	QMRF identifier (JRC Inventory): **To be entered by JRC**
	QMRF Title: **Plasma Protein Binding – LogK (IRFMN) v 1.0.0**
	Printing Date:*02*-**nov-2022**

1.QSAR identifier

1.1.QSAR identifier (title):

Plasma Protein Binding – LogK (IRFMN) v 1.0.0

1.2.Other related models:

Random forest model for predicting the fraction unbound (square root of fraction unbound) to plasma proteins

Random forest model for predicting the fraction unbound (square root of fraction unbound) of acidic drugs to plasma proteins

CORAL model for predicting the fraction unbound to plasma proteins (Square Root Transformation).

1.3.Software coding the model:

KNIME

Data Analysis, Machine Learning and Applications. Studies in Classification, Data Analysis, and Knowledge Organization

Prof. Dr. Michael Berthold, Michael.Berthold@uni-konstanz.de

https://www.knime.com/

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:

November 2022

2.2.QMRF author(s) and contact details:

[1] Domenico Gadaleta; IRCCS - Istituto di Ricerche Farmacologiche Mario Negri
domenico.gadaleta@marionegri.it

[2] Emilio Benfenati; IRCCS - Istituto di Ricerche Farmacologiche Mario Negri
emilio.benfenati@marionegri.it

2.3.Date of QMRF update(s):

2.4.QMRF update(s):

2.5.Model developer(s) and contact details:

[1] Cosimo Toma IRCCS - Istituto di Ricerche Farmacologiche Mario Negri
cosimo.toma@marionegri.it

[2] Domenico Gadaleta; IRCCS - Istituto di Ricerche Farmacologiche Mario Negri
domenico.gadaleta@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] Berthold, Michael R., et al. "KNIME-the Konstanz information miner: version 2.0 and beyond." AcM SIGKDD explorations Newsletter 11.1 (2009): 26-31. https://doi.org/10.1145/1656274.1656280

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

[3] Chemaxon (2017). JChem for Office (Excel). JChem for Office. Collaborative Drug Discovery, I. (2010). ChemCell - Cheminformatics Workflow Automation for Microsoft Excel https://chemaxon.com/

2.8.Availability of information about the model:

All parts of the model freely available except for the calculation of ionization state, that requires a ChemAxon academic license (https://chemaxon.com/)

2.9.Availability of another QMRF for exactly the same model:

3.Defining the endpoint - OECD Principle 1

3.1.Species:

Human

3.2.Endpoint:

QMRF 5. Toxicokinetics QMRF 5. 9. Toxicokinetics.Protein-binding

3.3.Comment on endpoint:

Collection of protein plasma binding in vivo data (logK) from different literature sources.

3.4.Endpoint units:

Adimensional

3.5.Dependent variable:

For modeling purposes the endpoint (fraction unbound, FU) was transformed as below:

Log K = log ((1- FU)/FU)

3.6.Experimental protocol:

Described in Obach et al., 2008, Drug Metab Dispos 36(7): 1385-1405

3.7.Endpoint data quality and variability:

4.Defining the algorithm - OECD Principle 2

4.1.Type of model:

The Plasma Protein Binding – LogK (IRFMN) v 1.0.0 is Random Forest model for predicting the fraction unbound (logK) to plasma proteins based on 489 substances.

4.2.Explicit algorithm:

Random Forest

Each tree was derived from a random sampling with replacement of training set data. The attributes for each tree were randomly selected from the initial pool of descriptors. The number of attributes of each tree was the square root of the initial number of descriptors.

The number of trees is 100.

Ionization state of compounds was defined before descriptors calculation and model derivation. Ionization state was calculated with JChem extension for Microsoft Excel provided by ChemAxon.

4.3.Descriptors in the model:

[1] C.024

[2] MATS5e

[3] C.

[4] T.O..O.

[5] SpMax_AEA.dm.

