Guide to BCF Model
version 2.1.11

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 extends the original CAESAR model, freely available at: http://www.caesar-project.eu/software/

1.2 Model details

Two models, Model A and Model B, have been used to build hybrid model, Model C. In the proposed approach, the outputs of the individual models (Model A and B) were used as inputs of the hybrid model. Model A was developed by Radial Basis Function Neural Networks (RBFNN) using an heuristic method to select the optimal descriptors; Model B was developed by RBFNN using genetic algorithm for the descriptors selection. RBFNN was used with a Matlab function for building the models. An in-house software made as a PC-Windows Excel macro was used to combine Models A and B within the Model C. Model A used an heuristic method to select the optimal descriptors and Model B used genetic algorithm for the descriptors selection. Full reference and details of the used formulas can be found in:

Zhao, C., Boriani, E., Chana, A., Roncaglioni,A., Benfenati, E. A new hybrid system of QSAR models for predicting bioconcentration factors (BCF). Chemosphere (2008), 73, 1701-1707.

Lombardo A, Roncaglioni A, Boriani E, Milan C, Benfenati E. Assessment and validation of the CAESAR predictive model for bioconcentration factor (BCF) in fish. Chemistry Central Journal (2010), 4 (Suppl 1).

The descriptors used are the following:
- Moriguchi octanol-water partition coefficient (MlogP).
- Moran autocorrelation of lag 5, weighted by atomic van der Waals volumes (MATS5V): molecular descriptor calculated from the molecular graph by summing the products of atom weights of the terminal atoms of all paths of the considered path length (the lag).
- Number of chlorine atoms (Cl-089), Cl attached to carbon (sp2).
- Second highest eigenvalue of Burden matrix, weighted by atomic polarizabilities (BEHp2).
- Geary autocorrelation of lag 5, weighted by atomic van der Waals volumes (GATS5V): molecular descriptor calculated from the molecular graph by summing the products of atom weights of the terminal atoms of all paths of the considered path length (the lag).
- Solvation connectivity index chi-0 (XOSolv): molecular descriptor designed for modeling solvation entropy and describing dispersion interactions in solution.
- Sum of all -Cl groups E-state values in molecule (SsCl).
- Absolute eigenvalues sum from electronegativity weighted distance matrix (Aeige).

The descriptors were calculated, in the original CAESAR version, by means of dragonX software and are now entirely calculated by an in-house software module in which they are implemented as described in: R. Todeschini and V. Consonni, Molecular Descriptors for Chemoinformatics, Wiley-VCH, 2009.

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 > 0.7	strongly similar compounds with known experimental value in the training set have been found
0.7 >= index > 0.65	only moderately similar compounds with known experimental value in the training set have been found
index <= 0.65	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 < 0.4	accuracy of prediction for similar molecules found in the training set is good
0.4 <= index <= 0.7	accuracy of prediction for similar molecules found in the training set is not optimal
index > 0.7	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 < 0.4	similar molecules found in the training set have experimental values that agree with the target compound predicted value
0.4 <= index <= 0.7	similar molecules found in the training set have experimental values that slightly disagree with the target compound predicted value
index > 0.7	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 < 0.8	the maximum error in prediction of similar molecules found in the training set has a moderate value, considering the experimental variability
index >= 0.8	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:

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

- **Descriptors noise sensitivity analysis**. This index checks whether the predicted compound falls in a reliable and stable descriptors space or not. A sequence of random scrambling (noise) is applied to the

descriptors calculated for the considered compound, and it is checked if the perturbation of descriptors lead to a significant change in the prediction; if the studied descriptors space is stable, these changes should be of little entity. After a large number of such random scrambling, a final index is calculated. Defined intervals are:

1 >= index > 0.8	predictions has a good response to noise scrambling, thus shows a good reliability
0.8 >= index > 0.5	predictions has a not so good response to noise scrambling, thus shows an uncertain reliability
index <= 0.5	predictions has a bad response to noise scrambling, thus shows a low reliability

- **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 = 1	descriptors for this compound have values inside the descriptor range of the compounds of the training set
index = 0	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 >= 0.7	predicted substance could be out of the Applicability Domain of the model
index < 0.7	predicted substance is out of the the Applicability Domain of the model

1.4 Structural Alerts for outliers

The model implements the detection of a set of Structural Alerts that have been found only in compounds that are outlier (labeled as SO). When such SO are found, the compound is declared out of the applicability domain. The SO for outlier compounds are the following:
- SO 01: 6 Cl atoms in the molecule
- SO 02: 2 t-butyl linked to aromatic
- SO 03: Si atom in the molecule
- SO 04: Sn atom in the molecule
- SO 05: O linked to aromatic and 3 Br/Cl linked to aromatic
- SO 06: Azo group liked to aromatic
- SO 07: 3 Nitro-groups linked to aromatic
- SO 08: Peroxide
- SO 09: Phosphinothioyl-oxy-imino
- SO 010: 10 F atoms in the molecule
- SO 011: Phosphorodithioate

