卒業論文

勾配情報を活用した
粒子群最適化による
**な大気的なパレート解

Chemoinformatics Using Feature Selection and Clustering
for Enzyme Commission Number Prediction
in Organic Synthesis

富山県立大学 工学部 情報システム工学科
2120019 柴原壮大
指導教員 奥原 浩之 教授
提出年月: 2024年2月

目 次

図一覧

表一覧

記号一覧

以下に本論文において用いられる用語と記号の対応表を示す.

用語	記号
分子フィンガープリント	MFP
反応差分フィンガープリント	RFP
クラスタ i	C_i
C_i に属するデータ集合	$\mathbf{x_i}$
クラスタ間の距離	$d(C_1, C_2)$
クラスタ内の要素間の距離	$d(\mathbf{x_1}, \mathbf{x_2})$
個数	n
次元数	p
p 次元観測ベクトル	$\boldsymbol{x}_j$
i 番目のユニット	m_i
ユニット数	k
i 番目ユニットの重心	$\boldsymbol{r}_i$
i 番目ユニットの重みベクトル	$\boldsymbol{\xi}_i$
$\boldsymbol{x}_j$ と $\boldsymbol{\xi}_i$ のユークリッド距離	$\|\|\boldsymbol{x}_j - \boldsymbol{\xi}_i\|\|$
$\|\|\boldsymbol{x}_j - \boldsymbol{\xi}_i\|\|$ を最小化する $\boldsymbol{\xi}_i$	$\boldsymbol{\xi}_c$
$\boldsymbol{\xi}_c$ を持つ勝者ユニット	m_c
近傍関数	$h(t)$
ユニット m_c の近傍領域	N_c
学習率係数	$\alpha(t)$
N_c の散らばりに関する調整関数	$\sigma^2(t)$
反応物 i の特性値	RT_i
生成物 i の特性値	PD_i
記述子 j の特性値変化量	cv_j
i 番目の反応式の特徴ベクトル	$\boldsymbol{DF}_i$
記述子 u，v 間の相関係数	s_{uv}
記述子 u の特性値平均	$\bar{cv}_u$

第1章

はじめに

§ 1.1　本研究の背景

近年, ケモインフォマティクスと呼ばれる，化学に関するデータを情報技術を用いて分析する分野が発展してきている．化合物の特性や構造を分析したり，化合物の特徴を抽出し，機械学習における分類や化学反応の設計や予測といったことが行われている．

現在，新型コロナウイルスの世界的な流行をはじめとする多種の影響によって，新薬開発のニーズが高まっている．2026 年までの間に，ケモインフォマティクス業界は，年平均成長率 13 %で市場が成長すると予想されていることからも [1]，ケモインフォマティクスの需要は日々拡大している．

有機合成分野においては，ケモインフォマティクスや機械学習などの技術を取り入れて，化学反応の設計や予測をする研究が増加している．一方で，目的の生成物を得るために使用する反応触媒に，グリーンケミストリーの観点から，環境適応型の酵素を用いることが世界の風潮となってきている．酵素に代表される生体触媒は，人工的な化学触媒に比べて環境にやさしく，化学反応をより効率的に進めることから，化学触媒の代わりに生体触媒を用いて合成を行う取り組みが増加している．実際，目的の化合物を生成するために従来では 10 ステップの合成を行っていたものを，生体触媒を取り入れることで 3 ステップまで短縮したという研究事例もある [2]．これらのことから，目的物生成のために酵素反応を取り入れたうえで，反応設計を行うこと，あるいは，特定の反応に対して生体触媒として最適な酵素を予測することも重要な要素の一つとなってきている．

情報科学の観点からとらえると，酵素を触媒として取り入れる際，反応物 (基質) に対して特定の生体酵素を加えれば目的の生成物が得られる．つまり，基質と生成物が決まった場合，それに対して最適な酵素を予測するというのは容易に見えるかもしれない．ところが，実際には基質特異性と呼ばれる，酵素が基質に対して高い反応性を示すかどうかという酵素の特性によって，問題が複雑になる．有機合成化学を研究していて，酵素に関する知識を持ち合わせていたり，経験が豊富であれば，どの酵素が使えそうかある程度予測ができるかもしれない．しかし，先ほど述べた基質特異性に加えて，酵素のタンパク質配列を参照したりと，遺伝子分野にかかわる部分もあり，有機合成の知識だけでは解決が難しい場合がある．

§ 1.2　本研究の目的

生体触媒（Biocatalyst）を用いた有機合成化学において，目的とする生成物を効率よく得るために，酵素のデータベースを参照したり，酵素の研究を行っている専門家と協力するなどして，最適な酵素候補の目途をつけるという手法が取られる場合がある．実際は，酵素にも同様の性質を持っていたり，複数の企業製品が存在していたりと，触媒候補が複数存在する場合があるため，スクリーニングなどの実験の試行錯誤を繰り返しながら，最終的に 1 つに絞られていく．ここで，酵素の候補を探索したり，新たな酵素を設計する際に，有機合成化学の研究者自身で，酵素候補を探索することができれば，次の実験のステップまでスムーズに進めることができると考えられる．つまり，目的生成物を得るために，最適な酵素を迅速に予測・設計してくれるようなツールが存在すればよい．

本研究では，反応式を与えた際，その反応を触媒するのに必要な酵素を予測するシステムを考える．前述のとおり，1 つの酵素に絞り込むためには，様々な条件が絡む実験を必要とするため，おおまかな予測という形になる．しかし，有機合成化学の知識内で手順を進めていくことが可能となるため，十分に有効性があると考えられる．

酵素は酵素番号 (Enzyme Commission numbers: EC 番号) とよばれる，4 組の数字の組み合わせからなる番号が割り振られており，どの反応を触媒し，どの結合・基質に反応するかによって分類されている [11] [12]．与えられた反応に対して，酵素 (EC 番号) を予測できれば，その EC 番号の酵素から何を選択するかという次のステップに進むことができる．

EC 番号の情報の中には，反応物から生成物への，その酵素を使った代表的な反応が記載されている．そこで，本研究では，酵素を予測するターゲットとなる反応式内の反応物から生成物，また，EC 番号の代表的な反応式内の反応物から生成物，それぞれの物理・化学的特性値の変化を比較し，類似性が最も高い反応の酵素番号を提示して，最適な酵素を予測する．

主な流れとして，化学・酵素データベースから酵素の EC 番号および，代表的な反応式の情報を取得し，EC 番号と反応式の対応表を作成する．次に各反応式を，反応物と生成物に分解する．ターゲットの反応式も同様に分解し，各化合物の構造をコンピュータ上で扱うための表現に変換する．その後，複数の化合物の物理・化学特性値を計算し，各反応式において反応物から生成物への特性値の変化量を求める．この複数の特性値変化量を要素にもつ多次元ベクトルを，反応式の特徴ベクトルとして表現する．最終的に特徴ベクトルの次元削減を行い，クラスタリングによって反応式の特徴ベクトルを 2 次元平面上に出力する．得られた結果から，ターゲットの反応式に対して，最も近い場所に位置する，反応式の EC 番号に登録されている酵素を，用いるべき最適な酵素として予測する．

§ 1.3　本論文の概要

本論文は次のように構成される．

第 1 章 本研究の背景と目的について説明した．背景では，ケモインフォマティクスの概要，有機合成において，生体触媒を用いることのメリットとその課題について述べた．目的では，目的の生成物を得る際に用いる，最適な酵素を予測するための，EC 番号を予測するシステムの概要について述べた．

第 2 章 有機合成，ケモインフォマティクス，および酵素の概要を述べる．また，本研究で用いるデータベースについて述べる．

第 3 章 化学データベースからの情報抽出，ケモインフォマティクスでにおける化合物の構造表現法，EC 番号予測の概要を述べる．また，クラスタリング手法について述べる．

第 4 章 提案手法についての説明，および手順について説明する．

第 5 章 提案手法による数値実験の概要，実験結果と考察を述べる．

第 6 章 まとめと今後の課題について述べる．

第2章

制約を考慮したPSO

§ 2.1　勾配系を考慮したPSO

群知能（Swarm Intelligence）は、鳥や魚、アリのコロニーなどの自然界の群れの行動に基づく最適化手法である。粒子群最適化（Particle Swarm Optimization: PSO）はその一つであり、ケネディによって社会的行動に基づいて開発された並列進化計算技術である [?] [?]。PSO は、群れの中の各粒子の最良の情報（pbest）と集団全体の最適値（gbest）を用いて、探索を行う確率的最適化手法であり、その収束特性には理論的根拠が不足している。

本研究では、PSO と勾配法を組み合わせたハイブリッド動的システムを提案し、より良い最適解を求めるための新しいアプローチを提示する。PSO に勾配法を組み込むことで、連続 PSO アルゴリズムの精度を向上させ、グローバルな情報に基づいた補間探索を実現する。

PSO の基本的な更新式は以下の通りである：

$$x_{k+1}^d = x_k^d + v_{k+1}^d \tag{2.1}$$

$$v_{k+1}^d = wv_k^d + c_1 r_1 (x_k^d - x_k^d) + c_2 r_2 (x_k^d - x_k^d) \tag{2.2}$$

ここで、x_k^d は粒子の位置、v_k^d は速度、w は慣性係数、c_1 と c_2 は学習係数、r_1 と r_2 はランダムな数値である。PSO の速度ベクトルは、pbest に向かうベクトル、gbest に向かうベクトル、そして過去の進行方向に基づくベクトルの合成によって決定される。

連続型 PSO（Continuous Particle Swarm Optimization: CPSO）アルゴリズムは、以下の更新式で表される：

$$\dot{X} = V \tag{2.3}$$

$$\dot{V} = -\beta V + \phi(X_{db} - X) + \gamma(X_{gb}^T - X) \tag{2.4}$$

ここで、ϕ はスケーリングパラメータ、β は減衰係数、γ は調整係数である。CPSO アルゴリズムは、PSO の確率的な動作を線形システムとして近似し、安定性の分析を行うことができる。

さらに、本研究では PSO と勾配情報を組み合わせた勾配 PSO を提案する。勾配 PSO では、以下の更新式が使用される：

$$\dot{X} = V \tag{2.5}$$

$$\dot{V} = -\beta V + Z \tag{2.6}$$

$$Z = \phi(X_{db} - X) + \gamma(X_{gb}^T - X) + \delta(X - X_{NN}) \tag{2.7}$$

ここで、X_{NN} はニューラルネットワークダイナミクスに由来する補正項であり、勾配情報を制御する役割を果たす。勾配 PSO は、PSO のグローバル探索能力とニューラルネットワークの局所的な最急降下法を組み合わせることで、より精密な探索を実現する。

§ 2.2　制約がある場合の PSO

制約条件付き PSO に関する研究では、連続時間 PSO アルゴリズムに制約条件を追加する方法が考察されている。連続時間系モデルにおいて、PSO の更新式は次のように表される：

$$\frac{d^2x_p(t)}{dt^2} + a\frac{dx_p(t)}{dt} = c\left[F_p(x_p(t),t) + C(x_p(t),t)\right] \tag{2.8}$$

ここで、F_p と C はそれぞれ粒子の位置と目標値に関する関数であり、制約条件は以下のようにモデル化される：

$$x_i = f_i(y_i) = \frac{q_i + p_i\exp(-y_i)}{1+\exp(-y_i)} \tag{2.9}$$

制約条件付き最適化問題を解決するために、変数変換モデルを導入し、連続時間モデルを離散化することでプログラムに実装可能な形に変換する。具体的には、以下の離散化式を用いる：

$$u_p(k+1) = (1-a\Delta T)u_p(k) + \Delta T v_p(k) \tag{2.10}$$

$$v_p(k+1) = v_p(k) + c\Delta T\left[F_p(u_p(k),k) + C(u_p(k),k) - \nabla E(u_p(k),k)\right] \tag{2.11}$$

ここで、ΔT はサンプリング時間、F_p と C は制約条件を考慮した関数であり、∇E は目的関数の勾配である。制約条件付きの PSO モデルは、特に多峰性問題よりも単峰性問題において効果的であり、計算コストを削減しつつ高精度な最適化を実現する。

実験結果として、Griewank 関数や Booth 関数を用いた比較では、提案手法が多峰性問題には適していない一方で、単峰性問題には有効であることが示された。特に、Booth 関数においては粒子数を減少させた場合でも良好な結果が得られ、制約条件付きの最適化も正しく行われることが確認された。

