KATE2020 Ecotoxicity Prediction System

Technical Document (Ver. 2024/03/25)

※ KATE2020 is a system to predict the following ecotoxicity of organic chemicals:

- 50% lethal concentration (LC50) in the fish acute toxicity test
- 50% effective concentration (EC50) in the *Daphnia magna* acute immobilization test
- 50% effective concentration (EC50) in the algal growth inhibition test
- No-observed-effect concentration (NOEC) in the fish early-life-stage toxicity test
- No-observed-effect concentration (NOEC) in the *Daphnia magna* reproduction test
- No-observed-effect concentration (NOEC) in the algal growth inhibition test

* Values predicted by KATE2020 cannot be used to satisfy the requirements for reporting ecotoxicity tests under the Law Concerning Examination of Chemical Substances and Regulation of Manufacturing, etc.

*Please use the predicted values as reference values for ecotoxicological effects of chemical substances.

Enquires should be directed to the email address indicated below.

Health and Environmental Risk Division, National Institute for Environmental Studies,

KATE Contact Desk

kate@nies.go.jp

KATE2020 Manual Revision History

Version	Date of issue	Revision history
0.99	June 28, 2023	KATE 2020 (Version 3.0) Technical Document
0.99.1	March 30, 2023	KATE 2020 (Version 4.0) Technical Document
0.99.2	March 25, 2024	KATE 2020 (Version 5.0) Technical Document

Table of Contents

Appendix

- Substructures
  https://kate3.nies.go.jp/nies/substructures.php

- Structure Classes
  https://kate3.nies.go.jp/nies/structure_classes.php

- QSAR Classes
  https://kate3.nies.go.jp/nies/qsar_classes.php

List of abbreviations

CDK: Chemistry Development Kit

A collection of free and open-source Java libraries for processing cheminformatics and bioinformatics. The kit can be used to search for the substructures of chemical substances and calculate their structures and physical properties[1].

EC50: 50% Effective Concentration

The concentration of a substance dissolved in test water expected to produce a certain effect in 50% of test organisms in a given population under a defined set of conditions.

KATE: KAshinhou Tool for Ecotoxicity

The ecotoxicity QSAR system researched and developed by the Health and Environmental Risk Division of the National Institute for Environmental Studies. It is pronounced as in the feminine name Kate.

KOWWIN™:

A program to estimate log P values of organic compounds included in the EPI Suite™ (Estimation Programs Interface: a tool intended for use in applications such as to quickly screen chemicals) developed by the U.S. EPA and its partners[2].

LC50: 50% Lethal Concentration

The concentration of a substance dissolved in test water which, on the basis of laboratory tests, is expected to kill 50% of a group of test species when administered as a single exposure.

log P: Logarithm of the octanol/water partition coefficient (Octanol/water partition coefficient)

A common logarithm of the ratio of concentrations of a chemical substance between 1-octanol and water in equilibrium. It represents the hydrophobicity of a chemical substance and does not account for ionization of the chemical substance in question[3].

NOEC: No Observed Effect Concentration

The highest tested concentration (also called maximum no-effect concentration) for which there is no statistically significant difference of effect ($p<0.05$) when compared to the control group. This concentration range is just below the LOEC (Least Observed Effect

Concentration) (reference: Japanese Ministry of the Environment, Environmental Risk Assessment of Chemical Substances Vol. 17, Chapter 1, Reference 2 Glossary [4])

OECD: Organisation for Economic Co-operation and Development

(Q) SAR: (Quantitative) Structure–Activity Relationships

The relationship between the structural characteristics or the physicochemical constant of a chemical substance and its biological activities (e.g., toxicity) is called the Structure–Activity Relationship (SAR), while the quantitative relationship is called the Quantitative Structure–Activity Relationship (QSAR). Both may be referred to collectively as (Q)SAR. SAR refers to, for example, an estimation of the toxicity level of a chemical based on the presence of a specific functional group. A model to quantitatively calculate the toxicity or other properties of a chemical based on the structure is called a QSAR model. (reference: Japanese Ministry of the Environment, Environmental Risk Assessment of Chemical Substances Vol. 17, Chapter 1, Reference 2 Glossary [4])

SMARTS: SMiles ARbitrary Target Specification

An identifier extended from SMILES to describe substructures.[5]

SMILES: Simplified Molecular Input Line Entry System

An identifier of line notation with printable characters to describe the molecular structures, etc., of chemical compounds.[6]

U.S. EPA: United States Environmental Protection Agency

1. Introduction

1.1 KATE (KAshinhou Tool for Ecotoxicity) System

KATE (kate.nies.go.jp) is an ecotoxicity QSAR system for which research and development was commissioned by the Japanese Ministry of the Environment (MoE) to the Health and Environmental Risk Division of the National Institute for Environmental Studies. KATE is a system to predict ecotoxicity based on the structure of chemical substances. The ecotoxicity values predicted by KATE2020 are as follows:

- 50% lethal concentration (LC50) in acute toxicity test for fish (OECD TG 203)[7]
- No-observed-effect concentration (NOEC) in fish early-life-stage toxicity test (OECD TG 210)
- 50% effective concentration (EC50) in the *Daphnia magna* acute immobilization test (OECD TG 202)[9]
- No-observed-effect concentration (NOEC) in the *Daphnia magna* reproduction test (OECD TG 211)[10]
- 50% effective concentration (EC50) and no-observed-effect concentration (NOEC) in the algal growth inhibition test (OECD TG 201)[11]
-

Data is input using SMILES notation, which can be obtained by CAS number search, or drawing using a chemical structure editor, in order to carry out a QSAR prediction by using a regression equation with log P as the descriptor.

The QSAR model for KATE2020 was constructed using ecotoxicity data obtained from tests conducted by the Japan MoE[12] (fish acute toxicity test, *Daphnia magna* acute immobilisation test, fish early-life-stage toxicity test, *Daphnia magna* reproduction test, and algal growth inhibition test), together with fish acute toxicity data from the U.S. EPA fathead minnow database[13].

1.2 Purpose of KATE 2020 Technical Document

This technical document explains the derivation of the KATE2020 QSAR model and its performance evaluation based on OECD QSAR validation principles[14]. Its target reader is those who have experience using KATE2020.

Please refer to the KATE2020 Operation Manual (https://kate2.nies.go.jp/nies/doc/KATEmanual_2020-e.pdf) for instructions on using KATE2020, its development history, and its update history.

1.3 OECD QSAR Validation Principles

The technical document describes the KATE2020 QSAR model in terms of the five OECD QSAR validation principles[14].

The QSAR model must satisfy the five OECD QSAR validation principles in order to ensure validity and reliability when applying the model to chemical substance regulations. They are as follows:

1. Endpoint definition
2. Unambiguous algorithms
3. Defined domain of applicability
4. Appropriate assessment of goodness-of-fit, robustness and predictivity
5. If possible, explanation of mechanism

1.4 log P

This system utilizes the log P prediction model KOWWIN™[5] copyrighted by the U.S. EPA, with its permission, in order to obtain log P values used for predicting the toxicity of chemical substances.

Users must comply with the KOWWIN™ licensing terms described below.

KOWWIN v1.69 (April 2015)

c 2000-2015 U.S. Environmental Protection Agency

KOWWIN is owned by the U.S. Environmental Protection Agency and is protected by copyright throughout the world.

Permission is granted for individuals to download and use the software on their personal and business computers.

> Users may not alter, modify, merge, adapt or prepare derivative works from the software. Users may not remove or obscure copyright, tradename, or proprietary notices on the program or related documentation.
>
> KOWWIN contained therein is a tradename owned by the U.S. Environmental Protection Agency.

1.5 Disclaimer

The accuracy of results predicted by the KATE system are not guaranteed. Please utilize this system as a tool to obtain reference information on the degree of ecotoxicological effects of chemical substances. The Ministry of the Environment and the National Institute for Environmental Studies do not guarantee the predicted ecotoxicity values provided by KATE and assume no responsibility whatsoever for any damages resulting from the use of ecotoxicity values predicted by KATE.

Further, values predicted by this system cannot currently be used to satisfy the requirements for reporting ecotoxicity tests under the Japanese Law Concerning Examination of Chemical Substances and Regulation of Manufacturing, etc.

For copyright information and instructions for linking to the site, please visit the Website Policy page in the KATE website (kate.nies.go.jp/spolicy-e.html).

1.6 Acknowledgements

KATE2020 uses results obtained from the following software and libraries. Here, we express appreciation to all who developed them.

- ☐ Open Babel[15]
- ☐ JSME Molecular Editor[16, 17]
- ☐ CDK (Chemistry Development Kit)[1, 18–21]
- ☐ KOWWIN™ (included in EPI Suite™)[2]

2. Endpoint Definition – OECD (Q) SAR Validation Principle 1

Here, we define the endpoints predicted by KATE2020.

2.1 Endpoints Predicted by KATE2020

KATE2020 predicts acute and chronic toxicities using the endpoints (toxicity indicators) shown in the following table to estimate the ecotoxicities of chemical substances.

Table 2-1: Endpoints predicted by KATE2020.

Predicted Toxicity Type		Species (Scientific Name)	Testing Method	Duration	Indicator
Organisms	Acute/ Chronic				
Fish	Acute	*Oryzias latipes*, *Pimephales promelas* *1	Fish acute toxicity test (OECD TG 203) [7]	96 h	LC50
Daphnid	Acute	*Daphnia magna*	Daphnia magna immobilization test (OECD TG 202) [9]	48 h	EC50
Alga	Acute	*Raphidocelis subcapitata* *2	Algal growth inhibition test (OECD TG 201) [11]	72 h	EC50
Fish	Chronic	*Oryzias latipes*	Fish early-life-stage toxicity test (OECD TG 210) [8]	Embryonic stage and 30 days after hatching *3	NOEC
Daphnid	Chronic	*Daphnia magna*	Daphnia magna reproduction test (OECD TG 211) [10]	21 d	NOEC
Alga	Chronic	*Raphidocelis subcapitata* *2	Algal growth inhibition test (OECD TG 201) [11]	72 h	NOEC

*1 If toxicity values are available for both Japanese medaka and fathead minnow, the Japanese medaka is used. When KATE was initially developed, MoE Japanese medaka test data alone were insufficient in terms of the number of substances covered to establish a predictive model exhibiting sufficient precision. Therefore, U.S. EPA fathead minnow test data were also employed and differences in predictions arising from interspecies differences in fish species acute QSAR traceable to Japanese medaka and fathead minnow were investigated. The findings confirmed that by establishing a QSAR formula using fathead minnow toxicity values, higher precision could be obtained compared to using data for Japanese medaka alone. Furthermore, for substances for which measured toxicity values for both Japanese medaka and fathead minnow could be obtained, the only substance for which the values differed by more than a factor of ten was dimethylamine (Studies on Regulatory based Assessment of Chemicals under Japanese Chemical Law, Fiscal 2009 Edition).

*2 May have been referred to previously as *Selenastrum capricornutum* or *Pseudokirchneriella subcapitata*, etc.

*3 Test periods for fish species early-life stage tests differ depending on the type of fish and incubation period. In the case of Japanese medaka used in ecotoxicity tests conducted by the MoE, the test period from the embryo stage to 30 days post-hatch is used.

2.2 Predicted Value Units

KATE outputs predicted values for toxicity indicators in units of mg/L.

2.3 Dependent Variable

KATE converts the toxicity indicator values from units of mg/L to units of mmol/L and uses the log of the reciprocal, i.e., log (1/toxicity value[mmol/L]) as the dependent variable.

