	QMRF identifier (JRC Inventory): **To be entered by JRC**
	QMRF Title: QSAR Model Reporting Format Prepared for KATE (Fish Acute)
	Printing Date:

1.QSAR identifier

1.1. QSAR identifier (title)

KATE2020 on NET for acute toxicity to Fish

1.2. Other related models

1.3. Software coding the model

Name: KATE2020 on NET

Description: Ecotoxicity prediction system consists of QSAR models

Contact: kate@nies.go.jp

WWW: https://kate2.nies.go.jp/

2.General information

2.1. Date of QMRF

14 January 2020

2.2. QMRF author(s) and contact details

Names: Hiroshi Yamamoto, Koji Imai, Yuto Itami, Ayako Furuhama, and Koichi Ohno

Affiliation: National Institute for Environmental Studies, Japan

Contact: 16-2 Onogawa, Tsukuba, Ibaraki, 305-8506, Japan

e-mail: kate@nies.go.jp

WWW: https://www.nies.go.jp/index-e.html

2.3. Date of QMRF update(s)

2.4. QMRF update(s)

2.5. Model developer(s) and contact details

Names: Hiroaki Shiraishi and Koji Imai

Affiliation: National Institute for Environmental Studies, Japan

Contact: 16-2 Onogawa, Tsukuba, Ibaraki, 305-8506, Japan

e-mail: kate@nies.go.jp

WWW: https://www.nies.go.jp/index-e.html

2.6. Date of model development and/or publication

3 February 2020

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

(1) KATE2020 on NET https://kate2.nies.go.jp/

(2) A Furuhama, T Toida, N Nishikawa, Y Aoki, Y Yoshioka, and H Shiraishi (2010). Development of an ecotoxicity QSAR model for the KAshinhou Tool for Ecotoxicity (KATE) system, March 2009 version. SAR and QSAR Environmental Research 21 (5-6) 403–413
http://dx.doi.org/10.1080/1062936X.2010.501815.

(3) A. Furuhama, K. Hasunuma, Y. Aoki, Y. Yoshioka and H. Shiraishi (2011). Application of chemical reaction mechanistic domains to an ecotoxicity QSAR model, KAshinhou Tool for Ecotoxicity (KATE). SAR and QSAR Environmental Research 22 (5-6) 505–523
http://dx.doi.org/10.1080/1062936X.2011.569944.

2.8. Availability of information about the model

The model is proprietary, but the training set and algorithm are available.

2.9. Availability of another QMRF for exactly the same model

None to date.

3.Defining the endpoint - OECD Principle 1

3.1. Species

Japanese medaka (*Oryzias latipes*) and Fathead minnow (*Pimephales promelas*)

3.2. Endpoint

Acute toxicity to fish (lethality)

3.3. Comment on endpoint

The model predicts the 50% lethal concentration (LC_{50}) values in the Fish Acute Toxicity Test (96-h exposure).

3.4. Endpoint units

mg/L

3.5. Dependent variable

All of the measured endpoint values were converted to mmol/L units, replaced with their reciprocal, and transformed to logarithmic units. The dependent variable is: $\log (1/LC_{50})$

3.6. Experimental protocol

Japanese medaka: OECD Test Guideline No. 203 test protocol

Fathead minnow: protocol described at Russom, C.L., S.P. Bradbury, S.J. Broderius, D.E. Hammermeister and R.A. Drummond. 1997. *Predicting modes of action from chemical structure: Acute toxicity in the fathead minnow (Pimephales promelas)*. Environmental Toxicology and Chemistry 16(5): 948-967.

3.7. Endpoint data quality and variability

The data are from the ecotoxicity tests conducted by the Ministry of the Environment, Government of Japan (Fish acute toxicity tests) according to good laboratory practice (GLP) and the Fish acute toxicity tests from the US Environmental Protection Agency fathead minnow database.

The former data are available at: http://www.env.go.jp/en/chemi/sesaku/aquatic_Mar_2019.pdf

Fathead minnow database is available at:

https://archive.epa.gov/med/med_archive_03/web/html/fathead_minnow.html

(Accessed on January 14, 2020)

4.Defining the algorithm - OECD Principle 2

4.1. Type of model

QSAR

4.2. Explicit algorithm

The predicted values are calculated with the following algorithm:

(1) Extract the information about the number of substructures, which we defined, in the target compound. (As for the definition of the substructures, see the spreadsheets "substructure_def" and "substructure_def_for_AD" in the "KATE2020_fish_acute.xlsx").

(2) Extract the adequate structure classes in the target compound from the information about the number of the substructure (see 4.5. (3)).

(3) Predict the ecotoxicity of the target compound using the descriptor (log P) and the QSAR equation (slope and intercept) provided for each structure class.

(4) Convert the unit of the predicted ecotoxicity from $\log (1/LC_{50}$ [mmol/L]) to [mg/L] using the molecular weight (see 3.5).

4.3. Descriptors in the model

Descriptor: log P   Octanol–water partition coefficient

For structural fragments, see the spreadsheets "substructure_def", "substructure_def_for_AD" and "structure_def" in the attachment file "KATE2020_fish_acute.xlsx".

4.4. Descriptor selection

See 4.2.

4.5. Algorithm and descriptor generation

(1) Log P can be input by the user, otherwise log P is calculated using KOWWIN from SMILES.

(2) The information about the number of substructures (see 4.2. (1)) is calculated using the JAVA application developed by NIES using Chemistry Development Kit (CDK) from inputted SMILES and the definition of the substructure.

(3) The adequacy to each structure class (1:Yes, 0:No) is evaluated by the information about the number of the substructures and the definition of the structure class (see the spreadsheet "structure_def" in the file "KATE2020_fish_acute.xlsx") in the KATE2020 on NET.

(4) Molecular weight of the target compound is calculated from SMILES by using Open Babel.

4.6. Software name and version for descriptor generation

Name: KOWWIN

Description: A program to estimate log P values of organic compounds. That is incorporated in EPI Suite™ (Estimation Programs Interface: a tool intended for use in applications such as to quickly screen chemicals), which was developed by the US EPA, etc.

KOWWIN v1.69 (April 2015)

KOWWIN™ (included in EPI Suite™)
• https://www.epa.gov/tsca-screening-tools/epi-suitetm-estimation-programinterface

Name: None
Description: An application for SMARTS substructure searching from SMILES that we have developed by using Chemistry Development Kit (CDK) version 2.0
Contact: kate@nies.go.jp
WWW: https://kate2.nies.go.jp/

Name: Open Babel
Description: The Open Source Chemistry Toolbox, version 2.3.1, 2011.
Contact: Contact not available
WWW: http://openbabel.org/wiki/Main_Page

4.7. Chemicals/Descriptors ratio

The ratio (column name: ratio1) for each QSAR class is described in the spreadsheet "qsar_data" in the attachment file "KATE2020_fish_acute.xlsx".

5.Defining the applicability domain - OECD Principle 3

5.1. Description of the applicability domain of the model

(1) Applicability domain of the model for the descriptor (log P) is established for each QSAR class, which we define as a group of compounds with specific structure classes, and is the range of the descriptor, "between the maximum and the minimum" (see the column named "Log P Range" of the spreadsheet "qsar_data" in the attached file "KATE2020_fish_acute.xlsx").

(2) Applicability domain of the model for the structure is also established for each QSAR class, and is provided from the list of substructure for the structure judgement of the chemicals in the training set which belongs to the QSAR class and in the class called narcotic group" (see the column "FragIDs_for_AD" of the spreadsheet "qsar_data" in the attached file "KATE2020_fish_acute.xlsx")

5.2. Method used to assess the applicability domain

The applicability domain of both descriptor and structure is assessed for each QSAR class into which the target compound is classified.

(1) If the log P value of the target compound is within the applicability domain of the descriptor as indicated in 5.1. (1), it is judged as "inside of the applicability domain" in terms of the descriptor.

(2) If the substructure for the structure judgement (see the substructure defined in the spreadsheet "substructure_def_for_AD" in the file "KATE2020_fish_acute.xlsx") possessed by the target compound is all included in the list of the structure judgement of the QSAR class (see 5.1.(2)), it is judged as "inside of the applicability domain" in terms of the structure.

For more details, see Furuhama et al. (2011) (see 2.7 (3)).

5.3. Software name and version for applicability domain assessment

Name: KATE2020 on NET

Contact: kate@nies.go.jp

WWW: https://kate2.nies.go.jp

5.4. Limits of applicability

See 5.1 and 5.2.

6.Internal validation - OECD Principle 4

6.1. Availability of the training set

It is available in supporting information.