第3章

ケモインフォマティクスと情報技術

§ 3.1　化学データベースからの情報抽出

Web サイト等から収集した大量の情報の中から，自然言語処理を用いて有用な情報を抽出するテキストマイニングにおいては，スクレイピングが用いられることがある．スクレイピングとは，Web サイトから文章をプログラミングによって自動取得する方法であり，効率的にデータを収集できる．一方でデータベースを管理している Web サイト等においては，独自のアプリケーション・プログラミング・インターフェース (Application Programming Interface: API) を備えている場合があり，指定された形式でプログラムを記述すれば，データベース上の情報を自動的に取得することができる．

化学データベースにも公式の API が公開されているものがいくつか存在する．KEGG では KEGG API [21], PubChem では POWER USER GATEWAY(PUG) [18] と呼ばれる API が公開されており，本節ではこの 2 つの API について説明する．

KEGG API の構成

KEGG API のフォーマットは以下のようになる [21]．<operation>の部分に上記の 7 つのいずれかを指定する，例えば，「list」を指定した場合．以下のフォーマットに従う．<dbentries>で目的のデータがある KEGG データベース名を指定する．例えば，「pathway」を指定することで，完成するリンク先へ行くと，各 Pathway のマップ番号と，Pathway 名の対応リストを取得できる．図 3.1 に番号と Pathway 名の対応表を示す．このように，「http://rest.kegg.jp/」以下の部分で指定された識別子を設定することで，データが保存されている URL に移動することができ，各プログラム言語で実装されている，リンク先の中身を取得するコードによって，必要なデータを取得することができる．

PUG

Common Gateway Interface(CGI) を経由して，PubChem のデータをプログラミングによって，取得する機能を提供するシステム [22]．データのやり取りは URL ではなく XML を用いる．XML によるリクエストを CGI へ送り，リクエストの内容が実行された後，結

```
http://rest.kegg.jp/<operation>/<argument>[/<argument2[/<argument3> ...]]
<operation> = info | list | find | get | conv | link | ddi
```

図 3.1: KEGG API の URL 構成 1

```
http://rest.kegg.jp/list/<dbentries>

<dbentries> = Entries of the following <database>
<database> = pathway | brite | module | ko | genome | <org> | vg | vp | ag |
    compound | glycan | reaction | rclass | enzyme | network | variant |
    disease | drug | dgroup | <medicus>
```

図 3.2: KEGG API の URL 構成 2

表 3.1: リンク先の対応表

path:map00010	Glycolysis / Gluconeogenesis
path:map00020	Citrate cycle (TCA cycle)
path:map00030	Pentose phosphate pathway
path:map00040	Pentose and glucuronate interconversions
path:map00051	Fructose and mannose metabolism
path:map00052	Galactose metabolism
path:map00053	Ascorbate and aldarate metabolism
path:map00061	Fatty acid biosynthesis
path:map00062	Fatty acid elongation
path:map00071	Fatty acid degradation
path:map00073	Cutin, suberine and wax biosynthesis
path:map00100	Steroid biosynthesis
path:map00120	Primary bile acid biosynthesis
path:map00121	Secondary bile acid biosynthesis
path:map00130	Ubiquinone and other terpenoid-quinone biosynthesis
path:map00140	Steroid hormone biosynthesis
path:map00190	Oxidative phosphorylation
path:map00195	Photosynthesis
path:map00196	Photosynthesis - antenna proteins
path:map00220	Arginine biosynthesis
path:map00230	Purine metabolism
path:map00232	Caffeine metabolism
path:map00240	Pyrimidine metabolism
path:map00250	Alanine, aspartate and glutamate metabolism
path:map00253	Tetracycline biosynthesis
path:map00254	Aflatoxin biosynthesis
path:map00260	Glycine, serine and threonine metabolism
path:map00261	Monobactam biosynthesis
path:map00270	Cysteine and methionine metabolism
path:map00280	Valine, leucine and isoleucine degradation

果が XML で返信される仕組みとなっている．例として，CID1 と CID99 の化合物の構造を SDF ファイル形式の gzip 圧縮でダウンロードする場合，図 3.4 のような XML 構造のリクエスト応答となる．PubChem ではアクセス簡略化のため，PUG-SOAP と PUG-REST というシステムが実装されている．本研究では PUG-REST を用いるため，PUG-REST について説明する．

PUG-REST

PUG や PUG-SOAP で用いられている XML 形式の記述を必要とせず, 簡単な記述方でデータを取得することができる API．PUG-REST のリクエストは以下のような URL で表記される [18]．<input specification>はさらに<domain>/<namespace>/<identifiers>で構成されており, 何のデータを取ってくるのかを定める．<domain>では, substance, compound, assay などの対象とするデータベースを指定する．また, <namespace>では CID(cid) や化合物名 (name), 分子式 (formura) 等を指定し, <identifiers>では, CID の番号, 化合物名・分子式の文字列といった, <namespace>に対する具体的な名前を指定する．<operation specification>では<input specification>で指定したデータ保管場所にアクセスした際, どのような操作を所望しているのかを記述する．例えば, <input specification>で CID 番号の情報を記述している状態で, synonyms を指定するとその CID 番号の化合物名に対する同義語のリストが返される．同様のケースで，<compound property>で property/XXX,YYY,$\cdots$,ZZZ/ を指定すると，その化合物の物性値や化学的特性値を複数取得することができる．<output

```
https://pubchem.ncbi.nlm.nih.gov/rest/pug/<input specification>/
<operation specification>/[<output specification>][?<operation_options>]

<input specification> = <domain>/<namespace>/<identifiers>
<operation specification> = record | <compound property> | synonyms | sids |
    cids | aids | assaysummary | classification | <xrefs> | description |
    conformers
<output specification> = XML | ASNT | ASNB | JSON | JSONP [ ?callback=<
    callback name> ] | SDF | CSV | PNG | TXT
```

図 3.3: PUG-REST のリクエスト

図 3.4: PUG における XML 応答の例 [18]

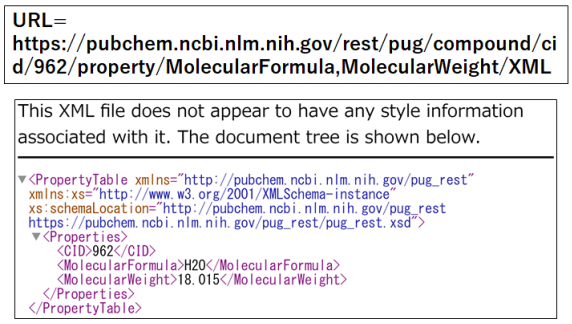


図 3.5: 水 (H_2O) の情報を取得するリクエスト URL とデータ取得結果

specification>の部分では取得したいデータをどのような形式で出力するかを指定する．基本的には，<input specification>/<operation specification>/<output specification>の部分を指定すれば良く，例として，水 (CID968) の分子式 (MolecularFormula) と分子量 (MolecularWeight) を XML で取得した場合を図 3.5 に示す．

§ 3.2　化合物の構造表現法と EC 番号予測手法

化合物同士の構造比較について述べる前に，ケモインフォマティクスで一般的に使われている化合物の構造表現について説明する．

MOL ファイル

化合物の構造情報を記したテキスト形式のファイル．「.mol」の拡張子で保存されることが多い．ファイル内には結合している原子と各原子の 3 次元座標リストやどの原子同士が結びついているかのリストが記述されている．通常の構造式と mol ファイルを比較したものを図 3.6 に示す

SDF ファイル

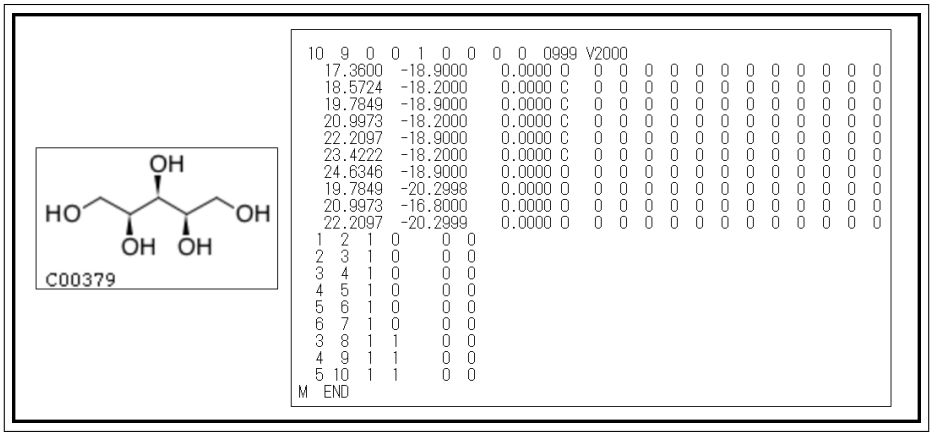


図 3.6: KEGG COMPOUND で取得できる構造式と MOL ファイルの例

MOL ファイルを複数統合した拡張子「.sdf」のファイル．2 つ以上の分子の MOL ファイルをデータベースから同時に入手する際は，この形式となることが多い．

SMILES

化合物の構造を文字列で表したもの．以下の規則に従って文字列に変換していく [23].

1. 原子は元素記号で表し，2 文字で区別がつきにくい原子 (Nb と NB 等) は [ ] で囲む
2. 水素原子は省略する
3. 隣接する原子は隣に記す
4. 二重結合は =，三重結合は#で表し，単結合・芳香族結合は省略する (芳香族原子は小文字の c などで表記する)
5. イオンなどで結合がない部分は「.」で分ける
6. 構造が分岐する箇所は ( ) で表記する
7. 環構造は切断して切断箇所を記すとともに (C1 など)，鎖錠構造で表す．

以上の規則に基づいて作成されたものを generic SMILES と呼ぶが，以下の規則を加えたものを isomeric SMILES と呼ぶ

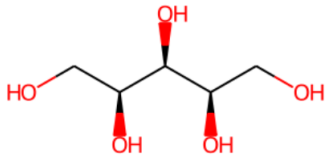
1. 同位体 (例えば炭素) がある場合 [13C] という表記にする
2. 立体異性体を区別するための絶対配置を「@」または「@@」で表現する
3. 二重結合などで生じる幾何異性を「/」と「¥」で表す

フィンガープリント

化合物の構造や特徴をビット列で表現したもの．構造のどの部分に注目するか，または性質などで種類がある．例えば，MACCS Keys のフィンガープリント [24] では，166 種の特徴的な構造を化合物が持っているかどうかを 0 と 1 で表現している．フィンガープリントは主に化合物同士の類似性比較で用いられる．

化合物の数値化 (特徴ベクトル)

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




```
Chem.MolToSmiles(Xylitol)
```

```
'OC[C@H](O)[C@@H](O)[C@H](O)CO'
```

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

```
152.14600000000002
```

図 3.7: rdkit を用いた化合物の情報

```
import pubchempy
pubchempy.get_properties([ 'MolecularFormula','MolecularWeight',
                           'IsomericSmiles' ], 'glucose',
                          'name', as_dataframe=True)
```

CID	MolecularFormula	MolecularWeight	IsomericSMILES
5793	C6H12O6	180.16	C([C@@H]1[C@H]([C@@H]([C@H](C(O1)O)O)O)O)O

図 3.8: PubChempy でグルコースの情報を取得した結果

機械学習で様々な予測を行うためには，化合物を数値化して表現する必要がある．その方法として，前述のフィンガープリントではビット列で化合物を数値化しているが，物性値を特徴として用いられることも多い．一般的に複数の物性値が用いられ，多次元の特徴ベクトルとして化合物の特徴を表現する．これらの構造情報や物性値で化合物の特徴を表したものは記述子と呼ばれている．

RDKit [25] を用いた化合物のデータ化

RDKit は Python 提供されている，化合物の構造を扱うライブラリである．SDF ファイルや MOL ファイルを読み込んで構造式の画像を出力したり，SMILES やフィンガープリントに変換することができる．RDKit では読み込んだ構造式から，化合物の記述子を計算することができるため，化合物同士の類似性を評価したり，機械学習に発展させることができる．例として，化合物の分子量を意味する MolWt を知りたい場合，MOL ファイルから読み込んだ化合物のインスタンスを生成し，RDKit の Descriptors クラスにある MolWt メソッドに生成したインスタンスを渡すことで，MolWt が計算され出力される．図 3.7 に，rdkit を用いて化合物の構造式と SMILES を出力した様子，および化合物の MolWt を計算した結果を示す．

