CASE STUDY: QSAR calculation of octanol-water partition coefficient log Kow

31/05/2023, LIFE CONCERT REACH Web-Seminars - (Q)SAR Models under REACH: Practical Examples

Andrzej Szymoszek, Ph.D.
knoell Germany GmbH
aszymoszek@knoell.com

CONCERTREACH
CONCERTING EXPERIMENTAL DATA
AND IN SILICO MODELS FOR REACH

LIFE17 GIE/IT/000461

TABLE OF CONTENTS

01 **Introduction on models and current case study**

02 Running models and analysis of results

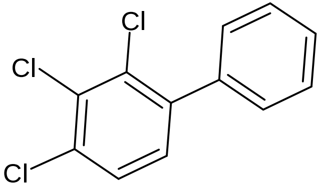
03 Documentation of QSAR results in IUCLID for REACH dossier preparation

04 Summary and conclusions

Case study description

Aim: Prediction of octanol-water partition coefficient log Kow using **VEGA QSAR models** and **documentation in IUCLID**

Target molecule: 2,3,4-Trichlorobiphenyl

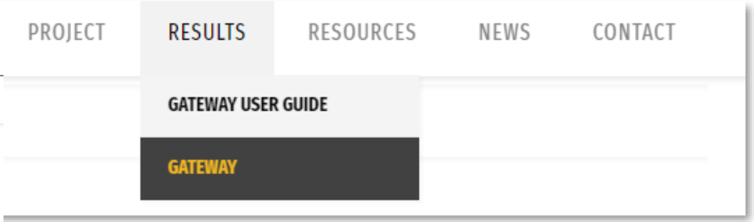


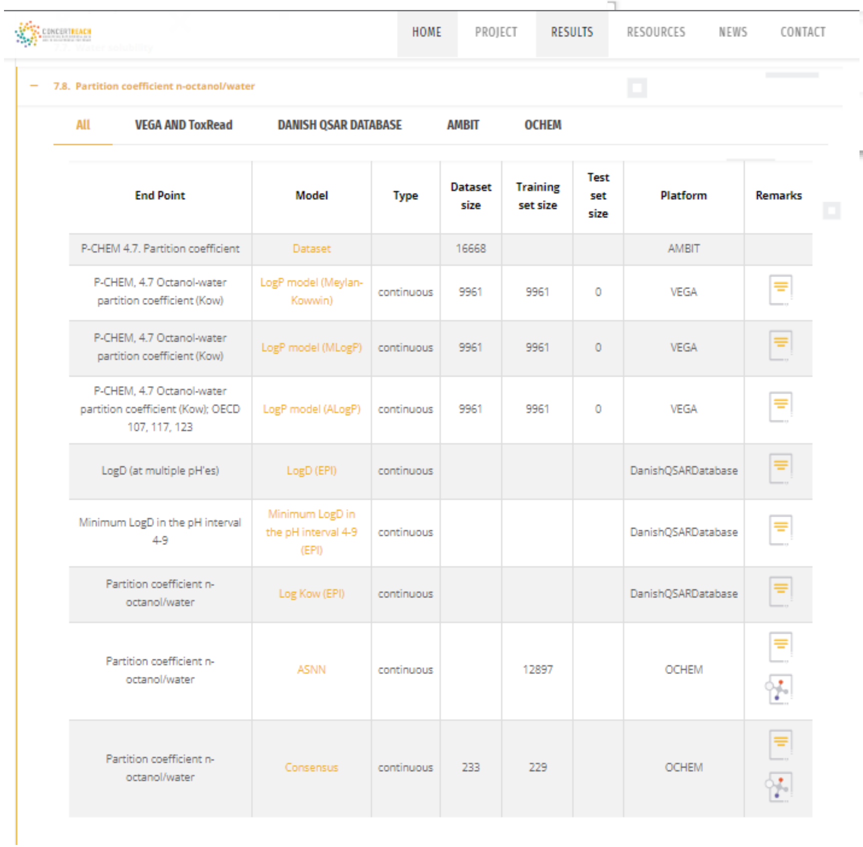
Models: VEGA – Meylan/KOWWIN, MLogP, ALogP. QSAR Model Reporting Formats (QMRFs) are available.

Input data: SMILES notation - c1ccc(cc1)c2ccc(c(c2Cl)Cl)Cl

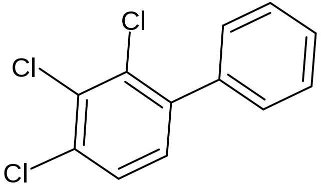
Models for log Kow in the CONCERT REACH gateway

https://www.life-concertreach.eu/results/results-gateway/



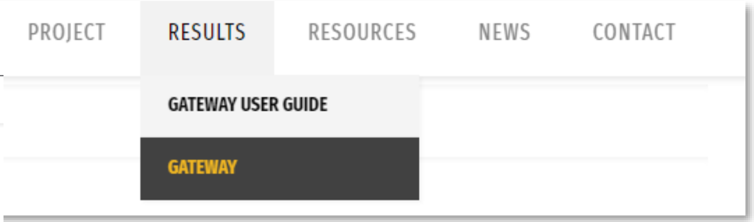


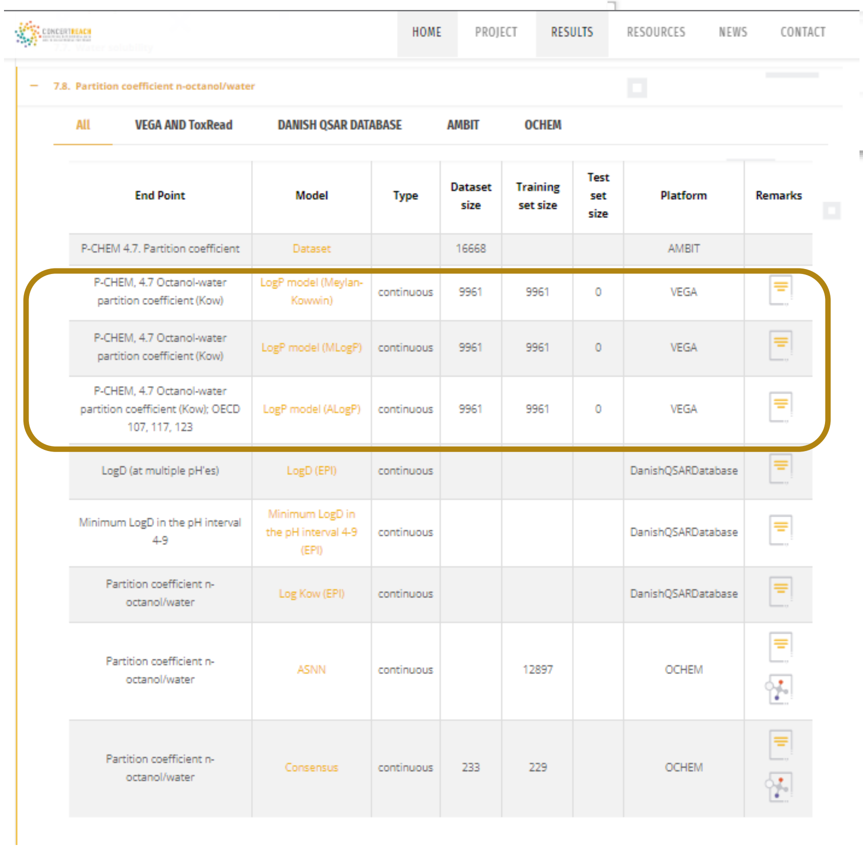
End Point	Model	Type	Dataset size	Training set size	Test set size	Platform	Remarks
P-CHEM 4.7. Partition coefficient	Dataset		16668			AMBIT	
P-CHEM, 4.7 Octanol-water partition coefficient (Kow)	LogP model (Meylan-Kowwin)	continuous	9961	9961	0	VEGA	☰
P-CHEM, 4.7 Octanol-water partition coefficient (Kow)	LogP model (MLogP)	continuous	9961	9961	0	VEGA	☰
P-CHEM, 4.7 Octanol-water partition coefficient (Kow); OECD 107, 117, 123	LogP model (ALogP)	continuous	9961	9961	0	VEGA	☰
LogD (at multiple pH'es)	LogD (EPI)	continuous				DanishQSARDatabase	☰
Minimum LogD in the pH interval 4-9	Minimum LogD in the pH interval 4-9 (EPI)	continuous				DanishQSARDatabase	☰
Partition coefficient n-octanol/water	Log Kow (EPI)	continuous				DanishQSARDatabase	☰
Partition coefficient n-octanol/water	ASNN	continuous		12897		OCHEM	☰
Partition coefficient n-octanol/water	Consensus	continuous	233	229		OCHEM	☰