2.4 Training Data Set Endpoints

The KATE2020 training data set toxicity values are based on the findings of ecotoxicity tests carried out by the Japan MoE (fish acute toxicity test, *Daphnia magna* acute immobilisation test, fish early-life stage toxicity test, *Daphnia magna* reproduction test, algal growth inhibition test), and U.S. EPA fathead minnow database fish species acute toxicity test results.

Number of Substances

Table 2-2 lists the number of substances in KATE2020 for each predicted toxicity type that are part of the training data set*1, support chemicals*2 and substances not assigned with any QSAR class*3 (the number of substances in KATE2020 for all QSAR classes. A single substance is counted as one even if it is assigned with multiple QSAR classes).

Table 2-2: Number of substances for each predicted toxicity type

		Acute			Chronic		
		Fish	Daphnid	Alga	Fish	Daphnid	Alga
Training Set*1	MOE Data	361	436	315	32	316	401
	US EPA Data	498					
Support Chemical*2	MOE Data	206	135	214	0	54	115
	US EPA Data	1					
Chemical that is not classified into any of QSAR classes of KATE2020*3	MOE Data	8	16	31	0	12	19
	US EPA Data	8					
Total		1095	1082	487	560	32	382

*1 Substances used to formulate QSAR class regression formulae.

*2 Data with inequality sign (limit tests etc.) or outliers (toxicity test data whose reliability cannot be verified etc.), and substances for which log P is greater than six are not included in QSAR class regression formulae. Further, log P values estimated by KOWWIN™ are used for all substances.

*3 Substances that do not correspond to any KATE2020 QSAR class are designated as “Unclassified class” when predictions are being made.

3. Algorithm – OECD (Q)SAR Validation Principle 2

Here, we describe the KATE2020 algorithm. Section 3.1 presents an outline, while the subsequent sections go into more detail.

3.1 Algorithm Outline

KATE2020 is a linear regression-based QSAR system that predicts the toxicity of chemical substances (organic compounds).

The bio membrane permeability of chemical substances, their bioaccumulation, and toxicity are thought to be correlated and based on this assumption, log P is used as a descriptor.

The Japanese MoE's ecotoxicological effect test data and the U.S. EPA's fathead minnow acute test data are used as training data set. Classification is carried out based on a decision tree with the feature values of the presence and number of substructures contained in each substance.

KATE2020 employs 18 major categories and the other categories as shown in Table 3–1. Each category has a characteristic substructure, while substances are also categorized into reactive and unreactive classes based on the presence or absence of substructures considered to be highly reactive and highly toxic. Further conditional branches (presence of halogens, presence of aromatic atoms, etc.) are used for segmentation of structure classes (refer to Chapter 3, Section 3.5). QSAR classes (refer to Chapter 3, Section 3.6) classes are assigned to structure classes and regression equations are formulated for each QSAR class. Some QSAR classes are classified as being specifically for toxicity prediction for specific groups of organisms. Substances possessing substructures with low reactivity that only exhibit narcotic effects are categorized as the narcotic group class. Substructures exhibiting specific effects such as skin sensitization[22] are primarily used to evaluate the applicability domain of structures (refer to Chapter 4, Section 4.2 A)[23], although some are used for classification.

Molecular chemical structure data for substances to be predicted are inputted using the alphanumeric SMILES notation[6], while substructure/s and their number/s are obtained by using the CDK cheminformatics tool CDK[1]. Substructures are defined beforehand using SMARTS notation[5], which is an extended notation of SMILES. Please refer to documentation on the Internet for a list of substructures (https://kate2.nies.go.jp/nies/substructures.php).

QSAR Class assignment is done according to extracted substructures based on the aforementioned decision tree. Multiple classes may be assigned per predicted toxicity

type (if the query chemical contains substructures that do not match any classes existing in KATE, the chemicals is assigned to Unclassified class and toxicities are not predicted). Subsequently, the predicted toxicity value for each class is generated based on log P estimated by KOWWIN™[2]. While the log P value can be manually entered by the user, log P values of all training set chemicals are estimated by KOWWIN™.

Table 3–1: KATE2020 Major Classification List

	Classification Category	Substructure ID	Substructure Name	SMARTS
1	acid	3034	carboxylic acid C(=O)O	[#6;$([#6](=[#8])([#6])[#8H1])]
		4760	-SO3H, Sulfonic Acid, sulfo-, -sulfonic a	[C,c,O]S(=O)(=O)[O;$([OH1]),$(O[Na,Li,K])]
2	alcohol	3046	alcohol COH	[#6;$([#6][#8;H]);!X3;v4]
3	aldehyde, ketone	3031	ketone CC(=O)C	[#6;$([#6](=[#8])([#6])[#6])]
		3036	aldehyde	[#6H1;$([#6](=[#8])[#6])]
4	ester	3032	ester CC(=O)OC	[#6;$([#6](=[#8])([#6])[#8][#6]);!$([#6](=[#8])([#6])[#8][#6]=[O,S,N])]
		3145	acetal	[#6X4;$([#6]([#8])[#8])]
5	ether	3044	ether general	[#8H0;!$([#8]C=[O,N,S]);!$([#8]C[#8]);$([#8]([#6])[#6])]
6	phenol	3047	phenol cOH	[OX2H][cX3;$(c1ccccc1)]
7	amine primary	3100	amine CNH2	[#7X3H2;!$([#7][*v6]);!$(N[#6](~[#7,#8,#16]))]
8	amine sec, tert	3110	amine CNH1	[#7v3X3H1;!$([#7][!#6]);!$([#7][*v6]);!$([#7][#6](~[#7,#8,#16]));!$([#7r5H1](a)a)]
		3120	amine CNH0	[#7v3X3H0;!$([#7][!#6]);!$([#7][*v6]);!$([#7][#6](~[#7,#8,#16]))]
9	aromatic n	4911	aromatic n	[n]
10	hydrazine	3210	NN, hydrazine general, not in ring	[Nv3X3R0;$([N][Nv3X3])]
11	nitrile	3104	nitril C#N	[N;$([#7X1]#[#6X2])]
12	amide	3123	amide	[#7v3X3;$([#7][C](=O)[!N;!S;!O;!P]),$([#7][CH1](=O));!$(N~[N,S,O]);!$(N(C=[N,O,S])C=O)]
13	carbamate	3041	carbamate general NC(=O,S)O,S	[#7v3X3;$([#7][#6R0](=[O,S])[O,S])]
14	nitro	3231	nitro aromatic	[N;$(N(c)(=O)=O),$([N+](c)(=O)[O-])]
15	phosphorus	5018	Phosphorus [P]	[#15]
16	sulfur	5016	Sulfur [S]	[#16]
17	halogen	4507	halogen	[F,Cl,Br,I]
18	heteroatomatic	4911	aromatic n	[n]
		4912	aromatic o	[o]
		4913	aromatic s	[s]
	Others	3106	azo N=N	[NX2;$(N=N)]
		4541	epoxide monocyclic	[#8r3;$([#8]1[#6R1][#6R1])]
		3108	imino C=N-, guanidine	[#7v3X2;$(N=[Cv4X3]);!$(N[N,O,S])]
		etc		
		.		
		.		
		.		

	contains oxygen
	contains nitrogen
	contains oxygen and nitrogen
	contains phoshorus
	contains sulfur
	others

The toxicity prediction flow is shown below (refer to Figure 3–1).

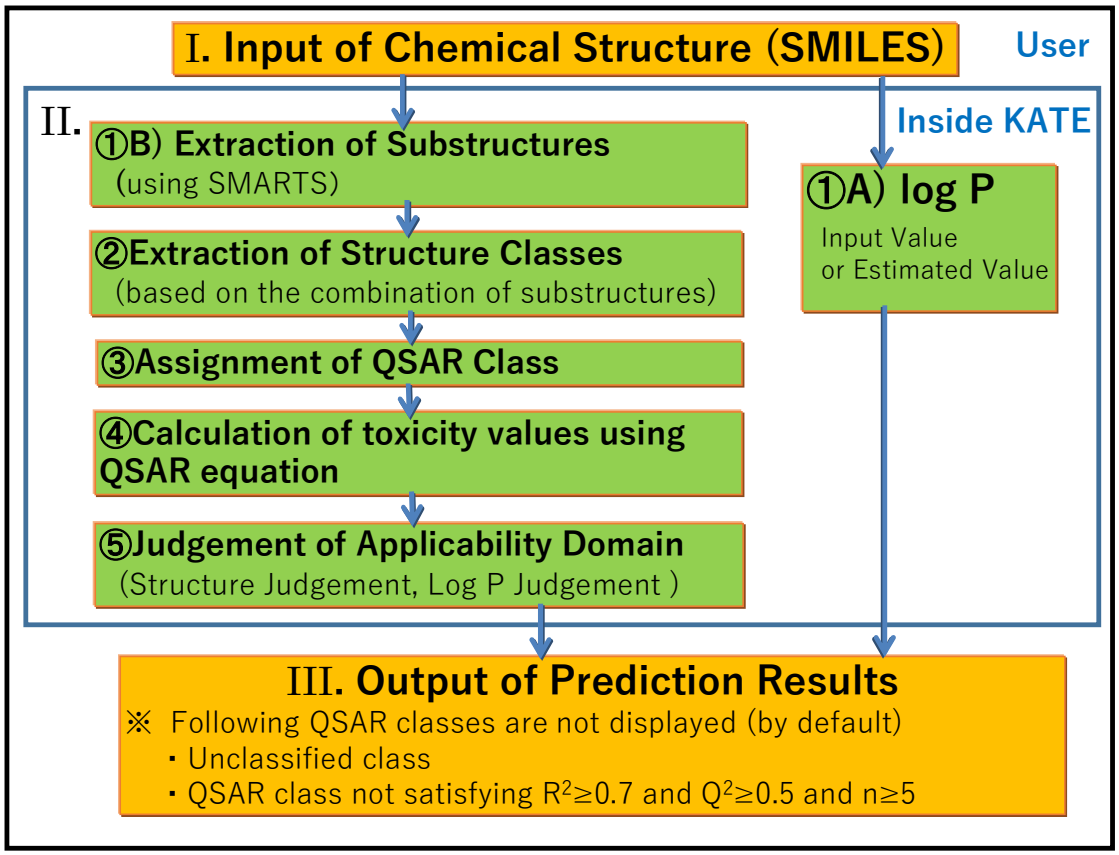

Figure 3-1: KATE2020 toxicity prediction flow

Ⅰ. The user inputs the structure of the chemical substance (using SMILES notation) and the log P value (optional).

Ⅱ. Processes on the QSAR equation assignment, predicted toxicity values, and the applicability domain judgement.

① For the query chemical (entered chemical substance)

A) log P value is estimated (or the input value is used).

B) substructure/s is/are extracted.

② Structure class[*1] is extracted based on the combination of the extracted substructures.

③ For each predicted toxicity type, a QSAR class[*2] corresponding to the structure class is assigned (multiple classes may be assigned).

④ For each assigned QSAR class, the toxicity value is calculated using the QSAR equation[*3].

⑤ The applicability domain is judged (log P judgement and structure judgement).

*1 Classification defined by a combination of each substructure number condition AND/OR (Section 3.5 Structure class extraction)

*2 Classification defined based on the structure of the chemical for each predicted toxicity type.

*3 Model constructed based on training set data included in the QSAR class. Here log P is used as the descriptor for a simple regression equation.

