

1. はじめに
2. 有機合成と酵素
3. 機械学習による  
EC 番号予測
4. 提案手法
5. 実験結果並びに  
考察
6. おわりに

# 有機合成に最適な酵素候補提示のための 特徴量エンジニアリングによる EC 番号予測

EC Number Prediction Using Feature Engineering  
to Present Optimal Enzyme Candidates  
in Organic Synthesis

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

N212, 10:00-10:30, Tuesday, February 13, 2024.

# 1. はじめに

2/18

## 1.1 研究背景

有機合成化学において、生体触媒の効率性や環境面から化学反応の設計に酵素を生体触媒として利用される機会が増加している。酵素は EC 番号によって分類されており、代謝経路の解析や新たな酵素反応設計のため、機械学習で EC 番号を予測し、酵素の性質を特定する研究が行われている。

## 1.2 本研究の目的

有機合成に用いる酵素を探索する実験コストや時間削減のため、化学反応に最適な酵素候補を EC 番号として予測できる EC 番号予測手法を開発する。

### 1. 代謝経路の解明 = 生体の機能の解明

未知のタンパク質配列

[ MAKLLLLIFGVFIFVNSQAQTFPTILEKHN · · · ]

どんな性質か知る  
時間 大  
コスト 大

まず大まかに  
知りたい

?

### 2. 新たな化合物の設計 = 医薬品など

酵素(生体触媒)

効率よく反応  
環境にやさしい

A + B → C

どの酵素最適か？  
時間 大  
コスト 大

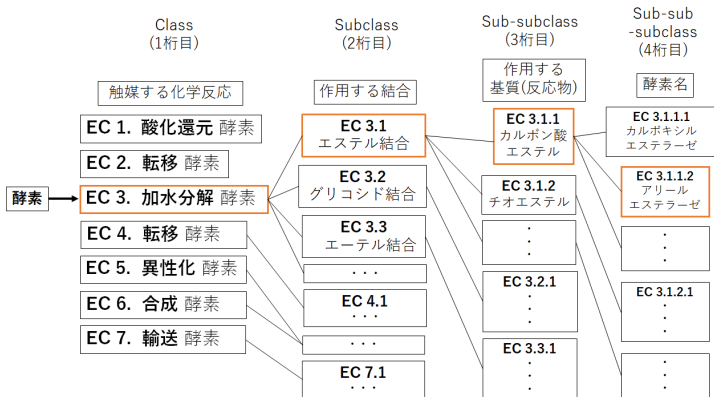
候補絞りたい

?

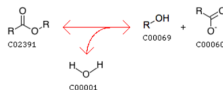
## 2. 酵素と EC 番号

酵素を 4 組の数字 (EC 〇. 〇. 〇. 〇) の組み合わせで分類したもの.  
EC 番号の機械学習予測 = 酵素候補の絞り込み

1. はじめに
2. 有機合成と酵素
3. 機械学習によるEC 番号予測
4. 提案手法
5. 実験結果並びに考察
6. おわりに



例. EC 3.1.1.1 の酵素を用いた反応



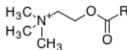
# 3.1 計算機上における化学反応の表現法

4/18

## 計算機上で化学反応を表現する各種方法

1. はじめに
2. 有機合成と酵素
3. 機械学習によるEC 番号予測
4. 提案手法
5. 実験結果並びに考察
6. おわりに

構造式



計算機表現

SMILES

\*C(=O)OCC[N+](C)(C)C

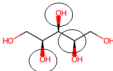
フィンガープリント

(100001011 ···) =

化合物の構造を表現

例1)官能基の有無

例2)分子の結合関係



物理・化学的特性値

分子量, 電荷, 疎水性, etc.

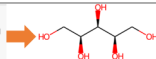
例) 化合物A: (100, 8.32, -0.23, ···)

→化学反応の表現(特性値ベクトル)

RDKit

化学のデータ分析モジュール(Python)  
→210種類の特性値(記述子)

```
from rdkit import Chem
Xylitol = Chem.MolFromMolFile('Xylitol.mol')
```



```
from rdkit.Chem import Descriptors
print("SMILES: " + Chem.MolToSmiles(Xylitol))
print("分子量: " + str(Descriptors.MolWt(Xylitol)))
print("LogP: " + str(Descriptors.MolLogP(Xylitol)))
print("TPSA: " + str(Descriptors.TPSA(Xylitol)))
```

