

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
- 参考文献

# 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

2/28

## 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

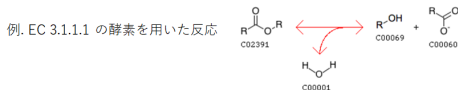
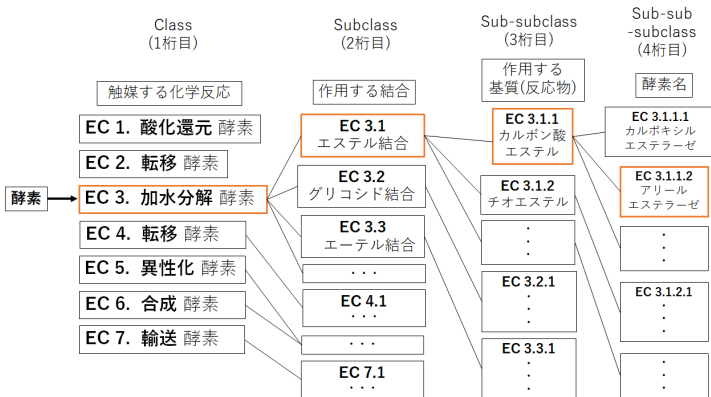
?

## 2. 酵素と EC 番号

3/28

酵素を 4 組の数字 (EC ○. ○. ○. ○) 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
- 参考文献

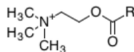


# 3.1 計算機上における化学反応の表現法

4/28

## 計算機上で化学反応を表現する各種方法

構造式



RDKit

化学のデータ分析モジュール(Python)

210種類の記述子

- ・特性値 125種
- ・部分構造のバイナリ値 85種



- ・合成材料探索
- ・生体反応の予測

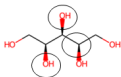
計算機表現

SMILES

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

フィンガープリント

(100001011 · · ·) =



化合物の構造表現

- 手法1)部分構造の有無
- 手法2)分子の結合関係

物理・化学的特性値

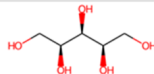
分子量, 親油性, 電荷分布, etc.

例) 化合物A: (100, -0.23, 8.32, · · · )  
→ 化学反応の表現 (特性値ベクトル)

```
from rdkit import Chem
```

```
Xylitol = Chem.MolFromMolFile("Xylitol.mol")
```

```
Xylitol
```



```
from rdkit.Chem import Descriptors
```

```
print(f"SMILES: {Chem.MolToSmiles(Xylitol)}")
```

```
print(f"分子量: {Descriptors.MolWt(Xylitol)}")
```

```
print(f"親油性: {Descriptors.MolLogP(Xylitol)}")
```

```
print(f"電荷分布: {Descriptors.BCUT2D_CHGI(Xylitol)}")
```

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

```
分子量: 152.14600000000002
```

```
親油性: -2.9462999999999995
```

```
電荷分布: 2.221860407854264
```

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
- 参考文献

## 機械学習での分類と回帰の違い

### 分類モデル

データ( $X_i$ 次元)を適切なクラスに分類する

学習データ

正解1



0~9

10クラスに分類

正解6



正解0



正解5



分類モデル

予測  
(正しいクラスに分類されるか)

テストデータ



正解2 正解3

### 回帰モデル

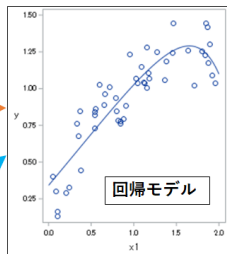
目的変数Yと説明変数 $X_1, X_2, \dots$ の関係を推定

データ  
( $X_i$ 次元)

学習  
データ

予測  
当てはまりの良さ  
(決定係数など)

