	***QMRF identifier (JRC Inventory):* To be entered by JRC**
	***QMRF Title:* Aromatase activity model (IRFMN) - v 1.0.1**
	Printing Date: October 2022

1.QSAR identifier

1.1.QSAR identifier (title):

Aromatase activity model (IRFMN) - v 1.0.1

1.2.Other related models:

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:

October 2022

2.2.QMRF author(s) and contact details:

[1] Cosimo Toma Istituto di Ricerche Farmacologiche Mario Negri - IRCSS Via Mario Negri 2, 20156 Milano, Italy cosimo.toma@marionegri.it https://www.marionegri.it/

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

2.3.Date of QMRF update(s):

NA

2.4.QMRF update(s):

NA

2.5.Model developer(s) and contact details:

[1] Cosimo Toma Istituto di Ricerche Farmacologiche Mario Negri - IRCSS Via Mario Negri 2, 20156 Milano, Italy cosimo.toma@marionegri.it https://www.marionegri.it/

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

2.6.Date of model development and/or publication:

2018

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

[1] Benfenati E, Roncaglioni A, Lombardo A, Manganaro A. Integrating QSAR, Read-Across, and Screening Tools: The VEGAHUB Platform as an Example. Advances in Computational Toxicology; Springer; 2019. p. 365-81.

[2] Chemistry Development Kit (CDK) http://sourceforge.net/apps/mediawiki/cdk/index.php?title=Main_Page

[3] Malot, C., VSURF: An R Package for Variable Selection Using Random Forests. The R Journal 2015, 7, 19-33.

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:

Human

3.2.Endpoint:

IN VITRO Endocrine disruptor mammalian sceeening - in vitro Tox21_Aromatase_Inhibition (activity test)

3.3.Comment on endpoint:

This assay is based on Aromatase Breast cancer cell line (MCF-7 aro) Cell-based assay, and measures the inhibition of the conversion of testosterone to estradiol catalyzed by aromatase. The control used for this assay is Letrozole (IC50 =9.44 ± 1.4 nM (n =27)).

3.4.Endpoint units:

AC50

3.5.Dependent variable:

For modeling purposes the Endpoint is transformed in molar log10.

3.6.Experimental protocol:

The procedure of curation of data involved the analysis of purity (only samples labeled with purity A are kept) and the outcome of the test (all the assays labeled as inconclusive are discarded). All the samples with additional flags are discarded as well. For each compound several assays could be performed. Only compounds with the same outcome are kept. AC50 of the compounds was the mean of each sample. If the span between AC50 was too large, the compounds are excluded.

SMILES have been retrieved with an in house data curation workflow. This is a semi-automated procedure that addresses automatic chemical data retrieval (i.e., SMILES) from different, orthogonal web based databases, by using two different identifiers, chemical name and CAS registration number. Records were scored based on the coherence of information retrieved from different web sources (DSStox, Cactvs, Pubchem and Chemid).

Data curation procedure performed to top scored records. The procedure includes removal of inorganic and organometallic compounds and mixtures, neutralization of salts, removal of duplicates, checking of tautomeric forms. SMILES have been normalized with VEGA notation. One of the compounds (Fortretamine) is incorrectly recognized as aromatic so it has been manually corrected. From the original dataset here are the compounds maintained/rejected, with the relative reason:

Maintain (with duplicates): 3401

Rejected (mixtures): 3

Rejected (inorganic or unusual elements): 10

Checked manually: 119

Rejected (missing/ambiguous): 67

3.7.Endpoint data quality and variability:

The experimental values are then added to the 3401 compounds. Duplicates are merged in order to keep a unique value for each structure. The experimental value for the duplicates has been maintained only if the Assay outcome was coherent among all members (rejected 6 compounds). The final dataset has 3254 compounds, with 281 active agonists, 170 active antagonists, and 2803 inactive.

4.Defining the algorithm - OECD Principle 2

4.1.Type of model:

The Aromatase activity model (IRFMN) - v 1.0.1 model is a random Forest developed on 3254 molecules retrieved in Tox21 database.

4.2.Explicit algorithm:

Weighted Random Forest

Data sampling for each tree was done with replacement, and the default number of randomly chosen descriptors at each split was the square root, these attributes being different for each tree. A weighted importance has given to the errors related to conazoles.

