	QMRF identifier (JRC Inventory): **To be entered by JRC**
	QMRF Title: **ACD/Percepta QSAR for the intrinsic solubility of organic compounds in water (log S0)**
	Printing Date: **13-ott-2017**

1.QSAR identifier

1.1.QSAR identifier (title):

ACD/Percepta QSAR for the intrinsic solubility of organic compounds in water (log S0)

1.2.Other related models:

ACD/Percepta QSAR for the solubility of organic compounds in pure water (log Sw). (It should be noted that references reported in section 2.7 are referred to this QSAR model)

1.3.Software coding the model:

ACD/Percepta 2016

ACD/Labs, Inc. 110 Yonge Street, 14th floor, Toronto, Ontario, Canada M5C 1T4

http://acdlabs.com/products/percepta/

2.General information

2.1.Date of QMRF:

October 2017

2.2.QMRF author(s) and contact details:

[1]Andrius Sazonovas ACD/Labs, Inc. A.Mickeviciaus g. 29, LT-08117, Vilnius, Lithuania. Phone: +370 5 262 40 32; fax: +370 5 262 37 28 andrius.sazonovas@acdlabs.com http://www.acdlabs.com

[2]Pranas Japertas ACD/Labs, Inc. A.Mickeviciaus g. 29, LT-08117, Vilnius, Lithuania. Phone: +370 5 262 40 32; fax: +370 5 262 37 28 pranas.japertas@acdlabs.com http://www.acdlabs.com

[3]Remigijus Didziapetris ACD/Labs, Inc. A.Mickeviciaus g. 29, LT-08117, Vilnius, Lithuania. Phone: +370 5 262 40 32; fax: +370 5 262 37 28 remigijus.didziapetris@acdlabs.com http://www.acdlabs.com

[4]Kiril Lanevskij ACD/Labs, Inc. A.Mickeviciaus g. 29, LT-08117, Vilnius, Lithuania. Phone: +370 5 262 40 32; fax: +370 5 262 37 28 kiril.lanevskij@acdlabs.com http://www.acdlabs.com

[5]Justas Dapkunas ACD/Labs, Inc. A.Mickeviciaus g. 29, LT-08117, Vilnius, Lithuania. Phone: +370 5 262 40 32; fax: +370 5 262 37 28 justas.dapkunas@acdlabs.com http://www.acdlabs.com

[6]Simona Kovarich S-IN Soluzioni Informatiche Srl Via G. Ferrari, 14, I-36100 Vicenza (Italy) simona.kovarich@s-in.it www.s-in.it

2.3.Date of QMRF update(s):

2.4.QMRF update(s):

2.5.Model developer(s) and contact details:

ACD/Labs, Inc. A.Mickeviciaus g. 29, LT-08117, Vilnius, Lithuania. Phone: +370 5 262 40 32; fax: +370 5 262 37 28 vilnius@acdlabs.com http://www.acdlabs.com

2.6.Date of model development and/or publication:

2012.11.23

2.7.Reference(s) to main scientific papers and/or software package:

[1]Japertas, P., Sazonovas, A., Didziapetris, R., and Petrauskas, A., Similarity based assessment of model applicability domain and quantitative evaluation of the reliability of the prediction. Abstr. Paper. Am. Chem. Soc., 2008 , 235, Meeting abstract 271-COMP

http://oasys2.confex.com/acs/235nm/techprogram/P1160866.HTM
[2]Japertas, P., Sazonovas, A., Didziapetris, R., and Petrauskas, A., Similarity based correction for the predictions of compounds physicochemical properties. Abstr. Paper. Am. Chem. Soc., 2008, 235, Meeting abstract 247 - MEDI
http://oasys2.confex.com/acs/235nm/techprogram/P1160811.HTM
[3]Japertas, P., Didziapetris, R., and Petrauskas, A., Fragmental methods in the design of new compounds. Applications of Advanced Algorithm Builder, Quant. Struct-Act. Relat., 2002; 21, 23-37
http://www3.interscience.wiley.com/journal/93521415/abstract

2.8.Availability of information about the model:

Model is proprietary, however the compounds used to derive the model and their experimental data are available within the corresponding software products.

