

Alessandra Roncaglioni,  
Cosimo Toma,  
Giovanna Lavado,  
Kristijan Vukovic,  
Domenico Gadaleta,  
and  
Emilio Benfenati

Istituto di Ricerche  
Farmacologiche  
“Mario Negri”

via G. La Masa 19

Milan - Italy

# Modeling quantitative acute oral systemic toxicity based on a k-Nearest Neighbor (k-NN) algorithm



# Introduction

---

## Laboratory of Environmental Chemistry and Toxicology

- Modeling exercise based on the dataset released by the NICEATM team
- 2 months, 5 people involved using different tools and algorithms (more results on posters)
- Common internal procedure for the dataset curation



# Data preparation: LD<sub>50</sub> values

## Aggregation of entries

- De-salting
- Duplicated structures

## Data variability analysis

- Removed entries:
  - with SD  $\geq 0.5$  for data point values
  - Conflicting classes

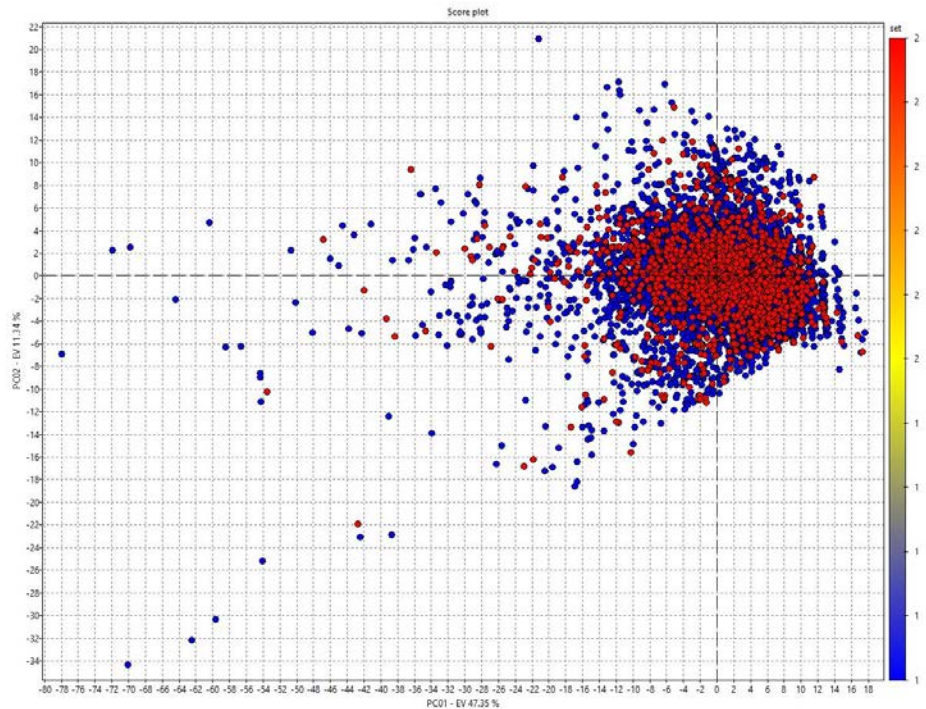
## Final dataset: 8476 unique structures

- Median 1st quantile LD<sub>50</sub> (mmol/kg)
- Univocal label for classification

Count (single point)	Canonical_QSARr	Salt_Solvent	Mean	1st Quantile median	Standard Deviation
4	<chem>COP(=S)(OC)SCN1N=NC2C=CC=CC=2C1=O</chem>		-0.71	-1.84	1.90
4	<chem>NC1CCCCC1</chem>	<chem>Cl(1), C(C)(=O)O(1)</chem>	0.25	-0.95	0.87
2	<chem>CC[P+](C1C=CC=CC=1)(C1C=CC=CC=1)C1C=CC=CC=1</chem>	<chem>[Br-](1), [I-](1)</chem>	-0.50	-0.72	0.32
5	<chem>NCCO</chem>		1.72	1.45	0.30

# Dataset preparation

- Training and validation set splitting (80%/20%)
- Indigo fingerprints calculated in KNIME and compared using the Tanimoto score
- stratified sampling based on clusters obtained with the k-Means clustering algorithm
- nearly-common dataset to all the 5 modeled endpoints



