_oral\_TDLo | 1.04 (4.0) | 1.05 (5.0) | 1.09 (6.0) | 1.01 (1.5) | 1.01 (1.5) | 1.17 (7.0) | 1.68 (8.0) | 1.03 (3.0) |
| 59.man\_oral\_TDLo | 1.11 (5.0) | 1.13 (6.0) | 1.04 (1.5) | 1.06 (3.5) | 1.04 (1.5) | 1.20 (7.0) | 1.57 (8.0) | 1.06 (3.5) |
| **Average RMSE (Rank)** | **0.69 (5.0)** | **0.70 (5.5)** | **0.67 (4.0)** | **0.66 (3.2)** | **0.65 (2.1)** | **0.78 (7.0)** | **0.87 (7.6)** | **0.61 (1.7)** |

Tables D1 and D2 present the R2 and RMSE values for all algorithms across the 59 toxicity endpoints. Each row represents the model performance for a specific endpoint, with values in parentheses indicating the performance rank. A rank of 1 denotes the best performance, 2 denotes the second-best, and so forth. If multiple algorithms achieve equal performance, an average rank is assigned. The last row of each table summarizes the average performance and rank of all models across the 59 toxicity endpoints.

#### Section E Friedman test and Nemenyi test for Tables D1 and D2

To further validate the effectiveness of the proposed method, the Friedman test and Nemenyi test [4] were conducted to assess the performance differences between the proposed method and comparison algorithms. The Friedman test, serving as a non-parametric statistical test, evaluated whether significant differences were observed in the prediction performance of multiple models across multiple datasets by analyzing rank data.

Consider N datasets and K algorithms, where , denotes the rank of the *j-th* algorithm on the *i-th* dataset, and represents the average rank of the *j-th* algorithm across all datasets (Equation (1)). The initial calculation for the Friedman test is described by Equation (2), whereas Iman et al. [4] considered this formula to be overly conservative and proposed an improved version, represented by Equation (3).

(1)

(2)

(3)

Using Table D1 of Section D as an example, with K=8 and N=59, the average rank of each algorithm is provided in the last row of Table D1. The Friedman test calculations for Table D1 are presented in Equations (4) and (5).

(4)

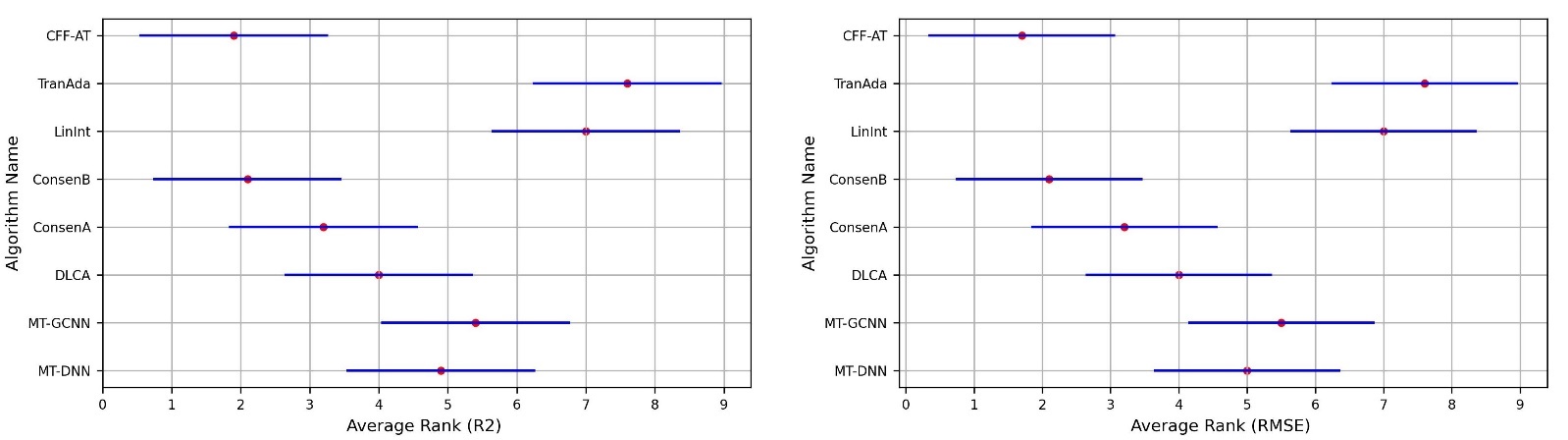
(5)