Ⅲ. Prediction Output

If the query chemical cannot be assigned to any of QSAR classes for a predicted toxicity type, it is assigned to the unclassified class.
By default, the unclassified class and QSAR classes that do not satisfy statistical criteria ($R^2 \geq 0.7$, $Q^2 \geq 0.5$ and $n \geq 5$)*4 are not shown.

*4 R^2, Q^2, and n are the coefficient of determination, the internal validation indicator (Leave-one-out method), and the number of data points in the training set, respectively, and they are calculated beforehand for each QSAR class.

As an actual example, Fig. 3–2 shows the prediction flow of the chemical substance 1-pyridin-3-ylethanone.

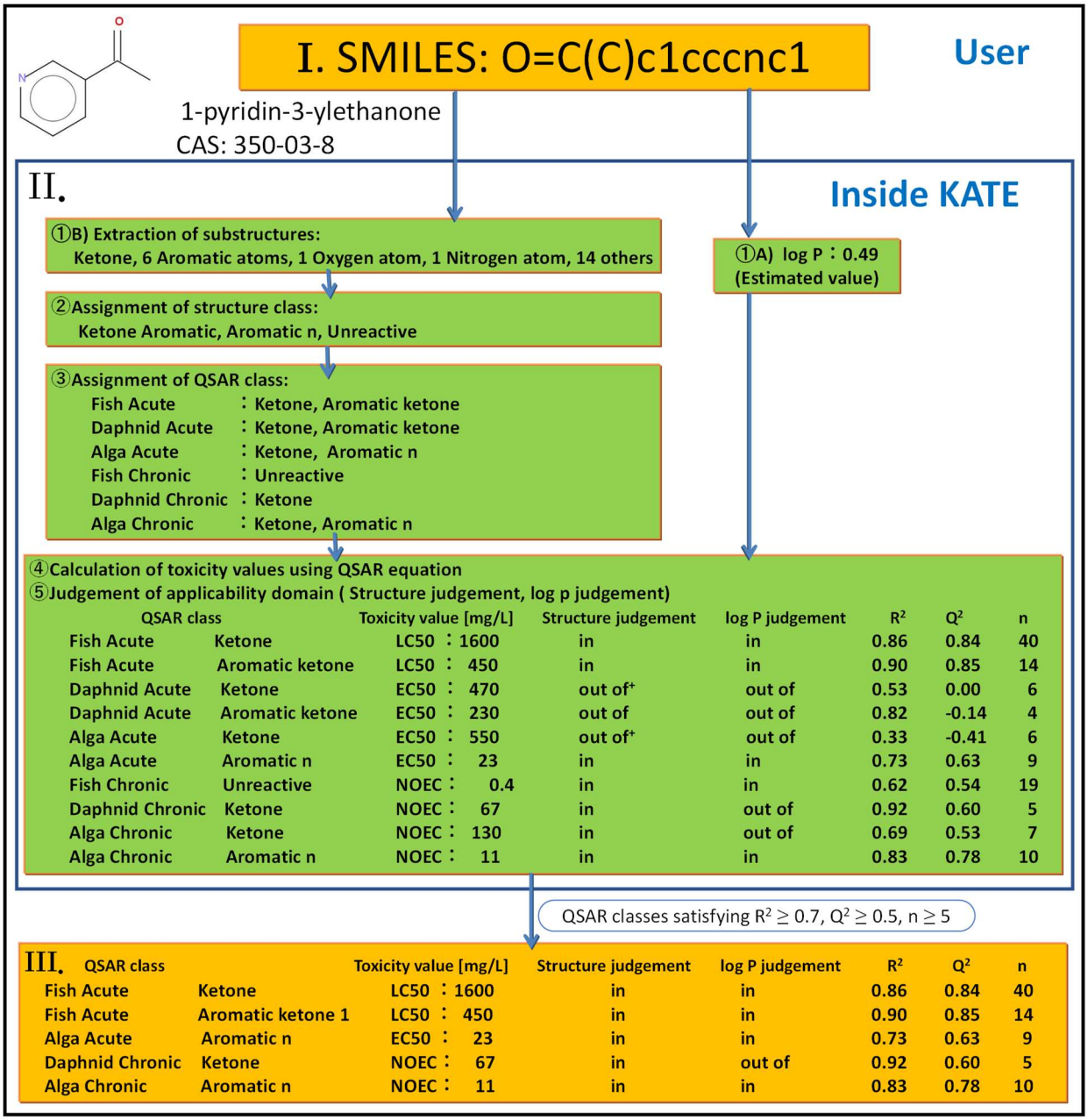

Figure 3–2: KATE2020 QSAR prediction flow (example)

※ The actual names of the Ketone, Aromatic ketone, Aromatic n and Unreactive in Figure 3–2 used in KATE2020 are as follows.

Ketone: COns_X ketone unreactive

Aromatic Ketone: COS_X ketone unreactive aromatic

Aromatic n: CNOS_X aromatic n unreactive

Unreactive: CNO_X unreactive Fish chronic, w/ N,O

In the following, we explain the flow of toxicity predictions in KATE2020.

3.2 Input of Chemical Substance(s)

The user either inputs SMILES directly, or converts a drawn structural formula, the CAS number, or the substance name to SMILES.

3.3 Obtaining log P Values

The log P value for the query chemical used to calculate the predicted toxicity value is determined by the following order of priority.

1. log P value entered by user (when user enters a value)
2. log P predicted value (when no value is entered by user)

KOWWIN™[2], licensed by the U.S. EPA, is used to estimate the log P value (refer to Chapter 1, Section 1.6).

3.4 Extraction of Substructures

The number of substructures contained in the query chemical is calculated based on the substructure definition list (Table 3–2). Each row defines a single substructure in terms of SMARTS notation (Table 3–3).

Table 3–2: Substructure list (partial)

Substructure ID	Substructure Name	SMARTS
…		
3031	Ketone	[#6;$([#6](=[#8])([#6])[#6])]
3032	Ester	[#6;$([#6](=[#8])([#6])[#8][#6]);!$([#6](=[#8])([#6])[#8][#6]=[O,S,N])]
3033	Carbonate	[#6;$([#6](=[#8])([#8][#6])[#8][#6])]
…		

Refer to the following link for a list of all substructures.

https://kate2.nies.go.jp/nies/substructures.php

The CDK library[1] is used to calculate the number of substructures using SMARTS.

Table 3–3: List of substructures contained in substance to be predicted

(SMILES: Case of O=C(C)c1cccnc1)

Substructure ID	Substructure Name	SMARTS	Count
3001	elements other than CX	[!#6;!#9;!#17;!#35;!#53]	2
3002	elements other than CNX	[!#6;!#7;!#9;!#17;!#35;!#53]	1
3003	elements other than COX	[!#6;!#8;!#9;!#17;!#35;!#53]	1
3004	elements other than CSX	[!#6;!#16;!#9;!#17;!#35;!#53]	2
3009	elements other than COSX	[!#6;!#8;!#16;!#9;!#17;!#35;!#53]	1
3014	elements other than CnosX	[$([!#6;!F;!Cl;!Br;!I;!n;!s;!o]),$([n+])]	1
3022	Carbon	[#6]	7
3030	carbonyl C=O	[#6;$([#6](=[#8]))]	1
3031	ketone CC(=O)C	[#6;$([#6](=[#8])([#6])[#6])]	1
3059	C=O w/o electron donated o-, p-Nv3X3	[C;$(C=O);!$(C(=O)c1c([Nv3X3])ccc c1);!$(C(=O)c1ccc([Nv3X3])cc1)]	1
4504	>C=O or >C=S (sPilot4)	[CX3]=[OX1,SX1]	1
4543	MF:not C,c,O,F	[!C;!c;!O;!F]	1
4892	MF: not CHO (kPilotO)	[!C;!c;!O]	1
4893	MF: not CHOP	[!C;!c;!O;!P]	1
4910	aromatic	[a]	6
4911	aromatic n	[n]	1
5007	Nitrogen [N,n]	[#7]	1
5008	Oxygen [O,o]	[#8]	1

Substructure ID's beginning with "5" indicate substructures that are also used for structure judgement (Refer to Chapter 4, Section 4.1 A, Structure Applicability Domain).

3.5 Assignment of Structure Class

Based on the structure class definition (Figure 3-3 shows the structure class definition for ketone as an example), assign a structure class that matches the query substance. For example, O=C(C)c1cccnc1 has one or more substructure 3031 from Table 3-3 but does not match the conditions for R_00001 (with any of substructures 3036, 3053, 3054, 3055, 3056, 3174, 4515, 4791, 6112) and G1_00010 (with one of the substructures 3034, 4760), so it is determined to be ketone unreactive. And as it does not have substructure 3011, it is assigned a structure class of G1_21025. In total, four structure classes are assigned for this predicted substance, as shown in Table 3-4.

Structure ID	Description	Decision tree
-	ketone CC(=O)C	ID:3031 > 0
-	ketone reactive	├ R_00001 = true
G1_21028	CNO_X ketone reactive	│ └ ID:3006 = 0
G1_42027	CNO_X aldehyde, ketone	│ └ ID:3173 > 0
-	ketone unreactive	└ R_00001 = false and G1_00010 = false
G1_21025	COns_X ketone unreactive	└ ID:3011 = 0
G1_21029	COS_X ketone unreactive aromatic	├ ID:4910 > 0
G1_21027	COns_X ketone unreactive aliphatic	└ ID:4910 = 0

Figure 3–3 Definition of structure class for ketone

Please refer to the following link for the list of structure classes.

https://kate2.nies.go.jp/nies/structure_classes.php

Clicking on a value in the Decision tree column of a structure class definition displays the definition of the structure class and substructure in question. Figure 3-4 shows an example of clicking on "R_00001 = false and G1_00010 = false".

R_00001 (ID:3036 > 0 or ID:3053 > 0 or ID:3054 > 0 or ID:3055 > 0 or ID:3056 > 0 or ID:3174 > 0 or ID:4515 > 0 or ID:4791 > 0 or ID:6112 > 0)

FragID	Substructure Name	Count	SMARTS
3036	aldehyde	0	[#6H1;$([#6](=[#8])[#6])]
3053	carbonyl a-halo, ab-unsaturated	0	[$([#6R0](=O)[#6H1,#6H2][Cl,Br,I,F]),$([#6](=O)[#6]=[#6;H2]),$([#6](=O)[#6]=[#6;H1][a]),$([a][#6](=O)[#6]=[#6][#6](=O)[a]),$([#6R0](=O)[#6H2][#6R0](=O))]
3054	O-alkenyl,alkynyl,halogen, ketene	0	[$([#8X2](C=O)[#6]=,#[#6]),$([#8X2][#6H1,#6H2][#6]=,#[#6]),$([#8X2][#6H1,#6H2][#6H1,#6H2][#6]=[#6H2]),$([#8X2][#6H1,#6H2][#6H1,#6H2][#6]#[#6]),$([#8X2][#6H1,#6H2][#6H1,#6H2][Cl,Br,I]),$([#8X2][#6H1,#6H2][#6][F]),$([#8X2][#6H1,#6H2][#6r3][Or3])]
3055	CH2=CHC(=O)O	0	[$([#6H2]=[#6H1][#6](=O)[OH]),$([#6H2]=[#6H1][#6](=O)[O][#6])]
3056	CH2=CC(=O)O	0	[$([#6H2]=[#6H0][#6](=O)[OH]),$([#6H2]=[#6H0][#6](=O)[O][#6])]
3174	carbonyl a-C=O C=CC(=O)-a	0	[$([#6v4X3][#6](=O)[#6]=[#6][#6](=O)[#6v4X3])]
4515	o-disubstitued pix2, pi-n-pi	0	cOC(=O)[c;R][c;R]C(=O)Oc
4791	Pro: a,b-unsaturated C=O (amine elimination)	0	[C;X4;$([CH2][N;X3]),$([CH2][OH;X2,SH;X2]),$([CH2][Cl,Br,I])][C;X4;$([C;H1]),$([C;H2])][C;R0]=O
6112	pyrethroids-b	0	aa(a)[C,O]aaaCOC(=O)C

G1_00010 (ID:3034 > 0 or ID:4760 > 0)

FragID	Substructure Name	Count	SMARTS
3034	carboxylic acid C(=O)O	0	[#6;$([#6](=[#8])([#6])[#8H1])]
4760	-SO3H, Sulfonic Acid, sulfo-, -sulfonic acid	0	[C,c,O]S(=O)(=O)[O;$([OH1]),$(O[Na,Li,K])]

Figure 3–4 Definition of structure class R_00001 and G1_00010

Table 3–4: Structure classes matching with query chemical (partial)

(in the case of SMILES: O=C (C) c1cccnc1)

Structure class ID	Category	Structure class name
G1_21025	Ketone	COns_X ketone unreactive
G1_21029	Ketone	COS_X ketone unreactive aromatic
G1_25002	Hydrocarbon	CNO_X unreactive Fish Chronic, w/ N,O
GA_22075	aromatic n	aromatic n reactive Alga

The first letter/s (and in one case a combination of letter and number) in each structure class ID refer to the following structural types.