6.2. Available information for the training set

Chemical Name: Yes

SMILES: Yes

CAS RN: Yes

InChI: No

MOL file: No

Formula: No

6.3. Data for each descriptor variable for the training set

No

6.4. Data for the dependent variable for the training set

The data for the dependent variable for each QSAR class are available in the spreadsheet "training_set" in the attachment file "KATE2020_fish_acute.xlsx".

6.5. Other information about the training set

The number of the training set for each QSAR class is described on the column named "n" in the spreadsheet "qsar_data" in the attachment file "KATE2020_fish_acute.xlsx".

6.6. Pre-processing of data before modelling

The endpoint data LC_{50} [mg/L] was converted into log ($1/LC_{50}$ [mmol/L]).

6.7. Statistics for goodness-of-fit

The following statistical values for each QSAR class are obtained and shown in the spreadsheet "qsar_data" in the attachment file "KATE2020_fish_acute.xlsx".

R^2: square of correlation coefficient

RMSE: Root Mean Squared Error

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

The following statistical value for each QSAR class is obtained and shown in the spreadsheet "qsar_data" in the attachment file "KATE2020_fish_acute.xlsx".

Q^2: square of correlation coefficient

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

No

6.10. Robustness - Statistics obtained by Y-scrambling

No

6.11. Robustness - Statistics obtained by bootstrap

No

6.12. Robustness - Statistics obtained by other methods

No

7.External validation - OECD Principle 4

7.1. Availability of the external validation set

    Available for some QSAR classes, and available in supporting information.

7.2. Available information for the external validation set

    Chemical Name: No

    SMILES: No

    CAS RN: No

    InChI: No

    MOL file: No

    Formula: 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

    No

7.6. Experimental design of test set

    No

7.7. Predictivity - Statistics obtained by external validation

    No

7.8. Predictivity - Assessment of the external validation set

    No

7.9. Comments on the external validation of the model

    No

8.Providing a mechanistic interpretation - OECD Principle 5

8.1. Mechanistic basis of the model

Each QSAR class is a group of compounds with specific structure class. Structure class is defined by the information about the substructures (see 4.2) and not directly relevant to the mode of actions. There are a few classes which can be relevant to the mode of actions. For example, QSAR ID: 12899941 (in the spreadsheet of "qsar_data") is a group of chemicals with structure class: GD_28999 (narcotic group), and considered to exert narcotic actions. This structure class was created with the integration of the multiple structure classes including aliphatic hydrocarbons, aliphatic and aromatic ethers, aliphatic and aromatic ketones, and alcohols which are considered to have simple narcotic actions.

8.2. A priori or a posteriori mechanistic interpretation

A priori mechanistic interpretation.

8.3. Other information about the mechanistic interpretation

No

9.Miscellaneous information

9.1. Comments

This model can output multiple estimations based on the adequacy to the structure classes (see 4.2). There is neither recommended value nor guideline to select one right now and the users should select by themselves based on R2, Q2, the shape of the regression line (slope and the number of reference chemicals), and the applicability domain based on descriptor (log P) and structure (structure judgement).

9.2. Bibliography

[1] OECD Guideline of Testing of Chemicals, No. 203 (17th July 1992) - Fish, Acute Toxicity Test.

9.3. Supporting information

1. KATE2020_fish_acute.xlsx

The name and content of each spreadsheet are as follows:

substructure_def: definition of substructures

substructure_def_for_AD: definition of substructures (also used for the applicability domain)

structure_def: definition of structure classes

qsar_data: data for each QSAR class (QSAR equation etc.)

training_set: data of the training set

10.Summary (JRC QSAR Model Database)

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

	*QMRF identifier (JRC Inventory):***To be entered by JRC**
	QMRF Title: QSAR Model Reporting Format Prepared for KATE (Fish Chronic)
	Printing Date:

1.QSAR identifier

1.1. QSAR identifier (title)

KATE2020 on NET for chronic toxicity to Fish

1.2. Other related models

1.3. Software coding the model

Name: KATE2020 on NET

Description: Ecotoxicity prediction system consists of QSAR models

Contact: kate@nies.go.jp

WWW: https://kate2.nies.go.jp/

2.General information

2.1. Date of QMRF

14 January 2020

2.2. QMRF author(s) and contact details

Names: Hiroshi Yamamoto, Koji Imai, Yuto Itami, Ayako Furuhama, and Koichi Ohno

Affiliation: National Institute for Environmental Studies, Japan

Contact: 16-2 Onogawa, Tsukuba, Ibaraki, 305-8506, Japan

e-mail: kate@nies.go.jp

WWW: https://www.nies.go.jp/index-e.html

2.3. Date of QMRF update(s)

2.4. QMRF update(s)

2.5. Model developer(s) and contact details

Names: Hiroaki Shiraishi and Koji Imai

Affiliation: National Institute for Environmental Studies, Japan

Contact: 16-2 Onogawa, Tsukuba, Ibaraki, 305-8506, Japan

e-mail: kate@nies.go.jp

WWW: https://www.nies.go.jp/index-e.html

2.6. Date of model development and/or publication

3 February 2020

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

(1) KATE2020 on NET https://kate2.nies.go.jp/

(2) A Furuhama, T Toida, N Nishikawa, Y Aoki, Y Yoshioka, and H Shiraishi (2010). Development of an ecotoxicity QSAR model for the KAshinhou Tool for Ecotoxicity (KATE) system, March 2009 version. SAR and QSAR Environmental Research 21 (5-6) 403–413
http://dx.doi.org/10.1080/1062936X.2010.501815.

(3) A. Furuhama, K. Hasunuma, Y. Aoki, Y. Yoshioka and H. Shiraishi (2011). Application of chemical reaction mechanistic domains to an ecotoxicity QSAR model, KAshinhou Tool for Ecotoxicity (KATE). SAR and QSAR Environmental Research 22 (5-6) 505–523
http://dx.doi.org/10.1080/1062936X.2011.569944.

2.8. Availability of information about the model

The model is proprietary, but the training set and algorithm are available.

2.9. Availability of another QMRF for exactly the same model

None to date.

3.Defining the endpoint - OECD Principle 1

3.1. Species

Japanese medaka (*Oryzias Latipes*)

3.2. Endpoint

Long-term toxicity to fish (early-life stage)

3.3. Comment on endpoint

The model predicts the no-observed-effect concentration (NOEC) values in the Fish Early-Life Stage Toxicity Test.

3.4. Endpoint units

mg/L

3.5. Dependent variable

All of the measured endpoint values were converted to mmol/L units, replaced with their reciprocal, and transformed to logarithmic units. The dependent variable is: log (1/NOEC)

3.6. Experimental protocol

OECD Test Guideline No. 210 test protocol

3.7. Endpoint data quality and variability

The data are from the ecotoxicity tests conducted by the Ministry of the Environment, Government of Japan (Fish early-life stage toxicity tests) according to good laboratory practice (GLP).

This data are available at: http://www.env.go.jp/en/chemi/sesaku/aquatic_Mar_2019.pdf

(Accessed on January 14, 2020)

4.Defining the algorithm - OECD Principle 2

4.1. Type of model

QSAR

4.2. Explicit algorithm

The predicted values are calculated with the following algorithm:

(1) Extract the information about the number of substructures, which we defined, in the target compound. (As for the definition of the substructures, see the spreadsheets "substructure_def" and "substructure_def_for_AD" in the "KATE2020_fish_chronic.xlsx").

(2) Extract the adequate structure classes in the target compound from the information about the number of the substructure (see 4.5. (3)).

(3) Predict the ecotoxicity of the target compound using the descriptor (log P) and the QSAR equation (slope and intercept) provided for each structure class.

(4) Convert the unit of the predicted ecotoxicity from log (1/NOEC [mmol/L]) to [mg/L] using the molecular weight (see 3.5).

4.3. Descriptors in the model

Descriptor: log P   Octanol–water partition coefficient

For structural fragments, see the spreadsheets "substructure_def", "substructure_def_for_AD" and "structure_def" in the attachment file "KATE2020_fish_chronic.xlsx".

4.4. Descriptor selection

See 4.2.

4.5. Algorithm and descriptor generation

(1) Log P can be input by the user, otherwise log P is calculated using KOWWIN from SMILES notation.

(2) The information about the number of substructures (see 4.2. (1)) is calculated using the JAVA application developed by NIES using Chemistry Development Kit (CDK) from inputted SMILES and the definition of the substructure.

(3) The adequacy to each structure class (1:Yes, 0:No) is evaluated by the information about the number of the substructures and the definition of the structure class (see the spreadsheet "structure_def" in the file "KATE2020_fish_chronic.xlsx") in the KATE2020 on NET.

(4) Molecular weight of the target compound is calculated from SMILES by using Open Babel.