2.9.Availability of another QMRF for exactly the same model:

None to date

3.Defining the endpoint - OECD Principle 1

3.1.Species:

Not applicable

3.2.Endpoint:

QMRF 1. Physical Chemical Properties QMRF 1. 3. Water solubility

3.3.Comment on endpoint:

The logarithm of a compound intrinsic solubility in water at 25°C.

3.4.Endpoint units:

log(mol/l) or log(mmol/ml)

3.5.Dependent variable:

LogS0

3.6.Experimental protocol:

The dataset used to develop the reported model has been compiled from a great number of different sources, including books [4-6] as well as various articles from peer-reviewed scientific journals. Articles reporting the models of solubility in pure water (LogSw) by other authors were the predominant type among analyzed literature, meaning that each publication contained larger collections of experimental data (usually in the order of tens or hundreds compounds) compiled from corresponding original experimental articles.

In general, the data sources used to collect the data covering a wide variety of experimental protocols used to determine log Sw values reported within them. This includes the classical yet very labor-intensive saturation shake-flask log Sw determination method involving lengthy equilibration periods, as well as methods aiming at some high-througput needs like turbidimetric method, a reverse-phase HPLC based 96-well plate assay adaptation of the shake-flask methodology, various potentiometric methods ranging from the standard one to one of the latest developments called dissolution template titration (DTT), as well as the real high-throughput fast UV plate spectrophotomer methods involving aqueous dilution or co-solvent

utilization strategies. Estimated detection limits of the listed methods range from 1 g/mL down to 5 ng/mL. For a comprehensive overview of the experimental log Sw measurement techniques please see [7].

Original LogSw data had been converted to LogS0 prior to modeling.

3.7.Endpoint data quality and variability:

log Sw is a relatively easily measured property. As a result the experimental data quality, which is usually inversely proportional to the complexity of the experiment, is reasonably good. Independent external studies show that the error between the log Sw measurements performed by different laboratories using the same protocol (reproducibility) can be expected to be within 0.5 logarithmic units [8,9]. The LogSw data is converted to LogS0 data using a set of mass balance equations involving the predicted information about the distribution of various ionized forms of the compound at a certain pH. Since the accuracy of pKa predictions for the conventional structures is generally regarded as acceptable, it is assumed that any additional errors introduced into the data by this conversion step is negligible.

The characteristics of the entire dataset compiled for the task of this model development is:

No. of compounds = 6807

Min. Value = -12.79

Max. Value = 2.06

Std. Dev. = 2.13

Skewness = -0.69

4.Defining the algorithm - OECD Principle 2

4.1.Type of model:

QSAR

4.2.Explicit algorithm:

GALAS algorithm

Global linear baseline QSAR + local similarity based corrections

GALAS (Global, Adjusted Locally According to Similarity) models consist of two parts: (1) a global linear model, and (2) local corrections based on the analysis of global model performance for the most similar compounds from the training set.

The global QSAR was developed using PLS in combination with bootstrapping technique. This method implies random compound sampling from the initial training set, i.e. generation of new “training sub-sets”. Each of the sampled sub-sets is of the same size as the initial training set, however, random manner of their population results in some compounds being selected more than once, others being omitted. This procedure is performed 100 times and an independent PLS model is derived for every sub-set.

Each of those PLS models is based on 2D fragmental descriptors:
log LC50 = $SUM_{[i=1..n]}(a_i*f_i) + c$
where $f_i$is the number of occurences of the i-th fragment in
a molecule, a_i- its statistical coefficient, and c -
intercept.

As a result, each global QSAR model actually represents an ensemble of 100 PLS models, providing each compound with a vector of 100 log LC50 predictions, each based on a slightly different sub-set of the initial training set. It is defined that two compounds with similar trends in the variation patterns of the 100 value vectors predicted by a global QSAR model are considered similar in terms of the analyzed property, i.e. the differences in the compound sets used to parameterize each of 100 PLS models, constituting a baseline model, affect estimations for the two compounds in a similar way. The correlation coefficient of the two vectors is called an Individual Similarity Index between two compouds (SI_i). An analogous definition of the "property-specific" or dynamic similarity was first used by Tetko and his co-workers [10-14] and this method has been recently used in the analysis of the acute toxicity and CYP3A4 inhibition data [15-16].
With the available robust similarity measure, it becomes possible to analyse the performance of the baseline QSAR model in the local chemical environment of a query molecule represented by the most similar compounds in the training set. In case any systematic errors are encountered for sufficiently similar compounds, a local correction (*Delta*) is calculated. Later on it is possible to train the model quickly and efficiently using new experimental data by just adding it to this second similarity correction calculation procedure, without the time costly baseline model re-training.

