新規フィンガープリント及び機械学習による魚類慢性生態毒性値の予測手法開発

高田 道義^{a*}、林 彬勒^b、寺田 昭彦^{a**}、細見 正明^a

^a東京農工大学大学院工学府 応用化学専攻(〒184-0012 東京都小金井市中町2-24-16)

^b産業技術総合研究所 安全科学研究部門(〒305-0053 茨城県つくば市小野川16-1)

(Received November 25, 2019; Accepted December 19, 2019)

動物試験の削減を目的とした化学物質における生態毒性値の予測手法として、官能基で化学物質を分類後、線形回帰で毒性値を予測する方法が一般的である(従来法)。従来法は官能基を複数持つ化学物質に対し複数の結果を出力し、金属化合物や電解質は扱えない。著者らは化学物質をクラスタリングし、生じたクラスタとの類似度から毒性を予測する手法の開発を行ってきたが、慢性生態毒性に関しては精度が従来法に比べて劣っていた。これは、急性に比べ慢性毒性試験のデータ数が少ない事に加え、作用機序が多様であることが原因と推定された。本研究では多様性への対策として、生態毒性用のフィンガープリントの開発を行った。更にこの新しいフィンガープリントと既存の特徴量を併用し、機械学習モデルの結果を特徴量として利用する方法を用いた生態慢性毒性の予測手法を開発した。従来法の課題を克服し、同等以上の予測精度を得た。

キーワード: 生態毒性予測、機械学習、アンサンブル学習、AIST-MeRAM、ECOSAR

1. 序論

化学物質の環境に対する影響(生態毒性)の評価を目的に動物試験が実施されている。一方で、欧州を中心とした近年の動物愛護主義の台頭により、*in silico*による生態毒性予測手法が注目されている。従来の生態毒性予測手法は、専門家が化学物質の官能基を基準に専門家が手動で分類し、生

物濃縮性の指標となるLog Kowや炭素数など、化学物質の記述子を説明変数(特徴量)とした直線回帰モデルで生態毒性を予測する手法が主流である(従来法)。これらの従来法として、米国のECOLOGICAL Structure-Activity Relationship (ECOSAR) [1]、EUのQSAR Toolbox[2]、日本のKATE[3]などが利用されている。この直線回帰モデルは、複数の官能基を持つ化学物質に対して複数の予測結果を提示する、あるいは金属元素及び電離した物質(以後、金属化合物)の予測ができない課題がある。

* s183801z@st.go.tuat.ac.jp

** akte@st.go.tuat.ac.jp

著者らは毒性試験データベース(AIST-MeRAM DB[4]; Multi-Purpose ecological Risk Assessment and Management Tool)のデータを元に、①急性生態毒性値予測手法の開発、②慢性生態毒性値予測手法の開発、③説明可能な機械学習モデルの開発、の3項目に関して検討を行ってきた。一方で、②慢性生態毒性値予測に関しては、現時点で信頼性の高い手法は開発されておらず、大きな課題として残されている。

急性生態毒性の予測を目的に開発した手法は生態毒性値既知の化学物質に対してフィンガープリントにより特徴量を抽出後、特徴量を基準に化学物質をクラスタリングし、生成したクラスタとの類似度を指標として生態毒性値未知物質の毒性値予測を行う方法である[5]。これにより、従来法の課題を克服するとともに、急性生態毒性の予測に関しても、従来法と同等以上の精度を得た。

一方で、この開発手法は慢性生態毒性の予測精度が従来法に比べて劣っていた。慢性生態毒性の教師データが少ない事が一つの原因である。また、エラの破壊など呼吸器系への障害が主因の急性生態毒性に比べ、作用機序の多くは解明されていないため証明は困難だが、慢性生態毒性は作業機序の多様性が高いと推定される事も、開発手法の予測精度が従来法に劣った原因と考えられた。

多様性への対策として、本研究では生態毒性予測用の開発したフィンガープリントを開発した。この開発したフィンガープリントは一般的なフィンガープリント単独では分離困難な、個別の金属元素や農薬の分離が可能である。

