# 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

February 2, 2024

# 1. はじめに





## 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
？

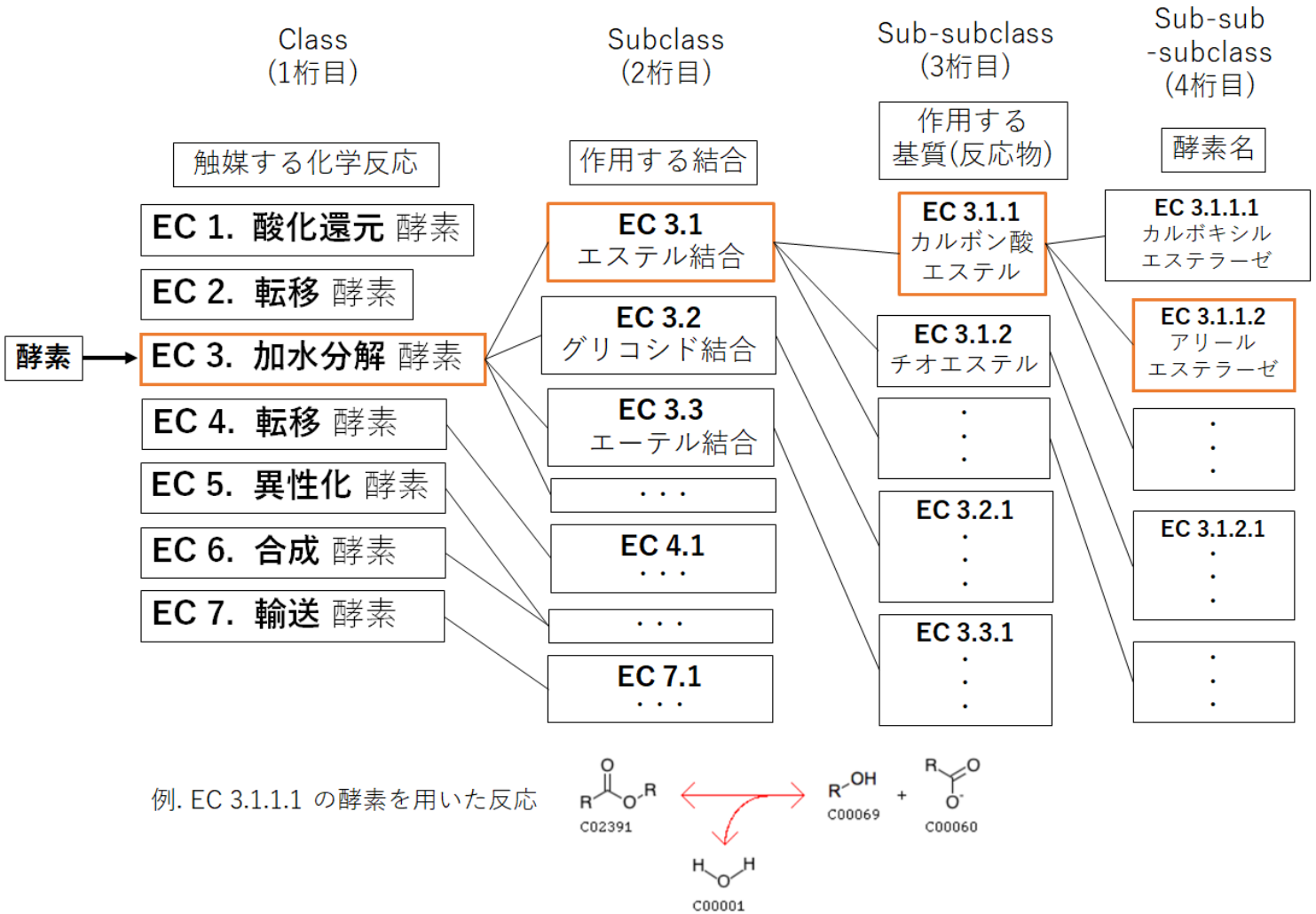
# 2. 酵素と EC 番号





酵素を 4 組の数字 (EC ○. ○. ○. ○) 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



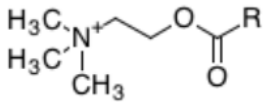
# 3.1 化合物の構造表現法





## 計算機上で化学反応を表現する各種方法

**構造式**

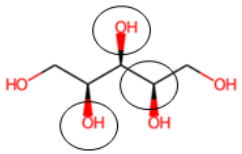


**計算機表現**

**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**

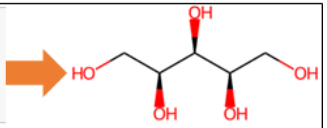
分子量，電荷，疎水性，etc.

