化学反応に最適な酵素の EC 番号予測における精度向上のための特徴量エンジニアリング

Feature engineering to improve accuracy in predicting the EC number of the best enzyme for a chemical reaction.

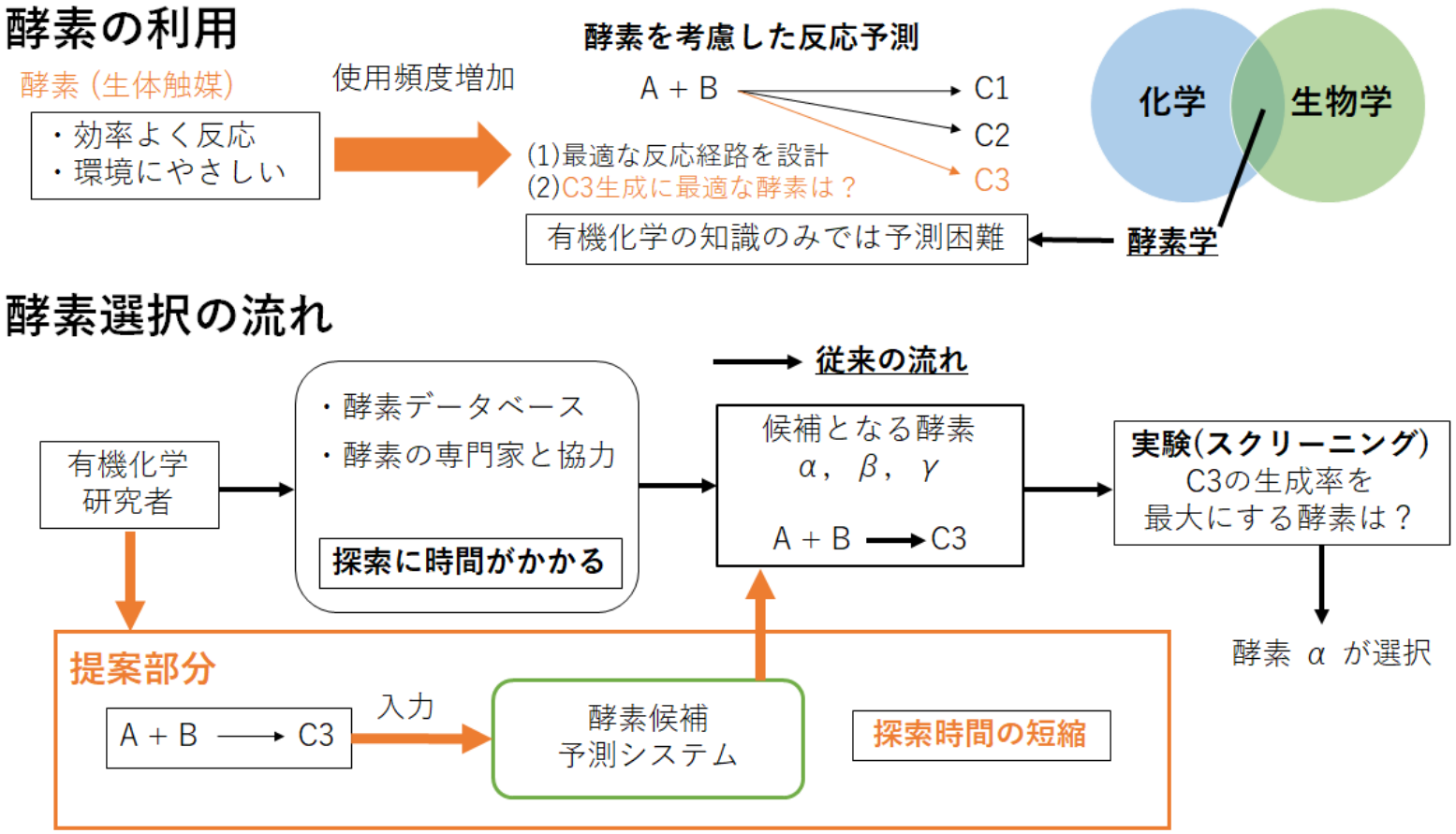
武藤 克弥 (Katsuya Muto)
u255018@st.pu-toyama.ac.jp

富山県立大学大学院 電子・情報工学専攻 情報基盤工学部門

N211, 17:40-18:00, Tuesday, December 6, 2022.

