	QMRF identifier (JRC Inventory): **To be entered by JRC**
	QMRF Title: **Skin Sensitization Model (CAESAR) version 2.1.7**
	Printing Date: April, 2022

1.QSAR identifier

1.1.QSAR identifier (title):

Skin Sensitization Model (CAESAR) version 2.1.7

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:

31/01/2017

2.2.QMRF author(s) and contact details:

[1]Emilio Benfenati Istituto di Ricerche Farmacologiche Mario Negri - IRCSS Via Mario Negri 2,20156 Milano, Italy emilio.benfenati@marionegri.it https://www.marionegri.it/

[2]Gianluca Selvestrel Istituto di Ricerche farmacologiche Mario negri - IRCCSgianluca.selvestrel@marionegri.it

2.3.Date of QMRF update(s):

1) 06/02/2020;

2) April 2022

2.4.QMRF update(s):

1) Modification in sections 2.2 and section 9.3

2) Modification in sections 1.1, 2.9, 3.7, 5.1, 5.2, 6.5, 6.7, 7.5, 7.6, 7.7 and 7.8.

2.5.Model developer(s) and contact details:

[1]Qasim Chaudhry Food & Environment Research Agency, Sand Hutton, York qasim.chaundhry@fera.qsi.gov.uk

[2]Nadège Piclin BioChemics Consulting, 111 Bld. Duhamel du Monceau nadege.piclin@biochemics-consulting.com

[3]Jane Cotterill Food & Environment Research Agency, Sand Hutton, York jane.cotterill@fera.qsi.gov.uk

[4]Marco Pintore BioChemics Consulting, 111 Bld. Duhamel du Monceau marco.pintore@biochemics-consulting.com

[5]Nick R Price Food & Environment Research Agency, Sand Hutton, York nick@technologyforgrowth.co.uk

[6]Jacques R Chrétien BioChemics Consulting, 111 Bld. Duhamel du Monceau jacues.chretien@biochemics-consulting.com

[7]Alessandra Roncaglioni Istituto di Ricerche Farmacologiche Mario Negri – IRCCS alessandra.roncaglioni@marionegri.it

2.6.Date of model development and/or publication:

July 2010

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

[1] Chaudhry, Q., Piclin, N., Cotterill, J., Pintore, M., Price, N. R., Chrétien, J. R. and Roncaglioni, A.(2010). Global QSAR models of skin sensitisers for regulatory purposes., Chem Cent J 4 Suppl 1

S5 https://bmcchem.biomedcentral.com/articles/10.1186/1752-153X-4-S1-S5

[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:

NA

3.Defining the endpoint - OECD Principle 1

3.1.Species:

CBA mice

3.2.Endpoint:

Skin sensitisation on mouse (local lymph node assay model) OECD 429

3.3.Comment on endpoint:

Skin sensitizers are substances able to elicit an allergic response following contact with the skin, termed allergic contact dermatitis (ACD) in humans. Molecules have been classified as “sensitizer” or “non sensitizer”

3.4.Endpoint units:

Adimensional

3.5.Dependent variable:

The model consists in an Adaptive Fuzzy Partition (AFP) based on 8 descriptors

3.6.Experimental protocol:

OECD 429 Test. The methods described here are based on the use of in vivo radioactive labelling to measure an increased number of proliferating cells in the draining auricular lymph nodes.

3.7.Endpoint data quality and variability:

Gerberick GF, Ryan CA, Kern PS, Schaltter H, Dearman RJ, Kimber I, Patlewicz GY, Basketter DA: Compilation of historical local node data for evaluation of skin sensitization alternative methods. Dermatitis. 2005, 16 (4): 157-202.

All the chemical structures were manually checked deleting for example doubtful compounds, mixture, inorganic compounds and tautomers. The final dataset is composed of 209 mono- constituent organic compounds. The dataset was randomly split into training and test set with respectively the 80% (167) and the 20% (42) of the compounds.

4.Defining the algorithm - OECD Principle 2

4.1.Type of model:

The model consists in an Adaptive Fuzzy Partition (AFP) based on 8 descriptors. The AFP produces as output two values that represent the belonging degree respectively to the sensitizer and non-sensitizer classes. The input compound is assigned to the class having this degree value higher than 0.5, unless the difference between the values of the two degrees is lower than the threshold of 0.001; in this case, the

belonging to one class or the other is not sure, thus no prediction is made. The descriptors were calculated, in the original model, by means of MDL and DragonX software and are now entirely calculated by an in-house software module in which they are implemented as described in [2] in point 9.2: R. Todeschini and V. Consonni, Molecular Descriptors for Chemoinformatics, Wiley-VCH, 2009

