1.QSAR identifier

1.1. QSAR identifier (title):

Zebrafish embryo AC50 (IRFMN/CORAL) - v. 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:

Sept 2022

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):

NA

2.4. QMRF update(s):

NA

2.5. Model developer(s) and contact details:

[1] Andrey Toropov Istituto di Ricerche Farmacologiche Mario Negri IRCCS andrey.toropov@marionegri.it

[2] Alla Toropova Istituto di Ricerche Farmacologiche Mario Negri IRCCS alla.toropova@merionegri.it

2.6. Date of model development and/or publication:

October 2017

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

[1] Lavado, G., Gadaleta, D., Toma, C., Golbamaki, A., Toropov, A., Toropova, A., Marzo, M., Baderna, D., Arning, J., & Benfenati, E. (2020). Zebrafish AC$_{50}$ modelling: (Q)SAR models to predict developmental toxicity in zebrafish embryo. Ecotoxicology and environmental safety, 202, 110936

[2] Development of Monte Carlo Approaches in Support of Environmental Research. Toropova A.P.,Toropov A.A., Benfenati E., Rallo R., Leszczynska D., Leszczynski J. (2017). In: Roy K. (eds)Advances in QSAR Modeling. Challenges and Advances in Computational Chemistry and Physics,vol 24. Springer, Cham

[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:

Zebrafish embryo (Danio rerio)

3.2. Endpoint:

Developmental toxicity towards zebrafish Half-maximal activity concentration AC50

3.3. Comment on endpoint:

NA

3.4. Endpoint units:

AC50 in microM

3.5. Dependent variable:

Log AC50

3.6. Experimental protocol:

NA

3.7. Endpoint data quality and variability:

The used data set is part of the Computational Toxicology Research Program of the U.S. EPA (https://www.epa.gov/chemical-research/toxicity-forecasting). The used data set and the published developmental toxicity data for zebrafish was retrieved from Padilla et al. (2012).

4.Defining the algorithm - OECD Principle 2

4.1. Type of model:

The Zebrafish embryo AC50 model was built on 170 substances with Half-maximal activity concentration AC50 values retrieved on EPA. This model is a regression one

4.2. Explicit algorithm:

CORAL (http://www.insilico.eu/coral) was used to develop regression QSAR model from SMILES-based optimal descriptors. A CORAL mathematical model describes the relationship between an endpoint (dependent variable)and relevant SMILES attributes (independent variable), as shown in the equation: Endpoint = C0 +C1 *DCW (T, N) where C0 and C1 are the intercept and slope for the relationship, and DCW (T, N)is the combination of SMILES-based attributes, each associated with a correlation weight (CW).CWs are determined with the Monte Carlo algorithm in an iterative procedure that aims to optimize a target function (TF). The TF is calculated as shown in the equation: TF = R + R' – |R-R'| 0.01 where R and R' are the correlation coefficients between DCW(T, N) and the endpoints for TS and ITS. This procedure is defined as a balance of correlations (BC). The TF is a function of the CWs and is optimized by iteratively modifying them. In the first part of the optimization, CWs are incremented bya value Dstart. This increment is repeated as long as these was a corresponding improvement of the TF. When no further improvement is observed, the Dstart value is modified to Dstart,1 = -0.5 (Dstart)for subsequent iterations. Dstart is iteratively modified each time that an increment of CWs fails to correspond to an increment of TF, until | Dstart | is lower than a threshold value (Dprecession).log(AC 50)= 1.1053641 + 0.0272434 * DCW(1,15)

4.3. Descriptors in the model:

These SMILES-based attributes can be described as in the following Equation: DCW (T*,N*) = CW(Sk) + CW (SSk) + CW (SSSk) where Sk, SSk and SSSk are SMILES attributes defined by asequence of atoms and bonds present in the SMILES string. Sk represents single elements, SSktwo elements combined and SSSk three elements combined. Attributes with a positive CW areconsidered promoters of an increase of the endpoint value, while attributes with a negativecorrelation weights are considered promoters of a decrease

4.4. Descriptor selection:

N is the number of epochs of Monte Carlo for optimization of the target function TF, and T is a threshold used to classify SMILES attributes as rare or not rare. An attribute is defined as rare if it is found in the SMILES of the CS less than T times. Rare SMILES attribute values were set to zero so they were not involved in the modeling. T and N are set to optimize the statistical performance for the CS. 4.Defining the algorithm - OECD Principle 2

For this model, parameters were set as follows: T = 1; N = 35; Dstart= 0.5; Dprecession = 0.1

4.5. Algorithm and descriptor generation:

Simplified Molecular Input Line Entry System (SMILES) notation describes the structure of a chemical using linear strings in place of the classical bi- or tri- dimensional representation. CORAL breaks the SMILES strings of the TS compounds into small components (SMILES-based attributes). Each SMILES-based attributes check the presence of particular characters (or combinations of characters) within the SMILES

4.6. Software name and version for descriptor generation:

CORAL-2017 http://www.insilico.eu/coral

4.7. Chemicals/Descriptors ratio:

170 chemicals / 361 SMILES attributes (NB, these are not descriptors)

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 and 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.

