	QMRF identifier (JRC Inventory): **To be entered by JRC**
	QMRF Title: **Carcinogenicity model (CAESAR) v 2.1.10**
	Printing Date: **June 1 2022**

1.QSAR identifier

1.1.QSAR identifier (title):

Carcinogenicity model (CAESAR) v 2.1.10

1.2.Other related models:

NA

1.3.Software coding the model:

VEGA (https://www.vegahub.eu/)

The VEGA software provides QSAR models to predict tox, ecotox, environ, phys-chem and toxicokinetic properties of chemical substances.

emilio.benfenati@marionegri.it

2.General information

2.1.Date of QMRF:

June 2022

2.2.QMRF author(s) and contact details:

[1] Manuela Pavan S-In Soluzioni Informatiche Soluzioni Informatiche Srl Via Ferrari 14, I-36100Vicenza manuela.pavan@s-in.it http://www.s-in.it

[2]Simona Kovarich S-In Soluzioni Informatiche Soluzioni Informatiche Srl Via Ferrari 14, I-36100Vicenza simona.kovarich@s-in.it http://www.s-in.it

[3] Erika Colombo – Istituto di ricerche farmacologiche Mario Negri – erika.colombo@marionegri.it

2.3.Date of QMRF update(s):

NA

2.4.QMRF update(s):

NA

2.5.Model developer(s) and contact details:

[1] Natalja Fjodorova National Institute of Chemistry, Hajdrihova 19, SI-1001 Ljubljana, Slovenia

[2] Marjan Vrako National Institute of Chemistry, Hajdrihova 19, SI-1001 Ljubljana, Slovenia

[3] Marjana Novi National Institute of Chemistry, Hajdrihova 19, SI-1001 Ljubljana, Slovenia

[4] Alberto Manganaro RCCS-Istituto di Ricerche Farmacologiche Mario Negri Via La Masa 19, 20156 Milano, Italy alberto.manganaro@marionegri.it

2.6.Date of model development and/or publication:

2010

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

[1] Natalja Fjodorova, Marjan Vrako, Marjana Novi, Alessandra Roncaglioni and Emilio Benfenati. New public QSAR model for carcinogenicity. Chemistry Central Journal 2010, 4(Suppl1): S3 doi:10.1186/1752-153X-4-S1-S3 http://www.journal.chemistrycentral.com/content/4/S1/S3

[2] Benfenati E, Manganaro A, Gini G. VEGA-QSAR: AI inside a platform for predictive toxicology

Proceedings of the workshop "Popularize Artificial Intelligence 2013", December 5th 2013, Turin, Italy

Published on CEUR Workshop Proceedings Vol-1107

2.8.Availability of information about the model:

The model is non-proprietary and the training set is available.

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

Another QMRF is not available.

3.Defining the endpoint - OECD Principle 1

3.1.Species:

Male or female rats

3.2.Endpoint:

TOX 7.7. Carcinogenicity (in vivo)**3.3.Comment on endpoint:**

Carcinogenicity is a very complex biochemical phenomenon involving processes at the cellular level. The carcinogenicity of a substance depends on its molecular structure and a certain number of phenomena which are only partially known. Typically, the carcinogenic process involves one or more processes, showing a relationship with the mutagenic potential of a substance, but other processes are possible for carcinogens which are non mutagenic

The dataset is based on the classification based on TD50 (which is the dose that produces an

increase of 50% of the tumors in animals) value for rat; the label carcinogen is if TD50 is positive, otherwise is not carcinogen

3.4.Endpoint units:

Adimensional

3.5.Dependent variable:

The dependent variable is cancerogenic effect on rat, as binary classification: 0 (non-carcinogen), 1 (carcinogen)

3.6.Experimental protocol:

Experimental values for carcinogenicity were taken fromhttp://www.epa.gov/ncct/dsstox/sdf_cpdbas.html

3.7.Endpoint data quality and variability:

The initial dataset of 1481 chemicals was taken from Distributed Structure-Searchable Toxicity (DSSTox) Public Database Network http://www.epa.gov/ncct/dsstox/sdf_cpdbas.html which was built from the Lois Gold Carcinogenic Database (CPDB). The initial dataset has been cleaned of all incorrect structures, ambiguous or mixed structures, polymers, inorganic compounds, metallo-organic compounds, salts, complexes and compounds without well-defined structure.

The obtained data and structures of chemicals

were cross-checked by at least two partners using the following online databases: ChemFinder [47], ChemIDPlus [48] and PubChem Compound [49]. The final data set of 806 chemicals, with their ID number, chemical name, CASRN and corresponding binary carcinogenicity classes

4.Defining the algorithm - OECD Principle 2

4.1.Type of model:

This model is based on Counter Propagation Neural Network CP ANN

4.2.Explicit algorithm:

Counter Propagation Artificial Neural Network (CP ANN)

CP ANN consists of two layers of neurons arranged in a two-dimensional rectangular matrix. For more information see Fjodorova et al. 2010 [1]

4.3.Descriptors in the model:

[1]PW5 path/walk 5 - Randic shape index

[2]D/Dr06 distance/detour ring index of order 6

[3]MATS2p Moran autocorrelation - lag 2 / weighted by atomic polarizabilities

[4]EEig10x Eigenvalue 10 from edge adj. matrix weighted by edge degrees

[5]ESpm11 Spectral moment 11 from edge adj. matrix weighted by edge degrees

[6]ESpm09 Spectral moment 09 from edge adj. matrix weighted by dipole moments

[7]GGI2 topological charge index of order 2

[8]JGI6 mean topological charge index of order 6

[9]nRNNOx number of N-nitroso groups (aliphatic)

[10]nPO4 number of phosphates / thiophosphates

[11]N-067 number of Al2-NH atom centered fragments

[12]N-078 number of Ar-N=X / X-N=X atom centered fragments

4.4.Descriptor selection:

Descritpor selection was performed using cross correlation matrix, multicolinearity and fisher ratio techniques. As a result descriptors space was reduced from 835 to 12 descriptors listed in 4.3

4.5.Algorithm and descriptor generation:

The descriptors were calculated, in the original model, by means of dragonX software and are now entirely calculated by an in-house software module in which they are implemented as described in Todeschini & Consonni (2009) [2]

4.6.Software name and version for descriptor generation:

VEGA

https://www.vegahub.eu

https://www.vegahub.eu/portfolio-item/sarpy/

4.7.Chemicals/Descriptors ratio:

645 chemicals/ 12 descriptors = 53.75

5.Defining the applicability domain - OECD Principle 3

5.1.Description of the applicability domain of the model:

The Applicability Domain (AD) is assessed using the original algorithm implemented within VEGA. An overall AD index is calculated, based on a number of parameters, which relate to the results obtained on similar chemicals within the training and test sets.

If 1 ≥ AD index ≥ 0.8, the predicted substance is into the Applicability Domain of the model. It corresponds to good reliability of prediction.

If 0.8 > AD index ≥ 0.6, the predicted substance could be out of the Applicability Domain of the model. It corresponds to moderate reliability of prediction.

If AD index < 0.6, the predicted substance is out of the Applicability Domain of the model and corresponds to low reliability of prediction.

5.2.Method used to assess the applicability domain:

The Applicability Domain and chemical similarity is measured with the algorithm developed for VEGA. Full details are in the VEGA website (www.vegahub.eu), including the open access paper describing it [3]. The AD also evaluates the correctness of the prediction on similar compounds (accuracy), the consistency between the predicted value for the target compound and the experimental values of the similar compounds, the range of the descriptors, and the presence of unusual fragments, using atom centred fragments.

Similar molecules with known experimental value:

This index takes into account how similar are the first two most similar compounds found. Values near 1 mean that the predicted compound is well represented in the dataset used to build the model, otherwise the prediction could be an extrapolation. Defined intervals are:

If 1 ≥ index > 0.80, strongly similar compounds with known experimental value in the training set have been found.

If 0.80 ≥ index > 0.6, only moderately similar compounds with known experimental value in the training set have been found.

If index ≤ 0.6, no similar compounds with known experimental value in the training set have been found.

Accuracy (average error) of prediction for similar molecules:

This index takes into account the classification accuracy in prediction for the two most similar compounds found. Values near 1 mean that the predicted compounds fall in an area of the model's space where the model gives reliable predictions (no misclassifications), otherwise the lower is the value, the worse the model behaves. Defined intervals are:

If 1 ≥ index > 0.90, accuracy of prediction for similar molecules found in the training set is good

If 0.9 ≥ index > 0.5, accuracy of prediction for similar molecules found in the training set is not optimal

If index ≤ 0.5, accuracy of prediction for similar molecules found in the training set is not adequate

Concordance for similar molecules:

This index takes into account the difference between the predicted value and the experimental values of the two most similar compounds. Values near 0 mean that the prediction made disagrees with the values found in the model's space, thus the prediction could be unreliable. Defined intervals are:

If 1 ≥ index > 0.90, molecules found in the training set have experimental values that agree with the target compound predicted value

If 0.9 ≥ index > 0.5,, similar molecules found in the training set have experimental values that slightly disagree with the target compound predicted value

If index ≤ 0.5, similar molecules found in the training set have experimental values that completely disagree with the target compound predicted value

Atom Centered Fragments similarity check: This index takes into account the presence of one or more fragments that aren't found in the training set, or that are rare fragments. First order atom centered fragments from all molecules in the training set are calculated, then compared with the first order atom centered fragments from the predicted compound; then the index is calculated as following: a first index RARE takes into account rare fragments (those who occur less than three times in the training set), having value of 1 if no such fragments are found, 0.85 if up to 2 fragments are found, 0.7 if more than 2 fragments are found; a second index NOTFOUND takes into account not found fragments, having value of 1 if no such fragments are found, 0.6 if a fragments is found, 0.4 if more than 1 fragment is found. Then, the final index is given as the product RARE * NOTFOUND. Defined intervals are:

If index = 1, all atom centered fragment of the compound have been found in the compounds of the training set

If 1 > index ≥ 0.7, some atom centered fragment of the compound have not been found in the compounds of the training set or are rare fragments

