Guide to BCF Meylan Model version 1.0.3

Table of Contents

1. Model explanation

1.1 Introduction

The model provides a quantitative prediction of bioconcentration factor (BCF) in fish, given in log(L/kg). It is implemented inside the VEGA online platform, accessible at:http://www.vega-qsar.eu/ The model implements the Meylan model, as described in EPI Suite BCFBAF module: http://www.epa.gov/oppt/exposure/pubs/episuite.htm

1.2 Model details

The model is based on the method proposed by Meylan et al (Meylan W.M., Howard P.H., Boethling R.S. et al. Improved Method for Estimating Bioconcentration / Bioaccumulation Factor from Octanol/Water Partition Coefficient. 1999, Environ. Toxicol. Chem. 18(4): 664-672) et implemented in the EPI Suite BCFBAF module (http://www.epa.gov/oppt/exposure/pubs/episuite.htm). The model provides a BCF prediction based on different regression equations or fixed values, selected on the basis of an initial classification between ionic and non-ionic compounds, and on the value of the predicted logP value.

For the purpose of the model, ionic compounds include carboxylic acids, sulfonic acids and salts of sulfonic acids, and charged nitrogen compounds (nitrogen with a +5 valence such as quaternary ammonium compounds). All other compounds are classified as non-ionic. The logP prediction is provided by the VEGA logP model.

The original dataset from EPI Suite has been taken, then processed and cleared from duplicates and compounds provided with structure that had problems. The final dataset has 662 compounds.

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). Note that when the experimental value for the given compound is found, the Applicability Domain indices are calculated only considering this value, without taking into account the first *n* similar compounds.

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 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:

1 >= index > 0.9	strongly similar compounds with known experimental value in the training set have been found
0.9 >= index > 0.75	only moderately similar compounds with known experimental value in the training set have been found
index <= 0.75	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 error in prediction for the two 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:

index < 0.5	accuracy of prediction for similar molecules found in the training set is good
0.5 <= index <= 1.0	accuracy of prediction for similar molecules found in the training set is not optimal
index > 1.0	accuracy of prediction for similar molecules found in the training set is not adequate

- **Concordance with similar molecules (average difference between target compound prediction and experimental values of 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 agrees with the experimental values found in the model's space, thus the prediction is reliable. Defined intervals are:

index < 0.5	similar molecules found in the training set have experimental values that agree with the target compound predicted value
0.5 <= index <= 1.0	similar molecules found in the training set have experimental values that slightly disagree with the target compound predicted value
index > 1.0	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 falls in an area of the model's space where the model gives reliable predictions without any outlier value. Defined intervals are:

index < 0.5	the maximum error in prediction of similar molecules found in the training set has a low value, considering the experimental variability

0.5 <= index < 1.0	the maximum error in prediction of similar molecules found in the training set has a moderate value, considering the experimental variability
index >= 1.0	the maximum error in prediction of similar molecules found in the training set has a high value, considering the experimental variability

- **LogP reliability**. This index takes into account the reliability of the logP value used in the model. Note that the Meylan BCF model is strongly based on the logP prediction of the compound, thus this index is highly relevant for the assessment of the final prediction. The reliability of the logP value comes from the assessment of the VEGA LogP model (that provides the used logP value), which is also provided in the "Prediction summary" section of the report. Defined intervals are:

index = 1	reliability of logP value used by the model is good
index = 0.7	reliability of logP value used by the model is not optimal
index = 0	reliability of logP value used by the model is not adequate

- **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

- **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 > 0.75	predicted substance could be out of the Applicability Domain of the model
index <= 0.75	predicted substance is out of the the Applicability Domain of the model

1.6 Model statistics

Following, statistics obtained applying the model to its original dataset:

- Training set: n = 516; R^2 = 0.80; RMSE = 0.55
- Test set: n =146; R^2 = 0.79; RMSE = 0.66

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 warning 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). Following, all information related to the prediction are reported (the calculated logP, the reliability of the calculated logP, the classification of the given compound as ionic or non-ionic). The prediction and the experimental value (if available) are given in log(L/kg), the same prediction expressed in L/kg is also provided. Note that if some problems were encountered while processing the molecule structure, some warning 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 non-bioaccumulative, logBCF value is less than 2.7


Compound could be bioaccumulative, logBCF value is more than 2.7 and less than 3.3


Compound is bioaccumulative, logBCF value is more than 3.3


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.

4.2 – Reasoning: Analysis of molecular descriptors

Here it is reported an analysis on the fundamental descriptor for the BCF model, LogP, made of two charts. The first one is a scatter plot of LogP against response values for all compounds of the training set, and the LogP value against the predicted value for the studied compound. The second

one is a scatter plot of LogP against response values only for the three most similar compounds in the training set where red dot is the value of the studied compound, black outlined circles represents experimental values of compounds from training set, black dots represents predicted value of the same compound; the size of the circle is proportional to the similarity to the studied compound.

3. Differences from previous versions

3.1 VEGA model history

3.1.1 Version 1.0.0

First official release published in the VEGA platform.

3.1.2 Version 1.0.2

This version is updated with the new calculation core (1.1.1) based on a new release of the CDK libraries (1.4.9). These updates can influence the calculation, so there could be some changes in the predictions produced.
The new calculation core implements a new version of the algorithm used for calculating the similarity index. This means that the list of similar molecules given as part of the applicability domain evaluation will often be different from the ones produced by older releases of the model. Furthermore, the applicability domain index (ADI) itself and the final assessment could often be different.
Model statistics in the current guide have been updated with the new values.
Some thresholds for the applicability domain sub-indices have been revised to obtain better performances. Also, the values for the conservative intervals have been revised on the basis of the new applicability domain results.
The section “Analysis of molecular descriptors” (LogP) has been added in the PDF output .

3.1.3 Version 1.0.3

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.