Guide to LogP Model
version 1.1.4

Table of Contents

1. Model explanation

1.1 Introduction

The model provides a quantitative prediction of water/octanol partition coefficient (LogP). It is implemented inside the VEGA online platform, accessible at: http://www.vega-qsar.eu/

1.2 Model details

The model is based on the Atom/Fragment Contribution (AFC) method from the work of Meylan and Howard (Meylan, W.M. and P.H. Howard, Atom/fragment contribution method for estimating octanol-water partition coefficients. 1995, J. Pharm. Sci. 84: 83-92.), as implemented in the EPI Suite KOWWIN module (http://www.epa.gov/oppt/exposure/pubs/episuite.htm). The calculated model has a lower bound of -5.0 log units (all predictions lower than this value are set to -5.0). A dataset of compounds with experimental logP values has been built starting from the original dataset provided in EPI suite. The set has been processed and cleared from compounds that were replicated or that had problems with the provided molecule structure. The final dataset has 9,961 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).

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

- **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

1.4 Model statistics

On the pruned training set from EPI Suite KowWin module (9,961 compounds), the logP model has the following statistics:

- Training set: n = 9961; R^2 = 0.86; RMSE = 0.76

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 values of the two logP descriptors). 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 has a low logP value (less than 3.0)


Compound has a high logP value (more than 3.0 and less than 8.0)


Compound has a very high logP value (more than 8.0)


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.

3. Differences from previous versions

3.1 VEGA model history

3.1.1 Version 1.0.1

First official release published in the VEGA platform.

3.1.2 Version 1.0.2

Dataset has been revised, several duplicate compounds were removed. New dataset consists of 2524 molecules. Statistics in the current document have been updated.

3.1.3 Version 1.0.3

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).
The final assessment has been fixed, in previous version a bug occurred (the final assessment was not consistent with the AD assessment reported in the following sections)
There are NO changes in prediction values, but as similarity is changed and a bug fixed, some differences in AD assessment can be found.

3.1.4 Version 1.0.4

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.

3.1.5 Version 1.1.0

This version is a full update, the logP prediction method is changed from previous version. The logP calculation is based on Meylan method, but ALogP and MLogP values, as calculated in previous versions, are still provided.

3.1.6 Version 1.1.2

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.
A lower bound of -5.0 log units for calculated logP values has been introduced.

3.1.7 Version 1.1.3

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.
A further check of structures and experimental data has been performed, resulting in the removal of some compounds from the original dataset (10,005 compounds) which had incosistent experimental data.