1.5 Other Structural Alerts

Other relevant Structural Alerts have been studied and proposed for reasoning, each one is related to a class of chemicals that have a particular BCF behavior (they are labeled as SR). The relevant SR are the following, given with the full explanation of the behavior they are bound to:
- SR 01: O=Cc1ccccc1 moiety; this SA has been found only in non-bioaccumulative compounds (24 chemicals), even when the logP value was higher than 3.
- SR 02: Carbonyl residue; this SA has been found to be present in a very large (112) number of non-bioaccumulative compounds, even when the logP value was higher than 3.
- SR 03: O-P=O residue; this SA has been found only in non-bioaccumulative compounds (45 chemicals), even when the logP value was higher than 3.
- SR 04: Thiobenzene residue; this SA has been found only in non-bioaccumulative compounds (39 chemicals), even when the logP value was higher than 3.
- SR 05: Tertiary amine; this SA has been found to be present in a large number of non-bioaccumulative compounds (28), even when the logP value was higher than 3.
- SR 06: Triazole ring; this SA has been found to be present in a number of non- bioaccumulative compounds (16), even when the logP value was higher than 3.
- SR 07: Clc1ccccc1c1ccc(Cl)cc1 moiety; this SA has been found only in bioaccumulative compounds (15 chemicals). The high lipophylicity of this moiety increases the bioaccumulative behavior.
- SR 08: C1cc(Oc2ccccc2)ccc1Cl; this SA has been found only in bioaccumulative compounds (9 chemicals). The high lipophylicity of this moiety increases the bioaccumulative behavior.
- SR 09: Clc1cc(c2ccccc2)c(Cl)cc1; this SA has been found only in bioaccumulative compounds (15 chemicals). The high lipophylicity of this moiety increases the bioaccumulative behavior.

Furthermore, another set of Structural Alerts for polar groups (labeled as PG) is used for reasoning purpose: usually, the presence of one or more polar groups is related to high hydrophilicity. These SAs have been divided into 3 groups, starting from more relavant (under the aspect of polarity); they are searched in a progressive way, so that if some SAs of the first group are found, no more groups are searched, otherwise the reasearch proceed with the second group, and so on. The group are the following:

First group:
- PG 01: COOH group.
- PG 02: SO3H group.
- PG 03: PO3 group.
- PG 04: PO2S group.
- PG 05: POS2 group.

Second group:
- PG 06: OH group.
- PG 07: NH2 group.
- PG 08: CS2 group.

Third group:
- PG 09: >C=O group

1.6 Model statistics

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

- Training set: n = 378; R^2 = 0.82; RMSE = 0.34
- Test set: n = 95; R^2 = 0.78; RMSE = 0.38

Furthermore, the statistics for the test set considering the Applicability Domain (AD) index is here reported; the AD index is used, as in the final model's assessment, in order to divide results in three groups (into AD, possibly out of AD, out of AD), showing that compounds considered into AD have better performance than the others:

- Test set with AD index greater than 0.85 (compounds into the AD):
  n = 43; R^2 = 0.84; RMSE = 0.28
- Test set with AD index between 0.85 and 0.7 (compounds could be out of AD):
  n = 26; R^2 = 0.63; RMSE = 0.49
- Test set with AD index lower than 0.7 (compounds out of the AD):
  n = 26; R^2 = 0.86; RMSE = 0.45

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 predicted values of the two sub-models, the calculated logP). 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)

2 – Possible use and uncertainty

Here is reported a classification for two relevant thresholds (3.3 and 3.7 log units). To the given prediction is associated a conservative interval, if this adjusted value falls under the given threshold the compound can be safely classified under the threshold. Intervals are determined on the basis of the AD index value, for each threshold the original BCF dataset has been studied and each interval defined as the minimum value to be added to the prediction in order to obtain no false negative classification. If compound is outside the applicability domain, no confidence interval is available. In the following these intervals are reported:

	ADI = 1	1 < ADI <= 0.85	0.85 < ADI <= 0.7	ADI < 0.7
For 3.3 threshold	0.6 log units	0.6 log units	0.8 log units	n.a
For 3.7 threshold	0.5 log units	0.9 log units	0.9 log units	n.a

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 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.

4.2 – Reasoning: Analysis of molecular descriptors

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

The VEGA model has several differences that can lead to prediction that can be slightly different from the ones produced by the CAESAR model. Mainly, descriptors are no longer calculated by means of third party software thus, even if the algorithm definition is the same, their values could be different for some molecules. Please note that also the algorithm for the calculation of similarity is slightly different from the one implemented in CAESAR.

3.2 VEGA model history

3.2.1 Version 2.1.7

First official release published in the VEGA platform.

3.2.2 Version 2.1.8

Some minor code updates, mainly due to changes in the core.
Explicit codes for Structural Alerts have been added (SO, SR, PG).
There are NO changes in prediction values and AD assessment.

3.2.3 Version 2.1.9

This version is updated with the new calculation core (1.0.26) where similarity algorithm is slightly changed. The new version considers halogen atoms are really similar, especially Chlorine and Bromine atoms are considered almost the same. The main difference with previous algorithm can be thus seen just for halogenated compounds.
A more precise check for similarity has been introduced for the extraction of experimental values, in order to avoid mismatches (as the similarity index is based on fingerprints, there are some rare cases in which a value equal to 1 does not points to a exactly isomorph compound).
Some minor bugs in the procedure for reading molecule structures have been fixed; some compounds, previously not loaded, could now be correctly processed.
Structural Alerts for polar groups have been revised and extended.
There are NO changes in prediction values, but as similarity is changed some small differences in AD assessment can be found.

3.2.4 Version 2.1.10

This version is updated with the new calculation core (1.0.27), that generates a graphically renewed PDF report. In this version, the propositions for prediction and assessment are changed, but there are NO changes in their values.

3.2.5 Version 2.1.11

This version is updated with the new calculation core (1.0.28). A problem in the section “possible use and uncertainty” has been fixed, the assessment for the 3.7 thresholds was in some cases wrong. In this version there are NO changes in the final prediction and assessment.