- Beginning with A: Acidic structure
- Beginning with C: Carbon-containing structure
- Beginning with R: Reactive structure
- Beginning with U: Unreactive structure
- Beginning with GF: Structure set as QSAR class related to fish
- Beginning with GD: Structure set as QSAR class related to daphnids
- Beginning with GA: Structure set as QSAR class related to alga
- Beginning with GFD: Structure set as QSAR class related to fish and daphnids
- Beginning with G1: Structure set as QSAR class related to one of either fish, daphnids, or alga

3.6 Assignment of QSAR Class

Based on the structural classes contained in the query substance, a corresponding QSAR class is assigned to each predicted toxicity type (Table 3-5).

* A single substance can be assigned multiple QSAR classes for a particular predicted toxicity type. The substance is regarded as unclassified if it cannot be categorized under any QSAR class.

Please refer to the following link for a list of all QSAR classes.
https://kate2.nies.go.jp/nies/qsar_classes.php

Table 3–5: QSAR class list for assigning to substance to be predicted

(in case of SMILES: O=C(C)c1cccnc1)

QSAR ID	QSAR class	predicted toxicity type	structure class ID
12102541	COns_X ketone unreactive	Fish Acute	G1_21025
12102941	COS_X ketone unreactive aromatic fish	Fish Acute	G1_21029
22102541	COns_X ketone unreactive	Daphnid Acute	G1_21025
22102941	COS_X ketone unreactive aromatic	Daphnid Acute	G1_21029
32102541	COS_X ketone unreactive	Alga Acute	G1_21025
32207541	aromatic n reactive Alga	Alga Acute	GA_22075
12500251	CNO_X unreactive Fish Chronic, w/ N,O	Fish Chronic	G1_25002
22102551	COns_X ketone unreactive	Daphnid Acute	G1_21025
32102551	COns_X ketone unreactive	Alga Chronic	G1_21025
32207551	aromatic n reactive Alga	Alga Chronic	GA_22075

The prefix of each QSAR class name in Table 3–5, such as COS_X, shows which elements may be in substances of this QSAR class. Upper case alphabet signifies both aliphatic and aromatic, while lower case alphabet signifies only aromatic. For example, COS_X means that both aliphatic and aromatic carbon, oxygen, sulfur, and halogens can be included, while CXnos means that aliphatic and aromatic carbon and halogens, and aromatic nitrogen, oxygen, and sulfur may be included. Therefore, substances containing aliphatic nitrogen are not classified as QSAR classes beginning with COS_X or CXnos.

QSAR IDs (QSAR class ID) are following the rules below.

- The ID consists of 8 digits.
- The first digit is 1 for fish species, 2 for *Daphnia magna*, and 3 for algal species.
- The second to sixth digits match the last five digits of the structure class ID.
- The seventh digit is 4 for acute and 5 for chronic.
- The eighth digit indicates the descriptor. In the current release, only 1: log P is used.

Figure 3–5 shows a simplified flow from substructure extraction to QSAR class assignment when the SMILES of the query chemical is O=C(C)c1cccnc1.

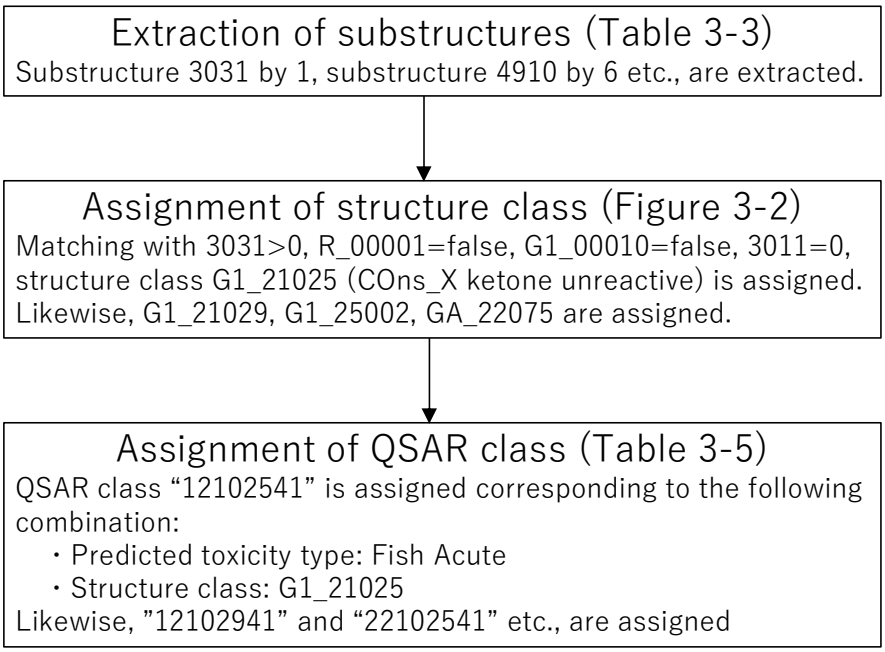

Figure 3–5: Flow from calculating the number of substructures to QSAR class assignment (in case of SMILES: O=C (C) c1cccnc1)

3.7 Calculation of Toxicity Value Using QSAR Equation

For each assigned QSAR class, log P of the query chemical and the precalculated slope and intercept of the relevant QSAR class*2 (Table 3–6) are entered into the following QSAR equation to calculate log (1/toxicity value[mmol/L]). Then, the molecular weight*1 of the query chemical is employed to convert to units of toxicity value[mg/L].

$$\log (1/\text{toxicity value[mmol/L]}) = \text{slope} \times \log P + \text{intercept} \quad \ldots (1)$$

*1 Open Babel [15] is used to calculate the molecular weight of the substance to be predicted.

*2 The slope and the intercept are obtained by a simple regression of the descriptor log P for the training data set of the applicable QSAR class and the dependent variable log (1/toxicity value[mmol/L]). log P values for all training sets use estimated values by KOWWIN™.

3.8 Calculation of Prediction Interval (reference information)

For each assigned QSAR class, substituting the x (log P) value of the query chemical and the statistical value t_{95},n,V_e,$\bar{x}$, Σ^{-1} (Tables 3–6, 3–7) calculated beforehand for the applicable QSAR class gives a 95% confidence level prediction interval for log (1/toxicity value[mmol/L]) defined below. Subsequently, using the molecular weight of the query chemical, the lower and upper limits of the prediction interval of Equation (2) are converted into toxicity values [mg/L].

$$\text{95\% predicted interval} = [\log(1/\text{toxicity value[mmol/L]}) - dy,\ \log(1/\text{toxicity value[mmol/L]}) + dy] \quad \ldots (2)$$

Here,

$$dy = t_{95} \times \sqrt{(1 + 1/n + D^2/(n-1)) \times V_e} \quad \ldots (3)$$

$$D^2 = (x - \bar{x})^{\mathrm{T}} \Sigma^{-1} (x - \bar{x}) \quad \ldots (4)$$

Table 3–6: Statistical values required for calculation of predicted intervals

Variable	Explanation	Formula
n	Number of training set data for the relevant QSAR class	
$\bar{x}$	Average descriptor value for the training set data of the relevant QSAR class	$\sum x_i/n$ x_i : log P value of the i-th training set data
Σ^{-1}	Inverse of the covariance matrix of the descriptors (Here, the inverse of the variance since it is a simple regression)	$n/\sum (x_i - \bar{x})^2$ x_i : log P value of the i-th training set data $\bar{x}$: Mean value of log P of training set data
V_e	Residual variance	$\sum (y_i - \hat{y}_i)^2/(n - p - 1)$ y_i : Measured toxicity value of the i-th training set data $\hat{y}_i$: Predicted toxicity value for the i-th training set data p : Number of descriptors (here 1 since it is a single regression)
t_{95}	t-value at 5% significance level (two-tailed test) for the degrees of freedom of the QSAR class	

Table 3–7: Example of statistical values required for calculating QSAR equation slope, intercept, and prediction interval

QSAR ID	slope	intercept	n	$\bar{x}$	Σ^{-1}	V_e	t_{95}
12100241	0.847	-1.270	30	2.188	0.618	0.178	2.05
12101741	0.784	-1.397	25	2.803	0.963	0.033	2.07

3.9 Calculation of Similarity (reference information)

The degree of similarity of the substance to be predicted is calculated using a fingerprint* (here, the Tanimoto Coefficient using the PubChem fingerprint [24]) for each substance (training data set and support chemicals) assigned to a QSAR class. Degree of similarity is not required information for predictions but can be used as reference information. The fingerprints for all substances included in KATE2020 are calculated beforehand.

The degree of similarity T between bitstream X (fingerprint of chemical included in QSAR class) and bitstream Y (fingerprint of the query chemical) is calculated using the formula below.

$$T = N_c / (N_x + N_y - N_c) \quad \ldots (5)$$

Here,

Nx : Bit count of 1 in X

Ny : Bit count of 1 in Y

Nc : Bit count of 1 common to X and Y

The numerical value is between 0 and 1, and the closer the value is to 1, the more similar is the substance to be predicted. Figure 3-6 illustrates the similarity calculation

*A fixed length bitstream represents chemical substance data, each bit expresses the presence or absence of a particular feature of the chemical substance: 1 means the chemical substance features the property designated by that bit, while 0 means it does not possess that property.

Example of a fingerprint notation:

100100100000000011001011000010101000011 0

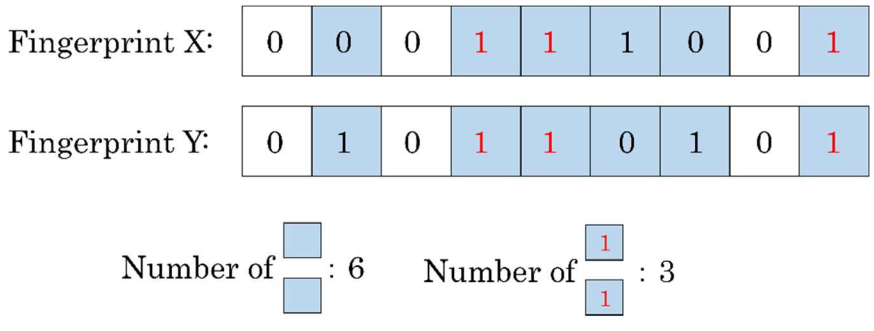

Figure 3–6: Image of similarity (Tanimoto coefficient) calculation

The similarity can be verified using the QSAR class detailed view screen (Verify QSAR screen). Please refer to the KATE2020 Operating Manual† [6. QSAR Class Information Detailed View (Verify QSAR screen)] for more information regarding the Verify QSAR screen.

