| | |
|---|---|
| | ***QMRF identifier (JRC Inventory):*** **To be entered by JRC** |
| | ***QMRF Title:*** **VEGA Ready Biodegradation model** |
| | ***Printing Date: May 30, 2022*** |
| | |

## 1.QSAR identifier

### 1.1.QSAR identifier (title):

VEGA Ready Biodegradation model (version 1.0.10)

### 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:

30 May 2022

### 2.2.QMRF author(s) and contact details:

Emilio Benfenati IRCCS - Istituto di Ricerche Farmacologiche Mario Negriemilio.benfenati@marionegri.i

### 2.3.Date of QMRF update(s):

NA

### 2.4.QMRF update(s):

NA

### 2.5.Model developer(s) and contact details:

[1]Anna Lombardo IRCCS - Istituto di Ricerche Farmacologiche Mario Negri anna.lombardo@marionegri.it

[2]Fabiola Pizzo IRCCS - Istituto di Ricerche Farmacologiche Mario Negri fabiola.pizzo@marionegri.it

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

[4]Alberto Manganaro IRCCS - Istituto di Ricerche Farmacologiche Mario Negri alberto.manganaro@marionegri.it

[5]Thomas Ferrari Politecnico di Milano, Dipartimento di Elettronica, Informazione e Bioingegneria thedataconspiracy@gmail.com

[6]Giuseppina Gini Politecnico di Milano, Dipartimento di Elettronica, Informazione e Bioingegneria gini@elet.polimi.it

### 2.6.Date of model development and/or publication:

The model was published in 2014 (see 2.7).

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

[1] Lombardo A, Pizzo F, Benfenati E, Manganaro A, Ferrari T, Gini G (2014). A new in silico classification model for ready biodegradability, based on molecular fragments. Chemosphere. 108,10–16 http://dx.doi.org/10.1016/j.chemosphere.2014.02.073

[2] 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:

No

## 3.Defining the endpoint - OECD Principle 1

### 3.1.Species:

Activated sludge (according to OECD 301C Modified MITI (I) Test).

### 3.2.Endpoint:

QMRF 2. 3.a. Persistence: Biodegradation. Ready/not ready biodegradability OECD 301C Modified MITI (I) Test

### 3.3.Comment on endpoint:

Ready biodegradability tests are screening tests in with a high concentration of the test substance is used and ultimate bio degradation is measured by non-specific parameters under aerobic conditions. In the OECD 301C Modified MITI (I) test, a substance can be considered ready biodegradable if 60% of the substance is mineralized in 28 days (in terms of ThOD)

### 3.4.Endpoint units:

Adimensional

### 3.5.Dependent variable:

Ready Biodegradable/Not Ready Biodegradable.

For modelling purpose four ready biodegradability classes were used: ready biodegradable, possible ready biodegradable, not ready biodegradable and possible not ready biodegradable. To perform prediction, possible ready biodegradable were considered as ready biodegradable and possible not ready biodegradable were considered as not ready biodegradable.

### 3.6.Experimental protocol:

OECD 301C Modified MITI (I) Test. It measures the oxygen uptake in a period of 28 days of a solution of the substance inoculated with activated sludge under aerobic conditions c.f. OECD TG 301 (5)

### 3.7.Endpoint data quality and variability:

For the development and testing of the model the datasets was extracted from the OECD QSAR toolbox v 2.0 and from the BIOWIN 5 and 6 models (in EPISuiteTM) and then combined (see ref. 1, section 9.2). The dataset is described in ref 2 (section 9.2). To validate the model data were extracted from Cheng et al., 2012 (see ref 3, section 9.2). 17 compounds were found with non-concordant data among the OECD QSAR toolbox v 2.0 and the BIOWIN 5 and 6 models (i.e. not classified in the same class: Ready Biodegradable or Not-Ready Biodegradable) and were excluded. All the common data with Cheng et al., 2012 were eliminated and not checked to verify non-concordant data because the data from Cheng et al 2012 were used only as external validation set.