Using the degrees of freedom (K-1) and (K-1)\*(N-1) in the F distribution, the threshold value F(7,406) at α=0.05 was 2.03, as found in the standard statistical tables. Under the null hypothesis, the Friedman test assumes that all algorithms perform equally. However, the statistical result from Equation (5) exceeded the threshold value of 2.03, allowing us to reject the null hypothesis and conclude that there were significant differences in R2 performance across all models for the 59 toxicity endpoints at a 95% confidence level. Additionally, the Friedman test was performed for Table D2 of Section D, demonstrating significant differences in the RMSE performance among all models. The calculation process was not repeated here.

When the null hypothesis of the Friedman test was rejected, a follow-up Nemenyi test was conducted to analyze the performance differences between pairwise algorithms. A significant statistical difference exists if the difference in their average ranks exceeds the Critical Difference (CD) calculated using Equation (6). Referring to Table D1, CD at was computed as . Using the last row of Table D1 as an example, the rank difference between the proposed algorithm CFF-AT and the comparison algorithm TranAda was , indicating a significant performance difference between these two methods in terms of R2.

(6)

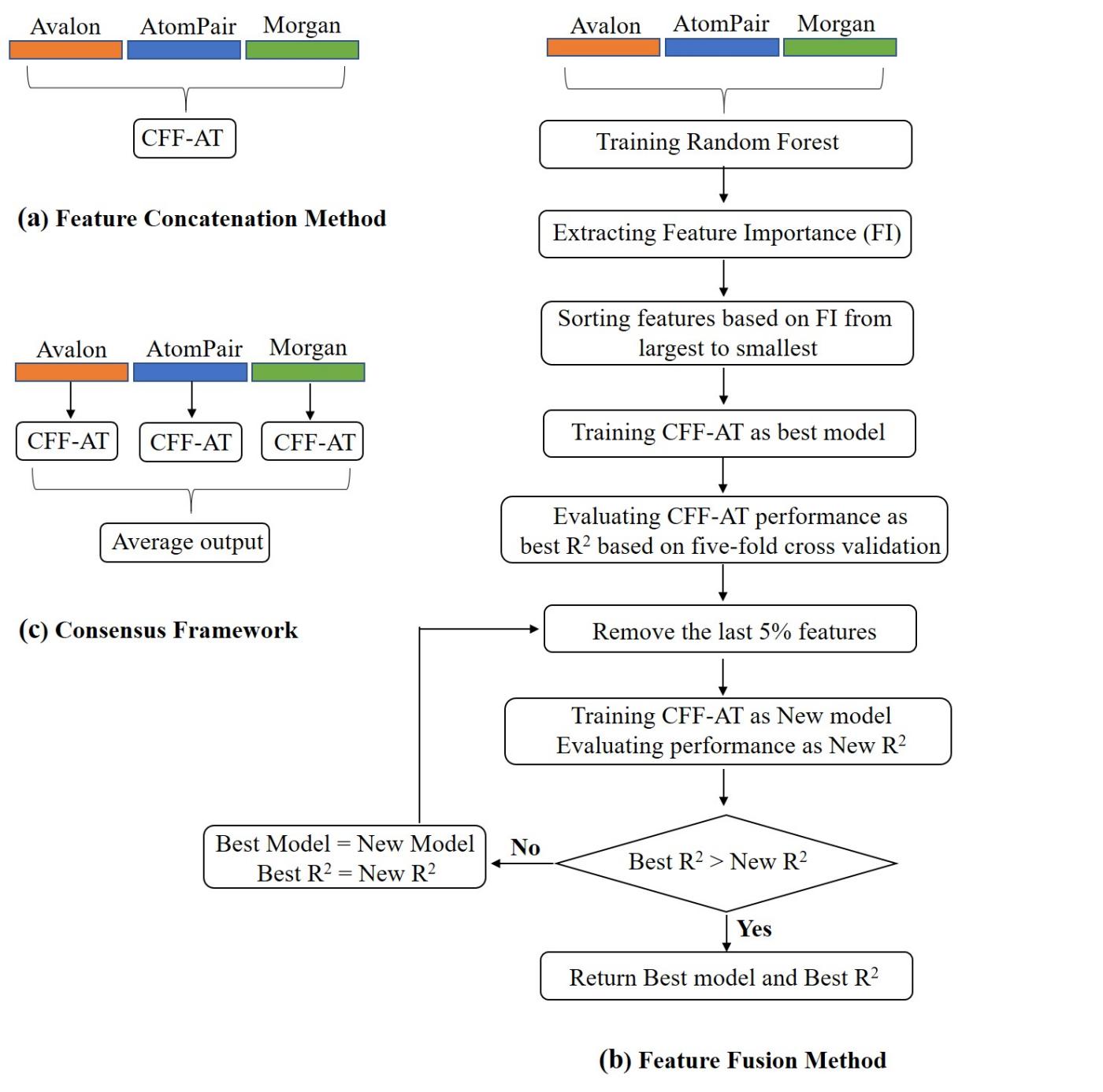
|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Table E1 Critical values for the two-tailed Nemenyi test** | | | | | | | | | |
| Number of algorithms | 2 | 3 | 4 | 5 | 6 | 7 | **8** | 9 | 10 |
|  | 1.960 | 2.343 | 2.569 | 2.728 | 2.850 | 2.949 | **3.031** | 3.102 | 3.164 |
|  | 1.645 | 2.052 | 2.291 | 2.459 | 2.589 | 2.693 | 2.780 | 2.855 | 2.920 |