4.6. Software name and version for descriptor generation

Name: KOWWIN

Description: A program to estimate log P values of organic compounds. That is incorporated in EPI Suite™ (Estimation Programs Interface: a tool intended for use in applications such as to quickly screen chemicals), which was developed by the US EPA, etc.

KOWWIN v1.69 (April 2015)

KOWWIN™ (included in EPI Suite™)
• https://www.epa.gov/tsca-screening-tools/epi-suitetm-estimation-programinterface
Name: None
Description: An application for SMARTS substructure searching from SMILES that we have developed by using Chemistry Development Kit (CDK) version 2.0
Contact: kate@nies.go.jp
WWW: https://kate2.nies.go.jp/

Name: Open Babel
Description: The Open Source Chemistry Toolbox, version 2.3.1, 2011.
Contact: Contact not available
WWW: http://openbabel.org/wiki/Main_Page

4.7. Chemicals/Descriptors ratio

The ratio (column name: ratio1) for each QSAR class is described in the spreadsheet "qsar_data" in the attachment file "KATE2020_fish_chronic.xlsx".

5.Defining the applicability domain - OECD Principle 3

5.1. Description of the applicability domain of the model

(1) Applicability domain of the model for the descriptor (log P) is established for each QSAR class, which we define as a group of compounds with specific structure classes, and is the range of the descriptor, "between the maximum and the minimum" (see the column named "Log P Range" of the spreadsheet "qsar_data" in the attached file "KATE2020_fish_chronic.xlsx").

(2) Applicability domain of the model for the structure is also established for each QSAR class, and is provided from the list of substructure for the structure judgement of the chemicals in the training set which belongs to the QSAR class and in the class called "narcotic group" (see the column "FragIDs_for_AD" of the spreadsheet "qsar_data" in the attached file "KATE2020_fish_chronic.xlsx")

5.2. Method used to assess the applicability domain

The applicability domain of both descriptor and structure is assessed for each QSAR class into which the target compound is classified.

(1) If the log P value of the target compound is within the applicability domain of the descriptor as indicated in 5.1.(1), it is judged as "inside of the applicability domain" in terms of the descriptor.

(2) If the substructure for the structure judgement (see the substructure defined in the spreadsheet "substructure_def_for_AD" in the file "KATE2020_fish_chronic.xlsx") possessed by the target compound is all included in the list of the structure judgement of the QSAR class (see 5.1.(2)), it is judged as "inside of the applicability domain" in terms of the structure.

For more details, see Furuhama et al. (2011) (see 2.7 (3)).

5.3. Software name and version for applicability domain assessment

Name: KATE2020 on NET

Contact: kate@nies.go.jp

WWW: https://kate2.nies.go.jp

5.4. Limits of applicability

See 5.1 and 5.2.

6.Internal validation - OECD Principle 4

6.1. Availability of the training set

It is available in the supporting information.

6.2. Available information for the training set

Chemical Name: Yes

SMILES: Yes

CAS RN: Yes

InChI: No

MOL file: No

Formula: No

6.3. Data for each descriptor variable for the training set

No

6.4. Data for the dependent variable for the training set

The data for the dependent variable for each QSAR class are available in the spreadsheet "training_set" in the attachment file "KATE2020_fish_chronic.xlsx".

6.5. Other information about the training set

The number of the training set for each QSAR class is described on the column named "n" in the spreadsheet "qsar_data" in the attachment file "KATE2020_fish_chronic.xlsx".

6.6. Pre-processing of data before modelling

The endpoint data NOEC [mg/L] was converted into log (1/NOEC [mmol/L]).

6.7. Statistics for goodness-of-fit

The following statistical values for each QSAR class are obtained and shown in the spreadsheet "qsar_data" in the attachment file "KATE2020_fish_chronic.xlsx".

R^2: square of correlation coefficient

RMSE: Root Mean Squared Error

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

The following statistical value for each QSAR class is obtained and shown in the spreadsheet "qsar_data" in the attachment file "KATE2020_fish_chronic.xlsx".

Q^2: square of correlation coefficient

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

No

6.10. Robustness - Statistics obtained by Y-scrambling

No

6.11. Robustness - Statistics obtained by bootstrap

No

6.12. Robustness - Statistics obtained by other methods

No

7.External validation - OECD Principle 4

7.1. Availability of the external validation set

No.

7.2. Available information for the external validation set

Chemical Name: No

SMILES: No

CAS RN: No

InChI: No

MOL file: No

Formula: 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

No

7.6. Experimental design of test set

No

7.7. Predictivity - Statistics obtained by external validation

No

7.8. Predictivity - Assessment of the external validation set

No

7.9. Comments on the external validation of the model

No

8.Providing a mechanistic interpretation - OECD Principle 5

8.1. Mechanistic basis of the model

Each QSAR class is a group of compounds with specific structure class. Structure class is defined by the information about the substructures (see 4.2) and not directly relevant to the mode of actions. There are a few classes which can be relevant to the mode of actions. For example, QSAR ID: 12899951 (in the spreadsheet of "qsar_data") is a group of chemicals with structure class: GD_28999 (narcotic group), and considered to exert narcotic actions. This structure class was created with the integration of the multiple structure classes including aliphatic hydrocarbons, aliphatic and aromatic ethers, aliphatic and aromatic ketones, and alcohols which are considered to have simple narcotic actions.

8.2. A priori or a posteriori mechanistic interpretation

A priori mechanistic interpretation.

8.3. Other information about the mechanistic interpretation

No

9.Miscellaneous information

9.1. Comments

This model can output multiple estimations based on the adequacy to the structure classes (see 4.2). There is neither recommended value nor guideline to select one right now and the users should select by themselves based on R2, Q2, the shape of the regression line (slope and the number of reference chemicals), and the applicability domain based on descriptor (log P) and structure (structure judgement).

9.2. Bibliography

[1] OECD Guideline for Testing of Chemicals, No. 210 (26th July 2013) - Fish, Early-Life Stage Toxicity Test.

9.3. Supporting information

1. KATE2020_fish_chronic.xlsx

The name and content of each spreadsheet are as follows:

substructure_def: definition of substructures

substructure_def_for_AD: definition of substructures (also used for the applicability domain)

structure_def: definition of structure classes

qsar_data: data for each QSAR class (QSAR equation etc.)

training_set: data of the training set

10.Summary (JRC QSAR Model Database)

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

	*QMRF identifier (JRC Inventory):***To be entered by JRC**
	QMRF Title: QSAR Model Reporting Format Prepared for KATE (Daphnid Acute)
	Printing Date:

1.QSAR identifier

1.1. QSAR identifier (title)

KATE2020 on NET for acute toxicity to Daphnid

1.2. Other related models

1.3. Software coding the model

Name: KATE2020 on NET

Description: Ecotoxicity prediction system consists of QSAR models

Contact: kate@nies.go.jp

WWW: https://kate2.nies.go.jp/

2.General information

2.1. Date of QMRF

14 January 2020

2.2. QMRF author(s) and contact details

Names: Hiroshi Yamamoto, Koji Imai, Yuto Itami, Ayako Furuhama, and Koichi Ohno

Affiliation: National Institute for Environmental Studies, Japan

Contact: 16-2 Onogawa, Tsukuba, Ibaraki, 305-8506, Japan

e-mail: kate@nies.go.jp

WWW: https://www.nies.go.jp/index-e.html

2.3. Date of QMRF update(s)

2.4. QMRF update(s)

2.5. Model developer(s) and contact details

Names: Hiroaki Shiraishi and Koji Imai

Affiliation: National Institute for Environmental Studies, Japan

Contact: 16-2 Onogawa, Tsukuba, Ibaraki, 305-8506, Japan

e-mail: kate@nies.go.jp

WWW: https://www.nies.go.jp/index-e.html

2.6. Date of model development and/or publication

3 February 2020

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

(1) KATE2020 on NET https://kate2.nies.go.jp/

(2) A Furuhama, T Toida, N Nishikawa, Y Aoki, Y Yoshioka, and H Shiraishi (2010). Development of an ecotoxicity QSAR model for the KAshinhou Tool for Ecotoxicity (KATE) system, March 2009 version. SAR and QSAR Environmental Research 21 (5-6) 403–413 http://dx.doi.org/10.1080/1062936X.2010.501815.

(3) A. Furuhama, K. Hasunuma, Y. Aoki, Y. Yoshioka and H. Shiraishi (2011). Application of chemical reaction mechanistic domains to an ecotoxicity QSAR model, KAshinhou Tool for Ecotoxicity (KATE). SAR and QSAR Environmental Research 22 (5-6) 505–523 http://dx.doi.org/10.1080/1062936X.2011.569944.

2.8. Availability of information about the model

The model is proprietary, but the training set and algorithm are available.

2.9. Availability of another QMRF for exactly the same model

None to date.