SMILES: OC[C@H](O)[C@H](O)[C@H](O)CO

分子量: 152.14600000000002

LogP: -2.9462999999999995

TPSA: 101.15

## 3.2 EC 番号予測手法<sup>1, 2, 3</sup>

5/18

### EC 番号予測の目的

酵素探索の短縮: 既存データで学習&予測精度向上 → (将来的) 未知データ適用

(1) タンパク質配列<sup>1</sup>      予測範囲: 1桁目~4桁目  
画像処理(CNN), 自然言語処理(Transformer)ベースの予測

MAKLLLLNSQAQTFPTILEKHN · · ·

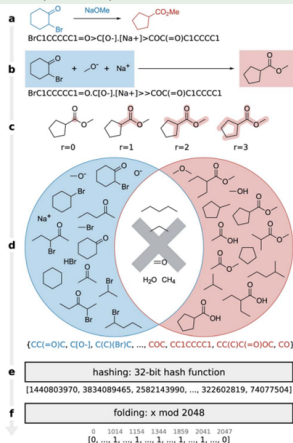
(2) 化合物の物理・化学的特性値<sup>2</sup>      予測範囲  
分子量, 電荷, 疎水性など      1桁目~3桁目  
データ数小

(3) 差分反応フィンガープリント<sup>3</sup>      予測範囲  
SMILESをハッシュ化      1桁目~3桁目  
→ 2048次元のバイナリベクトル      データ数大

EC 番号反応式: 反応物 1 + 反応物 2  $\longleftrightarrow$  生成物 1 + 生成物 2  
→  $RFP = FP_{\text{生成物 1} + \text{生成物 2}} - FP_{\text{反応物 1} + \text{反応物 2}}$

反応物 → 生成物の変化  
= フィンガープリントの変化

有機合成目線



<sup>1</sup>Naoki Watanabe et al., 2023.

<sup>2</sup>Diogo A. R. S. Latino et al., 2009.

<sup>3</sup>Daniel Probstl., 2023.

SMILESの2値ベクトル化<sup>3</sup>

### 3.3 機械学習と特徴量エンジニアリング

6/18

#### ランダムフォレスト (RF)<sup>4</sup> による EC 番号分類

各ノードで情報利得 (IG) を最大にする記述子  $f$  と分割閾値を決定

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

$D_P$  : 上位ノードに属するデータ

$f$  : ノード分割に用いる特徴量

$D_{left}, D_{right}$  : 下位ノードに属するデータ

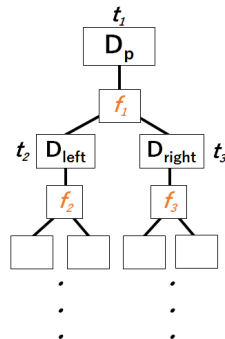
$N_P, N_{left}, N_{right}$  : 上位, 下位ノードのデータ数

ノード  $t$  のジニ不純度 :

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

$p(i|t)$  : クラス  $i$  の割合

$c$  : クラス数



<sup>4</sup>Leo Breiman., 2001.

### 3.3 機械学習と特徴量エンジニアリング

7/18

#### ラッパー法による記述子選択 (SequentialFeatureSelector(SFS)<sup>5</sup>)

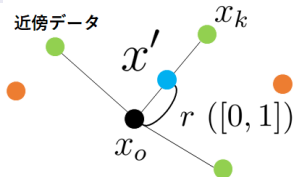
【Step Forward 法】 210 種から分類精度を高める記述子組合せを選択

- ① 記述子  $n$  ( $1 \leq n \leq 209$ ) 個から 1 つ選択し,  $210 - n$  種類の分類モデルを作成
- ② Macro F1-Score が最も高いモデルの記述子組合せを選択
- ③ 指定した記述子数になるまで 1 と 2 を繰り返す.

#### SMOTE<sup>6</sup> によるオーバーサンプリング

EC 番号データ： 多数クラスと少数クラスの間データ差大きい

→ (多数クラスに比べ) 少数クラスの正分類が難しい



$K = 3$

1.  $x_o$  の近傍データ点を  $K$  個選択
2.  $K$  個から 1 個 ( $x_k$ ) 選択
3.  $x_o$  と  $x_k$  間に新データ ( $x'$ ) を生成

$$x' = x_o + r(x_k - x_o)$$

少数クラスを閾値数までオーバーサンプリング

<sup>5</sup> Mlxtend.feature selection,

[http://rasbt.github.io/mlxtend/api\\_subpackages/mlxtend.feature\\_selection/](http://rasbt.github.io/mlxtend/api_subpackages/mlxtend.feature_selection/)

<sup>6</sup> Nitesh V. Chawla et al., 2002.

