生態毒性予測システム「KATE2020」
技術文書（2023年3月30日版）

※ KATE2020 は、化学物質の生態毒性に関する

- 魚類急性毒性試験における半数致死濃度（LC50）
- ミジンコ急性遊泳阻害試験における半数影響濃度（EC50）
- 藻類生長阻害試験における半数影響濃度（EC50）
- 魚類初期生活段階毒性試験における無影響濃度（NOEC）
- ミジンコ繁殖試験における無影響濃度（NOEC）
- 藻類生長阻害試験における無影響濃度（NOEC）

を予測するシステムです。

※ **本システムで得られた予測結果は、「 化学物質の審査及び製造等の規制に関する法律」に基づく届出に必要な生態毒性試験結果として利用することはできません。**
化学物質の生態毒性影響の程度についての参考としてご利用ください。

ご質問等がございましたら、下記までお問い合わせ下さい。
国立研究開発法人 国立環境研究所 環境リスク・健康領域　KATE担当
kate@nies.go.jp

KATE2020 技術文書 改訂履歴

バージョン	発行日	改訂履歴
第0.99版	2022年3月30日	KATE2020（version3.0）向け技術文書
第0.99.1版	2023年3月30日	KATE2020（version4.0）向け技術文書

目次

Web 上の関連資料　Appendix
https://kate.nies.go.jp/data/Appendix.html

- 物質データー覧　Chemicals Data
  https://kate.nies.go.jp/data/Chemicals_Data.html

- 部分構造一覧　Substructures
  https://kate.nies.go.jp/data/Substructures.html

- 構造判定用部分構造一覧　Substructures for judgement
  https://kate.nies.go.jp/data/Substructures_for_judgement.html

- 構造クラス一覧　Structure Classes
  https://kate.nies.go.jp/data/Structure_Classes.html

- QSAR クラス一覧　QSAR Classes
  https://kate.nies.go.jp/data/QSAR_Classes.html

略語一覧

CDK: Chemistry Development Kit

プログラミング言語JAVAで書かれた、ケモインフォマティクスやバイオインフォマティクスのためのオープンソースソフトウェア。化学物質の部分構造の抽出や物性の計算等を行うことができる[1] 。

EC50: 50% Effective Concentration（半数影響濃度）

試験水に溶解した化学物質などによって、半数（50%）の試験生物に対して影響を与えると考えられる濃度。

KATE: KAshinhou Tool for Ecotoxicity

国立環境研究所 環境リスク・健康研究領域において研究・開発された生態毒性QSARシステム」の通称。「ケイト」と読む。

KOWWIN™:

USEPAなどが開発しているEPI Suite™ (Estimation Programs Interface: 化学物質を迅速にスクリーニングするためのアプリケーションなどで使用されることを目的としたツール) に含まれる化学物質のlog P推定プログラム[2] 。

LC50: 50% Lethal Concentration（半数致死濃度）

試験水に溶解した化学物質などによって、半数（50%）の試験生物を死亡させる濃度。

log P: The logarithm of the octanol/water partition coefficient

（オクタノール/水分配係数）

ある化学物質について、1-オクタノールと水の２つの溶媒中の平衡状態における濃度比を常用対数で表したもの。化学物質の疎水性を表す指標とされている。log Pは対象物質のイオン化を無視した数値である[3] 。

NOEC: No Observed Effect Concentration（無影響濃度）

最大無影響濃度、最大無作用濃度ともいう。対照区と比較して統計的に有意な（有害） 影響が認められなかった最高濃度であり、LOEC （最小影響濃度）のすぐ下の濃度区である(環境省 化学物質の環境リスク評価 第１７巻 第１編 参考２ 用語集 [4] 等より。)

OECD: Organisation for Economic Co-operation and Development（経済協力開発機構）

(Q)SAR: (Quantitative) Structure-Activity Relationships（(定量的)構造活性相関）

化学物質の構造上の特徴又は物理化学定数と生物学的活性（毒性等）の相関関係を構造活性相関（SAR: Structure-Activity Relationship）といい、定量的なものを定量的構造活性相関（QSAR: Quantitative Structure-Activity Relationship）という。両者を併せて(Q)SAR と記載することもある。構造活性相関は、例えば、特定の官能基の有無から物質の有害性の多寡を推測することを指し、構造を手掛かりに毒性等を定量的に算出する仕組みをいわゆるQSAR モデルと呼ぶ(環境省 化学物質の環境リスク評価 第１７巻 第１編 参考２ 用語集 [4] 等より。)

SMARTS: SMiles ARbitrary Target Specification

SMILESを拡張した、部分構造を表現するための識別子 [5] 。

SMILES: Simplified Molecular Input Line Entry System

化合物の分子構造等を印刷可能な文字で線形表記した識別子 [6]。

US EPA: United States Environmental Protection Agency（米国環境保護庁）

第１章　はじめに

1.1 生態毒性予測システム「KATE」：KAshinhou Tool for Ecotoxicity とは

KATE（https://kate.nies.go.jp）は環境省の請負業務（平成 16 年度から令和 3 年度）として、国立環境研究所 環境リスク・健康研究領域において研究・開発された生態毒性 QSAR システムです。KATE は化学物質の構造をもとに生態毒性を予測します。KATE2020 で予測結果が得られる毒性値は以下の通りです。

- 魚類急性毒性試験 (OECD TG 203) [7]における半数致死濃度（LC50）
- 魚類初期生活段階毒性試験 (OECD TG 210) [8]における無影響濃度（NOEC）
- ミジンコ急性遊泳阻害試験 (OECD TG 202) [9]における半数影響濃度（EC50）
- ミジンコ繁殖試験 (OECD TG 211) [10]における無影響濃度（NOEC）
- 藻類生長阻害試験 (OECD TG 201) [11]における半数影響濃度（EC50）及び無影響濃度（NOEC）

化学物質の CAS 番号検索や構造式エディタを用いた作図等を用いて SMILES 記法による入力を行い、log P を記述子とする回帰式によって QSAR 予測を行います。

KATE2020 の QSAR モデル構築に当たっては、環境省が実施した生態毒性試験結果[12]（魚類急性毒性試験、ミジンコ急性遊泳阻害試験、魚類初期生活段階毒性試験、ミジンコ繁殖試験、藻類生長阻害試験）及び US EPA のファットヘッドミノー・データベースの魚類急性毒性試験結果[13]を用いています。

1.2 KATE2020 技術文書の目的について

本技術文書は、KATE2020 の QSAR モデル導出に関する説明やモデルの性能評価について、OECD の QSAR バリデーション原則[14]に基づいて説明するものであり、KATE2020 を使用したことのある方を対象としています。

KATE2020 の使用方法、開発の経緯や更新履歴等については、KATE2020 操作マニュアル（https://kate2.nies.go.jp/nies/doc/KATEmanual_2020.pdf）を参照してください。

1.3 OECD QSAR バリデーション原則について

本技術文書では、KATE2020 の QSAR モデルについて、OECD QSAR バリデーション 5 原則[14] に沿って説明しています。OECD QSAR バリデーション 5 原則とは、OECD が 2004 年に策定した、QSAR モデルを化学物質の規制に適用する際に正当性・信頼性を確保するために、QSAR モデルが満たすべき５つの原則です。

1. エンドポイントの定義
2. 曖昧さのないアルゴリズム
3. 適用領域の定義
4. 適合度、頑健性、予測性の適切な評価
5. 可能ならば、メカニズムに関する解釈

1.4 log P について

本システムは、化学物質の毒性を予測する際に使用する log P として、US EPA が著作権を有する log P 予測モデル KOWWIN™[5]を US EPA の許諾を得て使用しています。

利用者は下記に示す KOWWIN™ 使用許諾条件について遵守してください。

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 免責事項

KATE の予測結果は十分な予測精度を保証できるものではありません。本システムは、化学物質の生態毒性影響の程度についての参考情報を得るためのツールの一つとしてご利用ください。環境省および国立環境研究所は KATE による毒性予測値を保証するものではなく、また、KATE による毒性予測値の使用により生じた損害については一切の責任を負いません。

また、現時点では KATE による毒性予測結果を「化学物質の審査及び製造等の規制に関する法律（化審法）」に基づく届出に必要な生態毒性試験結果に代替するものとして利用することはできません。

著作権、リンク等については、KATE ウェブサイト内のサイトポリシー（https://kate.nies.go.jp/spolicy.html）をご覧ください。

1.6 謝辞

KATE2020 は下記のソフトウェアまたはライブラリからの結果を使用させていただいております。ここに記して謝意を表します。

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

第２章　エンドポイントの定義　– OECD (Q)SAR バリデーション原則 1

ここでは、KATE2020で予測するエンドポイントについての定義を行います。

2.1 KATE2020 で予測するエンドポイント

KATE2020 では、各予測毒性タイプに対して、化学物質の生態毒性に関する以下のエンドポイント（毒性の指標等）により毒性予測を行います。

表２－１　KATE2020 で予測するエンドポイント

予測毒性タイプ		生物種（学名）	試験法	試験期間	毒性の指標
生物群	急性/慢性				
魚類	急性	ヒメダカ（*Oryzias latipes*）およびファットヘッドミノー（*Pimephales promelas*） *1	魚類急性毒性試験 (OECD テストガイドライン 203) [7]	96 h	LC50
ミジンコ	急性	オオミジンコ（*Daphnia magna*）	ミジンコ急性遊泳阻害試験 (OECD テストガイドライン 202) [9]	48 h	EC50
藻類	急性	*Raphidocelis subcapitata* *2	藻類生長阻害試験 (OECD テストガイドライン 201) [11]	72 h	EC50
魚類	慢性	ヒメダカ（*Oryzias latipes*）	魚類初期生活段階毒性試験 (OECD テストガイドライン 210) [8]	胚期およびふ化後 30 日間*3	NOEC
ミジンコ	慢性	オオミジンコ（*Daphnia magna*）	ミジンコ繁殖試験 (OECDテストガイドライン211) [10]	21 d	NOEC
藻類	慢性	*Raphidocelis subcapitata* *2	藻類生長阻害試験 (OECD テストガイドライン 201) [11]	72 h	NOEC