**Fig. E1** Performance rank difference between pairwise algorithms under R2 and RMSE.

Figure E1 illustrates the performance rank differences among all algorithms for clarity. The Y-axis represents the algorithm names, and the x-axis shows the rank values, where smaller values indicate better performance. The red dots correspond to the rank values of each algorithm, and the length of each line is twice that of CD. A significant performance difference was observed when the horizontal distance between the two red dots exceeded half of the line length. As shown in Figure E1, the performance of most of the comparison algorithms lagged significantly behind, except for ConsenB, whose average performance was relatively close to that of CFF-AT.

#### Section F Exploration of multi-view data modeling



**Fig. F1** Overview of multi-view data modeling approach.

This experiment explored the modeling performance of CFF-AT with multi-view data using three methods, as illustrated in Figure F1. Method (a), Feature Concatenation [5], concatenates the representations of all samples in Avalon, AtomPair, and Morgan feature spaces to train a single model. Method (b), Feature Fusion, extends method (a) by first concatenating the three feature types to train a random forest, extracting the feature importance values (FIV), and reordering features based on FIV. It then iteratively removed the bottom 5% of the features with the lowest ranks and tested for performance improvement until convergence. Method (b) [6] effectively combines iterative feature dimensionality reduction with CFF-AT. Method (c), Consensus Framework [7], trains separate models for each of the three feature types and averages their outputs within the overall framework.

Table F1 compares the performance of the three multi-view modeling methods. Model 1 represents the performance of CFF-AT using a single view (Avalon) as a benchmark. Model 2 corresponds to the Feature Concatenation method, where the performance slightly decreases, likely owing to the feature redundancy caused by concatenation. Model 3, corresponding to the Feature Fusion method, improves the performance by iteratively eliminating redundant features, ultimately using only 10% of the fused features. These highlights significant information redundancy in the feature spaces of Avalon, AtomPair, and Morgan. Model 4 represents the consensus framework, which achieved the best performance by averaging the outputs of the three CFF-AT models trained on each feature type.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Table F1** Performance comparison of three multi-view modeling methods | | | | | |
| ID | Model Name | Input Feature Type | Num of features used | R2 | RMSE |
| 1 | CFF-AT | Avalon | 1024 | 0.61 | 0.61 |
| 2 | CFF-AT-Concatenation | Avalon, AtomPair, Morgan | 1024\*3=3072 | 0.60 | 0.61 |
| 3 | CFF-AT-Fusion | Avalon, AtomPair, Morgan | 1024\*3\*0.1=307 | 0.62 | 0.60 |
| 4 | CFF-AT-Consensus | Avalon, AtomPair, Morgan | 1024\*3=3072 | **0.64** | **0.57** |

#### Section G Supplementary experiments during the revision process

1. **Scaffold-CV**

Compared with 5-fold cross-validation (R2 = 0.61, RMSE = 0.61), when using Scaffold-CV with reserving 20% of the samples as the test set, the model performance does significantly decline (R2 = 0.42, RMSE = 0.75). The main reason is that cross-validation divides the training set and testing set based on the data distribution, allowing the model to make effective predictions. However, Scaffold-CV divides based on the molecular skeleton, and its data partitioning strategy is stricter. The distribution pattern of the testing set may rarely be reflected in the training set, deteriorating model performance. Additionally, all the published literatures based on this dataset just report results using cross-validation methods. To enable effective comparison with the comparison algorithms, we mainly report the results of cross-validation in the manuscript.