PubChemPy [26]

PUG REST を用いて PubChem のデータを取得するための Python ライブラリ．化合物名や CID を引数にして，対象化合物の物性値や SMILES を取得することができる．例として，グルコースの分子式，分子量，IsomericSMILES を取得した結果を図 3.8 に示す．

EC 番号予測手法

上記で紹介した構造表現法や化学・酵素データベースを用いて，酵素反応の予測や分類を行う研究が多く行われている．2.2 で示した通り，酵素は EC 番号によって管理されているが，同時にその酵素を用いた代表的な化学反応が反応式として登録されている．ここで，代表的な化学反応とは，生体内など自然界で起こる反応を指し，各 EC 番号に 1 つまたは複数登録されている．例として KEGG ENZYME の EC3.1.1.2 では代表的な反応式 3 種が

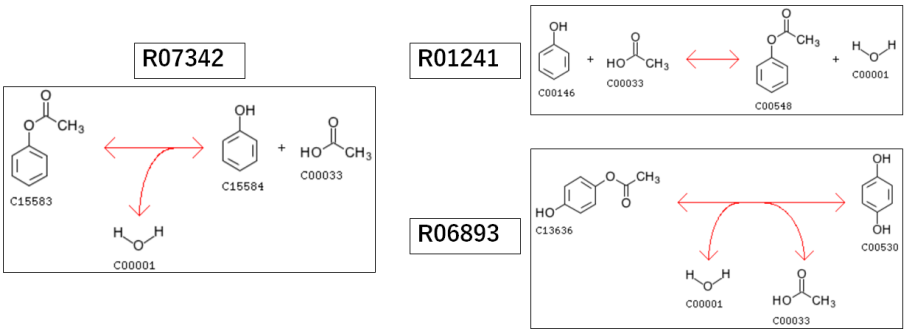


図 3.9: EC3.1.1.2 の代表的な反応式

R 番号として表記されており，図 3.9 のような化学反応となる．これらの反応式に対して，EC 番号の分類問題を考え，より多くの反応式を正しい EC 番号に分類できるように分類モデルを検討する研究が主に行われている．以下で 2 つの手法を示す．

EC 番号予測手法の 1 つとして，アミノ酸配列の類似性を用いるものがある．酵素はタンパク質であるため，アミノ酸配列で表現される．アミノ酸配列の類似性に基づいて，該当する EC 番号を予測する．もう 1 つの手法として，基質と生成物の構造に着目したものがある．構造をフィンガープリントなどで表したものや [27]，構造として特徴的な部分の化学変化に注目したもの [28] などがある．

フィンガープリントを用いる手法では，各基質と生成物を分子の部分構造 (フラグメント) に着目したフィンガープリントで表している．その後，基質フィンガープリントから生成物フィンガープリントを引いた反応差分フィンガープリントを定義する．そして，EC 番号が正解ラベルとして与えられている反応差分フィンガープリントとのユークリッド距離を求め，最小距離となるものの EC 番号を割り当てるという方法を用いている．例えば，KEGG REACTION の R00005 に登録されている反応式 $C01010+C00001 <=> 2C00011+2C00014$ に対して，各分子の分子フィンガープリントを MFP として，反応差分フィンガープリント RFP を以下のように定義している．

$$RFP_{R00005} = MFP_{C01010+C00001} - MFP_{2C00011+2C00014} \tag{3.1}$$

構造の特徴的な部分の化学変化を用いる手法では，RDM パターンと呼ばれる，基質と生成物の各構造に対して，反応中心原子 (R atom)，その近傍の原子で異なっている領域 (D atom) と一致している領域 (M atom) を定義している．EC 番号の基質と生成物の RDM パターンと，入力した反応の RDM パターンの類似性を比較することで，入力反応の EC 番号を予測している．

§ 3.3　クラスタリング手法

本研究では 2 つのクラスタリング手法を用いるが，それに伴いここではクラスタリングについて述べる．クラスタリングは観測されたデータのみを扱う教師なし学習の一つで，特定の基準に従い類似しているデータどうしでクラスタを形成し，分類する手法である．データが 1 つのクラスタのみに属するクラスタリングをハードクラスタリングと呼ばれており，

種類によっては，複数のクラスタに属することを許容するソフトクラスタリングも存在する．クラスタリングは主に階層的クラスタリングと非階層的クラスタリングに分けられる．階層的クラスタリングではさらに，分割型のものと凝集型のものに分けられる．分割型ではデータを全て 1 つのクラスタとみなしたのち，細かいクラスタに分割していく手法である．凝集型では，データそれぞれを 1 つのクラスタとみなし，特定の基準にしたがって複数データが属するクラスタを形成する．複数データを持つクラスタ同士も連結され，新たなクラスタを形成し，指定したクラスタ数になるまで繰り返される．

階層型では凝集型が主に用いられ，以下では，凝集型におけるクラスタを形成していく基準について述べる．なお，クラスタ C_1，C_2 に属するデータの集合をそれぞれ $\mathbf{x_1}$,$\mathbf{x_2}$，$\mathbf{x_1}$ と $\mathbf{x_2}$ の距離を $d(\mathbf{x_1}, \mathbf{x_2})$ としたときのクラスタ間の距離を $d(C_1, C_2)$ とする [29].

最短距離法

2 つのクラスタ内のデータどうしで，最も距離が近い組を基準として，新しきクラスターを作成する．計算量は少ないが，外れ値に弱いとされている．

$$d(C_1, C_2) = \min_{\mathbf{x_1} \in C_1, \mathbf{x_2} \in C_2} \{d(\mathbf{x_1}, \mathbf{x_2})\} \tag{3.2}$$

最長距離法

最短距離法に対して，最も距離が遠い組を基準としたもの，外れ値には弱いが，クラスタサイズが一定になる傾向がある．

$$d(C_1, C_2) = \max_{\mathbf{x_1} \in C_1, \mathbf{x_2} \in C_2} \{d(\mathbf{x_1}, \mathbf{x_2})\} \tag{3.3}$$

群平均法

2 つのクラスタ内の要素同士の距離を合計し，各クラスタサイズで割った平均を基準としたもの．外れ値の影響が少なく．クラスタが帯状に並ぶ鎖効果が起こりにくいとされている．

$$d(C_1, C_2) = \frac{1}{|C_1||C_2|} \sum_{x_1 \in C_1} \sum_{x_2 \in C_2} d(x_1, x_2) \tag{3.4}$$

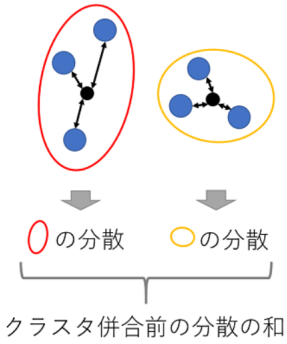
ウォード法

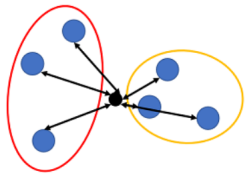
あらかじめ 2 つのクラスタを結合し，結合したクラスタ内の重心に対する，データの分散 $E(C_1 \cup C_2)$ に対して，結合前の各クラスタ内のデータの分散 $E(C_i)$ を引いた差が，最小となるクラスタのペアを結合する方法．計算量は多くなるものの，分類感度が良いとされ，階層的クラスタリングで最も用いられている．ウォード方のイメージを図 3.10 に示す．クラスタ C_1 の重心を $\boldsymbol{c_i}$ として，以下のように表される．

$$\boldsymbol{c_i} = \sum_{\boldsymbol{x} \in C_i} \frac{\boldsymbol{x}}{|C_i|} \tag{3.5}$$

$$E(C_i) = \sum_{\boldsymbol{x} \in C_i} d(\boldsymbol{x}, \boldsymbol{c_i})^2 \tag{3.6}$$

$$d(C_1, C_2) = E(C_1 \cup C_2) - E(C_1) - E(C_2) \tag{3.7}$$





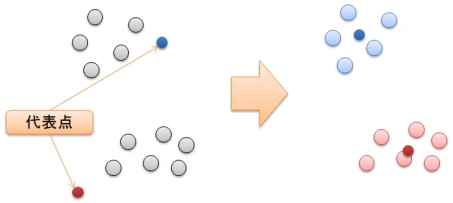


図 3.11: k-means 法のイメージ [31]

図 3.10: ウォード法のイメージ [30]

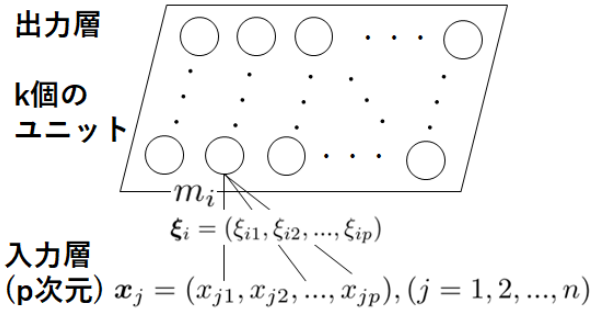


図 3.12: SOM における入力データのマッピング

非階層的クラスタリング

非階層的クラスタリングでは，あらかじめクラスタリング数を決めておき，各手法で定められている基準にしたがって，データを分類する．非階層的クラスタリングの手法をいくつか以下に示す．

k-means 法

データに対して，ランダムにクラスタを割り振り，重心に基づいてクラスタを再構成していく手法．以下の手順に沿ってクラスタリングを行う．

1. 最初に指定した k 個のクラスタリングに，データ点をランダムに割り振る
2. 各クラスタ内の各データに対して重心を計算し，データ点が最短距離にある重心のクラスタに属するように，データ点へのクラスタを振り直す．
3. 振り直しで全てのデータ点のクラスタが固定されるまで，上記の手順を繰り返す．

自己組織化マップ (Self-Organizing Map: SOM) [32]

多次元データを低次元にマッピングし，可視化するクラスタリング手法．以下そのアルゴリズムを示す [33]．n 個の p 次元観測ベクトル $\boldsymbol{x}_j = (x_{j1}, x_{j2}, ..., x_{jp}), (j = 1, 2, ..., n)$ を，ユニット $m_i (i = 1, 2, ..., k)$ で構成された，2 次元平面上に写像する．図 3.12 にそのイメージを示す．このとき各ユニットの重心を $\boldsymbol{r}_i = (r_{i1}, r_{i2})$ とし，これを m_i の位置ベクトルとする．さらに，各ユニットは，重みベクトル $\boldsymbol{\xi}_i = (\xi_{i1}, \xi_{i2}, ..., \xi_{ip}), (i = 1, 2, ..., k)$ を持っているとする．ここで，$\boldsymbol{x}_j$，m_i をそれぞれ，入力層，出力層と呼び，次の手順によって出力層を更新する．(ただし，$\boldsymbol{\xi}_i$ はランダムな値で初期化を行う)