[6] J_D.Dt
[7] GATS1i
[8] MLOGP
[9] ALOGP
[10] CATS2D_00_LL
[11] CATS2D_00_PP
[12] MLOGP2
[13] N.
[14] AMW
[15] SpMax2_Bh.p.
[16] nCsp2
[17] PCD
[18] F01.C.N.
[19] Eta_betaP
[20] Ui
[21] P_VSA_p_3
[22] nBM
[23] vtotalcharge
[24] P_VSA_i_2

4.4.Descriptor selection:

The initial pool included 3850 2D descriptors calculated with Dragon 7.0. Descriptors were filtered based on 1) Variance, 2) Absolute Pair Correlation and 3) VSURF (R package).

4.5.Algorithm and descriptor generation:

Feature selection was based on training set chemicals. Descriptors were pruned by constant and semi-constant vales (i.e. standard deviation < 0.01), then if a couple of descriptors was characterized by an absolute pair correlation greater than 90%, the descriptor with the highest pair correlation with all the other descriptors was removed. Optimal subsets of descriptors for modeling were obtained with the R package VSURF. The algorithm consists in a three step variable selection based on the logic underpinning the random forest (RF) algorithm (i.e. permutation importance and out-of-bag error). The first step eliminates irrelevant descriptors according to the permutation-based RF score of importance and a user-defined threshold. The second step finds important descriptors closely related to the response variable (interpretation step) and the third step (prediction step) identifies a sufficient parsimonious set of important descriptors leading to a good prediction of the response variables.

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

4.7.Chemicals/Descriptors ratio:

391/24

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.

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

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

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:

6.6.Pre-processing of data before modelling:

SMILES were retrieved using different chemicals identifiers (chemical name, CAS number) checking data from different web sources (i.e., EPA's CompTox database, ChemIDPlus, PubChem) by mean of an automated in-house tool. SMILES were stripped of their counterions and neutralized.

Chemicals were checked for removal of inorganic chemicals and mixtures, and for correction of inaccurate SMILES codes with the help of chemical databases.

Descriptors were centered and autoscaled before modeling.

Fraction unbound data were converted to logK for modeling purposes.

6.7.Statistics for goodness-of-fit:

After the implementation in VEGA: n 391, RMSE 0.50, R2 0.81

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

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

Method: 5-fold cross-validation:

$R^2 = 0.61$

RMSE = 0.72

6.10.Robustness - Statistics obtained by Y-scrambling:

6.11.Robustness - Statistics obtained by bootstrap:

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

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:

Activity sampling method. The entire dataset was sorted based on activity and divided in equal sized bins For each bin, the 80% of chemicals were assigned to the training set while the 20% was assigned to the validation set.

7.7.Predictivity - Statistics obtained by external validation:

r^2= 0.68

RMSE= 0.65

Percentage of chemicals within the applicability domain = 98%

After the implementation in VEGA:

Test set: n 98, RMSE 0.70, R2 0.62

Test set in AD: n 24, RMSE 0.38, R2 0.79

Test set could be out of AD: n 43, RMSE 0.59, R2 0.66

Test set out of AD: n 31, RMSE 0.99, R2 0.47

7.8.Predictivity - Assessment of the external validation set:

The activity sampling method allowed to design a validation set that chemically representative of training set compounds.

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:

Plasma protein binding is heavily influenced by lipophilicity and ionization of compounds.

8.2.A priori or a posteriori mechanistic interpretation:

8.3.Other information about the mechanistic interpretation:

9.Miscellaneous information

9.1.Comments:

9.2.Bibliography:

[1] Obach, R. S., F. Lombardo and N. J. Waters (2008). "Trend analysis of a database of intravenous pharmacokinetic parameters in humans for 670 drug compounds." Drug Metab Dispos 36(7): 1385-1405.

[2] Hastie, T. (2008). Tibshirani, R. and Friedman, J.(2009): The elements of statistical learning. Data mining, inference, and prediction, Springer, New York, ISBN.

[3] Kuhn, M. and K. Johnson (2013). Applied Predictive Modeling, Springer-Verlag New York**:** XIII, 600.

[4] 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

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