1. **Y-randomization experiments**

We conducted 30 runs of Y-randomization experiments on the proposed method. The steps are as follows:

(a) Under the Avalon feature space, set the original modeling result of the proposed method as the benchmark (R2 = 0.61).

(b) Keep the feature matrix X unchanged and randomly shuffle the element values of the Y matrix for each toxicity endpoint.

(c) Use the proposed method to model the new dataset and perform 5-fold cross-validation to obtain the new performance.

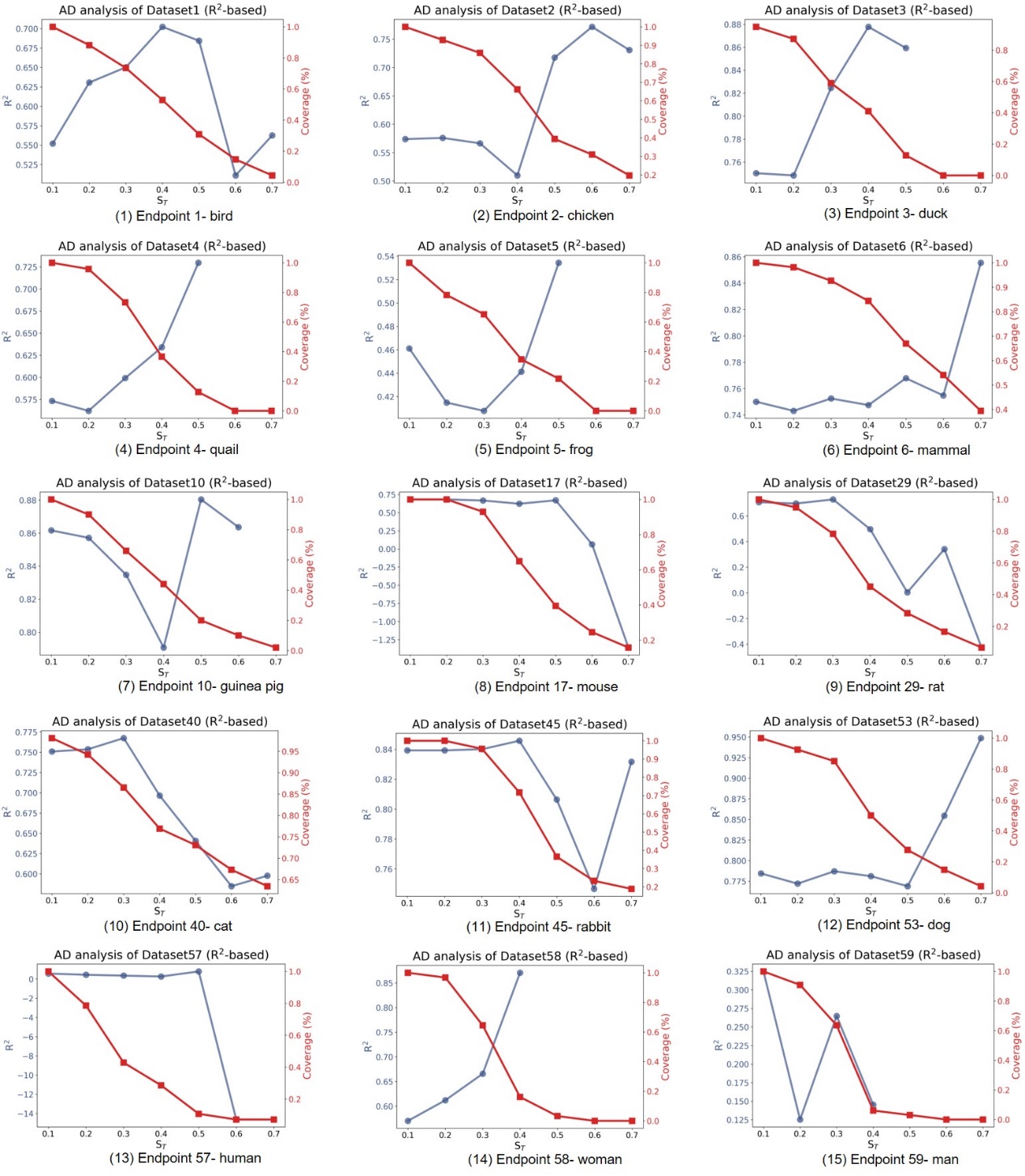
(d) Repeat steps (b) and (c), and check the number of times that the new performance exceeds the benchmark performance after conducting 30 runs of Y-randomization experiments.