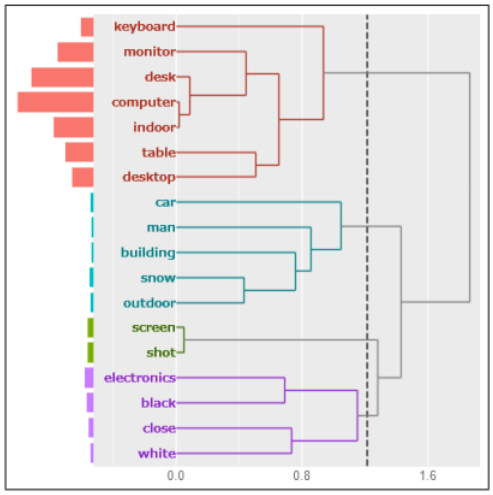


図 3.13: 階層的クラスタリングによる行動識別

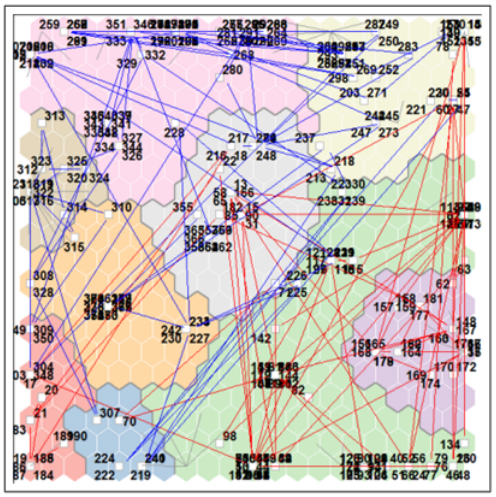


図 3.14: SOM を用いた行動時系列分析

1. $j = 1$ から n までの順に，各 $\boldsymbol{x}_j$ に対してユークリッド距離 $||\boldsymbol{x}_j - \boldsymbol{\xi}_i||$ を求める．
2. $||\boldsymbol{x}_j - \boldsymbol{\xi}_i||$ を最小値にする $\boldsymbol{\xi}_i$ を $\boldsymbol{\xi}_c$ と置く．この $\boldsymbol{\xi}_c$ を持つユニットを勝者ユニット m_c と呼び，勝者ユニット m_c とその近傍のユニットが持つ重みベクトルを次のように更新する．

$$\begin{cases} \boldsymbol{\xi}_i = \boldsymbol{\xi}_i + h(t)\{\boldsymbol{x}_j - \boldsymbol{\xi}_i\} & i \in N_c \\ \boldsymbol{\xi}_i = \boldsymbol{\xi}_i & i \notin N_c \end{cases} \tag{3.8}$$

N_c は m_c の近傍領域を表し，m_c と N_c に含まれる m_i が $\boldsymbol{x}_j$ に近くなるように更新される．また，$h(t)$ は以下で定義される近傍関数であり，m_c が最も $\boldsymbol{x}_j$ に近づくように働きかける．ただし，$\alpha(t)$ を学習率係数 (学習回数 t の増加とともに減少)，$\sigma^2(t)$ は N_c の散らばりに関する調整関数とする．

$$h(t) = \alpha(t) \exp\left[\frac{-||\boldsymbol{r}_c - \boldsymbol{r}_i||}{2\sigma^2(t)}\right] \tag{3.9}$$

3. j で更新した $\boldsymbol{\xi}_i$ を記憶した状態で，$j + 1$ として 1,2 を繰り返す．
4. 3 までを 1 回の学習とし，指定した回数まで学習を行う
5. 学習後，ユークリッド距離 $min||\boldsymbol{x}_j - \boldsymbol{\xi}_i||$ を満たす $\boldsymbol{\xi}_c$ を持つ勝者ユニット m_c に $\boldsymbol{x}_j$ をマッピングする

クラスタリングを用いた研究例として，ヒトの行動パターンを解析し，行動識別を行ったものがある [34]．まず，画像認識 API を用いて，視界に映っている物体を認識し，その物体名をテキストデータに出力している．次に，テキストマイニングデータのクラスタリングを行うソフトウェア KH Corder [35] を用いて，物体名の同時出現頻度に関して，階層的クラスタリングを行っている．それによって，クラスタ内に含まれる物体名から行動全体のイベント性を分析している．また，SOM によるクラスタリングも行われている．観測されたデータから順番に SOM の 2 次元マップ上にプロットしていき，プロット点を線で結んでいくことで，行動の時系列を作り，複数の測定における行動の類似性を分析している．階層的クラスタリングと SOM を用いた行動分析の様子を図 3.13 および図 3.14 に示す．

第4章

提案手法

§ 4.1　特性値変化量を用いたEC番号予測

医薬品などの新規化合物を開発する分野において，それに必要な有機合成を効率的かつ，なるべく環境に負荷を与えない形で行えるほうが望ましい．その点，生体触媒の酵素を用いると，反応物の特定の部位だけの選択的合成，反応の効率化など，グリーンケミストリーの優れた反応となるため，酵素を用いる機会が増加している．それに伴い特定の反応を行うために最適な酵素を選択することも重要となってきている．一方で，基質特異性などの酵素の性質は，生物分野に関わる内容であるため，有機合成の知識のみでは解決が難しい．酵素研究の専門家と協力して，または，酵素データベースなどを参照して最適な酵素候補の目途をつけ，その後のスクリーニングなどで，1つの酵素に絞っていく．ここで，目的とする反応情報を与えた際に，酵素候補を予測するシステムがあれば，酵素候補の探索にかかる時間を著しく短縮することができ，次のスクリーニングの段階まで研究をスムーズに進めることができる．その様子を図4.1に示す．また，酵素はEC番号で分類されており，EC番号には生物の体内で起こる酵素を用いた代表的な反応が反応式として記載されている．EC番号の酵素を用いた様々な生物由来の酵素製品が開発されているが，EC番号を予測することで，スクリーニング候補として，そのEC番号内の酵素に絞り込むことができる．

そこで，本研究では，ターゲットとなる反応式を与えた際に最適なEC番号を予測する．すなわち，EC番号内の多数の酵素を生体触媒として最適な酵素候補として提示する．予測の方法として，対象とする反応式 (ターゲット反応式) とEC番号の代表的な反応式 (EC反応式) の構造変化を比較する．図4.2に比較のイメージを示す．ターゲット反応式で目的とする生成物は，逆合成的な思考でどの反応物を用いれば得られるのか分かっている．ここで，ターゲット反応式における反応物から生成物への構造変化が，EC反応式における反応物から生成物の構造変化に類似しているならば，EC反応式で用いられている酵素をターゲットで使用することで，反応の効率が上がり，高い収率でターゲット生成物が得られるという仮定を置く．これは化学の分野で用いられている類似性の概念 [36] に基づいている．

反応による構造変化を，反応式の類似性の評価指標とした理由として，ターゲット反応式の化合物と，各EC反応式の化合物どうしの比較では反応式の類似性を正確に評価できないためである．例えば，ターゲット反応式の反応物が，あるEC反応式の反応物に最も類似していると評価されたとしても，生成物は異なるEC反応式の生成物に最も類似されていると，評価される可能性がある．また，反応物や生成物は複数ある場合が多いため，より反応式の類似性を評価することが難しくなる．そのような理由から反応による構造変化を類似性の比較として用いる．

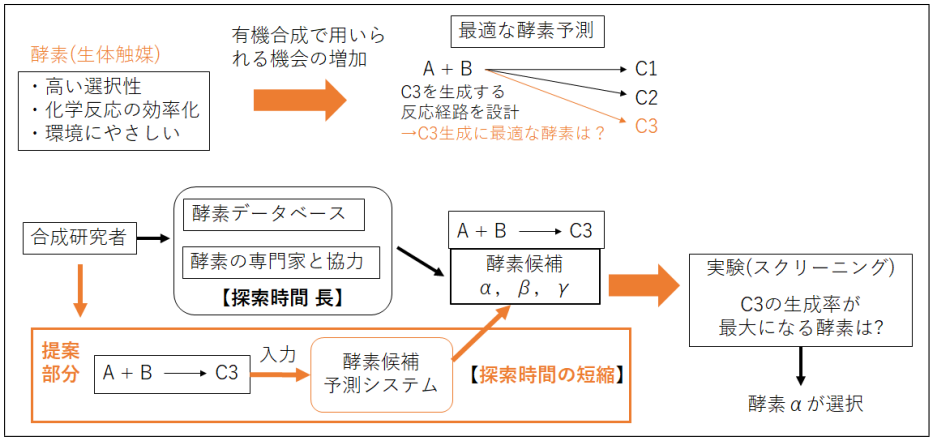


図 4.1: 従来の酵素探索と提案する酵素探索の比較

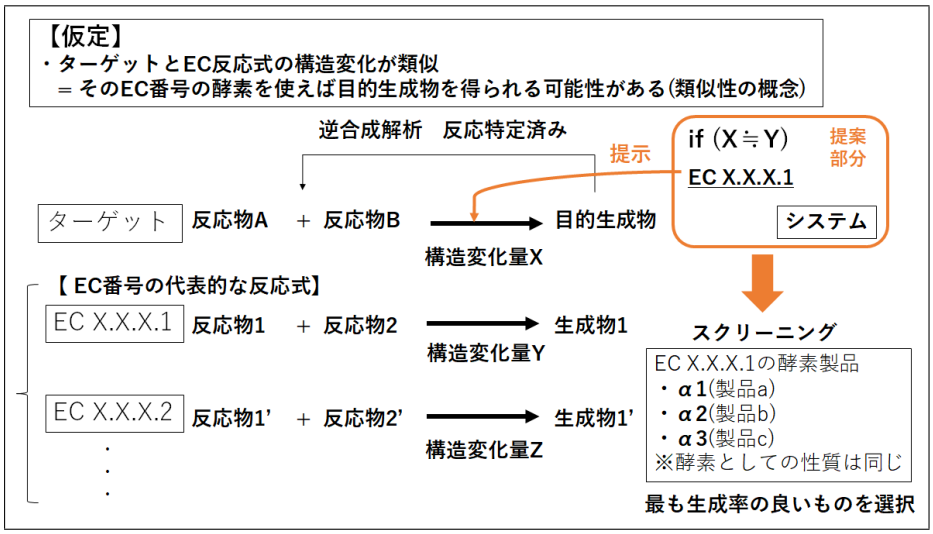


図 4.2: 反応式の類似性比較

反応物から生成物への構造変化を捉える特徴として，記述子を用いた物性値・化学特性値の変化量 (特性値変化量) を用いる．従来研究では，反応物から生成物に変化する際の，構造記述子の変化が，反応変化の特徴として用いられている [27]．ここでは，反応物の生成物の部分構造に注目したフィンガープリントを求め，その差分を比較することで，類似する EC 反応式の酵素を予測している．しかし，フィンガープリントには様々な種類があり，それぞれ化合物のどのような特徴を説明しているのかが異なっている．つまり，1 つのフィンガープリントでは反応変化の特徴を全てとらえるのは難しい．一方で，物理・化学的な特性値を表す記述子も，化合の構造を表現する指標として考えられる．RDKit では 208 種類の特性値に関する記述子が実装されており，読み込んだ分子構造式から簡単に特性値を計算できる．そのため，RDKit 記述子を多数用いて，多次元の特性値変化量を要素に持つ特徴ベクトルを求めることで，ターゲット反応式と EC 反応式の反応時の構造変化を表現する．

差分フィンガープリントと同様に特性値変化量を以下のように定義する．各反応の反応物と生成物の個数をそれぞれ 2 個としたとき，反応物 i の特性値を RT_i，生成物 i の特性値を PD_i とする．このとき，各 n 種の記述子に対する特性値変化量 $cv_j(j = 1, 2, ..., n)$ およ

表 4.1: 各反応式に対する記述子ごとの特性値

	記述子 1	記述子 2	$\cdots$	記述子 n
$\boldsymbol{DF}_1$	cv_{11}	cv_{12}	$\cdots$	cv_{1n}
$\boldsymbol{DF}_2$	cv_{21}	cv_{22}	$\cdots$	cv_{2n}
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
$\boldsymbol{DF}_m$	cv_{m1}	cv_{m2}	$\cdots$	cv_{mn}

び，m 個の反応式の特徴ベクトル $\boldsymbol{DF}_i(i = 1, 2, ..., m)$ を以下のように表す．

$$cv_j = (PD_1 + PD_2) - (RT_1 + RT_2) \tag{4.1}$$

$$\boldsymbol{DF}_i = (cv_{i1}, cv_{i2}, ...cv_{ij}, ..., cv_{in}) \tag{4.2}$$

これらをもとに，表 4.1 のような $i \times j$ の各反応式の特性値表を作成する．行ラベルはターゲット反応式 T と EC 番号，列ラベルは記述子名となる．

特徴ベクトル比較のために必要となる化合物などのデータは，KEGG と PubChem から収集する．これらのデータベースを用いる理由として 2 つ挙げられる．1 つ目は，API でデータを取得するフォーマットが整っていることである．API によって必要となるデータを簡単に取得できることは，プログラミングで自動収集するシステムの，開発のしやすさにつながり，効率的なデータ収集を行える．2 つ目はリンクによってデータベースどうしの行き来がしやすい点にある．異なるデータベースへの参照リンクが多いほど，多種多様なデータを収集をしたり，1 つのデータベース内では見られないデータ間の関係を得ることができる．必要となるデータを API で取得し，集めたデータ関係を分析する，または，新たなデータ関係を見出すデータベースを構築することも可能となる．

KEGG では図 4.3 のように，KEGG ENZYME，KEGG REACTION，KEGG COMPOUND 間で，リンクによって EC 番号から R 番号，R 番号から C 番号とたどることができる．この関係をもとに EC 番号と代表的な反応式な反応式を構成する各化合物の ID を取得する [40].

