1.QSAR identifier

1.1.QSAR identifier (title):

Adipose tissue:blood model (INERIS) - v. 1.0.1

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:

07-10-2020

2.2.QMRF author(s) and contact details:

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):

No update

2.4.QMRF update(s):

No update

2.5.Model developer(s) and contact details:

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:

2019

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

[1] Cappelli CI, Manganelli, Toma C, Emilio Benfenati E, Mombelli E. Prediction of the Partition Coefficient between Adipose Tissue and Blood for Environmental Chemicals: from Single QSAR Models to an Integrated Approach, Mol. Inf. 2021, 40, 2000072, DOI: 10.1002/minf.202000072

[2] 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.

2.8.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:

Rat.

3.2.Endpoint:

Adipose tissue:blood partition coefficient (Kab).

3.3.Comment on endpoint:

Key-endpoint to predict the bioaccumulation and the pharmacokinetics in humans and animals, since other organ: blood affinities can be estimated as a function of this parameter

3.4.Endpoint units:

Kab

3.5.Dependent variable:

Log (Kab).

3.6.Experimental protocol:

OECD Test No. 417: Toxicokinetics, (2010).

3.7.Endpoint data quality and variability:

101 in vivo data of Kab measured in rats retrieved from one review and several paper in literature ([5]-[10])

The dataset contains mono-constituent organic chemicals belonging to different categories and uses: drugs, plant protection products, polychlorinated biphenyls, volatile organic compounds.

All chemicals' names were converted in SMILES using the CIR and REST node of KNIME v 3.5.1; CAS were retrieved form ChemIDplus and PubChem. Consistence among the CAS numbers, the chemical names and chemical structures of all substances were checked. All structures were standardized and normalized. All duplicates were removed.

The experimental data (in vivo tissue: plasma partition coefficients) were converted in adipose tissue: blood partition coefficient by dividing the partition coefficients determined in plasma by the blood-to-plasma ration. Due to the lack of experimental values of the blood-to-plasma ratio, we considered this value equal to 0.55 for acid drugs and equal to 1 for the remaining chemicals. All the values of KAB were changed to their base-10 logarithms.

Dataset were split into training (63) and test (38) according to three criteria:

- The covered range of the experimental activity
- The chemical structures representativeness (using PCA on Padel descriptors)
- A balanced distribution between training and test set chemicals with respect to their ionisation state

4.Defining the algorithm - OECD Principle 2

4.1.Type of model:

Model is based on Random Forest (RF) approach.

4.2.Explicit algorithm:

Machine Learning Algorithim for Regression

The number of trees selected for the RF were finely tuned by identifying the onset of the plateau of the curve describing the Q2LOO as a function of the number of trees.

4.3.Descriptors in the model:

Six PaDEL-Descriptors v. 2.21 PaDEL compute all the 2D descriptors:

ALogp2

ATSC1s

BCUTp-1l

minHBd

XLogP

WTPT-5

4.4.Descriptor selection:

The initial pool of descriptors (1,407descriptors) was pruned to eliminate non-informative descriptors:

1) with near zero variance;

2) with constant values for the 89% of the compounds.

3) highly correlated to other descriptors (cut-off of the pair correlation equal to 80%.

After this pruning step a total number of 262 descriptors was obtained.

VSURF v 1.0.4 was used as variables selection method. The VSURF algorithm was applied as a function of different number of trees(50, 100,200,300,400and500) and the final RF model adopted the combination of descriptors at prediction step that were associated with the lowest out-of-bag error.

4.5.Algorithm and descriptor generation:

Random Forest (RF)

4.6.Software name and version for descriptor generation:

The "randomForest" v. 4.6.14 in R v. 3.4.3

The "randomForest" package provides an R interface to the Fortran programs by Breiman and Cutlert (http://www.stat.berkeley.edu/users/breiman/).

4.7.Chemicals/Descriptors ratio:

63/6 = 10

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 > 0.85, the predicted substance is regarded 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 regarded out of the Applicability Domain of the model. It corresponds to “low reliability” of prediction.

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:

$$\frac{\sum_c^k |exp_c - pred_c|}{k}$$

where exp_c is the experimental value of the c-*th* molecule in the training set and $pred_c$ is the c-*th* molecule predicted value by the model.

Concordance index (*IdxConcordance*) is calculated as:

$$\frac{\sum_c^k \left|exp_c - pred_{target}\right|}{k}$$

where exp_c is the experimental value of the c-*th* molecule in the training set and $pred_{target}$ is the predicted value for the input target molecule.

Max Error index (*IdxMaxError*) is calculated as:

$$max(|exp_c - pred_c|)$$

where exp_c is the experimental value of the c-*th* molecule in the training set and $pred_{target}$ is the predicted value for the input target molecule, evaluated over the k molecules.