Models for log Kow in the CONCERT REACH gateway

https://www.life-concertreach.eu/results/results-gateway/

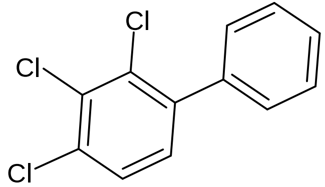




– 7.8. Partition coefficient n-octanol/water

All | VEGA AND ToxRead | DANISH QSAR DATABASE | AMBIT | OCHEM

End Point	Model	Type	Dataset size	Training set size	Test set size	Platform	Remarks
P-CHEM 4.7. Partition coefficient	Dataset		16668			AMBIT	
P-CHEM, 4.7 Octanol-water partition coefficient (Kow)	LogP model (Meylan-Kowwin)	continuous	9961	9961	0	VEGA	
P-CHEM, 4.7 Octanol-water partition coefficient (Kow)	LogP model (MLogP)	continuous	9961	9961	0	VEGA	
P-CHEM, 4.7 Octanol-water partition coefficient (Kow); OECD 107, 117, 123	LogP model (ALogP)	continuous	9961	9961	0	VEGA	
LogD (at multiple pH'es)	LogD (EPI)	continuous				DanishQSARDatabase	
Minimum LogD in the pH interval 4-9	Minimum LogD in the pH interval 4-9 (EPI)	continuous				DanishQSARDatabase	
Partition coefficient n-octanol/water	Log Kow (EPI)	continuous				DanishQSARDatabase	
Partition coefficient n-octanol/water	ASNN	continuous		12897		OCHEM	
Partition coefficient n-octanol/water	Consensus	continuous	233	229		OCHEM	



VEGA Models for log Kow

- Meylan/KOWWIN v1.1.5: VEGA implementation of EPISUITE KOWWIN. Regression equation is based on the hydrophobicity contribution of 120 atom types.It is an implementation of the atom fragment contribution (AFC) method described by Meylan et al., 1995. It is a “reductionist” approach and it was developed via multiple linear regressions of reliable, experimental log P values.

- MLogP v1.0.1: VEGA implementation of the multiple linear regression developed by Moriguchi et al. (1992; 1995) that relates 13 structural parameters with the experimental log P values of 1230 compounds with different structures

- ALogP v1.0.1: VEGA implementation of the Ghose-Crippen-Viswanadhan regression equation based on the hydrophobicity contribution of 120 atom types.

TABLE OF CONTENTS

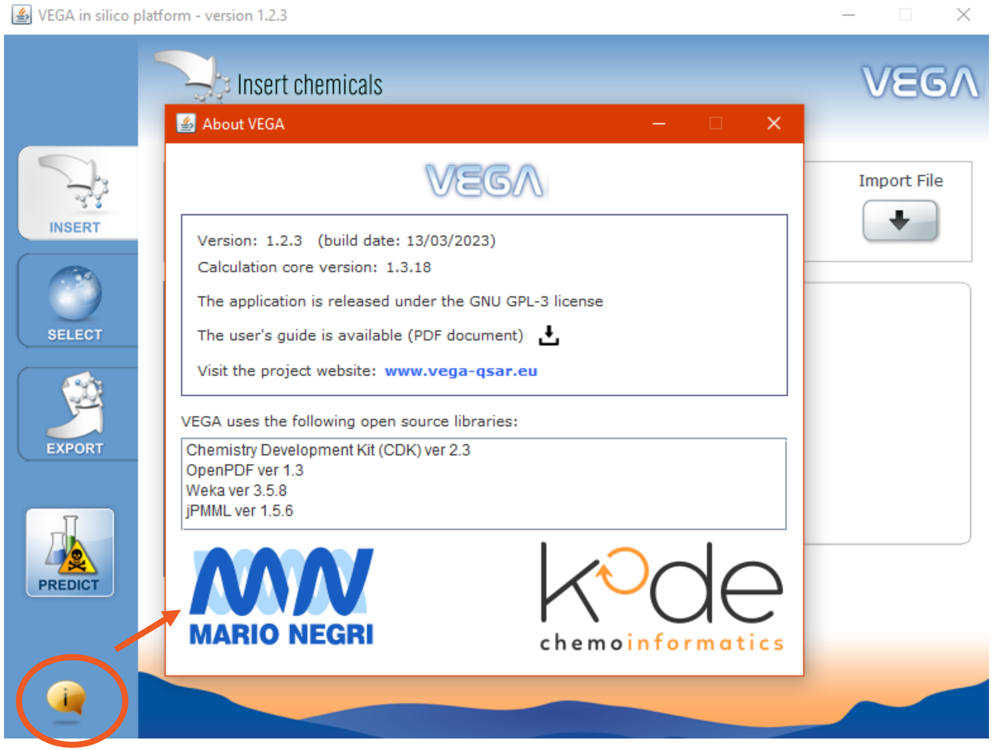
01 Introduction on models and current case study

02 **Running models and analysis of results**

03 Documentation of QSAR results in IUCLID for REACH dossier preparation

04 Summary and conclusions

VEGA: introduction



Version: 1.2.3 (build date: 13/03/2023)

Calculation core version: 1.3.18

The application is released under the GNU GPL-3 license

The user's guide is available (PDF document)

Visit the project website: www.vega-qsar.eu

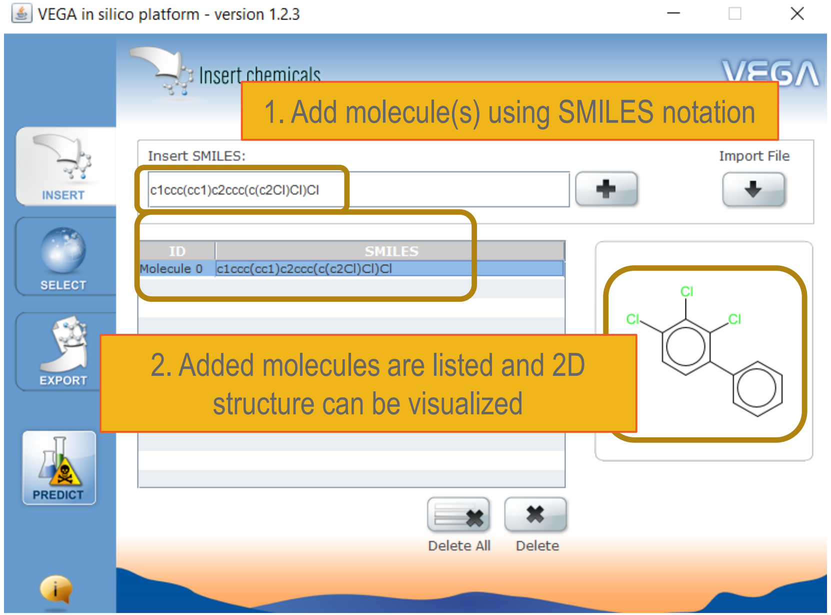
VEGA uses the following open source libraries:

Chemistry Development Kit (CDK) ver 2.3
OpenPDF ver 1.3
Weka ver 3.5.8
jPMML ver 1.5.6

VEGA: Virtual models for Evaluating the properties of chemicals within a Global Architecture

- Developed mainly by Mario Negri Institute (Milan) and Kode s.r.l. (Pisa)
- **Free platform** developed based on contributions from EU projects
- Includes **more than 100 statistical and knowledge-based (Q)SAR models** for the prediction of (eco)toxicity, environmental fate and physico-chemical properties of chemicals.