前回は機械学習モデルとして類似度を用いたが、今回は作用機序の多様性に対する対策から、特徴量を直接使用して予測する手法の開発に取り組んだ。

化学物質から抽出した一般的な特徴量と、開発した開発したフィンガープリントを併用して特徴量を増加させ、更に複数の機械学習モデルにおける予測結果を特徴量として使用する方法(スタッキング法、アンサンブル学習の一種)を用いた方法を開発した。この開発手法は慢性生態毒性に関して、重複回答や金属化合物の予測が困難といった従来法の課題を克服し、更に従来法より高い予測精度を得た。

今回の研究により、線形QSAR以外の慢性生態毒性予測手法を開発すると共に、生態毒性値の予測に有用な新規フィンガープリントを開発する事ができた。

2. 手法

2.1 データセット

AIST-MeRAM データベースから37,599の慢性毒性データ(LC20/EC20に該当)を抽出した。その際の化学物質数は2,177であった。抽出した化学物質に対して、MACCS Keys[6]、Extended-Connectivity Fingerprint (ECFP) [7]、Klekota-Roth Fingerprint[8]、及び後述する毒性予測用の開発したフィンガープリントを作成した。

また、Mol Log P (Log Kow)やLog Sなどを含む14種の記述子を計算し(表1)、前述のフィンガープリント(以後FPと省略)と合わせて特徴量とした。

表1 今回使用した記述子

No.	Descriptors
1	Mol Log P [9]
2	Calculated Log S [10]
3	Molecular Weight
4	The Number of Rotatable Bonds
5	Heavy Atom Count
6	The Number of Aromatic Heavy Atoms
7	The Number of Atoms
8	The Number of Aromatic Heavy Atoms Divided the Number of Atoms
9	Topological Molecular Polar Surface Area [11]
10	The Number of Proton Donors
11	The Number of Proton Acceptors
12	The Number of SP ³ Hybridized Carbons Divided Total Carbon Count [12]
13	The Number of Aromatic Carbocycles
14	The Number of Aromatic Heterocycles.

2.2 毒性予測用フィンガープリントの開発

今回開発した毒性予測用FP (以後、開発したFP)における各ビットの内訳を表2に示す。開発したFPは62のビットで構成され、金属元素(52個)、6価クロム、電解質、有機ハロゲン、有機金属(4種、56~59)、農薬(3種、60~62)など、生態毒性に寄与する構造の有無を判定する。

一般に有機ヒ素化合物は無機ヒ素化合物より毒性が低い傾向があるため、無機ヒ素と有機ヒ素の判定ビットを加えた。一方、この開発したFPだけでは一般的な有機物の毒性構造を識別できないため、MACCS Keys、ECFP、Klekota-Roth Fingerprintなどの有機物に対応した従来のFPと併用を前提に設計した。

表2 新規フィンガープリントのビット構成

Bit Position	Components
1-52	Metal compounds; Li, Be, Na, Mg, Al, K, Ca, Sc, Ti, V, Cr, Mn, Fe, Co, Ni, Cu, Zn, Ga, Ge, As, Se, Rb, Sr, Y, Zr, Nb, Mo, Tc, Ru, Rh, Pd, Ag, Cd, In, Sn, Sb, Te, Cs, Ba, Hf, Ta, W, Re, Os, Ir, Pt, Au, Hg, Tl, Pb, Bi, Po
53	Hexavalent Chromium Compound
54	Ionization Compound
55	Organohalogen Compound
56	Organoarsenic Compound
57	Organolead Compound
58	Organotins Compound
59	Organic Mercury Compound
60	Carbamate Insecticide
61	Organophosphorus Compound
62	Diphenyl Ether Derivative

2.3 開発したFPの性能評価

開発したFPの性能評価を目的に、開発したFPを他のFPや記述子と組み合わせて使用した際の、慢性生態毒性値に対する予測性能を比較した。評価用の機械学習モデルとしてRandom Forestを採用した。Random Forestは特徴量を用いて決定木を複数作成し、複数の決定木が出力した値の平均値を用いて予測する手法である。学習の際に特徴量をランダムに選択する特徴があるため、仮にFPの併用等で重複した特徴量があっても、高い予測精度を維持できると推定される、代表的な機械学習モデルである。