4.3.Descriptors in the model:

[1]SpMax2_Bh(m)

[2]SpMax2_Bh(p)

[3]SpMaxA_AEA(dm)

[4]piPC08

[5]piPC09

[6]MLOGP

[7]P_VSA_MR_6

[8]X5sol

[9]ATSC4p

[10]Eig03_EA(bo)

[11]SsssN

[12]SM6_B(m)

[13]nRNR2

[14]ATS5s

[15]SssNH

[16]H-049

[17]F02[C-N]

[18]SM15_EA(ed)

4.4.Descriptor selection:

From 3850 2D descriptors has been applied: Variance filter, Correlation filter, VSURF. Feature selection has then applied only on the training set. After discarding all the descriptors that correlated over 0.95 (Pearson) with at least another, a Feature Selection has been performed with R package VSURF [3]. VSURF is a three-steps variable selection based the RF algorithm. The first step is based on evaluating variable importance of descriptors selecting a certain threshold. The second step finds important descriptors highly closely related to the response variable (interpretation step) and the third step (prediction step) evaluates the number of descriptors that are sufficient to have good performance. In order to treat unbalancing of dataset, the parameter strata (vector for stratified sampling) is added to assure that all the classes of the dataset are passed in every tree grown.

4.5.Algorithm and descriptor generation:

Descriptors are generated starting with SMILES with Dragon 7.0

4.6.Software name and version for descriptor generation:

Dragon 7.0

Calculation of several sets of molecular descriptors from molecular geometries (topological, geometrical, WHIM, 3D-MoRSE, molecular profiles, etc.)

Kode srl. Via Nino Pisano, 14 56122 Pisa (PI) - Italy, info@kode-solutions.net www.kode-solutions.net

https://chm.kode-solutions.net/products_dragon.php

4.7.Chemicals/Descriptors ratio:

3254/18

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.

ADI is defined in this way for this QSAR model´s predictions:

If 1 ≥ AD index > 0.80, the predicted substance is regarded in the Applicability Domain of the model. It corresponds to “good reliability” of prediction.

If 0.80 ≥ 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. It corresponds to “low reliability” of prediction.

5.2. Method used to assess the applicability domain:

The Applicability domain 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 [1]. 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.

These indices are defined in this way for this QSAR model:

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 index < 0.6, 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 ≥ 0.8, 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 index < 0.6, molecules found in the training set have experimental values that agree with the target compound predicted value

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

If index ≥ 0.8, 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

5.3.Software name and version for applicability domain assessment:

VEGA (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: Yes

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:

Dataset n = 2602

6.6.Pre-processing of data before modelling:

Descriptors are centered and autoscaled before modeling

6.7.Statistics for goodness-of-fit:

Accuracy 0.94, MCC 0.74, n 2602

Pred/Ref	Active Antagonist	Active agonist	Inactive
Active Antagonist	82	0	0
Active agonist	0	128	1
Inactive	54	95	2241

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

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:

3

6.12.Robustness - Statistics obtained by other methods:

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:

7.6.Experimental design of test set:

7.7.Predictivity - Statistics obtained by external validation:

Test: n 652, accuracy 0.93, MCC 0.68

Pred/Ref	Active Antagonist	Active agonist	Inactive
Active Antagonist	14	0	0
Active agonist	1	32	1
Inactive	19	25	560

Test set in AD: n 559, Accuracy 0.98, MCC 0.90

Pred/Ref	Active Antagonist	Active agonist	Inactive
Active Antagonist	13	0	0
Active agonist	0	31	0
Inactive	5	4	506

Test set could be out of AD: n 63, Accuracy 0.71, MCC 0.19

Pred/Ref	Active Antagonist	Active agonist	Inactive
Active Antagonist	1	0	0
Active agonist	1	0	0
Inactive	5	12	44

Test set out of AD: n 30, accuracy 0.37, MCC 0.008

Prediction	Active Antagonist	Active agonist	Inactive
Active Antagonist	0	0	0
Active agonist	0	1	1
Inactive	9	9	10

7.8.Predictivity - Assessment of the external validation set:

The validation set is well representative of training set.

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: The model is statistical one

8.2.A priori or a posteriori mechanistic interpretation: a posteriori

8.3.Other information about the mechanistic interpretation:

9.Miscellaneous information

9.1.Comments:

9.2.Bibliography:

[1] M. Floris, A. Manganaro, O. Nicolotti, R. Medda, G. Mangiatordi, and E. Benfenati, “A generalizable definition of chemical similarity for read-across,” Journal of Cheminformatics, vol. 6, Oct. 2014, doi: 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