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')
Xylitol
```



```
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 [123]





## 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[1]　**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

MAKLLLNSQAQTFPTILEKHN・・・

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

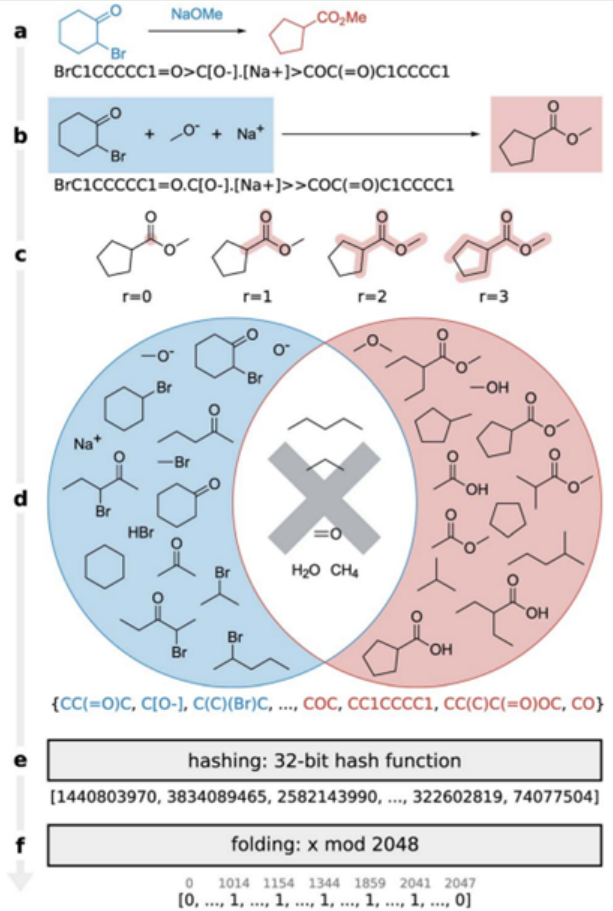
(3)差分反応フィンガープリント[3]
SMILESをハッシュ化
→2048次元のバイナリベクトル
予測範囲
1桁目～3桁目
**データ数大**

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

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

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




SMILESの2値ベクトル化[3]

[1]Naoki Watanabe et al., 2023.
[2]Diogo A. R. S. Latino et al., 2009.
[3]Daniel Probstl., 2023.

# 4.1 提案手法の概要





## 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 208 種類の特性値変化量を計算

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, ・・・, 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 ⟶ C + D

**特徴ベクトル：(A + B) - (C + D)**

**= (10, 2, -6.23, ・・・, 1, 0, -1, ・・・)**

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**

RDKit特性値(記述子)

特徴ベクトル↓

| | 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. |

# 4.2 提案手法 特徴量 (記述子) 選択





RDKit 記述子の中から必要なものだけを選択する

## ラッパー法による記述子選択 (SequentialFeatureSelector(SFS)[4])

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.