VEGA: running predictions



VEGA in silico platform - version 1.2.3

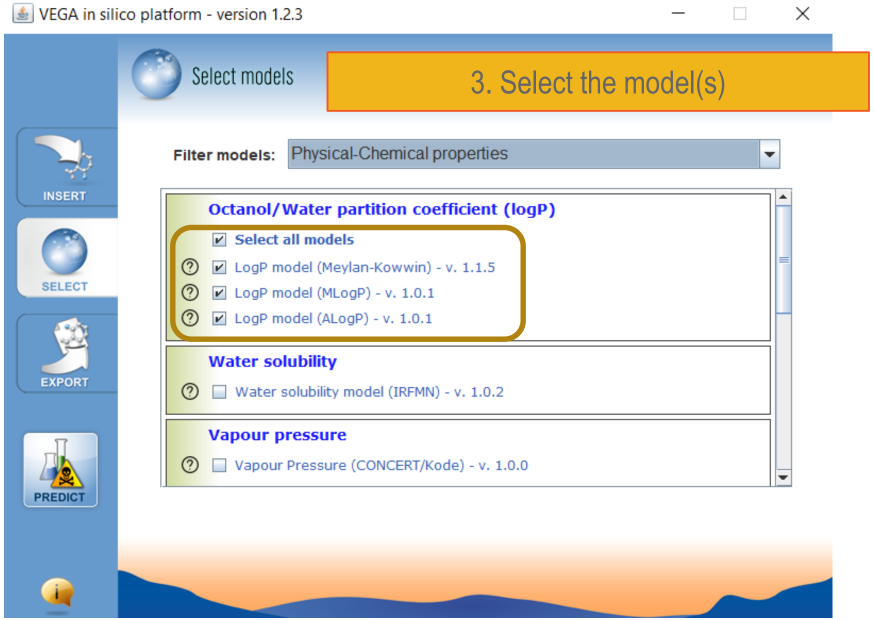
Insert chemicals

1. Add molecule(s) using SMILES notation

Insert SMILES: c1ccc(cc1)c2ccc(c(c2Cl)Cl)Cl

ID	SMILES
Molecule 0	c1ccc(cc1)c2ccc(c(c2Cl)Cl)Cl

2. Added molecules are listed and 2D structure can be visualized



VEGA in silico platform - version 1.2.3

Select models

3. Select the model(s)

Filter models: Physical-Chemical properties

Octanol/Water partition coefficient (logP)

- ☑ Select all models
- ☑ LogP model (Meylan-Kowwin) - v. 1.1.5
- ☑ LogP model (MLogP) - v. 1.0.1
- ☑ LogP model (ALogP) - v. 1.0.1

Water solubility

- ☐ Water solubility model (IRFMN) - v. 1.0.2

Vapour pressure

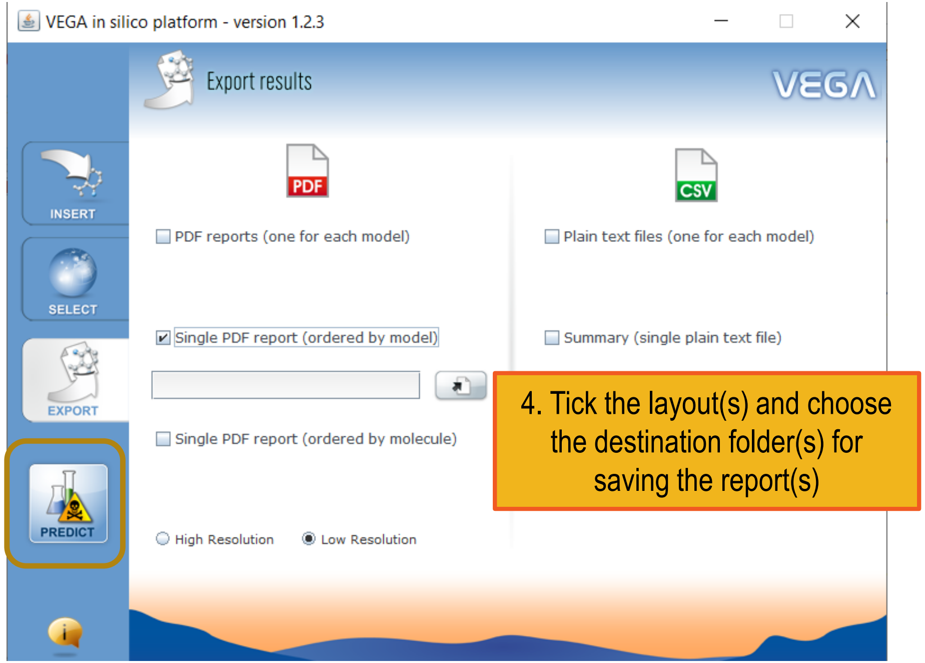
- ☐ Vapour Pressure (CONCERT/Kode) - v. 1.0.0

VEGA: running predictions

Full PDF reports:

- prediction(s) results
- applicability domain
- experimental data of the target (if any)
- most similar substances
- other supporting info (if any)

5. Click on «Predict»



☐ PDF reports (one for each model)

☐ Plain text files (one for each model)

☑ Single PDF report (ordered by model)

☐ Summary (single plain text file)

☐ Single PDF report (ordered by molecule)

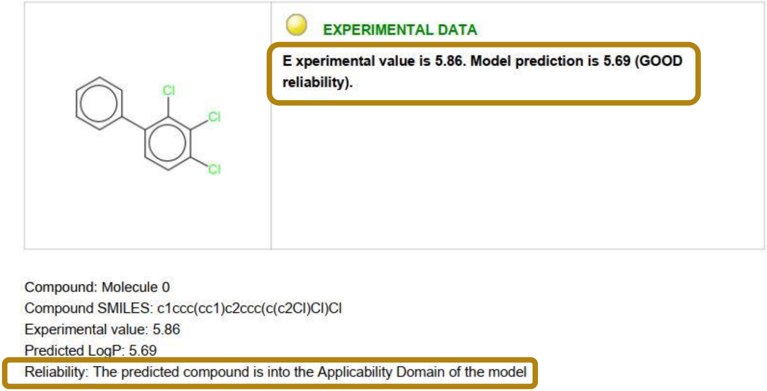
○ High Resolution ● Low Resolution

4. Tick the layout(s) and choose the destination folder(s) for saving the report(s)

Simplified text reports

(useful for excel import)

Determination of log Kow: Meylan/KOWWIN



EXPERIMENTAL DATA

E xperimental value is 5.86. Model prediction is 5.69 (GOOD reliability).

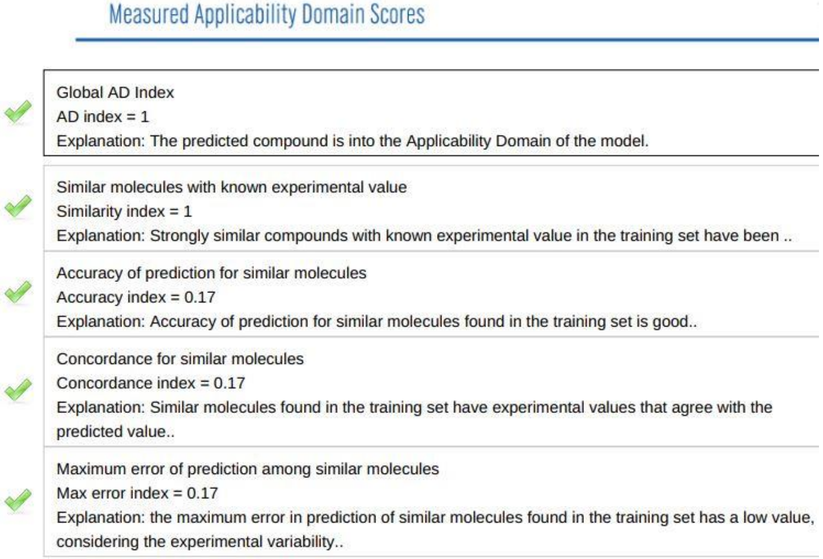
Compound: Molecule 0

Compound SMILES: c1ccc(cc1)c2ccc(c(c2Cl)Cl)Cl

Experimental value: 5.86

Predicted LogP: 5.69

Reliability: The predicted compound is into the Applicability Domain of the model