# Data distribution

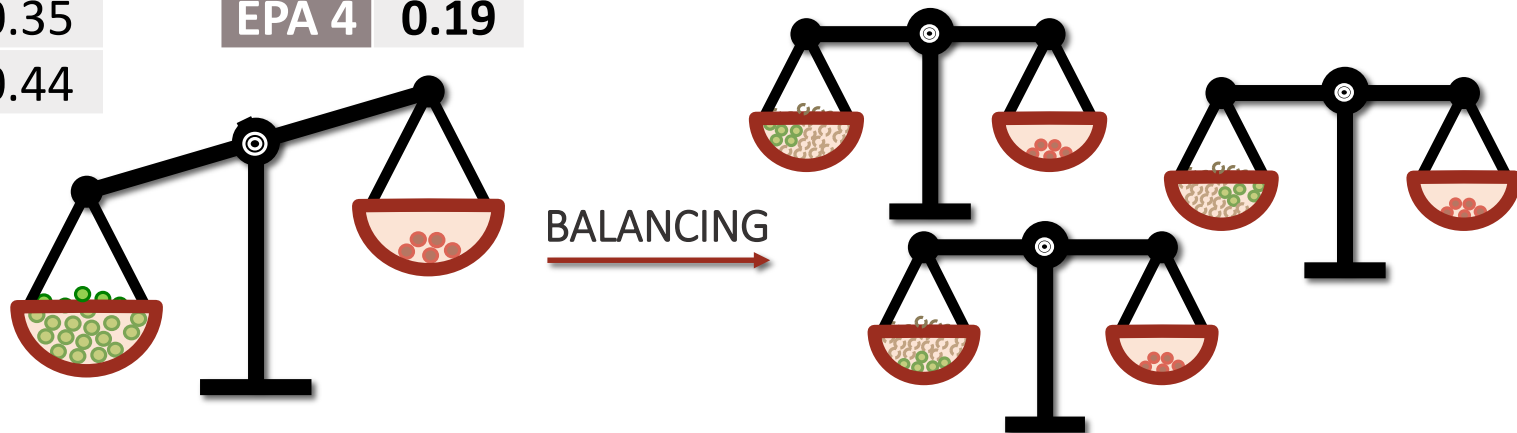
	n	Min logLD50 (mmol/kg)	Max logLD50 (mmol/kg)	Mean	St.dev	Skewness
TS	5029	-4.66	2.71	0.44	0.90	-0.98
VS	1251	-3.70	2.38	0.44	0.90	-1.05

	%
GHS 1	<b>0.02</b>
GHS 2	<b>0.06</b>
GHS 3	<b>0.13</b>
GHS 4	0.35
GHS 5	0.44

	%
EPA 1	<b>0.09</b>
EPA 2	<b>0.22</b>
EPA 3	0.50
EPA 4	<b>0.19</b>

	%
nT	0.43
T	0.57

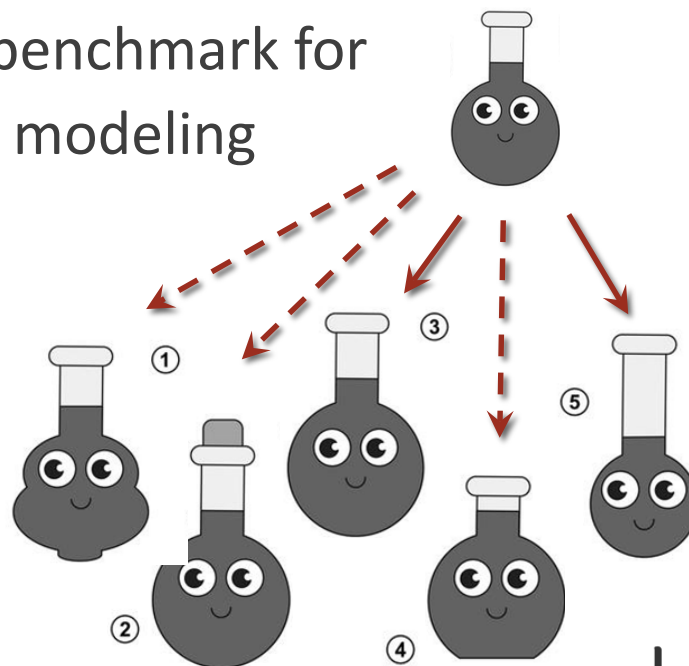
	%
vT	<b>0.92</b>
not vT	0.08



# Modeling approach

---

- K-nn algorithm as implemented into in house software (istKNN by A. Mangano, Kode srl)
- Simple method, internal benchmark for other more sophisticated modeling techniques (see posters)
- Allows a read-across like approach