† https://kate2.nies.go.jp/nies/doc/KATEmanual_2020-e.pdf

4. Definition of Applicability Domain – OECD (Q) SAR Validation Principle 3

Here, we explain the applicability domain of the KATE2020 QSAR model and the method for determining the applicability domain.

4.1 Applicability Domain

The applicability domain of KATE2020 is described in terms of A) structure applicability domain and B) log P applicability domain.

A) Applicability domain for structure

The applicability domain for structure is set for each QSAR class and substructures are assigned based on the list of substructures for structure judgement (Table 4–1) for training set data and chemical data with inequality sign (log P judgement for the applicable QSAR class (explained in 4.2 B) is within the applicability domain only).

Table 4–1: List of substructures for structure judgement for each QSAR class (partial)

QSAR ID	List of Substructures for structure judgement
...	
12100241	5004,5007,5008,5016,5022,5023,5028,5081,5500
22100741	5074, 5075
32100541	5020,5021,5067,5155
...	

Please refer to the following link for the substructure list used in structure judgement.

kate.nies.go.jp/data/Substructures_for_judgement.html

B) Applicability domain for log P

The following log P ranges (Table 4–2) for each QSAR class are calculated beforehand in order to judge log P (explained in 4.2 B).

- log P range 1: log P minimum value and maximum value of training set data included in the applicable QSAR class (required for the judgement of "in" and "out of")
- log P range 2: log P minimum value and maximum value of training set data and chemical data with inequality sign included in the applicable QSAR class (required for the judgement of "out of+)

The log P applicability domain is given based on the aforementioned log P range 1.

Table 4–2: log P range for each QSAR class

QSAR ID	log P Range 1	log P Range 2
…		
12100241	0.09 ~ 4.84	0.09 ~ 4.84
22100741	3.77 ~ 6.17	3.77 ~ 6.89
32100541	1.46 ~ 3.48	1.46 ~ 4.63
…		

4.2 Applicability Domain Judgement Method

KATE2020 judges whether the predicted toxicity value of the query chemical is within the range where it can be applied. Judgement by A) structure and by B) log P are conducted and if both are within the applicability domains, the predicted toxicity value is judged to be within the applicability domain.

A) Structural Judgement

KATE2020 compares substructures used for structure judgement to judge whether the structure of a substance to be predicted is within the applicability domain of the relevant QSAR class (Figure 4–1). There are three possible results as explained below. Structures judged to be within the applicability domain of a particular QSAR class are classified as either "in" or "in (conditionally)."

- in: Within applicability domain

  When all substructures for structure judgement contained in the query chemical are included in the substructure list (refer to Table 4–1) for

structure judgement of substances*1 for the applicable QSAR class (pink area of Figure 4–1), or when the query chemical possesses none of the substructures for structure judgement.

・ in (conditionally): Conditionally within applicability domain

While not satisfying the conditions for "in," all substructures for structure judgement for the substance to be predicted are either included in the substructure list for structure judgement for the applicable QSAR class or included in the substructure list for structure judgement *1 of Narcotic Group*2 (pink and yellow areas of Figure 4–1).

・ out of : Outside applicability domain

Cases where the conditions for neither "in" nor "in (conditionally)" are satisfied. In other words, when the query chemical possesses substructures for structure judgement that are neither included in the substructure list for structure judgement for the applicable QSAR class, nor included in the Narcotic Group*2 class (structures in the gray area of Figure 4–1).

*1 Here, data with inequality signs whose log P judgement is "in" (within the applicability domain) are included.

*2 Low reactivity, and baseline toxicity (narcotic effect) not based on a specific bioactivity. QSAR classes containing chemicals such as aliphatic hydrocarbons, sulfoxides, aliphatic and aromatic esters, aliphatic and aromatic ketones, and alcohols, whose toxicity can be explained by simple narcotic effect, exist for each predicted toxicity type. Narcotic group class is defined to include such QSAR classes.

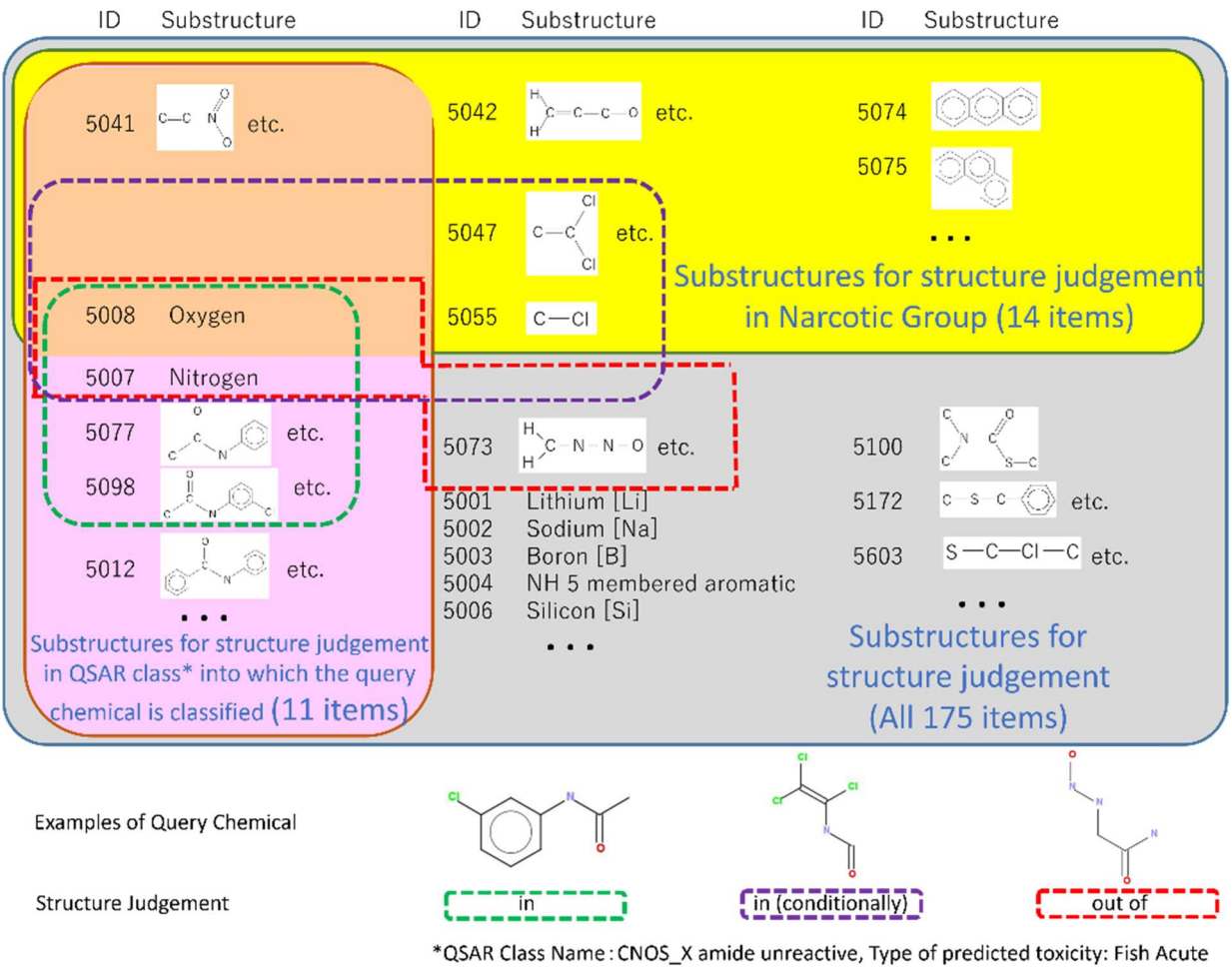

Figure 4–1: Structural judgement example

B) log P Judgement

KATE2020 judges whether the log P value of the query chemical lies between the log P minimum value and maximum value (the range shown in Table 4–2 for the applicable QSAR class) for all training data set (support chemicals data is not included) for the applicable QSAR class in order to judge whether the value lies within the applicability domain. Note that all substances with log P>6.0* are outside the applicability domain for KATE2020 (a modification for the KATE2020 version).

*The threshold value of 6.0 set here was selected based on a comprehensive consideration of the following: the cut-off value for predicted acute toxicity in ECOSAR is 5.0 (some set at 6.4); the cut-off value for predicted chronic toxicity value is set at 8.0; the upper limit of linearity between log P and log BCF (for example, in Dimitrov et al. SAR QSAR Environ Res., 13, 177-184, 2010, the upper limit is 6.1–6.5); and the upper limit of the applicability domain for measurement of log P for highly hydrophobic (log P>4) substances using the HPLC method (prescribed by OECD Test Guideline 117) is 6.

- in: Within applicability domain (refer to Figure 4–2).
- out of : Outside applicability domain. Excluding the case of "out of+" explained in the following (refer to Figure 4–3).
- out of+ : Outside applicability domain. However, the log P value of the query chemical lies between the log P minimum value and maximum value (log P range 2 for each QSAR class row in Table 4–2) for all substances including the training data set and support chemicals for the applicable QSAR class (refer to Figure 4–4).

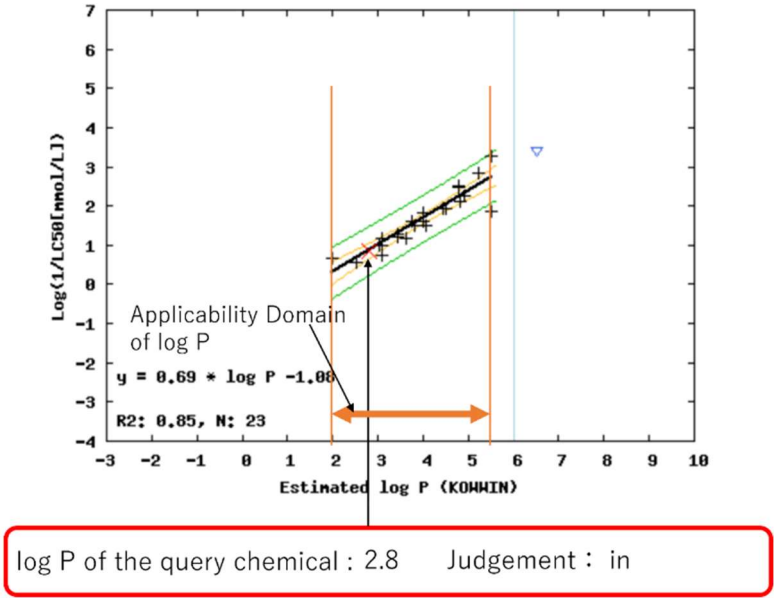

Figure 4–2: log P judgement example (in)