テスト  
データ



回帰モデル

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_i X_i$$

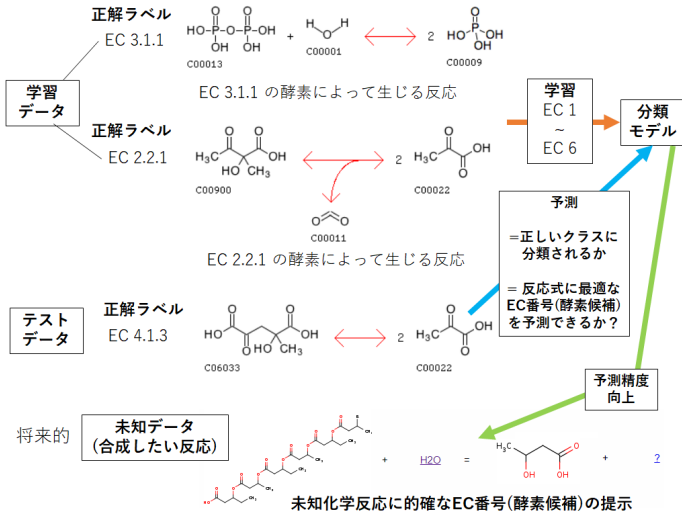
画像

<https://ex-ture.com/blog/2021/01/08/freehand-accuracy/>

<https://www.n-insight.co.jp/niblog/20190917-1351/>

## 本研究のクラス分類

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
- 参考文献



## 分類精度の評価指標

クラス1	クラス2	合計
980	20	1000
正解	正解	
900/980	0/20	



Accuracy  
 $900/1000 = 0.9$

	実際クラス1	実際クラス2	計
予測クラス1	900	80	980
予測クラス2	20	0	20
計	920	80	1000
Precision (正確性)	0.97826087	0	
Recall(感度)	0.91836735	0	
F1-Score (調和平均)	0.94736842	0	

Precision, Recall, F1-Score で  
適切に評価

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
- 参考文献

## 3.2 EC 番号予測手法<sup>1, 2</sup>

8/28

### EC 番号予測の目的

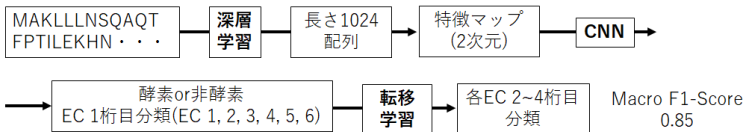
酵素探索の短縮: 既存データで学習&予測精度向上 → (将来的) 未知データ適用

#### (1) タンパク質配列<sup>1</sup>

予測範囲 1~4桁目

→ 代謝経路の解析

画像処理(CNN) を用いた予測



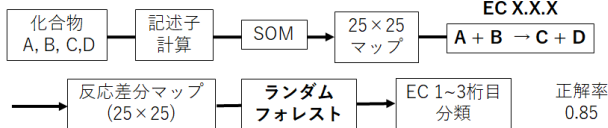
#### (2) 化合物の物理・化学的特性値<sup>2</sup>

55種の記述子(結合特性, 電荷など)を用いた予測

予測範囲: 1~3桁目

データ数 約7,500

有機合成  
目線

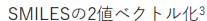


<sup>1</sup>Naoki Watanabe et al., 2023.

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



反応物から生成物の変化による予測



<sup>4</sup>Daniel Probst., 2023.

分野理解：論文内の参考文献をさかのぼる

その分野の現状と課題の把握：最新の手法を探す

分子間の結合関係に関するフィンガープリント (Q-N Hu., 2012)

EC 番号反応式：反応物 1 + 反応物 2  $\longleftrightarrow$  生成物 1 + 生成物 2

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

予測範囲：1~3桁目  
データ数 約5,000

Accuracy:  
0.92

Google Scholar 被引用検索

「Assignment of EC Numbers to Enzymatic Reactions with Reaction Difference Fingerprints (2012)」

【最新手法】差分反応フィンガープリント (D Probst., 2023)

予測範囲：1~3桁目  
データ数 約80,000

Accuracy:  
0.93  
0.95(データ数約7000)

Macro F1-Score  
0.77  $\pm$  0.01

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  
参考文献

# 先行研究の調査方法

11/28

- はじめに
  - 有機合成と酵素
  - 機械学習によるEC番号予測
  - 提案手法
  - 実験結果並びに考察
  - おわりに
- 参考文献

Google Scholar

Assignment of EC Numbers to Enzymatic Reactions with Reaction Difference

記事

期間指定なし  
2024 年以降  
2023 年以降  
2020 年以降  
期間を指定...