Measured Applicability Domain Scores

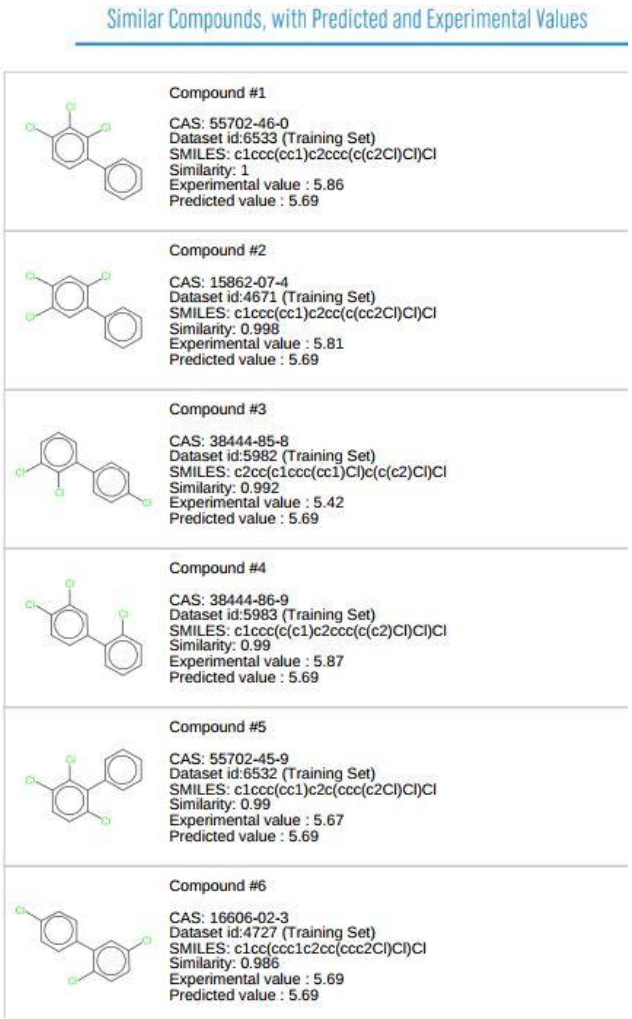
✓ Global AD Index
AD index = 1
Explanation: The predicted compound is into the Applicability Domain of the model.

✓ Similar molecules with known experimental value
Similarity index = 1
Explanation: Strongly similar compounds with known experimental value in the training set have been ..

✓ Accuracy of prediction for similar molecules
Accuracy index = 0.17
Explanation: Accuracy of prediction for similar molecules found in the training set is good..

✓ Concordance for similar molecules
Concordance index = 0.17
Explanation: Similar molecules found in the training set have experimental values that agree with the predicted value..

✓ Maximum error of prediction among similar molecules
Max error index = 0.17
Explanation: the maximum error in prediction of similar molecules found in the training set has a low value, considering the experimental variability..



Similar Compounds, with Predicted and Experimental Values

Compound #1

CAS: 55702-46-0
Dataset id:6533 (Training Set)
SMILES: c1ccc(cc1)c2ccc(c(c2Cl)Cl)Cl
Similarity: 1
Experimental value : 5.86
Predicted value : 5.69

Compound #2

CAS: 15862-07-4
Dataset id:4671 (Training Set)
SMILES: c1ccc(cc1)c2cc(c(cc2Cl)Cl)Cl
Similarity: 0.998
Experimental value : 5.81
Predicted value : 5.69

Compound #3

CAS: 38444-85-8
Dataset id:5982 (Training Set)
SMILES: c2cc(c1ccc(cc1)Cl)c(c(c2)Cl)Cl
Similarity: 0.992
Experimental value : 5.42
Predicted value : 5.69

Compound #4

CAS: 38444-86-9
Dataset id:5983 (Training Set)
SMILES: c1ccc(c(c1)c2ccc(c(c2)Cl)Cl)Cl
Similarity: 0.99
Experimental value : 5.87
Predicted value : 5.69

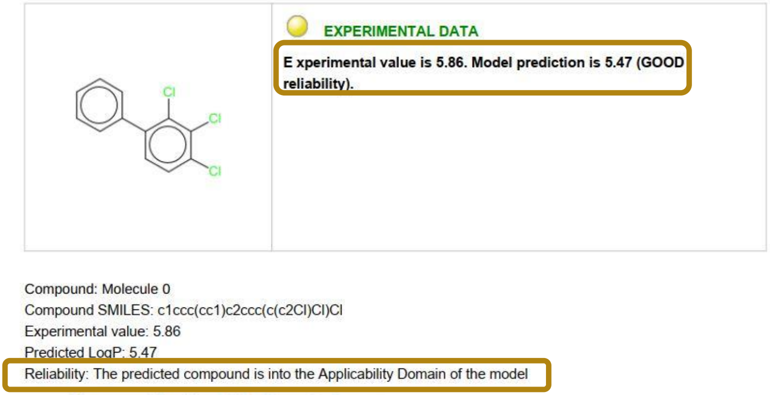
Compound #5

CAS: 55702-45-9
Dataset id:6532 (Training Set)
SMILES: c1ccc(cc1)c2c(ccc(c2Cl)Cl)Cl
Similarity: 0.99
Experimental value : 5.67
Predicted value : 5.69

Compound #6

CAS: 16606-02-3
Dataset id:4727 (Training Set)
SMILES: c1cc(ccc1c2cc(ccc2Cl)Cl)Cl
Similarity: 0.986
Experimental value : 5.69
Predicted value : 5.69

Determination of log Kow: MLogP



EXPERIMENTAL DATA

E xperimental value is 5.86. Model prediction is 5.47 (GOOD reliability).

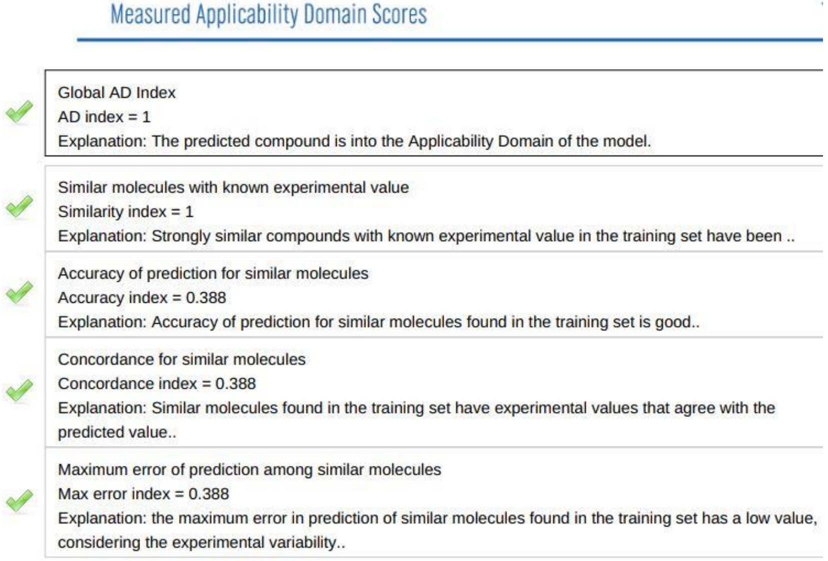
Compound: Molecule 0

Compound SMILES: c1ccc(cc1)c2ccc(c(c2Cl)Cl)Cl

Experimental value: 5.86

Predicted LogP: 5.47

Reliability: The predicted compound is into the Applicability Domain of the model



Measured Applicability Domain Scores

✓ Global AD Index

AD index = 1

Explanation: The predicted compound is into the Applicability Domain of the model.

✓ Similar molecules with known experimental value

Similarity index = 1

Explanation: Strongly similar compounds with known experimental value in the training set have been ..

✓ Accuracy of prediction for similar molecules

Accuracy index = 0.388

Explanation: Accuracy of prediction for similar molecules found in the training set is good..