3.Defining the endpoint - OECD Principle 1

3.1. Species

Daphnia magna

3.2. Endpoint

Short-term toxicity to Daphnid (immobilization)

3.3. Comment on endpoint

The model predicts the 50% effective concentration (EC_{50}) values in the *Daphnia* sp. Acute Immobilization Test (48-h exposure).

3.4. Endpoint units

mg/L

3.5. Dependent variable

All of the measured endpoint values were converted to mmol/L units, replaced with their reciprocal, and transformed to logarithmic units. The dependent variable is: $\log (1/EC_{50})$

3.6. Experimental protocol

OECD Test Guideline No. 202 test protocol

3.7. Endpoint data quality and variability

The data are from the ecotoxicity tests conducted by the Ministry of the Environment, Government of Japan (*Daphnia* sp. immobilization tests) according to good laboratory practice (GLP).

This data are available at: http://www.env.go.jp/en/chemi/sesaku/aquatic_Mar_2019.pdf

(Accessed on January 14, 2020)

4.Defining the algorithm - OECD Principle 2

4.1. Type of model

QSAR

4.2. Explicit algorithm

The predicted values are calculated with the following algorithm:

(1) Extract the information about the number of substructures, which we defined, in the target compound. (As for the definition of the substructures, see the spreadsheets "substructure_def" and "substructure_def_for_AD" in the "KATE2020_daphnid_acute.xlsx").

(2) Extract the adequate structure classes in the target compound from the information about the number of the substructure (see 4.5. (3)).

(3) Predict the ecotoxicity of the target compound using the descriptor (log P) and the QSAR equation (slope and intercept) provided for each structure class.

(4) Convert the unit of the predicted ecotoxicity from log ($1/EC_{50}$ [mmol/L]) to [mg/L] using the molecular weight (see 3.5).

4.3. Descriptors in the model

Descriptor: log P   Octanol–water partition coefficient

For structural fragments, see the spreadsheets "substructure_def", "substructure_def_for_AD" and "structure_def" in the attachment file "KATE2020_daphnid_acute.xlsx".

4.4. Descriptor selection

See 4.2.

4.5. Algorithm and descriptor generation

(1) Log P can be input by the user, otherwise log P is calculated using KOWWIN from SMILES notation.

(2) The information about the number of substructures (see 4.2. (1)) is calculated using the JAVA application developed by NIES using Chemistry Development Kit (CDK) from inputted SMILES and the definition of the substructure.

(3) The adequacy to each structure class (1:Yes, 0:No) is evaluated by the information about the number of the substructures and the definition of the structure class (see the spreadsheet "structure_def" in the file "KATE2020_daphnid_acute.xlsx") in the KATE2020 on NET.

(4) Molecular weight of the target compound is calculated from SMILES by using Open Babel.

4.6. Software name and version for descriptor generation

Name: KOWWIN

Description: A program to estimate log P values of organic compounds. That is incorporated in EPI Suite™ (Estimation Programs Interface: a tool intended for use in applications such as to quickly screen chemicals), which was developed by the US EPA, etc.

KOWWIN v1.69 (April 2015)

KOWWIN™ (included in EPI Suite™)
• https://www.epa.gov/tsca-screening-tools/epi-suitetm-estimation-programinterface

Name: None
Description: An application for SMARTS substructure searching from SMILES that we have developed by using Chemistry Development Kit (CDK) version 2.0
Contact: kate@nies.go.jp
WWW: https://kate2.nies.go.jp/

Name: Open Babel
Description: The Open Source Chemistry Toolbox, version 2.3.1, 2011.
Contact: Contact not available
WWW: http://openbabel.org/wiki/Main_Page

4.7. Chemicals/Descriptors ratio

The ratio (column name: ratio1) for each QSAR class is described in the spreadsheet "qsar_data" in the attachment file "KATE2020_daphnid_acute.xlsx".

5.Defining the applicability domain - OECD Principle 3

5.1. Description of the applicability domain of the model

(1) Applicability domain of the model for the descriptor (log P) is established for each QSAR class, which we define as a group of compounds with specific structure classes, and is the range of the descriptor, "between the maximum and the minimum" (see the column named "Log P Range" of the spreadsheet "qsar_data" in the attached file "KATE2020_daphnid_acute.xlsx").

(2) Applicability domain of the model for the structure is also established for each QSAR class, and is provided from the list of substructures for the structure judgement of the chemicals in the training set which belongs to the QSAR class and in the class called "narcotic group" (see the column "FragIDs_for_AD" of the spreadsheet "qsar_data" in the attached file "KATE2020_daphnid_acute.xlsx")

5.2. Method used to assess the applicability domain

The applicability domain of both descriptor and structure is assessed for each QSAR class into which the target compound is classified.

(1) If the log P value of the target compound is within the applicability domain of the descriptor as indicated in 5.1. (1), it is judged as "inside of the applicability domain" in terms of the descriptor.

(2) If the substructure for the structure judgement (see the substructure defined in the spreadsheet "substructure_def_for_AD" in the file "KATE2020_daphnid_acute.xlsx") possessed by the target compound is all included in the list of the structure judgement of the QSAR class (see 5.1.(2)), it is judged as "inside of the applicability domain" in terms of the structure. For more details, see Furuhama et al. (2011) (see 2.7 (3)).

5.3. Software name and version for applicability domain assessment

Name: KATE2020 on NET

Contact: kate@nies.go.jp

WWW: https://kate2.nies.go.jp

5.4. Limits of applicability

See 5.1 and 5.2.

6.Internal validation - OECD Principle 4

6.1. Availability of the training set

It is available and attached as supporting information.

6.2. Available information for the training set

Chemical Name: Yes

SMILES: Yes

CAS RN: Yes

InChI: No

MOL file: No

Formula: No

6.3. Data for each descriptor variable for the training set

No

6.4. Data for the dependent variable for the training set

The data for the dependent variable for each QSAR class are available in the spreadsheet "training_set" in the attachment file "KATE2020_daphnid_acute.xlsx".

6.5. Other information about the training set

The number of the training set for each QSAR class is described on the column named "n" in the spreadsheet "qsar_data" in the attachment file "KATE2020_daphnid_acute.xlsx".

6.6. Pre-processing of data before modelling

The endpoint data EC_{50} [mg/L] was converted into log $(1/EC_{50}$ [mmol/L]).

6.7. Statistics for goodness-of-fit

The following statistical values for each QSAR class are obtained and shown in the spreadsheet "qsar_data" in the attachment file "KATE2020_daphnid_acute.xlsx".

R^2: square of correlation coefficient

RMSE: Root Mean Squared Error

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

The following statistical value for each QSAR class is obtained and shown in the spreadsheet "qsar_data" in the attachment file "KATE2020_daphnid_acute.xlsx".

Q^2: square of correlation coefficient

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

No

6.10. Robustness - Statistics obtained by Y-scrambling

No

6.11. Robustness - Statistics obtained by bootstrap

No

6.12. Robustness - Statistics obtained by other methods

No

7.External validation - OECD Principle 4

7.1. Availability of the external validation set

Available for some QSAR classes, and in supporting information.

7.2. Available information for the external validation set

Chemical Name: No

SMILES: No

CAS RN: No

InChI: No

MOL file: No

Formula: 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

No

7.6. Experimental design of test set

No

7.7. Predictivity - Statistics obtained by external validation

No

7.8. Predictivity - Assessment of the external validation set

No

7.9. Comments on the external validation of the model

No

8.Providing a mechanistic interpretation - OECD Principle 5

8.1. Mechanistic basis of the model

Each QSAR class is a group of compounds with specific structure class. Structure class is defined by the information about the substructures (see 4.2) and not directly relevant to the mode of actions. There are a few classes which can be relevant to the mode of actions. For example, QSAR ID: 22899941 (in the spreadsheet of "qsar_data") is a group of chemicals with structure class: GD_28999 (narcotic group), and considered to exert narcotic actions. This structure class was created with the integration of the multiple structure classes including aliphatic hydrocarbons, aliphatic and aromatic ethers, aliphatic and aromatic ketones, and alcohols which are considered to have simple narcotic actions.

8.2. A priori or a posteriori mechanistic interpretation

A priori mechanistic interpretation.

8.3. Other information about the mechanistic interpretation

No

9.Miscellaneous information

9.1. Comments

This model can output multiple estimations based on the adequacy to the structure classes (see 4.2). There is neither recommended value nor guideline to select one right now and the users should select by themselves based on R2, Q2, the shape of the regression line (slope and the number of reference chemicals), and the applicability domain based on descriptor (log P) and structure (structure judgement).

9.2. Bibliography

[1] OECD Guideline for Testing of Chemicals, No. 202 (13th April 2004) - *Daphnia* sp., Acute Immobilization Test.