また、各種FPや記述子を組み合わせて評価することで、慢性生態毒性の予測に適した組み合わせに関しても検討を実施した。

2.4 機械学習モデルの予測性能比較

表3に示す各種機械学習モデルの比較を行った。各機械学習モデルにおけるハイパーパラメータの調整にはBayesian optimization[13]を用いた。深層学習のグループ(4種)は4層で構成される多層パーセプトロンを共通の出力層とした。この層に関しては、ノード数、ドロップアウト層、 L^1 あるいは L^2 正則化の導入によりグリッドサーチにて最適化を行った。深層学習における特徴量抽出層に関しては、ハイパーパラメータ調整、層の数及び組み合わせに関してグリッドサーチで最適化を行った。オプティマイザ

ーにはAdam[14]、活性化関数にはRectified Linear Unit (ReLU) [15]を用いた。

深層学習モデルの精度改善を目的に、溶解度予測で二系統入力を行った事例を参考に[16]、特徴量を二系統入力するモデルに関しても検討を行った。

更に、これらの一般的な機械学習モデルに併せて、図1に示すスタッキング法を組み込んだ慢性生態毒性予測手法(以後、スタッキング法を用いた開発モデルと省略)により、予測精度の比較評価を行った。一般的なスタッキング法は各種機械学習モデルの出力結果を特徴量として用いる方法だが、この方法では予測精度が元の特徴量と機械学習モデルの組み合わせに及ばなかった。スタッキング法を用いた開発モデルは、一般的なスタッキング法に、次元削減で作成した特徴量と元の特徴量を追加することで精度を高めた方法である。各種機械学習モデルの組み合わせはグリッドサーチで最適化を行った。

表3に示す機械学習モデルに対する評価方法としては、2.3で最も高精度だった全特徴量の組み合わせ(開発したFP、MACCS Keys、ECFP、Klekota-Roth Fingerprint及び14種の記述子を併用した特徴量)を説明変数、慢性生態毒性値のデータを目的変数として、10-fold交差検証により予測精度の比較を行った。評価指標には相関係数及びRoot Mean Square Error (RMSE)を使用した。

表3 今回検討した機械学習モデル

Type of Algorithm	The Name of the Model
Decision Tree	Random Forest
Bootstrap Aggregating	Extremely Randomized Trees
	Regularized Greedy Forest
Decision Tree	XGBoost
Gradient Boosting	LightGBM
	Multiple Linear Regression
Linear Regressions	Stochastic Gradient Descent
	Ridge Regression
	Convolutional Neural Network
Deep Neural Network	Dilated Convolution
	LSTM
	Dense Neural Network
Matrix Factorization	Factorization machine
	K-nearest Neighbor
Others	Partial Least Squares
	Support Vector Regression
Ensemble Learning	Stacking Method

図1 スタッキング法を組み込んだ慢性生態毒性値予測モデル

2.5 ベンチマークテスト

前項で実施した機械学習モデルの比較では、スタッキング法を用いた開発モデルが慢性生態毒性値の予測精度が最も高い機械学習モデルであった。AIST-MeRAMデータベースから、信頼性の高いデータのうち、25%に相当する42個の慢性毒性データをランダムに選択し、ホールダアウト法にて、スタッキング法を用いた開発モデルのベンチマークテストを実施した。比較対象には、本分野で代表的なアプリケーションであるECOSARを使用した。

3. 結果と考察

3.1 開発したFPの性能評価及び最適化

表4にFP及び記述子を組み合わせで作成した特徴量による慢性生態毒性予測の精度をまとめた。精度の指標に用いたRMSEは値が低いほど、相関係数は値が高いほど生態毒性予測の精度が良好であることを示す。

開発したFPを除いた一般的な特徴量を比較すると、予測精度はMACCS Keys、記述子、Klekota-Roth、ECFP4の順となった。これら一般的な特徴量と開発したFPを組み合わせると、ECFP、Klekota-Roth Fingerprint、並びに記述子に対して開発したFPは慢性生態毒性値の予測精度を向上させる効果が確認された。特に記述子と相性が良く、MACCS Keysに近い水準まで予測精度が向上した。一方で、MACCS

Keysと開発したFPの組み合わせでは精度改善効果はなかった。これは、MACCS Keysは金属元素を周期表のグループとして区別が可能であり[6]、この慢性生態毒性データセット用いた予測では、開発したFPを組み合わせても、差別化が困難であったと推定された。