✓ Concordance for similar molecules

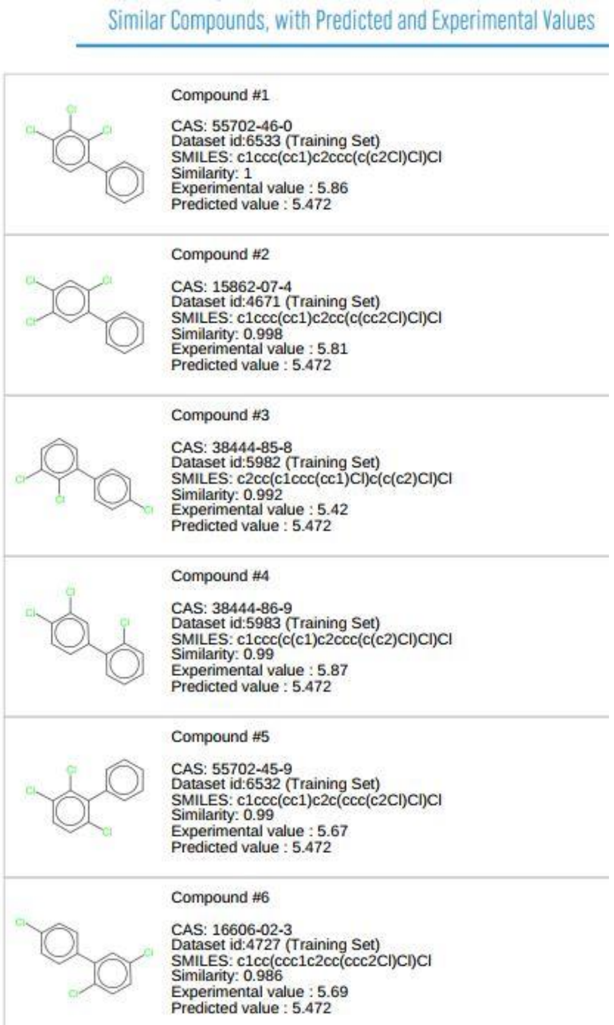
Concordance index = 0.388

Explanation: Similar molecules found in the training set have experimental values that agree with the predicted value..

✓ Maximum error of prediction among similar molecules

Max error index = 0.388

Explanation: the maximum error in prediction of similar molecules found in the training set has a low value, considering the experimental variability..



Similar Compounds, with Predicted and Experimental Values

Compound #1

CAS: 55702-46-0
Dataset id:6533 (Training Set)
SMILES: c1ccc(cc1)c2ccc(c(c2Cl)Cl)Cl
Similarity: 1
Experimental value : 5.86
Predicted value : 5.472

Compound #2

CAS: 15862-07-4
Dataset id:4671 (Training Set)
SMILES: c1ccc(cc1)c2cc(c(cc2Cl)Cl)Cl
Similarity: 0.998
Experimental value : 5.81
Predicted value : 5.472

Compound #3

CAS: 38444-85-8
Dataset id:5982 (Training Set)
SMILES: c2cc(c1ccc(cc1)Cl)c(c(c2)Cl)Cl
Similarity: 0.992
Experimental value : 5.42
Predicted value : 5.472

Compound #4

CAS: 38444-86-9
Dataset id:5983 (Training Set)
SMILES: c1ccc(c(c1)c2ccc(c(c2)Cl)Cl)Cl
Similarity: 0.99
Experimental value : 5.87
Predicted value : 5.472

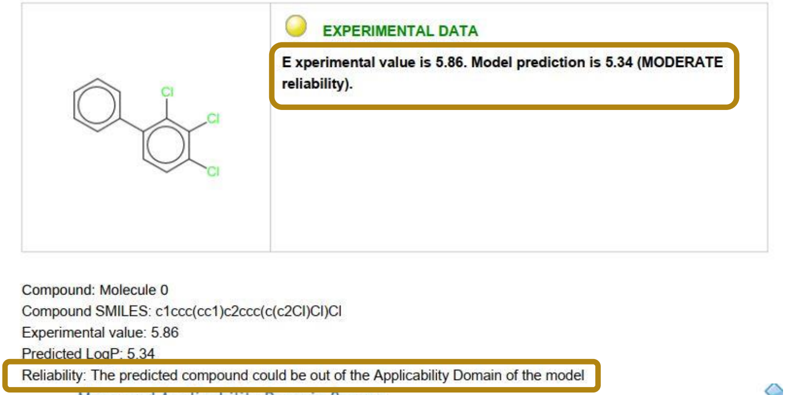
Compound #5

CAS: 55702-45-9
Dataset id:6532 (Training Set)
SMILES: c1ccc(cc1)c2c(ccc(c2Cl)Cl)Cl
Similarity: 0.99
Experimental value : 5.67
Predicted value : 5.472

Compound #6

CAS: 16606-02-3
Dataset id:4727 (Training Set)
SMILES: c1cc(ccc1c2cc(ccc2Cl)Cl)Cl
Similarity: 0.986
Experimental value : 5.69
Predicted value : 5.472

Determination of log Kow: ALogP



EXPERIMENTAL DATA

E xperimental value is 5.86. Model prediction is 5.34 (MODERATE reliability).

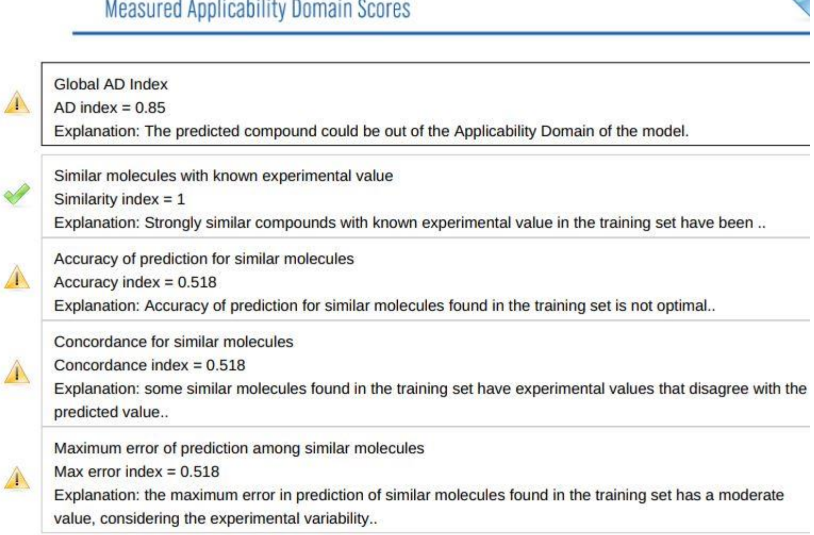
Compound: Molecule 0

Compound SMILES: c1ccc(cc1)c2ccc(c(c2Cl)Cl)Cl

Experimental value: 5.86

Predicted LogP: 5.34

Reliability: The predicted compound could be out of the Applicability Domain of the model



Measured Applicability Domain Scores

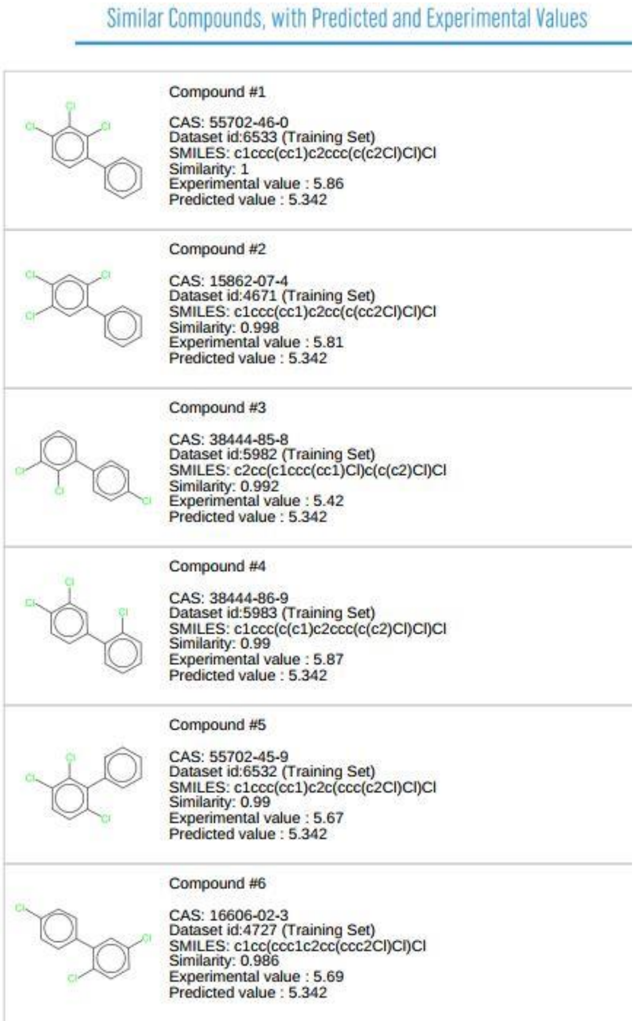
⚠ Global AD Index
AD index = 0.85
Explanation: The predicted compound could be out of the Applicability Domain of the model.