9.3. Supporting information

1. KATE2020_daphnid_acute.xlsx

   The name and content of each spreadsheet are as follows:

   substructure_def: definition of substructures

   substructure_def_for_AD: definition of substructures (also used for the applicability domain)

   structure_def: definition of structure classes

   qsar_data: data for each QSAR class (QSAR equation etc.)

   training_set: data of the training set

10.Summary (JRC QSAR Model Database)

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

	QMRF identifier (JRC Inventory): **To be entered by JRC**
	QMRF Title: QSAR Model Reporting Format Prepared for KATE (Daphnid Chronic)
	Printing Date:

1.QSAR identifier

1.1. QSAR identifier (title)

KATE2020 on NET for chronic toxicity to Daphnid

1.2. Other related models

1.3. Software coding the model

Name: KATE2020 on NET

Description: Ecotoxicity prediction system consists of QSAR models

Contact: kate@nies.go.jp

WWW: https://kate2.nies.go.jp/

2.General information

2.1. Date of QMRF

14 January 2020

2.2. QMRF author(s) and contact details

Names: Hiroshi Yamamoto, Koji Imai, Yuto Itami, Ayako Furuhama, and Koichi Ohno

Affiliation: National Institute for Environmental Studies, Japan

Contact: 16-2 Onogawa, Tsukuba, Ibaraki, 305-8506, Japan

e-mail: kate@nies.go.jp

WWW: https://www.nies.go.jp/index-e.html

2.3. Date of QMRF update(s)

2.4. QMRF update(s)

2.5. Model developer(s) and contact details

Names: Hiroaki Shiraishi and Koji Imai

Affiliation: National Institute for Environmental Studies, Japan

Contact: 16-2 Onogawa, Tsukuba, Ibaraki, 305-8506, Japan

e-mail: kate@nies.go.jp

WWW: https://www.nies.go.jp/index-e.html

2.6. Date of model development and/or publication

3 February 2020

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

(1) KATE2020 on NET https://kate2.nies.go.jp/

(2) A Furuhama, T Toida, N Nishikawa, Y Aoki, Y Yoshioka, and H Shiraishi (2010). Development of an ecotoxicity QSAR model for the KAshinhou Tool for Ecotoxicity (KATE) system, March 2009 version. SAR and QSAR Environmental Research 21 (5-6) 403–413
http://dx.doi.org/10.1080/1062936X.2010.501815.

(3) A. Furuhama, K. Hasunuma, Y. Aoki, Y. Yoshioka and H. Shiraishi (2011). Application of chemical reaction mechanistic domains to an ecotoxicity QSAR model, KAshinhou Tool for Ecotoxicity (KATE). SAR and QSAR Environmental Research 22 (5-6) 505–523
http://dx.doi.org/10.1080/1062936X.2011.569944.

2.8. Availability of information about the model

The model is proprietary, but the training set and algorithm are available.

2.9. Availability of another QMRF for exactly the same model

None to date.

3.Defining the endpoint - OECD Principle 1

3.1. Species

Daphnia magna

3.2. Endpoint

Long-term toxicity to Daphnid (reproduction output)

3.3. Comment on endpoint

The model predicts the no-observed-effect concentration (NOEC) values in the *Daphnia magna* Reproduction Test (21-d exposure).

3.4. Endpoint units

mg/L

3.5. Dependent variable

All of the measured endpoint values were converted to mmol/L units, replaced with their reciprocal, and transformed to logarithmic units. The dependent variable is: log (1/NOEC)

3.6. Experimental protocol

OECD Test Guideline No. 211 test protocol

3.7. Endpoint data quality and variability

The data are from the ecotoxicity tests conducted by the Ministry of the Environment, Government of Japan (*Daphnia magna* reproduction tests) according to good laboratory practice (GLP).

This data are available at: http://www.env.go.jp/en/chemi/sesaku/aquatic_Mar_2019.pdf

(Accessed on January 14, 2020)

4.Defining the algorithm - OECD Principle 2

4.1. Type of model

QSAR

4.2. Explicit algorithm

The predicted values are calculated with the following algorithm:

(1) Extract the information about the number of substructures, which we defined, in the target compound. (As for the definition of the substructures, see the spreadsheets "substructure_def" and "substructure_def_for_AD" in the "KATE2020_daphnid_chronic.xlsx").

(2) Extract the adequate structure classes in the target compound from the information about the number of the substructure (see 4.5. (3)).

(3) Predict the ecotoxicity of the target compound using the descriptor (log P) and the QSAR equation (slope and intercept) provided for each structure class.

(4) Convert the unit of the predicted ecotoxicity from log (1/NOEC [mmol/L]) to [mg/L] using the molecular weight (see 3.5).

4.3. Descriptors in the model

Descriptor: log P   Octanol–water partition coefficient

For structural fragments, see the spreadsheets "substructure_def", "substructure_def_for_AD" and "structure_def" in the attachment file "KATE2020_daphnid_chronic.xlsx".

4.4. Descriptor selection

See 4.2.

4.5. Algorithm and descriptor generation

(1) Log P can be input by the user, otherwise log P is calculated using KOWWIN from SMILES notation.

(2) The information about the number of substructures (see 4.2. (1)) is calculated using the JAVA application developed by NIES using Chemistry Development Kit (CDK) from inputted SMILES and the definition of the substructure.

(3) The adequacy to each structure class (1:Yes, 0:No) is evaluated by the information about the number of the substructures and the definition of the structure class (see the spreadsheet "structure_def" in the file "KATE2020_daphnid_chronic.xlsx") in the KATE2020 on NET.

(4) Molecular weight of the target compound is calculated from SMILES by using Open Babel.

4.6. Software name and version for descriptor generation

Name: KOWWIN

Description: A program to estimate log P values of organic compounds. That is incorporated in EPI Suite™ (Estimation Programs Interface: a tool intended for use in applications such as to quickly screen chemicals), which was developed by the US EPA, etc.

KOWWIN v1.69 (April 2015)

KOWWIN™ (included in EPI Suite™)
• https://www.epa.gov/tsca-screening-tools/epi-suitetm-estimation-programinterface
Name: None
Description: An application for SMARTS substructure searching from SMILES that we have developed by using Chemistry Development Kit (CDK) version 2.0
Contact: kate@nies.go.jp
WWW: https://kate2.nies.go.jp/

Name: Open Babel
Description: The Open Source Chemistry Toolbox, version 2.3.1, 2011.
Contact: Contact not available
WWW: http://openbabel.org/wiki/Main_Page

4.7. Chemicals/Descriptors ratio

The ratio (column name: ratio1) for each QSAR class is described in the spreadsheet "qsar_data" in the attachment file "KATE2020_daphnid_chronic.xlsx".

5.Defining the applicability domain - OECD Principle 3

5.1. Description of the applicability domain of the model

(1) Applicability domain of the model for the descriptor (log P) is established for each QSAR class, which we define as a group of compounds with specific structure classes, and is the range of the descriptor, "between the maximum and the minimum" (see the column named "Log P Range" of the spreadsheet "qsar_data" in the attached file "KATE2020_daphnid_chronic.xlsx").

(2) Applicability domain of the model for the structure is also established for each QSAR class, and is provided from the list of substructure for the structure judgement of the chemicals in the training set which belongs to the QSAR class and in the class called "narcotic group" (see the column "FragIDs_for_AD" of the spreadsheet "qsar_data" in the attached file "KATE2020_daphnid_chronic.xlsx")

5.2. Method used to assess the applicability domain

The applicability domain of both descriptor and structure is assessed for each QSAR class into which the target compound is classified.

(1) If the log P value of the target compound is within the applicability domain of the descriptor as indicated in 5.1.(1), it is judged as "inside of the applicability domain" in terms of the descriptor.

(2) If the substructure for the structure judgement (see the substructure defined in the spreadsheet "substructure_def_for_AD" in the file "KATE2020_daphnid_chronic.xlsx") possessed by the target compound is all included in the list of the structure judgement of the QSAR class (see 5.1.(2)), it is judged as "inside of the applicability domain" in terms of the structure.

For more details, see Furuhama et al. (2011) (see 2.7 (3)).

5.3. Software name and version for applicability domain assessment

Name: KATE2020 on NET

Contact: kate@nies.go.jp

WWW: https://kate2.nies.go.jp

5.4. Limits of applicability

See 5.1 and 5.2.

6.Internal validation - OECD Principle 4

6.1. Availability of the training set

It is available and attached as supporting information.

6.2. Available information for the training set

Chemical Name: Yes

SMILES: Yes

CAS RN: Yes

InChI: No

MOL file: No

Formula: No

6.3. Data for each descriptor variable for the training set

No

6.4. Data for the dependent variable for the training set

The data for the dependent variable for each QSAR class are available in the spreadsheet "training_set" in the attachment file "KATE2020_daphnid_chronic.xlsx".

