Guide to Ready Biodegradability Model version 1.0.9

Table of Contents

1. Model explanation

1.1 Introduction

The model is based on the OECD TG 301C - modified MITI -I test data and provides a qualitative evaluation (binary classification) of ready biodegradability properties. It has been developed using Sarpy software, by Istituto di Ricerche Farmacologiche Mario Negri and Politecnico di Milano.

1.2 Model details

The model has been built as a set of rules, extracted from the training set with Sarpy software. The final set of fragments obtained come from a work that involved both a statistical part and an expert-based part. Seven different rules-sets of fragments related to ready biodegradability activity were obtained: rules-set 1 (non-readily biodegradable fragments with high specificity), rules-set 2 (non-readily biodegradable fragments with balanced performance), rules-set 3 (readily biodegradable fragments with high specificity), rules-set 4 (readily biodegradable fragments with balanced performance), rules-set 5 (readily biodegradable fragments extracted from unknown), rules-set 6 (non-readily biodegradable expert-based fragments), rules-set 7(ready biodegradable expert-based fragments).

The overall model is conservative, and in case of the presence of conflicting fragments the prediction is for non readily biodegradability. The logical scheme of the model comes directly from a chemical reasoning: a substance is always considered non biodegradable if at least one fragment related to non biodegradability is found, even if easily biodegradable fragments are found; this means that a part of the compound is anyway persistent.

The above mentioned rules have been grouped in four main sets: rules for non biodegradable compounds with high specificity and with balanced performances, rules for biodegradable compounds with high specificity and with balanced performances. On the basis of these sets, a compound is predicted as: “NON readily biodegradable” if at least one rule for non biodegradable compounds with high specificity has been found; “possible NON readily biodegradable” if at least one rule for non biodegradable compounds with balanced performances has been found; “readily biodegradable” if at least one rule for biodegradable compounds with high specificity has been found; “possible readily biodegradable” if at least one rule for biodegradable compounds with balanced performances has been found. If no rule has been found at all, the model is not able to provide a prediction.

1.3 Applicability Domain

The applicability domain of predictions is assessed using an Applicability Domain Index (ADI) that has values from 0 (worst case) to 1 (best case). The ADI is calculated by grouping several other indices, each one taking into account a particular issue of the applicability domain. Most of the indices are based on the calculation of the most similar compounds found in the training and test set of the model,

calculated by a similarity index that consider molecule's fingerprint and structural aspects (count of atoms, rings and relevant fragments).

For each index, including the final ADI, three intervals for its values are defined, such that the first interval corresponds to a positive evaluation, the second one corresponds to a suspicious evaluation and the last one corresponds to a negative evaluation.

Following, all applicability domain components are reported along with their explanation and the intervals used. Note that for purpose of evaluating accuracy and concordance indices, prediction of "possible ready biodegradable" and "possible non ready biodegradable" are considered as "ready biodegradable" and "non ready biodegradable".

- **Similar molecules with known experimental value**. This index takes into account how similar are the first three 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:

1 >= index > 0.85	strongly similar compounds with known experimental value in the training set have been found
0.85 >= index > 0.7	only moderately similar compounds with known experimental value in the training set have been found
index <= 0.7	no similar compounds with known experimental value in the training set have been found

- **Accuracy of prediction for similar molecules**. This index takes into account the error in prediction for the three most similar compounds found. Values near 0 mean that the predicted compounds falls in an area of the model's space where the model gives reliable predictions, otherwise the greater is the value, the worse the model behaves. Defined intervals are:

1 >= index > 0.8	accuracy of prediction for similar molecules found in the training set is good
0.8 >= index > 0.5	accuracy of prediction for similar molecules found in the training set is not optimal
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 three 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:

1 >= index > 0.8	similar molecules found in the training set have experimental values that agree with the predicted value
0.8 >= index > 0.5	some similar molecules found in the training set have experimental values that disagree with the predicted value
index <= 0.5	similar molecules found in the training set have experimental values that disagree with the 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:

index = 1	all atom centered fragment of the compound have been found in the compounds of the training set
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
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

- **Global AD Index**. The final global index takes into account all the previous indices, in order to give a general global assessment on the applicability domain for the predicted compound. Defined intervals are:

1 >= index >= 0.8	predicted substance is into the Applicability Domain of the model
0.8 > index >= 0.65	predicted substance could be out of the Applicability Domain of the model
index < 0.65	predicted substance is out of the the Applicability Domain of the model