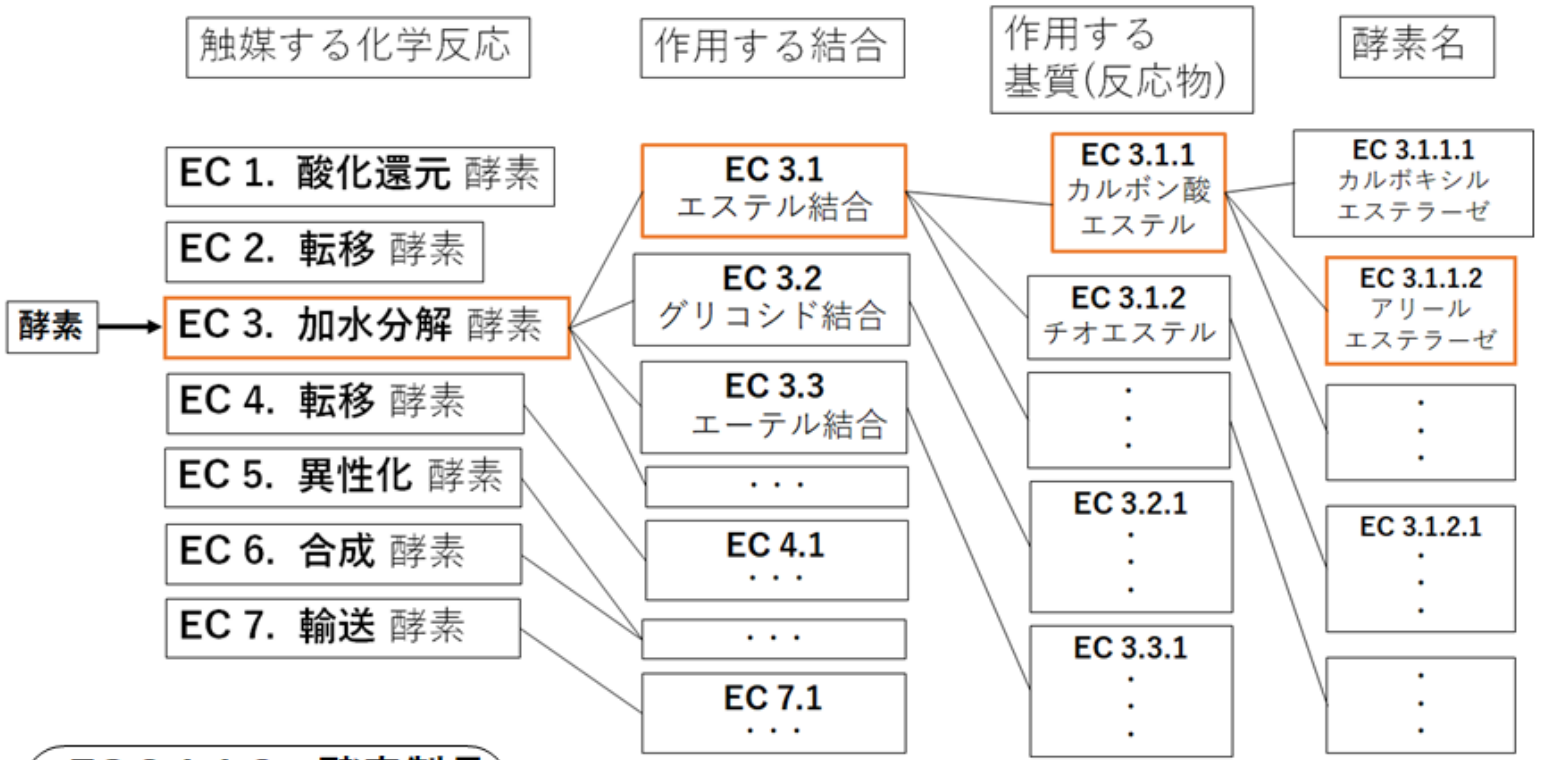
1. はじめに

近年，新型コロナウイルスなどの影響で新薬開発の需要が高まっており，効率的に目的物を生成するため，酵素を用いることが重要視されるようになった．



2. 酵素番号 (EC 番号)

酵素を 4 組の数字 (EC ○. ○. ○. ○) の組み合わせで分類したもの.



EC 3.1.1.2の酵素製品

- 酵素 α
- 酵素 β
- 酵素 γ

EC番号を予測することで → **スクリーニング**

最適な酵素候補を絞り込める

1 つの酵素を選択

3. EC 番号予測の先行研究

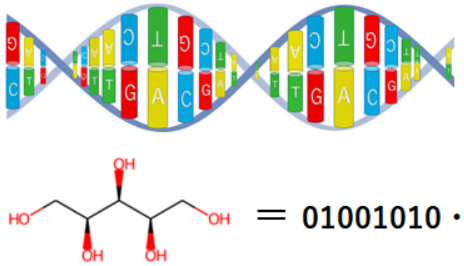
予測で用いる特徴量としてタンパク質配列 [1] や構造的性質 [2,3] が用いられている

(1)タンパク質配列 = 配列の類似性で分類

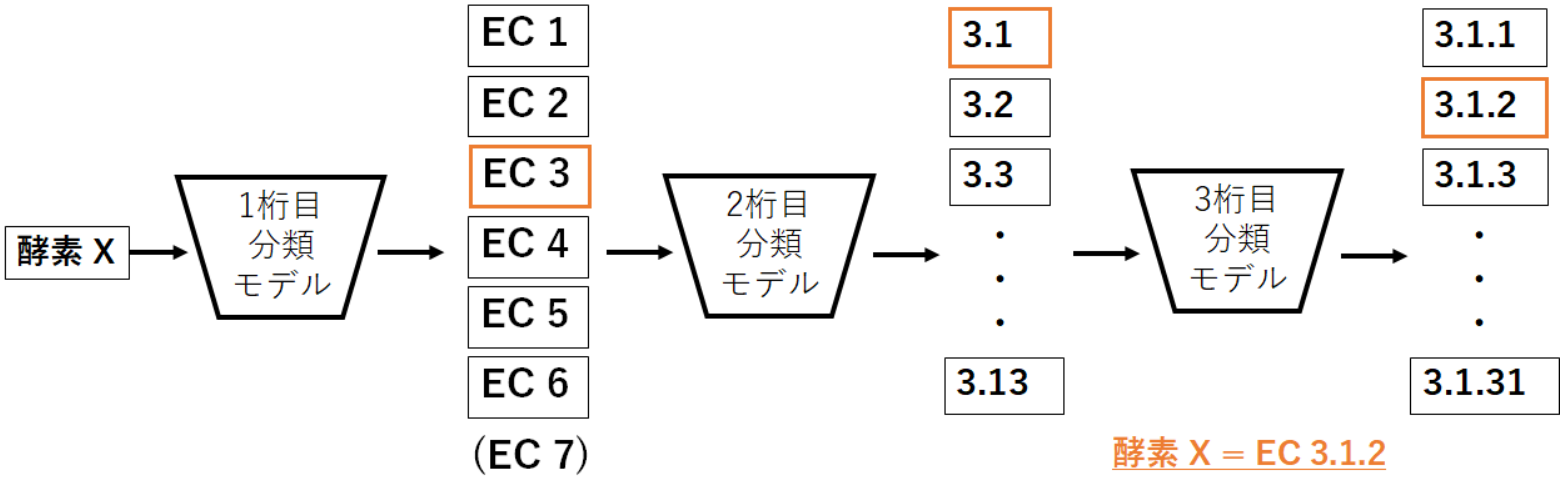
(2)構造的特徴
＝化合物の構造情報を用いる

化学反応を
取り扱いやすい

(3)提案手法
=化合物の物理・化学的な性質を用いる



＝ 01001010・・・



[1] Jae Yong Ryu et al., 2019.
[2] Diogo A. R. S. Latino et al., 2009.
[3] Yoshihiko Matsuta et al., 2013.