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 are in the VEGA website (www.vegahub.eu), including the open access paper describing it [1]. The 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 < 0.6, accuracy of prediction for similar molecules found in the training set is good

If 1.2 > index ≥ 0.6, accuracy of prediction for similar molecules found in the training set is not optimal

If index ≥ 1.2, 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 1.2 > index ≥ 0.6, similar molecules found in the training set have experimental values that slightly disagree with the target compound predicted value

If index ≥ 1.2, 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 < 0.6, the maximum error in prediction of similar molecules found in the training set has a low value, considering the experimental variability

If 1.2 > index ≥ 0.6, the maximum error in prediction of similar molecules found in the training set has a moderate value, considering the experimental variability

If index ≥ 1.2, 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:

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: 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:

All

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

All

6.5. Other information about the training set:

NA

6.6. Pre-processing of data before modelling:

The used data set is part of the Computational Toxicology Research Program of the U.S. EPA(https://www.epa.gov/chemical-research/toxicity-forecasting). The useddata set and the published developmental toxicity data for zebrafish was retrieved from Padilla et al. (2012). The simplified molecular input line entry system (SMILES) in the data set have been cleaned and standardized with software istMolBase(https://chm.kode-solutions.net/products_istmolbase.php). After a meticulous check of each chemical structure, 10 compounds were removed because they were mixtures or had disconnected molecular structures. Tobuild the model 16 compounds with low frequency distribution of AC50have been removed because considered noise. The fnal data set contains170 compounds. The data set was split into four groups: training TS (37%), invisible training ITS (39%), calibration CS (12%) and validation VS (12%) sets. In CORAL software, TS, ITS, CS and VS are identified as +, -, # and *,respectively. In the implementation, the validation set is the test set in VEGA andthe other sets form the whole training set

6.7. Statistics for goodness-of-fit:

Training set (TS): R2= 0.7437 RMSE = 0.361

Invisible Training set (ITS): R2= 0.7436 RMSE = 0.395

Calibration set (CS): R2= 0.6577 RMSE = 0.385

After the implementation in VEGA:

Training set: n 150, R2 0.70, RMSE 0.38

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

Training set (TS): Q2= 0.728

Invisible Training set (ITS): Q2= 0.726

Calibration set (CS): Q2= 0.592

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: Yes

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:

NA

7.6. Experimental design of test set:

The initial dataset was randomly divided into a training set (TS) of 64compounds, an invisible training set (ITS) of 66 compounds, a calibration set (CS) of 20 compounds, and a validation set (VS) of 20compounds. The TS was used for regression model’s derivation. The ITS was used as “inspector” during model derivation, to confirm (or reject) predictivity of the model for substances which were not involved directly to the optimization process. The CS detected the beginning of overfitting by verifying the increase of the correlation between descriptors and endpoint during the optimization process, until improvements were no longer observed. The VS is the final estimator of the predictive potential of the model.

7.7. Predictivity - Statistics obtained by external validation:

Validation set (VS): R2= 0.697 RMSE=0.386

After the implementation in VEGA:

Test set: n 20, R2 0.69, RMSE 0.38

Test set in AD: n 5, R2 0.16, RMSE 0.46

Test set could be out AD: n 8, R2 0.62, RMSE 0.43

Test set out AD: n 7, R2 0.84, RMSE 0.20

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

8.3. Other information about the mechanistic interpretation:

NA

9.Miscellaneous information

9.1. Comments:

NA

9.2. Bibliography:

Padilla, S., Corum, D., Padnos, B., Hunter, D. L., Beam, A., Houck, K. A., Sipes, N., Kleinstreuer, N.,Knudsen, T., Dix, D. J., & Reif, D. M. (2012). Zebrafish developmental screening of the ToxCastTMPhase I chemical library. Reproductive Toxicology, 33, 174-187. http://www.sciencedirect.com/science/article/pii/S0890623811004011

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