	QMRF identifier (JRC Inventory): **To be entered by JRC**
	QMRF Title: **Estrogen Receptor-mediated effect (IRFMN/CERAPP) (version 1.0.1)**
	Printing Date: **June 7 2022**

1.QSAR identifier

1.1. QSAR identifier (title):

Estrogen Receptor-mediated effect (IRFMN/CERAPP) (version 1.0.1)

1.2. Other related models:

NA

1.2. 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 IRCCS-Istituto di Ricerche Farmacologiche Mario Negri Via La Masa 19, 20156Milano, Italy emilio.benfenati@marionegri.it https://www.marionegri.it/

[2] Erika Colombo IRCCS-Istituto di Ricerche Farmacologiche Mario Negri Via La Masa 19, 20156Milano, Italy erika.colombo@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 RCCS-Istituto di Ricerche Farmacologiche Mario Negri Via La Masa 19,20156 Milano, Italy alberto.manganaro@marionegri.it

[2] Alessandra Roncaglioni Istituto di ricerche farmacologiche Mario Negri - IRCSS AlessandraRoncaglioni alessandra.roncaglioni@marionegri.it

2.6. Date of model development and/or publication:

23 february 2016

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

[1]CERAPP: Collaborative Estrogen Receptor Activity Prediction Project https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4937869/

[2] Integrated Model of Chemical Perturbations of a Biological Pathway Using 18 In Vitro High-Throughput Screening Assays for the Estrogen Receptor https://www.ncbi.nlm.nih.gov/pubmed/26272952

[3] 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:

Homo sapiens (in vitro)

3.2. Endpoint:

Endocrine disruptor estrogen receptor-mediated effect

3.3. Comment on endpoint:

effect of compound on Estrogen receptor (classification)

3.4. Endpoint units:

Model is a classification so there is no units, possible results should be Active/non-active

3.5. Dependent variable:

The dependent variable is Endocrine effect, as binary classification: Active/Possible Active/NON-Active/Possible NON-active

3.6. Experimental protocol:

3.7. Endpoint data quality and variability:

The model has been built as a set of rules, extracted with Sarpy software from a dataset obtained from a collection of high-quality estrogen receptor (ER) signaling data (1529 chemicals screened across 18 high-throughput screening assays integrated into a single score) from the ToxCast program [2].

4.Defining the algorithm - OECD Principle 2

4.1. Type of model:

Fragments-based model

4.2. Explicit algorithm:

Model match structure by fragment extracted by SARPy software

4.3. Descriptors in the model:

The model is a structure-based model and does not make use of descriptors

4.4. Descriptor selection:

NA

4.5. Algorithm and descriptor generation:

The Sarpy software has been used with a cross-validated procedure, ending with the extraction of two sets of rules (structural alerts) related to ER-mediacted effect activity and inactivity (for a total of59 rules). These rules have been further divided, according to their statistical significance, into a sub-set of rules with strong statistical evidence and another one of rules with weaker evidence. These rules are expressed SMARTS representing molecular fragments. If at least one rule for activity is matching with the given compound, a “Active” or “Possible active” prediction is given, depending on the statistical evidence of the rule. If no active rules are found, but at least one rule for non-activity is matching with the given compound, a “NON-Active” or “Possible NON-active” prediction is given, depending on the statistical evidence of the rule. If no rules are matching at all, no prediction is provided.

4.6. Software name and version for descriptor generation:

SARpy softwareSARpy software (for secondary and tertiary fragment)

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

4.7. Chemicals/Descriptors ratio:

1529 chemicals/59 structural alerts

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 ≥ 0.8, the predicted substance is 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 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 > 0.8, strongly similar compounds with known experimental value in the training set have been found

If 0.8 ≥ 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 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 > 0.8, accuracy of prediction for similar molecules found in the training set is good

If 0.8 ≥ index > 0.6, accuracy of prediction for similar molecules found in the training set is not optimal

If index ≤ 0.6, 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 > 0.8, similar molecules found in the training set have experimental values that agree with the predicted value

If 0.8 ≥ index > 0.6, some similar molecules found in the training set have experimental values that disagree with the predicted value

If index ≤ 0.6, 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 atoms centered fragments of the compound have not been found in the compounds of the training set or are rare fragments

5.3. Software name and version for applicability domain assessment:

VEGA (www.vegahub.eu)

5.4. Limits of applicability:

VEGA provides a quantitative value for the prediction of each substance. This helps the user to identify potential critical aspects, which are indicated. Similar compounds are shown.

6.Internal validation - OECD Principle 4

6.1. Availability of the training set:

Yes

6.2. Available information for the training set:

CAS RN: No

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:

No

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

No

6.5. Other information about the training set:

NA

6.6. Pre-processing of data before modelling:

NA

6.7. Statistics for goodness-of-fit:

After the implementation in VEGA:

n = 1529, not predicted = 241, Accuracy 0.97, Sensitivity 0.85, Specificity 0.97, MCC 0.70. TP 60, TN 1179, FP 38, FN 11.

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:

No

7.2. Available information for the external validation set:

NA

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

NA

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

NA

7.5. Other information about the external validation set:

A further set has been used as an external validation, consisting of data from the Japanese METI database, from which all compounds already found in the present training set have been removed, resulting in a total of 568 compounds

7.6. Experimental design of test set:

NA

7.7. Predictivity - Statistics obtained by external validation:

External set: n 568, not predicted: 71, , Accuracy 0.74, Specificity 0.70, Sensitivity 0.77

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:

With Estrogen Receptor-mediated effect (IRFMN/CERAPP) you can predict the mediated effect on estrogen receptor

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]CERAPP: Collaborative Estrogen Receptor Activity Prediction Project
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4937869/
[2] Integrated Model of Chemical Perturbations of a Biological Pathway Using 18 In Vitro High-Throughput Screening Assays for the Estrogen Receptor https://www.ncbi.nlm.nih.gov/pubmed/26272952
[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 datasets 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