関連性で並べ替え  
日付順に並べ替え

すべての言語  
英語と日本語のページを検索

すべての種類  
総説論文

☐ 特許を含める  
☒ 引用部分を含める

Assignment of EC numbers to enzymatic reactions with reaction difference fingerprints  
QIN Hu, H Zhu, X Li, M Zhang, Z Deng, X Yang, Z Deng  
PloS one, 2012 - journals.plos.org

The EC numbers represent enzymes and enzyme genes (genomic information), but they are also utilized as identifiers of enzymatic reactions (chemical information). In the present work (ECAssigner), our newly proposed reaction difference fingerprints (RDF) are applied to assign EC numbers to enzymatic reactions. The fingerprints of reactant molecules minus the fingerprints of product molecules will generate reaction difference fingerprints, which are then used to calculate reaction Euclidean distance, a reaction similarity

さらに表示 ▾

☆ 保存 引用 **被引用数: 47** 関連記事 全 11 ページ 80

この検索の最上位の結果を表示しています。 検索結果をすべて見る

Assignment of EC numbers to enzymatic reactions with reaction difference fingerprints  
Symmetry-informed geometric representation for molecules, proteins, and crystalline materials  
Li Xia, Xu Z, Li Z, Zhang C, Qian ... - Advances in ... proceedings.neurips.cc  
Artificial intelligence for scientific discovery has recently generated significant interest within the machine learning and scientific communities, particularly in the domains of chemistry ... ☆ 保存 引用 被引用数: 10 関連記事 全 5 ページ 80

Artificial intelligence in synthetic chemistry: achievements and prospects  
B. Babin, T. M. M. ... - Russian Chemical ... 2017 - iopscience.iop.org  
The review is devoted to the achievements in analysis of information on chemical reactions using machine learning methods. Four large areas that actively use these methods are ... ☆ 保存 引用 被引用数: 63 関連記事 全 9 ページ 80

Turnover number predictions for kinetically uncharacterized enzymes using machine and deep learning  
A. K. ... Y. Roussel, X. P. Hu, N. A. Liebrand ... - Nature ... 2023 - nature.com  
The turnover number  $k_{cat}$ , a measure of enzyme efficiency, is central to understanding cellular physiology and resource allocation. As experimental  $k_{cat}$  estimates are unavailable ... ☆ 保存 引用 被引用数: 15 関連記事 全 11 ページ 80

Carboxylic ester hydrolases in bacteria: Active site, structure, function and application  
C. Shi, T. D. Kim, K. K. Kim ... - Crystalline ... 2019 - mdpi.com  
Carboxylic ester hydrolases (CEHs), which catalyze the hydrolysis of carboxylic esters to produce alcohol and acid, are identified in three domains of life. In the Protein Data Bank ... ☆ 保存 引用 被引用数: 30 関連記事 全 7 ページ 80

Generation of an atlas for commodity chemical production in *Escherichia coli* and a novel pathway prediction algorithm, GEM-Path  
M. Campondonio, B. A. Andrews, J. A. Aspar ... - Metabolic ... 2014 - Elsevier  
The production of 79% of the current drug molecules and 39% of all chemicals could be achieved through bioprocessing (Kordule and Sawaya, 2009). To accelerate the transition ... ☆ 保存 引用 被引用数: 107 関連記事 全 8 ページ 80

↑ より最新の手法が出ている可能性

## 3.3 機械学習と特徴量エンジニアリング

12/28

### ランダムフォレスト (RF)<sup>5</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

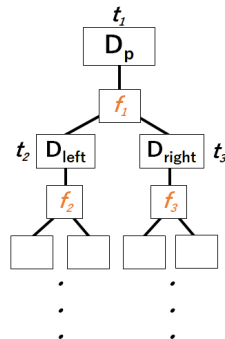
$N_P, N_{left}, N_{right}$  : 上位, 下位ノードのデータ数

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>5</sup>Leo Breiman., 2001.

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
- 参考文献

クラス1 980	クラス2 20	・ クラス1(多数クラス) の方が予測しやすい
正解 900/980	正解 0/20	・ クラス2(少数クラス) を適切に予測しにくい

Accuracy  
0.9

## 解決策



- ・ 機械学習手法の変更
- ・ **少数クラスを増やす**  
(オーバーサンプリング)
- ・ 多数クラスを減らす  
(アンダーサンプリング)

## オーバーサンプリング手法

- ・ ランダムに少数クラスを増やす  
→ 過学習の危険
- ・ 同じ少数クラスの間  
データを増やす(**SMOTE**)