## 4.Defining the algorithm - OECD Principle 2

### 4.1.Type of model:

SAR built using both statistically-related fragments extracted using SARpy (see ref 4, section 9.2) and expert-based fragments

### 4.2.Explicit algorithm:

see ref 1, section 9.2

The classification scheme is based on a flow chart (see ref. 1, section 9.2). Basically if at least one fragment linked with non-ready biodegradability and with high specificity is found, the chemical is classified as non-ready biodegradable, otherwise, if no highly specific fragments are found but at least one fragment linked with non-ready biodegradability and lower specificity is found, the compound is classified as possible not ready biodegradable. If no fragments linked with not ready biodegradability are found but fragments linked

with ready biodegradability are found, the compound is classified as Ready Biodegradable (if highly specific fragments are found) or possible Ready Biodegradable (if only lower specific fragments are found). If no fragments are found, the compound is classified as unknown

### 4.3. Descriptors in the model:

[1]1,2-dichlorobenzene c1ccc(c(c1)Cl)Cl Fragment in ruleset 1: highly specific fragments linked with non-ready biodegradability

[2]1-ethyl-3-methylbenzene c1cc(cc(c1)CC)C Fragment in ruleset 1: highly specific fragments linked with non-ready biodegradability

[3]1,2-dichloroethane C(C(Cl))Cl Fragment in ruleset 1: highly specific fragments linked with non-ready biodegradability

[4]2-chloroaniline Nc1ccc(cc1Cl) Fragment in ruleset 1: highly specific fragments linked with non-ready biodegradability

[5]diphenylmethanone c1ccccc1C(=O)c2ccccc2 Fragment in ruleset 1: highly specific fragments linked with non-ready biodegradability

[6]dimethoxyphosphinic acid O=P(OC)(OC)O Fragment in ruleset 1: highly specific fragments linked with non-ready biodegradability

[7]cyclohex-4-ene-1,2-dicarbaldehyde O=CC1CC=CCC1C(=O) Fragment in ruleset 1: highly specific fragments linked with non-ready biodegradability

[8]benzene-1,3-diamine Nc1ccc(c(N)c1) Fragment in ruleset 1: highly specific fragments linked with non-ready biodegradability

[9]1-bromopropane CCCBr Fragment in ruleset 1: highly specific fragments linked with non-ready biodegradability

[10]3-chlorophenol Oc1ccc(c(c1)Cl) Fragment in ruleset 1: highly specific fragments linked with non-ready biodegradability

[11]fluorine F Fragment in ruleset 1: highly specific fragments linked with non-ready biodegradability

[12](1-phenylethyl)benzene c1ccc(cc1)C(c2ccccc2)C Fragment in ruleset 1: highly specific fragments linked with non-ready biodegradability

[13]methoxy(sulfanylidene)phosphinous acid P(=S)(OC)O Fragment in ruleset 1: highly specific fragments linked with non-ready biodegradability

[14]N-phenylaniline c1ccc(cc1)Nc2ccc(cc2) Fragment in ruleset 1: highly specific fragments linked with non-ready biodegradability

[15]naphthalen-1-amine c1ccc2c(c1)cccc2N Fragment in ruleset 1: highly specific fragments linked with non-ready biodegradability

[16]1-methylnaphthalene Cc1cccc2ccccc12 Fragment in ruleset 1: highly specific fragments linked with non-ready biodegradability

[17]benzyldimethylamine C(c1cccc(c1))N(C)C Fragment in ruleset 1: highly specific fragments linked with non-ready biodegradability

[18]2-methylnonane CCCCCCCC(C)C Fragment in ruleset 1: highly specific fragments linked with non-ready biodegradability

[19]1,3-benzothiazole c1nc2ccccc2s1 Fragment in ruleset 1: highly specific fragments linked with non-ready biodegradability

[20]tin [Sn] Fragment in ruleset 1: highly specific fragments linked with non-ready biodegradability

[21]methanimine C=N Fragment in ruleset 1: highly specific fragments linked with non-ready biodegradability

[22]1,4-diethylbenzene c1cc(ccc1C(C))C(C) Fragment in ruleset 1: highly specific fragments linked with non-ready biodegradability

[23]propoxybenzene CCCOc1ccccc1 Fragment in ruleset 1: highly specific fragments linked with non-ready biodegradability

[24]2-ethyl-1-methoxyhexane O(C)CC(CC)CCCC Fragment in ruleset 1: highly specific fragments linked with non-ready biodegradability

[25]2-chlorobenzaldehyde O=Cc1c(cccc1Cl) Fragment in ruleset 1: highly specific fragments linked with non-ready biodegradability