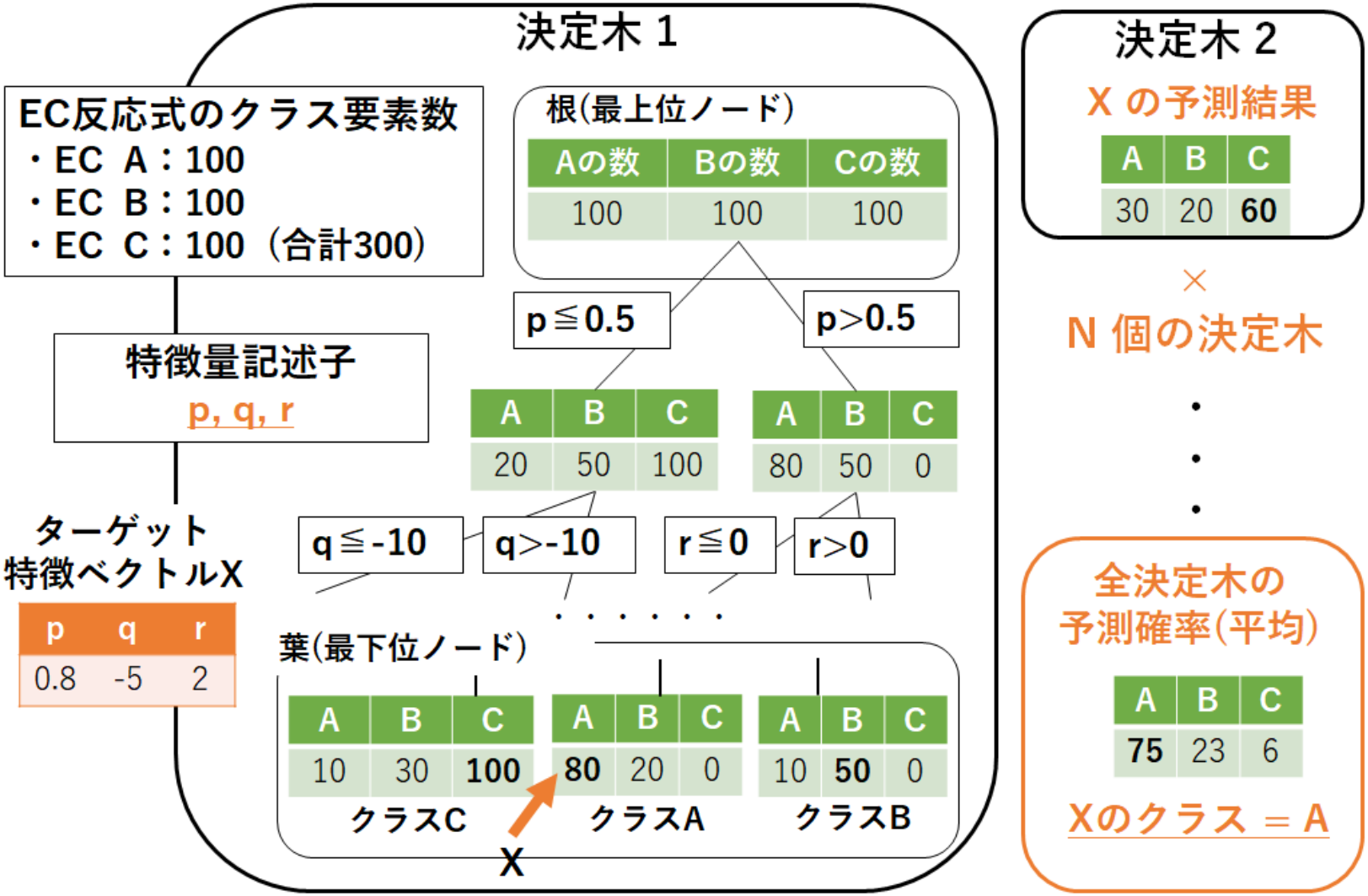
モデルの評価基準： F1 スコア平均 (層化 5 分割交差検証)

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

# 4.3 提案手法 ランダムフォレスト (RF) による EC 番号分類 (1)







# 4.3 提案手法ランダムフォレスト (RF) 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

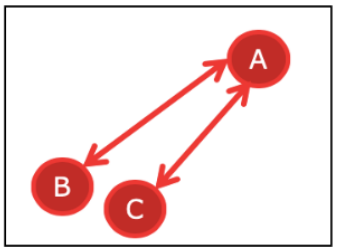
$$I_{imp}(t) = \sum_{i=1}^{c} p(i|t)(1 - p(i|t)) = 1 - \sum_{i=1}^{c} p(i|t)^2$$