After the above experiments, no performance of Y-randomization experiments exceeded the benchmark performance. Therefore, it can be considered that the p-value corresponding to the experiment (0/30) is much smaller than 0.05, proving that the proposed method has truly discovered the mapping relationship between the compounds and the toxicity values.

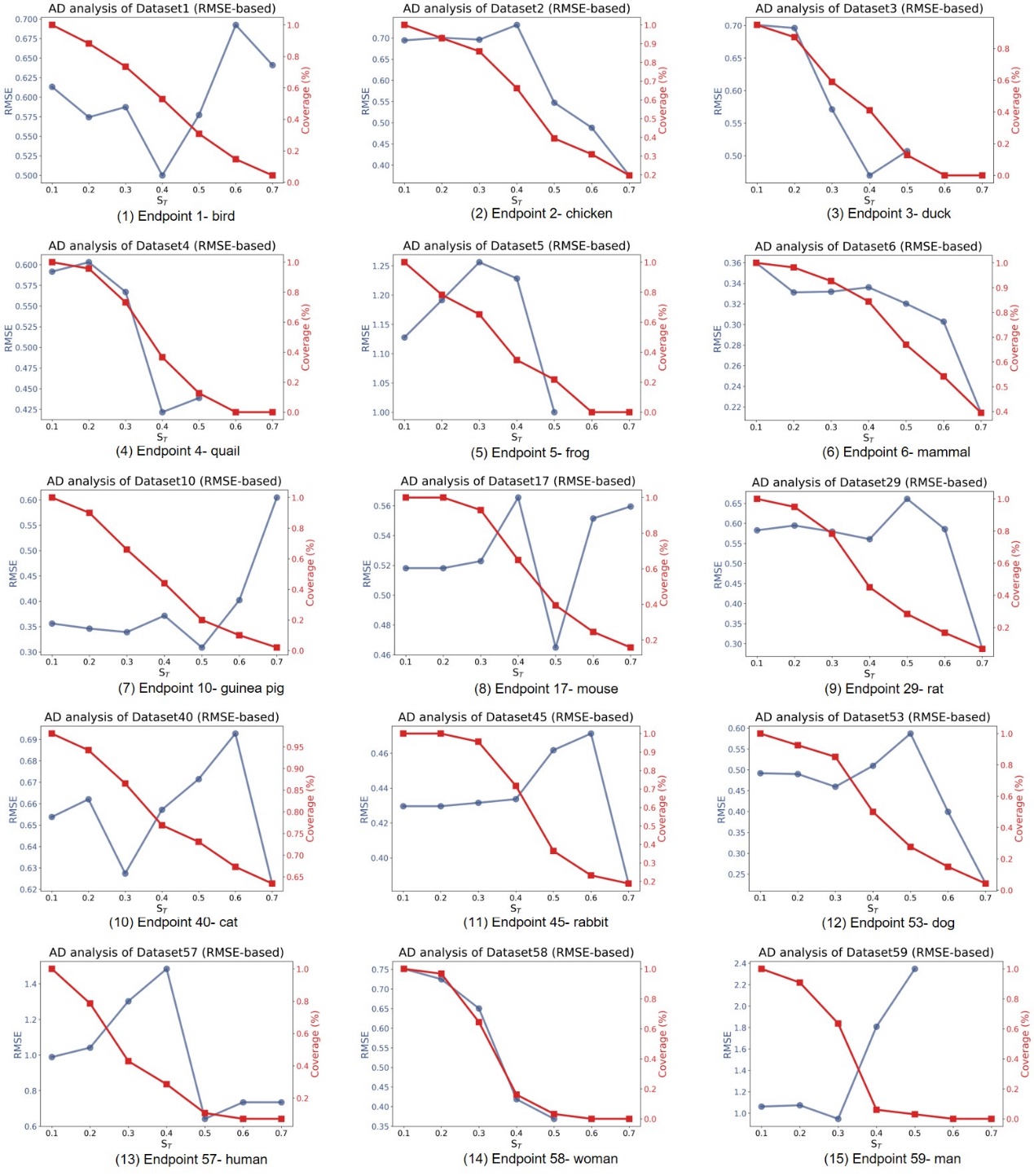
1. **Applicability domain analysis**

The reliability of a model’s prediction depends on whether the queried compound falls within the Applicability Domain (AD) defined by the model[8-10]. Following the study of Lou et al. [8] and Jain et al.[1], we quantify the AD of our model using Tanimoto similarity based on molecular Avalon fingerprints. The AD threshold is defined as the range from 0.1 to 0.7 for evaluation. For each test compound, we identify its *k* nearest neighbors (*k*=10) in the training set based on Tanimoto similarity scores. The test compound is classified as in-AD if the average similarity of these top *k* neighbors exceeds the calculated threshold .

The AD of the model was assessed to identify the chemical space where it achieves reliable predictions. By evaluating a range of AD thresholds from 0.1 to 0.7, we analyzed the corresponding variations