# 4.1 提案手法の概要

8/18

## RDKit 特性値を用いた EC 番号予測

反応物から生成物に変化するときの 210 種類の特性値変化量を計算

1. はじめに
2. 有機合成と酵素
3. 機械学習による EC 番号予測
4. 提案手法
5. 実験結果並びに考察
6. おわりに

特性値(RDKit)    分子量, 電荷, 疎水性    官能基1, 官能基2, 官能基3

化合物     $A = (100, 8.32, -0.23, \dots, 1, 0, 0, \dots)$

210種

$B = (99, 9.32, -6.23, \dots, 0, 0, 0, \dots)$

$C = (100, 8.32, -0.23, \dots)$

$D = (89, 7.32, 0, \dots)$

反応式:  $A + B \longrightarrow C + D$

特徴ベクトル:  $(A + B) - (C + D)$   
 $= (10, 2, -6.23, \dots, 1, 0, -1, \dots)$

210次元

### 有意性

- (2)特性値ベース  
 +  
 (3)フィンガープリント(一部)の組み合わせ

### (3)との比較

- ・ 1~3桁目の予測(同様)
- ・ 構造情報 + 物理化学情報

→ 化学反応時の特徴をより詳細化

特徴ベクトル  
↓

RDKit特性値(記述子)

	MaxEStateIndex	MinEStateIndex	MinAbsEStateIndex	qed
4.1.1.74	-7.449074	-2.629630	-0.064815	-0.360209 -1.
1.2.1.8	-0.197403	1.307870	0.405116	-0.079826 0.
2.5.1.85	0.593569	0.196239	-2.312624	0.488055 -4.
1.4.1.4	0.234718	-0.413194	0.418052	0.389325 0.
1.1.1.3	-0.155930	-0.317778	0.059255	0.016389 -2.
...	...	...	...	...
4.4.1.13	-3.236897	-0.282721	-1.272102	-0.270358 5
2.3.1.-	-0.286151	0.039395	0.454936	-0.386083 0.
2.3.1.57	-0.286151	0.039395	0.454936	-0.386083 0.



## 4.1 提案手法の概要

9/18

### 特徴ベクトルの多クラス分類

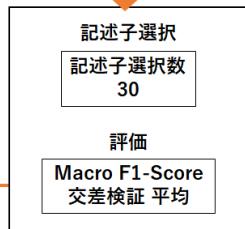
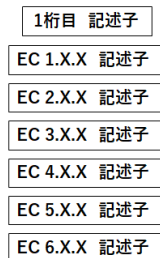
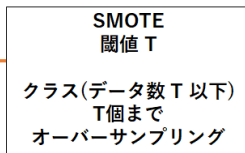
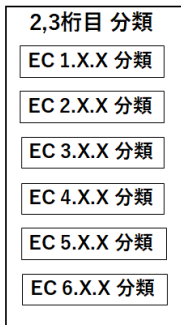
EC 番号の 1~3 桁目までを多クラス分類

→ EC 番号 1 桁目 + EC T.X.X ( $T = 1, 2, \dots, 6$ ) の分類 (記述子選択) を実施

1. はじめに
2. 有機合成と酵素
3. 機械学習による EC 番号予測
4. 提案手法
5. 実験結果並びに考察
6. おわりに

学習データ (80%)

検証用学習データ  
(交差検証)



1桁目 分類

記述子選択(最終)