*1 ヒメダカとファットヘッドミノーの両方について毒性値がある場合、ヒメダカの方を優先的に利用しています。KATE 構築の当初、環境省のヒメダカ試験結果だけでは物質数が少なく精度良い予測式が構築できなかったため、US EPA のファットヘッドミノー試験結果を含めて、魚類急性の QSAR におけるヒメダカ・ファットヘッドミノー生物種間に起因する差異の検討が行われました。その結果、ファットヘッドミノーの毒性値を用いて QSAR 式を作成することで、ヒメダカのみのから作成するよりも精度が良くなることを確認しました。また、ヒメダカ及びファットヘッドミノー両方の実測毒性値が得られている物質について、毒性値に差異があるか確認を行った結果、10 倍以上異なる物質はジメチルアミンのみであった（平成 21 年度化審法審査支援等検討調査報告書）。

*2 過去には *Selenastrum capricornutum* や *Pseudokirchneriella subcapitata* 等の名称が用いられていたことがあります。

*3 魚類初期生活段階試験は魚種やふ化日数によって試験期間が異なりますが、環境省が実施した生態毒性試験で用いられているヒメダカでは「胚期およびふ化後 30 日間」となっています。

2.2 予測値の単位

KATE では、毒性の指標の予測値を[mg/L]の単位で出力します。

2.3 従属変数

KATE では、毒性の指標の値の単位を[mg/L]から[mmol/L]に変換し、その逆数の log を取った log(1/毒性値[mmol/L])を従属変数として使用しています。

2.4 トレーニングセットデータのエンドポイントについて

KATE2020 のトレーニングセットデータの毒性値は、環境省が実施した生態毒性試験結果（魚類急性毒性試験、ミジンコ急性遊泳阻害試験、魚類初期生活段階毒性試験、ミジンコ繁殖試験、藻類生長阻害試験）結果及び US EPA のファットヘッドミノー・データベースの魚類急性毒性試験結果を用いています。

物質数について

KATE2020 における、各予測毒性タイプでのトレーニングセットデータ*1、サポートケミカル*2 およびどの QSAR クラスにも分類されない物質*3 の数は表２－２のようになります（KATE2020 での全ての QSAR クラスを対象にした物質数。１つの物質は、複数の QSAR クラスに分類されていても１つと数えます）。

表２－２　予測毒性タイプに対する物質数

		急性			慢性		
		魚類	ミジンコ	藻類	魚類	ミジンコ	藻類
トレーニングセット	環境省データ	366	451	318	33	347	427
	US EPA データ	506					
サポートケミカル	環境省データ	206	131	201	1	59	90
	US EPA データ	0					
どの QSAR クラスにも分類されない物質	環境省データ	9	15	32	0	12	19
	US EPA データ	8					
合計		1095	597	551	34	418	536

*1 QSARクラスの回帰式の構築に使用される物質。

*2 不等号付きデータ（限度試験等）、外れ値（毒性試験結果の信頼性が確認できない等）のデータおよびlog Pが6より大きくQSARクラスの回帰式に含まれない物質。なお、全物質のlog PはKOWWIN™による推定値を使用します。

*3 KATE2020のどのQSARクラスにも該当しない物質は予測時にUnclassifiedクラスに分類されます。

物質データの詳細については以下を参照ください。
https://kate.nies.go.jp/data/Chemicals_Data.html

第３章 アルゴリズム　– OECD (Q)SAR バリデーション原則 2

ここでは、KATE2020のアルゴリズムについて説明します。3.1で概要を、3.2以降で詳細を説明します。

3.1 アルゴリズムの概要

KATE2020は化学物質（有機化合物）の毒性を予測する線形回帰に基づくQSARです。化学物質の生体膜透過性や生物体内への蓄積性と毒性に相関があると考えられることより、log Pを記述子としています。

環境省の生態影響試験データおよびUS EPAのファットヘッドミノー急性試験データをトレーニングセットデータとして使用します。各物質に含まれる部分構造の有無と個数を特徴量とする決定木により、クラス分類を行っています。

KATE2020には表３－１に示す18の大分類およびその他の分類があります。いずれの分類も特有の部分構造を持っており、各大分類に関して反応性が高く毒性が強くなると考えられる部分構造の有無によりReactiveなクラスとUnreactiveなクラスに分類します。そしてさらなる条件分岐（ハロゲンの有無、芳香族原子の有無等）によって細分化された構造クラス（３章 3.5 参照）に対してQSARクラス（３章 3.6 参照）が割り当てられており、各QSARクラスに回帰式が立てられています。一部のQSARクラスは特定の生物群の毒性予測に特化するようにクラス分類しています。反応性が低く麻酔作用のみを有すると考えらえる部分構造を持つ場合には、Narcotic groupクラスにも分類されます。皮膚感作性[22]などの特殊活性に関係する部分構造は、主に構造に関する適用領域の判定（４章 4.2 A) 参照）に使用され[23]、一部はクラス分類にも使用されます。

予測する化学物質の情報は分子の化学構造を英数字で文字列化したSMILES記法[6]で入力し、ケモインフォマティクスツールCDK[1]を用いて部分構造とその個数を取得します。部分構造はSMILESの拡張表記であるSMARTS記法[5]によって事前に定義されています。部分構造の一覧はWeb上の関連資料（https://kate.nies.go.jp/data/Substructures.htmlおよびhttps://kate.nies.go.jp/data/Substructures_for_judgement.html）を参照してください。抽出された部分構造をもとに上述した決定木によってクラスの割り当てを行います。一つの予測毒性タイプに対して複数のクラスが割り当てられることもあります（KATEのどのクラスにも合致しない部分構造をもつ物質はUnclassifiedクラスに分類され、毒性予測が行われません）。その後、KOWWIN™ [2]によって推定されたlog Pをもとに、割り当てられた各クラスに対して予測毒性値を出力します。log Pの値をユーザが手動で入力することも可能ですが、トレーニングセットデータのlog Pは全てKOWWIN™による推定値を使用しています。

表３－１　KATE2020 の大分類一覧

	大分類	キーとなる部分構造ID	部分構造名	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		
		.		
		.		
		.		

	酸素を含むクラス
	窒素を含むクラス
	酸素と窒素を含むクラス
	リンを含むクラス
	硫黄を含むクラス
	その他のクラス

毒性予測の流れは下記の通りです（図３－１ 参照）。

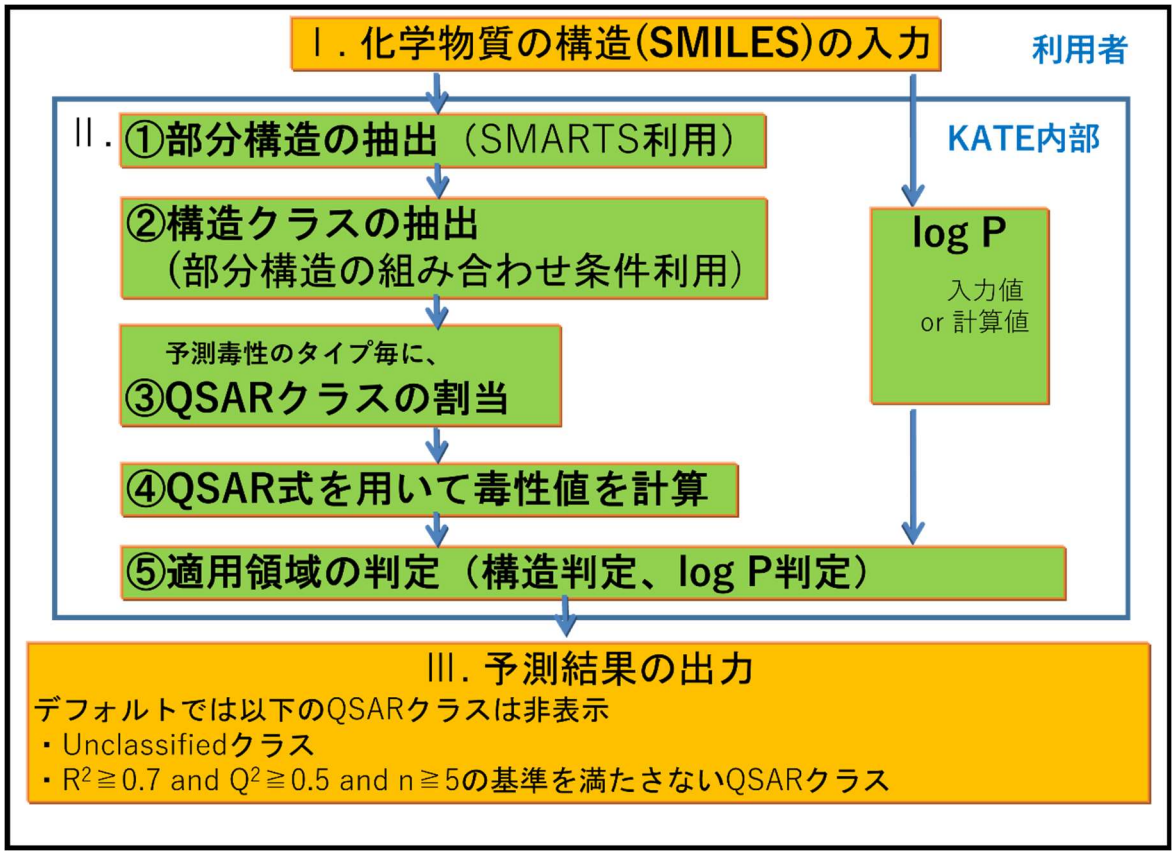


図３－１　KATE2020の毒性予測の流れ

Ⅰ. 利用者による化学物質の構造（SMILES記法による）の入力、log P値の入力（任意）

Ⅱ. QSAR式、毒性予測値の決定、及び適用領域判定に関する処理

① 予測対象物質（入力された化学物質）に対して、部分構造を抽出、及びlog P値を計算（又は入力値の利用）

② 複数の部分構造の組み合わせによる構造クラス*1を抽出

③ 予測毒性タイプごとに、構造クラスに対応するQSARクラス*2を割当（複数のクラスに割当てられる場合もある）

④ 割り当てられたQSARクラスごとに、QSAR式*3 を用いて毒性値を計算

⑤ 適用領域の判定（log P判定と構造判定）

*1 各部分構造の個数条件のAND/ORによる組み合わせにより定義した分類（１５ページ「３章3.5 構造クラスの抽出」参照）

*2 各予測毒性タイプでの物質の構造に基づいて定義した分類。

*3 QSARクラスに含まれるトレーニングセットデータで形成されたモデル。ここではlog Pを記述子とする単回帰式。

Ⅲ．予測結果の出力

予測対象物質が何らかの予測毒性タイプでどのQSARクラスにも分類されなかった場合、Unclassifiedクラスに割当てられます。
デフォルトではUnclassified クラスと統計値の基準（$R^2 \geqq 0.7$、 $Q^2 \geqq 0.5$および $n \geqq 5$）を満たさないQSARクラス*4は非表示にしています。