4.3.Descriptors in the model:

Fragmental descriptors Dimentionless (occurence count) Fixed set of fragmental descriptors, based on the expanded list of Platt's type fragments (see [17]). A fixed and relatively small set of fragments was used due to the specifics of the employed modeling methodology. In order for the correlation between two compound vectors of log S0 predictions coming from a baseline QSAR model to be representative of compound similarity in terms of the analyzed property, these vectors have to be parameterized using exactly the same set of fragmental descriptors. This prevents the use of any sort of automated fragmentation routines (atom based, isolating carbon based, chain based, etc.) that result in a dynamic set of fragments depending on the training set structures. They leave the possibility that for any query structure from outside the training set the same rules will yield certain new fragments not encountered in the training set molecules which is not compatible with the main condition just mentioned. On the other hand, it is equally important for the model to be able to identify any new structural features of a query molecule that were not present in the training set compounds. I.e., the fixed fragment set cannot be constructed based on the analysis of the training set either, or in general any molecule set whatsoever. Because in that case any new structural features not present in that database would be eventually ignored. As a result, the fragmental descriptor set is based on the general knowledge and considerations regarding all possible chemical

structures rather than a finite dataset and include all the fragments, even those that are not detected in the training set molecules at all.

4.4.Descriptor selection:

No special descriptor selection techniques had been used to reduce the initial descriptor pool (e.g., excluding statistically insignificant or intercorrelated variables) due to the specifics of employed modeling methodology. Any potential negative influence of insignificant fragments would be remedied by the use of PLS method, but their presence is necessary for providing the so called "dynamic similarity" measure between the molecules. For this purpose, even “blank” fragments (with zero occurrence count) should remain, as these would allow detecting new structural features of a query molecule that were not present in the training set, and would thus decrease its similarity coefficient to training set molecules.

4.5.Algorithm and descriptor generation:

The generation of the descriptor matrix following the outlined approach constituted counting the occurences of any of the pre-defined fragments in the training set molecules. This procedure as well as all the subsequent statistical analysis were performed using Algorithm Builder 1.8 software.

4.6.Software name and version for descriptor generation:

Algorithm Builder 1.8

ACD/Labs, Inc. 110 Yonge Street, 14th floor, Toronto, Ontario, Canada M5C 1T4.

http://www.acdlabs.com

4.7.Chemicals/Descriptors ratio:

12.6 (4764 chemical in the training set, 379 descriptors).

5.Defining the applicability domain - OECD Principle 3

5.1.Description of the applicability domain of the model:

Applicability domain of the model is defined based on the training set compounds. This procedure takes into account the following two aspects:

* Similarity of the tested compound to the training set. No reliable predictions can be made if we have no similar compounds in the training set;

* Consistence of the experimental values with regard to the baseline model for similar compounds. Even if we do have similar compounds in the dataset the quality of prediction could be lower if that data cannot be reproduced by the baseline model. It does not matter what the reason for this inconsistency – experimental variability or sudden change in mechanism of action because of slight structural changes – in any case it indicates possible problems when trying to give accurate predictions.

5.2.Method used to assess the applicability domain:

The two aspects mentioned in Section 5.1 receive their quantitative assessment in terms of Similarity Index (SI) and Data-Model Consistency Index (DMCI). The SI, evaluating how distant the query structure is from the whole training set, is calculated by weighted averaging of all the

individual Similarity Indices (SI_i) for the test molecule and each of the 5 most similar compounds from the training set. DMCI is calculated by comparing the differences between experimental and global QSAR predicted values for the 5 most similar compounds and the suggested similarity correction value (*Delta*) for the test compound, calculated by averaging these differences. The more individual differences are scattered around the calculated average (*Delta*), the more inconsistent are the data for the similar compounds with regards to the global QSAR model.