Bayesian optimizationによるハイパーパラメータの最適化前後の予測精度に関し、表4に示す。ハイパーパラメータの最適化により、RMSEで2程度、相関係数で0.03程度の精度向上が確認された。一方で、最適化を行っても、各種FPや記述子など予測精度の順位に変化はなかった。

今回用いた全ての特徴量(記述子、ECFP4、Klekota-Roth Fingerprint、MACCS Keys、開発したFP)を組み合わせた方法が、慢性生態毒性値の予測精度が最も良好であったことから(表4 All Feature Setの下段)、この組み合わせを以後の検討における特徴量として採用した。

表4 新規フィンガープリントによる慢性生態毒性値予測精度の改善効果

Evaluation method	Hyperparameters	RMSE [mg/L]		Correlation Coefficient	
		Default	Optimized	Default	Optimized
Descriptors	Only Feature	20.2	17.8	0.611	0.644
	with New FP	19.8	16.9	0.619	0.660
ECFP4	Only Feature	22.0	20.0	0.574	0.606
	with New FP	20.9	18.7	0.600	0.631
Klekota-Roth Fingerprint	Only Feature	22.2	19.7	0.578	0.610
	with New FP	20.4	18.2	0.611	0.638
MACCS Keys	Only Feature	19.7	16.8	0.617	0.661
	with New FP	19.5	17.0	0.621	0.657
All Feature Set	Only Feature	16.7	14.4	0.665	0.707
	with New FP	16.3	14.3	0.672	0.709

3.2 機械学習モデルの比較

全ての特微量(記述子、ECFP4、Klekota-Roth、MACCS Keys、および開発したFP)を使用した際の、慢性生態毒性値に対する予測精度の比較結果を表5に示す。この中で、スタッキング法を用いた開発モデルの予測精度が最も高かった(表5 No. 19)。

スタッキング法を用いた開発モデルを除いた、16種の一般的なモデルの中で最も高精度なのはXGboost[17]で、次点がRandom Forestであった(表5 No. 4、1)。スタッキング法を用いた開発モデルは、最終判定でRandom Forestを採用している(図1)。これは、Random Forestが今回比較した結果の中で2番目に高精度であることに加え、2.3で説明したように、特微量をランダムに選択する手法であることから、スタッキング法を用いた開発モデルで懸念される、重複した特微量による過学習を抑制可能と考えて採用した。

各種アルゴリズムのグループで比較すると、決定木のグループは他のグループよりも予測精度が高かった(表5 No. 1~5)。決定木のグループにおけるブートストラップ(表5 No. 1、2)と勾配ブースティング(表5 No. 3-5)を比べると、ほぼ同等の性能だった。

線形回帰のグループは、今回の研究で最も精度が低く、重回帰は特に顕著であった(表5 No. 6-8)。重回帰を使用した教師データと検証データの予測精度差が大きく、過学習が強く発生している事が示唆された。一方、リッジ回帰は、線形回帰の中で最も高精度であった。リッジ回帰の持つ罰則項により、過学習が抑制された事が主因と推定される。

深層学習のグループは、共通の層として構築した多層パーセプトロン単独に比べ、Long Short-Term Memory (LSTM) [18]で特微量を抽出しても予測精度の改善効果はなかった(表5 No. 11-12)。一方で、Dilated Convolution[19]とConvolutional Neural Network (CNN) [20]は予測精度が改善されたが、決定木系のグループには予測精度が及ばなかった(表5 No.9、10、12)。LSTMで予測精度は改善せず、CNNで予測精度が改善していることから、配列の並び順は優位性がないが、毒性に寄与する特微量のパターンが存在する事を示唆していると考えている。CNNの精度改善を目的に、特微量抽出層の増加、及びCNNで利用頻度の高いBatch Normalization[21]の導入を試みたが、予測精度は逆に低下した。特にCNNにおける畳み込み層、あるいはプーリング層の後にBatch Normalizationを入れると、化学物質の毒性値を予測

表5 各種機械学習モデルにおける慢性生態毒性予測精度の比較結果(10-fold 交差検証)