6.5. Other information about the training set

The number of the training set for each QSAR class is described on the column named "n" in the spreadsheet "qsar_data" in the attachment file "KATE2020_daphnid_chronic.xlsx".

6.6. Pre-processing of data before modelling

The endpoint data NOEC [mg/L] was converted into log (1/NOEC [mmol/L]).

6.7. Statistics for goodness-of-fit

The following statistical values for each QSAR class are obtained and shown in the spreadsheet "qsar_data" in the attachment file "KATE2020_daphnid_chronic.xlsx".

R^2: square of correlation coefficient

RMSE: Root Mean Squared Error

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

The following statistical value for each QSAR class is obtained and shown in the spreadsheet "qsar_data" in the attachment file "KATE2020_daphnid_chronic.xlsx".

Q^2: square of correlation coefficient

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

No

6.10. Robustness - Statistics obtained by Y-scrambling

No

6.11. Robustness - Statistics obtained by bootstrap

No

6.12. Robustness - Statistics obtained by other methods

No

7.External validation - OECD Principle 4

7.1. Availability of the external validation set

No.

7.2. Available information for the external validation set

Chemical Name: No

SMILES: No

CAS RN: No

InChI: No

MOL file: No

Formula: 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

No

7.6. Experimental design of test set

No

7.7. Predictivity - Statistics obtained by external validation

No

7.8. Predictivity - Assessment of the external validation set

No

7.9. Comments on the external validation of the model

No

8.Providing a mechanistic interpretation - OECD Principle 5

8.1. Mechanistic basis of the model

Each QSAR class is a group of compounds with specific structure class. Structure class is defined by the information about the substructures (see 4.2) and not directly relevant to the mode of actions. There are a few classes which can be relevant to the mode of actions. This structure class was created with the integration of the multiple structure classes including aliphatic hydrocarbons, aliphatic and aromatic ethers, aliphatic and aromatic ketones, and alcohols which are considered to have simple narcotic actions.

8.2. A priori or a posteriori mechanistic interpretation

A priori mechanistic interpretation.

8.3. Other information about the mechanistic interpretation

No

9.Miscellaneous information

9.1. Comments

This model can output multiple estimations based on the adequacy to the structure classes (see 4.2). There is neither recommended value nor guideline to select one right now and the users should select by themselves based on R2, Q2, the shape of the regression line (slope and the number of reference chemicals), and the applicability domain based on descriptor (log P) and structure (structure judgement).

9.2. Bibliography

[1] OECD Guideline for Testing of Chemicals, No. 211 (2nd October 2012) - *Daphnia magna* Reproduction Test.

9.3. Supporting information

1. KATE2020_daphnid_chronic.xlsx

The name and content of each spreadsheet are as follows:

substructure_def: definition of substructures

substructure_def_for_AD: definition of substructures (also used for the applicability domain)

structure_def: definition of structure classes

qsar_data: data for each QSAR class (QSAR equation etc.)

training_set: data of the training set

10.Summary (JRC QSAR Model Database)

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

	*QMRF identifier (JRC Inventory):***To be entered by JRC**
	QMRF Title: QSAR Model Reporting Format Prepared for KATE (Alga Acute)
	Printing Date:

1.QSAR identifier

1.1. QSAR identifier (title)

KATE2020 on NET for acute toxicity to Alga

1.2. Other related models

1.3. Software coding the model

Name: KATE2020 on NET

Description: Ecotoxicity prediction system consists of QSAR models

Contact: kate@nies.go.jp

WWW: https://kate2.nies.go.jp/

2.General information

2.1. Date of QMRF

14 January 2020

2.2. QMRF author(s) and contact details

Names: Hiroshi Yamamoto, Koji Imai, Yuto Itami, Ayako Furuhama, and Koichi Ohno

Affiliation: National Institute for Environmental Studies, Japan

Contact: 16-2 Onogawa, Tsukuba, Ibaraki, 305-8506, Japan

e-mail: kate@nies.go.jp

WWW: https://www.nies.go.jp/index-e.html

2.3. Date of QMRF update(s)

2.4. QMRF update(s)

2.5. Model developer(s) and contact details

Names: Hiroaki Shiraishi and Koji Imai

Affiliation: National Institute for Environmental Studies, Japan

Contact: 16-2 Onogawa, Tsukuba, Ibaraki, 305-8506, Japan

e-mail: kate@nies.go.jp

WWW: https://www.nies.go.jp/index-e.html

2.6. Date of model development and/or publication

3 February 2020

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

(1) KATE2020 on NET https://kate2.nies.go.jp/

(2) A Furuhama, T Toida, N Nishikawa, Y Aoki, Y Yoshioka, and H Shiraishi (2010). Development of an ecotoxicity QSAR model for the KAshinhou Tool for Ecotoxicity (KATE) system, March 2009 version. SAR and QSAR Environmental Research 21 (5-6) 403–413 http://dx.doi.org/10.1080/1062936X.2010.501815.

(3) A. Furuhama, K. Hasunuma, Y. Aoki, Y. Yoshioka and H. Shiraishi (2011). Application of chemical reaction mechanistic domains to an ecotoxicity QSAR model, KAshinhou Tool for Ecotoxicity (KATE). SAR and QSAR Environmental Research 22 (5-6) 505–523 http://dx.doi.org/10.1080/1062936X.2011.569944.

2.8. Availability of information about the model

The model is proprietary, but the training set and algorithm are available.

2.9. Availability of another QMRF for exactly the same model

None to date.

3.Defining the endpoint - OECD Principle 1

3.1. Species

Pseudokirchneriella subcapitata

3.2. Endpoint

Short-term toxicity to alga (inhibition of the exponential growth rate)

3.3. Comment on endpoint

The model predicts the 50% effective concentration (EC_{50}) values in the Alga Growth Inhibition Test (72-h exposure).

3.4. Endpoint units

mg/L

3.5. Dependent variable

All of the measured endpoint values were converted to mmol/L units, replaced with their reciprocal, and transformed to logarithmic units. The dependent variable is: $\log (1/EC_{50})$

3.6. Experimental protocol

OECD Test Guideline No. 201 test protocol

3.7. Endpoint data quality and variability

The data are from the ecotoxicity tests conducted by the Ministry of the Environment, Government of Japan (Alga growth inhibition tests) according to good laboratory practice (GLP).

This data are available at: http://www.env.go.jp/en/chemi/sesaku/aquatic_Mar_2019.pdf

(Accessed on January 14, 2020)

4.1. Type of model

QSAR

4.2. Explicit algorithm

The predicted values are calculated with the following algorithm:

(1) Extract the information about the number of substructures, which we defined, in the target compound. (As for the definition of the substructures, see the spreadsheets "substructure_def" and "substructure_def_for_AD" in the "KATE2020_alga_acute.xlsx").

(2) Extract the adequate structure classes in the target compound from the information about the number of the substructure (see 4.5. (3)).

(3) Predict the ecotoxicity of the target compound using the descriptor (log P) and the QSAR equation (slope and intercept) provided for each structure class.

(4) Convert the unit of the predicted ecotoxicity from log ($1/EC_{50}$ [mmol/L]) to [mg/L] using the molecular weight (see 3.5).

4.3. Descriptors in the model

Descriptor: log P   Octanol–water partition coefficient

For structural fragments, see the spreadsheets "substructure_def", "substructure_def_for_AD" and "structure_def" in the attachment file "KATE2020_alga_acute.xlsx".

4.4. Descriptor selection

See 4.2.

4.5. Algorithm and descriptor generation

(1) Log P can be input by the user, otherwise log P is calculated using KOWWIN from SMILES notation.

(2) The information about the number of substructures (see 4.2. (1)) is calculated using the JAVA application developed by NIES using Chemistry Development Kit (CDK) from inputted SMILES and the definition of the substructure.

(3) The adequacy to each structure class (1:Yes, 0:No) is evaluated by the information about the number of the substructures and the definition of the structure class (see the spreadsheet "structure_def" in the file "KATE2020_alga_acute.xlsx") in the KATE2020 on NET.

(4) Molecular weight of the target compound is calculated from SMILES by using Open Babel.

4.6. Software name and version for descriptor generation

Name: KOWWIN

Description: A program to estimate log P values of organic compounds. That is incorporated in EPI Suite™ (Estimation Programs Interface: a tool intended for use in applications such as to quickly screen chemicals), which was developed by the US EPA, etc.

