



Gadaleta et al.:

A *k*-NN Algorithm for Predicting Oral Sub-Chronic Toxicity in the Rat

Supplementary Data

Tab. S1: Refused predictions

The number and the percentage of refused predictions are reported for cross-validation on the TS (containing 254 compounds). Each criterion is considered to be applied together with all the prior criteria.

Criterion	Number of refused predictions (%)				
	<i>k</i> =2	<i>k</i> =3	<i>k</i> =4	<i>k</i> =5	<i>k</i> =6
(1)	123 (48%)	123 (48%)	123 (48%)	123 (48%)	123 (48%)
(2)	28 (11%)	28 (11%)	28 (11%)	28 (11%)	28 (11%)
(3)	16 (6%)	28 (11%)	34 (13%)	34 (13%)	36 (14%)
(4)	19 (7%)	24 (9%)	20 (8%)	20 (8%)	20 (8%)
Total predictions	68 (28%)	51 (21%)	49 (20%)	49 (20%)	47 (19%)

(1) Chemicals without any neighbor with similarity ≥ 0.85 cannot be predicted.

(2) Chemicals with a single neighbor with similarity ≥ 0.85 but with similarity < 0.90 cannot be predicted.

(3) Prediction is rejected if the difference among experimental values of nearest neighbors is greater than 1 log unit.

(4) If the nearest neighbor of a target chemical has an error in cross-validation greater than 1 log unit, the prediction of the target chemical is rejected.


Tab. S2: Performance of *k*-NN models

Statistics and predicted compounds are reported for both cross-validation on the TS (containing 254 compounds) and external validation on the VS (containing 179 compounds). Data relate to the maximum admitted difference between experimental values for the neighbors (Δ , see section 2.3) by 0.75 and 0.50 log units.

<i>k</i>	Training set (TS)									
	$\Delta < 0.75$					$\Delta < 0.50$				
	2	3	4	5	6	2	3	4	5	6
q^2	0.741	0.829	0.831	0.831	0.831	0.784	0.856	0.859	0.856	0.856
RMSE	0.424	0.362	0.365	0.367	0.367	0.409	0.356	0.360	0.364	0.364
Predicted compounds	47	36	34	33	33	41	31	30	29	29
Prediction rate	19%	14%	13%	13%	13%	16%	12%	12%	11%	11%
<i>k</i>	Validation set (VS)									
	$\Delta < 0.75$					$\Delta < 0.50$				
	2	3	4	5	6	2	3	4	5	6
r^2	0.564	0.598	0.707	0.707	0.707	0.637	0.772	0.779	0.779	0.779
RMSE	0.620	0.620	0.643	0.643	0.643	0.684	0.637	0.661	0.661	0.661
Predicted compounds	22	17	15	15	15	16	14	13	13	13
Prediction rate	12%	9%	8%	8%	8%	9%	8%	7%	7%	7%

<http://dx.doi.org/10.14573/altex.1405091s>



	QMRF identifier (JRC Inventory): To be entered by JRC
	QMRF Title: kNN models for sub-chronic (90-days) oral repeated dose toxicity (RDT) in rats
	Printing Date: 30-giu-2014

1. QSAR identifier

1.1. QSAR identifier (title):

kNN models for sub-chronic (90-days) oral repeated dose toxicity (RDT)
in rats

1.2. Other related models:

1.3. Software coding the model:

2. General information

2.1. Date of QMRF:

26 June 2014

2.2. QMRF author(s) and contact details:

[1]Domenico Gadaleta Dipartimento di Farmacia Scienze del Farmaco, Università degli Studi di Bari "Aldo Moro", Via E. Orabona 4, 70125 Bari, Italy domenico.gadaleta@uniba.it

[2]Fabiola Pizzo Laboratory of Environmental Chemistry and Toxicology, IRCCS-Istituto di Ricerche Farmacologiche Mario Negri, Via La Masa 19, 20156 Milano, Italy fabiola.pizzo@marionegri.it

[3]Anna Lombardo Laboratory of Environmental Chemistry and Toxicology, IRCCS-Istituto di Ricerche Farmacologiche Mario Negri, Via La Masa 19, 20156 Milano, Italy anna.lombardo@uniba.it

[4]Angelo Carotti Dipartimento di Farmacia Scienze del Farmaco, Università degli Studi di Bari "Aldo Moro", Via E. Orabona 4, 70125 Bari, Italy angelo.carotti@uniba.it

[5]Sylvia E. Escher Fraunhofer Institute for Toxicology and Experimental Medicine (ITEM), Nikolai-Fuchs-Strasse 1, 30625 Hannover, Germany.