4.2. Explicit algorithm:

See fer.1, section 9.2

The classification scheme is based on binary classification (see ref. 1, section 9.2). Based on eight fragments, an in house Adaptive Fuzzy Partition model (AFP)[4]. The AFP model for skin sensitisation was built on the training set by using the following parameters: maximal number o frules for each chemical activity = 30; minimal number of compounds for a given rule = 2; maximal number of cuts for each axis = 5. The trapezoidal parameters used were: p/wi= 1.25 and q/wi= 0.45.The AFP method allocates degrees of membership of the different classes for each compound within a 0 to 1 range. Then, a compound is attributed to a given class if its degree of membership is greater than 0.5. The percentage of compounds correctly predicted is computed by comparing their experimental and predicted classes

4.3. Descriptors in the model:

[1]nN: Number of nitrogen atoms

[2]GNar: Narumi geometric topological index

[3]MDDD: Mean Distance Degree Deviation

[4]X2v: Valence connectivity index chi-2

[5]EEig10r: Eigenvalue 10 from the edge adjacency matrix weighted by resonance integrals

[6]GGI8: Topological charge index of order 8

[7]nCconj: Number of non aromatic conjugated C(sp2)[8]O-058: Atom-centred fragment =O

4.4. Descriptor selection:

See [1] and [2] in point 9.2.

To develop robust and reliable models the descriptor space was reduced by extracting the most significant variables. All variables were normalized into -1+1 range and variable selection was performed with a hybrid selection algorithm (HSA). This method combines a genetic algorithm (GA) with a stepwise regression [3]. A stepwise approach was combined with GA in order to reach local convergence as it is quick and adapted to find solutions in "promising" areas already identified. To prevent over-fitting and a poor generalization, a cross validation procedure was included in the algorithm during the selection procedure. Thus, the dataset was randomly divided into training and validation sets in such a way that the fitness score of each chromosome was derived from the combination of the scores of the training and validation sets.

The following parameters were used in the data processing of the sensitisation data set:

- fuzzy parameters: weighting coefficient was set equal to 1.5, tolerance convergence was equal to 0.001, number of iterations was 30 and cluster number was 6;

- genetic parameters: chromosome number used was 10, chromosome size was equal to the total number of descriptors used; initial active descriptors in each chromosome was 8,

- crossover point number was 1, percentage of rejections was set at 0.1, percentage of crossover was 0.8, percentage of mutation was 0.05, number of generations was set at 10;

- stepwise parameters: ascending coefficient was 0.02, descending coefficient was -0.02

4.5. Algorithm and descriptor generation:

see ref 1, section 9.2

4.6. Software name and version for descriptor generation:

MDL and DragonX software

4.7. Chemicals/Descriptors ratio:

167 chemicals (training set)/8 descriptors = 21

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.
Indices are calculated on the first k = 2 most similar molecules, each having S_k similarity value with the target molecule.

Similarity index (*IdxSimilarity*) is calculated as:

$$\frac{\sum_k S_k}{k} \times (1 - Diam^2)$$

where *Diam* is the difference in similarity values between the most similar molecule and the *k*-th molecule.

Accuracy index (*IdxAccuracy*) is calculated as:

$$\frac{\sum_c log\ (1 + S_c)}{\sum_k log\ (1 + S_k)}$$

where the molecules with *c* index are the subset of the *k* molecules where the prediction of the model matches with the experimental value of the molecule.

Concordance index (*IdxConcordance*) is calculated as:

$$\frac{\sum_c log\ (1 + S_c)}{\sum_k log\ (1 + S_k)}$$

where the molecules with *c* index are the subset of the *k* molecules where the experimental value of the molecule matches with the prediction made for the target molecule.

ACF contribution (*IdxACF*) index is calculated as

$$ACF = rare \times missing$$

where: *rare* is calculated on the number of fragments found in the molecule and found in the training set in less than 3 occurences as following: if the number is 0, *rare* is set to 1.0; if the number is 1, *rare* is set to 0.6; otherwise *rare* is set to 0.4

missing is calculated on the number of fragments found in the molecule and never found in the training set as following: if the number is 0, *missing* is set to 1.0; if the number is 1, *missing* is set to 0.6; otherwise *missing* is set to 0.4

Descriptors Range (*IdxDescRange*) index is calculated as 1.0 if all molecular descriptors used in the prediction fall within the range of descriptors used in the whole training set, 0.0 otherwise.

AD final index is calculated as following:

$$ADI = (IdxSimilarity^{0.5} \times IdxAccuracy^{0.25} \times IdxConcordance^{0.25}) \times IdxACF \times IdxDescRange$$

If 1 ≥ AD index ≥ 0.8, the predicted substance is regarded to be in 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 regarded 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 are measured with the algorithm developed for VEGA. Full details are in the VEGA website (www.vegahub.eu), including the open access paper describing it [5]. 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.