評価値  
未更新 4回目直前の  
記述子組合せ

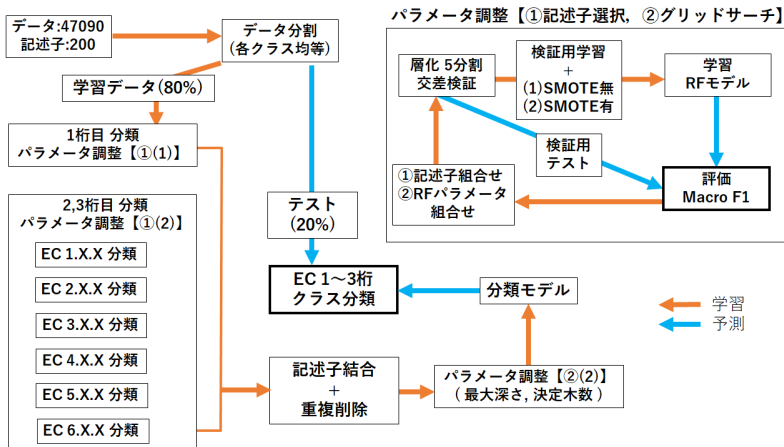
## 4.2 EC 番号予測モデルの構築と予測

10/18

### モデル作成・予測手順

記述子選択 (7 回分) 結合→重複削除&グリッドサーチで分類モデル作成

1. はじめに
2. 有機合成と酵素
3. 機械学習による EC 番号予測
4. 提案手法
5. 実験結果並びに考察
6. おわりに



## 4.3 提案手法の実装と流れ

11/18

### デモ動画による説明

1. はじめに
2. 有機合成と酵素
3. 機械学習による  
EC 番号予測
4. 提案手法
5. 実験結果並びに  
考察
6. おわりに

## 5.1 数値実験の概要

12/18

### 数値実験の流れ

【予備実験 1】 SMOTE 適用前と適用後に対するクラス分類精度の比較

【予備実験 2】 記述子選択, パラメータ調整 (本実験用)

【本実験】 EC1~3 桁までの多クラス分類

#### 【予備実験1】 RF × 記述子選択

##### SMOTE 未適用

クラス	データ数	クラス	データ数	クラス	データ数
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

##### SMOTE 適用

層化5分割交差検証

検証用学習データにSMOTE

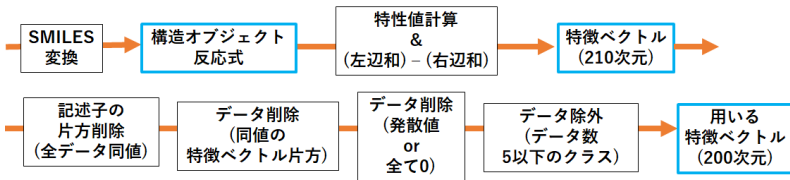
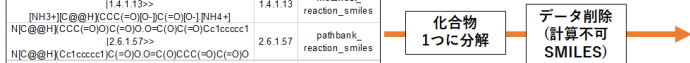
1. はじめに
2. 有機合成と酵素
3. 機械学習による EC 番号予測
4. 提案手法
5. 実験結果並びに考察
6. おわりに

## 化学反応データの取得と整形

4つのデータベース (Rhea, BRENDA, MetaNetX, PathBank) からなる SMILES データセット<sup>7</sup>を使用

rxn	ec	source
CC(=O)C(=O)[O-] [4.1.1.74>> CC=O.O=C=O	4.1.1.74	brenda_ reaction_smiles
NC(=O)CC[C@H](NH3+)[C-](O)[O-] [1.4.1.13>> [NH3+][C-@H](C)CC(=O)[O-].[C-](O)[O-].[NH4+]	1.4.1.13	metanex_ reaction_smiles
N[C@@H](C)CCC(=O)O[C-](=O)O.O=C(O)(C)[O-]Cc1ccccc1 [2.6.1.57>>	2.6.1.57	pathbank_ reaction_smiles
N[C@@H](C)C1cccc1[C-](O)O.O=C(O)CCC(=O)C(=O)O COc1cc(C=C/C(C)=O)CCC[N+](C)(C)Cc2OC)c1O.O [3.1.1.49>> COc1cc(C=C/C(C)=O)[O-] )Cc2OC)c1O.C[N+](C)(C)CCO.[H+]	3.1.1.49	rhea_ reaction_smiles
...	...	...

元のデータセット



<sup>7</sup>Daniel Probstl., 2023.

## 5.2 実験結果と考察

14/18

### 予備実験 1 結果

#### EC 3 (20 クラス, 962 データ) 2,3 桁目の多クラス分類比較

SMOTE 未適用

	precision	recall	f1-score	support
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
weighted avg	0.91	0.92	0.91	193