PubChem Compound では化合物の特性情報など KEGG にはない情報が記載されており，CID で管理されている．さらに，CID は PubChem Substance において SID とともに併記されていることが多く，SID は KEGG COMPOUND の化合物情報にリンクとして表記されている．これによって，R 番号の C 番号で書かれた反応式からそれぞれの化合物の詳細情報を得ることができる．今回は PubChem から化合物の SMILES 情報を取得し，C 番号と SID・CID の対応によって SMILES 形式の反応式と，EC 番号の対応表を作成する．図 4.4 に C 番号と SID の対応関係を表す． SMILES 形式の化合物を RDKit で読み込むことで，化合物の構造オブジェクトに変換できる．それにより，構造式をコンピュータ上で表現するとともに，RDKit の記述子を用いて特性値を計算し，化合物を数値で表現する．各反応式において，生成物と反応物の差分を取り，作成した SMILES 反応式と EC 番号の対応表より，EC 番号と各反応式の特性値変化量の特徴ベクトルを取得する．

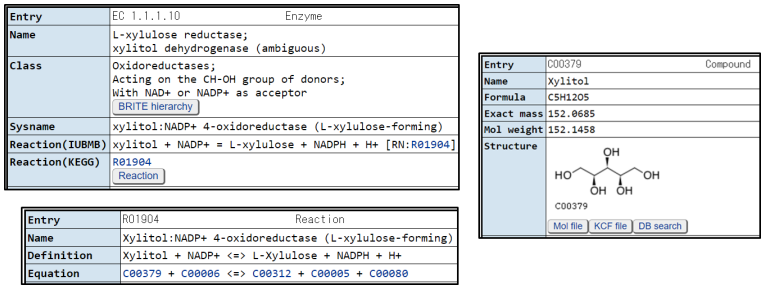


図 4.3: EC 番号，R 番号，C 番号の参照 [14](一部抜粋)

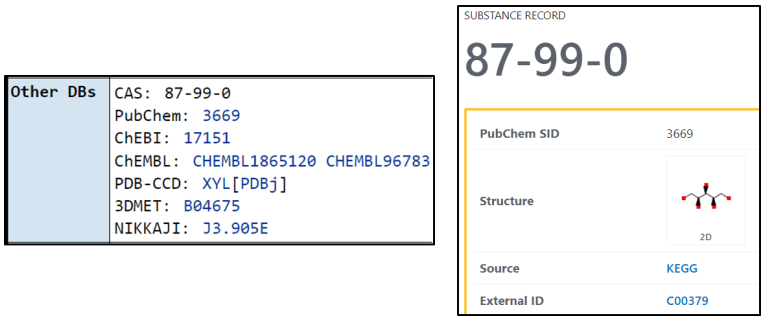


図 4.4: KEGG の C 番号と PubChemSID の対応 [17](一部抜粋)

§ 4.2　凝集型クラスタリングによる次元削減

反応変化の特徴として，特性値変化量からなる多次元の特徴ベクトルを用いるが，次元のサイズが大きすぎることが，問題となるケースがある．一般的には多重共線性や次元の呪いに絡んでくる．多重共線性とは，説明変数間に高い相関があるときに起きる現象で，汎化性能や分類精度の低下の原因とされている．次元の呪いは，用いる変数が多い場合に起こり，過学習の原因とされる問題である．今回のケースでは相関の高い記述子のペアが存在すると，同じような記述子が存在することになり，構造変化の特徴の一部として，他の記述子よりも重みづけが大きくなると考えられる．そのため，多重共線性の問題を解決しつつ，次元の削減も同時に行う．

多重共線性を解決するためには，相関の高いペアの変数に対して，どちらか片方を取り除く方法が取られることが多い．しかし，誤って重要な変数を除去していまう可能性や 3 個以上の変数間の高い相関には対処できない等の問題がある．そのため，相関に基づき，記述子間で凝集性クラスタリングを行うことで多重共線性をなくす方法を用いる [37]．ここでは，最長距離法をクラスタ間の距離としてクラスタリングを行う．図 4.5 にそのイメージを示す．

最初の段階では，記述子どうしのマージが行われ，要素数が 2 つのクラスタが形成される．次にクラスタどうしのマージとなり，最長距離法を用いるが，このとき異なるクラスタの記述子間で最も相関の低いペアに注目する．それらのペアの中で相関が最も高いペアのクラスタどうしは，クラスタ間での相関が最も高い関係と考えられる．つまり，最長距離法を用いることで，多数の記述子間の相関を考慮した，多重共線性の対策となる．次元削減としては，同クラスタ内の記述子を合成した合成記述子を作成することで，相関の高

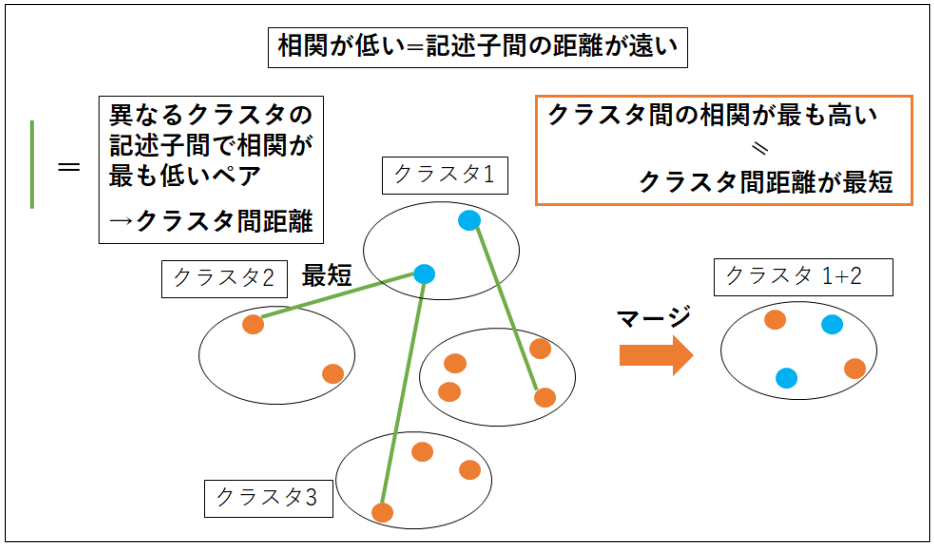


図 4.5: 最長距離法による記述子のクラスタリング

表 4.2: 相関係数の逆数を要素に持つ距離行列

	記述子 0	記述子 2	···	記述子 n
記述子 1	0	$1/s_{12}$	···	$1/s_{1n}$
記述子 2	$1/s_{21}$	0	···	$1/s_{2n}$
⋮	⋮	⋮	⋱	⋮
記述子 n	$1/s_{n1}$	$1/s_{n2}$	···	0

い複数の記述子を新しい 1 つの記述子として表現する．

クラスタプログラムは Python の sklearn に実装されている凝集性クラスタリングである，AgglomerativeClustering ライブラリ [38] を用いる．記述子 u，v 間の相関係数を s_{uv} としたとき，以下のように表される．

$$s_{uv} = \frac{\sum_{i=1}^{m}(cv_{iu} - c\bar{v}_u)(cv_{iv} - c\bar{v}_v)}{\sqrt{\sum_{i=1}^{m}(cv_{iu} - c\bar{v}_u)}\sqrt{\sum_{i=1}^{m}(cv_{iv} - c\bar{v}_v)}} \text{(ただし，} c\bar{v}_u\text{，} c\bar{v}_v \text{は記述子 } u,v \text{ の特性値平均)} \tag{4.3}$$

s_{uv} に対して，逆数を取った，$1/s_{uv}$ を記述子間の距離とし，Python で表 4.2 のような距離行列を作成してクラスタリングする．ここでは，相関係数が 1 となる要素を 0 としている．AgglomerativeClustering では，入力データとして通常の特徴ベクトルだけでなく，距離行列を用いることができ，記述子間でマージするときの閾値を指定することができる．今回は相関係数 $s_{ij} \geqq 0.9$ すなわち，$1/s_{ij} \leqq 1/0.9 \approx 1.11$ で記述子をマージする．クラスター間でのクラスタリングにおいて，最遠距離法を用いたとき，クラスター間距離 $d(C_1, C_2)$ は，式 3.3 より以下のようになる．

$$d(C_1, C_2) = \max_{\mathbf{u}\in C_1, \mathbf{v}\in C_2} \frac{1}{s_{uv}} \qquad \left(\text{ただし，} \frac{1}{s_{uv}} \leqq 1.11\right) \tag{4.4}$$

これらを用いて次の手順で記述子間のクラスタリングを行う．

1. $1/s_{ij} \leqq 1.11$ を満たす，記述子のペアにおいて，互いの距離が最短となるものをマージする．
2. $1/s_{ij} \leqq 1/0.9 \approx 1.11$ となる記述子ペアが存在するクラスタ間で，$d(C_1, C_2)$ が最小となるクラスタ C_1，C_2 をマージする．条件を満たす記述子ペアが存在しなくなるまで繰り返す．
3. クラスタリングを終了後，クラスタ番号とそのクラスタに所属する記述子の対応表を取得する．
4. 同クラスタ内の各記述子の特性値列に対し標準化，および平均化を行い，合成記述子を新たな記述子として利用する．

表 4.1 において，同クラスタの記述子同士をまとめ，合成記述子 clusterX(X はクラスタ番号) と置き換えることで，次元削減を行う．

§ 4.3　SOM による反応式クラスタリング

次元削減した特徴ベクトルを用いて，反応式間の類似度を比較する手法として，SOM によるクラスタリングを行う．SOM を用いることの利点として，2 つ挙げられる．1 つ目は，低次元空間への可視化が可能になる点である．高次元の特徴ベクトルの場合，反応式どうしの位置関係が把握しにくいが，2,3 次元まで圧縮することで，その関係を把握することが可能となる．2 つ目として，クラスタリングによる類似比較が挙げられる．類似度を比較する手法として，コサイン類似度や相関係数等が用いられるが，複数の反応式間の類似度を調べたいときには，直感的な理解が難しい場合がある．その際に，クラスタリングを用いることによって，全ての反応式の類似性を把握することができる．これらのことから，ターゲット反応式の近くに分布する類似性の高い EC 反応式を複数同時に確認できる他，他の反応式どうしの類似性も見ることができるようになる．

用いる SOM のプログラムとして，KH Corder で出力される SOM の R 言語ファイルを参考に作成された，R 言語使用のソースコードを用いる [34]．入力するデータは次元削減後の各反応式の特徴ベクトルを全体に対して標準化したものを用いる．SOM のプログラム中には R 言語のパッケージとして実装されている som を使用している [39]．プロット点のラベルはターゲット (T) と反応式の EC 番号を扱う．

ユニット数は 400(20×20) であり，ユニットの形状を六角形とする．学習は大まかな順位付けを行う段階と収束段階の 2 段階に分けて行う (KH Coder3 リファレンス・マニュアルに記載)．今回は 1 段階目 1,000 回，2 段階目は 200,000 とした．SOM の実行後，各勝者ユニット上に反応式の特徴ベクトルがマッピングされ，色分けによる凝集型クラスタリングが実行される．このクラスタリングはユークリッド距離によるウォード法によって行われ，今回はクラスタ数 9 で色分けされる．

第5章

実験結果並びに考察

§ 5.1　数値実験の概要

本研究の実験の流れについて説明する．まず, 化合物の特徴ベクトルを求めるため, KEGG と PubChem から各反応式の情報を取得する．次に，反応式内の反応物・生成物の SMILES を出力する．さらに，RDKit にある 208 種の記述子を用いて，化合物の物理・化学特性値を計算し，特性値変化量を求めることで，ターゲット反応式と EC 反応式を，208 次元の特徴ベクトルで表現する．さらに，不適切な値を含む記述子を除外し，凝集型クラスタリングによって相関の高い記述子同士をまとめて，新たな合成記述子を作成することで，次元削減を行う．最後に，SOM によって反応式をクラスタリングし，ターゲット反応式に対して適切な酵素を予測する．