## アンダーサンプリング手法

- ・ 多数クラスを減らす  
(多数クラスの重要データ**削除の危険**)

### 3.3 機械学習と特徴量エンジニアリング

14/28

#### ラッパー法による記述子選択 (SequentialFeatureSelector(SFS))<sup>6</sup>

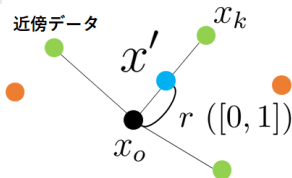
【Step Forward 法】 210 種から分類精度を高める記述子組合せを選択

- 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>7</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でデータ差大きい

→ (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>6</sup> Mlxtend.feature selection, [http://](http://rasbt.github.io/mlxtend/api_subpackages/mlxtend.feature_selection/)

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

<sup>7</sup> Nitesh V. Chawla et al., 2002.

# 4.1 提案手法の概要

15/28

## 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用いる

特性値(RDKit) 分子量, 親油性, 電荷分布 部分構造フラグ1, フラグ2, フラグ3

化合物 A = (100, -0.23, 8.32, . . . , 1, 0, 0, . . . )

210種

B = (99, 9.32, -6.23, . . . , 0, 0, 0, . . . )

C = (100, 8.32, -0.23, . . . )

D = (89, 7.32, 0, . . . )

反応式:  $A + B \rightarrow C + D$

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

210次元

### 提案手法の有意性

物理・化学特性値 + フィンガープリント(FP)の組み合わせ

### 【最新手法】(3)差分反応FPとの比較

化学反応の特徴をより詳細に捉える

特性値  $\longleftrightarrow$  FP  
反応特徴の充実 学習コスト高 構造情報のみ 学習コスト低(高次元)

RDKit  
特性値 125種  
FP(部分構造の有無) 85種

学習コスト抑制  
+  
化学反応の説明力向上

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.

- はじめに
  - 有機合成と酵素
  - 機械学習による EC 番号予測
  - 提案手法
  - 実験結果並びに考察
  - おわりに
- 参考文献

## 4.1 提案手法の概要

16/28

### 特徴ベクトルの RF 多クラス分類

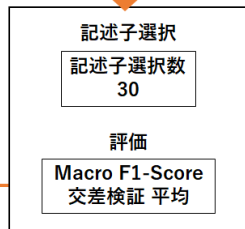
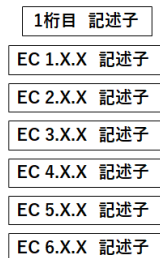
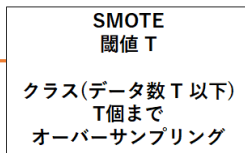
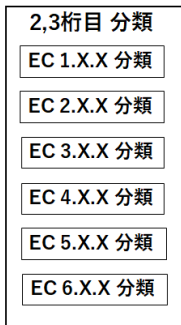
EC 番号の 1~3 桁目までを多クラス分類