SMOTE適用

	precision	recall	f1-score	support
3.1.1.	1.00	0.96	0.98	25
3.1.2.	1.00	1.00	1.00	12
3.1.3.	0.97	0.94	0.95	31
3.1.4.	1.00	1.00	1.00	6
3.1.6.	0.75	1.00	0.86	3
3.1.7.	1.00	1.00	1.00	2
3.13.1.	0.50	0.50	0.50	2
3.2.1.	1.00	0.96	0.98	26
3.2.2.	0.71	1.00	0.83	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.	1.00	1.00	1.00	5
3.5.4.	0.88	0.78	0.82	9
3.5.5.	1.00	1.00	1.00	2
3.5.99.	0.67	1.00	0.80	2
3.6.1.	0.89	0.89	0.89	19
3.7.1.	0.88	1.00	0.93	7
3.8.1.	1.00	0.67	0.80	3
accuracy			0.94	193
macro avg	0.86	0.88	0.87	193
weighted avg	0.94	0.94	0.94	193

1. はじめに
2. 有機合成と酵素
3. 機械学習による  
EC 番号予測
4. 提案手法
5. 実験結果並びに  
考察
6. おわりに

## 5.2 実験結果と考察

15/18

### 予備実験 2 結果

【記述子選択】 足し合わせ × 重複削除で 93 種選択

【最適パラメータ】 最大深さ 90, 決定木数 300

SMOTE増加数

合計の2%

～

最多クラス数

EC 1.X.X計	クラス数	SMOTE 増加数	1.1.1	1.14.13	1.2.1	...	1.4.3	...	1.23.5
6380	64	3%(191)	1745	761	666	...	156	...	5
EC 2.X.X計	クラス数	SMOTE 増加数	2.7.8	2.3.1	2.1.1	...	2.6.1	...	2.7.3
23160	24	2%(463)	10074	7309	2797	...	280	...	5
EC 3.X.X計	クラス数	SMOTE 増加数	3.1.1	3.1.3	3.6.3	...	3.1.4	...	3.3.1
5377	27	10%(538)	2277	589	508	...	104	...	5
EC 4.X.X計	クラス数	SMOTE 増加数	4.1.1	4.2.1	4.1.2	...	4.1.3	...	4.6.1
1878	14	20%(376)	1037	361	106	...	32	...	5
EC 5.X.X計	クラス数	SMOTE 増加数	5.5.1	5.3.1	5.3.3	...	5.4.3	...	5.1.1
273	12	最多(80)	80	46	44	...	12	...	5
EC 6.X.X計	クラス数	SMOTE 増加数	6.2.1	6.3.2	6.3.4	6.3.5	6.3.1	6.4.1	6.1.2
604	7	最多(266)	266	233	30	29	22	18	6

記述子選択  
(本実験用)

	選択記述子数
EC X	19
EC 1.X.X	27
EC 2.X.X	20
EC 3.X.X	28
EC 4.X.X	21
EC 5.X.X	13
EC 6.X.X	15



記述子結合  
+  
重複削除



93種

## 5.2 実験結果と考察

16/18

### 本実験結果 (EC 1 桁～3 桁の多クラス分類)

84 種の記述子でグリッドサーチしたモデルに Test データを適用

	EC2	EC1	EC3	EC4	EC6	EC5	Total
<b>Train</b>	23160	6380	5377	1878	604	273	37672
<b>Test</b>	5789	1601	1345	462	154	67	9418
						<b>ALL</b>	47090

クラス数 148  
テストデータ 9418

ECクラス	データ数	Precision	Recall	Macro F1-Score
EC 1.X.X	1601	0.80	0.78	0.78
EC 2.X.X	5789	0.83	0.81	0.81
EC 3.X.X	1345	0.81	0.87	0.83
EC 4.X.X	462	0.86	0.85	0.84
EC 5.X.X	67	0.66	0.71	0.68
EC 6.X.X	154	0.96	0.75	0.81
合計	9418			
Macro Average		0.81	0.80	0.79
Weighted Average		0.96	0.95	0.95
Accuracy		0.95		

1. はじめに
2. 有機合成と酵素
3. 機械学習による EC 番号予測
4. 提案手法
5. 実験結果並びに考察
6. おわりに



## 5.2 実験結果と考察

17/18

### 考察

- ・選ばれた各 EC クラスの記述子
  - どのような種類の記述子が集まっているのか分析
  - 分析結果〜〜
  - フィンガープリントの手法では得られないもの

1. はじめに
2. 有機合成と酵素
3. 機械学習による EC 番号予測
4. 提案手法
5. 実験結果並びに考察
6. おわりに

### おわりに

- SMOTE によるオーバーサンプリングの有効性を示した
- 多クラス分類で得られた記述子の分析  
→さらに優れたモデルの構築

### 今後の課題

- EC 番号 4 桁目までの予測
- 実際実験に利用