Type of Algorithm	The Name of the Model	Correlation Coefficient		RMSE [mg/L]	
		Training	Test	Training	Test
Decision Tree Bootstrap Aggregating	Random Forest	0.966	0.709	0.00	0.00
	Extremely Randomized Trees	0.977	0.707	0.00	0.00
	Regularized Greedy Forest	0.896	0.699	0.00	0.00
Decision Tree Gradient Boosting	XGBoost	0.953	0.709	0.00	0.00
	LightGBM	0.970	0.703	0.00	0.00
Linear Regressions	Multiple Linear Regression ^a	0.915	0.195	0.00	- ^b
	Stochastic Gradient Descent	0.953	0.567	0.00	0.00
	Ridge Regression	0.835	0.669	0.00	0.00
Deep Neural Network	Convolutional Neural Network	0.990	0.658	0.00	0.00
	Dilated Convolution	0.994	0.645	0.00	0.00
	LSTM	0.976	0.592	0.00	0.00
	Dense Neural Network	0.997	0.663	0.00	0.00
Matrix Factorization	Factorization machine	0.859	0.670	0.00	0.00
	K-nearest Neighbor	0.803	0.631	0.00	0.00
Others	Partial Least Squares	0.832	0.641	- ^b	- ^b
	Support Vector Regression	0.708	0.638	0.00	0.00
Ensemble Learning	Stacking Method	0.994	0.716	0.00	0.00

^a このモデルは目的変数が特微量よりも多い必要があるため、ビット数の多い Klekota-Roth Fingerprint を除外した。

^b 1,000,000,000 以上を示した。

できないケースが頻発した。また、深層学習における特徴量の入力部分を2つにすることで、CNNと元の特徴量を併用、あるいはCNNとLSTMを併用する手法を試みたが、多層パーセプトロン単独よりも精度が低下した(表5 No. 13-14)。

3.3 ベンチマークテスト結果

スタッキング法を用いた開発モデルと従来法の代表的なアプリケーションであるECOSARによる生態毒性値の予測に対するベンチマークテスト結果を表6にまとめた。また、表7に慢性生態毒性値予測結果、図2に予測対象の化学物質における構造式の代表例を示す。スタッキング法を用いた開発モデルはECOSARと比べて慢性生態毒性値の予測精度が良好である事に加え、予測可能な化学物質の数がECOSARの36に対して42と、金属化合物を含む全ての化学物質を予想可能であった(表7 No. 9-10)。更に、ECOSARが20個の重複回答を行っているのに対し、スタッキング法を用いた開発モデルは全て単一回答であり、重複回答はなかった。

表6 ベンチマークテストの集計結果

Validation Metrics	ECOSAR	Stacking Method
Number of Predicted Chemical Substances	36/42	42/42
Number of Duplicated Results	20	0
Number of All Correct Answers/ Total Answers	34/62	25/42
Accuracy Ratio	55%	60%

図2 ベンチマークテスト対象の構造式抜粋(10/42)

表7 ベンチマークテスト結果の抜粋(10/42).

No.	CAS No.	Chemical substance	ECOSAR Class	Type	Tox value [mg/L]	Result ^a
1	4170-30-3	2-Butenal	Vinyl/Allyl/Propargyl Aldehydes	ECOSAR	0.01	×
			-	Stacking	0.44	✓
			-	DB value ^b	0.49	-
2	2104-64-5	P-Phenylphosphonic acid O-ethyl O-(4-nitrophenyl) ester	Esters, Monothiophosphate	ECOSAR	0.015	✓
			-	Stacking	0.28	×
			-	DB value	0.0041	-
3	21725-46-2	2-[[4-Chloro-6-(ethylamino)-1,3,5-triazin-2-yl]amino]-2-methylpropanenitrile	Triazines, Aromatic	ECOSAR	3.0600	✓
			-	Stacking	2.36	✓
			-	DB value	0.90	-
4	13171-21-6	Dimethyl-(diethylamino-1-chlorocrotonyl)-phosphate	Esters	ECOSAR	7.92	✓
			Esters, Phosphates	ECOSAR	0.02	×
			Vinyl/Allyl/Propargyl Halides	ECOSAR	140.00	✓
			Acrylamides	ECOSAR	0.87	×
			-	Stacking	1.43	×
5	2303-17-5	Triallate	-	DB value	20.00	-
			Vinyl/Allyl/Propargyl Halides	ECOSAR	0.02	×
			Thiocarbamates, Mono	ECOSAR	0.057	✓
6	25311-71-1	Isufenphos metabolite	-	Stacking	0.38	✓
			-	DB value	0.560	-
			Esters	ECOSAR	0.083	✓
			Esters, Monothiophosphate	ECOSAR	0.018	×
7	314-40-9	5-Bromo-6-methyl-3-(1-methylpropyl)-2,4(1H,3H)-pyrimidinedione	-	Stacking	0.26	✓
			-	DB value	0.65	-
			Vinyl/Allyl/Propargyl Halides	ECOSAR	52.12	✓
8	52645-53-1	3-(2,2-Dichloroethenyl)-2,2-dimethylcyclopropanecarboxylic acid, (3-Phenoxyphenyl)methyl ester	Carbonyl Ureas	ECOSAR	0.34	×
			-	Stacking	16.65	✓
			-	DB value	29.00	-
			Esters	ECOSAR	0.000853	✓
			Vinyl/Allyl/Propargyl Halides	ECOSAR	0.000010	×
9	1333-82-0	Chromium oxide (CrO ₃)	Pyrethroids	ECOSAR	0.000017	×
			-	Stacking	0.00	×
			-	DB value	0.0075	-
10	4717-38-8	19-Norpregna-1,3,5(10)-trien-20-yne-3,17-diol	-	ECOSAR	-	-
			-	Stacking	9.20	×
			-	DB value	0.03	-