The final prediction Reliability Index is calculated as a product of the aforementioned two indices:

RI = SI * DMCI

Both SI and DMCI are scaled to vary from 0 to 1, so the resulting RI also varies in this range. Lower values suggest a compound being further from the Model Applicability Domain and the prediction less reliable (low SI or low DMCI either alone or in combination can be the reason). On the other hand, high RI values indicate an increasing confidence about the quality of the prediction (both SI and DMCI have to be high to yield such a result).

5.3.Software name and version for applicability domain assessment:

ACD/Percepta
ACD/Labs, Inc. 110 Yonge Street, 14th floor, Toronto, Ontario, Canada M5C 1T4
http://acdlabs.com/products/percepta/

5.4.Limits of applicability:

Reliability Index < 0.3: NOT reliable predictions
Reliability Index in the range 0.3-0.5: borderline reliability of prediction
Reliability Index in the range 0.5-0.75: moderate reliability of prediction
Reliability Index >0.75: high reliability of prediction

6.Internal validation - OECD Principle 4

6.1.Availability of the training set:

Yes

6.2.Available information for the training set:

The training set is available through the software products listed here, but is not attached to the form itself
CAS RN: No
Chemical Name: No
Smiles: No
Formula: No
INChI: No
MOL file: No

6.3.Data for each descriptor variable for the training set:

No

6.4.Data for the dependent variable for the training set:

No

6.5.Other information about the training set:

The statistics of the training set data:

No. of compounds = 4764

Min. Value = -12.04

Max. Value = 2.00

Std. Dev. = 2.14

Skewness = -0.71

6.6.Pre-processing of data before modelling:

Inorganic compounds have been excluded.

6.7.Statistics for goodness-of-fit:

Statistics provided for the fraction of the training set that falls within the applicability domain of the model (RI > 0.3 - see Section 5.4)

N_{RI}>0.3 = 4651 (i.e. 97.6% of the training set compounds)

R^2 = 0.846

MAE = 0.622

RMSE = 0.842

F = 25489.1 (Fisher's F-statistics)

6.8.Robustness - Statistics obtained by leave-one-out cross-validation:

N/A

6.9.Robustness - Statistics obtained by leave-many-out cross-validation:

N/A

6.10.Robustness - Statistics obtained by Y-scrambling:

N/A

6.11.Robustness - Statistics obtained by bootstrap:

N/A

6.12.Robustness - Statistics obtained by other methods:

N/A

7.External validation - OECD Principle 4

7.1.Availability of the external validation set:

Yes

7.2.Available information for the external validation set:

The validation set is available through the software products listed here, but is not attached to the form itself

CAS RN: No

Chemical Name: No

Smiles: No

Formula: No

INChI: No

MOL file: No

7.3.Data for each descriptor variable for the external validation set:

No

7.4.Data for the dependent variable for the external validation set:

No

7.5.Other information about the external validation set:

The statistics of the validation set data:

No. of compounds = 2043

Min. Value = -12.79

Max. Value = 2.06

Std. Dev. = 2.11

Skewness = -0.65

7.6.Experimental design of test set:

Random splitting of the initial dataset into the training and validation sets at ~70%:30% ratio.

7.7.Predictivity - Statistics obtained by external validation:

Statistics provided for the fraction of the validation set that falls within the applicability domain of the model (RI > 0.3 - see Section 5.4)

$N_{RI>0.3}$ = 2002 (i.e. 98.0% of all the validation set compounds)

R^2 = 0.846; MAE = 0.611; RMSE = 0.829; F = 10962.0 (Fisher's F-statistics)

Analysis of the subsets of the higher quality results:

$N_{RI>0.5}$ = 1737 (i.e. 85.0% of all the validation set compounds)

R^2 = 0.867; MAE = 0.564; RMSE = 0.769; F = 11267.5 (Fisher's F-statistics)

$N_{RI>0.75}$ = 612 (i.e. 30.0% of all the validation set compounds)

R^2 = 0.927; MAE = 0.429; RMSE = 0.586;F = 7768.7 (Fisher's F-statistics)

7.8.Predictivity - Assessment of the external validation set:

As can be seen from the results of the Section 7.7, 98% of the validation set is within the Applicability Domain of the reported model, and more than 80% of the compounds obtain predictions of moderate and high reliability.

