**Consensus models to predict oral rat acute toxicity and validation on a dataset coming from the industrial context**

Filippo Lunghinia,b, Gilles Marcoua\*, Philippe Azamb, Dragos Horvatha, Remi Patouxb, Erik Van Miertb, Alexandre Varneka\*

*aLaboratory of Chemoinformatics, University of Strasbourg, 4 Rue Blaise Pascal, 67081, Strasbourg, France; bSolvay S.A., 20 Rue Marcel Etienne Sembat, 69190, St. Fons, France*

\* Corresponding author. email: g.marcou@unistra.fr; varnek@unistra.fr; phone no.: +33-68851304

**SUPPLEMENTARY INFORMATION**

**Section 1 – Model generation**

***QSAR Toolbox relevant datasets for oral acute toxicity***

The following databases have been mined for relevant data.

* ECOTOX. United States Environmental Protection Agency US-EPA. Endpoints: Bioaccumulation aquatic, Bioaccumulation terrestrial, Aquatic toxicity, Terestrial toxicity, Acute toxicity, Repeated dose toxicity
* Food TOX Hazard EFSA. European Food Safety Authority (EFSA). Endpoints: Human Health Hazards, Ecotoxicological Information.
* Toxicity Japan MHLW. Ministry of Health, Labour and Welfare, Japan. Endpoints: LD50, Chromosome aberration, Gene mutation.
* ZEBET database. Federal Institute for Consumer Health Protection and Veterinary Medicine (BgVV). Endpoints: IC50, LD50

***Consensus decisions***

If the concordance of predictions score is higher than the established threshold the decision is rejected. For the continuous LD50 model, a 0.5 threshold for the Median Absolute Deviation (MAD) was chosen.

For multi-class model, the threshold of 0.71 was selected based on the calculation of the entropy (S) of the predictions among the 5 classes (entropy formula in Equation S1).

***Equation S1*** *– Entropy calculation formula*

Where *Pi* is the probability of the compound to be predicted as the given toxicity class *i*. If all the individual models’ predictions in consensus fall into one single GHS class, then the entropy is 0; if the class repartition is totally random (i.e. probability for each class is equal to 0.2), then entropy reaches a maximum level of 1.609. 0.71 was chosen as cut-off values as at this level the most voted descend below the 50 % of the total votes.

***Employed ISIDA’s descriptor spaces***

Table S1 lists all the employed descriptor spaces. The nomenclature of ISIDA descriptors is here briefly described [7].

* T = type of fragmentation. Where 1 = sequences of atom, 3 = sequences of atoms and bonds, 4 = atom centred fragments based on sequences of atoms, 6 = atom centred fragments based on sequences of atoms and bonds, 7 = atom centred fragments based on sequences of atoms (fixed length), 9 = atom centred fragments based on sequences of atoms and bonds (fixed length), 10 = triplets
* L and U = minimum and maximum length
* FF = Force Field coloration
* AP = Atom pairs (all constitutional details of a sequence are removed and only the number of constitutive atoms is given)

***Table S1*** *listing all the employed descriptor spaces with the respective number of fragments.*

|  |  |
| --- | --- |
| No. Of fragments | Descriptor space name |
| 2871 | t10l3u6 |
| 4401 | t10l3u6 |
| 6816 | t1l2u9 |
| 11914 | t1l2u9 |
| 14037 | t3l2u8 |
| 30955 | t3l2u8 |
| 1241 | t3l2u8\_AP |
| 1732 | t3l2u8\_AP |
| 5351 | t3l3u6 |
| 9888 | t3l3u6 |
| 980 | t3l3u7\_AP |
| 1368 | t3l3u7\_AP |
| 5074 | t3l4u6 |
| 9453 | t3l4u6 |
| 5641 | t4l2u4\_AP |
| 23500 | t4l2u4\_AP |
| 714 | t4l3u3\_AP |
| 1552 | t4l3u3\_AP |
| 3258 | t6l2u3 |
| 10560 | t6l2u3 |
| 3359 | t7l2u5\_AP |
| 7907 | t7l2u5\_AP |
| 8826 | t7l3u5 |
| 31623 | t7l3u5 |
| 7516 | t7l4u5 |
| 28266 | t7l4u5 |
| 387 | t9l2u2 |
| 767 | t9l2u2 |
| 2851 | t9l2u3 |
| 8188 | t9l2u3 |
| 2205 | t9l2u3\_AP |
| 5605 | t9l2u3\_AP |
| 2482 | t9l3u3 |
| 7455 | t9l3u3 |
| 1836 | t9l3u3\_AP |
| 4872 | t9l3u3\_AP |
| 5099 | t9l4u4 |
| 22491 | t9l4u4 |
|  |  |
| Average | 7974.763 |

**Section 2 – Database comparison**

Figure S1 shows the distribution pattern for the continuous LD50 property. The distribution pattern is similar for all the databases, with the most clustering of compounds within -1.6 / 0 pLD50 values.

***Figure S1*** *– pLD50 continuous property distribution*

The pLD50 values for common compounds shared between the databases were compared. Table S1 reports a series of scatterplot of a pairwise comparison between the NICEATM database vs. the others. The number of shared compounds is showed in brackets. Correlation coefficients are almost equal to 1 when comparing NICEATM vs. QSARToolbox and TEST EPA; while for ECHA and the Industrial set they are much lower (≈ 0.73).

***Table S2*** *– pLD50 values pairwise comparison for common compounds*

|  |  |
| --- | --- |
| **NICEATM vs. QSARToolbox (# 5996)** | **NICEATM vs. TEST EPA (# 5944)** |
|  |  |
|  | |
| **NICEATM vs. ECHA (# 968)** | **NICEATM vs. Industrial set (# 468)** |
|  |  |

Tanimoto pairwise comparison of the datasets: Industrial/Industrial; Public/ Public; Public /Industrial. The comparison has been made using the same DS selected for GTM.

The rate of compound pairs with Tanimoto similarity less than 0.2 is much more frequent comparing industrial with public set, supporting the structural differences between the two sources of data.

Average similarities:

Industrial/Industrial: 0.307

Public / Public: 0.512

Public / Industrial: 0.378

***Figure S2.*** *Datasets' Tanimoto similarity distributions comparison*

**Section 3 – ColorAtom examples**

Table S3 reports several molecules showing the same coloration pattern with the same functional groups associated with enhanced toxicity.

***Table S3*** *– ColorAtom examples. Colors are directly referred to the modelled property (i.e. pLD50 values). Red color means that the atom contributes to decrease its value (lowering the toxicity); while blue means an increase of its value (i.e. increasing the toxicity).*

|  |  |
| --- | --- |
| **Diazo- and Triazo- substructures** | |
|  |  |
| CAS 1516-67-2  LD50 = 5 mg/kg | CAS 7203-90-9  LD50 = 362 mg/Kg |
|  |  |
| CAS 20241-03-6  LD50 = 300 mg/kg | CAS 40643-36-5  LD50 = 466 mg/kg |
|  |  |
| CAS 41798-82-7  LD50 = 212 mg/kg |  |
|  | |
| **Halogenated benzene substructures** | |
|  |  |
| CAS 19408-74-3  LD50 = 0.8 mg/kg | CAS 36518-74-8  LD50 = 430 mg/kg |
|  |  |
| CAS 50585-41-6  LD50 = 0.1 mg/kg |  |