→ 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
- 参考文献

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



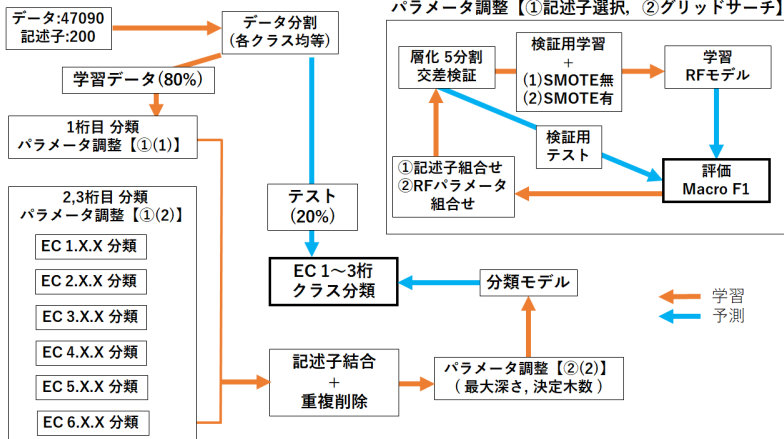
## 4.2 EC 番号予測モデルの構築と予測

17/28

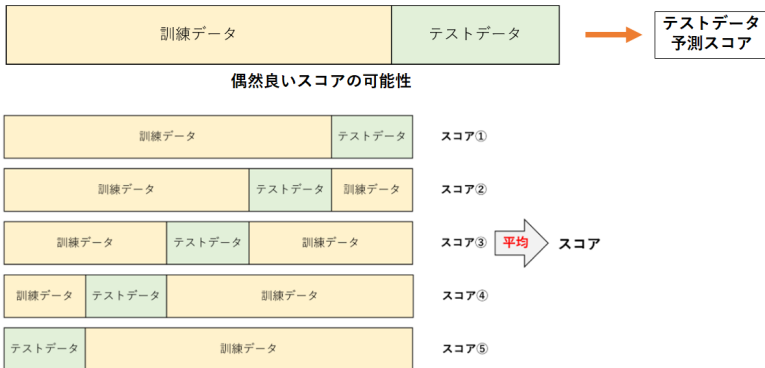
### モデル作成・予測手順

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
- 参考文献



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
- 参考文献

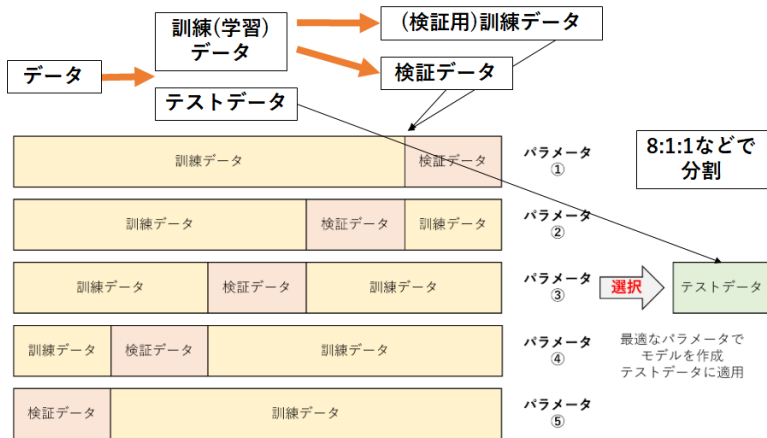


「学習データ」「テストデータ」の役に分けて平均評価(信頼性向上)

画像

[https://www.codexa.net/cross\\_validation/](https://www.codexa.net/cross_validation/)

## パラメータ調整の場合の交差検証



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
- 参考文献

## 4.3 提案手法の実装と流れ

20/28

### 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  
参考文献

## 5.1 数値実験の概要

21/28

### 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

【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

最多クラス(155)まで  
オーバーサンプリング

層化5分割交差検証

検証用学習データにSMOTE

- はじめに
  - 有機合成と酵素
  - 機械学習によるEC 番号予測
  - 提案手法
  - 実験結果並びに考察
  - おわりに
- 参考文献

## 5.1 数値実験の概要

22/28

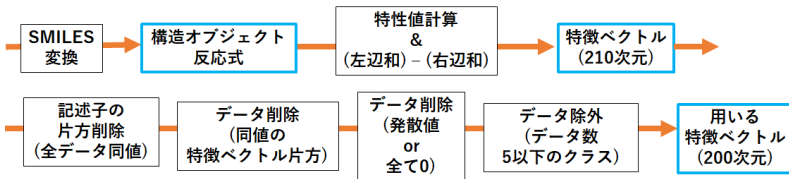
### 化学反応データの取得と整形

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>8</sup> を使用

rxn	ec	source
<chem>CC(=O)C(=O)[O-]</chem> [4.1.1.74>> <chem>CC=O.O=C=O</chem>	4.1.1.74	brenda_ reaction_smiles
<chem>NC(=O)CC[C@H]([NH3+])C(=O)[O-].O</chem> [1.4.1.13>> <chem>[NH3+][C@@H](X)CCC(=O)[O-]C(=O)[O-].[NH4+]</chem>	1.4.1.13	metanetx_ reaction_smiles
<chem>N[C@@H](X)CCC(=O)O.C(=O)O.O=C(O)C(=O)Cc1ccccc1</chem> [2.6.1.57>> <chem>N[C@@H](X)Cc1ccccc1.C(=O)O.O=C(O)CCC(=O)C(=O)O</chem>	2.6.1.57	pathbank_ reaction_smiles
<chem>CCOc1cc/C=C/C(=O)OCC[N+](X)(X)C(C)C(OC)c1O.O</chem> [3.1.1.49>> <chem>CCOc1cc/C=C/C(=O)[O-]</chem>	3.1.1.49	rhea_ reaction_smiles
<chem>]cc(OC)c1O.C[N+](X)(X)C(C)CCO.[H+]</chem> ...	...	...