1.4 Structural alerts for non biodegradable compounds

The model first check the set of structural alerts related to non ready biodegradability. If some are found from the first set, the compound is predicted as non ready biodegradable, while if only fragments from the second rule-set are found it is predicted as suspicious non ready biodegradable. This means that not only the model provides a prediction, but also a degree of reliability, coming from the statistical quality of the fragments found. The SAs are the following:

Fragments with high specificity:
- NBs1 (1,2-dichlorobenzene), defined by the SMARTS: c1ccc(c(c1)Cl)Cl
- NBs2 (1-ethyl-3-methylbenzene), defined by the SMARTS: c1cc(cc(c1)CC)C
- NBs3 (1,2-dichloroethane), defined by the SMARTS: C(C(Cl))Cl
- NBs4 (2-chloroaniline), defined by the SMARTS: Nc1ccc(cc1Cl)
- NBs5 (diphenylmethanone), defined by the SMARTS: c1ccccc1C(=O)c2ccccc2
- NBs6 (dimethoxyphosphinic acid), defined by the SMARTS: O=P(OC)(OC)O
- NBs7 (cyclohex-4-ene-1,2-dicarbaldehyde), defined by the SMARTS: O=CC1CC=CCC1C(=O)
- NBs8 (benzene-1,3-diamine), defined by the SMARTS: Nc1ccc(c(N)c1)
- NBs9 (1-bromopropane), defined by the SMARTS: CCCBr

- NBs10 (3-chlorophenol), defined by the SMARTS: Oc1ccc(c(c1)Cl)
- NBs11 (fluorine), defined by the SMARTS: F
- NBs12 ((1-phenylethyl)benzene), defined by the SMARTS: c1ccc(cc1)C(c2ccccc2)C
- NBs13 (methoxy(sulfanylidene)phosphinous acid), defined by the SMARTS: P(=S)(OC)O
- NBs14 (N-phenylaniline), defined by the SMARTS: c1ccc(cc1)Nc2ccc(cc2)
- NBs15 (naphthalen-1-amine), defined by the SMARTS: c1ccc2c(c1)cccc2N
- NBs16 (1-methylnaphthalene), defined by the SMARTS: Cc1cccc2ccccc12
- NBs17 (benzyldimethylamine), defined by the SMARTS: C(c1cccc(c1))N(C)C
- NBs18 (2-methylnonane), defined by the SMARTS: CCCCCCCC(C)C
- NBs19 (1,3-benzothiazole), defined by the SMARTS: c1nc2ccccc2s1
- NBs20 (tin), defined by the SMARTS: [Sn]
- NBs21 (methanimine), defined by the SMARTS: C=N
- NBs22 (1,4-diethylbenzene), defined by the SMARTS: c1cc(ccc1C(C))C(C)
- NBs23 (propoxybenzene), defined by the SMARTS: CCCOc1ccccc1
- NBs24 (2-ethyl-1-methoxyhexane), defined by the SMARTS: O(C)CC(CC)CCCC
- NBs25 (2-chlorobenzaldehyde), defined by the SMARTS: O=Cc1c(cccc1Cl)
- NBs26 (1-ethyl-2-methylbenzene), defined by the SMARTS: c1ccc(c(c1)CC)C
- NBs27 (3-chloroaniline), defined by the SMARTS: Nc1ccc(c(c1)Cl)
- NBs28 ((2-methoxyethyl)(propyl)amine), defined by the SMARTS: CCCNCCOC
- NBs29 (2,4-dimethylpent-1-ene), defined by the SMARTS: C=C(C)CC(C)(C)
- NBs30 (bromobenzene), defined by the SMARTS: c1ccc(cc1)Br
- NBs31 (methylcarbamic acid), defined by the SMARTS: O=C(O)NC
- NBs32 (1,1,1-trichloroethane), defined by the SMARTS: CC(Cl)(Cl)Cl
- NBs33 (chloroethene), defined by the SMARTS: C(=CCl)
- NBs34 (dithioperoxol), defined by the SMARTS: SS

Fragments with balanced performance:
- NBb1 (benzylbenzene), defined by the SMARTS: C(c1ccccc1)c2ccccc2
- NBb2 (1-chloro-2-methylbenzene), defined by the SMARTS: Cc1ccccc1Cl
- NBb3 (naphthalene), defined by the SMARTS: c1ccc2c(c1)cccc2
- NBb4 (bromine), defined by the SMARTS: Br
- NBb5 (3-methylaniline), defined by the SMARTS: Nc1cccc(c1)C
- NBb6 (hydroxylamine), defined by the SMARTS: ON
- NBb7 (chlorobenzene), defined by the SMARTS: c1ccc(c(c1)Cl)
- NBb8 (benzenesulfonic acid), defined by the SMARTS: c1ccc(cc1)S(=O)(=O)O
- NBb9 (pentan-2-amine), defined by the SMARTS: C(N)(C)CC(C)
- NBb10 (diazene), defined by the SMARTS: N=N
- NBb11 (halogenated ring structure), defined by the SMARTS: [R][Cl,F,Br,I]