in predictive performance and the coverage rates of compounds within the AD in some representative endpoints of each specie (Fig. G1 and G2). As the threshold increases, the coverage of molecules in the AD decreases sharply, while prediction accuracy improves. For instance, in Dataset2 (chicken-oral-LD50, the second item of Fig.G1 and G2), when the threshold increases, R2 rises from 0.57 to 0.73 and RMSE decreases from 0.69 to 0.38. This demonstrates that stricter AD thresholds effectively screen high-confidence predictions by narrowing the chemical space. By balancing prediction accuracy and coverage, *ST*=0.5 was selected as the globally optimal parameter, achieving 39.44% in-AD coverage while maintaining strong predictive performance, thus ensuring both model reliability and practical utility. A similar trend was observed across the other endpoints.



**Fig G1**. Performance metrics (R²) of samples in AD under varying thresholds. The x-axis represents the AD threshold ST. The left y-axis (blue lines) indicates metric values, while the right y-axis (red lines) denotes the proportion of extracted samples relative to the total.



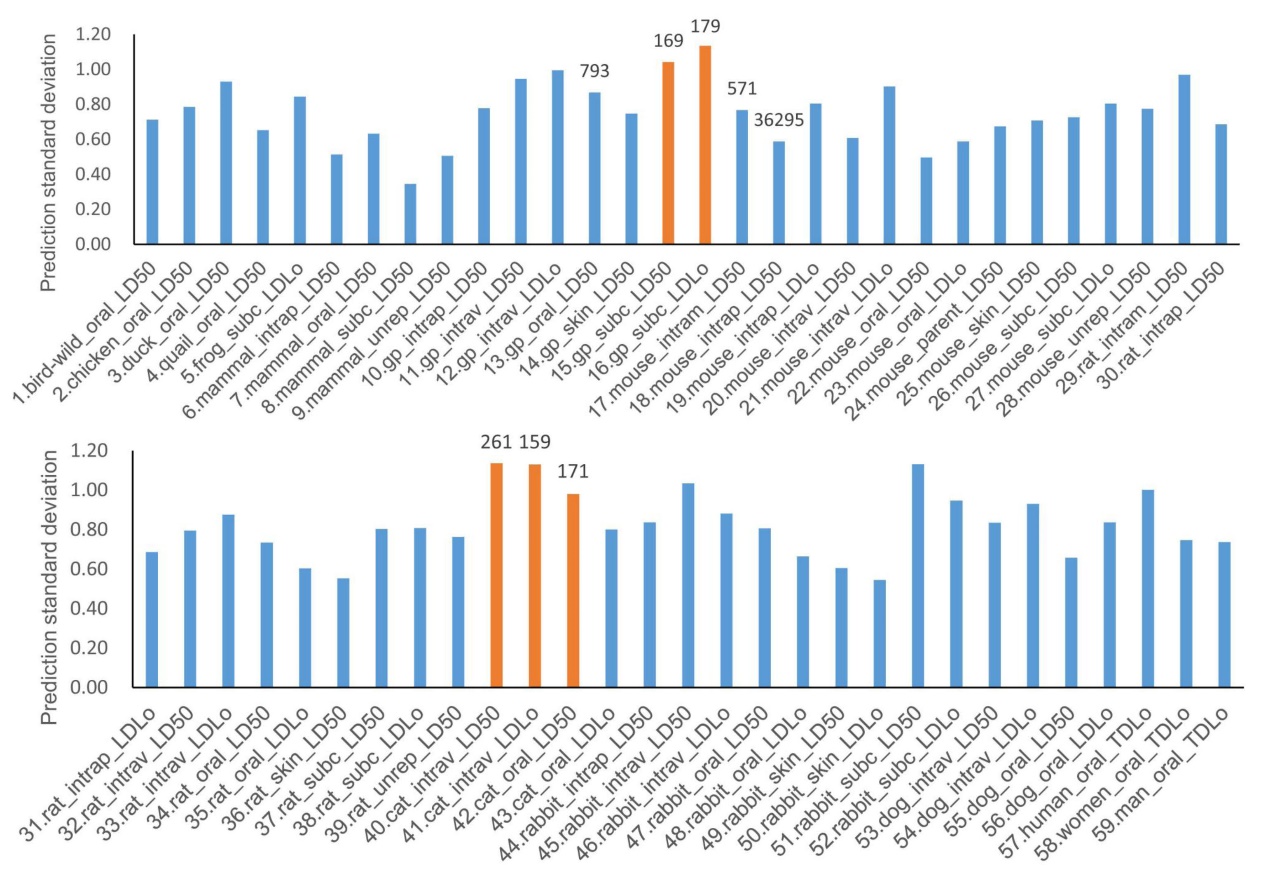
**Fig G2**. Performance metrics (RMSE) of samples in AD under varying thresholds.