4.1 提案手法の概要

(1) ターゲット反応式と EC 番号反応式 [4] における、反応物から生成物に変化する際の物理・化学的特性値の変化を比較

(2) 最も類似する EC 番号反応式の EC 番号を最適な酵素候補として予測

【類似性の概念】
化学反応の性質が類似＝同様の反応が起こる
→同じ酵素を使用しても同様の効果が得られる可能性が高い

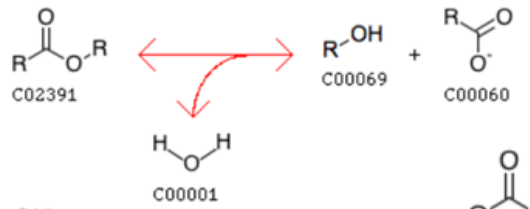
ターゲット（実験的反応）

反応物A ＋ 反応物B ⟷ 生成物 ＋ 副生成物

特性値変化 T【（生成物+副生成物）-（反応物A+反応物B）】

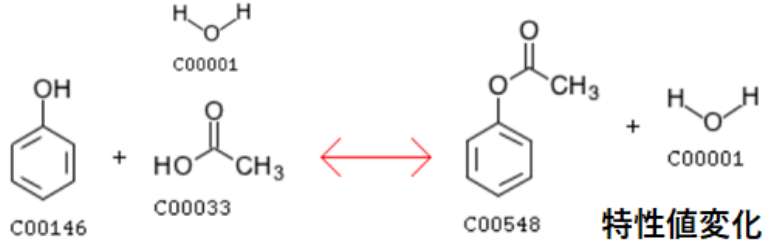
EC番号反応式（自然界に存在する反応）→KEGG[4] などに記載

EC 3.1.1.1 (x 5種)



特性値変化 A

EC 3.1.1.2 (x 3種)



特性値変化 B

T ≒ A

ターゲットに EC 3.1.1.1の酵素

[4] KEGG: Kyoto Encyclopedia of Genes and Genomes, https://www.genome.jp/kegg/kegg_ja.html

4.2 先行研究との比較

(1) 先行研究： 構造情報 (フィンガープリント (FP)) の変化 [5]

(2) 提案手法： 化合物の物理・化学的性質の変化量 [6]

FPの例：化合物がどの部分構造(166種)を持っているかを判定したバイナリベクトル

EC 番号反応式： 反応物 1 + 反応物 2 ⟷ 生成物 1 + 生成物 2

$$\rightarrow RFP = FP_{\text{生成物 1+生成物 2}} - FP_{\text{反応物 1+反応物 2}}$$

(1)RFP：反応差分フィンガープリント
→FPの変化 = 化合物の構造変化

→KEGGの(自然由来の)反応に対してクラス分類
改善点：化学的な特徴表現力の拡張

(2)提案手法：物理・化学的性質の変化 (特性値変化量)
→208種の物理・化学記述子 (RDKit[6] ライブラリ)
→構築の容易さ
→KEGGだけでなく、実験的な反応に対しても予測

V_j: 記述子 j の特性値 $(j = 1, 2, ..., n)$

CV_{ij}: 反応式 i の特性値変化量 $(i = 1, 2, ..., m)$

$$CV_j = V_j(\text{生成物 1} + \text{生成物 2}) - V_j(\text{反応物 1} + \text{反応物 2})$$

$\boldsymbol{DF}_i$: EC 番号反応式 i の特徴ベクトル

$$\boldsymbol{DF}_i = (CV_{i1}, CV_{i2}, \cdots, CV_{ij}, \cdots, CV_{in})$$

($\boldsymbol{DF}_T$: ターゲット反応式)

特性値変化量

表1. EC番号(ターゲット)反応式の特性値変化量

	記述子 1 (分子量変化)	記述子 2 (疎水性変化)	⋯	記述子 n (電荷量変化)
$\boldsymbol{DF}_T$	CV_{T1}	CV_{T2}	⋯	CV_{Tn}
$\boldsymbol{DF}_1$	CV_{11}	CV_{12}	⋯	CV_{1n}
$\boldsymbol{DF}_2$	CV_{21}	CV_{22}	⋯	CV_{2n}
⋮	⋮	⋮	⋱	⋮
$\boldsymbol{DF}_m$	CV_{m1}	CV_{m2}	⋯	CV_{mn}

[5] Qian-Nan Hu et al., 2012.
[6] The RDKit Documentation, https://www.rdkit.org/docs/GettingStartedInPython.html

4.3 特徴選択による記述子削減

ラッパー法による記述子選択 (SequentialFeatureSelector(SFS))[7]