具体的なデータ整理，前処理，および分析条件について以下で説明していく．

ターゲット反応式と提案手法の評価方法

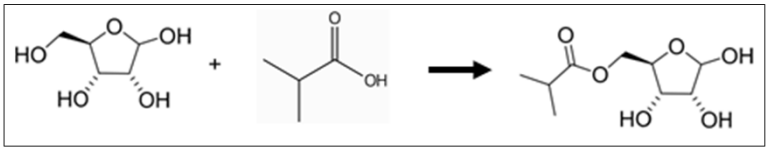
今回は，モルヌピラビルを生成する過程における，1 ステップ目の合成の反応式に焦点を当てる [2]．図 5.1 にターゲット反応式を示す．ターゲット 2 が本来行われた合成であり，リボース (左辺第 1 項) の第一級アルコール部分を選択的にエステル化する反応である．ここでは，8 つの酵素製品に対する，生成物のアッセイ収率を調べるための，スクリーニングを行っている．最終的に Novozym435 の酵素製品が一番優れた結果となったが，これは BRENDA によると EC3.1.1.3 に分類される酵素とされている．特性値変化量が，酵素番号予測を行うために，十分な特徴を備えている場合，この反応式をターゲットとして他の EC 反応式とともに SOM によるクラスタリング行えば，EC3.1.1.3 の反応式がターゲット 2 の付近に位置すると考えられる．

一方で，ターゲット 2 の反応は，通常では起こりえない特殊な反応である．初めに別の反応を試したのち，生成物の収率を上げるために，等価体として類似の性質を持つ化合物に置き換えた，ターゲット 2 を用いたと考えられる．ターゲット 1 も EC3.1.1.3 の酵素を用いた場合に起こりうる反応であり，初めに行った別の反応としてターゲット 1 を仮定する．EC 反応式とそれぞれのターゲット反応式の類似性を SOM のクラスタリングで可視化し，基準として EC3.1.1.3 反応式がターゲットに対してどの場所に位置するかで提案手法を評価する．

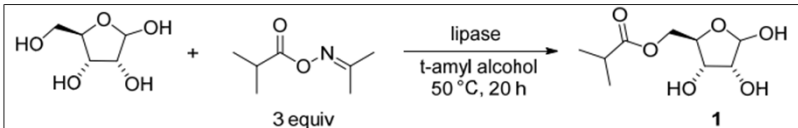
比較対象となる EC 反応式

今回の予測は比較する EC 反応式をあらかじめ絞ったうえで行う．ターゲットの反応はエステル加水分解の逆反応となるエステル化反応のため，用いる酵素として，EC3.1.1 の加

ターゲット1



ターゲット2



Enzyme Name	5-isobutyryl ribose assay yield%
Novozym 435 (Candida antarctica lipase B)	65
IMMTLL-T2-150 (Thermomyces lanuginosus lipase)	40
IMMRES-T2-150 (Resinase HT lipase)	38
IMMLIPX-T2-150 (Lipex 100 L lipase)	56
IMML51-T2-150 (Novozymes 51032)	61
IMMP6-T2-250 (protease from Bacillus licheniformis)	11
Lipozyme RM IM	10
CDX IMB-103	33

図 5.1: ターゲット反応式

水分解酵素が適当であると考えられる．これは，加水分解が一般的には可逆的な反応であり，加水分解酵素でエステル化も可能であるためである．したがって，EC3.1.1 反応式もエステル化する方向 (右辺を反応物，左辺を生成物とする) でターゲットと比較を行う．一方で，全ての EC 反応式に対して比較することも重要であると考えられる．しかし，EC3.1.1 以外の番号で類似していると認識される可能性を考慮し，今回は，EC3.1.1 と認識された場合を仮定し，そのうえで提案した記述子変化量とその記述子選択によって，適切な酵素を予測できるか検証する．

データの対応表取得と整理

まず，KEGG の ID や反応式を取得するソースコードを用いて [40]，EC3.1.1 に属する EC 番号，対応する R 番号，C 番号で構成された R 番号の反応式を取得し，それぞれの対応表を作成した．同様に PubChem からは C 番号と CID，SID の対応表を取得した．次に C 番号と CID または SID を参照し，PubChemPy によって，各反応式の反応物と生成物の SMILES を取得することを試みた．しかし，C 番号に対する CID がまだ登録されていない化合物や，SID を引数にして，PubChemPy から SMILES を取得できないなどの問題が発生した．そこで，PubChem の SID で検索したリンク内で SDF ファイルを入手し，それを RDKit で読み込み SMILES に変換した．また，ターゲットの SMILES は SciFinder^n で入手した MOL ファイルを RDKit で変換することで取得した．それらの SMILES をまとめて，ターゲットおよび EC 番号に対する，各反応物・生成物の SMILES 対応表を作成した．以下，表 5.1，表 5.2, および表 5.3 にそれぞれの対応表を示す．

対応表の前処理 1

得られた SMILES 対応表には KEGGC COMPOUND に登録されていない (番号が新しい) 化合物，あるいは登録されているが，構造式が記載されていない化合物が存在する．そ

表 5.1: EC 番号と KCID

	ENZYME	left1	left2	right1	right2	right3
0	3.1.1.22	C04546	C00001	C01089	None	None
1	3.1.1.20	C01572	C00001	C01424	None	None
2	3.1.1.40	C02868	C00001	C01839	None	None
3	3.1.1.33	C02655	C00001	C00031	C00033	None
4	3.1.1.6	C01883	C00001	C00069	C00033	None
...	...	...	...	...	...	...
173	3.1.1.111	C18125	C00001	C22237	C00162	None
174	3.1.1.115	C22218	C00001	C22219	None	None
175	3.1.1.117	C22373	C00001	C22374	C00069	None
176	3.1.1.118	C01194	C00001	C22400	C00162	None
177	3.1.1.118	C00416	C00001	C03974	C00162	None

表 5.2: 各化合物 ID 対応表

	cid	pubchem_SID	pubchem_CID
1	C00001	3303	962
2	C00002	3304	5957
3	C00003	3305	5893
4	C00004	3306	439153
5	C00005	3307	5884
...	...	...	...
18594	C22269	405226444	6365572
18595	C22272	405226445	11788398
18596	C22273	405226446	11411510
18597	C22274	405226447	135567131
18598	C22275	405226448	44468216

表 5.3: EC 番号と SMILES の対応表

	ENZYME	left1	left2	right1	right2	right3
0	3.1.1.22	C[C@@H](O)CC(=O)O[C@H](C)CC(=O)O	[H]O[H]	C[C@@H](O)CC(=O)O	N	N
1	3.1.1.20	O=C(O)c1cc(O)c(O)c(OC(=O)c2cc(O)c(O)c(O)c2)c1	[H]O[H]	O=C(O)c1cc(O)c(O)c(O)c1	N	N
2	3.1.1.40	Cc1cc(OC(=O)c2c(C)cc(O)cc2O)cc(O)c1C(=O)O	[H]O[H]	Cc1cc(O)cc(O)c1C(=O)O	N	N
3	3.1.1.33	CC(=O)OC[C@H]1O[C@@H](O)[C@H](O)[C@@H](O)[C@@H]1O	[H]O[H]	OC[C@H]1OC(O)[C@H](O)[C@@H](O)[C@@H]1O	CC(=O)O	N
4	3.1.1.6	*OC(C)=O	[H]O[H]	*O	CC(=O)O	N
...	...	...	...	...	...	...
173	3.1.1.111	*C(=O)OCC(O)COP(=O)(O)OC[C@H](N)C(=O)O	[H]O[H]	N[C@@H](COP(=O)(O)OCC(O)CO)C(=O)O	*C(=O)O	N
174	3.1.1.115	O=C1OC[C@](O)(CO)[C@H]1O	[H]O[H]	O=C(O)[C@H](O)C(O)(CO)CO	N	N
175	3.1.1.117	N	[H]O[H]	N	*O	N
176	3.1.1.118	*C(=O)OC[C@H](COP(=O)(O)O[C@H]1[C@H](O)[C@@H](...	[H]O[H]	N	*C(=O)O	N
177	3.1.1.118	*C(=O)OCC(COP(=O)(O)O)OC(*)=O	[H]O[H]	*C(=O)O[C@H](CO)COP(=O)(O)O	*C(=O)O	N

のため，対応表内の SMILES の項に空白となる部分が発生するため，その項を含む反応式は除外した．また，ターゲットは反応物 2 個，生成物 1 個の組み合わせであるため，EC 反応式はターゲット同様の組み合わせにする．これは，提案手法の特性値変化量の計算に用いられる化合物の数が増加または減少することで，構造変化とは別の要因による変化が影響すると考えられるためである．つまり，ターゲットの物理・化学特性値の純粋な変化と比較して，化合物の多寡による特性値の変化も追加されると推測されるためである．以上の理由から，ほとんどの反応式に含まれている H_2O を除外した場合の，反応物が 2 個，生成物が 1 個の組み合わせとなる EC3.1.1 反応式 113 種のみを採用した．

物理・化学特性値および記述子変化量の計算

ターゲット+113 種の SMILES 対応表を元に，RDKit の rdkit.chem.descriptor から 208 種の記述子名を取得し，反応式 1，反応式 2，生成物それぞれの場合で特性値を計算した．その後，特性値変化量を求め，図 5.4 のような 208 次元ベクトルを持つ反応式の表を作成した．

対応表の前処理 2

図 5.4 の表から，nan 値や発散している要素を持つ記述子を除外した．また，全ての反応式において等しい特性値を持つ記述子を除外し，最終的に 128 種類 (ターゲット 1)，または 129 種類 (ターゲット 2) の記述子で次元削減を行う．

表 5.4: 各反応式の特性値変化量

	MaxEStateIndex	MinEStateIndex	MaxAbsEStateIndex	MinAbsEStateIndex	qed	MolWt	HeavyAtomMolWt	ExactMolWt	NumValenceElectrons
Target	-8.378152	0.949632	-8.378152	-0.144028	-0.330982	-18.015	-15.999	-18.010565	-8
3.1.1.33	-7.632875	0.794822	-7.632875	-1.064815	-0.343138	-18.015	-15.999	-18.010565	-8
3.1.1.6	-6.597222	0.946759	-6.597222	-0.949074	-0.409219	-18.015	-15.999	-18.010565	-8
3.1.1.1	-6.486111	0.972222	-6.486111	-0.675926	-0.331106	-17.007	-15.999	-17.003288	-8
3.1.1.7 3.1.1.8	-7.085822	0.351574	-7.085822	-0.914074	-0.484689	-18.015	-15.999	-18.010565	-8
...	...	...	...	...	...	...	...	...	...
3.1.1.106	-8.896201	0.794521	-8.896201	-0.839784	-0.462056	-18.015	-15.999	-18.010565	-8
3.1.1.113	-6.747917	0.372685	-6.747917	-0.872685	-0.398840	-18.015	-15.999	-18.010565	-8
3.1.1.112	-7.033650	0.317731	-7.033650	-0.979769	-0.421762	-18.015	-15.999	-18.010565	-8
3.1.1.111	-8.902683	0.535378	-8.902683	-0.657129	-0.360318	-18.015	-15.999	-18.010565	-8
3.1.1.118	-8.839073	0.535378	-8.839073	-0.575822	-0.317022	-18.015	-15.999	-18.010565	-8

表 5.5: 記述子間の相関係数に基づくクラスタリング結果 (ターゲット 1)