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: null

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:

The training set is composed of 167 substances, 133 sensitizer and 34 NON-sensitizer

6.6.Pre-processing of data before modelling:

All the chemical structures were manually checked deleting doubtful compounds, mixture, inorganic compounds and tautomers

6.7.Statistics for goodness-of-fit:

Statistics on the training set

Accuracy: 91% Sensitivity: 95% Specificity: 74%

TP 127, TN 27, FP 7, FN 6

Considering the low number of non-sensitizing compounds, the measure of specificity may be uncertain

Statistics on the test set

The test set is composed of42 substances, 34 sensitizer and 8 NON-sensitizer

Accuracy: 93% Sensitivity: 97% Specificity: 75%

TP 33, TN 6, FP 2, FN 1

Considering the low number of non-sensitizing compounds, the measure of specificity may be uncertain

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:

7.External validation - OECD Principle 4

7.1.Availability of the external validation set:

Yes, on request.

7.2.Available information for the external validation set:

CAS RN: Yes

Chemical Name: No

Smiles: Yes

Formula: No

INChI: No

MOL file: No

NanoMaterial: null

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:

Under the Life project CONCERT REACH (https://www.life-concertreach.eu/) a big dataset was collected from several public sources, listed below:

-SKINSENS DB (https://cwtung.kmu.edu.tw/skinsensdb/search)

Missed CAS number were retrieved from ChemSpider (to avoid a redundancy with the sources checked by the workflow) or from NICEATM DB. Compounds identified with “Formulation” as name or with composed CAS number (xxx/yyy/) were removed because they are mixtures.

-Alves et al., 2015. Predicting chemically-induced skin reactions. Part I: QSAR models of skin sensitization and their application to identify potentially hazardous compounds. Toxicol Appl Pharmacol. 2015 Apr 15;284(2):262-72. doi: 10.1016/j.taap.2014.12.014. Epub 2015 Jan 3.

-NICEATM LLNA database 2013. National Toxicology Program. https://ntp.niehs.nih.gov/iccvam/methods/immunotox/niceatm-llnadatabase-23dec2013.xls Identified with acronym NA including 1060 quantitative EC3% values (multiple values for the same compound)

-VEGA MODEL SKIN SENSITIZATION CAESAR dataset

-VEGA MODEL SKIN SENSITIZATION IRFMN-JRC dataset

-Jaworska et al., 2015. Bayesian integrated testing strategy (ITS) for skin sensitization potency assessment: a decision support system for quantitative weight of evidence and adaptive testing strategy. Arch Toxicol (2015) 89:2355–2383

-Natsch et al., 2015. Predicting Skin Sensitizer Potency Based on In Vitro Data from KeratinoSens and Kinetic Peptide Binding: Global Versus Domain-Based Assessment. TOXICOLOGICAL SCIENCES, 143(2), 2015, 319–332.

-Strickland et al., 2017. Multivariate models for prediction of human skin sensitization hazard. J. Appl. Toxicol. 2017; 37: 347–360.

-QSAR Toolbox (extraction 2019) using the database SKIN SENSITIZATION and DATA EXTRACTOR for LLNA.

After a further check of the data (i.e. duplicates, multi-constituent and UVCB substances not eliminated), the final database of skin sensitization includes 623 compounds with univocal LLNA assessment: 178 non sensitizers, 445 sensitizers.

7.6.Experimental design of test set:

7.7.Predictivity - Statistics obtained by external validation:

ADI thresholds were applied to calculate the performances of the model on the external validation dataset (453 compounds).

143 compounds in AD (ADI>=0.8)

Sensitivity	Specificity	Accuracy	MCC
0.94	0.25	0.78	0.25

TP 104, TN8, FP 24, FN 7

66 compounds could be out AD (0.8> ADI ≥ 0.6)

Sensitivity	Specificity	Accuracy	MCC
0.94	0.22	0.74	0.23

TP 45, TN 4, FP 14, FN 3

244 compounds out of AD (ADI< 0.6)

Sensitivity	Specificity	Accuracy	MCC
0.74	0.35	0.57	0.10

TP 103, TN37, FP 68, FN 36

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:

A posteriori

8.3.Other information about the mechanistic interpretation:

NA

9.Miscellaneous information

9.1.Comments:

NA

9.2.Bibliography:

[1]Chaudhry, Q., Piclin, N., Cotterill, J., Pintore, M., Price, N. R., Chrétien, J. R. and Roncaglioni,A.(2010). Global QSAR models of skin sensitisers for regulatory purposes., Chem Cent J https://bmcchem.biomedcentral.com/articles/10.1186/1752-153X-4-S1-S5

[2]R. Todeschini and V. Consonni, Molecular Descriptors for Chemoinformatics, Wiley-VCH, 2009https://onlinelibrary.wiley.com/doi/book/10.1002/9783527628766

[3]Ros F, Pintore M, Chretien JR. (2002) Molecular description selection combining geneticalgorithms and fuzzy logic: application to database mining procedures. Chemom Intell Lab Syst. 63 15-26

[4]Ros F., Taboureau O., Pintore M., J.R.Chrétien (2003) Development of predictive models by adaptive fuzzy partitioning. Application to compounds active on the central nervous system. Chemometrics and Intelligent Laboratory Systems 67 29-50

[5] 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:

Training and test sets 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