	QMRF identifier (JRC Inventory): **To be entered by JRC**
	QMRF Title: **Estrogen Receptor Relative Binding Affinity Model v- 1.0.2**
	Printing Date: **June 7 2022**

1.QSAR identifier

1.1.QSAR identifier (title):

Estrogen Receptor Relative Binding Affinity Model v- 1.0.2

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]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]Alessandra Roncaglioni Istituto di Ricerche Farmacologiche Mario Negri - IRCSS Via Mario Negri2, 20156 Milano, Italy alessandra.roncaglioni@marionegri.it https://www.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]Alberto Manganaro Istituto di Ricerche Farmacologiche Mario Negri - IRCSS Via Mario Negri 2,20156 Milano, Italy alberto.manganaro@marionegri.it https://www.marionegri.it/

[2]Alessandra Roncaglioni Istituto di Ricerche Farmacologiche Mario Negri - IRCSS Via Mario Negri2, 20156 Milano, Italy alessandra.roncaglioni@marionegri.it https://www.marionegri.it/

2.6.Date of model development and/or publication:

2008

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

[1] A. Roncaglioni, N. Piclin, M. Pintore, E. Benfenati, "Binary classification models for endocrine disrupter effects mediated through the estrogen receptor", SAR and QSAR in Environmental Research (2008), 19, 7-8 https://www.tandfonline.com/doi/abs/10.1080/10629360802550606

[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:

Human

3.2.Endpoint:

Toxicology Estrogen Receptor Relative Binding Affinity

3.3.Comment on endpoint:

Experimentally determined values of human ER alpha for receptor binding assay (RBA), expressed as percentage of activity using 17-estradiol as reference

3.4.Endpoint units:

percentage Activity (% Activity)

3.5.Dependent variable:

Relative Binding Affinity (RBA)

3.6.Experimental protocol:

The tested substance is added to a system where radio-labeled reference hormone binds to a prescribed quantity of hormone receptor. The chemical concentration that inhibits 50% of the binding of the reference hormone to the receptor is measured and defined as IC50. Then, Relative Binding Affinity (RBA) between IC50 values of the chemical and natural hormone (E2) is defined as the endpoint when the IC50 concentration of natural hormone is set at 100.

3.7.Endpoint data quality and variability:

As a source of activity data, the Japanese METI database was used [2]. It contains experimentally determined values of human ER alpha for both receptor binding (RBA) and reporter gene (RA) assays, expressed as percentage of activity using 17-estradiol as reference. It represents a heterogeneous dataset of compounds, including natural and synthetic steroids, drugs and chemical contaminants such as pesticides, PCBs, and phthalates. To develop binary classification models in this study any detectable activity in the test was associated with the ‘active’ class while those compounds with no detectable activity were labeled ‘inactive’.

Chemical structures were sketched and for salts the free acid or basis form was used. Adopting a very simple approach only a 2D configuration of the molecules was used, while the 3D conformation and steric configuration were ignored. For this reason, 2D duplicates of different 3D isomers were included only once, verifying that the associated activity class was comparable for all possible forms sharing the same 2D structure (only two structures did not satisfy this requirement and were discarded). A further 100 compounds whose RBA values were not determined were excluded to consider both the endpoints on a similar basis. The final dataset comprised 806 single 2D structures, with the majority of the compounds considered inactive.

The dataset was split into training (656 chemicals) and test (150 chemicals).

4.Defining the algorithm - OECD Principle 2

4.1.Type of model:

QSAR classification model based on a classification and regression tree (CART).Compounds with any detected activity were labeled as positive while compounds with lack of ant activity where labelled as negative

4.2.Explicit algorithm:

Classification and regression tree (CART)uses the methodology of tree building as a hierarchical classification method. CART formulates simple ‘if/then’ rules for binary recursive partitioning of all the objects into smaller subgroups, where the compounds belonging to the dataset, represented by a ‘node’ in

a decision tree, can be split into only two groups. Thus, each node can be split into two new ‘branches’. The goal of this process is to maximize homogeneity of the values of the dependent variable Y in the various subgroups taking into account the fact that different relationships may hold among variables in different parts of the data. The cost for the errors was considered equally important for the two classes, while the a priori probability for the two classes was set as equal. Automatic stepwise selection of variables was used in CART. All splits are ranked and the variable that achieves the highest purity at root is selected. The Gini measure for node impurity was used. The performances on the validation set were used to efficiently select the well dimensioned tree; then, the best tree was evaluated on the test set. In this study we used the CART algorithm as implemented in Statistica software

4.3.Descriptors in the model:

[1]X2v Valence connectivity index chi-2

[2]MATS6m Moran autocorrelation – lag 6 / weighted by atomic masses

[3]MATS8v Moran autocorrelation – lag 8 / weighted by atomic van der Waals volumes

[4]MATS5p Moran autocorrelation – lag 5 / weighted by atomic polarizabilities

[5]BEH2e Highest eigenvalue no. 2 of Burden matrix/weighted by atomic Sanderson electronegativities

[6]BEH1p Highest eigenvalue no. 1 of Burden matrix/weighted by atomic polarizabilities

[7]nArOH Number of aromatic hydroxyls

[8]MlogP Moriguchi octanol-water partition coeff. (log P)

4.4.Descriptor selection:

CART

4.5.Algorithm and descriptor generation:

Statistica

4.6.Software name and version for descriptor generation:

Statistica, version 6.1,StatSoft Italia Srl, Vigonza, Italy, 2003https://statistica.software.informer.com/6.1/

4.7.Chemicals/Descriptors ratio:

82

5.Defining the applicability domain - OECD Principle 3

5.1.Description of the applicability domain of the model:

The 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 and is defined in this way for this QSAR model´s predictions.

If 1 ≥ AD index ≥ 0.85, the predicted substance is in the Applicability Domain of the model. It corresponds to “good reliability of prediction”.

If 0.85 > AD index ≥ 0.7, the predicted substance could be out of the Applicability Domain of the model. It corresponds to “moderate reliability of prediction”.

If AD index < 0.7, 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 in the VEGA website (www.vegahub.eu), including the open access paper describing it [3]. The VEGA 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.
Information on these indices is given below:

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 > 0.85, strongly similar compounds with known experimental value in the training set have been found

If 0.85 ≥ index > 0.7, only moderately similar compounds with known experimental value in the training set have been found

If index ≤ 0.7, no similar compounds with known experimental value in the training set have been found

Accuracy 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 > 0.9, 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 > 0.9, similar molecules found in the training set have experimental values that agree with the predicted value

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

If index ≤ 0.5, similar molecules found in the training set have experimental values that disagree with the 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 < 0.6, 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

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:

Training n= 656

6.6.Pre-processing of data before modelling:

NA

6.7.Statistics for goodness-of-fit:

Accuracy 0.86, Sensitivity 0.87, Specificity 0.85, MCC 0.70. TP 203, TN 356, FP 64, FN 31

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: 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:

Test n= 150

7.6.Experimental design of test set:

Kohonen map

7.7.Predictivity - Statistics obtained by external validation:

Test set: Accuracy = 0.84; Specificity = 0.88; Sensitivity = 0.78, MCC 0.65. TP 42, TN 84, FP 12, FN 12.

Test set in AD: n = 71, Accuracy 0.92, Specificity 0.93, Sensitivity 0.89, MCC 0.82. TP 25, TN 40, FP 3, FN 3.

Test set could be out of AD: n = 34, Accuracy 0.88, Sensitivity 0.90, Specificity 0.88, MCC 0.74. TP 9, TN 21, FP 3, FN 1

Test set out of AD: n = 45, Accuracy 0.69, Sensitivity 0.5, Specificity 0.79, MCC 0.30. TP 8, TN 23, FP 6, FN 23

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 based on descriptors analysis. nArOH Number of aromatic hydroxyls accounts for the ability of the molecules to form hydrogen bonds with aminoacids in the binding pocket of estrogen receptor while MlogP Moriguchi octanol-water partition coeff. is useful to model the affinity with hydrophobic regions of the binding activity

8.3.Other information about the mechanistic interpretation:

NA

9.Miscellaneous information

9.1.Comments:

NA

9.2.Bibliography:

[1] A. Roncaglioni, N. Piclin, M. Pintore, E. Benfenati, "Binary classification models for endocrinedisrupter effects mediated through the estrogen receptor", SAR and QSAR in EnvironmentalResearch (2008), 19, 7-8 https://www.tandfonline.com/doi/abs/10.1080/10629360802550606

[2] METI, Ministry of Economy Trade and Industry, Japan. Current status of testing methods development for endocrine disruptors. 6th meeting of the task force on endocrine disrupters testing and assessment (EDTA), 24–25June 2002, Tokyo, Japan, 2002

[3] Floris M, Manganaro A, Nicolotti O, Medda R, Mangiatordi GF, Benfenati E. A generalizable definition of chemical similarity for read-across. J Cheminform. 2014 Oct 18;6(1):39. doi: 10.1186/s13321-014-0039-1. PMID: 25383097; PMCID: PMC4212147.

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