元のデータセット



<sup>8</sup>Daniel Probstl., 2023.

## 5.2 実験結果と考察

23/28

### 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

- はじめに
  - 有機合成と酵素
  - 機械学習による  
EC 番号予測
  - 提案手法
  - 実験結果並びに  
考察
  - おわりに
- 参考文献

## 5.2 実験結果と考察

24/28

### 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

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
- 参考文献

SMOTE増加数

合計の2%  
～  
最多クラス数

↓  
学習時間  
&  
精度増加  
↓  
トレードオフ

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

25/28

### 本実験結果 (EC 1 桁～3 桁の多クラス分類)

93 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

学習 データ計	クラス数	SMOTE 増加数	2.7.8	2.3.1	...	4.1.1	1.14.13	...	3.6.4	2.7.3
37672	148	最少 × 100	10074	7309	...	1037	761	...	5	5

RFパラメータ調整  
(最大深さ, 決定木数)



Best分類器

- ・ 最大深さ=90
- ・ 決定木数=300



ECクラス	テスト 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  
参考文献

## 5.2 実験結果と考察

26/28

### 考察

- EC 5 の予測精度最も低い  
→他のクラスの酵素反応と類似性が高い＝誤分類されやすい<sup>9</sup>
- Macro F1-Score 0.79  
→ (3) 差分反応 FP と同等  
→パラメータ調整時間, 利用可能データ数の制限  
= 利用データ, 学習手法の改善
- 各 7 回で選択された記述子  
記述子の特徴を分析. 類似する記述子をまとめる (次元削減)
- 選択時, 記述子の重要度を評価して重み付けする手順の追加
- 必要最低限の記述子の絞り込み  
→長さ 1000 以上のフィンガープリントと組合せ (省学習コスト + 説明力向上)
- 別のクラス分類手法の検討 (XGBoost, LightGBM などの導入)

<sup>9</sup>Daniel Probstl., 2023.

## おわりに

- 有機合成に最適な酵素を EC 番号として提示する機械学習モデルを開発
- 酵素反応をより詳細に捉えるための、フィンガープリントと物理・化学特性値を組み合わせた手法を提案
- フィンガープリント差分手法と同程度の予測精度となり、提案手法の改善が求められる
- 選択された記述子組合せの特徴分析が必要

## 今後の課題

- EC 番号 1~4 桁目までの予測手法の開発  
→ 3 桁目よりもさらに詳細な分類手法や記述子組合せの利用
- 実際の有機合成でのモデル使用  
→ 現実的な実験条件や予測誤差のフィードバック

## 参考文献

- ① Naoki Watanabe, Masaki Yamamoto, Masahiro Murata, Yuki Kuriya, and Michihiro Araki, "EnzymeNet: residual neural networks model for Enzyme Commission number prediction", *Bioinformatics Advances*, Vol. 3, No. 1, 2023.
- ② Latino Diogo ARS and Aires-de-Sousa Joao, "Assignment of EC Numbers to Enzymatic Reactions with MOLMAP Reaction Descriptors and Random Forests", *Journal of chemical information and modeling*, Vol. 49, No. 7, pp. 1839-1846, 2009.
- ③ Daniel Probst, Philippe Schwaller, and Jean-Louis Reymond, "Reaction classification and yield prediction using the differential reaction fingerprint DRFP", *Digital discovery*, Vol. 1, No. 2, pp. 91-97, 2022.
- ④ Daniel Probst, "An explainability framework for deep learning on chemical reactions exemplified by enzyme-catalysed reaction classification", *Journal of Cheminformatics*, Vol. 15, No. 1, pp. 113, 2023.
- ⑤ Leo Breiman., "Random Forests", *Machine Learning*, Vol 45, pp. 5-32, 2001.
- ⑥ Chawla Nitesh V, Bowyer Kevin W, Hall Lawrence O, and Kegelmeyer W. Philip, "SMOTE: synthetic minority over-sampling technique", *Journal of artificial intelligence research*, Vol 16, pp. 321-357, 2002.