*4 R^2, Q^2, nはそれぞれ決定係数、内部バリデーションの指標（Leave-one-out法）、トレーニングセットデータ数であり、各QSARクラスに対してあらかじめ計算されています。

具体的な例として、化学物質 1-pyridin-3-ylethanone を予測したときの予測フローを**図３－２**に示します。

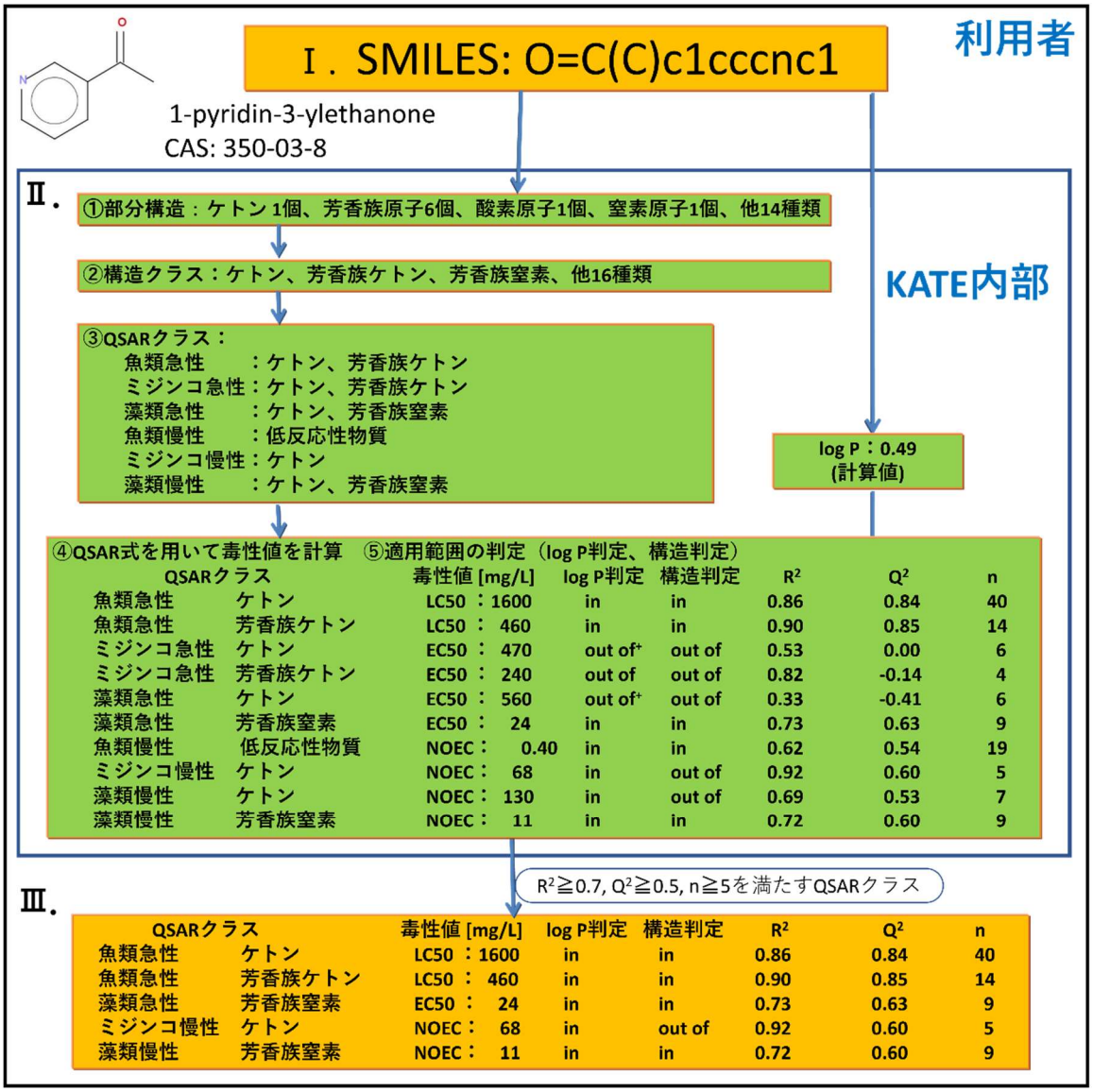


図３－２　KATE2020のQSAR予測フロー（例）

※　図３－２中の、ケトン、芳香族ケトン、芳香族窒素、低反応性物質の、KATE2020における実際の名称は以下のようになっています。

ケトン	: COS_X ketone unreactive
芳香族ケトン	: COS_X ketone unreactive aromatic
芳香族窒素	: CNOS_X aromatic n unreactive
低反応性物質	: CNO_X unreactive (Fish chronic), excl. (CnosX w/o n+)

以降でKATE2020における毒性予測の流れの詳細について、順番に説明していきます。

3.2 ユーザによる化学物質の入力

ユーザが SMILES を直接入力する、または描画した構造式や CAS 番号や物質名から SMILES に変換します。

3.3 log P 値の取得

毒性予測値の計算に使用する予測対象物質のlog Pの値は以下の優先順位で決定します。

1. log Pのユーザ入力値（ユーザによる入力値がある場合に利用される）
2. log Pの推定値（ユーザによる入力値がない場合に利用される）

log P推定値の計算にはUS EPAから許諾を得たKOWWIN™ [2]を利用します。

3.4 部分構造の抽出

部分構造定義一覧（表３－２。各行が１つの部分構造に対応し、SMARTS 記法により定義されています。）を元に、予測対象物質に含まれる部分構造の個数を計算します（表３－３）。

表３－２　部分構造一覧（一部のみ）

部分構造ID	部分構造名	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])]
…		

全ての部分構造一覧については、以下を参照ください。
https://kate.nies.go.jp/data/Substructures.html

SMARTS を利用した部分構造個数計算には CDK ライブラリ[1]を利用しています。

表３－３　予測対象物質に含まれる部分構造一覧
（SMILES: O=C(C)c1cccnc1 の場合）

部分構造ID	部分構造名	SMARTS	個数
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])cccc1);!$(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

部分構造 ID の先頭が 5 で始まるものは構造判定用部分構造（構造判定のために使用される部分構造）としても使用されます（「４章 4.1 A) 構造の適用領域」参照）

3.5 構造クラスの抽出

構造クラス一覧（表３－４。各行が１つの構造クラスに対応し、IDCode* 記法により定義されています。）を元に、予測対象物質に合致する構造クラスを抽出します（表３－５）。

* IDCode は、KATE2020 の構造クラスを表現するために独自に定義した識別子です。今まで定義されている構造クラスと部分構造の個数（構造クラスについては抽出されたものを１、抽出されてないものを０と数える）の and/or 条件で線形表記しています。一度定義した構造クラスは、それ以降に定義する構造クラスの IDCode の中に使用できます（例：表３－４の構造クラス G1_00010 は、それ以降で定義される構造クラス U_10030 の IDCode の中で使用されている）。

IDCode の例：4500,>0;6055,>2/6055,<5/4328,=0|F_00007,>0

ここで「/」,「|」,「;」,「$」はそれぞれ「AND」,「OR」,「AND」,「OR」の条件を意味します（ただし、演算は「/」,「|」,「;」,「$」の順で行います）。上記の例では、以下のように()で囲まれたかのように解釈されます。

4500,>0 AND ((6055,>2 AND 6055,<5 AND 4328,=0) OR F_00007,>0)

上記の「4500,>0」は ID: 4500 の部分構造が 1 個以上存在することを意味します。

例えば、表３－３より予測対象物質が部分構造 3032 を 2 個含み、表３－５より予測対象物質に対して構造クラス R_00001 と G1_00010 が抽出されてないことから、予測対象物質が表３－４の IDCode「3032,>0,/R_00001,=0,/G1_00010,=0,/」の条件式に合致するため、それに対する構造クラス U_10030 が抽出されます。

表３－４　構造クラス一覧（一部のみ）

構造クラスID	大分類	構造クラス名	IDCode
…			
G1_00010	acid	oxoacid [C,c]CO2-, [C,c,O]SO3-	3034,>0,|4760,>0,
…			
R_00001	aldehyde,ketone	C=O reactive (ketone, ester, acid)	3053,>0,|3054,>0|3055,>0,|3056,>0,|4515,>0|3054,>0,|4791,>0,|3036,>0,|6112,>0,|3174,>0,|
F_00007	halogen	halogen not amine, phenol, ...	4507,>0,/4918,=0,/4927,=0,/4917,=0,/3130,=0,
G1_21031	methacrylate	CO_X methacrylate	G1_00031,>0,/3003,=0,/
…			
U_10030	ester	ester unreactive	3032,>0,/R_00001,=0,/G1_00010,=0,/
…			

構造クラス一覧については以下を参照ください。
https://kate.nies.go.jp/data/Structure_Classes.html

表３－５　予測対象物質に合致する構造クラス（一部のみ）
（SMILES: O=C(C)c1cccnc1 の場合）

構造クラスID	大分類	構造クラス名	IDCode
G1_21025	aldehyde,ketone	CO_X ketone unreactive	U_00025,>0,/3011,=0,/
G1_21029	aldehyde,ketone	CO_X ketone unreactive aromatic	G1_21025,>0,/4910,>0,/
G1_22005	aromatic n	CNOSX basic aromatic n reactive	3013,=0,/S_00006,>0,/
G1_25002	Other	CNO_X unreactive (Fish Chronic), excl. (CnosX w/o n+)	U_00002,>0,/3014,>0,/3006,=0,/
G1_41025	aldehyde,ketone	ketone unreactive	U_00025,>0,/3003,>0,/
G1_41026	aldehyde,ketone	ketone unreactive, excl. (CO_X, C(=O)CCN, nitro)	G1_41025,>0,/4791,=0, /4259,=0,/
GA_22075	aromatic n	aromatic n reactive (alga)	SA_00006,>0,/5095,=0,/6983,=0,/6981,=0,/
R_00006	aromatic n	basic aromatic n	3030,>0,|3100,>0,|3104,>0,|3102,>0,|3120,>0,|4511,>0,|3119,>0,|4731,>0,|4813,>0,|