KOWWIN v1.69 (April 2015)

c 2000-2015 U.S. Environmental Protection Agency

KOWWIN™ (included in EPI Suite™)
• https://www.epa.gov/tsca-screening-tools/epi-suitetm-estimation-programinterface

Name: None
Description: An application for SMARTS substructure searching from SMILES that we have developed by using Chemistry Development Kit (CDK) version 2.0
Contact: kate@nies.go.jp
WWW: https://kate2.nies.go.jp/

Name: Open Babel
Description: The Open Source Chemistry Toolbox, version 2.3.1, 2011.
Contact: Contact not available
WWW: http://openbabel.org/wiki/Main_Page

4.7. Chemicals/Descriptors ratio

The ratio (column name: ratio1) for each QSAR class is described in the spreadsheet "qsar_data" in the attachment file "KATE2020_alga_acute.xlsx".

5.Defining the applicability domain - OECD Principle 3

5.1. Description of the applicability domain of the model

(1) Applicability domain of the model for the descriptor (log P) is established for each QSAR class, which we define as a group of compounds with specific structure classes, and is the range of the descriptor, "between the maximum and the minimum" (see the column named "Log P Range" of the spreadsheet "qsar_data" in the attached file "KATE2020_alga_acute.xlsx").

(2) Applicability domain of the model for the structure is also established for each QSAR class, and is provided from the list of substructure for the structure judgement of the chemicals in the training set which belongs to the QSAR class and in the class called "narcotic group" (see the column "FragIDs_for_AD" of the spreadsheet "qsar_data" in the attached file "KATE2020_alga_acute.xlsx")

5.2. Method used to assess the applicability domain

The applicability domain of both descriptor and structure is assessed for each QSAR class into which the target compound is classified.

(1) If the log P value of the target compound is within the applicability domain of the descriptor as indicated in 5.1.(1), it is judged as "inside of the applicability domain" in terms of the descriptor.

(2) If the substructure for the structure judgement (see the substructure defined in the spreadsheet "substructure_def_for_AD" in the file "KATE2020_alga_acute.xlsx") possessed by the target compound is all included in the list of the structure judgement of the QSAR class (see 5.1.(2)), it is judged as "inside of the applicability domain" in terms of the structure.

For more details, see Furuhama et al. (2011) (see 2.7 (3)).

5.3. Software name and version for applicability domain assessment

Name: KATE2020 on NET

Contact: kate@nies.go.jp

WWW: https://kate2.nies.go.jp

5.4. Limits of applicability

See 5.1 and 5.2.

6.Internal validation - OECD Principle 4

6.1. Availability of the training set

It is available in the supporting information.

6.2. Available information for the training set

Chemical Name: Yes

SMILES: Yes

CAS RN: Yes

InChI: No

MOL file: No

Formula: No

6.3. Data for each descriptor variable for the training set

No

6.4. Data for the dependent variable for the training set

The data for the dependent variable for each QSAR class are available in the spreadsheet "training_set" in the attachment file "KATE2020_alga_acute.xlsx".

6.5. Other information about the training set

The number of the training set for each QSAR class is described on the column named "n" in the spreadsheet "qsar_data" in the attachment file "KATE2020_alga_acute.xlsx".

6.6. Pre-processing of data before modelling

The endpoint data EC_{50} [mg/L] was converted into log $(1/EC_{50}$ [mmol/L]).

6.7. Statistics for goodness-of-fit

The following statistical values for each QSAR class are obtained and shown in the spreadsheet "qsar_data" in the attachment file "KATE2020_alga_acute.xlsx".

R^2: square of correlation coefficient

RMSE: Root Mean Squared Error

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

The following statistical value for each QSAR class is obtained and shown in the spreadsheet "qsar_data" in the attachment file "KATE2020_alga_acute.xlsx".

Q^2: square of correlation coefficient

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

No

6.10. Robustness - Statistics obtained by Y-scrambling

No

6.11. Robustness - Statistics obtained by bootstrap

No

6.12. Robustness - Statistics obtained by other methods

No

7.External validation - OECD Principle 4

7.1. Availability of the external validation set

No.

7.2. Available information for the external validation set

Chemical Name: No

SMILES: No

CAS RN: No

InChI: No

MOL file: No

Formula: 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

No

7.6. Experimental design of test set

No

7.7. Predictivity - Statistics obtained by external validation

No

7.8. Predictivity - Assessment of the external validation set

No

7.9. Comments on the external validation of the model

No

8.Providing a mechanistic interpretation - OECD Principle 5

8.1. Mechanistic basis of the model

Each QSAR class is a group of compounds with specific structure class. Structure class is defined by the information about the substructures (see 4.2) and not directly relevant to the mode of actions. There are a few classes which can be relevant to the mode of actions. For example, QSAR ID: 32899941 (in the spreadsheet of "qsar_data") is a group of chemicals with structure class: GD_28999 (narcotic group), and considered to exert narcotic actions. This structure class was created with the integration of the multiple structure classes including aliphatic hydrocarbons, aliphatic and aromatic ethers, aliphatic and aromatic ketones, and alcohols which are considered to have simple narcotic actions.

8.2. A priori or a posteriori mechanistic interpretation

A priori mechanistic interpretation.

8.3. Other information about the mechanistic interpretation

No

9.Miscellaneous information

9.1. Comments

This model can output multiple estimations based on the adequacy to the structure classes (see 4.2). There is neither recommended value nor guideline to select one right now and the users should select by themselves based on R2, Q2, the shape of the regression line (slope and the number of reference chemicals), and the applicability domain based on descriptor (log P) and structure (structure judgement).

9.2. Bibliography

[1] OECD Guideline for Testing of Chemicals, No. 201 (23rd March 2006) - Annex 5 corrected: 28th July 2011) - Freshwater Alga and Cyanobacteria, Growth Inhibition Test.

9.3. Supporting information

1. KATE2020_alga_acute.xlsx

The name and content of each spreadsheet are as follows:

substructure_def: definition of substructures

substructure_def_for_AD: definition of substructures (also used for the applicability domain)

structure_def: definition of structure classes

qsar_data: data for each QSAR class (QSAR equation etc.)

training_set: data of the training set

10.Summary (JRC QSAR Model Database)

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

	QMRF identifier (JRC Inventory): **To be entered by JRC**
	QMRF Title: QSAR Model Reporting Format Prepared for KATE (Alga Chronic)
	Printing Date:

1.QSAR identifier

1.1. QSAR identifier (title)

KATE2020 on NET for chronic toxicity to Alga

1.2. Other related models

1.3. Software coding the model

Name: KATE2020 on NET

Description: Ecotoxicity prediction system consists of QSAR models

Contact: kate@nies.go.jp

WWW: https://kate2.nies.go.jp/

2.General information

2.1. Date of QMRF

14 January 2020

2.2. QMRF author(s) and contact details

Names: Hiroshi Yamamoto, Koji Imai, Yuto Itami, Ayako Furuhama, and Koichi Ohno

Affiliation: National Institute for Environmental Studies, Japan

Contact: 16-2 Onogawa, Tsukuba, Ibaraki, 305-8506, Japan

e-mail: kate@nies.go.jp

WWW: https://www.nies.go.jp/index-e.html

2.3. Date of QMRF update(s)

2.4. QMRF update(s)

2.5. Model developer(s) and contact details

Names: Hiroaki Shiraishi and Koji Imai

Affiliation: National Institute for Environmental Studies, Japan

Contact: 16-2 Onogawa, Tsukuba, Ibaraki, 305-8506, Japan

e-mail: kate@nies.go.jp

WWW: https://www.nies.go.jp/index-e.html

2.6. Date of model development and/or publication

3 February 2020

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

(1) KATE2020 on NET https://kate2.nies.go.jp/

(2) A Furuhama, T Toida, N Nishikawa, Y Aoki, Y Yoshioka, and H Shiraishi (2010). Development of an ecotoxicity QSAR model for the KAshinhou Tool for Ecotoxicity (KATE) system, March 2009 version. SAR and QSAR Environmental Research 21 (5-6) 403–413 http://dx.doi.org/10.1080/1062936X.2010.501815.

(3) A. Furuhama, K. Hasunuma, Y. Aoki, Y. Yoshioka and H. Shiraishi (2011). Application of chemical reaction mechanistic domains to an ecotoxicity QSAR model, KAshinhou Tool for Ecotoxicity (KATE). SAR and QSAR Environmental Research 22 (5-6) 505–523 http://dx.doi.org/10.1080/1062936X.2011.569944.

2.8. Availability of information about the model

The model is proprietary, but the training set and algorithm are available.

2.9. Availability of another QMRF for exactly the same model

None to date.

3.Defining the endpoint - OECD Principle 1

3.1. Species

Pseudokirchneriella subcapitata

3.2. Endpoint

Short-term toxicity to alga (inhibition of the exponential growth rate)

3.3. Comment on endpoint

The model predicts the no-observed-effect concentration (NOEC) values in the Alga Growth Inhibition Test (72-h exposure).