0	0	1	1	2	2	2	2	2	2	2	2	2	2	2	2	2
Kappa2	Kappa3	fr_Al_COO	fr_COO	NumValeI	Chi0n	Chi0v	Chi1n	Chi1v	Chi2n	Chi2v	Kappa1	LabuteAS	SMR_VSA	SlogP_VS	NumRotaI	MolMR
3	3	4	4	5	5	6	6	6	7	7	8	8	8	9	9	9
NumAliph	RingCoun	FpDensit	FpDensit	SMR_VSA	SlogP_VS	SMR_VSA	VSA_ESta	fr_C_O	NumAliph	NumSatu	fr_Ar_OH	fr_phenol	fr_phenol	fr_alkyl_h	fr_ketone	fr_lactone
10	11	12	12	13	14	14	15	15	15	16	16	16	16	17	17	17
fr_ester	fr_ether	MaxEStat	MaxAbsE	NumSatu	fr_NH1	fr_NH2	fr_amide	VSA_ESta	fr_allylic	VSA_ESta	NumAron	NumAron	fr_benzen	MolWt	HeavyAtc	ExactMolWt
17	17	17	17	17	17	17	18	18	18	19	20	21	22	23	24	25
Chi0	Chi1	Chi3n	Chi3v	Chi4n	Chi4v	HeavyAtc	SMR_VSA	TPSA	NOCount	NHOHCo	EState_V	EState_V	NumHete	fr_COO2	FractionC	VSA_EState4
26	27	28	29	30	31	32	33	34	35	36	36	36	36	36	37	38
VSA_ESta	VSA_ESta	NumHDo	fr_bicycli	fr_C_O_n	fr_metho	fr_Ar_CO	VSA_ESta	SlogP_VS	fr_unbrch	SlogP_VS	NumAliph	NumSatu	fr_NH0	fr_piperd	EState_V	fr_Al_OH
38	39	40	41	42	43	44	45	46	47	48	49	50	51	52	53	54
fr_Al_OH	VSA_ESta	Ipc	PEOE_VS	SlogP_VS	PEOE_VS	HallKierA	PEOE_VS	EState_V	PEOE_VS	fr_ArN	PEOE_VS	EState_V	SMR_VSA	EState_V	PEOE_VS	EState_VSA10
55	56	57	58	59	60	61	62	63	64	65	66	67	68	69	70	71
EState_V	PEOE_VS	qed	VSA_ESta	PEOE_VS	fr_aldehy	EState_V	FpDensit	MinAbsE	SlogP_VS	VSA_ESta	EState_V	BalabanJ	PEOE_VS	PEOE_VS	PEOE_VS	SMR_VSA6
72	73	74	75	76	77	78	79									
SlogP_VS	PEOE_VS	BertzCT	PEOE_VS	SMR_VSA	MolLogP	NumHAc	MinEStateIndex									

§ 5.2 実験結果と考察

特徴ベクトルの次元削減

相関係数の逆数である距離行列を入力とした，最遠距離法の凝集型クラスタリングによる次元削減を行った．各ターゲットの場合でそれぞれ 18 個のクラスタが形成され，ターゲット 1 では 80 次元，ターゲット 2 では 82 次元の特徴ベクトルとなった．表 5.5 に，ターゲット 1 の場合のクラスタ番号と，そのクラスタに含まれる記述子名の対応表を示す．12 番のクラスタに所属する記述子「MaxEStateIndex」と「MaxAbsEStateIndex」は相関係数が 1 であるが，ともにマージされていることが分かる．また，記述子名が類似している記述子が，同じクラスタに属している傾向があることが分かる．表 5.6 にはターゲット 1 の場合に，クラスタリングでマージされた記述子，および次元削減の結果を示す．合成された記述子は，クラスタ番号 X を後ろにつけた「clusterX」で表示されている．また，ターゲットを T とし，EC 番号の下 1 桁のみ表示している，ピリオド以下の番号は，EC 番号の代表の反応式が，複数ある場合の区別に用いられている．アンダーバーで区切られているものは，その反応式が複数の EC 番号間で重複している場合の区別となっている．

SOM による反応式のクラスタリング結果

SOM のプログラムによって，反応式をクラスタリングするとターゲット 1，ターゲット 2 でそれぞれ図 5.2 のようになった．E は EC3.1.1 以外の反応式を表す．ターゲット 1 では青色のクラスタが離れているが，これは，SOM が本来は 2 次元平面を円柱状に丸め，さらに

表 5.6: 次元削減後におけるターゲット 1 と EC 反応式の特徴ベクトル

	cluster0	cluster1	cluster2	cluster3	cluster4	cluster5	cluster6	cluster7	cluster8	cluster9	...	PEOE_VSA10	SMR_VSA6	SlogP_VSA10
T	-0.358029	-0.842985	0.204247	4.571328	0.622585	0.207469	-1.696273	2.341995	0.173683	-0.133043	...	0.796801	0.043121	-0.1269
33	-0.223667	-0.842985	0.219299	-0.103504	0.568775	0.207469	0.045521	-0.141173	0.173683	-0.133043	...	0.796801	0.043121	-0.1269
6	3.495807	-0.842985	0.161480	-0.103504	-0.360529	0.207469	0.004312	-0.141173	0.173683	-0.133043	...	0.046341	0.043121	-0.1269
1	3.518113	1.079686	0.207055	-0.103504	-0.497256	0.207469	0.009803	-0.141173	0.173683	-0.133043	...	0.046341	0.043121	-0.1269
7_8	-0.214983	-0.842985	0.219101	-0.103504	0.596503	0.207469	0.036318	-0.141173	0.173683	-0.133043	...	0.796801	0.043121	-0.1269
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
106.1	-0.233937	-0.842985	0.213692	-0.103504	0.310413	0.207469	0.083963	-0.141173	0.173683	-0.133043	...	-0.646994	0.043121	-0.1269
113	-0.251825	-0.842985	0.217973	-0.103504	0.596561	0.207469	0.012651	-0.141173	0.173683	-0.133043	...	0.046341	0.043121	-0.1269
112	-0.182730	-0.842985	0.219101	-0.103504	0.501004	0.207469	0.031504	-0.141173	0.173683	-0.133043	...	0.046341	0.043121	-0.1269
111	-0.280504	1.079686	0.215345	-0.103504	-0.970931	0.207469	0.055048	-0.141173	0.173683	-0.133043	...	-1.333272	0.043121	-0.1269
118	-0.266600	1.079686	0.223487	-0.103504	-1.377476	0.207469	0.075928	-0.141173	0.173683	-0.133043	...	0.046341	0.043121	-0.1269

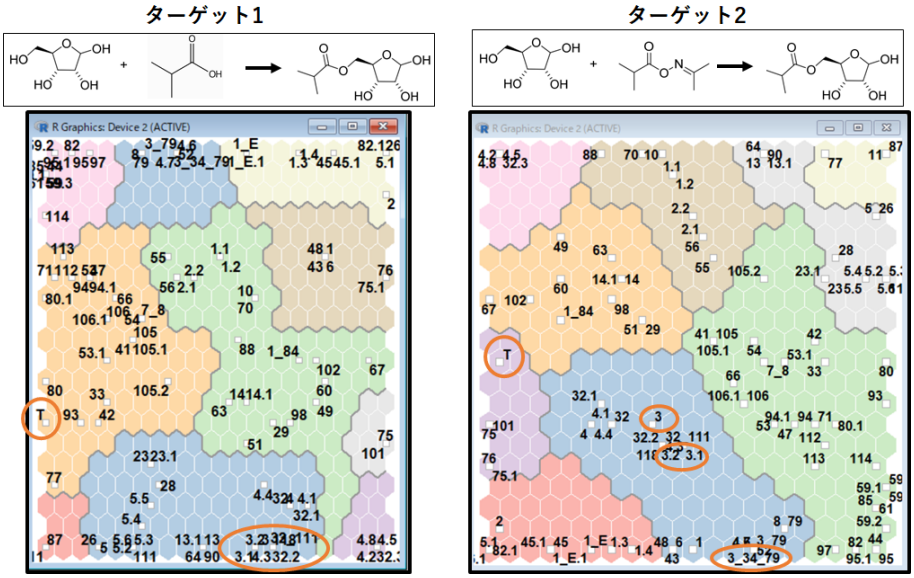


図 5.2: SOM による反応式のクラスタリング結果

円柱を曲げて切り口をつないだトーラス型の形状をしており，2 次元平面にした際に分裂したものと考えられる．ターゲット 1(T) と同じクラスタに属し，かつ付近に位置する EC 反応式として，EC3.1.1.80，EC3.1.1.93 という結果となった．これらのターゲット反応式と比較すると図 5.3 のようになった．EC 3.1.1.93 の右辺 (反応物) である C00191 は，リボースと同じ糖類に分類されるグルコースの構造が含まれている．また，ターゲット 2(T) において同じクラスタに属するのは，EC 3.1.1.75(2 つ)，EC 3.1.1.76，EC 3.1.1.101 となった．図 5.4 にこれらの反応式を示す．各反応式中には同様の構造が n 個連なる重合体が多く含まれる結果となった．一方で，EC3.1.1.3 の代表反応式は他 EC 番号の重複を含めて 4 種採用しているが，いずれの結果においても全てターゲットと異なるクラスタに属していた．

考察 1

ターゲット 2 を用いた反応式クラスタリングにおいて，特性値変化が類似しているものとして，本実験では，あらかじめ用いられた酵素が分かっているターゲットを使用した．提案手法によって優れた予測が行われているかを検証するため，その EC 番号反応式が最も類似しているものとして，認識されるかを確認した．結果として，目的としていた EC 3.1.1.3

反応式はSOMのマップ上において，ターゲット付近に位置せず，最も類似しているものとして提示されなかった．原因として，以下の4つが挙げられる．

1つ目は反応式中の係数を反映していなかったことが挙げられる．用いられる全ての化合物の比率を平等にした場合でも，特性値変化量は構造変化の特徴として機能すると考えられる．しかし，反応式中の1つの化合物が用いられる分子数だけ特性値を上乗せすることで，より化学的に構造変化を捉えられると考えられる．

2つ目は，ターゲットの反応において，提案した特徴変化量では，捉えきれていない要因が多く影響している点である．モルヌピラビルの論文のサポート資料 [41] では，tert-アミルアルコールを溶媒として用いており，50 ℃で 20 時間振とうを行うことでターゲットの生成物を生成している．また，用いた反応物の分量なども異なっている．一方で，EC 反応式は生物の体内等で起こる反応であり，基本的には有機溶媒等を用いない．実験に用いた試薬や，溶媒，実験環境，配合比率などの様々な要因によって，天然に起こる反応との差異を発生させていると考えられる．特性値変化量だけでなく，反応以外の要因も考慮した特徴作成が必要である．

3つ目は，相関係数に基づくクラスタリングで，同一クラスタに存在する記述子を合成した際に，構造変化に重要な特徴の影響力を弱めてしまった可能性が挙げられる．今回は，相関の高い記述子ペアに対し，片方を除外することで重要な記述子を誤って削除するのを避けるため，クラスタ内の記述子どうしで標準化，および平均化を行った．しかし，やはり重要な記述子と重要でない記述子が混ざったクラスタが存在し，平均化によって，重要な記述子の説明力を薄めてしまった可能性がある．改善策としては，相関係数に応じて形成されたクラスタ内で，記述子の重要度に基づいて，重みづけをする手法を提案できれば，適切な記述子選択と次元削減ができると考えられる．

4つ目として，次元削減後に用いられた記述子は80種とまだまだ多く，不必要な記述子によって構造変化を上手く説明できなかったことが考えられる．今回は主に多重共線性の対策としての手法を提案したが，少数かつ構造変化を十分に説明できるような，記述子の組み合わせを提案する手法を検討していきたい．

考察 2

ターゲット 2 と同クラスタの反応式において，多数の重合体が含まれていた点について，ターゲット化合物が重合体の特性と類似点が多い可能性がある．また，EC3.1.1.3 よりも