[26]1-ethyl-2-methylbenzene c1ccc(c(c1)CC)C Fragment in ruleset 1: highly specific fragments linked with non-ready biodegradability

[27]3-chloroaniline Nc1ccc(c(c1)Cl) Fragment in ruleset 1: highly specific fragments linked with non-ready biodegradability

[28](2-methoxyethyl)(propyl)amine CCCNCCOC Fragment in ruleset 1: highly specific fragments linked with non-ready biodegradability

[29]2,4-dimethylpent-1-ene C=C(C)CC(C)(C) Fragment in ruleset 1: highly specific fragments linked with non-ready biodegradability

[30]bromobenzene c1ccc(cc1)Br Fragment in ruleset 1: highly specific fragments linked with non-ready biodegradability

[31]methylcarbamic acid O=C(O)NC Fragment in ruleset 1: highly specific fragments linked with non-ready biodegradability

[32]1,1,1-trichloroethane CC(Cl)(Cl)Cl Fragment in ruleset 1: highly specific fragments linked with non-ready biodegradability

[33]chloroethene C(=CCl) Fragment in ruleset 1: highly specific fragments linked with non-readybiodegradability

[34]dithioperoxol SS Fragment in ruleset 1: highly specific fragments linked with non-ready biodegradability

[35]benzylbenzene C(c1ccccc1)c2ccccc2 Fragment in ruleset 2: lower specific fragments linked with non-ready biodegradability

[36]1-chloro-2-methylbenzene c1ccc(c(c1)C)Cl Fragment in ruleset 2: lower specific fragments linked with non-ready biodegradability

[37]naphthalene c1cccc2ccccc12 Fragment in ruleset 2: lower specific fragments linked with non-ready biodegradability

[38]bromine Br Fragment in ruleset 2: lower specific fragments linked with non-ready biodegradability

[39]3-methylaniline Nc1ccc(c(c1)C) Fragment in ruleset 2: lower specific fragments linked with non-ready biodegradability

[40]hydroxylamine ON Fragment in ruleset 2: lower specific fragments linked with non-ready biodegradability

[41]chlorobenzene c1ccc(c(c1))Cl Fragment in ruleset 2: lower specific fragments linked with non-ready biodegradability

[42]benzenesulfonic acid c1ccc(cc1)S(=O)(=O)O Fragment in ruleset 2: lower specific fragments linked with non-ready biodegradability

[43]pentan-2-amine C(N)(C)CC(C) Fragment in ruleset 2: lower specific fragments linked with non-ready biodegradability

[44]2-hydroxyethyl acetate O=C(OCCO)C Fragment in ruleset 3: highly specific fragments linked with ready biodegradability

[45]propanedial O=CCC(=O) Fragment in ruleset 3: highly specific fragments linked with ready biodegradability

[46]N-(2-hydroxyethyl)formamide O=C(N(CCO)) Fragment in ruleset 3: highly specific fragments linked with ready biodegradability

[47]1-propoxynonane O(CCCCCCCCC)CCC Fragment in ruleset 3: highly specific fragments linked with ready biodegradability

[48]2,6-dimethylhepta-1,5-diene CC(=C)CCC=C(C)C Fragment in ruleset 3: highly specific fragments linked with ready biodegradability

[49]2-methoxybutane CCC(OC)C Fragment in ruleset 3: highly specific fragments linked with ready biodegradability

[50](dodecyloxy)phosphonous acid P(O)(O)OCCCCCCCCCCCC Fragment in ruleset 3: highly specific fragments linked with ready biodegradability

[51]benzyl formate O=C(OCc1ccccc1) Fragment in ruleset 3: highly specific fragments linked with ready biodegradability

[52]6-oxohexanoic acid O=C(O)CCCCC(=O) Fragment in ruleset 3: highly specific fragments linked with ready biodegradability

[53]docosane CCCCCCCCCCCCCCCCCCCCCC Fragment in ruleset 3: highly specific fragments linked with ready biodegradability

[54]ridecan-1-ol OCCCCCCCCCCCCC Fragment in ruleset 4: lower specific fragments linked with ready biodegradability

[55]butan-2-one O=C(C)CC Fragment in ruleset 4: lower specific fragments linked with readybiodegradability