1. **Uncertainty quantification**

Uncertainty Quantification (UQ) mainly addresses some critical challenges in high-dimensional dynamical systems and mathematical systems [11]. In the current research field of Machine Learning (ML), there are very few studies related to the UQ of deep forest, and there is a lack of general tools for UQ of ML predictions [12]. Apart from some methods that can directly provide confidence intervals of predictions, such as linear and polynomial models or Bayesian neural networks, most machine learning methods are unable to provide confidence intervals related to predictions. In response to the concerns raised by the reviewers, we conducted the UQ analysis from the following two aspects.

1. In ML, uncertainty can be simply classified into data uncertainty and model uncertainty. Data uncertainty is caused by the inherent and inevitable noise in the data, and it cannot be reduced by collecting more data. Our research data comes from existing literatures, based on which many published studies have been directly proposed. To maintain fair performance comparisons with published studies, we do not analyze data uncertainty.
2. Regarding model uncertainty, some studies [13, 14] have pointed out that "In deep ensembles methods, model uncertainty is often estimated by looking at an ensemble of trained models, the spread of predictions obtained from different models is then used as an estimate of model uncertainty". The proposed method is based on the modification of the deep forest framework. Its output layer is the ensemble of multiple forests, which can be described as the fusion of the ensemble of decision trees. Therefore, we can calculate the Prediction Standard Deviation (PSB) of all decision trees on the output layer in the testing set, and representing the average of all decision trees’ PSBs as the model's uncertainty. On 59 toxicity endpoints, we plotted the trend of PSB changes (Figure 3). A higher PSB indicates that the prediction uncertainty of the model at that toxicity endpoint is higher.

In Fig G3, the y-axis in the above figure represents the PSB, and the x-axis corresponds to all toxicity endpoints. The yellow bars indicate the toxicity endpoints with larger PSBs, and the numbers above represent the size of these endpoints. It can be observed that the endpoints with larger PSB are primarily caused by small training samples. We can use the prediction results ± standard deviation to reflect the fluctuation range of the predictions. For example, when the R2 performance is higher and the standard deviation is smaller, it indicates that the prediction reliability of the model is higher.



**Fig G3**. Predicted standard deviation (PSB) for all toxicity endpoints. The x-axis represents the toxicity endpoint. The y-axis indicates PSB, the yellow bars represent the toxicity endpoints with relatively large PSBs. The number above the bar indicates the number of samples corresponding to the endpoint.