✓ Similar molecules with known experimental value
Similarity index = 1
Explanation: Strongly similar compounds with known experimental value in the training set have been ..

⚠ Accuracy of prediction for similar molecules
Accuracy index = 0.518
Explanation: Accuracy of prediction for similar molecules found in the training set is not optimal..

⚠ Concordance for similar molecules
Concordance index = 0.518
Explanation: some similar molecules found in the training set have experimental values that disagree with the predicted value..

⚠ Maximum error of prediction among similar molecules
Max error index = 0.518
Explanation: the maximum error in prediction of similar molecules found in the training set has a moderate value, considering the experimental variability..



Similar Compounds, with Predicted and Experimental Values

Compound #1

CAS: 55702-46-0
Dataset id:6533 (Training Set)
SMILES: c1ccc(cc1)c2ccc(c(c2Cl)Cl)Cl
Similarity: 1
Experimental value : 5.86
Predicted value : 5.342

Compound #2

CAS: 15862-07-4
Dataset id:4671 (Training Set)
SMILES: c1ccc(cc1)c2cc(c(cc2Cl)Cl)Cl
Similarity: 0.998
Experimental value : 5.81
Predicted value : 5.342

Compound #3

CAS: 38444-85-8
Dataset id:5982 (Training Set)
SMILES: c2cc(c1ccc(cc1)Cl)c(c(c2)Cl)Cl
Similarity: 0.992
Experimental value : 5.42
Predicted value : 5.342

Compound #4

CAS: 38444-86-9
Dataset id:5983 (Training Set)
SMILES: c1ccc(c(c1)c2ccc(c(c2)Cl)Cl)Cl
Similarity: 0.99
Experimental value : 5.87
Predicted value : 5.342

Compound #5

CAS: 55702-45-9
Dataset id:6532 (Training Set)
SMILES: c1ccc(cc1)c2c(ccc(c2Cl)Cl)Cl
Similarity: 0.99
Experimental value : 5.67
Predicted value : 5.342

Compound #6

CAS: 16606-02-3
Dataset id:4727 (Training Set)
SMILES: c1cc(ccc1c2cc(ccc2Cl)Cl)Cl
Similarity: 0.986
Experimental value : 5.69
Predicted value : 5.342

Determination of log Kow: summary of VEGA results

Model	Meylan/KOWWIN	MLogP	ALogP
Predicted log Kow	5.69	5.47	5.34
Deviation from experimental value 5.86	0.17	0.39	0.52
Applicability domain compliance	In	In	Could be out
Performance on 6 most similar molecules	6x good	6x good	4x good 2x moderate

The better the compliance with the model applicability domain, the more precise the result.

VEGA: important remarks

- Full documentation of all models is available, as a **QMRF**

- Supporting information (**AD compliance, similar molecules**) is provided, allowing expert evaluation

Relevant for REACH dossier preparation in IUCLID

- AD compliance is affected by identified similar molecules from the training or validation set

- Automated AD compliance check is not perfect, user expert critical check is helpful

  ➢ This affects other tools as well, including commercial ones

A novel tool called VERA has been developed, aiming also at improving similarity evaluation and AD compliance check (Presentation 17.05.)

TABLE OF CONTENTS

01 Introduction on models and current case study

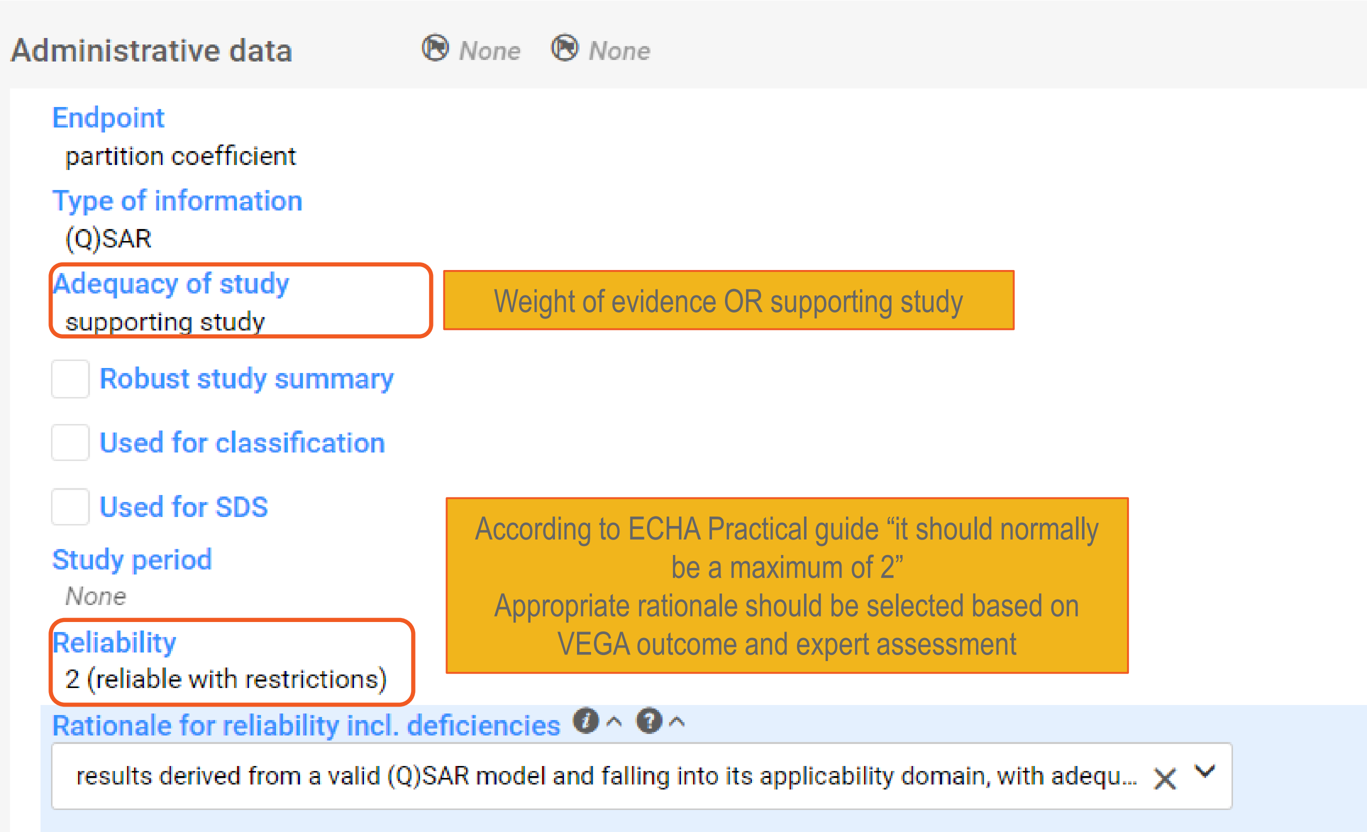
02 Running models and analysis of results

03 **Documentation of QSAR results in IUCLID for REACH dossier preparation**

04 Summary and conclusions

QSAR results in IUCLID

VEGA (example: Meylan/KOWWIN) outcome reported according to ECHA Practical guide “How to use and report (Q)SARs” Version 3.1 – July 2016