[56]3-methoxyprop-1-ene C(OC)C=C Fragment in ruleset 4: lower specific fragments linked with ready biodegradability

[57]2-methylhept-2-ene C(C)CCC=C(C)C Fragment in ruleset 4: lower specific fragments linked with ready biodegradability

[58]methyl propanoate O=C(OC)CC Fragment in ruleset 4: lower specific fragments linked with ready biodegradability

[59]butyl formate C(=O)OCCCC Fragment in ruleset 4: lower specific fragments linked with ready biodegradability

[60]1-ethoxybutane CCOCCCC Fragment in ruleset 4: lower specific fragments linked with ready biodegradability

[61]octan-1-ol OCCCCCCCC Fragment in ruleset 4: lower specific fragments linked with ready biodegradability

[62]tridecane CCCCCCCCCCCCC Fragment in ruleset 4: lower specific fragments linked with ready biodegradability

[63]propanoic acid CCC(=O)O Fragment in ruleset 4: lower specific fragments linked with ready biodegradability

[64]benzoic acid O=C(O)c1ccc(cc1) Fragment in ruleset 4: lower specific fragments linked with ready biodegradability

[65]butan-1-ol OCCCC Fragment in ruleset 4: lower specific fragments linked with ready biodegradability

[66]acetamide O=C(N)C Fragment in ruleset 4: lower specific fragments linked with ready biodegradability

[67]acetaldehyde O=CC Fragment in ruleset 4: lower specific fragments linked with ready biodegradability

[68]ethane-1,2-diol OCCO Fragment in ruleset 4: lower specific fragments linked with ready biodegradability

[69]propan-1-amine NCCC Fragment in ruleset 5: lower specific fragments linked with ready biodegradability extracted from unknown

[70]sulfanone S(=O) Fragment in ruleset 5: lower specific fragments linked with ready biodegradability extracted from unknown

[71]heptanes CCCCCCC Fragment in ruleset 5: lower specific fragments linked with ready biodegradability extracted from unknown

[72]anisole O(c1ccccc1)C Fragment in ruleset 5: lower specific fragments linked with ready biodegradability extracted from unknown

[73]butane CCCC Fragment in ruleset 5: lower specific fragments linked with ready biodegradability extracted from unknown

[74](chloromethyl)benzene C1(ccccc1)C(Cl) Fragment in ruleset 5: lower specific fragments linked with ready biodegradability extracted from unknown

[75]diazene N=N Fragment in ruleset 6: expert-based fragments linked with non-ready biodegradability

[76]halogenated ring structure [R][Cl,F,Br,I] Fragment in ruleset 6: expert-based fragments linked with non-ready biodegradability

[77]carbonyl bound to aromatic structure [a][C;D2]=O Fragment in ruleset 7: expert-based fragments linked with ready biodegradability

[78]formonitrile C#N Fragment in ruleset 7: expert-based fragments linked with ready biodegradability

### 4.4. Descriptor selection:

Full details are explained in the literature (see ref. 1 in section 9.2). Briefly, SARpy was run on the training set to extract rulesets1-4. All the rules were checked removing all the fragments that had a likelihood ration < 2 and/or a true positive rate (TP) < 70%. The rules with TP = 100% were considered highly specific, the other balanced. SARpy was run again on the compounds of the training set that could not be predicted using rule sets 1-4. 2 rulesets were extracted (one linked with ready and one with not ready biodegradable compounds). After the check (performed with the same rules explained above) only one ruleset remained, ruleset 5. The compounds of the training set were checked and 2 new rulesets were extracted (one linked with ready and one with not-ready biodegradable compounds): rulesets 6 and 7. These last fragments were all confirmed with studies from the literature

### 4.5. Algorithm and descriptor generation:

Fragments both statistically or expert based. Details provided in section 4.3

### 4.6. Software name and version for descriptor generation:

SARpy is a general software that automatically extracts knowledge from a dataset and detects the molecular structural fragments associated with the activity of interest

### 4.7. Chemicals/Descriptors ratio:

582 chemicals (training set)/78 fragments = 7.5

## 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 ≥ 0.8, the predicted substance is regarded in the Applicability Domain of the model. It corresponds to “good reliability” of prediction.