7.9.Comments on the external validation of the model:

Correlation coefficients and other statistical parameters for the training and validation set compounds falling within the applicability domain of the model are in a good agreement.

8.Providing a mechanistic interpretation - OECD Principle 5

8.1.Mechanistic basis of the model:

The only mechanistic consideration utilized in model building is the use of a linear regression method (PLS) and the fragmental descriptors. In other words it is assumed that the final predicted value

is composed of a linear combination of all the contributions of structural moieties making up the test molecule. Although very basic, this consideration is one of the most fundamental ones, even the name of (Q)SAR methods implies that the main determinant of all the properties of a compound is its structure. Quite obviously fragments are the best and really firsthand descriptors of a chemical structure.

8.2.A priori or a posteriori mechanistic interpretation:

A posteriori solubility model interpretation based on the coefficients of utilized fragmental descriptors is a bit more complicated compared to octanol-water partition coefficient models. In the latter case the distinction between lipophilicity increasing and reducing fragments is pretty straightforward. For example fragments with high content of electronegative atoms introducing polar bonds into the molecule and thus facilitating its interaction with water phase are universally treated as lipophilicity reducing factors. Whereas in case of solubility the same considerations sometimes can be misleading and some fragments with high electronegative atom content can actually end up exhibiting solubility reducing tendencies as well. It is especially likely to be true for larger fragments capable of representing possible multi-point interactions within the crystal lattice resulting in a tighter molecular packing and consequently lower solubility.

However majority of small heteroatom containing fragments (especially permanently charged or ionizable ones) can still be considered as solubility enhancing quite safely. And on the other end, the effect of carbon-rich non-polar substituents on solubility is as invariable as it is in case of log Kow - in this particular case it is always a solubility decreasing feature. In this regard the analysis of coefficients yields reasonably consistent results.

The top ten fragmental descriptors with positive coefficients are the following:

Phosphinic amide residue = 0.817; 2,6-dimethyl aromatic amine = 0.763; Activated aliphatic alcohol (string electron withdrawing group at alpha- position) = 0.751; Three-membered heterocycle = 0.741; Hydrazine fragment = 0.572; N or S mustard fragment = 0.541; Quaternary ammonium = 0.536; Any positive permanent charge = 0.531; Sulfonamide fragment = 0.504; Activated phenol (strong electron withdrawing group at meta-position) = 0.485.

Among the groups with the largest negative impact on solubility, the absolute majority of them can be clearly expected to be in this group just as it has been discussed at the beginning, e.g.:

Stereohindrance in the form of two bulk branched aliphatic substituents in both orto- positions of a phenolic group = -1.446; 1,3-dimethylcyclobutane scaffold = -0.953; n-Nonyl chain = -0.714; Fused 6:6:6 scaffold = -0.663; Carbonylguanidine fragment = -0.648; Imide = -0.608; 1,1,2-trimethylcyclopropane scaffold = -0.607; Steroid scaffold = -0.579; Fused 6:5:6 scaffold = -0.539; 1,2-difluoroethylene fragment = -0.538

Further similar examples can be established as well.Note: the average of all 379 statistical coefficients is -0.042

8.3.Other information about the mechanistic interpretation:

N/A

9.Miscellaneous information

9.1.Comments:

Together with the prediction, ACD/Percepta displays up to 5 most similar structures from the training set along with experimental results and references. The similarity is measured in terms of “property-specific” and structural similarity.