Administrative data None None

Endpoint
partition coefficient

Type of information
(Q)SAR

Adequacy of study
supporting study

□ Robust study summary

□ Used for classification

□ Used for SDS

Study period
None

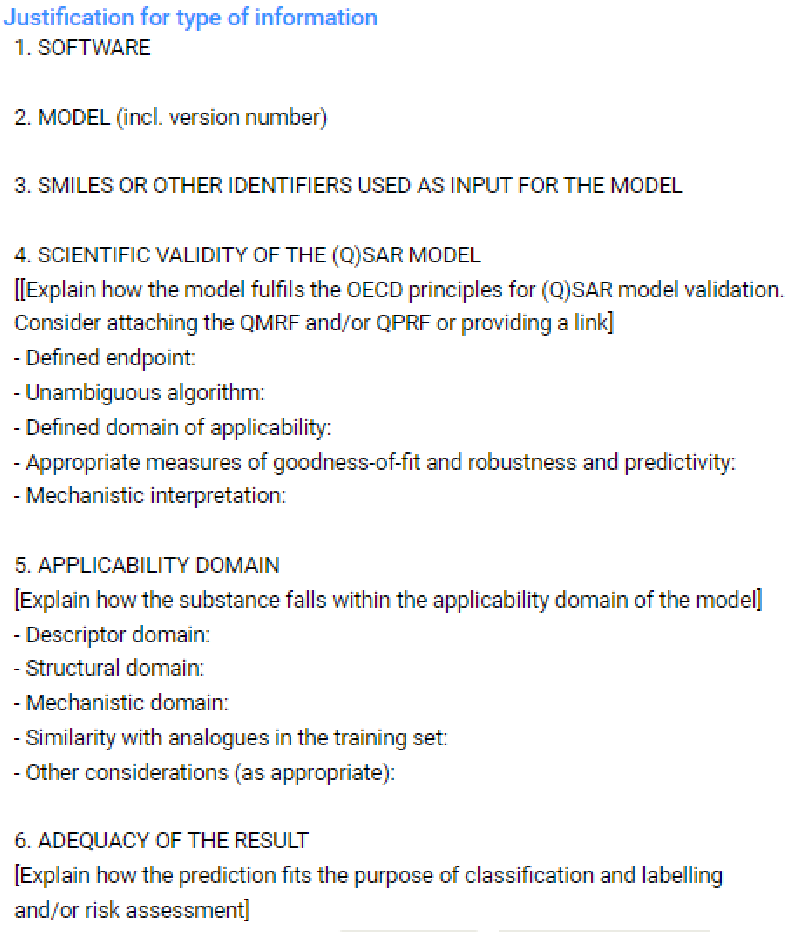
Reliability
2 (reliable with restrictions)

Rationale for reliability incl. deficiencies
results derived from a valid (Q)SAR model and falling into its applicability domain, with adequ...

Weight of evidence OR supporting study

According to ECHA Practical guide “it should normally be a maximum of 2”
Appropriate rationale should be selected based on VEGA outcome and expert assessment

(Q)SAR results in IUCLID



Justification for type of information

1. SOFTWARE

2. MODEL (incl. version number)

3. SMILES OR OTHER IDENTIFIERS USED AS INPUT FOR THE MODEL

4. SCIENTIFIC VALIDITY OF THE (Q)SAR MODEL
[[Explain how the model fulfils the OECD principles for (Q)SAR model validation. Consider attaching the QMRF and/or QPRF or providing a link]
- Defined endpoint:
- Unambiguous algorithm:
- Defined domain of applicability:
- Appropriate measures of goodness-of-fit and robustness and predictivity:
- Mechanistic interpretation:

5. APPLICABILITY DOMAIN
[Explain how the substance falls within the applicability domain of the model]
- Descriptor domain:
- Structural domain:
- Mechanistic domain:
- Similarity with analogues in the training set:
- Other considerations (as appropriate):

6. ADEQUACY OF THE RESULT
[Explain how the prediction fits the purpose of classification and labelling and/or risk assessment]

VEGA v1.2.3

Log P model (Meylan/Kowwin) v1.1.5

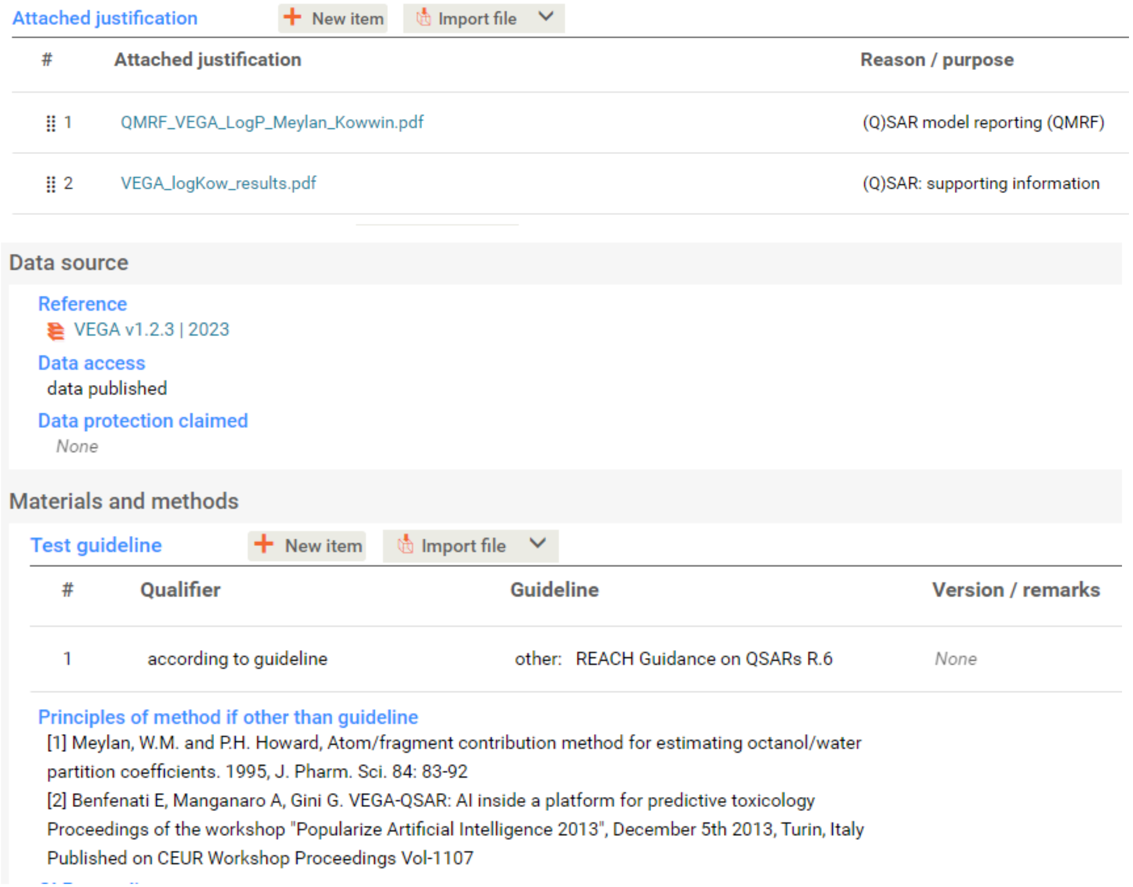
c1ccc(cc1)c2ccc(c(c2Cl)Cl)Cl

QMRF can be attached (next slide) and referenced here

VEGA report can be attached and used as reference. However, if expert assessment is performed, it can be described here.