If index < 0.7, a prominent number of atom centered fragments of the compound have not been found in the compounds of the training set or are rare fragments

Model descriptors range check:

This index checks if the descriptors calculated for the predicted compound are inside the range of descriptors of the training and test set. The index has value 1 if all descriptors are inside the range, 0 if at least one descriptor is out of the range. Defined intervals are:

Index = TRUE, descriptors for this compound have values inside the descriptor range of the compounds of the training set

Index = FALSE, descriptors for this compound have values outside the descriptor range of the compounds of the training set

Model assignment reliability:

This index checks if the two neural network output values (positive and non-positive) lead to an unreliable prediction; when the difference between these two values is lower than 0.1, the neuron where the predicted compound falls can not provide a good classification, thus the index is set to 0. Otherwise the index is set to 1

If index = 1, model class assignment is well defined

If index = 0, model class assignment is uncertain

Neural map neurons concordance:

This index checks the concordance of the predicted compound with the experimental values of the other compounds that falls int the same neuron. The index is built considering two sub-indices: Population (the number of compounds found in the neuron) and Concordance (the number of compounds in the neuron that have experimental value matching with current prediction divided by the number of compounds in the neuron). Low values mean that the predicted compound falls in a zone of the neural network that has no experimental compounds, or that has experimental compounds with eterogeneous experimental values, thus leading to a low reliability of the prediction.

If index = 1, predicted value agrees with experimental values of training set compounds laying in the same neuron

If index = 0.75, predicted value disagrees with experimental values of training set compounds laying in the same neuron

If index = 0.5, predicted substance falls into a neuron that is populated by no compounds of the training set

5.3.Software name and version for applicability domain assessment:

VEGA

Included in the VEGA software and automatically displayed when running the model

emilio.benfenati@marionegri.it

https://www.vegahub.eu/

5.4.Limits of applicability:

The model is not applicable to inorganic chemicals and substances containing unusual elements (i.e., different from C, O, N, S, P, Cl, Br, F, I). Salts can be predicted only if converted to the neutralized form.

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

NanoMaterial: No

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

All

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

All

6.5.Other information about the training set:

n 645

6.6.Pre-processing of data before modelling:

NA

6.7.Statistics for goodness-of-fit:

Training set: n = 645, Balanced Accuracy 0.88, Sensitivity 0.89, Specificity 0.68, MCC 0.75. TP 298, TN 268, FP 44, FN 35

Accuracy 0,87: Specificity 0,86; Sensitivity 0.89;

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

NA

6.9.Robustness - Statistics obtained by leave-many-out cross-validation:

NA

6.10.Robustness - Statistics obtained by Y-scrambling:

NA

6.11.Robustness - Statistics obtained by bootstrap:

NA

6.12.Robustness - Statistics obtained by other methods:

NA

7.External validation - OECD Principle 4

7.1.Availability of the external validation set:

Yes

7.2.Available information for the external validation set:

CAS RN: Yes

Chemical Name: Yes

Smiles: Yes

Formula: No

INChI: No

MOL file: No

NanoMaterial: No

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

All

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

All

7.5.Other information about the external validation set:

n: 161

7.6.Experimental design of test set:

NA

7.7.Predictivity - Statistics obtained by external validation:

Test set: n = 161, Balanced Accuracy 0.66, Sensitivity 072, Specificity 0.60, MCC 0.32. TP 64, TN 43, FP 29, FN 25

Test set in AD: n = 55, Balanced Accuracy 0.90, Sensitivity 0.94, Specificity 0.87, MCC 0.81. TP 30, TN 20, FP 3, FN 2.

Test set could be out of AD: n = 39, Balanced Accuracy 0.72, Sensitivity 0.68, Specificity 0.76, MCC 0.44. TP 15, TN 13, FP 4, FN 7

Test set out of AD: n = 67, Balanced Accuracy 0.43, Sensitivity 0.54, Specificity 0.31, MCC -0.15. TP 19, TN 10, FP 22, FN 16

7.8.Predictivity - Assessment of the external validation set:

NA

7.9.Comments on the external validation of the model:

NA

8.Providing a mechanistic interpretation - OECD Principle 5

8.1.Mechanistic basis of the model:

NA

8.2.A priori or a posteriori mechanistic interpretation:

NA

8.3.Other information about the mechanistic interpretation:

NA

9.Miscellaneous information

9.1.Comments:

NA

9.2.Bibliography:

[1]Natalja Fjodorova, Marjan Vrako, Marjana Novi, Alessandra Roncaglioni and Emilio Benfenati.New public QSAR model for carcinogenicity. Chemistry Central Journal 2010, 4(Suppl1):S3doi:10.1186/1752-153X-4-S1-S3 http://www.journal.chemistrycentral.com/content/4/S1/S3

[2]R. Todeschini and V. Consonni, Molecular Descriptors for Chemoinformatics, Wiley-VCH, 2009

[3] Floris, M., Manganaro, A., Nicolotti, O. et al. A generalizable definition of chemical similarity for read-across. J Cheminform 6, 39 (2014). https://doi.org/10.1186/s13321-014-0039-1

9.3.Supporting information:

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

All available dataset are present in the model inside the VEGA software.

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