1.5 Structural alerts for biodegradable compounds

If no fragments related to non biodegradability are found, but some related to biodegradability are found, an analogous prediction is provided: ready biodegradable (if fragments from the first set) or suspicious ready biodegradable (if only fragments from the second set are found). If no matching fragments have been found at all, the compound is considered non predictable (not assignable). The

SAs are the following:

Fragments with high specificity:
Bs1 (2-hydroxyethyl acetate), defined by the SMARTS: O=C(OCCO)C
Bs2 (propanedial), defined by the SMARTS: O=CCC(=O)
Bs3 (N-(2-hydroxyethyl)formamide), defined by the SMARTS: O=C(N(CCO))
Bs4 (1-propoxynonane), defined by the SMARTS: O(CCCCCCCCC)CCC
Bs5 (2,6-dimethylhepta-1,5-diene), defined by the SMARTS: C=C(C)CCC=C(C)C
Bs6 (2-methoxybutane), defined by the SMARTS: CCC(OC)C
Bs7 ((dodecyloxy)phosphonous acid), defined by the SMARTS: P(O)(O)OCCCCCCCCCCCC
Bs8 (benzyl formate), defined by the SMARTS: O=C(OCc1ccccc1)
Bs9 (6-oxohexanoic acid), defined by the SMARTS: O=C(O)CCCCC(=O)
Bs10 (docosane), defined by the SMARTS: CCCCCCCCCCCCCCCCCCCCCC

Fragments with balanced performance:
Bb1 (tridecan-1-ol), defined by the SMARTS: OCCCCCCCCCCCCC
Bb2 (butan-2-one), defined by the SMARTS: O=C(C)CC
Bb3 (3-methoxyprop-1-ene), defined by the SMARTS: C(OC)C=C
Bb4 (2-methylhept-2-ene), defined by the SMARTS: C(C)CCC=C(C)C
Bb5 (methyl propanoate), defined by the SMARTS: O=C(OC)CC
Bb6 (butyl formate), defined by the SMARTS: C(=O)OCCCC
Bb7 (1-ethoxybutane), defined by the SMARTS: CCOCCCC
Bb8 (octan-1-ol), defined by the SMARTS: OCCCCCCCC
Bb9 (tridecane), defined by the SMARTS: CCCCCCCCCCCCC
Bb10 (propanoic acid), defined by the SMARTS: CCC(=O)O
Bb11 (benzoic acid), defined by the SMARTS: O=C(O)c1ccc(cc1)
Bb12 (butan-1-ol), defined by the SMARTS: OCCCC
Bb13 (acetamide), defined by the SMARTS: O=C(N)C
Bb14 (acetaldehyde), defined by the SMARTS: O=CC
Bb15 (ethane-1,2-diol), defined by the SMARTS: OCCO
Bb16 (propan-1-amine), defined by the SMARTS: NCCC
Bb17 (sulfanone), defined by the SMARTS: S(=O)
Bb18 (heptanes), defined by the SMARTS: CCCCCCC
Bb19 (anisole), defined by the SMARTS: O(c1ccccc1)C
Bb20 (butane), defined by the SMARTS: CCCC
Bb21 ((chloromethyl)benzene), defined by the SMARTS: c1(ccccc1)C(Cl)
Bb22 (carbonyl bound to aromatic structure), defined by the SMARTS: [a][C;D2]=O
Bb23 (formonitrile), defined by the SMARTS: C#N

1.6 Model statistics

Following, statistics obtained applying the model to its original dataset (for these statistics, the experimental/prediction of NON readily biodegradable has been considered as the positive prediction):

- Training set: n = 486; Accuracy = 0.92; Specificity = 0.95; Sensitivity = 0.90
- Non predicted compounds in the training set: n = 96
- Test set: n = 120; Accuracy = 0.82; Specificity = 0.87; Sensitivity = 0.77
- Non predicted compounds in the test set: n = 26

2. Model usage

2.1 Input

The model accepts as input two molecule formats: SDF (multiple MOL file) and SMILES. All molecules found as input are preprocessed before the calculation of molecular descriptors, in order to obtain a standardized representation of compound. For this reason, some cautions should be taken.