構造クラス ID 先頭のアルファベットの意味は下記の通りです。

- 先頭が A で始まるもの：Acid に関係する構造
- 先頭が C で始まるもの：Carbon を含む構造
- 先頭が R で始まるもの：Reactive に関係する構造
- 先頭が U で始まるもの：Unreactive に関係する構造
- 先頭が GF で始まるもの：Fish に関連する QSAR クラスが設定されている構造
- 先頭が GD で始まるもの：Daphnid に関連する QSAR クラスが設定されている構造
- 先頭が GA で始まるもの：Alga に関連する QSAR クラスが設定されている構造
- 先頭が GFD で始まるもの：Fish と Daphnid に関連する QSAR クラスが設定されている構造
- 先頭が G1 で始まるもの：Fish, Daphnid, Alga のいずれかに関連する QSAR クラスが設定されている構造

3.6 QSAR クラスの割当

QSAR クラス定義一覧（表３－６。各行が１つの QSAR クラスに対応し、予測毒性タイプと構造クラスにより定義されています。）を基に、予測対象物質に対する QSAR クラスを各予測毒性タイプに割当てます（表３－７）。

具体的には、各物質に対して予測毒性タイプごとに、QSAR クラス定義一覧（表３－６）中の、下記の条件を満たす行の QSAR クラスを取得します†*。

① 当該予測毒性タイプが、「予測毒性タイプ」列と一致する。

② 当該物質に対して抽出された構造クラスのIDが、「構造クラスID」列と一致する。

* １つの物質が同じ予測毒性タイプに対して、複数の QSAR クラスに割当てられることもあります。どの QSAR クラスにも分類されなかった場合、Unclassified クラスとなります。

例えば、表３－５より予測対象物質が構造クラス U_10030 を含み、表３－６の QSAR クラス一覧より予測毒性タイプが Daphnid Chronic で構造クラスが U_10030 である組み合わせが存在するので、その予測毒性タイプと構造クラスに対応する QSAR クラス 22103051 が割当てられます。

表３－６　QSAR クラス一覧（一部のみ）

QSAR ID	QSARクラス名	予測毒性タイプ	構造クラスID
12103041	COS_X ester unreactive	Fish Acute	G1_21030
12120341	C_X HC aliphatic w/o X	Fish Acute	G1_21203
12303441	CNOS_X amine aromatic less toxic	Fish Acute	G1_23034
…			
22100741	C_X aromatic HC w/o X, R3=0	Daphnid Acute	G1_21007
22101041	C_X aromatic w/o X, fused R=0	Daphnid Acute	G1_21010
22103041	COS_X ester unreactive	Daphnid Acute	G1_21030
…			
32100741	C_X aromatic HC w/o X, R3=0	Alga Acute	G1_21007
…			
21003051	ester unreactive	Daphnid Chronic	U_10030
…			

全ての QSAR クラスについては以下を参照ください。
https://kate.nies.go.jp/data/QSAR_Classes.html

表３－６の QSAR クラス名の先頭の「COS_X」等は、その QSAR クラスの物質に含まれてもよい元素を示しており、大文字のアルファベットは脂肪族・芳香族の両方、小文字のアルファベットは芳香族のみを意味します。例えば、「COS_X」は脂肪族・芳香族の炭素、酸素、硫黄およびハロゲンを含んでも良いことを意味し、「CXnos」は脂肪族・芳香族の炭素、ハロゲン、および芳香族の窒素・酸素・硫黄を含んでも良いことを意味します。したがって、脂肪族の窒素を含む物質は、「COS_X」もしくは「CXnos」で始まる QSAR クラスには分類されないことになります。

表３－７　予測対象物質に対して割当てられる QSAR クラス一覧
（SMILES: O=C(C)c1cccnc1 の場合）

QSAR ID	QSARクラス名	予測毒性タイプ	構造クラスID
12102541	COS_X ketone unreactive	Fish Acute	G1_21025
12102941	COS_X ketone unreactive aromatic fish	Fish Acute	G1_21029
22102541	COS_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	CNOS_X basic aromatic n unreactive	Alga Acute	GA_22075
12500251	CNO_X unreactive (Fish chronic), excl. (CnosX w/o n+)	Fish Chronic	G1_25002
22102551	COS_X ketone unreactive	Daphnid Acute	G1_21025
32102551	COS_X ketone unreactive	Alga Chronic	G1_21025
32207551	CNOS_X basic aromatic n unreactive	Alga Chronic	GA_22075

QSAR ID（QSAR クラスの ID）の付け方のルールとしては以下のようになっています。

- 8 桁の数字で構成される。
- 1 桁目は 1: 魚類、2: ミジンコ、3: 藻類を表す。
- 2～6 桁目は構造クラス ID の下 5 桁と一致する。
- 7 桁目は 4: 急性、5: 慢性を表す。
- 8 桁目は記述子を表す。現在の公開版では 1: log P 推定値のみ使用している。

予測対象物質の SMILES が CC(=O)c1cccnc1 である場合の、部分構造の抽出から QSAR クラス割当までの流れを簡単にまとめると図３－３のようになります。

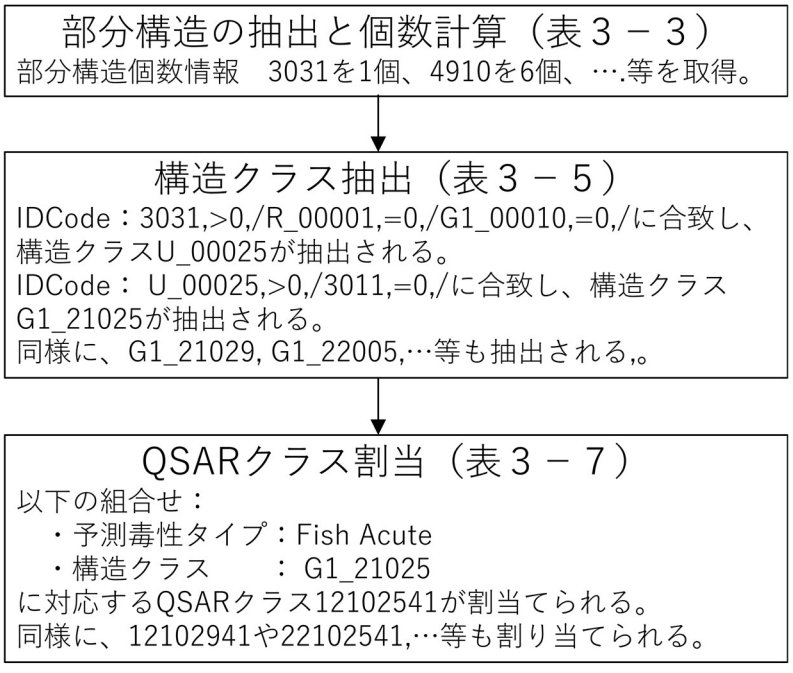


図３－３　部分構造個数計算から QSAR クラス割当までの流れ
（SMILES: O=C(C)c1cccnc1 の場合）

3.7 QSAR 式による毒性値計算

割当てられたQSARクラスごとに、以下のQSAR式に予測対象物質のlog P、および当該QSARクラスに対して予め計算された傾きと切片*2（表３－８）を代入することにより、log(1/毒性値[mmol/L])を計算し、その後、予測対象物質の分子量*1を用いて毒性値[mg/L]に単位変換します。

$$\log(1/\text{毒性値[mmol/L]}) = \text{傾き} \times \log P + \text{切片} \quad \cdots (1)$$

*1 予測対象物質の分子量の計算にはOpen Babel [15] を利用します。

*2 傾きと切片は、当該QSARクラスのトレーニングセットデータの記述子 log Pと従属変数 log(1/毒性値[mmol/L])に対する単回帰により取得します。トレーニングセットのlog Pは全てKOWWIN™による推定値を利用しています。

3.8 予測区間の計算（参考情報）

割当てられたQSARクラスごとに、予測対象物質のx (=log P)の値、および当該QSARクラスに対して予め計算された統計値 t_{95},n,V_ε,$\bar{x}$, Σ^{-1}（表３－８, ３－９）を代入することにより、下記で定義されるlog(1/毒性値[mmol/L])の信頼水準95%予測区間を計算します。その後、予測対象物質の分子量を用いて、(2)式の予測区間の下限値と上限値を毒性値[mg/L]に単位変換します。

$$95\%\text{予測区間} = [\log(1/\text{毒性値[mmol/L]}) - dy,\ \log(1/\text{毒性値[mmol/L]}) + dy] \quad \cdots (2)$$

ここで、

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

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

表３－８　予測区間の計算に必要な統計値

変数	説明	式
n	当該QSARクラスのトレーニングセットデータ数	
$\bar{x}$	当該QSARクラスのトレーニングセットデータに対する記述子の平均値	$\sum x_i/n$ x_i：i番目のトレーニングセットデータのlog P値
Σ^{-1}	記述子の共分散行列の逆行列 （ここでは単回帰なので分散の逆数）	$n/\sum(x_i-\bar{x})^2$ x_i：i番目のトレーニングセットデータのlog P値 $\bar{x}$：トレーニングセットデータのlog Pの平均値
V_ε	残差分散	$\Sigma(y_i-\hat{y}_i)^2/(n-p-1)$ y_i: i番目のトレーニングセットデータの実測毒性値 $\hat{y}_i$: i番目のトレーニングセットデータの予測毒性値 p: 記述子の数（ここでは単回帰なので1）
t_{95}	当該QSARクラスの自由度に対する5%有意水準(両側検定）でのt値	

表３－９　QSAR式の傾きと切片、**および予測区間の計算に必要な統計値の例**

QSAR ID	傾き	切片	n	$\bar{x}$	Σ^{-1}	V_ε	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 類似度の計算（参考情報）