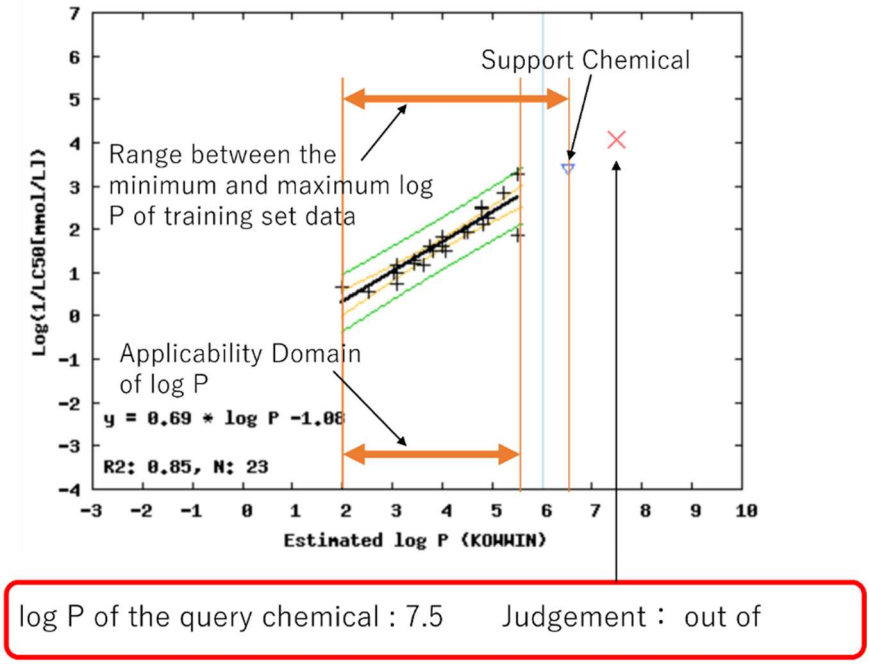

Figure 4–3: log P judgement example (out of)

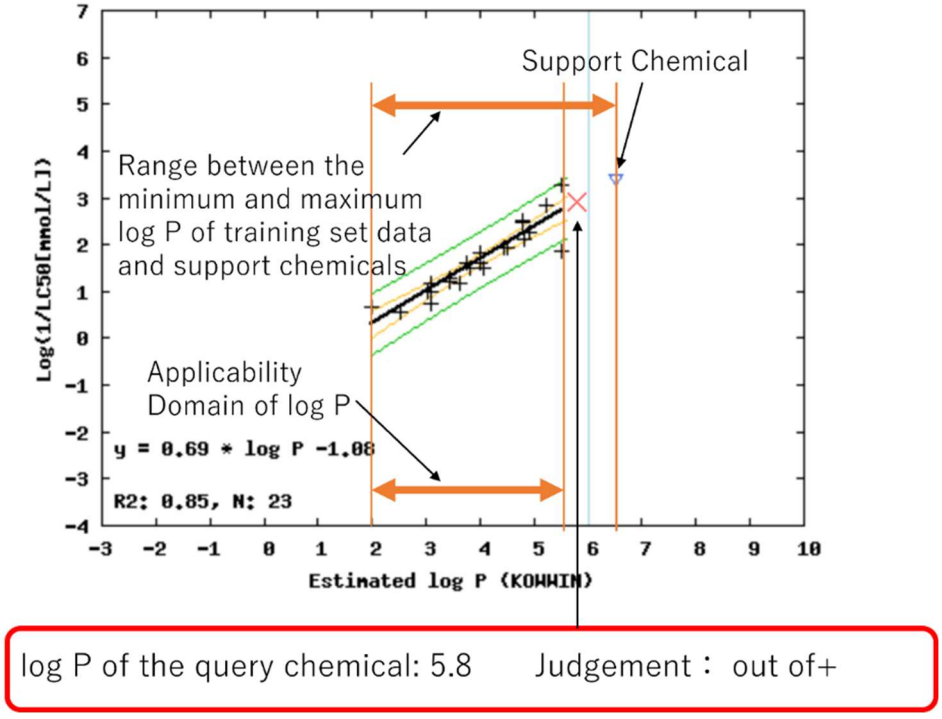

Figure 4–4: log P judgement example (out of+)

5. Validation – OECD (Q) SAR Validation Principle 4

5.1 Internal Validation

Here, the goodness-of-fit (to what extent the model explains responses in training set data, in other words, variances of predicted toxicity values) and robustness (the stability of predictions if one or more data is removed from a training set) of KATE2020 QSAR models are assessed.

5.1.1 Indicators used for internal validation

Internal validation uses the indicators listed in Table 5–1.

Table 5–1: Internal validation indicators

indicator	Explanation
R^2	An index of goodness of fit (coefficient of determination), taking a value between 0 and 1, the closer to 1, the better the fit. $R^2=1-\frac{\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}{\sum_{i=1}^{n}(y_i-\bar{y})^2}$ y_i : Measured toxicity value of the i-th training set data $\hat{y}_i$: Predicted toxicity value of the i-th training set data $\bar{y}$: Average of measured toxicity values of training set data n : Number of training set data
Q^2	Robustness index (leave-one-out method); the closer it is to 1, the more robust it is, and it can be negative. $Q^2=1-\frac{\sum_{i=1}^{n}(y_i-\hat{y}_{i,-i})^2}{\sum_{i=1}^{n}(y_i-\bar{y})^2}$ y_i : Measured toxicity value of the i-th training set data $\hat{y}_{i,-i}$: Predicted toxicity values from regression equations constructed with the remaining training set data, excluding the i-th training set data $\bar{y}$: Average of measured toxicity values of training set data n : Number of training set data

5.1.2 Internal validation results

Table 5–2 shows the internal validation results for QSAR classes that meet statistical criteria ($R^2 \geqq 0.7$, $Q^2 \geqq 0.5$ and $n \geqq 5$) [25–30].

Table 5–2: List of QSAR classes meeting statistical criteria

(sorted by *n* value for each organism and acute/chronic)

QSAR Class	Organics	Acute/Chronic	R2	Q2	n
narcotic group Fish Acute	Fish	Acute	0.87	0.87	154
CNOS_X halogen unreactive	Fish	Acute	0.76	0.75	95
phenol unreactive unhindered	Fish	Acute	0.88	0.87	58
phenol unreactive unhindered w/o bisphenol, HRAC Ea	Fish	Acute	0.87	0.86	57
CO_X alcohol unreactive w/o EO Fish	Fish	Acute	0.89	0.88	46
CO_X ether unreactive	Fish	Acute	0.87	0.86	44
COns_X ketone unreactive	Fish	Acute	0.84	0.82	41
CNOS_X aromatic n unreactive excl. triazine Fish	Fish	Acute	0.76	0.74	39
phenol unreactive unhindered w/o X	Fish	Acute	0.89	0.88	37
CNO_X ester unreactive	Fish	Acute	0.72	0.68	37
Cnos_X heteroaromatic unreactive	Fish	Acute	0.84	0.81	30
CNO_X nitro mono unreactive	Fish	Acute	0.74	0.70	28
COns_X ketone unreactive aliphatic	Fish	Acute	0.88	0.86	27
amine primary unreactive NH2=1 aliphatic	Fish	Acute	0.84	0.81	26
C_X hydrocarbon unreactive aliphatic w/ X, excl. Halomethane	Fish	Acute	0.87	0.86	24
C_X hydrocarbon unreactive aromatic fused R=0 w/o X	Fish	Acute	0.85	0.79	24
amine primary unreactive aromatic w/ NO2,SO	Fish	Acute	0.82	0.79	24
CO_X primary alcohol	Fish	Acute	0.92	0.9	22
CNOS_X amine sec,tert w/o n	Fish	Acute	0.91	0.89	21
Cnos_X heteroaromatic unreactive Fish, Daphnid	Fish	Acute	0.82	0.78	21
CNO_X amide unreactive	Fish	Acute	0.80	0.76	21
C_X hydrocarbon unreactive aliphatic w/o X	Fish	Acute	0.73	0.68	21
CNO_X aldehyde normal aromatic	Fish	Acute	0.85	0.81	19
CNOS_X amine aromatic w/ aliphatic carbon	Fish	Acute	0.77	0.66	19
CNO_X amine sec,tert unreactive aliphatic	Fish	Acute	0.91	0.86	16
CN_X amine sec,tert unreactive aliphatic	Fish	Acute	0.73	0.64	16
CN_X nitrile unreactive	Fish	Acute	0.87	0.84	15
COS_X ketone unreactive aromatic	Fish	Acute	0.90	0.85	14
CNOS_X amine primary reactive w/o ortho,para-OH,NH2	Fish	Acute	0.83	0.76	14
amine primary unreactive NH2 >1	Fish	Acute	0.82	0.75	14
CN_X amine sec,tert unreactive aromatic	Fish	Acute	0.84	0.78	12
CNOS_X acid unreactive	Fish	Acute	0.77	0.63	12
Cnos_X heteroaromatic reactive Fish	Fish	Acute	0.72	0.59	12

QSAR Class	Organics	Acute/Chronic	R2	Q2	n
n+, N+	Fish	Acute	0.75	0.61	11
CNOS_X aromatic n reactive excl. nitrile	Fish	Acute	0.73	0.53	11
phenol reactive w/o ortho,para-OH,NH2, w/o nitro	Fish	Acute	0.75	0.64	10
urea unreactive	Fish	Acute	0.95	0.87	9
amide reactive excl. C=O,S,N	Fish	Acute	0.92	0.85	8
CNOS_X carbamate unreactive Fish	Fish	Acute	0.9	0.68	8
ester reactive methacrylate	Fish	Acute	0.76	0.54	8
CS_X sulfide unreactive	Fish	Acute	0.70	0.59	8
CN_X nitrile unreactive aliphatic	Fish	Acute	0.97	0.95	7
COS_X methacrylate	Fish	Acute	0.87	0.66	7
COS_X thiol	Fish	Acute	0.96	0.92	6
CNOSP_X phosphorus unreactive	Fish	Acute	0.94	0.87	6
CNOS_X N-hetero unreactive w/o amine, aldoxime, carbamate	Fish	Acute	0.79	0.55	6
CO_X alcohol unreactive w/ EO	Fish	Acute	0.98	0.97	5
C_X hydrocarbon unreactive halomethane	Fish	Acute	0.94	0.85	5
CNOS_X amine tert unreactive w/ C=O	Fish	Acute	0.88	0.61	5
phenol unreactive bisphenol	Fish	Acute	0.87	0.63	5
narcotic group Daphnid Acute	Daphnid	Acute	0.71	0.70	83
CNOS_X halogen unreactive	Daphnid	Acute	0.86	0.84	44
phenol unreactive unhindered	Daphnid	Acute	0.82	0.78	28
phenol unreactive unhindered w/o bisphenol, HRAC Ea	Daphnid	Acute	0.8	0.76	27
phenol unreactive unhindered w/o X	Daphnid	Acute	0.87	0.83	19
CNOS_X aromatic n unreactive Daphnid	Daphnid	Acute	0.85	0.82	17
C_X hydrocarbon unreactive aromatic fused R=0 w/o X	Daphnid	Acute	0.8	0.74	17
CO_X ether unreactive	Daphnid	Acute	0.83	0.74	15
C_X hydrocarbon unreactive aliphatic w/ X	Daphnid	Acute	0.82	0.77	14
CO_X alcohol unreactive w/o EO Daphnid	Daphnid	Acute	0.78	0.72	14
amine primary unreactive NH2 >1	Daphnid	Acute	0.71	0.60	12
Cnos_X heteroaromatic unreactive	Daphnid	Acute	0.94	0.90	11
phenol unreactive hindered	Daphnid	Acute	0.76	0.64	11
CNO_X amine sec mono w/o n Daphnid	Daphnid	Acute	0.75	0.58	9
CNO_X ester unreactive Daphnid	Daphnid	Acute	0.93	0.86	8
CNO_X nitro mono unreactive Daphnid	Daphnid	Acute	0.85	0.74	8
C_X hydrocarbon unreactive aliphatic w/ X, excl. gem,vic-Cl,TCE	Daphnid	Acute	0.98	0.97	7

