VEGA

Guide to Persistence in soil Model version 1.0.0

Table of Contents

1. Model explanation

1.1 Introduction

The model is based on the half-lives test data and provides a qualitative evaluation (four classes) of persistence property in the soil compartment. It has been developed using an ensemble of *k*-NN modelling and a set of alerts extracted with Sarpy software, by Istituto di Ricerche Farmacologiche Mario Negri.

1.2 Model details

The model has been built by the integration of a *k*-NN model and a set of structural alerts extracted with Sarpy software and by human experts with the support of the istChemFeat application (developed by Kode srl, http://chm.kode-solutions.net), both developed on a dataset of 568 compounds collected from Gouin, T., Cousins, I., Mackay, D., "Comparison of two methods for obtaining degradation half-lives", *Chemosphere* 56, 2004, 531-535, Gramatica, P., Papa, E., "Screening and ranking of POPs for Global Half-Life: QSAR approaches for prioritization based on molecular structure", *Environ. Sci. Technol.* 41, 2007, 2833-9, Linders J.B.H.J., Jansma J.W., Mensink B.J.W.G., Otermann K., "Pesticedes: Beneaction or Pandora's box? A synopsis of the environmental aspects or 351 pesticides." RIVM Report 679101014, 1994 and USGS (Prioritizing Pesticide Compounds for Analytical Methods Development, 2012). The *k*-NN model has been implemented as described in:

A. Manganaro, F. Pizzo, A. Lombardo, A. Pogliaghi, E. Benfenati, "Predicting persistence in the sediment compartment with a new automatic software based on the k-Nearest Neighbor (*k*-NN) algorithm", *Chemosphere* (2015), accepted paper.

This model has been built with the istKNN application (developed by Kode srl, http://chm.kode-solutions.net) and it is based on the similarity index developed inside the VEGA platform; the index takes into account several structural aspects of the compounds, such as their fingerprint, the number of atoms, of cycles, of heteroatoms, of halogen atoms, and of particular fragments (such as nitro groups). The index value ranges from 1 (maximum similarity) to 0. On the basis of this structural similarity index, the four compounds from the dataset resulting most similar to the chemical to be predicted are taken into account; compounds with a similarity value lower than 0.75 are discarded, and if only one compound remains available for prediction, it is kept only if it has a similarity value higher than 0.8. If no compounds fall under these conditions, no prediction is provided. The prediction is calculated as the most representative class in the compound selected with the above mentioned procedure, considering their similarity index values as a weight so that the most similar compounds have an higher influence in the prediction.

The overall architecture provides firstly the prediction as calculated by the *k*-NN model. If some structural alerts are found, they do not change the prediction but modify the applicability domain value: if the alerts confirm the *k*-NN prediction, the applicability domain index (ADI) value increases, if the alert are in disagreement with the prediction the ADI value decreases. The alerts are anyway used to

provide a prediction if the *k*-NN model is not able to predict the compound.
The predicted value is provided as one of the following four classes based on the standard labelling of non persistent (nP), persistent (P) and very persistent (vP) compounds: nP, nP/P, P/vP, vP. The nP class defines compound with HL certainly below the threshold of 120 days. The vP class defines compound with HL certainly higher the threshold of 180 days. The remaining two classes indicate a not entirely certain situation: nP/P class represent compounds with HL values falling near the P threshold of 120 days; the P/vP class represent compounds with HL values falling near the vP threshold of 180 days.

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.

- **Similar molecules with known experimental value**. This index takes into account how similar are the most similar compounds used by the *k*-NN model. 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 > 0.8	strongly similar compounds with known experimental value in the training set have been found
0.8 >= index > 0.6	only moderately similar compounds with known experimental value in the training set have been found
index <= 0.6	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 classification accuracy in prediction for the most similar compounds used by the *k*-NN model . Values near 1 mean that the predicted compounds falls 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:

1 >= index > 0.9	accuracy of prediction for similar molecules found in the training set is good
0.9 >= 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 most similar compounds used by the *k*-NN model. 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 > 0.9	similar molecules found in the training set have experimental values that agree with the predicted value
0.9 >= 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

- **Structural alerts concordance**. This index takes into account the concordance between the prediction provided by the *k*-NN model and the alerts found. Defined values are:

index = 1	all alerts are related to experimental values in agreement with the prediction, thus confirming the *k*-NN output.
index = 0.9	no alerts have been found, thus it is not possible to confirm the *k*-NN output.
index = 0.85	no *k*-NN prediction is available and the final prediction is based only on the found alerts.
index = 0.7	one or more alerts are related to experimental values not in agreement with the prediction, thus conflicting with the *k*-NN output.

- **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 >= 0.85	predicted substance is into the Applicability Domain of the model
0.85 > 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 persistent compounds

The following SAs have been extracted from the original dataset and are related to compounds which are nP in the soil compartment:

- nP alert no. 1, defined by the SMARTS: O=C(OCCC)
- nP alert no. 2, defined by the SMARTS: O=C(CC)C
- nP alert no. 3, defined by the SMARTS: C(O)c1ccccc1C
- nP alert no. 4, defined by the SMARTS: Cc1c(ccc(c1))OCC
- nP alert no. 5, defined by the SMARTS: COC(=O)CO
- nP alert no. 6, defined by the SMARTS: C(O)C=C
- nP alert no. 7, defined by the SMARTS: C(=NOC(=O)NC)C
- nP alert no. 8, defined by the SMARTS: N(C)CCCl
- nP alert no. 9, defined by the SMARTS: CSCc1ccccc1
- nP alert no. 10, defined by the SMARTS: O(C)CCCl
- nP alert no. 11, defined by the SMARTS: [P]
- nP alert no. 12, defined by the SMARTS: C(CN(C)C)S
- nP alert no. 13, defined by the SMARTS: C(=S)N(C)
- nP alert no. 14, multiple primary alcohols, defined by the SMARTS: C[CH2]O
- nP alert no. 15, multiple esters (aromatic), defined by the SMARTS: AC(=O)O*
- nP alert no. 16, single oximes (aliphatic), defined by the SMARTS: AC(A)=NO*
- nP alert no. 17, esters (aliphatic), defined by the SMARTS: AC(=O)O[*;!H]
- nP alert no. 18, single aldehydes (aliphatic), defined by the SMARTS: A[CH1](=O)
- nP alert no. 19, single carboxylic acids (aliphatic), defined by the SMARTS: AC(=O)O
- nP alert no. 20, single (thio-) carbamates (aliphatic), defined by the SMARTS: A[O,S]C(=[O,S])N(A)A
- nP alert no. 21, single ketones (aromatic), defined by the SMARTS: aC(=O)*
- nP alert no. 22, single phosphates/thiophosphates, defined by the SMARTS: *[O,S]=P([O,S]*)([O,S])[O,S]*

1.5 Structural alerts for very persistent compounds

The following SAs have been extracted from the original dataset and are related to compounds which are vP in the soil compartment:

- vP alert no. 1, defined by the SMARTS: c1ccc(c(c1)c2ccccc2Cl)
- vP alert no. 2, defined by the SMARTS: c1c(Cl)cc2Oc3cc(Cl)cc(Cl)c3Oc2c1

1.6 Model statistics

Following, statistics obtained applying the *k*-NN model to its original dataset, with a leave-one-out approach (*k*-NN for each compound has been performed on the whole dataset without the compound itself):

- n = 568; Accuracy = 0.72
- Non predicted compounds: n = 9

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 nP

🟡 Compound is classified as nP/P

🟠 Compound is classified as P/vP

🔴 Compound is classified as vP

★☆☆ 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.