[6]Orazio Nicolotti Dipartimento di Farmacia Scienze del Farmaco, Università degli Studi di Bari "Aldo Moro", Via E. Orabona 4, 70125 Bari, Italy orazio.nicolotti@uniba.it

[7]Emilio Benfenati Laboratory of Environmental Chemistry and Toxicology, IRCCS-Istituto di Ricerche Farmacologiche Mario Negri, Via La Masa 19, 20156 Milano, Italy emilio.benfenati@marionegri.it

2.3. Date of QMRF update(s):

2.4. QMRF update(s):

2.5. Model developer(s) and contact details:

[1]Domenico Gadaleta Dipartimento di Farmacia Scienze del Farmaco, Università degli Studi di Bari "Aldo Moro", Via E. Orabona 4, 70125 Bari, Italy domenico.gadaleta@uniba.it

[2]Fabiola Pizzo Laboratory of Environmental Chemistry and Toxicology, IRCCS-Istituto di Ricerche Farmacologiche Mario Negri, Via La Masa 19, 20156 Milano, Italy fabiola.pizzo@marionegri.it

[3]Anna Lombardo Laboratory of Environmental Chemistry and Toxicology, IRCCS-Istituto di Ricerche Farmacologiche Mario Negri, Via La Masa 19, 20156 Milano, Italy anna.lombardo@marionegri.it

[4]Angelo Carotti Dipartimento di Farmacia Scienze del Farmaco, Università degli Studi di Bari "Aldo Moro", Via E. Orabona 4, 70125 Bari, Italy angelo.carotti@uniba.it

[5]Sylvia E. Escher Fraunhofer Institute for Toxicology and Experimental Medicine (ITEM), Nikolai-Fuchs-Strasse 1, 30625 Hannover, Germany



[6]Orazio Nicolotti Dipartimento di Farmacia Scienze del Farmaco, Università degli Studi di Bari “Aldo Moro”, Via E. Orabona 4, 70125 Bari, Italy orazio.nicolotti@uniba.it

[7]Emilio Benfenati Laboratory of Environmental Chemistry and Toxicology, IRCCS-Istituto di Ricerche Farmacologiche Mario Negri, Via La Masa 19, 20156 Milano, Italy emilio.benfenati@marionegri.it

2.6.Date of model development and/or publication:

2.7.Reference(s) to main scientific papers and/or software package:

[1]Sakuratani, Y., Zhang, H., Nishikawa, S., et al. (2013). Hazard Evaluation Support System (HESS) for predicting repeated dose toxicity using toxicological categories. SAR QSAR Env. Res. 24, 351-363. <http://dx.doi.org/10.1080/1062936X.2013.773375>

[2]De Julián-Ortiz, J. V., García-Domenech, R., Gálvez, L., et al. (2005). Predictability and prediction of lowest observed adverse effect levels in a structurally heterogeneous set of chemicals. SAR QSAR Environ. Res. 16, 263-272. <http://dx.doi.org/10.1080/10659360500036927>

[3]García-Domenech, R., de Julián-Ortiz, J. V., and Besalú, E. (2006) True prediction of lowest observed adverse effect levels. Mol. Diversity 10, 159–168 <http://dx.doi.org/10.1007/s11030-005-9007-z>

[4]Mazzatorta, P., Estevez, M. D., Coulet, M., et al. (2008). Modeling oral rat chronic toxicity. J. Chem. Inf. Model. 48, 1949-1954. <http://dx.doi.org/10.1021/ci8001974>

[5]Venkatapathy, R., Moudgal, C.J. and Bruce, R.M. (2004). Assessment of the oral rat chronic lowest observed adverse effect level model in TOPKAT, a QSAR software package for toxicity prediction. J. Chem. Inf. Comput. Sci. 44, 1623-1629. <http://dx.doi.org/10.1021/ci049903s>

2.8.Availability of information about the model:

Training set is available. Test set is under copyright and is not available.

2.9.Availability of another QMRF for exactly the same model:

None to date

3.Defining the endpoint - OECD Principle 1

3.1.Species:

Rattus norvegicus

3.2.Endpoint:

4. Human health effects 4.14 Repeated dose toxicity

3.3.Comment on endpoint:

Sub-chronic oral toxicity test: repeated dose 90-day oral (gavage, diet, drinking water) toxicity study in rodents [EC B.26, OECD 408].

3.4.Endpoint units:

mg/Kg body weight /day

3.5.Dependent variable:

logLO(A)EL

3.6.Experimental protocol:

The sub-chronic toxicity to rats was determined using the OECD 408 test guideline [ref 1, sect 9.2].

The experimental determination of LO(A)EL is closely dependent on the protocol employed. The test substance is administered daily at different doses to several groups of animals, one dose level per group, for a given period. During this period the animals are observed for signs



of toxicity. Then, the highest dose at which no adverse effects are noted (NO(A)EL) and the lowest dose at which an adverse effect is noted (LO(A)EL) are determined.