# Similarity Index (SI)

---

- accounting for different chemical and structural features of the molecules
- considering different relevant structural aspects, including the size of the molecules

$$SI = S(FP)^{0.4} * S(CD)^{0.35} * S(HE)^{0.1} * S(FG)^{0.15}$$

FP → Fingerprints

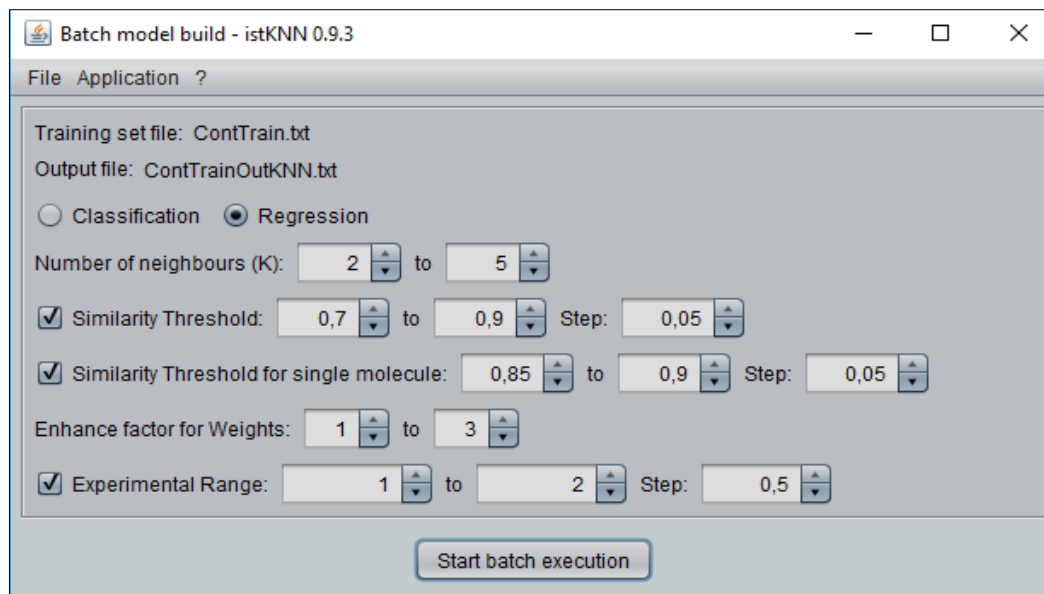
CD → Structural key with 35 Constitutional descriptors (MW, nr of skeleton atoms, etc.)

HE → Structural key with 11 Hetero-atoms descriptors

FG → Structural key with 154 Functional groups (specific chemical moieties)

- FP similarity → Maxwell-Pilliner index
- CD, HE, FG similarities → Bray-Curtis index