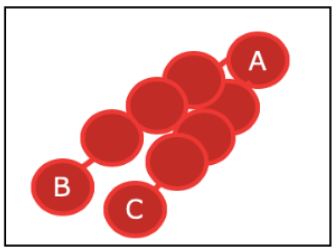
$p(i|t)$：ノード $t$ のクラス $i$ のデータ割合，$c$：ノード $t$ 内のクラス数

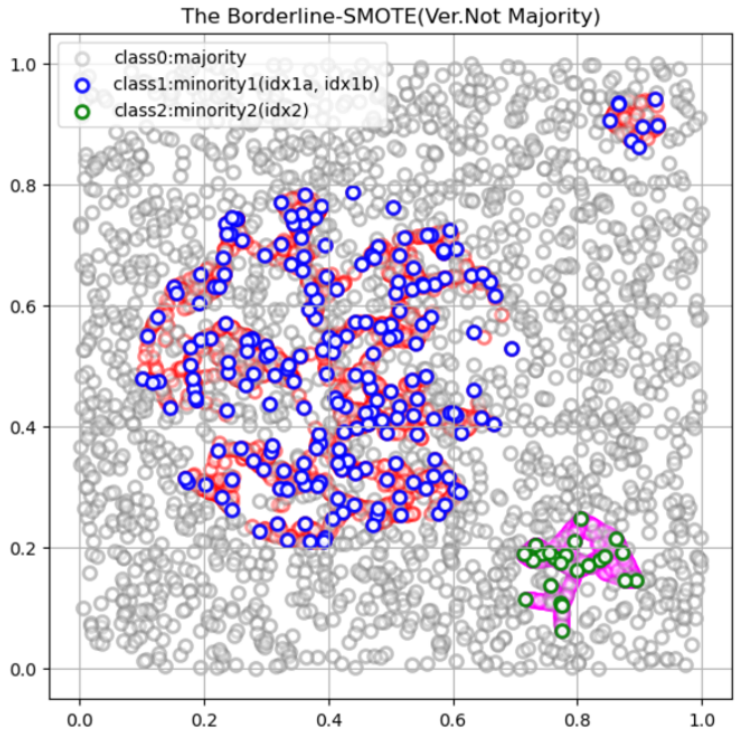
**パラメータ調整(グリッドサーチ)→各決定木の最大深さ、決定木数**

## SMOTE[5] によるオーバーサンプリング

不均衡なクラスデータを調整 [6,7]




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

[6] 【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/

[7] BorderlineSMOTE(Ver.Multiclass_Classification).ipynb., https://github.com/hkosho/pimientitosML/blob/main/

# 5.1 数値実験の概要





## 数値実験の流れ

3 種類実行

1. 従来の分類法 [8] と SMOTE 適用後の比較
2. 記述子選択
3. EC 1〜3 桁までの多クラス分類

[8] 武藤 克弥, 富山県立大学学位論文 2022

# 5.1 数値実験の概要





## 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 [9] を使用

[9] Daniel Probstl., 2023.