3.7. Endpoint data quality and variability:

The data were taken from three different databases: Munro and HESS (taken from OECD QAR Toolbox [ref 2, sect 9.2]) and EPA's IRIS (<http://cfpub.epa.gov/ncea/iris/index.cfm?fuseaction=iris.showSubstanceList>) When multiple data were available for the same compound, the lowest value has been retained. LO(A)EL data have not been extrapolated from a dose-response curve, and there is intrinsic uncertainty in the LO(A)EL experimental data that is related to the choice of administration doses and dose spacing. The spacing differed even in the same experiment and did not always seem to follow any specific pattern.

4. Defining the algorithm - OECD Principle 2

4.1. Type of model:

k-Nearest Neighbor

4.2. Explicit algorithm:

k-Nearest Neighbor

The 5 models are based on a k-Nearest Neighbors algorithm that can estimate the logLO(A)EL on the basis of a different number of similar compounds (from 2 to 6) in the Training set.

The prediction for target chemical is generated as the arithmetical mean of the experimental values of the k nearest neighbors in the Training set.

The similarity between chemicals is estimated by means of a combined similarity index (SI) resulting from the weighted combination of a binary fingerprint array and three non-binary structural keys based on topological descriptors as defined by the software Dragon. Fingerprint and structural keys are combined as follows:

$$SI(A,B) = Sb(FPa,FPb)^{Wfp} * Snb(CDa,CDb)^{Wcd} * Snb(HDa,HDb)^{Whd} * Snb(FGa,FGb)^{Wfg} \text{ where:}$$

A and B are two molecules to be compared; FPa, CDa, HDa, FGa, FPb, CDb, HDb, FGb are the Fingerprints, Constitutional Descriptors and Functional Groups keys respectively calculated on the two molecules A and B (see 4.3); Sb(Xa,Xb) is the result of the application of a binary similarity coefficient to two fingerprints Xa and Xb, where the resulting values are in the interval [0,1]; Snb(Xa,Xb) is the result of the application of a non-binary similarity coefficient to two descriptors based keys Xa and Xb, where the resulting values are in the interval [0,1]; Wfp, Wcd, Whd, Wfg are the relative weights of the four contributions, under the condition that the sum of the four weights is equal to 1.

4.3. Descriptors in the model:

[1] Extended Fingerprints Fingerprints as described in Daylight [ref 3, sect 9.2] with additional bits that take into account ring features

[2] Constitutional Descriptors (CD) key Structural key made of 35 constitutional descriptors, as defined in Dragon



[3]Heteroatom descriptors (HD) key Structural key made of 11 different hetero-atom counts, as defined in Dragon.

[4]Functional Groups (FG) keys Structural key made of 154 functional groups, as defined in Dragon

4.4.Descriptor selection:

4.5.Algorithm and descriptor generation:

4.6.Software name and version for descriptor generation:

Dragon

derivation of descriptors for structural keys

http://www.taletе.mi.it/products/dragon_description.htm

Chemistry Development Kit

fingerprints algorithm

<http://sourceforge.net/projects/cdk/>

VEGA

implementation of the similarity index

<http://www.vega-qsar.eu/>

4.7.Chemicals/Descriptors ratio:

5.Defining the applicability domain - OECD Principle 3

5.1.Description of the applicability domain of the model:

A target chemical must fulfill all the following conditions to be considered reliably predicted:

1. Nearest neighbors among the k selected for the prediction of the query compound should have a SI ≥ 0.85 otherwise they do not participate in the prediction stage.
 - 1a. If there are no neighbors matching at least this similarity threshold, the model does not provide a prediction value for the target compound.
 - 1b. If two or more neighbors fulfill this condition, the difference between the maximum and minimum experimental values among retained neighbors is considered. If this difference is < 1 log unit (all the neighbors have similar values), the target is predicted as the average of the neighbors' values, otherwise the model does not return any prediction.
 - 1c. If the prediction of the target is based on a single neighbor, SI ≥ 0.90 for obtaining a prediction (which is equal in this case to the experimental values of the neighbor).
2. The algorithm verifies how the target's nearest neighbor is predicted in LOO (leave-one-out) internal cross-validation.
 - 2a. If the error in internal prediction (resulting from LOO cross-validation) of the query's nearest neighbor is ≥ 1 log units, the query's prediction is rejected.
 - 2b. If the nearest neighbor is not predicted in LOO cross-validation because it lacks other nearest neighbors (SI ≥ 0.85) for the read-across,



the query's prediction is retained.

2c. If the neighbor has at least one other similar compound within the training set, but it is not predicted because it not fulfill others conditions (e.g., differences among experimental values for its similar compounds), the query prediction is rejected.