# Software setting



Batch model  
development



> 300 models  
optimizing 5  
parameters

Batch models on training set

Select best setting with  $R^2_{LOO}$   
and check VS with best model

Use this setting to run a new  
model on the entire dataset

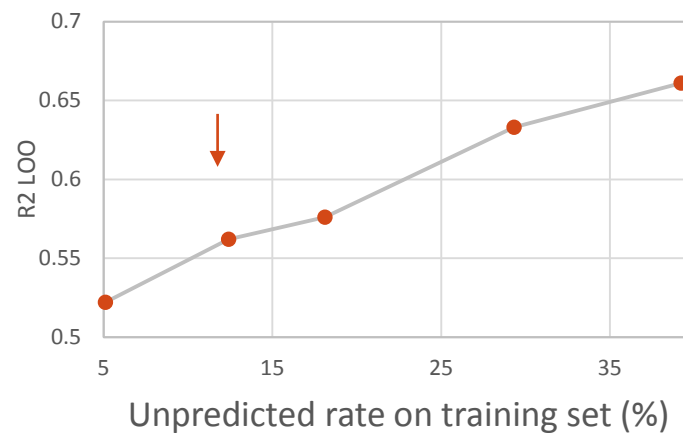
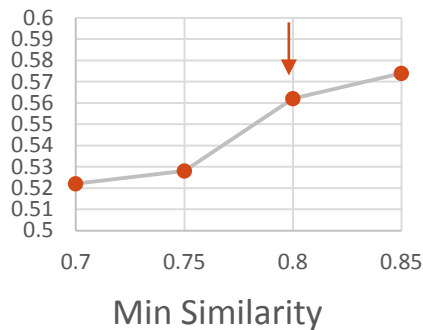
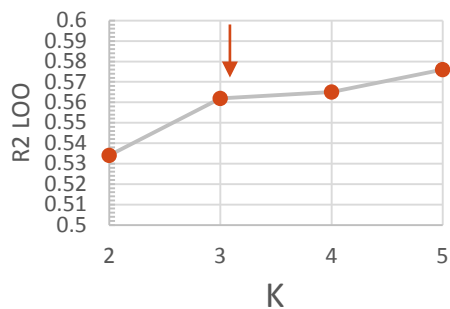


# Results

Model parameters						Training Set (TS)			Validation Set (VS)		
Model No.	K	Min Similarity	Min Similarity for single molecule	Enhance factor	Exper. range	R <sup>2</sup> (LOO)	RMSE (LOO)	Unpred. rate	R <sup>2</sup>	RMSE	Unpred. rate
135	3	0.8	0.85	3	2	0.56	0.59	13.2%	0.56	0.59	12.4%

## TS + VS

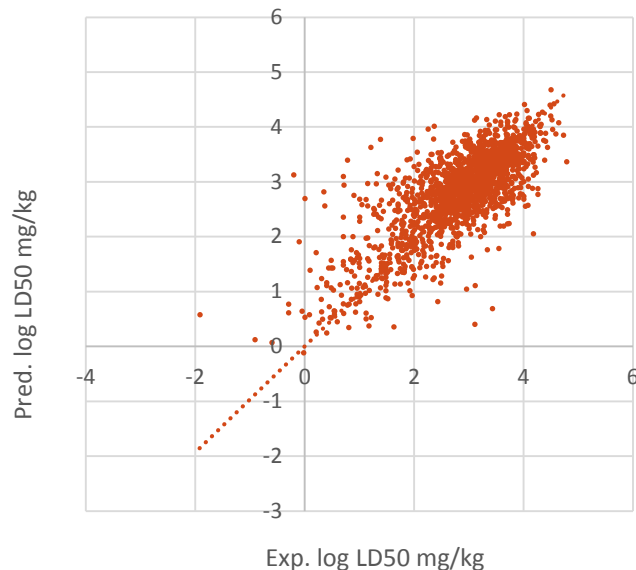
R <sup>2</sup> (LOO)	RMSE (LOO)	Unpred. rate
0.583	0.575	11.9%



# Results on the evaluation set

Evaluation set (n = 1865)

R2	RMSE	Unpred. rate
0.58	0.56	14%

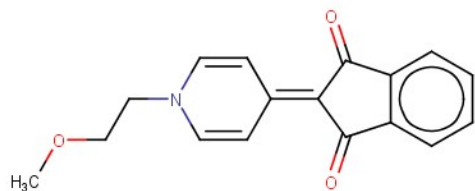


## Applicability domain

*Chemical domain:* Inorganic chemicals and chemicals with unusual elements (i.e., B and Se) are excluded during the data curation. Silicates are included.