Fingerprint* を用いることにより、割当てられた QSAR クラスに含まれる各物質（トレーニングセットデータおよびサポートケミカル）に対して、予測対象物質との類似度（ここでは、PubChem fingerprint [24] を用いた Tanimoto 係数）を計算します。類似度は予測結果として必須情報ではありませんが、参考情報として利用できます。KATE2020 に含まれる全物質の Fingerprint は予め計算されています。

QSAR クラスに含まれる物質のビット列 X と予測対象物質の Fingerprint のビット列 Y に対する類似度 T を以下の式により計算します。

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

ただし、

N_x : X の中で 1 になっているビット数
N_y : Y の中で 1 になっているビット数
N_c : X と Y で共通して 1 になっているビット数

数値は 0 と 1 の範囲を取り、予測対象物質との類似性が高いものほど 1 に近い値を取ります。

* 化学物質の情報を表現する固定長のビット列で、各ビットはそれぞれ化学物質の特徴の有無を示しており、1の場合は化学物質に当該ビットに対応する特徴があること、0の場合はないことを示しています。

Fingerprintの表記例：**1001001000000000110010110000101010000110**

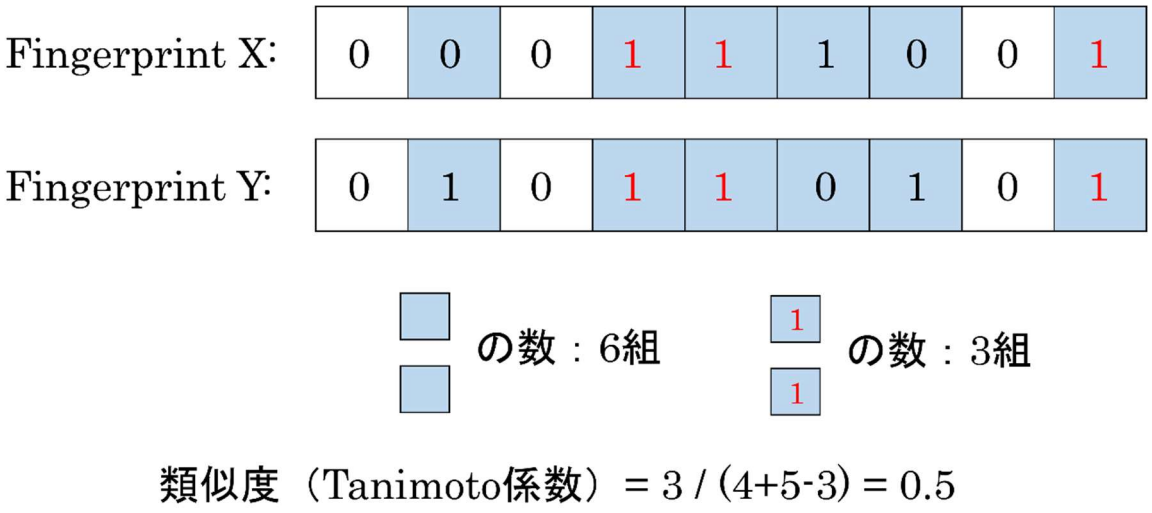


図３－４　類似度（Tanimoto 係数）計算のイメージ

類似度は QSAR クラスの詳細表示画面（Verify QSAR 画面）で確認することができます。Verify QSAR 画面については KATE2020 操作マニュアル† 「6. QSAR クラス情報の詳細表示（Verify QSAR 画面）」を参照してください。

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

第４章　適用領域の定義　– OECD (Q)SAR バリデーション原則 3

ここでは、KATE2020のQSARモデルにおける適用領域、および適用領域の判定方法について説明します。

4.1 適用領域

KATE2020では適用領域として、A) 構造の適用領域とB) log Pの適用領域があります。

A)　構造の適用領域

構造の適用領域はQSARクラスごとに設定されており、当該QSARクラスのトレーニングセットデータおよび不等号付きデータの物質（当該QSARクラスのlog P判定（4.2のBで説明）が適用領域内のもののみ）が持つ、構造判定用部分構造リスト（表４－１）によって与えられます。

表４－１　各 QSAR クラスの構造判定用部分構造のリスト（一部のみ）

QSAR ID	構造判定用部分構造のリスト
…	
12100241	5004,5007,5008,5016,5022,5023,5028,5081,5500
22100741	5074, 5075
32100541	5020,5021,5067,5155
…	

構造判定用部分構造一覧は以下を参照ください。
https://kate.nies.go.jp/data/Substructures_for_judgement.html

B)　log P の適用領域

log P判定（4.2のBで説明）のために、各QSARクラスに対して、以下のlog P範囲が予め計算されています（表４－２）。

- log P範囲１：当該QSARクラスのトレーニングセットデータのlog Pの最小値と最大値　…　”in”または“out of”の判定のために必要
- log P範囲２：当該QSARクラスのトレーニングセットおよびサポートケミカルのlog Pの最小値と最大値　…　”out of+”の判定のために必要

log Pの適用領域は、上記のlog P範囲１によって与えられます。

表４－２　各 QSAR クラスの log P 範囲

QSAR ID	log P範囲１	log P範囲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 適用領域の判定方法

KATE2020では、予測対象物質の予測毒性値が、予測結果として適用できる範囲内にあるかどうかを判定します。A) 構造による判定（構造判定）とB) log Pによる判定（log P判定）の２つを行い、両方とも適用領域内の場合に、KATE2020での予測毒性値が適用領域内と判定されます。

A） 構造判定

KATE2020では、「構造判定用部分構造」の比較により、予測対象物質の構造が当該QSARクラスの適用領域内であるかどうかについて判定します（図４－１）。判定結果には下記の３つの場合があり、「in」又は「 in (conditionally)」の場合に当該QSARクラスを「構造に関して適用領域内」と判定しています。

in：適用領域内

予測対象物質に含まれる「構造判定用部分構造」の全てが、当該 QSAR クラスに含まれる物質*1 が持つ「構造判定用部分構造リスト」（表４－１参照）に含まれる場合（図４－１におけるピンク色の範囲）、または予測対象物質が持つ部分構造に「構造判定用部分構造」が１つも含まれていなかった場合。

in (conditionally)：条件付き適用領域内

「in」の条件には合致しないが、予測対象物質の「構造判定用部分構造」全てが、当該 QSAR クラスの「構造判定用部分構造リスト」、あるいは Narcotic Group*2 に含まれる物質*1 が持つ「構造判定用部分構造リスト」に含まれる場合（図４－１におけるピンク色と黄色の部分）。

out of：適用領域外

「in」と「in (conditionally)」のいずれの条件にも合致しない場合。すなわち、予測対象物質の「構造判定用部分構造」に、当該 QSAR クラスと Narcotic Group*2 クラスが持つ「構造判定用部分構造リスト」に含まれない部分構造がある場合（図４－１において灰色部分の構造が含まれる場合）。

*1 ここでは、log P 判定が”in”（適用領域内）である不等号付きデータも含みます。

*2 反応性が低く特異的な生理活性作用に基づかないベースライン毒性（麻酔作用）。KATE2020 では、脂肪族炭化水素、スルホキシド、脂肪族・芳香族エーテル、脂肪族・芳香族ケトン、アルコールといった単純な麻酔作用のみで毒性が説明できると考えられる QSAR クラスが予測毒性タイプ毎に用意されており、これらをまとめた QSAR クラスが Narcotic Group として各予測毒性タイプで再定義されています。

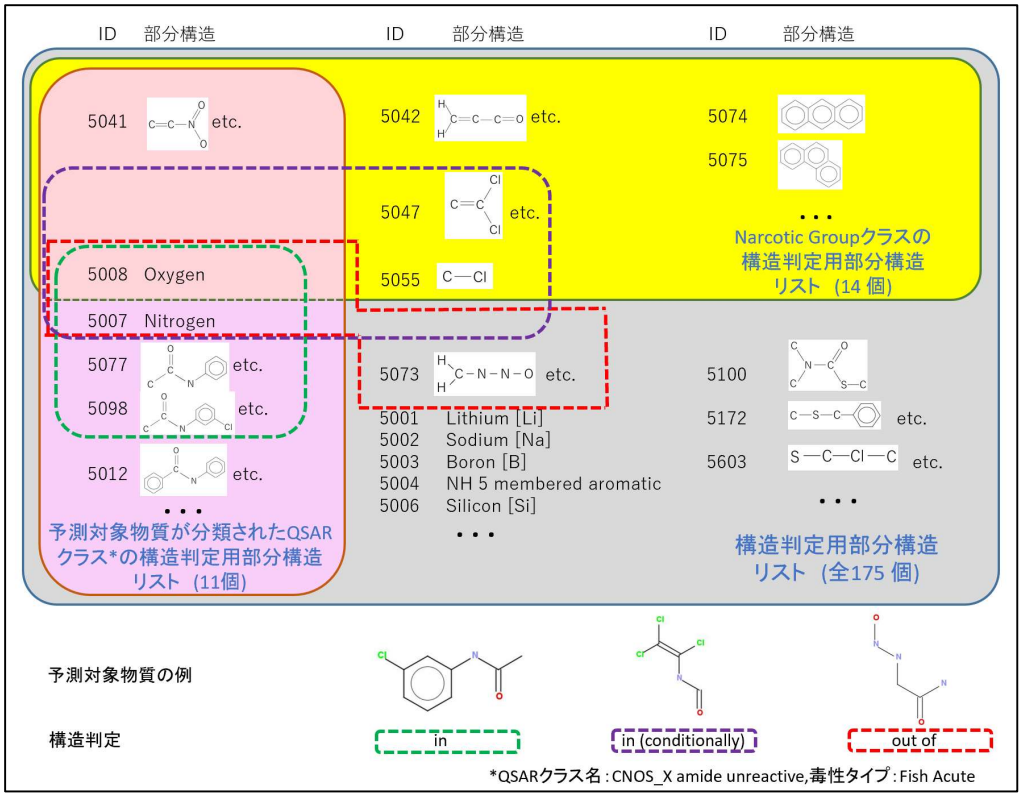


図４－１　構造判定の例

B） log P判定

KATE2020では、予測対象物質のlog P値が当該QSARクラスの全トレーニングセットデータ（サポートケミカルデータは含まない）のlog Pの最小値と最大値（表４－２の当該QSARクラス行の「log P範囲１」欄）の間にあるかどうかで適用領域内にあるかどうかを判定します。なお、KATE2020では、log P>6.0*の物質は全て適用領域外としています（KATE2020版における変更点）。