QSAR Class	Organics	Acute/Chronic	R2	Q2	n
Cnos_X heteroaromatic unreactive Fish, Daphnid	Daphnid	Acute	0.96	0.9	7
CN_X amine sec,tert unreactive aromatic	Daphnid	Acute	0.97	0.92	6
CO_X primary alcohol	Daphnid	Acute	0.95	0.76	6
CN_X amine sec,tert unreactive aliphatic	Daphnid	Acute	0.89	0.64	6
CNO_X amide unreactive Daphnid	Daphnid	Acute	0.88	0.70	6
n+, N+	Daphnid	Acute	0.87	0.77	6
CNOS_X N-hetero unreactive w/o amine, aldoxime, carbamate	Daphnid	Acute	0.78	0.63	6
ester reactive methacrylate	Daphnid	Acute	0.82	0.60	5
CNOS_X amine sec,tert multi-functional	Daphnid	Acute	0.79	0.53	5
CNO_X imide unreactive	Daphnid	Acute	0.78	0.59	5
narcotic group Alga Acute	Alga	Acute	0.76	0.74	52
phenol unreactive unhindered w/o bisphenol, HRAC Ea	Alga	Acute	0.8	0.77	26
aromatic n reactive Alga	Alga	Acute	0.78	0.72	10
CO_X ether unreactive excl. HRAC Ea Alga	Alga	Acute	0.92	0.82	9
CNOS_X sulfur reactive excl. disulfide Alga	Alga	Acute	0.86	0.54	8
CO_X alcohol unreactive w/o halogen, acid, EO	Alga	Acute	0.95	0.9	6
CNO_X ester unreactive Alga	Alga	Acute	0.94	0.85	6
CO_X primary alcohol	Alga	Acute	0.91	0.79	6
COS_X thiol	Alga	Acute	0.88	0.52	6
C_X hydrocarbon unreactive aliphatic w/ X, excl. Halomethane	Alga	Acute	0.97	0.91	5
Cnos_X heteroaromatic excl. pyridine Alga	Alga	Acute	0.83	0.52	5
narcotic group Fish Chronic	Fish	Chronic	0.82	0.75	12
Cnos_X unreactive Fish Chronic	Fish	Chronic	0.76	0.68	12
C_X hydrocarbon unreactive	Fish	Chronic	0.78	0.68	11
narcotic group Daphnid Chronic	Daphnid	Chronic	0.70	0.68	74
C_X hydrocarbon unreactive aromatic fused R=0 w/o X	Daphnid	Chronic	0.87	0.83	15
CNO_X amine sec,tert unreactive w/ N-Oxide,Nitroso	Daphnid	Chronic	0.81	0.74	15
CO_X alcohol unreactive w/o EO Daphnid	Daphnid	Chronic	0.82	0.75	14
CO_X ether unreactive	Daphnid	Chronic	0.88	0.76	10
CNO_X ester unreactive Daphnid	Daphnid	Chronic	0.84	0.73	8
CNO_X amide unreactive Daphnid	Daphnid	Chronic	0.83	0.74	8
Cnos_X heteroaromatic unreactive	Daphnid	Chronic	0.83	0.64	7
Cnos_X heteroaromatic unreactive Fish, Daphnid	Daphnid	Chronic	0.83	0.64	7
C_X hydrocarbon unreactive aliphatic w/ X, excl. gem,vic-Cl,TCE	Daphnid	Chronic	0.98	0.97	6

QSAR Class	Organics	Acute/Chronic	R2	Q2	n
COns_X ketone unreactive	Daphnid	Chronic	0.92	0.59	5
phenol unreactive unhindered w/o X, bisphenol, HRAC Ea	Alga	Chronic	0.72	0.65	18
CO_X ether unreactive excl. HRAC Ea Alga	Alga	Chronic	0.89	0.86	15
CNO_X nitro mono unreactive	Alga	Chronic	0.78	0.53	13
aromatic n reactive Alga	Alga	Chronic	0.77	0.70	11
CO_X alcohol unreactive w/o halogen, acid, EO	Alga	Chronic	0.87	0.81	10
amine primary unreactive NH2>1, Nv3 <3	Alga	Chronic	0.78	0.70	10
ester unreactive w/o acid	Alga	Chronic	0.88	0.79	9
CNOS_X sulfur reactive excl. disulfide Alga	Alga	Chronic	0.82	0.66	9
CNOSP_X phosphorus all	Alga	Chronic	0.74	0.64	9
CNO_X ester unreactive Alga	Alga	Chronic	0.90	0.79	8
COS_X thiol	Alga	Chronic	0.85	0.07	7
CNO_X amine sec,tert unreactive aliphatic	Alga	Chronic	0.90	0.79	6
CNOSP_X phosphorus unreactive	Alga	Chronic	0.95	0.86	5
CNOS_X N-hetero unreactive w/o amine, aldoxime, carbamate	Alga	Chronic	0.80	0.57	5

5.2 External Validation

Here, we assess KATE2020 QSAR model external predictivity performance (how well can new data that was not used in model development be predicted).

5.2.1 Data (test set) used in external validation

Toxicity values listed in OECD SIDS (Screening Information DataSet) were used as the test set for external validation. Data deviating from the standard test duration of ±24 hours and data with reliability (Klimisch code) of 3 or 4 were excluded, while test results for species not included in KATE2020 (e.g., *Danio rerio* and *Desmodesmus subspicatus*) were adopted. Further, if multiple test results existed for a single substance, the data of highest reliability was adopted. if multiple test results with the highest reliability existed, the geometric mean value of those data was taken.

Table 5–3 shows the number of test set substances for each predicted toxicity type and the data points where predicted values were obtained for KATE2020 classes. All predicted values are included, including where multiple predicted values were obtained for a single substance.

Table 5–3: Test set data points

	Acute			Chronic		
	Fish	Daphnid	Alga	Fish	Daphnid	Alga
Number of Chemicals	178	196	141	3	46	100
Number of Predicted Values	199	207	93	1	55	76

5.2.2Results

Of the data of Table 5–3, Table 5–4 shows the number and ratio of data points within the applicability domain (including "in (conditionally)" structure judgement) where the ratio of predicted value to measured value is less than one order of magnitude (1/10th to 10 times) and ess than two order of magnitude (1/100th to 100 times) for all QSAR classes. The data is also present in graph form in Figure 5–1. In the figure, inside the blue dotted line is the range of one order of magnitude, and the red single-pointed line the two order of magnitude.

Table 5–4: Number and ratio of data points where ratio of predicted to measured value is less than one order and two order of magnitude for all QSAR classes

	Acute						Chronic					
	Fish		Daphnid		Alga		Fish		Daphnid		Alga	
	Number	Ratio	Number	Ratio	Number	Ratio	Number	Ratio	Number	Ratio	Number	Ratio
Data falling within 1 order of magnitude	176	88%	147	71%	66	71%	0	0%	39	71%	51	67%
Data falling within 2 order of magnitude	196	98%	199	96%	89	96%	0	0%	53	96%	70	92%
All Data	199		207		93		1		55		76	

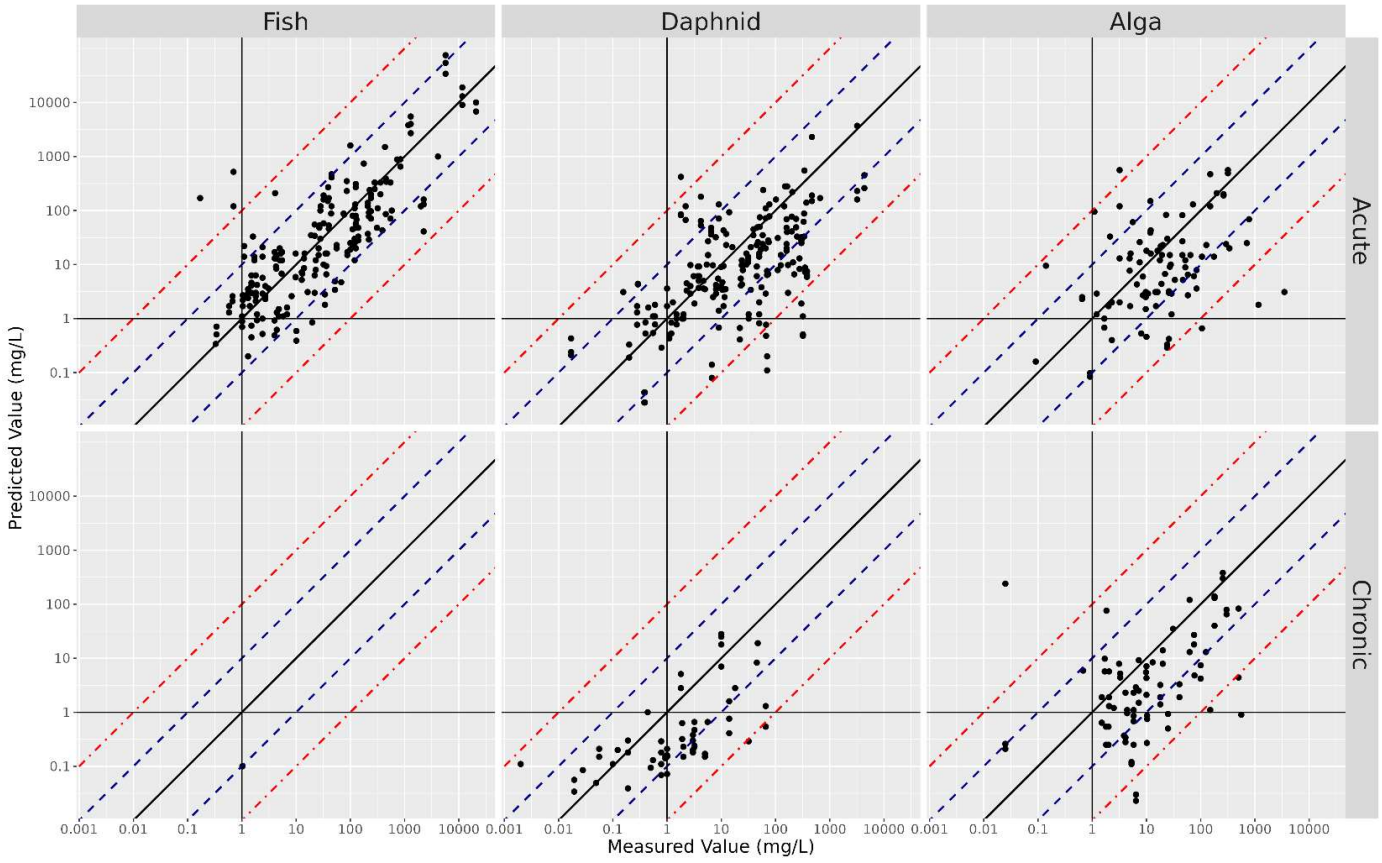

Figure 5–1: Predicted versus measured values for all QSAR classes

The number and ratio of data points assigned to QSAR classes that satisfy statistical standards (Table 5–2) that are within the applicability domain (log P judgement and structure judgement are both "in," including "in (conditionally)" structure judgement) where the ratio of predicted value to measured value is less than one order of magnitude and two order of magnitude for QSAR classes that satisfy statistical criteria are shown in Table 5–5 and graphically represented in Figure 5–2.