Expert assessment is needed

(Q)SAR results in IUCLID



Attached justification

#	Attached justification	Reason / purpose
1	QMRF_VEGA_LogP_Meylan_Kowwin.pdf	(Q)SAR model reporting (QMRF)
2	VEGA_logKow_results.pdf	(Q)SAR: supporting information

Data source

Reference
VEGA v1.2.3 | 2023

Data access
data published

Data protection claimed
None

Materials and methods

Test guideline

#	Qualifier	Guideline	Version / remarks
1	according to guideline	other: REACH Guidance on QSARs R.6	None

Principles of method if other than guideline
[1] Meylan, W.M. and P.H. Howard, Atom/fragment contribution method for estimating octanol/water partition coefficients. 1995, J. Pharm. Sci. 84: 83-92
[2] Benfenati E, Manganaro A, Gini G. VEGA-QSAR: AI inside a platform for predictive toxicology Proceedings of the workshop "Popularize Artificial Intelligence 2013", December 5th 2013, Turin, Italy Published on CEUR Workshop Proceedings Vol-1107

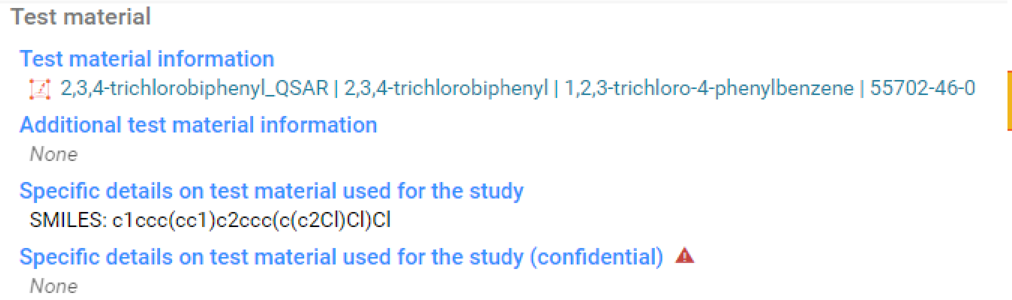
QPRF can also be attached, if prepared by the user

Basic information about the software and model are sufficient

Otherwise, the test guidelines used to generate the data for the training set

Information from QMRF section 2.7 - Reference(s) to main scientific papers and/or software package

(Q)SAR results in IUCLID



Test material

Test material information

2,3,4-trichlorobiphenyl_QSAR | 2,3,4-trichlorobiphenyl | 1,2,3-trichloro-4-phenylbenzene | 55702-46-0

Additional test material information

None

Specific details on test material used for the study

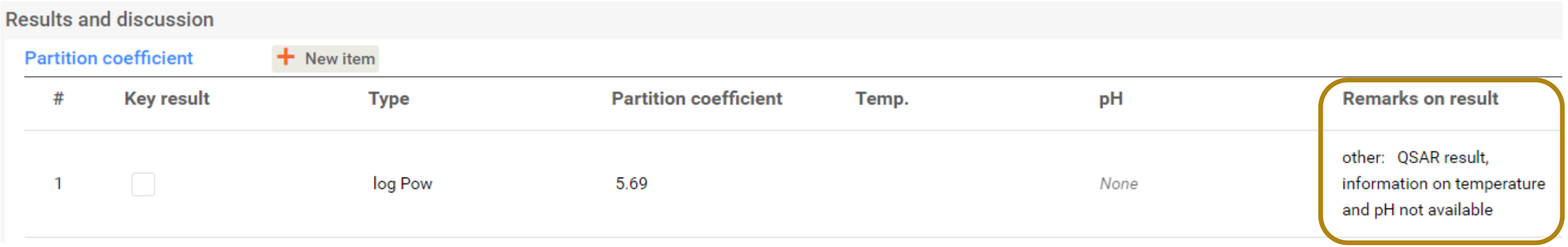
SMILES: c1ccc(cc1)c2ccc(c(c2Cl)Cl)Cl

Specific details on test material used for the study (confidential) ⚠

None

Test material must reflect the evaluated structure

If multiple constituents are assessed for one substance, the Practical Guide suggests preparation of separate entries



Results and discussion

Partition coefficient + New item

#	Key result	Type	Partition coefficient	Temp.	pH	Remarks on result
1	□	log Pow	5.69		*None*	other: QSAR result, information on temperature and pH not available

CONCERTREACH
CONCERTING EXPERIMENTAL DATA AND IN SILICO MODELS FOR REACH

LIFE17 GIE/IT/000461

TABLE OF CONTENTS

01 Introduction on models and current case study

02 Running models and analysis of results

03 Documentation of QSAR results in IUCLID for REACH dossier preparation

04 **Summary and conclusions**

Summary

- https://www.life-concertreach.eu/results/results-gateway/ The CONCERT REACH gateway is available; QSAR predictions are possible for REACH purposes

- QSAR prediction of log Kow using 3 VEGA models was presented and evaluated

- Preparation of a QSAR IUCLID entry for log Kow was shown, focusing on critical fields

Conclusions

- Applicability domain compliance is the most important factor which should be taken into account when evaluating the reliability of the QSAR results

The predictions may be used in the context of REACH:

- To cover the endpoint fully
- Together with other information (e.g. experimental data) as supporting data or part of WoE

CREDITS

Acknowledgement:

- The knoell Academy team
- Prof. Emilio Benfenati and the team, Mario Negri Institute, Milan
- The QSAR team at knoell
- The speakers of today and of 17 May
- The partners of the LIFE CONCERT REACH project

CONCERTREACH
CONCERTING EXPERIMENTAL DATA AND IN SILICO MODELS FOR REACH

LIFE17 GIE/IT/000461

ISTITUTO DI RICERCHE FARMACOLOGICHE MARIO NEGRI · IRCCS

big chem

kode from data to knowledge

knoell

DTU

Sviluppo chimica

Thanks for your attention

Questions?

Models for log Kow

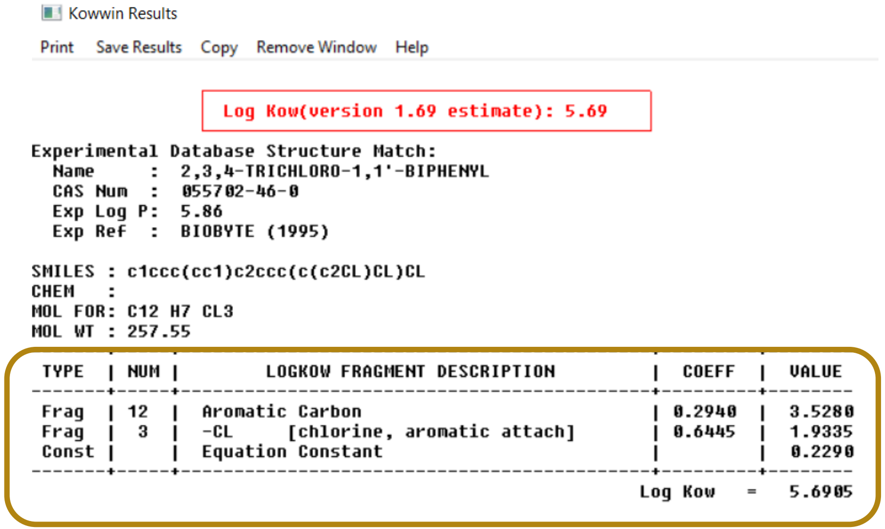
How to select appropriate model(s) for my substance?

- ***A priori*** selection is generally **not possible**
- However, **experience in using the models** might suggest which one could give more reliable results for certain type of substances (e.g., industrial chemicals, active substances, etc.)
- Information on **compliance** of the target molecule **with the applicability domain of the model**
- **Comparison with similar molecules** with available experimental results
- **Documentation (QMRF, QPRF)**
- **It is generally required to use multiple and different models for evaluating the same endpoint**

Expert analysis of the results and supporting information is needed

VEGA Meylan/KOWWIN vs EPI Suite KOWWIN

- Both models should provide the same result for any molecule.
- Advantages of VEGA: analysis of the compliance with applicability domain of the model is performed and reported. QMRF is available. The results can be directly compared to the results of other available QSAR models (MLogP, ALogP).
- Advantage of EPI Suite: the final result is explained in terms of contributions of single molecular fragments (more transparency)



```
          Log Kow(version 1.69 estimate): 5.69

Experimental Database Structure Match:
  Name      :  2,3,4-TRICHLORO-1,1'-BIPHENYL
  CAS Num   :  055702-46-0
  Exp Log P:  5.86
  Exp Ref   :  BIOBYTE (1995)

SMILES : c1ccc(cc1)c2ccc(c(c2CL)CL)CL
CHEM   :
MOL FOR: C12 H7 CL3
MOL WT : 257.55
-------+-----+--------------------------------------------+---------+---------
 TYPE  | NUM |        LOGKOW FRAGMENT DESCRIPTION         |  COEFF  |  VALUE
-------+-----+--------------------------------------------+---------+---------
 Frag  | 12  | Aromatic Carbon                            | 0.2940  |  3.5280
 Frag  |  3  | -CL     [chlorine, aromatic attach]        | 0.6445  |  1.9335
 Const |     | Equation Constant                          |         |  0.2290
-------+-----+--------------------------------------------+---------+---------
                                                   Log Kow   =   5.6905
```