* ここで閾値として設定した 6.0 は、ECOSAR において急性毒性予測値のカットオフ値が 5.0（一部は 6.4 に設定）、慢性毒性予測値のカットオフ値が 8.0 に設定されていること、log P と log BCF との線形性の上限（たとえば、Dimitrov et al. SAR QSAR Environ Res., 13, 177-184, 2010 では上限値として 6.1～6.5）、疎水性の高い（log P＞4）物質について log P を実測する HPLC 法（OECD Test Guideline 117 で規定）の適用範囲の上限が 6 であること等を総合して設定した。

in：適用領域内（図４－２参照）。
out of：適用領域外。ただし、下記のout of+になる場合は除く（図４－３参照）。
out of+：適用領域外。ただし、予測対象物質の log P 値は当該 QSAR クラスのトレーニングセットデータと参考情報であるサポートケミカルを併せた全物質の log P の最小値と最大値（表４－２の当該 QSAR クラス行の「log P 範囲２」）の内側に存在します（図４－４参照）。

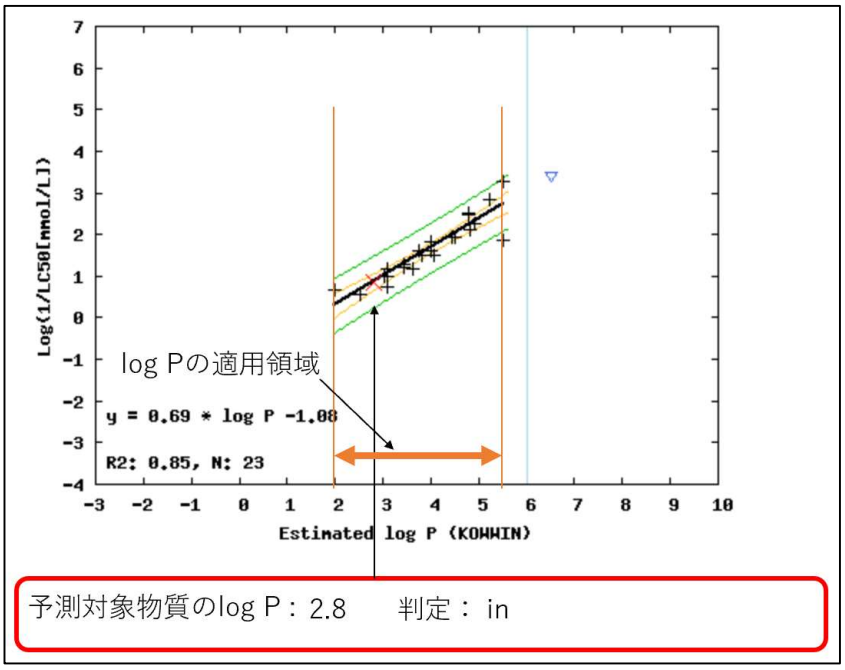
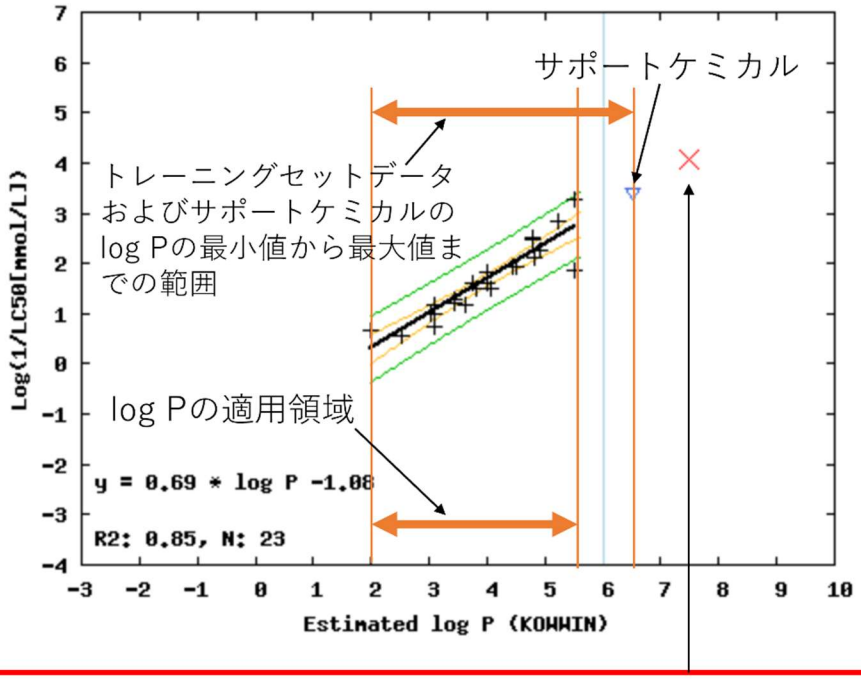
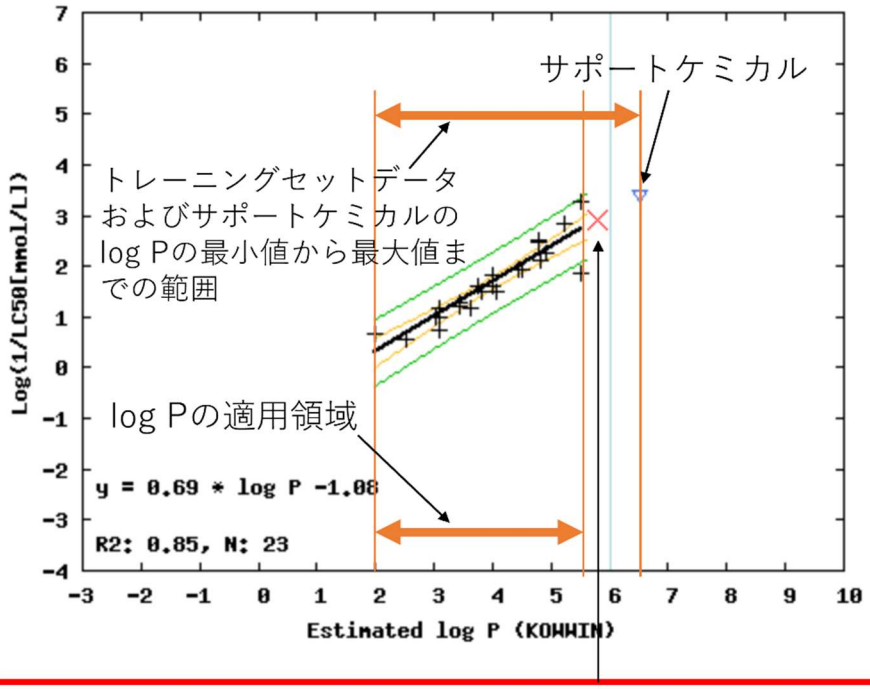


図４－２　log P判定の例（in）



予測対象物質のlog P：7.5　　判定： out of

図４－３　log P判定の例（out of）



予測対象物質のlog P：5.8　　判定： out of+

図４－４　log P判定の例（out of+）

第５章　バリデーション – OECD (Q)SAR バリデーション原則 4

5.1 内部バリデーション

ここでは、KATE2020のQSARモデルの適合度（モデルがトレーニングセットデータにおける応答、すなわち予測毒性値の分散をどれくらいよく説明しているか）および頑健性（トレーニングセットから一つ以上のデータを取り除いた時の予測の安定性）の評価結果について説明します。

5.1.1 内部バリデーションで使用する指標

内部バリデーションは表５－１の指標を利用しています。

表５－１　内部バリデーションの指標

指標	説明
R^2	適合度の指標（決定係数）。0から1の間の値を取り、１に近いほど適合度が高い。 $R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$ y_i：i番目のトレーニングセットデータの実測毒性値 $\hat{y}_i$：i番目のトレーニングセットデータの予測毒性値 $\bar{y}$：トレーニングセットデータの実測毒性値の平均値 n：トレーニングセットデータ数
Q^2	頑健性の指標（leave-one-out法）。１に近いほど頑健性が高く、負の値を取ることもある。 $Q^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_{i,-i})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$ y_i：i番目のトレーニングセットデータの実測毒性値 $\hat{y}_{i,-i}$：i番目のトレーニングセットデータを除いた残りのトレーニングセットデータで構築した回帰式による予測毒性値 $\bar{y}$：トレーニングセットデータの実測毒性値の平均値 n：トレーニングセットデータ数

5.1.2 内部バリデーションの結果

内部バリデーションの結果、統計値基準（$R^2 \geqq 0.7$ and $Q^2 \geqq 0.5$ and $n \geqq 5$）[25-30]を満たすQSARクラスを表５－２に示します。