^a 正答値に対して0.1から10 倍以内の範囲を ✓、それ以外を ×、-はデータなしを示す。

^b DB valueはデータベースの値(正答値)を示す。

4. 結論

一般的なフィンガープリントや記述子により抽出した特徴量に加え、開発した生態毒性用のフィンガープリントを併用して特徴量として用い、更にスタッキング法を組み込んだ機械学習モデルを用い

る事で、重複回答や金属化合物が予測できないといった、従来法の課題を克服しながら、慢性生態毒性値を従来法より良好な精度で予測することが可能となった。

線形QSAR以外の慢性生態毒性予測手法に加え、生態毒性値の予測に有用な新規フィンガープリントを開発する事ができた。

5. 付録

各種特徴量の抽出にはRDkit[22] 及びChemistry Development Kit[23] を用いた。XGboost[17]、LightGBM[24]、Factorization Machines[25]、Regularized Greedily Forest[26]はそれぞれ作者のライブラリを用いた。多層パーセプトロンや深層学習モデルの構築にはKeras[27]を使用し、その他の機械学習モデルはScikit-learn[28]を用いた。スタッキング法を用いた開発モデルの構築にはMlxtend[29]及びScikit-learn[28]を用いた。

文献および注

- [1] ECOSAR Predictive Model (URL=<https://www.epa.gov/tsca-screening-tools/ecological-structure-activity-relationships-ecosar-predictive-model>).
- [2] OECD QSAR Toolbox (URL=<http://www.oecd.org/env/ehs/risk-assessment/theoecdqsartoolbox.htm>).
- [3] 古濱 彩子, 白石 寛明, 日本化学会情報化学部会誌, **30**, 42–45, (2012).
- [4] AIST-MeRAM (URL=<https://en-meram.aist-riss.jp/>).
- [5] M. Takata, B.-L. LIN, M. Xue, A. Terada, and M. Hosomi, *Chemosphere*, **238**, 124604, (2019).
- [6] J. L. Durant, B. A. Leland, D. R. Henry, and J. G. Nourse, *J. Chem. Inf. Comput. Sci.*, **42**, 1273–1280, (2002).
- [7] D. Rogers and M. Hahn, *J. Chem. Inf. Model.*, **50**, 742–754, (2010).
- [8] J. Klekota and F. P. Roth, *Bioinformatics*, **24**, 2518–2525, (2008).
- [9] S. A. Wildman and G. M. Crippen, *J. Chem. Inf. Comput. Sci.*, **39**, 868–873, (1999).
- [10] J. S. Delaney, *J. Chem. Inf. Comput. Sci.*, **44**, 1000–1005, (2004).
- [11] P. Ertl, B. Rohde, and P. Selzer, *J. Med. Chem.*, **43**, 3714–3717, (2000).
- [12] F. Lovering, J. Bikker, and C. Humblet, *J. Med. Chem.*, **52**, 6752–6756, (2009).
- [13] P. I. Frazier, *arXiv preprint*, **2018**, 1–22, arXiv:1807.02811.
- [14] D. P. Kingma and J. L. Ba, *3rd Int. Conf. Learn. Represent. ICLR 2015 - Conf. Track Proc.*, **2015**, 1–15.
- [15] V. Nair and G. Hinton, *ICML 2010 - Proceedings, 27th Int. Conf. Mach. Learn.*, **2010**, 807–814.
- [16] 鈴木天音, 木倉悠一郎, 田中健一, *J. Comput. Aided Chem.*, **19**, 1–6, (2018).
- [17] T. Chen and C. Guestrin, in *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, **13-17-August-2016**, 785–794, (2016).
- [18] S. Hochreiter and J. Schmidhuber, *Neural Comput.*, **9**, 1735–1780, (1997).
- [19] F. Yu and V. Koltun, *arXiv preprint*, **2015**, ArXiv: 1511.07122.
- [20] Y. Lecun, L. Bottou, Y. Bengio, and P. Ha, *Proc. IEEE*, **November**, 1–46, (1998).
- [21] S. Ioffe and C. Szegedy, *32nd Int. Conf. Mach. Learn. ICML 2015*, **1**, 448–456, (2015).
- [22] RDkit (<https://www.rdkit.org/>)
- [23] E. L. Willighagen et al., *J. Cheminform.*, **9**, 1, 1–19, (2017).
- [24] G. Ke et al., *Nips '17*, **2017**, 9.