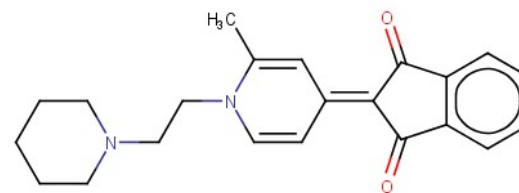
*Model domain:* No prediction if 1) too high experimental range of similar molecules or 2) no similar molecules present in the dataset

# Examples



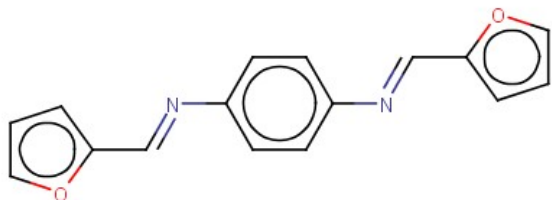
Sim. = 0.881

Exp. LD<sub>50</sub> = 350 mg/kg



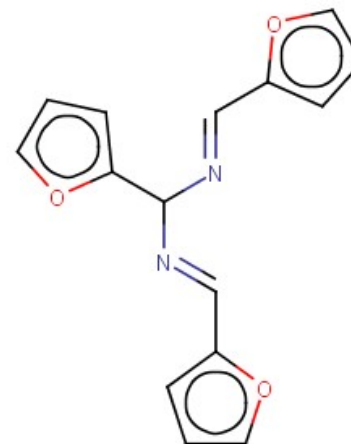
Exp. = 390 mg/kg

Pred. = 439 mg/kg



Sim. = 0.83

Exp. LD<sub>50</sub> = 1220 mg/kg

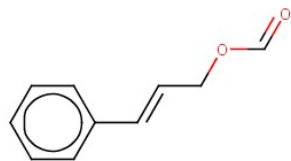


Exp. = 400 mg/kg

No Prediction

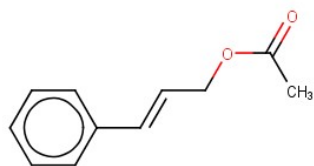
# Examples

---



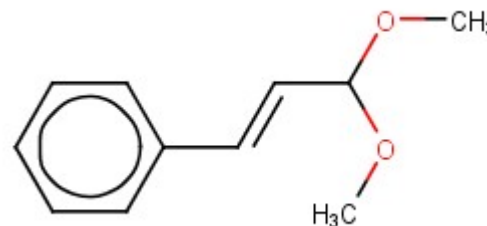
Sim. = 0.941

Exp. LD<sub>50</sub> = 2900 mg/kg



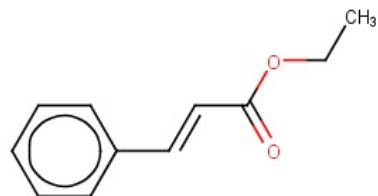
Sim. = 0.94

Exp. LD<sub>50</sub> = 3300 mg/kg



Exp. = 3700 mg/kg

Pred. = 3497 mg/kg

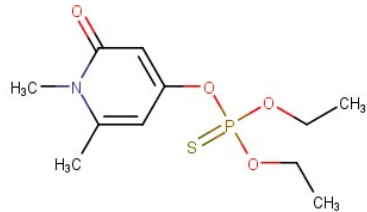


Sim. = 0.931

Exp. LD<sub>50</sub> = 4000 mg/kg

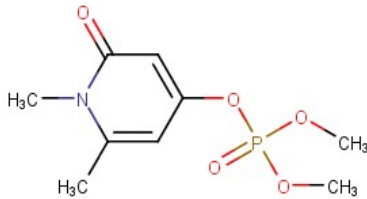


# Examples



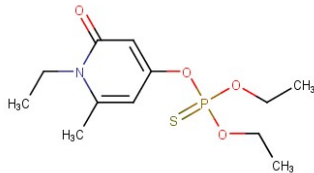
Sim. = 0.964

Exp. LD<sub>50</sub> = 5.97 mg/kg



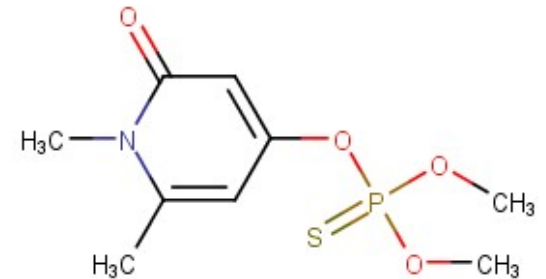
Sim. = 0.958

Exp. LD<sub>50</sub> = 2.91 mg/kg



Sim. = 0.946

Exp. LD<sub>50</sub> = 7070 mg/kg

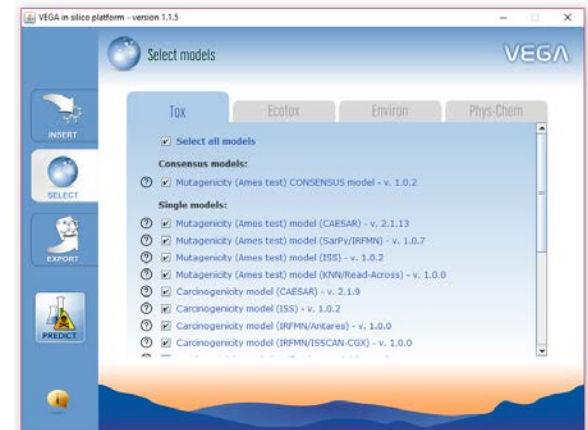


Exp. = 0.812 mg/kg

No Prediction