9.2.Bibliography:

[1]Japertas, P., Sazonovas, A., Didziapetris, R., and Petrauskas, A., Similarity based assessment of model applicability domain and quantitative evaluation of the reliability of the prediction. Abstr. Paper. Am. Chem. Soc., 2008 , 235, Meeting abstract 271-COMP
http://oasys2.confex.com/acs/235nm/techprogram/P1160866.HTM

[2]Japertas, P., Sazonovas, A., Didziapetris, R., and Petrauskas, A., Similarity based correction for the predictions of compounds physicochemical properties. Abstr. Paper. Am. Chem. Soc., 2008, 235, Meeting abstract 247 - MEDI
http://oasys2.confex.com/acs/235nm/techprogram/P1160811.HTM

[3]Japertas, P., Didziapetris, R., and Petrauskas, A., Fragmental methods in the design of new compounds. Applications of Advanced Algorithm Builder, Quant. Struct-Act. Relat., 2002; 21, 23-37
http://www3.interscience.wiley.com/journal/93521415/abstract

[4]The Merck Index. An Encyclopedia of Chemicals, Drugs, and Biologicals, O'Neil, M.J., Smith, A., Heckelman, P.E., Budavari, S., Eds. 13th Edition, Merck & Co., Inc., Whitehouse Station, NJ, 2001

[5]Therapeutic Drugs, Dolery, C., Ed. 2nd Edition, Churchill Livingstone, New York, NY, 1999

[6]Clarke's Isolation and Identification of Drugs, Moffat, A.C., Jackson, J.V., Moss, M.S., Widdop, B., Eds. 2nd Edition, The Pharmaceutical Press, London, 1986

[7]Avdeev, A., Absorption and Drug Development: Solubility, Permeability, and Charge State, John Wiley & Sons, Inc., Hoboken, NJ, 2003.

[8]Kishi, H. and Hashimoto, Y., Evaluation of the procedures for the measurement of water solubility and n-octanol/water partition coefficient of chemicals results of a ring test in Japan, Chemosphere, 1989, 18, 1749-1759.

[9]Liu, R. and So, S. S., Development of quantitative structure-property relationship models for early ADME evaluation in drug discovery. 1. Aqueous solubility. J. Chem. Inf. Comput. Sci., 2001, 41, 1633-1639.

[10]I.V. Tetko, Neural network studies. 4. Introduction to associative neural networks, J. Chem. Inf. Comput. Sci. 2002, 42, 717-728.

[11]I.V. Tetko and P. Bruneau, Application of ALOGPS to predict 1-octanol/water distribution coefficients, logP, and logD, of AstraZeneca in-house database, J. Pharm. Sci. 2004, 93, 3103-3110.

[12]I.V. Tetko and V.Y. Tanchuk, Application of associative neural networks for prediction of lipophilicity in ALOGPS 2.1 program, J. Chem. Inf. Comput. Sci. 2002, 42, 1136-1145.

[13]H. Zhu, A. Tropsha, D. Fourches, A. Varnek, E. Papa, P. Gramatica, T. Oberg, P. Dao, A. Cherkasov, and I.V. Tetko, Combinatorial QSAR modeling of chemical toxicants tested against Tetrahymena pyriformis, J. Chem. Inf. Model. 2008, 48, 766-784.

[14]I.V. Tetko, I. Sushko, A.K. Pandey, H. Zhu, A. Tropsha, E. Papa, T. Oberg, R. Todeschini, D. Fourches, and A. Varnek, Critical assessment of QSAR models of environmental toxicity against Tetrahymena pyriformis: focusing on applicability domain and overfitting by variable selection, J. Chem. Inf. Model. 2008, 48, 1733-1746.

[15]Sazonovas, A., Japertas, P., and Didziapetris, R., Estimation of reliability of predictions and model applicability domain evaluation in the analysis of acute toxicity (LD50), SAR QSAR Environ. Res. 2010, 21, 127-148.

[16]Didziapetris, R., Dapkunas, J., Sazonovas, A., and Japertas, P., Trainable structure–activity relationship model for virtual screening of CYP3A4 inhibition, J. Comput. Aided Mol. Des. 2010, in press, DOI: 10.1007/s10822-010-9381-1.

[17]J.A. Platts, D. Butina, M.H. Abraham, and A. Hersey, Estimation of molecular linear free energy relation descriptors using a group contribution approach, J. Chem. Inf. Comput. Sci. 1999, 39, 835-845.

9.3.Supporting information:

Training set(s)Test set(s)Supporting information

10.Summary (JRC Inventory)

10.1.QMRF number:

To be entered by JRC

10.2.Publication date:

To be entered by JRC

10.3.Keywords:

To be entered by JRC

10.4.Comments:

To be entered by JRC