**References**

[1] S. Jain, V. Siramshetty, V. Alves, E. Muratov, N. Kleinstreuer, A. Tropsha, M. Nicklaus, A. Simeonov, A. Zakharov, Large-Scale Modeling of Multispecies Acute Toxicity End Points Using Consensus of Multitask Deep Learning Methods, Journal of Chemical Information and Modeling, 61 (2021) 653–663. <https://10.1021/acs.jcim.0c01164>.

[2] M. Miller, G. Hazard, V. Hudson, C. Hilt, J. Fang, D. Mayer, L. Callahan, ChemIDplus: A free, web-based portal to a variety of compound-based information, Abstracts of Papers of the American Chemical Society, 226 (2003) U305–U305.

[3] P. Wexler, TOXNET: An evolving web resource for toxicology and environmental health information, Toxicology, 157 (2001) 3–10. <https://10.1016/s0300-483x(00)00337-1>.

[4] J. Demsar, Statistical comparisons of classifiers over multiple data sets, Journal of Machine Learning Research, 7 (2006) 1–30.

[5] L. Wu, J. Gao, Y. Zhang, B. Sui, Y. Wen, Q. Wu, K. Liu, S. He, X. Bo, A hybrid deep forest-based method for predicting synergistic drug combinations, Cell reports methods, 3 (2023) 100411. <https://10.1016/j.crmeth.2023.100411>.

[6] F. Xu, L.L. Lin, Z.H. Li, Q.Q. Hong, K.H. Liu, Q.Q. Wu, Q.D. Li, Y.H. Zheng, J. Tian, MRDFF: A deep forest based framework for CT whole heart segmentation, METHODS, 208 (2022) 48–58. <https://10.1016/j.ymeth.2022.10.005>.

[7] S. Jain, V.B. Siramshetty, V.M. Alves, E.N. Muratov, N. Kleinstreuer, A. Tropsha, M.C. Nicklaus, A. Simeonov, A.V. Zakharov, Large-Scale Modeling of Multispecies Acute Toxicity End Points Using Consensus of Multitask Deep Learning Methods, Journal of Chemical Information and Modeling, 61 (2021) 653–663. <https://10.1021/acs.jcim.0c01164>.

[8] S. Lou, Z. Yu, Z. Huang, H. Wang, F. Pan, W. Li, G. Liu, Y. Tang, In Silico Prediction of Chemical Acute Dermal Toxicity Using Explainable Machine Learning Methods, Chemical Research in Toxicology, 37 (2024) 513–524. <https://10.1021/acs.chemrestox.4c00012>.

[9] F. Sahigara, K. Mansouri, D. Ballabio, A. Mauri, V. Consonni, R. Todeschini, Comparison of Different Approaches to Define the Applicability Domain of QSAR Models, Molecules, 17 (2012) 4791–4810. <https://10.3390/molecules17054791>.

[10] G. Patlewicz, J. Fitzpatrick, Current and Future Perspectives on the Development, Evaluation, and Application of in Silico Approaches for Predicting Toxicity, Chemical Research in Toxicology, 29 (2016) 438–451. <https://10.1021/acs.chemrestox.5b00388>.

[11] S. Cheng, C. Quilodrán-Casas, S. Ouala, A. Farchi, C. Liu, P. Tandeo, R. Fablet, D. Lucor, B. Iooss, J. Brajard, D. Xiao, T. Janjic, W. Ding, Y. Guo, A. Carrassi, M. Bocquet, R. Arcucci, Machine Learning With Data Assimilation and Uncertainty Quantification for Dynamical Systems: A Review, Ieee-Caa Journal of Automatica Sinica, 10 (2023) 1361–1387. <https://10.1109/jas.2023.123537>.

[12] A. Olivier, D. Giovanis, B. Aakash, M. Chauhan, L. Vandanapu, M. Shields, UQpy: A general purpose Python package and development environment for uncertainty quantification, Journal of Computational Science, 47 (2020). <https://10.1016/j.jocs.2020.101204>.

[13] B. Lakshminarayanan, A. Pritzel, C. Blundell, Simple and Scalable Predictive Uncertainty Estimation using Deep Ensembles, 31st Annual Conference on Neural Information Processing Systems (NIPS), Long Beach, CA, 2017.

[14] L. Mentch, G. Hooker, Quantifying Uncertainty in Random Forests via Confidence Intervals and Hypothesis Tests, Journal of Machine Learning Research, 17 (2016).