分類モデルの予測精度を評価し，最高評価となる組み合わせの記述子を選択

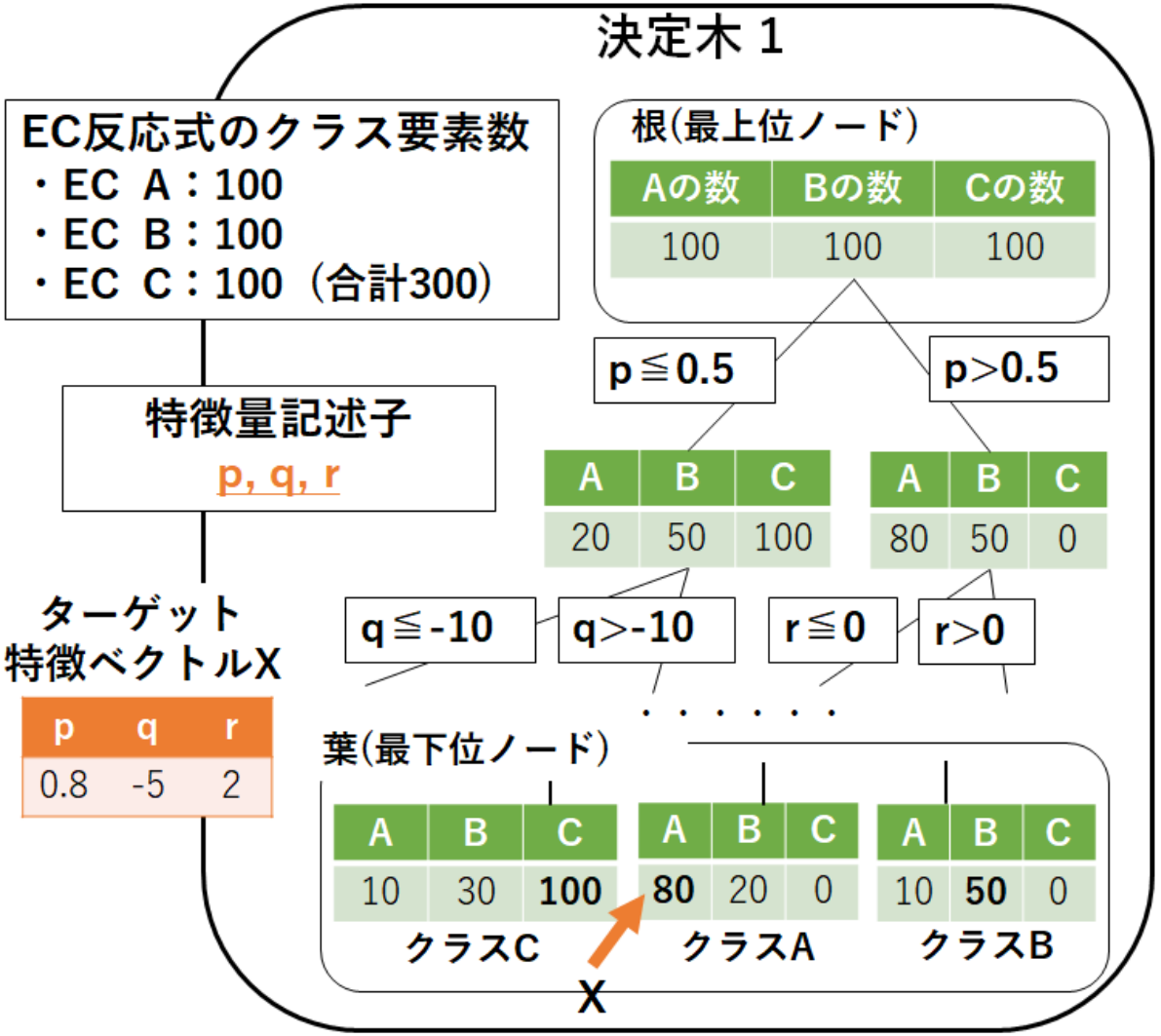
用いる手法: Step Forward 法

1. n 個の記述子から 1 つ選択し，n 種類の分類モデルを作成
2. 最も評価の高いモデルに用いられている，記述子を選択
3. $n-1$ 個の記述子から 1 つ選択し，先ほど選択されたモデルに追加することで，新たな分類モデルを作成
4. $n-1$ 個のモデルで最も評価の高いものに用いられている，記述子の組み合わせを選択
5. 指定した特徴数になるまで 3 と 4 を繰り返す.

モデルの評価基準： F1 スコア平均 (層化 5 分割交差検証)

[7] Mlxtend.feature selection, http://rasbt.github.io/mlxtend/api_subpackages/mlxtend.feature_selection/

4.3 ランダムフォレストによるEC番号分類(1)



4.3 ランダムフォレストによるEC番号分類(2)

各ノードで IG が最大となるように，特徴量 f とデータを分割する閾値を決定

$$IG(D_P, f) = I_{imp}(D_p) - \frac{N_{left}}{N_p} I_{imp}(D_{left}) - \frac{N_{right}}{N_p} I_{imp}(D_{right})$$

D_p：上位ノード内のデータ (特徴ベクトル)