If 0.8 >AD index ≥ 0.65, the predicted substance could be out of the Applicability Domain of the model. It corresponds to “moderate reliability” of prediction.

If AD index < 0.65, the predicted substance is regarded out of the Applicability Domain of the model. It corresponds to “low reliability” of prediction.

No ADI threshold was used to provide performance calculations with the exception of that for test set 2 (external validation set, c.f. point 7.7. below)

### 5.2. Method used to assess the applicability domain:

The AD 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 [6]. 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 1 ≥ index > 0.8 accuracy of prediction for similar molecules found in the training set is good

If 0.8 ≥ index > 0.5, accuracy of prediction for similar molecules found in the training set is not optimal

If index ≤ 0.5, 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 1 ≥ index > 0.8, molecules found in the training set have experimental values that agree with the target compound predicted value

If 0.8 ≥ index > 0.5, similar molecules found in the training set have experimental values that slightly disagree with the target compound predicted value

If index ≤ 0.5, 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 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

Included in the VEGA software and automatically displayed when running the model

emilio.benfenati@marionegri.it

https://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: null

### 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:

For some chemicals of the training, test and validation sets are available continuous values (Biological Oxygen Demand, BOD%) and binary values (Ready or not ready biodegradable), for other only the binary classification. The model is based on the binary values

### 6.6.Pre-processing of data before modelling:

Continuous values are converted in binary values: if BOD (28 d) is greater or equal to the 60% then the compound was classified as ready biodegradable, otherwise not-ready biodegradable. If BOD is available for period lower than 28d, only the values with at least the 60% of BOD were considered and the compound were classified as ready biodegradable. When multiple values were available, the compounds with values not in agreement were deleted; otherwise, the mean value was used to classify the compound. All the chemical structures were manually checked deleting doubtful compounds, mixture, inorganic compounds and tautomers.

The training set is composed of 582 mono constituent compounds (279 ready biodegradable, 303 not ready biodegradable)

### 6.7.Statistics for goodness-of-fit:

The possible ready biodegradable compounds were considered as active, and possible not ready biodegradable compounds were considered as not-active.

Accuracy = 92.2% Sensitivity = 94.8% Specificity = 89.8% MCC = 0.85 FP rate = 0.10 FN rate = 0.05

TP: 221, TN: 227, FP: 26, FN: 12

Not assigned: 96

There are compounds not assigned because of SARpy modelling: for these compounds no fragments are found

**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:**

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: null

**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:**

To test the model two datasets were used: a test set 1 of 146 mono constituent organic compounds (69 ready biodegradable and 77 not ready biodegradable), obtained from the same source of the training set (1 and 2, section 9.2), and a test set 2, i.e. an external set of 874 new compounds obtained from 1 and 3, section 9.2. This second, large set was available after the development of the model. All the data were pre-processed as explained in section 6.6.

**7.6.Experimental design of test set:**

The dataset was randomly split into training and test set with respectively the 80% and the 20% (test set 1) of the compounds

**7.7.Predictivity - Statistics obtained by external validation:**

To calculate the statistics, “possible ready biodegradable” compounds were considered as active, and “possible not ready biodegradable” compounds were considered as not-active Test set 1: Accuracy = 82% Sensitivity = 87 % Specificity = 77% MCC = 0.64

FP rate = 0.23 FN rate = 0.13

TP: 48, TN: 50, FP:15, FN:7, Not assigned: 26

Considering ADI thresholds:

Test set in AD: n 71, Sensitivity 100%, Specificity 87%, Accuracy 94%, MCC 0.89,

TP 41, TN 26, FP 4, FN 0

Test set Could be out of AD: n 26, Sensitivity 60%, Specificity 75%, Accuracy 69%, MCC 0.35,

TP 6, TN 12, FP 4, FN 4

Test set out of AD: n 23, Sensitivity 25%, Specificity 63%, Accuracy 57%, MCC -0.09,

TP 1, TN 12, FP 7, FN 3

Test set 2, i.e. External validation set: Accuracy = 76% Sensitivity = 73% Specificity = 84% MCC = 0.51

FP rate = 0.27 FN rate = 0.16, TP 173, TN 385, FP 142, FN 34.