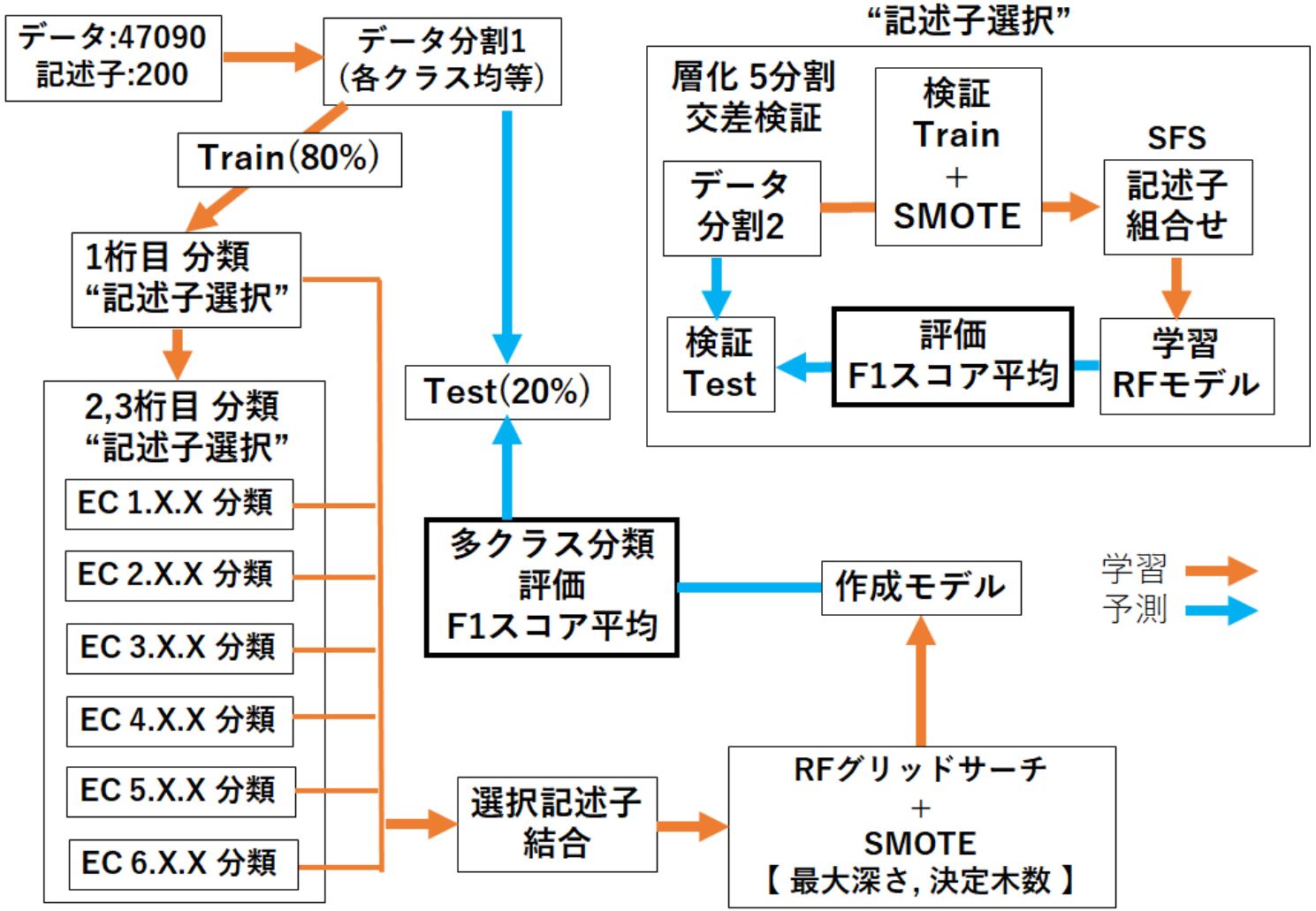
# 5.1 数値実験の概要





## モデル作成・予測手順

EC1, EC2, EC3, EC4, EC5, EC6 の多クラス分類を実行



# 5.2 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

卒業論文

| | 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適用後

| | 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 |

# 5.2. 実験２結果 記述子選択





## 特徴選択・パラメータ調整の結果

RF モデル各多クラス分類で 9〜27 種の記述子
足し合わせ × 重複削除で 84 種の記述子

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
|---|---|---|---|---|---|---|---|---|---|
| EC_1D | MolWt | NumValence | BCUT2D_M | BCUT2D_CHGH | HallKierAl | Ipc | Kappa3 | PEOE_VSA | SlogP_VSA2 |
| EC1 | MinEStateInd | BCUT2D_MR | BalabanJ | Chi4v | Kappa1 | Kappa3 | PEOE_VSA | SMR_VSA | SMR_VSA10 |
| EC2 | AvgIpc | Chi1 | Ipc | Kappa1 | PEOE_VSA | PEOE_VSA | SMR_VSA | SlogP_VSA | SlogP_VSA3 |
| EC3 | BCUT2D_MW | BCUT2D_MW | BCUT2D_L | Chi1 | Chi2n | Chi4v | Kappa1 | Kappa3 | PEOE_VSA12 |
| EC4 | BCUT2D_MW | BCUT2D_MW | BCUT2D_L | LabuteASA | SMR_VSA | SMR_VSA | VSA_ESta | RingCount | fr_Al_OH_noTert |
| EC5 | AvgIpc | BalabanJ | Chi1n | Chi2n | Chi3v | Chi4v | HallKierAl | VSA_ESta | NumHeteroatoms |
| EC6 | BCUT2D_MW | HallKierAlpha | Ipc | EState_VSA3 | VSA_ESta | FractionCS | fr_COO | fr_sulfide | fr_unbrch_alkane |
| | | | | | | | | | |
| | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 |
| EC_1D | SlogP_VSA3 | FractionCSP3 | fr_COO | end 12 | | | | | |
| EC1 | SMR_VSA3 | SMR_VSA7 | SlogP_VSA | EState_VSA1 | EState_VS | EState_VS | EState_VS | VSA_ESta | VSA_EState8 |
| EC2 | EState_VSA4 | VSA_EState1 | VSA_EStat | VSA_EState7 | VSA_ESta | VSA_ESta | NumAliph | fr_Al_OH | fr_C_S |
| EC3 | PEOE_VSA5 | PEOE_VSA9 | SMR_VSA1 | SlogP_VSA8 | VSA_ESta | NumArom | NumRotat | NumSatur | fr_C_O |