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:

$$ADI = IdxSimilarity \times IdxACF \times IdxDescRange$$

The initialADI index is the used together with the other sub-indices to calculate the final ADI, on the basis of the assessment class in which each sub-index falls:

IdxAccuracy ≥	IdxConcordance ≥	IdxMaxError ≥	InitialADI ≥	ADI
1.5	1.5	1.5	0.85	1.0
0.8	0.8	0.8	0.7	0.85
All other cases				0.7

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 [4]. 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 > 0.85, 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 (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 < 0.8, accuracy of prediction for similar molecules found in the training set is good

If 1.5 > index ≥ 0.8, accuracy of prediction for similar molecules found in the training set is not optimal

If index ≥ 1.5, 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 index < 0.8, molecules found in the training set have experimental values that agree with the target compound predicted value

If 1.5 > index ≥ 0.8, similar molecules found in the training set have experimental values that slightly disagree with the target compound predicted value

If index ≥ 1.5, similar molecules found in the training set have experimental values that completely disagree with the target compound predicted value

Maximum error of prediction among similar molecules:

This index takes into account the maximum error in prediction among the two most similar compounds. Values near 0 means that the predicted compounds fall in an area of the model's space where the model gives reliable predictions without any outlier value. Defined intervals are:

If index < 0.8, the maximum error in prediction of similar molecules found in the training set has a low value, considering the experimental variability

If 1.5 > index ≥ 0.8, the maximum error in prediction of similar molecules found in the training set has a moderate value, considering the experimental variability

If index ≥ 1.5, the maximum error in prediction of similar molecules found in the training set has a high value, considering the experimental variability

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

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: No

Smiles: Yes

Formula: No

INChI: No

MOL file: No

NanoMaterial: No

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

NA

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

NA

6.5.Other information about the training set:

Training set: n = 63

6.6.Pre-processing of data before modelling:

The RF model using the package “randomForest” v. 4.6.14 in R v. 3.4.3. The RF model was calibrated by using unscaled descriptors. The number of trees selected for the RF was finely tuned by identifying the onset of the plateau of the curve describing the Q2LOO as a function of the number of trees.

6.7.Statistics for goodness-of-fit:

We used only the leave-one-out approach

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

Q2 LOO = 0.73

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

Q2 LSO = 0.72 (10-fold leave-some-out)

6.10.Robustness - Statistics obtained by Y-scrambling:

Scrambled Q2 LSO = - 0.17 (10-fold leave-some-out)

6.11.Robustness - Statistics obtained by bootstrap:

NA

6.12.Robustness - Statistics obtained by other methods:

MAE LOO = 0.41 (Mean Absolute Error calculated for leave one out)

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: No

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:

Test set: n = 38

7.6.Experimental design of test set:

The test set was selected so that there were comparable proportions of acidic, basic, zwitterions and neutral chemicals between the two sets training and test sets. Number of neighbours (k) and Sahigara’s thresholds [11] (t) were optimized on the training set by setting a minimum acceptable coverage to 70%.

7.7.Predictivity - Statistics obtained by external validation:

Q2 F3 = 0.87 (External predictive ability according to [12])

Test set in AD: n = 12, RMSE 0.16, R2 0.98, MAE 0.14

Test set "Could be" out AD: n=8, RMSE 0.37, R2 0.84, MAE 0.27

Test set out AD: n=18, RMSE 0.68, R2 0.47, MAE 0.52

7.8.Predictivity - Assessment of the external validation set:

MAE ext = 0.24 (Mean Absolute Error for the test set)

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:

No assumption on the mechanism is done.

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]Prediction of the partition coefficient between adipose tissue and blood for environmental chemicals: from single QSAR models to an integrated approach

[2]Breiman, L. (2001). Random forests. Machine learning, 45(1), 5-32. https://link.springer.com/article/10.1023/A:1010933404324

[3]Genuer, R., Poggi, J. M., &Tuleau-Malot, C. (2018). VSURF: Variable Selection Using Random Forests. R package version 1.0.4. URL https://CRAN.R-project.org/package=VSURF https://journal.r-project.org/archive/2015/RJ-2015-018/RJ-2015-018.pdf

[3] OECD,Test No. 417:Toxicokinetics,OECDGuidelinesfor the Testing of Chemicals, https://www.oecd-ilibrary.org/environ-ment/test-no-417-toxicokinetics_9789264070882-en,2010

[4] 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

[5] P. Paixao,N. Aniceto,L. F. Gouveia,J. A. Morais,Eur. J. Pharm.Sci.2013,50, 526–5

[6] M. Pelekis,K. Krishnan,Regul.Toxicol.Pharmacol.2004,40,264–271

[7] S. Bjorkman,J. Pharm.Pharmacol.2002,54, 1237–1245.

[8] R. Jansson,U. Bredberg,M. Ashton,J. Pharm.Sci.2008,97,2324–2339.

[9]Y. E. Yun,C. A. Cotton,A. N. Edginton,J. Pharmacokinet.Pharmacodyn.2014,41, 1–14.

[10] R. Heredia-Ortiz,M. Bouchard,J. Pharmacokinet.Pharmacodyn.2013,40, 669–68[11] Sahigara, F., Ballabio, D., Todeschini, R. et al. Defining a novel k-nearest neighbours approach to assess the applicability domain of a QSAR model for reliable predictions. J Cheminform 5, 27 (2013). https://doi.org/10.1186/1758-2946-5-27

[12] Consonni, V., Ballabio, D., & Todeschini, R. (2009). Comments on the Definition of the Q2 Parameter for QSAR Validation. Journal of Chemical Information and Modeling, 49(7), 1669–1678. https://doi.org/10.1021/ci900115y

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