- [25] S. Rendle, *Proc. - IEEE Int. Conf. Data Mining, ICDM*, **2010**, 995–1000.
- [26] R. Johnson and T. Zhang, *IEEE Trans. Pattern Anal. Mach. Intell.*, **36**, 942–954, (2014).
- [27] François C. et al., *GitHub*, **2015**.
- [28] V. Pedregosa, F. and Varoquaux, G. et al., *J. Mach. Learn. Res.*, **12**, 2825–2830, (2011).
- [29] S. Raschka, *J. Open Source Softw.*, **3**, 638, (2018).

Predicting the Fish Chronic Ecotoxicity of Chemical Substance with New Ecotoxicity Fingerprint and Stacked Ensemble Method on Machine Learning

Michiyoshi Takata^{a*}, Bin-Le Lin^b, Akihiko Terada^{a**}, Masaaki Hosomi^a

^aDepartment of Chemical Engineering, Tokyo University of Agriculture and Technology, 2-24-16 Naka-cho, Koganei, Tokyo 184-8588, Japan

^bResearch Institute of Science for Safety and Sustainability, National Institute of Advanced Industrial Science and Technology (AIST), 16-1 Onogawa, Tsukuba City, Ibaragi, 305-8569 Japan.

The *in silico* method to predict the ecotoxicity of chemical substances for reducing animal testing has become attracted attention. A most common model for predicting ecotoxicity classify chemical substances empirically based on functional groups and then predict ecotoxicity with a linear regression by using a descriptor of a chemical substance such as Log Kow. But the conventional method outputs duplicate result for one chemical substance when it has multiple functional groups. Moreover, this method is not appropriate for predicting the ecotoxicity of metal compounds. To overcome these challenges, this study developed a new fingerprint as a feature set for machine learning, and a new prediction model with supervised machine learning for chronic ecotoxicity on fish. The new fingerprint extracts feature of a chemical substance by judging the existence of the structure contributing to ecotoxicities such as carbamate insecticide, organophosphorus pesticides, organic halogen, various metal elements, and hexavalent chromium. Moreover, we compared the accuracy for predicting chronic ecotoxicity on fish with various machine learning models by 10-fold cross-validation using this new fingerprint, general fingerprints, and descriptors together as a feature set. As a result, our developed method with the stacking ensemble was the most accurate in this study. This method improved accuracy by using the result of multiple machine learning algorithms as a part of a feature set. The result of the benchmark test show that the prediction accuracy of this method was better than conventional methods.

keywords: Ecotoxicity prediction; Machine learning; Ensemble learning; AIST-MeRAM; ECOSAR

* s183801z@st.go.tuat.ac.jp

** akte@st.go.tuat.ac.jp