Guide to BCF Read-Across
version 1.1.0

Table of Contents

1. Model explanation

1.1 Introduction

The model performs a read-across and 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/

1.2 Model details

The model performs a read-across on a dataset of 860 chemicals. This dataset has been made by Istituto di Ricerche Farmacologiche Mario Negri, merging experimental data from several reliable sources, including the original dataset of the CAESAR BCF model (note that experimental values may differ from the ones in the CAESAR BCF dataset, as this new dataset has been built including more sources). The read-across 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 0.7 are discarded, and if only one compound remains available for prediction, it is kept only if it has a similarity value higher than 0.75. If no compounds fall under these conditions, no prediction is provided. Furthermore, if the range of experimental values observed in the chosen molecules is higher than 3.5 log units, no prediction is provided. The estimated BCF value is calculated as the weighted average value of the experimental values of the chosen compounds, using their similarity values as weight. Their similarity values are raised to the power of 3 in order to enhance the weight of the most similar compounds in the calculated 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. For each index, including the final ADI, two intervals for its values are defined, such that the first interval corresponds to a positive evaluation, and the second one corresponds to a negative evaluation.
Following, all applicability domain components are reported along with their explanation:

- **Similar molecules with known experimental value**. This index takes into account how similar are the most similar compounds used by the KNN 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 > 0.75	strongly similar compounds with known experimental value in the training set have been found
0.75 >= 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 (average error) of prediction for similar molecules**. This index takes into account the error in prediction for the most similar compounds used by the KNN model. 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 < 0.6	accuracy of prediction for similar molecules found in the training set is good
0.6 <= index <= 1.2	accuracy of prediction for similar molecules found in the training set is not optimal
index > 1.2	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 most similar compounds used by the KNN model. 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 < 0.6	similar molecules found in the training set have experimental values that agree with the target compound predicted value
0.6 <= index <= 1.2	similar molecules found in the training set have experimental values that slightly disagree with the target compound predicted value
index > 1.2	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 most similar compounds used by the KNN model. 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 < 0.6	the maximum error in prediction of similar molecules found in the training set has a low value, considering the experimental variability
0.6 <= index < 1.2	the maximum error in prediction of similar molecules found in the training set has a moderate value, considering the experimental variability
index >= 1.2	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

- **Global AD Index**. The final global index takes into account the previous indices, in order to give a general global assessment on the applicability domain for the predicted compound. If at least one of the previous indices has a negative evaluation, the final global index will result in an assessment of unreliability; if all indices have positive evaluation, then the global index will result in an assessment of reliability. In both cases, the global index value is calculated as the average value of the similarity index for the three compounds taken into account for the read-across.

1.4 Model statistics

Following, statistics obtained applying the read-across prediction to its original dataset, with a leave-one-out approach (read-across for each compound has been performed on the whole dataset without the compound itself)

- n = 836; R^2 = 0.67; RMSE = 0.76
- Non predicted compounds: n = 24

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). The number of molecules used for the KNN prediction is reported, please note that these *k* molecules correspond to the first *k* molecules shown in the list of similar compounds (section 3.1). The prediction and the experimental value (if available) are given in log(L/kg). 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). Note that the first three compounds shown are the molecules used for the read-across.

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.

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.

3.1.3 Version 1.1.0

This version is a whole new implementation of the model (now based on the istKNN software), both predictions and applicability domain values can be remarkably different from previous versions.