表５－２　統計値基準を満たすQSARクラス一覧
（同じ生物群、急性/慢性の中では*n*の順）

QSARクラス	生物群	急性/慢性	R2	Q2	n
narcotic group Fish Acute	魚類	急性	0.87	0.87	154
CNOS_X halogen unreactive	魚類	急性	0.76	0.75	95
phenol unreactive unhindered	魚類	急性	0.88	0.87	58
phenol unreactive unhindered excl. Diphenylether HRAC Ea	魚類	急性	0.87	0.86	57
CO_X alcohol unreactive w/o EO Fish	魚類	急性	0.89	0.88	46
CO_X ether unreactive	魚類	急性	0.87	0.86	44
COns_X ketone unreactive	魚類	急性	0.84	0.82	41
CNOS_X aromatic n unreactive, excl. triazine Fish	魚類	急性	0.76	0.74	39
CNO_X ester unreactive	魚類	急性	0.72	0.68	37
phenol unreactive unhindered w/o X	魚類	急性	0.89	0.88	37
Cnos_X heteroaromatic unreactive	魚類	急性	0.84	0.81	30
CNO_X nitro mono unreactive	魚類	急性	0.74	0.7	28
COns_X ketone unreactive aliphatic	魚類	急性	0.88	0.86	27
amine primary unreactive NH2 =1 aliphatic	魚類	急性	0.84	0.81	26
C_X hydrocarbon unreactive aliphatic w/ X, excl. Halomethane	魚類	急性	0.87	0.86	24
C_X hydrocarbon unreactive aromatic w/o X, fused R=0	魚類	急性	0.85	0.79	24
amine primary unreactive aromatic w/ NO2,SO	魚類	急性	0.82	0.79	24
CO_X primary alcohol	魚類	急性	0.92	0.9	22
CNOS_X amine sec, tert w/o n	魚類	急性	0.91	0.89	21
CNO_X amide unreactive	魚類	急性	0.8	0.76	21
C_X hydrocarbon unreactive aliphatic w/o X	魚類	急性	0.73	0.68	21
Cnos_X heteroaromatic unreactive Fish, Daphnid	魚類	急性	0.82	0.78	21
CNOS_X amine aromatic w/ aliphatic carbon	魚類	急性	0.77	0.66	19
CNO_X aldehyde general aromatic	魚類	急性	0.85	0.81	19
CN_X amine sec, tert unreactive aliphatic	魚類	急性	0.73	0.64	16
CNO_X amine sec, tert unreactive aliphatic	魚類	急性	0.91	0.86	16
CN_X nitrile unreactive	魚類	急性	0.87	0.84	15
COS_X ketone unreactive aromatic	魚類	急性	0.9	0.85	14
amine primary unreactive NH2 >1	魚類	急性	0.82	0.75	14
CNOS_X amine primary reactive w/o ortho,para-OH,NH2	魚類	急性	0.83	0.76	14
CNOS_X acid unreactive	魚類	急性	0.77	0.63	12

QSARクラス	生物群	急性/慢性	R2	Q2	n
CN_X amine sec, tert unreactive aromatic	魚類	急性	0.84	0.78	12
Cnos_X heteroaromatic reactive Fish	魚類	急性	0.72	0.59	12
n+, N+	魚類	急性	0.75	0.61	11
CNOS_X aromatic n reactive, excl nitrile	魚類	急性	0.73	0.53	11
phenol reactive w/o ortho,para-OH,NH2, w/o nitro	魚類	急性	0.75	0.64	10
urea unreactive	魚類	急性	0.95	0.87	9
CNOS_X carbamate unreactive Fish	魚類	急性	0.9	0.68	8
amide reactive, excl. C=O,S,N	魚類	急性	0.92	0.85	8
CS_X sulfide unreactive	魚類	急性	0.7	0.59	8
ester reactive methacrylate	魚類	急性	0.76	0.54	8
COS_X methacrylate	魚類	急性	0.87	0.66	7
CN_X nitrile unreactive aliphatic	魚類	急性	0.97	0.95	7
CNOS_X N-hetero unreactive w/o amine, aldoxime, carbamate	魚類	急性	0.79	0.55	6
COS_X thiol	魚類	急性	0.96	0.92	6
CNOSP_X phosphorus unreactive	魚類	急性	0.94	0.87	6
CNOS_X amine tert unreactive w/ C=O	魚類	急性	0.88	0.61	5
CO_X alcohol unreactive w/ EO	魚類	急性	0.98	0.97	5
phenol unreactive bisphenol	魚類	急性	0.87	0.63	5
C_X hydrocarbon unreactive halomethane	魚類	急性	0.94	0.85	5
narcotic group Daphnid Acute	ミジンコ	急性	0.71	0.7	83
CNOS_X halogen unreactive	ミジンコ	急性	0.86	0.84	43
phenol unreactive unhindered	ミジンコ	急性	0.82	0.78	28
phenol unreactive unhindered excl. Diphenylether HRAC Ea	ミジンコ	急性	0.8	0.76	27
phenol unreactive unhindered w/o X	ミジンコ	急性	0.87	0.83	19
C_X hydrocarbon unreactive aromatic w/o X, fused R=0	ミジンコ	急性	0.8	0.74	17
CNOS_X aromatic n unreactive Daphnid	ミジンコ	急性	0.85	0.82	17
CO_X ether unreactive	ミジンコ	急性	0.83	0.74	15
CO_X alcohol unreactive w/o EO Daphnid	ミジンコ	急性	0.78	0.72	14
C_X hydrocarbon unreactive aliphatic w/ X	ミジンコ	急性	0.82	0.77	14
amine primary unreactive NH2 >1	ミジンコ	急性	0.71	0.6	12
phenol unreactive hindered	ミジンコ	急性	0.76	0.64	11
Cnos_X heteroaromatic unreactive	ミジンコ	急性	0.94	0.9	11
CNO_X amine sec mono w/o n Daphnid	ミジンコ	急性	0.75	0.58	9
CNO_X ester unreactive Daphnid	ミジンコ	急性	0.93	0.86	8

QSARクラス	生物群	急性/慢性	R2	Q2	n
CNO_X nitro mono unreactive Daphnid	ミジンコ	急性	0.85	0.74	8
C_X unreactive aliphatic w/ X, excl. gem,vic-Cl, TCE	ミジンコ	急性	0.98	0.97	7
Cnos_X heteroaromatic unreactive Fish, Daphnid	ミジンコ	急性	0.96	0.9	7
CNOS_X N-hetero unreactive w/o amine, aldoxime, carbamate	ミジンコ	急性	0.78	0.63	6
CNO_X amide unreactive Daphnid	ミジンコ	急性	0.88	0.7	6
CN_X amine sec, tert unreactive aromatic	ミジンコ	急性	0.97	0.92	6
CN_X amine sec, tert unreactive aliphatic	ミジンコ	急性	0.89	0.64	6
CO_X primary alcohol	ミジンコ	急性	0.95	0.76	6
CNOS_X amine sec, tert multi	ミジンコ	急性	0.79	0.53	5
n+, N+	ミジンコ	急性	0.95	0.89	5
CNO_X imide unreactive	ミジンコ	急性	0.78	0.59	5
ester reactive methacrylate	ミジンコ	急性	0.82	0.6	5
narcotic group Alga Acute	藻類	急性	0.76	0.74	52
phenol unreactive unhindered excl. Diphenylether HRAC Ea	藻類	急性	0.8	0.77	26
aromatic n reactive Alga	藻類	急性	0.8	0.75	11
CO_X ether unreactive excl. HRAC Ea Alga	藻類	急性	0.93	0.85	9
CO_X alcohol unreactive w/o halogen, acid, EO	藻類	急性	0.95	0.9	6
CNO_X ester unreactive Alga	藻類	急性	0.96	0.9	6
CO_X primary alcohol	藻類	急性	0.91	0.79	6
C_X hydrocarbon unreactive aliphatic w/ X, excl. Halomethane	藻類	急性	0.97	0.91	5
Cnos_X heteroaromatic excl. pyridine Alga	藻類	急性	0.83	0.52	5
narcotic group Fish Chronic	魚類	慢性	0.82	0.75	12
Cnos_X unreactive Fish Chronic	魚類	慢性	0.76	0.68	12
C_X hydrocarbon unreactive	魚類	慢性	0.78	0.68	11
narcotic group Daphnid Chronic	ミジンコ	慢性	0.7	0.68	74
C_X hydrocarbon unreactive aromatic w/o X, fused R=0	ミジンコ	慢性	0.87	0.84	15
CNO_X amine sec, tert unreactive w/ N-Oxide,Nitroso	ミジンコ	慢性	0.81	0.74	15
CO_X alcohol unreactive w/o EO Daphnid	ミジンコ	慢性	0.82	0.75	14
CO_X ether unreactive	ミジンコ	慢性	0.88	0.76	10
CNO_X ester unreactive Daphnid	ミジンコ	慢性	0.84	0.73	8
CNO_X amide unreactive Daphnid	ミジンコ	慢性	0.83	0.74	8
Cnos_X heteroaromatic unreactive	ミジンコ	慢性	0.83	0.64	7

QSARクラス	生物群	急性/慢性	R2	Q2	n
Cnos_X heteroaromatic unreactive Fish, Daphnid	ミジンコ	慢性	0.83	0.64	7
C_X unreactive aliphatic w/ X, excl. gem,vic-Cl, TCE	ミジンコ	慢性	0.98	0.97	6
COns_X ketone unreactive	ミジンコ	慢性	0.92	0.59	5
phenol unreactive unhindered	藻類	慢性	0.7	0.64	27
CNO_X amine sec, tert unreactive w/ N-Oxide,Nitroso	藻類	慢性	0.72	0.65	21
phenol unreactive w/o X, HRAC Ea	藻類	慢性	0.72	0.65	18
CO_X ether unreactive excl. HRAC Ea Alga	藻類	慢性	0.89	0.86	15
CNO_X nitro mono unreactive	藻類	慢性	0.78	0.54	13
aromatic n reactive Alga	藻類	慢性	0.77	0.7	11
CO_X alcohol unreactive w/o halogen, acid, EO	藻類	慢性	0.87	0.81	10
amine primary unreactive NH2 >1, Nv3 <3	藻類	慢性	0.78	0.7	10
CNOSP_X phosphorus all	藻類	慢性	0.74	0.64	9
CNO_X ester unreactive Alga	藻類	慢性	0.94	0.88	8
CNOS_X sulfur reactive excl. disulfide Alga	藻類	慢性	0.77	0.58	8
CNO_X amine sec, tert unreactive aliphatic	藻類	慢性	0.9	0.79	6
COS_X thiol	藻類	慢性	0.81	0.63	6
CNOSP_X phosphorus unreactive	藻類	慢性	0.95	0.86	5
amine sec, tert reactive w/ ortho,para-N,OH	藻類	慢性	0.89	0.76	5
CNOS_X N-hetero unreactive w/o amine, aldoxime, carbamate	藻類	慢性	0.8	0.57	5

5.2 外部バリデーション

ここでは、外部バリデーションによるKATE2020のQSARモデルに対する予測性能（モデルの開発には使われていない新たなデータをどれくらいよく予測できるか）の評価結果について説明します。

5.2.1 外部バリデーションに用いたデータ（テストセット）

外部バリデーションには、OECD SIDS (Screening Information DataSet) に掲載されている毒性値をテストセットとして利用しました。標準試験期間±24時間から逸脱しているものおよび信頼性（Klimischコード）が3もしくは4のものは除外し、KATE2020には含まれない生物種を用いた試験結果は採用しています。また、1物質に複数の試験結果が存在する場合、信頼性が高いデータを採用しています。