f：分割時に用いられる特徴量

$I_{imp}(t)$：ジニ不純度

D_{left}，D_{right}：分割先の下位ノード内のデータ

N_p，N_{left}，N_{right}：上位ノード，下位 (左右) ノードのデータ数

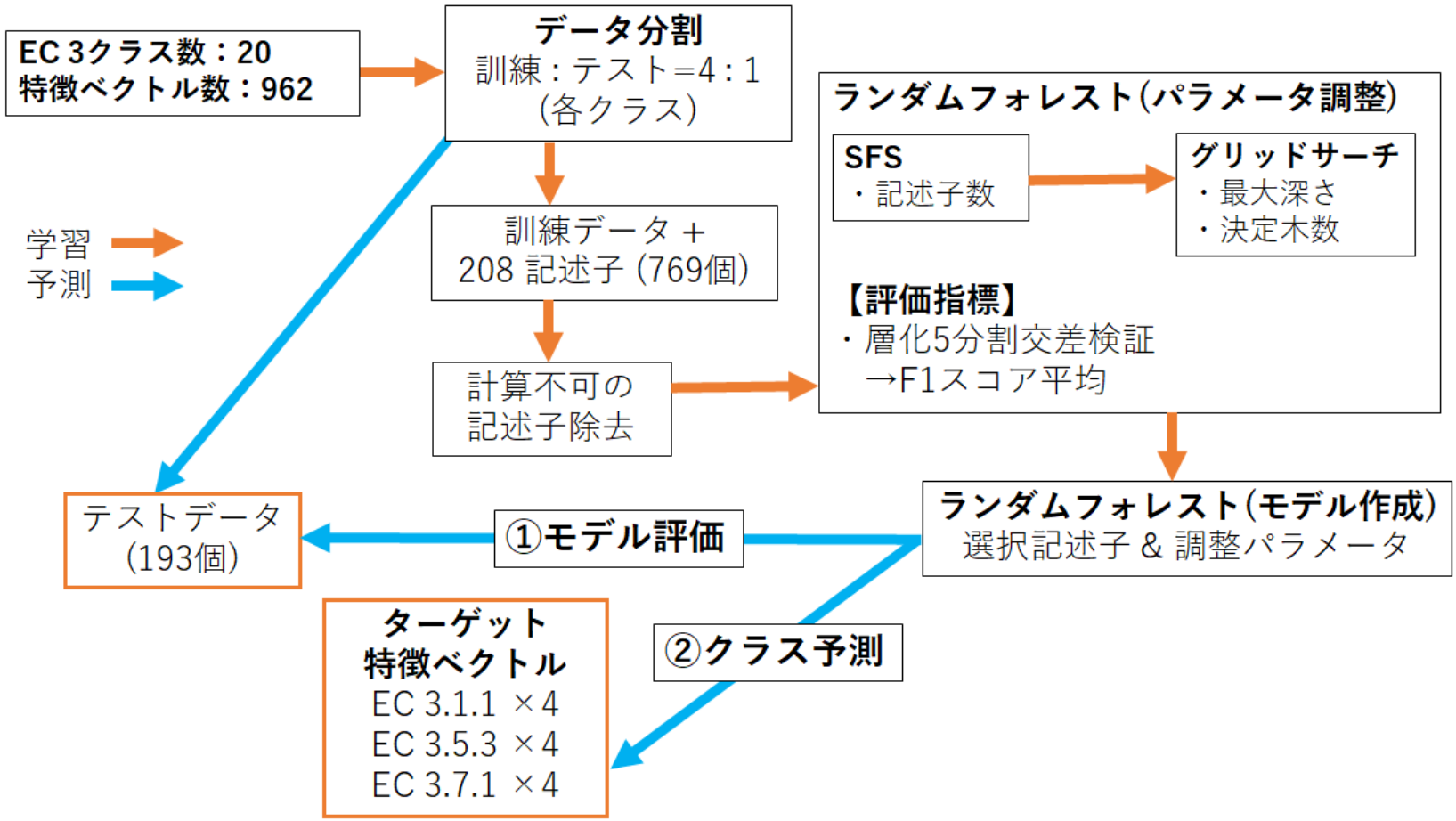
$$I_{imp}(t) = \sum_{i=1}^{c} p(i|t)(1 - p(i|t)) = 1 - \sum_{i=1}^{c} p(i|t)^2$$