- **Hydrogen atoms**. In SDF files, hydrogen atoms should be explicit. As some times SDF file store only skeleton atoms, and hydrogen atoms are implicit, during the processing of the molecule the system tries to add implicit hydrogens on the basis of the known standard valence of each atom (for example, if a carbon atoms has three single bonds, an hydrogen atom will be added such to reach a valence of four). In SMILES molecules, the default notation uses implicit hydrogen. Anyway please note that in some cases it is necessary to explicitly report an hydrogen; this happens when the conformation is not unambiguous. For example, when a nitrogen atom is into an aromatic ring with a notation like "cnc" it is not clear whether it corresponds to C-N=C or to C-[NH]-C, thus if the situation is the latter, it should be explicitly reported as "c[nH]c".

- **Aromaticity**. The system calculates aromaticity using the basic Hueckel rule. Note that each software for drawing and storing of molecules can use different approaches to aromaticity (for instance, commonly the user can choose between the basic Hueckel rule and a loose approach that lead to considering aromatic a greater number of rings). As in the input files aromaticity can be set explicitly (for instance, in SMILES format by using lowercase letters), during the processing of the molecule the system removes aromaticity from rings that don't satisfy the Hueckel rule. Please note that when aromaticity is removed from a ring, it is not always possible to rebuild the original structure in Kekule form (i.e. with an alternation of single and double bonds, like in the SMILES for benzene, C=1C=CC=CC1), in this case all bonds are set to single. Furthermore, please note that aromaticity detection is a really relevant issue, some molecular descriptors can have significantly different values whether a ring is perceived as aromatic or not. For this reason it is strongly recommended:
- Always use explicit hydrogens in SDF file.
- Avoid explicit aromaticity notation in original files; in this way, the perception of aromaticity is left to the preprocessing step and there is no chance of mistakes due to the transformation of rings that were set to aromatic in the original format but not recognized as aromatic in VEGA.

Note that when some modification of the molecule are performed during the preprocessing (e.g. adding of lacking hydrogens, correction of aromaticity), a warning is given in the remark field of the results.

2.2 Output

Results given as text file consist of a plain-text tabbed file (easily importable and processable by any spreadsheet software) containing in each row all the information about the prediction of a molecule. Note that if some problems were encountered while processing the molecule structure, some warnings

are reported in the last field (Remarks).

Results given as PDF file consists of a document containing all the information about the prediction. For each molecule, results are organized in sections with the following order:

1 – Prediction summary

Here is reported a depiction of the compound and the final assessment of the prediction (i.e. the prediction made together with the analysis of the applicability domain). Note that if some problems were encountered while processing the molecule structure, some warnings are reported in the last field (Remarks).

A graphical representation of the evaluation of the prediction and of its reliability is also provided, using the following elements:

 Compound is classified as readily biodegradable

 Compound is classified as possible readily biodegradable

 Compound is classified as possible NON readily biodegradable

 Compound is classified as NON readily biodegradable

 Prediction has low reliability (compound out of the AD)

 Prediction has moderate reliability (compound could be out of the AD)

 Prediction has high reliability (compound into the AD)

3.1 – Applicability Domain: Similar compounds, with predicted and experimental values

Here it is reported the list of the six most similar compounds found in the training and test set of the model, along with their depiction and relevant information (mainly experimental value and predicted value).

3.2 – Applicability Domain: Measured Applicability Domain scores

Here it is reported the list of all Applicability Domain scores, starting with the global Applicability Domain Index (ADI). Note that the final assessment on prediction reliability is given on the basis of the value of the ADI. For each index, it is reported its value and a brief explanation of the meaning of that value.

4.1 – Reasoning: Relevant chemical fragments and moieties

If some rare and/or missing Atom Centered Fragments are found, they are reported here with a depiction of each fragment.

If some relevant fragments are found (see section 1.4 and 1.5 of this guide), they are reported here (one for each page) with a brief explanation of their meaning and the list of the three most similar compounds that contain the same fragment. Note that if no relevant fragments are found, this section is not shown.

3. Differences from previous versions

3.1 VEGA model history

3.1.1 Version 1.0.8

First official release published in the VEGA platform.

3.1.2 Version 1.0.9

This version is updated with the new calculation core (1.2.0). This update can influence some calculation, in particular similarity evaluation, so there could be some changes in the applicability domain values produced.
The logic of the model has been fixed, as now no prediction is provided if no rules at all match with the given compound.