5.2.Method used to assess the applicability domain:

The parameters have been choosed in order to obtain good performance in LOO internal cross-vaidation on the Training set.

5.3.Software name and version for applicability domain assessment:

Micorsoft Office Professional Plus 2010
for performance evaluation and algorithm development

VEGA
for similarity calculation
<http://www.vega-qsar.eu/>

5.4.Limits of applicability:

see 5.1

6.Internal validation - OECD Principle 4

6.1.Availability of the training set:

Yes

6.2.Available information for the training set:

CAS RN: Yes
Chemical Name: No
Smiles: Yes
Formula: No
INChI: No
MOL file: No

6.3.Data for each descriptor variable for the training set:

Unknown

6.4.Data for the dependent variable for the training set:

All

6.5.Other information about the training set:

254 data point

6.6.Pre-processing of data before modelling:

Only values referred for sub-chronic toxicity studies (from 84 to 98 days) of oral exposure (gavage, diet, or drinking water) were taken into account. Only data related to studies on rats (*Rattus Norvegicus*) were considered. Data related to reproductive effects in females were rejected. Inorganic compounds, isomeric mixtures, metal complexes and the data related to mixtures of chemicals were rejected. Ionized structures were neutralized and counterions eliminated. The LO(A)EL numerical values were converted in a logarithmic scale.

6.7.Statistics for goodness-of-fit:

**6.8. Robustness - Statistics obtained by leave-one-out cross-validation:**

k; 2 ;3; 4; 5; 6

q2; 0.632; 0.747; 0.769; 0.764; 0.76

RMSE; 0.478; 0.405; 0.396; 0.400; 0.407

Predicted compounds; 68; 51; 49; 49; 47

Prediction rate; 27%; 20%; 19%; 19%; 19%

6.9. Robustness - Statistics obtained by leave-many-out cross-validation:**6.10. Robustness - Statistics obtained by Y-scrambling:****6.11. Robustness - Statistics obtained by bootstrap:****6.12. Robustness - Statistics obtained by other methods:****7. External validation - OECD Principle 4****7.1. Availability of the external validation set:**

No

7.2. Available information for the external validation set:

CAS RN: No

Chemical Name: No

Smiles: No

Formula: No

INChI: No

MOL file: No

7.3. Data for each descriptor variable for the external validation set:

Unknown

7.4. Data for the dependent variable for the external validation set:

No

7.5. Other information about the external validation set:

179 data point

7.6. Experimental design of test set:

RepDose database provided by Fraunhofer ITEM <http://www.fraunhofer-repdose.de/>

7.7. Predictivity - Statistics obtained by external validation:



k; 2; 3; 4; 5; 6

r²; 0.543; 0.552; 0.554; 0.675; 0.682

RMSE; 0.612; 0.659; 0.654; 0.627; 0.642

Predicted compounds; 27; 24; 23; 21; 20

Prediction rate; 15%; 13%; 13%; 12%; 11%

7.8. Predictivity - Assessment of the external validation set:

Statistics reported in 7.7 refers only to compounds that fulfill all the algorithm conditions (see 5.1)

7.9. Comments on the external validation of the model:

8. Providing a mechanistic interpretation - OECD Principle 5

8.1. Mechanistic basis of the model:

8.2. A priori or a posteriori mechanistic interpretation:

8.3. Other information about the mechanistic interpretation:

RDT is a composite endpoint, and the determination of LO(A)EL is a complex assessment that is related to a wide variety of toxic effects against different organs and tissues (i.e., nephrotoxicity, hepatotoxicity). Consequently, there are a plethora of different mechanisms of action that determines LO(A)EL. The herein *k*-NN algorithm predict toxicity by simply comparing the target compound to a restricted pool of structurally similar molecules, that plausibly generate toxicity with similar mechanisms. Thus, it does not try to create a correlation, by means of a mathematical equation, between structural features of a molecules and its toxicity.

9. Miscellaneous information

9.1. Comments:

9.2. Bibliography:

[1] OECD (1988). Test No. 408: Repeated Dose 90-day Oral Toxicity Study in Rodents, OECD Guidelines for the Testing of Chemicals, Section 4, OECD Publishing.

<http://dx.doi.org/10.1787/9789264070707-en>

[2] QSAR Toolbox <http://www.qsartoolbox.org/>

[3] Daylight Chemical Information System Inc.

<http://www.daylight.com/dayhtml/doc/theory/theory.finger.html>

9.3. Supporting information:

Training set(s) Test set(s) Supporting information

10. Summary (JRC QSAR Model Database)



10.1.QMRF number:

To be entered by JRC

10.2.Publication date:

To be entered by JRC

10.3.Keywords:

To be entered by JRC

10.4.Comments:

To be entered by JRC