$p(i|t)$：ノード t のクラス i のデータ割合，c：ノード t 内のクラス数

パラメータ調整(グリッドサーチ)→各決定木の最大深さ、決定木数

5.1 モデル作成・予測手順

EC3 クラスに限定 & 2・3 桁目までを分類予測



5.1 化学反応データの取得と整形

KEGG より EC3 クラスのデータを自動取得

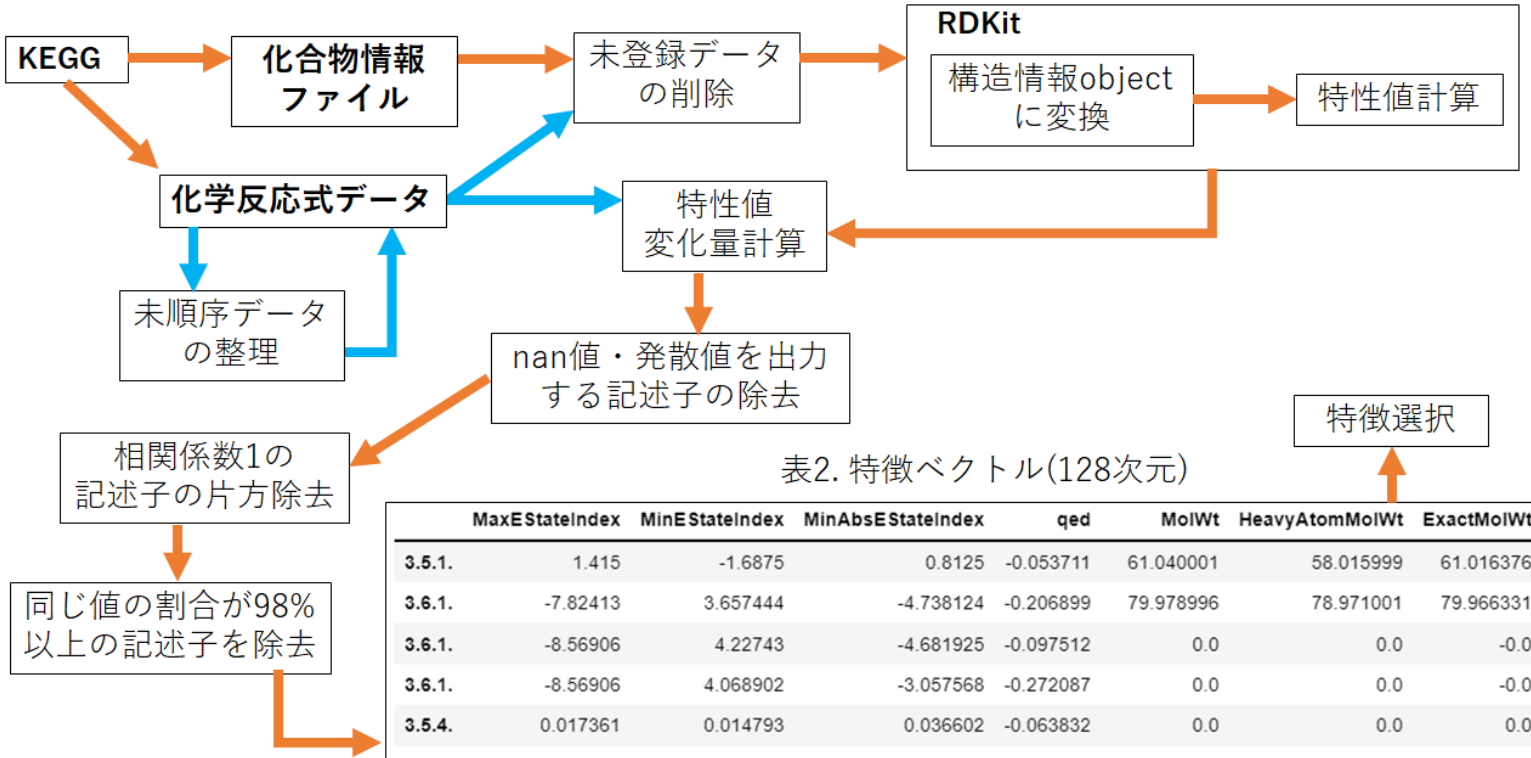


表2. 特徴ベクトル(128次元)

	MaxEStateIndex	MinEStateIndex	MinAbsEStateIndex	qed	MolWt	HeavyAtomMolWt	ExactMolWt	NumValer
3.5.1.	1.415	-1.6875	0.8125	-0.053711	61.040001	58.015999	61.016376	
3.6.1.	-7.82413	3.657444	-4.738124	-0.206899	79.978996	78.971001	79.966331	
3.6.1.	-8.56906	4.22743	-4.681925	-0.097512	0.0	0.0	-0.0	
3.6.1.	-8.56906	4.068902	-3.057568	-0.272087	0.0	0.0	-0.0	
3.5.4.	0.017361	0.014793	0.036602	-0.063832	0.0	0.0	0.0	
...	...	...	...	...	...	...	...	
3.1.1.	-6.828515	-0.668523	-1.021632	-0.087708	0.0	0.0	0.0	
3.2.1.	-8.720595	1.384253	-0.518534	-0.069449	0.0	0.0	-0.0	
3.2.1.	-8.708548	1.381274	-0.495988	-0.069449	0.0	0.0	-0.0	

5.1. 各 EC 3 クラスのデータ数

EC 反応式クラス：20，反応式数が 6 以上のクラスのみ採用

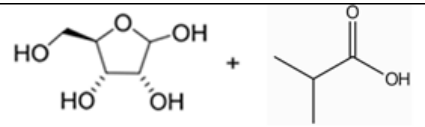
クラス名	反応式数	クラス名	反応式数	クラス名	反応式数
3.1.1	122	3.2.2	24	3.5.5	12
3.1.2	59	3.3.2	6	3.5.99	11
3.1.3	152	3.4.13	6	3.6.1	94
3.1.4	29	3.4.19	7	3.7.1	35
3.1.6	14	3.5.1.	155	3.8.1	16
3.1.7	8	3.5.3	25	3.13.1	9
3.2.1	131	3.5.4	47	合計	962

5.1 ターゲット反応式 (実験的反応) データの取得

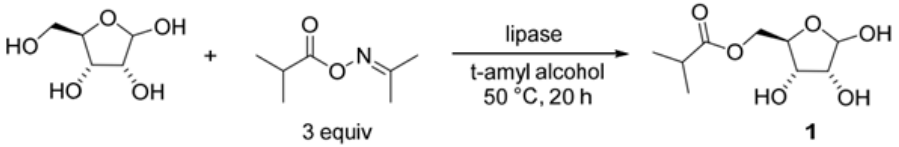
・KEGG に登録されていない実験的な反応に対し、モデルの予測精度を評価
・EC 3.1.1, EC 3.7.1, EC 3.5.3 からそれぞれ 4 つの反応がターゲット
→合成実験の文献 [8] や BRENDA[9] から取得

EC 3.1.1

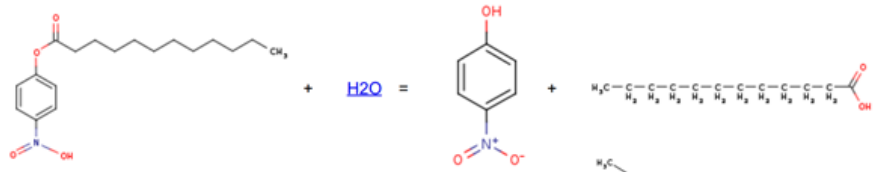
Target 1



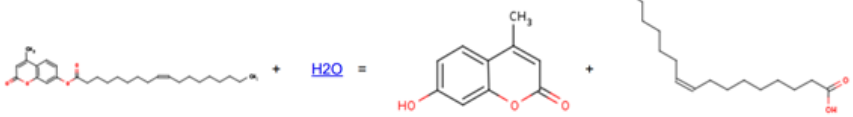
Target 2[8]



Target 3[9]



Target 4[9]



EC 3.7.1 Target 1 ～ Target 4

EC 3.5.3 Target 1 ～ Target 4

BRENDA記載の文献反応

[8] Supporting Information for: Evolving to an Ideal Synthesis of Molnupiravir, an Investigational Treatment for COVID - 19, 2020

[9] BRENDA, https://www.brenda-enzymes.org/index.php

5.2. 以前までの結果 1 (分類モデル生成・評価)