# Conclusions & perspectives

- k-NN demonstrated to be a simple method capable to obtain acceptable results
- Easily adapted for two prediction strategies:
  - automatic way for **screening** large inventories
  - read-across setting for allowing expert reasoning around available information for **hazard assessment**
- Easily implementable (freely available) in VEGA ([www.vegahub.eu](http://www.vegahub.eu))

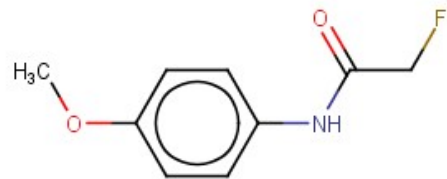


Thank you!

Questions?

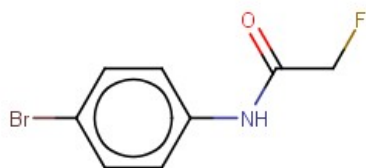
[alessandra.roncaglioni@marionegri.it](mailto:alessandra.roncaglioni@marionegri.it)

# Examples



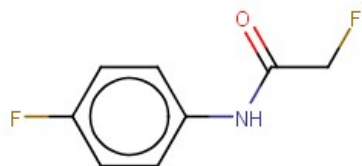
Sim. = 0.913

Exp. LD<sub>50</sub> = 10 mg/kg



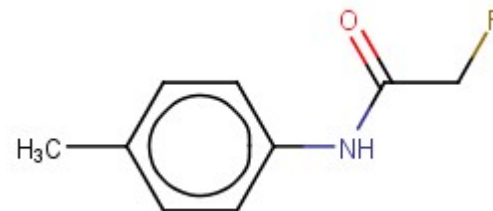
Sim. = 0.912

Exp. LD<sub>50</sub> = 29 mg/kg



Sim. = 0.931

Exp. LD<sub>50</sub> = 2 mg/kg

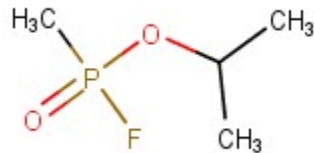


Exp. = 7 mg/kg

Pred. = 7.27 mg/kg

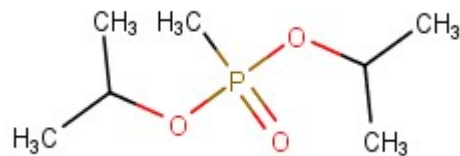


# Examples



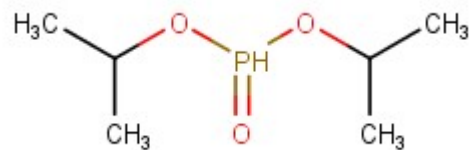
Sim. = 0.898

Exp. LD<sub>50</sub> = 0.55 mg/kg



Sim. = 0.877

Exp. LD<sub>50</sub> = 826 mg/kg

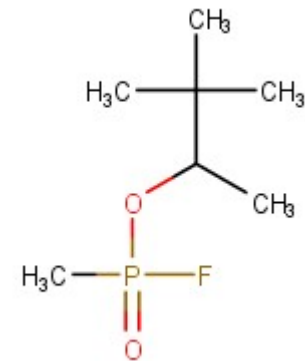


Sim. = 0.842

Exp. LD<sub>50</sub> = 1700 mg/kg



?



Exp. = 0.4 mg/kg

No Prediction