予測毒性タイプごとのテストセットの物質数およびKATE2020のクラスで予測値が得られたデータ数は表５－３のとおりです。1物質に複数の予測値が得られた場合はすべての予測値を含めています。

表５－３　テストセットのデータ数

	急性			慢性		
	魚類	ミジンコ	藻類	魚類	ミジンコ	藻類
物質数	178	196	141	3	46	100
予測値の数	197	207	65	1	56	77

5.2.2 結果

表５－３のデータについて、適用領域内（構造判定はin(conditionally)も含める）の予測値と実測値の差が1オーダー（1/10から10倍）の範囲に入る割合を表５－４に、予測値対実測値のグラフを図５－１に示します。

表５－４　全QSARクラスの予測値と実測値の差が１オーダーの範囲に入るデータ数と割合

	急性						慢性					
	魚類		ミジンコ		藻類		魚類		ミジンコ		藻類	
	データ数	割合	データ数	割合	データ数	割合	データ数	割合	データ数	割合	データ数	割合
予測値と実測値の差が１オーダーの範囲に入る	172	87%	150	72%	54	83%	0	0%	41	73%	52	68%
全データ	197		207		65		1		56		77	

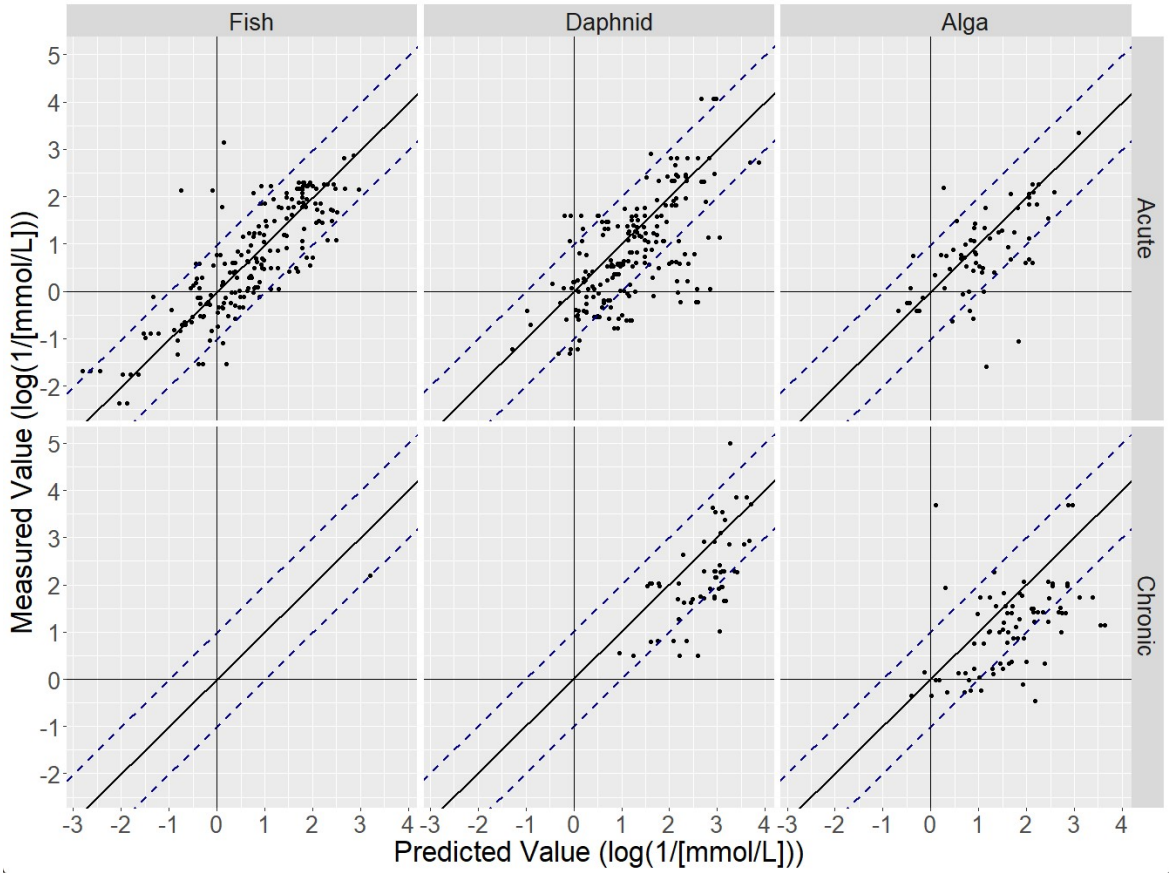


図５－１　全QSARクラスによる予測値対実測値のグラフ

統計値基準を満たすQSARクラス（表５－２参照）に割り当てられ、適用領域内（log P判定および構造判定がともにin）となったものについて（構造判定についてはin(conditionally)も含めます）、得られたデータ数および予測値と実測値の差が1オーダーに入るデータ数とその割合を表５－４に、予測値対実測値のグラフを図５－２に示します。

表５－５　統計値基準を満たすQSARクラスの予測値と実測値の差が１オーダーの範囲に入るデータ数と割合

	急性						慢性					
	魚類		ミジンコ		藻類		魚類		ミジンコ		藻類	
	データ数	割合	データ数	割合	データ数	割合	データ数	割合	データ数	割合	データ数	割合
予測値と実測値の差が１オーダーの範囲に入る	133	89%	79	75%	19	83%	0	NA	15	83%	16	67%
全データ	150		106		23		0		18		24	

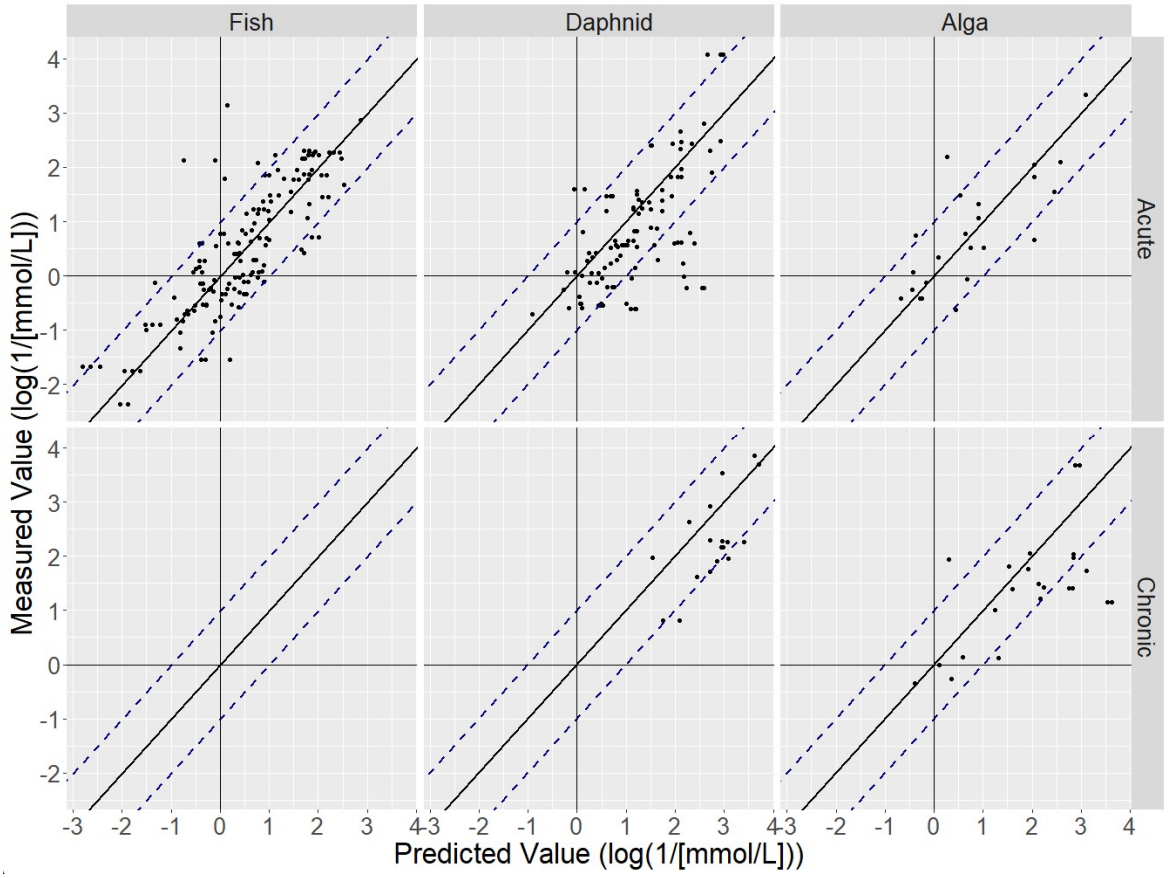


図５－２　統計値基準を満たすQSARクラスによる予測値対実測値のグラフ

第６章　メカニズムに関する解釈 – OECD (Q)SAR バリデーション原則 5

疎水性（細胞膜透過性）と毒性には、両者に対数を取ると線形関係があることが知られており[31]、KATE2020では説明変数をlog P、被説明変数を毒性値（影響濃度）とするQSAR式を構築しています。KATEの各QSARクラスは、特徴的な構造を持つ化学物質で構成されており、構造を特徴づける「構造クラス」は　部分構造の個数条件により定義されています（１５ページ ３章 3.5 参照）。反応性が高いと考えられる部分構造を持つ物質はReactiveな構造クラスに分類し、非特異的で高い反応性によって毒性を発現すると考えられます。そのような部分構造を持たない場合は反応性が低いと考えられ、Unreactiveな構造クラスに分類します [32]。一部の構造クラスには、特異的なメカニズムによって毒性を発現する部分構造が含まれます。構造クラスおよび部分構造の一覧については下記を参照してください。

構造クラス一覧：https://kate.nies.go.jp/data/Structure_Classes.html
部分構造一覧：https://kate.nies.go.jp/data/Substructures.html

参考文献

[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/
（上記URLはすべて2023年3月1日アクセス）

[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.github.io/cdk/1.5/docs/api/org/openscience/cdk/fingerprint/PubchemFingerprinter.html（2023年3月1日アクセス）

[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