Test set 2, External validation set in AD (as defined by the VEGA ADI approach with an in AD categorization when ADI $\geq$ 0.8): Accuracy = 81%Sensitivity = 76% Specificity = 91% MCC = 0.63

FP rate = 0.24 FN rate = 0.09, TP 147, TN 249, FP 80, FN 15 (for more details see ref 1 in section 9.2)

### 7.8.Predictivity - Assessment of the external validation set:

MCC values always > 50% and the other statistics prove that the model is highly predictive. The use of the applicability domain i.e. the VEGA ADI concept improves the results. Test set: 119 compounds (82%) have an Atom Centered Fragment evaluation of 1 (= all atom centered fragment of the compound have been found in the compounds of the training set). Test set 2, External set: 617 compounds (70.) have an Atom Centered Fragment evaluation of 1 (= all atom centered fragment of the compound have been found in the compounds of the training set)

### 7.9.Comments on the external validation of the model:

The test set is balanced (52.7% not ready biodegradable and 47.3%ready biodegradable). The test set 2, external set is not balanced (72.5% not ready biodegradable and 27.5% ready biodegradable). In both conditions the model seems highly predictive

## 8.Providing a mechanistic interpretation - OECD Principle 5

### 8.1.Mechanistic basis of the model:

The model is based on structural alerts. The majority of them are statistically based, and some have mechanistic explanation:

- Azo group. The initial step of the biodegradation of azo dyes is cleavage of the azo group. This reaction is catalysed by the enzyme azoreductase, which is inhibited by molecular oxygen and the Ready Biodegradability tests are conducted under aerobic condition. For this reason, this structural alert is considered associated to chemicals not ready biodegradable.

- All compounds containing halogen atoms. The presence of halogenated organic compounds in the environment is recent, so the enzymes that have evolved to metabolize these compounds are considered to be in a relatively early stage of development. The initial conversion of non-toxic compounds yields toxic products (e.g. the monooxygenase-catalysed oxidations of xenobiotics performed by various microorganisms). For this reason, this structural alert is considered associated to chemicals not ready biodegradable.

- Aromatic aldehydes (defined as a carbonyl group linked to any aromatic ring). Several reactions convert aromatic compounds into intermediates, which are subject to ring-cleavage and subsequent funnelling into the Krebs cycle. For this reason, this structural alert is considered associated to chemicals ready biodegradable.

- Nitrile group. They can be degraded by several strains of bacteria, fungi and plants under aerobic conditions. For this reason, this structural alert is considered associated to chemicals ready biodegradable.

For details and full reference, see ref.1 in section 9.2.

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

A posteriori mechanistic interpretation

### 8.3.Other information about the mechanistic interpretation:

For details and full reference, see ref. 1 in section 9.2

## 9.Miscellaneous information

### 9.1.Comments:

NA

### 9.2.Bibliography:

[1] Lombardo A, Pizzo F, Benfenati E, Manganaro A, Ferrari T, Gini G (2014). A new in silicoclassification model for ready biodegradability, based on molecular fragments. Chemosphere. 108,10-16 http://dx.doi.org/10.1016/j.chemosphere.2014.02.073

[2]Toropov AA, Toropova AP, Lombardo A, Roncaglioni A, De Brita N, Stella G, Benfenati E (2012).CORAL: the prediction of biodegradation of organic compounds with optimal SMILES-based descriptors. Central European Journal of Chemistry. 10(4), 1042-1048 DOI: 10.2478/s11532-012-0031-4

[3]Cheng F, Ikenaga Y, Zhou Y, Yu Y, Li W, Shen J, Du Z, Chen L, Xu C, Liu G, Lee PW, Tang Y(2012). In silico assessment of chemical biodegradability. Journal of Chemical Information Modeling.52(3), 655–669 doi: 10.1021/ci200622d

[4]Ferrari T, Gini G, Bakhtyari NG , Benfenati E (2011). Mining toxicity structural alerts from SMILES :a new way to derive structure-activity relationships. In: Proc. IEEE SSCI 2011: Symposium Serieson Computational Intelligence – CIDM 2011, 120–127. 10.1109/CIDM.2011.5949444

[5] OECD. Test Guideline No. 301: Ready Biodegradability. Paris: Organisation for Economic Co-operation and Development, 1992. https://www.oecd-ilibrary.org/environment/test-no-301-ready-biodegradability_9789264070349-en.

[6] 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 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