| EC4 | fr_COO2 | fr_NH1 | fr_aldehyde | fr_benzene | end 13 | | | | |
| EC5 | fr_C_O | fr_aldehyde | fr_ester | fr_ketone_Topli | end 13 | | | | |
| EC6 | end 9 | | | | | | | | |
| | | | | | | | | | |
| | 19 | 20 | 21 | 22 | 23 | 24 | 25 | 26 | 27 |
| EC_1D | end 12 | | | | | | | | |
| EC1 | VSA_EState9 | NumAliphatic | NumSatura | fr_Ar_COO | fr_Ar_OH | fr_NH2 | fr_SH | fr_aniline | fr_guanido |
| EC2 | fr_aldehyde | fr_epoxide | fr_ether | fr_imide | fr_lactone | fr_methox | fr_phos_ester | | end 25 |
| EC3 | fr_NH0 | fr_guanido | fr_hdrzone | fr_ketone | fr_ketone_ | fr_sulfide | end 24 | | |
| EC4 | end 13 | | | | | | | | |
| EC5 | end 13 | | | | | | | | |
| EC6 | end 9 | | | | | | | | |

# 5.2 実験3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 |
|---|---|---|---|---|---|---|---|
| Train | 23160 | 6380 | 5377 | 1878 | 604 | 273 | 37672 |
| Test | 5789 | 1601 | 1345 | 462 | 154 | 67 | 9418 |
| | | | | | | ALL | 47090 |

148 Class
9418 Test

| EC Class | Num of Data | precision average | recall average | f1-score macro average |
|---|---|---|---|---|
| EC 1.X.X | 1601 | 0.78 | 0.78 | 0.77 |
| EC 2.X.X | 5789 | 0.81 | 0.77 | 0.78 |
| EC 3.X.X | 1345 | 0.83 | 0.88 | 0.84 |
| EC 4.X.X | 462 | 0.81 | 0.85 | 0.81 |
| EC 5.X.X | 67 | 0.56 | 0.71 | 0.59 |
| EC 6.X.X | 154 | 0.93 | 0.76 | 0.82 |
| Total | 9418 | | | |
| macro average | | 0.79 | 0.80 | 0.78 |
| weighted average | | 0.95 | 0.95 | 0.95 |
| accuracy | | 0.95 | | |

# 5.2 考察





・EC 5 の予測精度が最も低い
→ EC 5 は他の EC 番号クラスとの類似性が高いことが知られている [10]

・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
→各 EC クラスを分類する際の手掛かりに
→フィンガープリントの手法では得られないもの

[10] Daniel Probstl., 2023.

## 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