Table 5–5: Number and ratio of data points where ratio of predicted to measured value is less than one order and two order of magnitude for QSAR classes satisfying statistical criteria

	Acute						Chronic					
	Fish		Daphnid		Alga		Fish		Daphnid		Alga	
	Number	Ratio	Number	Ratio	Number	Ratio	Number	Ratio	Number	Ratio	Number	Ratio
Data falling within 1 order of magnitude	136	90%	77	73%	25	69%	0	NA	10	77%	16	59%
Data falling within 2 order of magnitude	148	98%	102	96%	35	97%	0	NA	13	100%	24	89%
All Data	151		106		36		0		13		27	

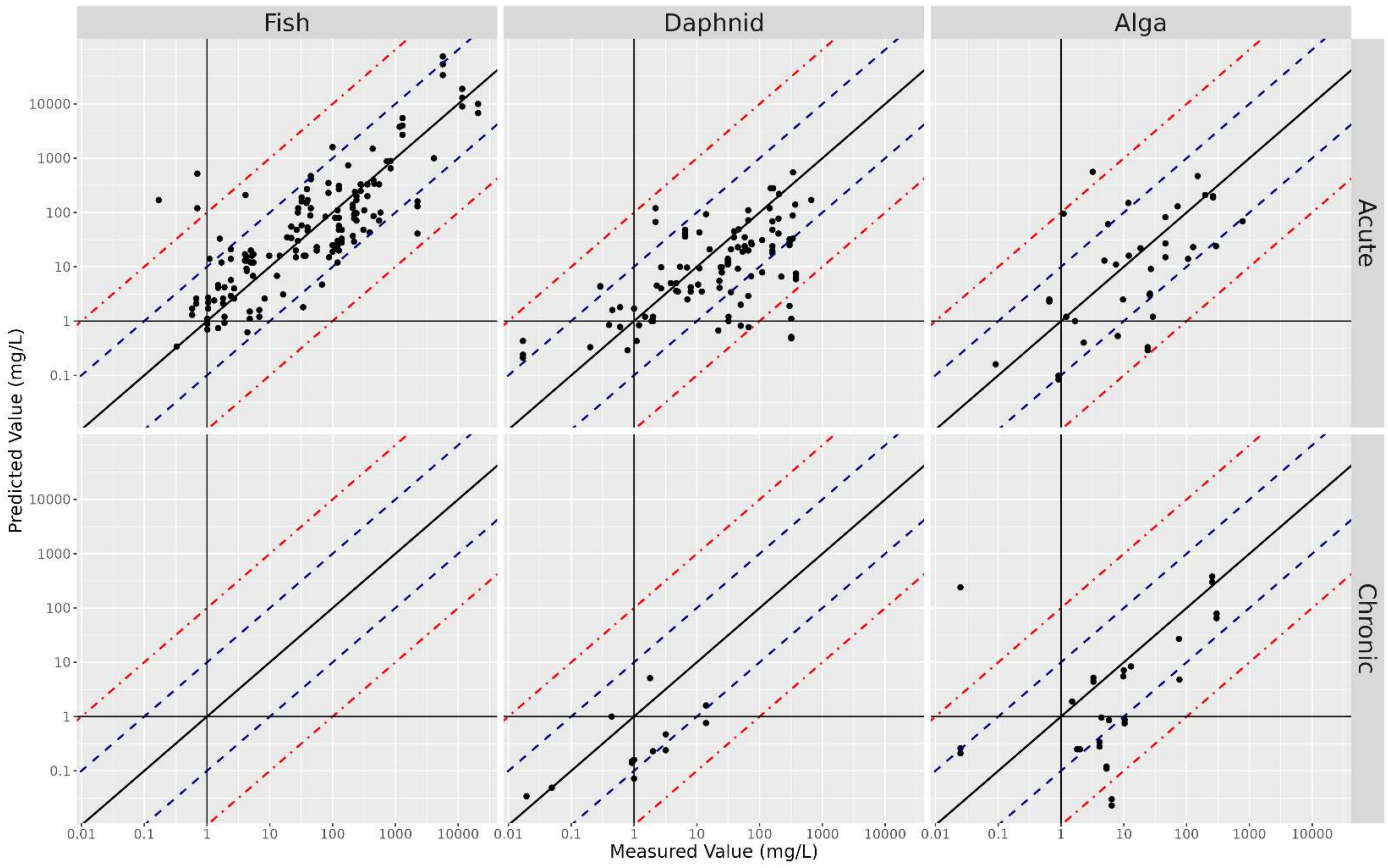

Figure 5–2: Predicted values versus measured values for QSAR classes satisfying statistical criteria

6. Interpretation of mechanism – OECD (Q) SAR Validation Principle 5

A linear relationship between the logarithms of hydrophilicity (membrane permeability) and toxicity has been reported[31], and KATE2020 constructs QSAR equations based on log P as an explanatory variable and toxicity value (effective concentration) as a dependent variable. Each QSAR class in KATE consists of chemical substances that possess characteristic structures, and each structure class is defined by the parameter for number of substructures (Chapter 3, Section 3.5). Substances possessing substructures considered to be highly reactive are classified as reactive structure classes and are believed to exhibit toxicity through their nonspecific high reactivity. Substances without such substructure are considered to have low reactivity and are classified as unreactive structure classes[32]. Some structure classes contain substructures that act via specific mechanisms. Please refer to the following for structure classes and substructure list.

Structure class list: https://kate2.nies.go.jp/nies/Structure_Classes.php
Substructure class list: https://kate2.nies.go.jp/nies/Substructures.php

References

[1] https: //cdk.github.io/

[2] https: //www.epa.gov/tsca-screening-tools/epi-suitetm-estimation-program-interface

[3] http: //www.eic.or.jp/ecoterm/?act=view&serial=295

[4] https: //www.env.go.jp/chemi/report/y052-[24]/1_ref2 terms.pdf

[5] https: //www.daylight.com/dayhtml/doc/theory/theory.smarts.html

[6] http: //www.daylight.com/smiles/index.html

[7] OECD Guidelines for the Testing of Chemicals, Test No. 203: Fish, Acute Toxicity Test (https: //www.oecd-ilibrary.org/environment/test-no-203-fish-acute-toxicity-test_9789264069961-en)

[8] OECD Guidelines for the Testing of Chemicals, Test No. 210: Fish, Early-life Stage Toxicity Test (https: //www.oecd-ilibrary.org/environment/test-no-210-fish-early-life-stage-toxicity-test_9789264203785-en)

[9] OECD Guidelines for the Testing of Chemicals, Test No. 202: Daphnia sp. Acute Immobilisation Test (https: //www.oecd-ilibrary.org/environment/test-no-202-daphnia-sp-acute-immobilisation-test_9789264069947-en)

[10] OECD Guidelines for the Testing of Chemicals, Test No. 211: *Daphnia magna* Reproduction Test (https: //www.oecd-ilibrary.org/environment/test-no-211-daphnia-magna-reproduction-test_9789264185203-en)

[11] OECD Guidelines for the Testing of Chemicals, Test No. 201: Freshwater Alga and Cyanobacteria, Growth Inhibition Test (https: //www.oecd-ilibrary.org/environment/test-no-201-alga-growth-inhibition-test_9789264069923-en)

[12] http: //www.env.go.jp/chemi/sesaku/01.html

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

[14] Guidance Document on the Validation of (Quantitative) Structure-Activity Relationship [(Q) SAR] Models, 2007 (https: //www.oecd.org/officialdocuments/publicdisplaydocumentpdf/?doclanguage=en&cote=env/jm/mono (2007) 2)

[15] http: //openbabel.org/wiki/Main_Page

[16] https: //jsme-editor.github.io/
(All of the website links listed above were accessed on March 1, 2023)

[17] Bienfait B, Ertl P (2013) JSME: a free molecule editor in JavaScript. *J Cheminform* 5 (24) . doi: 10.1186/1758-2946-5-24

[18] Willighagen E.L., Mayfield J.W., Alvarsson J, *et al* (2017) The Chemistry Development Kit (CDK) v2.0: atom typing, depiction, molecular formulas, and substructure searching. *J Cheminform* 9 (33) . doi: 10.1186/s13321-017-0220-4

[19] May J.W., Steinbeck C (2014) Efficient ring perception for the Chemistry Development Kit. *J Cheminform,* 6 (3) . doi: 10.1186/1758-2946-6-3

[20] Steinbeck C, Hoppe C, Kuhn S, Floris M, Guha R and Willighagen E.L. (2006) Recent Developments of the Chemistry Development Kit (CDK) - An Open-Source Java Library for Chemo- and Bioinformatics, *Curr. Pharm. Des,* 12 (17) , 2111-2120. doi: 10.2174/138161206777585274

[21] Steinbeck C, Han Y, Kuhn. S, Horlacher. O, Luttmann. E, and Willighagen E.L. (2003) The Chemistry Development Kit (CDK) : An open-source Java library for chemo- and bioinformatics, *J. Chem. Inf. Comput. Sci.* 43 (2) , 493-500. doi: 10.1021/ci025584y

[22] Aptula A.O., Patlewicz G, Roberts D.W. (2005) Skin sensitization: Reaction mechanistic applicability domains for structure–activity relationships, *Chem. Res. Toxicol.* 18 (9) , 1420–1426. doi: 10.1021/tx050075m

[23] Furuhama A, Hasunuma K, Aoki Y, Yoshioka Y, Shiraishi H (2011) Application of chemical reaction mechanistic domains to an ecotoxicity QSAR model, the KAshinhou Tool for Ecotoxicity (KATE) , *SAR QSAR Environ Res.*, 22 (5-6) , 505-523. doi: 10.1080/1062936X.2011.569944

[24] https: //cdk.githuRb.io/cdk/1.5/docs/api/org/openscience/cdk/fingerprint/PubchemFingerprinter.html (accessed on March 1, 2023)

[25] Golbraikh A, Tropsha A (2002) Beware of q2! *J. Mol. Graph. Model.* 20 (4) , 269–276. doi: 10.1016/s1093-3263 (01) 00123-1

[26] Eriksson L, Jaworska J, Worth A.P., Cronin M.T.D., McDowell R.M. (2003) Methods for Reliability and Uncertainty Assessment and for Applicability Evaluations of Classification- and Regression-Based QSARs, *Environ. Health Perspect.* 111 (10) , 1361-1375. doi: 10.1289/ehp.5758

[27] Tropsha A, Gramatica P (2003) The Importance of Being Earnest: Validation is the Absolute Essential for Successful Application and Interpretation of QSPR Models, *QSAR Comb. Sci.* 22 (1) , 69-77. doi: 10.1002/qsar.200390007

[28] ECHA (2016) Practical guide How to use and report (Q) SARs 3.1.

[29] Posthumus P, Slooff W (2001) RIVM report, Implementation of QSAR in ecotoxicological risk assessments

[30] Alexander D.L.J, Tropsha A (2015) Beware of R2: simple, unambiguous assessment of the prediction accuracy of QSAR and QSPR models, *J. Chem. Inf. Model.* 55 (7) , 1316–1322. doi: 10.1021/acs.jcim.5b00206

[31] Hansch C, Dunn W.J. (1972) Linear Relationships between Lipophilic Character and Biological Activity of Drugs, *J Pharm Sci.*, 61 (1) , 1-19. doi: 10.1002/jps.2600610102

[32] Verhaar H.J.M., van Leeuwen C.J., Hermens J.L.M. (1992) Classifying environmental pollutants, *Chemosphere*, 25 (4) , 471-491. doi: 10.1016/0045-6535 (92) 90280-5