パラメータ調整結果

→記述子数：27，各決定木の最大深さ：15, 決定木数 100 に決定

EC	要素数	適合率	再現率	F1 値	EC	要素数	適合率	再現率	F1 値
3.1.1	25	0.96	0.96	0.96	3.4.19	1	1.00	1.00	1.00
3.1.2	12	0.92	1.00	0.96	3.5.1	31	0.94	0.97	0.95
3.1.3	31	0.91	1.00	0.96	3.5.3	5	0.83	1.00	0.91
3.1.4	6	0.86	1.00	0.92	3.5.4	9	0.89	0.89	0.89
3.1.6	3	1.00	1.00	1.00	3.5.5	2	1.00	1.00	1.00
3.1.7	2	0.00	0.00	0.00	3.5.99	2	1.00	0.50	0.67
3.2.1	26	0.96	0.96	0.96	3.6.1	19	0.86	0.95	0.90
3.2.2	5	0.83	1.00	0.91	3.7.1	7	1.00	0.71	0.83
3.3.2	1	1.00	1.00	1.00	3.8.1	3	1.00	0.67	0.80
3.4.13	1	0.00	0.00	0.00	3.13.1	2	1.00	0.50	0.67
					合計	193			
					平均		0.85	0.80	0.81
					正解率				0.92

5.2. 以前までの結果 2

正解ラベル EC 3.1.1，EC 3.7.1，3.5.3 を持つターゲット (各 4 つ) のクラスを予測
→ 4 種のターゲットが異なるクラスに分類された

ターゲット (ラベル EC 3.1.1)

	1st	2nd	3rd
target1	3.1.1.	3.2.1.	3.7.1.
確率%	0.59	0.13	0.08
target2	3.2.1.	3.3.2.	3.2.2.
確率%	0.45	0.15	0.14
target3	3.1.1.	3.5.1.	3.7.1.
確率%	0.62	0.16	0.07
target4	3.1.1.	3.7.1.	3.1.2.
確率%	0.97	0.02	0.01

ターゲット (ラベル EC 3.7.1)

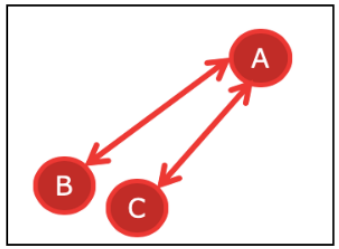
	1st	2nd	3rd
target1	3.7.1.	3.1.2.	3.5.1.
確率%	0.31	0.17	0.152
target2	3.2.2.	3.13.1.	3.5.4.
確率%	0.3495	0.210526	0.060526
target3	3.1.1.	3.7.1.	3.6.1.
確率%	0.29	0.22	0.11
target4	3.1.1.	3.5.1.	3.7.1.
確率%	0.28	0.24	0.13

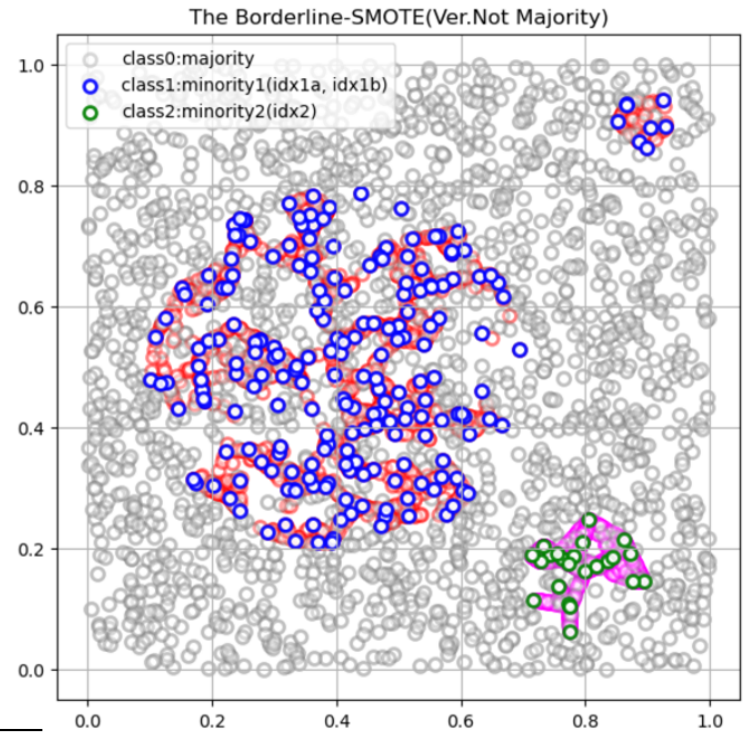
ターゲット (ラベル EC 3.5.3)

	1st	2nd	3rd
target1	3.5.3.	3.1.6.	3.5.99.
確率%	0.96	0.02	0.01
target2	3.5.3.	3.5.4.	3.5.99.
確率%	0.95	0.03	0.02
target3	3.5.3.	3.5.4.	3.5.99.
確率%	0.89	0.06	0.03
target4	3.5.3.	3.2.1.	3.1.1.
確率%	0.99	0.01	0.0

5.2. 提案手法の改善 1

SMOTE[10] を用いて，不均衡なクラスデータを調整 [11,12]





[10] Nitesh V. Chawla et al., 2002.

[11] 【リレー連載】わたしの推しノード –機械学習時代の申し子「SMOTE ノード」が不均衡データの壁を突破する，
https://www.ibm.com/blogs/solutions/jp-ja/spssmodeler-push-node-10/

[12] BorderlineSMOTE(Ver.Multiclass_Classification).ipynb.,
https://github.com/hkosho/pimientitosML/blob/main/%E3%80%90Pimientito's_ML-Lesson37%E3%80%91BorderlineSMOTE(Ver.Multiclass_Classification).ipynb

5.2. 提案手法の改善 2

最大データ数の EC 3.5.1 クラス (155 データ) に合わせてデータを増強

クラス名	反応式数	クラス名	反応式数	クラス名	反応式数
3.1.1	155	3.2.2	155	3.5.5	155
3.1.2	155	3.3.2	155	3.5.99	155
3.1.3	155	3.4.13	155	3.6.1	155
3.1.4	155	3.4.19	155	3.7.1	155
3.1.6	155	3.5.1.	155	3.8.1	155
3.1.7	155	3.5.3	155	3.13.1	155
3.2.1	155	3.5.4	155	合計	3100