3.4. Endpoint units

mg/L

3.5. Dependent variable

All of the measured endpoint values were converted to mmol/L units, replaced with their reciprocal, and transformed to logarithmic units. The dependent variable is: log (1/NOEC)

3.6. Experimental protocol

OECD Test Guideline No. 201 test protocol

3.7. Endpoint data quality and variability

The data are from the ecotoxicity tests conducted by the Ministry of the Environment, Government of Japan (Alga growth inhibition tests) according to good laboratory practice (GLP).

This data are available at: http://www.env.go.jp/en/chemi/sesaku/aquatic_Mar_2019.pdf

(Accessed on January 14, 2020)

4.Defining the algorithm - OECD Principle 2

4.1. Type of model

QSAR

4.2. Explicit algorithm

The predicted values are calculated with the following algorithm:

(1) Extract the information about the number of substructures, which we defined, in the target compound. (As for the definition of the substructures, see the spreadsheets "substructure_def" and "substructure_def_for_AD" in the "KATE2020_alga_chronic.xlsx").

(2) Extract the adequate structure classes in the target compound from the information about the number of the substructure (see 4.5. (3)).

(3) Predict the ecotoxicity of the target compound using the descriptor (log P) and the QSAR equation (slope and intercept) provided for each structure class.

(4) Convert the unit of the predicted ecotoxicity from log (1/NOEC [mmol/L]) to [mg/L] using the molecular weight (see 3.5).

4.3. Descriptors in the model

Descriptor: log P   Octanol–water partition coefficient

For structural fragments, see the spreadsheets "substructure_def", "substructure_def_for_AD" and "structure_def" in the attachment file "KATE2020_alga_chronic.xlsx".

4.4. Descriptor selection

See 4.2.

4.5. Algorithm and descriptor generation

(1) Log P can be input by the user, otherwise log P is calculated using KOWWIN from SMILES notation.

(2) The information about the number of substructures (see 4.2. (1)) is calculated using the JAVA application developed by NIES using Chemistry Development Kit (CDK) from inputted SMILES and the definition of the substructure.

(3) The adequacy to each structure class (1:Yes, 0:No) is evaluated by the information about the number of the substructures and the definition of the structure class (see the spreadsheet "structure_def" in the file "KATE2020_alga_chronic.xlsx") in the KATE2020 on NET.

(4) Molecular weight of the target compound is calculated from SMILES by using Open Babel.

4.6. Software name and version for descriptor generation

Name: KOWWIN

Description: A program to estimate log P values of organic compounds. That is incorporated in EPI Suite™ (Estimation Programs Interface: a tool intended for use in applications such as to quickly screen chemicals), which was developed by the US EPA, etc.

KOWWIN v1.69 (April 2015)

KOWWIN™ (included in EPI Suite™)
• https://www.epa.gov/tsca-screening-tools/epi-suitetm-estimation-programinterface
Name: None
Description: An application for SMARTS substructure searching from SMILES that we have developed by using Chemistry Development Kit (CDK) version 2.0
Contact: kate@nies.go.jp
WWW: https://kate2.nies.go.jp/

Name: Open Babel
Description: The Open Source Chemistry Toolbox, version 2.3.1, 2011.
Contact: Contact not available
WWW: http://openbabel.org/wiki/Main_Page

4.7. Chemicals/Descriptors ratio

The ratio (column name: ratio1) for each QSAR class is described in the spreadsheet "qsar_data" in the attachment file "KATE2020_alga_chronic.xlsx".

5.Defining the applicability domain - OECD Principle 3

5.1. Description of the applicability domain of the model

(1) Applicability domain of the model for the descriptor (log P) is established for each QSAR class, which we define as a group of compounds with specific structure classes, and is the range of the descriptor, "between the maximum and the minimum" (see the column named "Log P Range" of the spreadsheet "qsar_data" in the attached file "KATE2020_alga_chronic.xlsx").

(2) Applicability domain of the model for the structure is also established for each QSAR class, and is provided from the list of substructure for the structure judgement of the chemicals in the training set which belongs to the QSAR class and in the class called "narcotic group" (see the column "FragIDs_for_AD" of the spreadsheet "qsar_data" in the attached file "KATE2020_alga_chronic.xlsx")

5.2. Method used to assess the applicability domain

The applicability domain of both descriptor and structure is assessed for each QSAR class into which the target compound is classified.

(1) If the log P value of the target compound is within the applicability domain of the descriptor as indicated in 5.1.(1), it is judged as "inside of the applicability domain" in terms of the descriptor.

(2) If the substructure for the structure judgement (see the substructure defined in the spreadsheet "substructure_def_for_AD" in the file "KATE2020_alga_chronic.xlsx") possessed by the target compound is all included in the list of the structure judgement of the QSAR class (see 5.1.(2)), it is judged as "inside of the applicability domain" in terms of the structure.

For more details, see Furuhama et al. (2011) (see 2.7 (3)).

5.3. Software name and version for applicability domain assessment

Name: KATE2020 on NET

Contact: kate@nies.go.jp

WWW: https://kate2.nies.go.jp

5.4. Limits of applicability

See 5.1 and 5.2.

6.Internal validation - OECD Principle 4

6.1. Availability of the training set

It is available in the supporting information.

6.2. Available information for the training set

Chemical Name: Yes

SMILES: Yes

CAS RN: Yes

InChI: No

MOL file: No

Formula: No

6.3. Data for each descriptor variable for the training set

No

6.4. Data for the dependent variable for the training set

The data for the dependent variable for each QSAR class are available in the spreadsheet "training_set" in the attachment file "KATE2020_alga_chronic.xlsx".

6.5. Other information about the training set

The number of the training set for each QSAR class is described on the column named "n" in the spreadsheet "qsar_data" in the attachment file "KATE2020_alga_chronic.xlsx".

6.6. Pre-processing of data before modelling

The endpoint data NOEC [mg/L] was converted into log (1/NOEC [mmol/L]).

6.7. Statistics for goodness-of-fit

The following statistical values for each QSAR class are obtained and shown in the spreadsheet "qsar_data" in the attachment file "KATE2020_alga_chronic.xlsx".

R^2: square of correlation coefficient

RMSE: Root Mean Squared Error

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

The following statistical value for each QSAR class is obtained and shown in the spreadsheet "qsar_data" in the attachment file "KATE2020_alga_chronic.xlsx".

Q^2: square of correlation coefficient

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

No

6.10. Robustness - Statistics obtained by Y-scrambling

No

6.11. Robustness - Statistics obtained by bootstrap

No

6.12. Robustness - Statistics obtained by other methods

No

7.External validation - OECD Principle 4

7.1. Availability of the external validation set

No.

7.2. Available information for the external validation set

Chemical Name: No

SMILES: No

CAS RN: No

InChI: No

MOL file: No

Formula: 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

No

7.6. Experimental design of test set

No

7.7. Predictivity - Statistics obtained by external validation

No

7.8. Predictivity - Assessment of the external validation set

No

7.9. Comments on the external validation of the model

No

8.Providing a mechanistic interpretation - OECD Principle 5

8.1. Mechanistic basis of the model

Each QSAR class is a group of compounds with specific structure class. Structure class is defined by the information about the substructures (see 4.2) and not directly relevant to the mode of actions. There are a few classes which can be relevant to the mode of actions. This structure class was created with the integration of the multiple structure classes including aliphatic hydrocarbons, aliphatic and aromatic ethers, aliphatic and aromatic ketones, and alcohols which are considered to have simple narcotic actions.

8.2. A priori or a posteriori mechanistic interpretation

A priori mechanistic interpretation.

8.3. Other information about the mechanistic interpretation

No

9.Miscellaneous information

9.1. Comments

This model can output multiple estimations based on the adequacy to the structure classes (see 4.2). There is neither recommended value nor guideline to select one right now and the users should select by themselves based on R2, Q2, the shape of the regression line (slope and the number of reference chemicals), and the applicability domain based on descriptor (log P) and structure (structure judgement).

9.2. Bibliography

[1] OECD Guideline for Testing of Chemicals, No. 201 (23rd March 2006) - Annex 5 corrected: 28th July 2011) - Freshwater Alga and Cyanobacteria, Growth Inhibition Test.

9.3. Supporting information

1. KATE2020_alga_chronic.xlsx

The name and content of each spreadsheet are as follows:

substructure_def: definition of substructures

substructure_def_for_AD: definition of substructures (also used for the applicability domain)

structure_def: definition of structure classes

qsar_data: data for each QSAR class (QSAR equation etc.)

training_set: data of the training set

10.Summary (JRC QSAR Model Database)

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