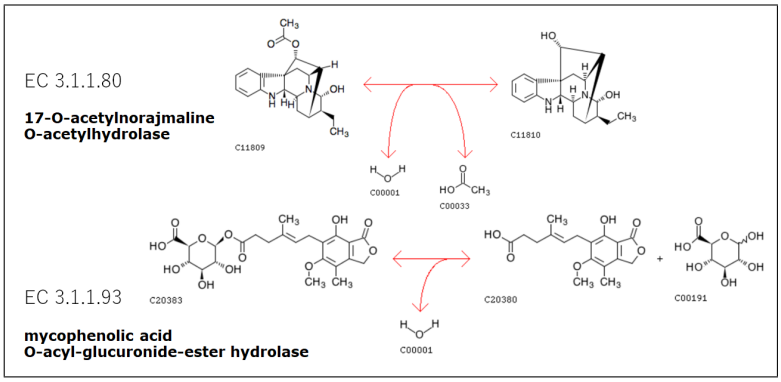


図 5.3: ターゲット 1 の近くに位置している反応式

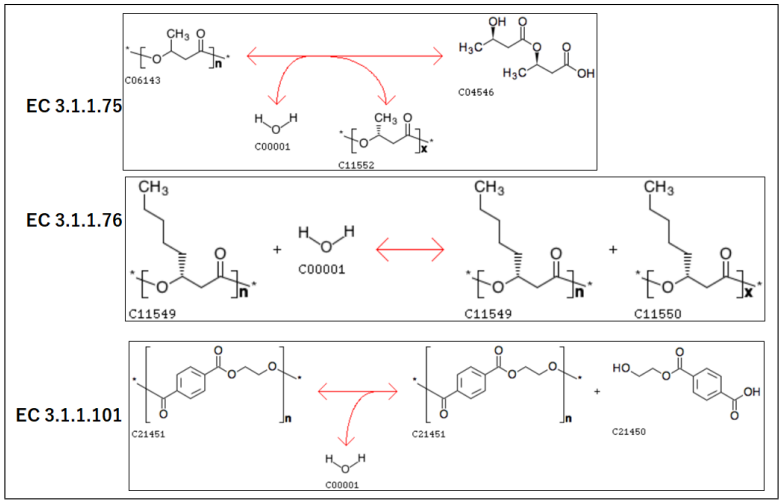


図 5.4: ターゲット 2 と同クラスタに属する反応式

ターゲットの近くに位置していた EC 番号について，実験に関する文献数が少なく，あまり知られていない酵素の場合，今後の検証実験等で，EC3.1.1.3 よりも優れた酵素であることが示される可能性があるため，これらについてさらなる分析が必要であると考えられる．

第6章

おわりに

近年，新型コロナウイルスになどの影響で，新薬開発の需要が高まり，化学反応の設計や予測を行う研究が発展を続けている．一方で，反応の効率化と環境面から，酵素の生体触媒を用いて合成が行われる機会が増えており，目的の反応に対して最適な酵素を予測することが重要視されている．しかし，基質特異性などの酵素の性質は生物分野にかかわるため，有機合成の知識のみでは解決が難しく，酵素研究の専門家協力する，または，酵素データベースを参照するなどして最適な酵素候補は探索されていた．

目的とする反応に対して，酵素候補を予測するシステムがあれば，次のステップである1つの酵素に絞るスクリーニングまでスムーズに進めることができる．また，酵素はEC番号と，生体内で自身が使用されて起こる代表的な反応の反応式で管理されている．

これらのことから，本研究では，ターゲットとなる反応式を与えた際に，EC番号の代表的な反応式と比較し，最も類似する反応式のEC番号を最適な酵素として予測する．予測の方法として，化合物の物理・化学的な特性値を計算し，反応物から生成物の差分を取った特性値変化量をもとに，反応式どうしの類似性を比較することを提案した．

KEGGやPubChemなどで必要とするデータを取得し，RDKitを用いて各反応に対して208種の特性値変化量を計算し，特徴ベクトルを作成した．その後，凝集型クラスタリングによって特徴ベクトルの次元を80次元まで削減し，SOMによって反応式のクラスタリングを行った．

2パターンでの検証を行い，1つ目では，ターゲットの反応式に対して，EC3.1.1.80，EC3.1.1.93 の反応式が，2つ目では，EC 3.1.1.75，EC 3.1.1.76，EC 3.1.1.101 が類似していると判定された．本来予測されるはずのEC 3.1.1.3を予測することはできなかったが，ターゲット反応式・EC反応式，または各反応式間で共通する特徴を確認することができた．

今後の課題として，化学反応時の構造変化を特徴としてより詳細に捉えるため，重要な記述子を残しつつ，さらに次元を削減していく手法を開発することが挙げられる．また，EC番号の反応式のクラス分類に着目し，最も精度よく分類できる記述子の組み合わせを特定したのち，ターゲットの反応式の酵素予測において検証していくことが考えられる．

謝辞

本研究を遂行するにあたり，多大なご指導と終始懇切丁寧なご鞭撻を賜った富山県立大学工学部電子・情報工学科情報基盤工学講座の奥原浩之教授，António Oliveira Nzinga René 講師に深甚な謝意を表します．そして，有機化学・酵素化学に関して貴重なご意見をいただいた工学部生物工学科酵素化学工学講座の浅野泰久教授，くすりのシリコンバレー TOYAMA 研究拠点化プロジェクトディレクター補佐の岩﨑源司博士 (薬学) に感謝申し上げます．最後になりましたが，多大な協力をしていただいた研究室の同輩諸氏に感謝致します．

2022 年 2 月

武藤 克弥

参考文献

[1] “ケモインフォマティクス市場、2021 年から 2026 年の間に CAGR13 %で成長見込み”, https://prtimes.jp/main/html/rd/p/000002048.000071640.html, 閲覧日 2022.1.6.

[2] Tamas Benkovics, John A. McIntosh, Steven M. Silverman, Jongrock Kong, Peter Maligres, Tetsuji Itoh, Hao Yang, Mark A. Huffman, Deeptak Verma, Weilan Pan, Hsing-I Ho, Jonathan Vroom, Anders Knight, Jessica Hurtak, William Morris, Neil A. Strotman, Grant Murphy, Kevin M. Maloney, and Patrick S. Fier1,, “Evolving to an Ideal Synthesis of Molnupiravir, an Investigational Treatment for COVID ‐ 19”, *ChemRxiv*, 2020.

[3] 北川勲，磯部稔, ”天然物化学・生物有機化学 I”. 朝倉書店, 2008. 3-4 ページ

[4] 西村淳, 樋口弘行, 大和武彦, ”有機合成化学入門 -基礎を理解して実践に備える” 丸善株式会社, 2010. 1 ページ

[5] “日本化学会・ケモインフォマティクス部会”, https://cicsj.csj.jp/, 閲覧日 2022.1.23.

[6] 中野裕太, 瀧川一学, “化学反応ネットワークにおける最適反応経路候補の列挙”, 情報処理学会研究報告, Vol. 122, No. 16, 2019.

[7] 佐藤寛子, “化学情報学 - 化学反応の系図と反応予測 -” 国立情報学研究所, 2003

[8] 藤波 美起登, 清野 淳司, “量子化学計算情報を記述子とした機械学習に基づく反応予測手法の開発”, *Journal of Computer Chemistry, Japan*, Vol. 15, No. 3, pp. 63-65, 2016.

[9] “酵素の化学”, http://www.sc.fukuoka-u.ac.jp/~bc1/Biochem/biochem5.htm, 閲覧日 2022.1.31.

[10] “酵素基質とは”, https://bizcomjapan.co.jp/iris-biotech/knowledge/substrate/, 閲覧日 2022.1.31.

[11] “新設された酵素分類 EC7 の和名提案について”, https://www.jbsoc.or.jp/notice/ec_translocase.html, 閲覧日 2022.1.15.

[12] 白兼孝雄, “酵素の分類と命名法”, JAS 情報, 2017

[13] “Enzyme Nomenclature”, https://iubmb.qmul.ac.uk/enzyme/, 閲覧日 2022.1.15.

[14] ”KEGG: Kyoto Encyclopedia of Genes and Genomes”, https://www.genome.jp/kegg/kegg_ja.html, 閲覧日 2022.1.17.

[15] ”CAS SciFinder^n” https://scifinder-n.cas.org/ 閲覧日 2022.1.23.

[16] "CAS SciFinder^n 検索ガイド" https://www.jaici.or.jp/scifinder-n/ref/sfn.pdf 閲覧日 2022.2.3.

[17] "PubChem", https://pubchem.ncbi.nlm.nih.gov/, 閲覧日 2022.1.17.

[18] "About PubChem", https://pubchemdocs.ncbi.nlm.nih.gov/about, 閲覧日 2022.2.6

[19] Sunghwan Kim, Paul A. Thiessen, Evan E. Bolton, Jie Chen, Gang Fu, Asta Gindulyte, Lianyi Han, Jane He, Siqian He, Benjamin A. Shoemaker, Jiyao Wang, Bo Yu, Jian Zhang, Stephen H. Bryant, "PubChem Substance and Compound databases", *Nucleic Acids Research*, Vol. 44, No. 1, pp. 1202-1213, 2016.

[20] "BRENDA The Comprehensive Enzyme Information System", https://www.brenda-enzymes.org/index.php, 閲覧日 2022.2.1

[21] "KEGG API", https://www.kegg.jp/kegg/rest/keggapi.html, 閲覧日 2022.2.1

[22] Sunghwan Kim, Paul A. Thiessen, Evan E. Bolton, Stephen H. Bryant, "PUG-SOAP and PUG-REST: web services for programmatic access to chemical information in PubChem", *Nucleic Acids Research*, Vol. 43, No. 1, pp. 605-611, 2015.

[23] "SMILES 記法は化学構造の線形表記法" https://future-chem.com/smiles-smarts/, 閲覧日 2022.1.27

[24] Joseph L. Durant, Burton A. Leland, Douglas R. Henry, and James G. Nourse, "Reoptimization of MDL Keys for Use in Drug Discovery", *American Chemical Society*, Vol. 7, No. 12, 2012.

[25] "The RDKit Documentation", https://www.rdkit.org/docs/GettingStartedInPython.html#list-of-available-descriptors, 閲覧日 2022.2.6

[26] "PubChemPy documentation", https://pubchempy.readthedocs.io/en/latest/#, 閲覧日 2022.2.6

[27] Qian-Nam Hu, Hui Zhu, Xiaobing Li, Manman Zhang, Zhe Deng, Xiaoyan Yang, and Zixin Deng, "Assignment of EC Numbers to Enzymatic Reactions with Reaction Difference Fingerprints", *J. Chem. Inf. Comput. Sci.*, Vol. 42, No. 6, 2002.

[28] Yoshihiro Yamanishi, Masahiro Hattori, Masaaki Kotera, Susumu Goto, Minoru Kanehisa, "E-zyme: predicting potential EC numbers from the chemical transformation pattern of substrate-product pairs", *Bioinformatics.*, Vol. 25, pp. 179-186, 2009.

[29] "クラスタリング (クラスター分析)", https://www.kamishima.net/jp/clustering/#bib_cutting, 閲覧日 2022.2.3

[30] "クラスタリングとは — 概要・手順・活用事例を紹介", https://ledge.ai/clustering/, 閲覧日 2022.2.3

[31] “クラスタリング手法の列挙 (一部)”, https://qiita.com/sotoattanito/items/b885ef2dd3fe11cb817d, 閲覧日 2022.2.8

[32] Teuvo KOHONEN, ”Self-organized formation of topologically correct feature map”, *Biological Cybernetics*, Vol. 43, pp. 59–69, 1982.

[33] 亀岡瑶, 宗像昌平, 八木圭太, 山本儀郎, “自己組織化マップによる顧客の分類とその可視化”, 計算機統計学, Vol. 29, No. 2, pp. 181-188, 2016.

[34] 福嶋 瑞希,”環境認識ライフログからの行動パターン解析による類似性・イベント検出”, 富山県立大学学位論文 2018.

[35] ”KH Coder” "http://khcoder.net/" 閲覧日 2022.1.30

[36] Mark A. Johnson, Gerald M. Maggiora, “Concepts and Applications of Molecular Similarity”, *Wiley*, New York, 1990.

[37] “[Python コード付き] 相関係数で変数選択したり変数のクラスタリングをしたりしてみましょう”, https://datachemeng.com/variable_selection_and_clustering_based_on_r/, 閲覧日 2022.1.29

[38] “sklearn.cluster.AgglomerativeClustering ”, https://scikit-learn.org/stable/modules/generated/sklearn.cluster.AgglomerativeClustering.html, 閲覧日 2022.2.3

[39] “Package ‘som’ ”, https://cran.r-project.org/web/packages/som/som.pdf, 閲覧日 2022.2.3

[40] ”KEGG API を用いてデータ取得”, https://rstudio-pubs-static.s3.amazonaws.com/472676_97a2c135b5704dc1b52f7759b73466e8.html#kegg-compound, 閲覧日 2022.12.28.

[41] “Supporting Information for: Evolving to an Ideal Synthesis of Molnupiravir, an Investigational Treatment for COVID - 19”, https://europepmc.org/api/fulltextRepo?pprId=PPR257265&type=FILE&fileName=EMS109513-supplement-Supporting_Information.pdf&mimeType=application/pdf, 閲覧日 2022.2.6