訓練データ
計2480 (各クラス124)

テストデータ
計620 (各クラス31)

5.2. 改善結果 1 (分類モデル生成・評価)

記述子数：14，各決定木の最大深さ：20, 決定木数 200 に決定

以前の方法

	precision	recall	f1-score	要素数
3.1.1.	0.96	0.96	0.96	25
3.1.2.	0.92	1.00	0.96	12
3.1.3.	0.91	0.94	0.92	31
3.1.4.	0.86	1.00	0.92	6
3.1.6.	1.00	1.00	1.00	3
3.1.7.	0.00	0.00	0.00	2
3.13.1.	1.00	0.50	0.67	2
3.2.1.	0.96	0.96	0.96	26
3.2.2.	0.83	1.00	0.91	5
3.3.2.	1.00	1.00	1.00	1
3.4.13.	0.00	0.00	0.00	1
3.4.19.	1.00	1.00	1.00	1
3.5.1.	0.94	0.97	0.95	31
3.5.3.	0.83	1.00	0.91	5
3.5.4.	0.89	0.89	0.89	9
3.5.5.	1.00	1.00	1.00	2
3.5.99.	1.00	0.50	0.67	2
3.6.1.	0.86	0.95	0.90	19
3.7.1.	1.00	0.71	0.83	7
3.8.1.	1.00	0.67	0.80	3
accuracy			0.92	193
macro avg	0.85	0.80	0.81	193

SMOTE 導入後

	precision	recall	f1-score	要素数
3.1.1.	1.00	0.94	0.97	31
3.1.2.	1.00	1.00	1.00	31
3.1.3.	0.97	1.00	0.98	31
3.1.4.	1.00	0.97	0.98	31
3.1.6.	1.00	1.00	1.00	31
3.1.7.	1.00	1.00	1.00	31
3.13.1.	1.00	1.00	1.00	31
3.2.1.	1.00	1.00	1.00	31
3.2.2.	0.97	1.00	0.98	31
3.3.2.	1.00	1.00	1.00	31
3.4.13.	1.00	1.00	1.00	31
3.4.19.	0.97	1.00	0.98	31
3.5.1.	1.00	0.94	0.97	31
3.5.3.	1.00	0.94	0.97	31
3.5.4.	0.97	0.97	0.97	31
3.5.5.	0.97	1.00	0.98	31
3.5.99.	0.94	1.00	0.97	31
3.6.1.	1.00	1.00	1.00	31
3.7.1.	0.94	1.00	0.97	31
3.8.1.	1.00	0.97	0.98	31
accuracy			0.99	620
macro avg	0.99	0.99	0.99	620

5.2. 改善結果2

複数の箇所で精度の向上が見られた

ターゲット(ラベルEC 3.1.1)

	1st	2nd	3rd
target1	3.1.1.	3.2.1.	3.7.1.
確率%	0.59	0.13	0.08
target2	3.2.1.	3.3.2.	3.2.2.
確率%	0.45	0.15	0.14
target3	3.1.1.	3.5.1.	3.7.1.
確率%	0.62	0.16	0.07
target4	3.1.1.	3.7.1.	3.1.2.
確率%	0.97	0.02	0.01

target1	3.1.1.	3.3.2.	3.2.1.
確率%	0.305	0.125	0.095
target2	3.1.1.	3.7.1.	3.2.1.
確率%	0.22	0.17	0.1435
target3	3.1.1.	3.6.1.	3.5.1.
確率%	0.385	0.16	0.14
target4	3.1.1.	3.7.1.	3.5.1.
確率%	0.8	0.055	0.055

ターゲット(ラベルEC 3.7.1)

	1st	2nd	3rd
target1	3.7.1.	3.1.2.	3.5.1.
確率%	0.31	0.17	0.152
target2	3.2.2.	3.13.1.	3.5.4.
確率%	0.3495	0.210526	0.060526
target3	3.1.1.	3.7.1.	3.6.1.
確率%	0.29	0.22	0.11
target4	3.1.1.	3.5.1.	3.7.1.
確率%	0.28	0.24	0.13

target1	3.7.1.	3.1.2.	3.1.1.
確率%	0.615	0.145	0.09
target2	3.13.1.	3.5.99.	3.7.1.
確率%	0.305	0.15	0.105
target3	3.7.1.	3.5.1.	3.13.1.
確率%	0.54	0.145	0.12
target4	3.7.1.	3.1.2.	3.1.1.
確率%	0.5	0.17	0.075

ターゲット(ラベルEC 3.5.3)

	1st	2nd	3rd
target1	3.5.3.	3.1.6.	3.5.99.
確率%	0.96	0.02	0.01
target2	3.5.3.	3.5.4.	3.5.99.
確率%	0.95	0.03	0.02
target3	3.5.3.	3.5.4.	3.5.99.
確率%	0.89	0.06	0.03
target4	3.5.3.	3.2.1.	3.1.1.
確率%	0.99	0.01	0.0

target1	3.5.3.	3.5.4.	3.5.99.
確率%	0.935	0.025	0.01
target2	3.5.3.	3.5.99.	3.5.4.
確率%	0.79	0.065	0.045
target3	3.5.3.	3.7.1.	3.1.1.
確率%	0.995	0.005	0.0
target4	3.5.3.	3.5.99.	3.1.1.
確率%	0.995	0.005	0.0

6 おわりに

おわりに

- SFS，ランダムフォレスト，SMOTE による実験データ (EC3 クラス) の 2,3 桁目予測手法を提案した

今後の改善

- 全クラスの 1 桁〜3 桁目予測への分類で拡張
  →取得したデータの整形に着手
- 4 桁目予測手法の開発
  →物理・化学記述子を用いた提案
- 実験データの補充
  →予測信頼性の向上
- 選択された記述子の重要度によるランク付け