修士論文

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

富山県立大学 工学研究科 電子・情報工学専攻
2255018 武藤 克弥
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	反応物 i の特性値	RT_i
反応差分フィンガープリント	RFP	生成物 i の特性値	PD_i
クラスタ i	C_i	記述子 j の特性値変化量	cv_j
C_i に属するデータ集合	$\mathbf{x_i}$	i 番目の反応式が持つ特徴ベクトル	$\boldsymbol{DF}_i$
クラスタ間の距離	$d(C_1, C_2)$	記述子 u，v 間の相関係数	s_{uv}
クラスタ内の要素間の距離	$d(\mathbf{x_1}, \mathbf{x_2})$	記述子 u の特性値平均	$\bar{cv}_u$
個数	n	特徴量	f
次元数	p	決定木上位ノード内のデータ	D_p
p 次元観測ベクトル	$\boldsymbol{x}_j$	上位ノードに対する下位の左 (右) ノード	D_{left}, D_{right}
i 番目のユニット	m_i	決定木上位ノードのデータ数	N_p
ユニット数	k	上位ノードに対する下位の左 (右) ノード	N_{left}，N_{right}
i 番目ユニットの重心	$\boldsymbol{r}_i$	D_p を含むノードにおける情報利得	$IG(D_P, f)$
i 番目ユニットの重みベクトル	$\boldsymbol{\xi}_i$	ノード t におけるクラス i のデータの割合	$p(i\|t)$
$\boldsymbol{x}_j$ と $\boldsymbol{\xi}_i$ のユークリッド距離	$\|\|\boldsymbol{x}_j - \boldsymbol{\xi}_i\|\|$	ノード t 内のクラス数	c
$\|\|\boldsymbol{x}_j - \boldsymbol{\xi}_i\|\|$ を最小化する $\boldsymbol{\xi}_i$	$\boldsymbol{\xi}_c$	ノード t におけるジニ不純度	$I_G(t)$
$\boldsymbol{\xi}_c$ 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	$\sigma^2(t)$		

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，最適な酵素候補を探索する場合がある．その後，スクリーニングなどの実験によって，試行錯誤を繰り返しながら 1 つの酵素に絞っていく．ここで，酵素候補を探索する代わりに酵素候補を予測するシステムがあれば，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．

そのため，反応式を与えた際に，その反応を触媒するのに最適な酵素候補を予測するシステムを考える．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 [12] [13]．EC 番号に登録されている酵素には様々な生物由来のものが存在し，製品として開発されている．与えられた反応に対して，EC 番号を予測することで，その番号内の酵素からどの種類を選択するかという，次のステップに進むことができる．また，EC 番号には，そこに登録されている酵素を用いた，反応物から生成物に変化する際の代表的な反応が記載されている．

それらを踏まえて本研究では，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，それぞれの化合物の構造変化を比較し，類似性が最も高い反応の EC 番号を提示することで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，各化合物の構造を計算機上で扱うための表現法に変換する．そして，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，反応式の構造変化を表す特徴ベクトルとして定義する．その後，EC 番号の上位桁を予測する手法，および下位桁を予測する手法の 2 つを用いて予測を行う．上位桁の予測では教師あり学習を用いて分類モデルを作成し，ターゲットの番号を予測する．また，下位桁の予測では，クラスタリングによって反応式の特徴ベクトルを 2 次元平面上に出力し，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 番号を予測結果として提示する．この 2 つの予測手法を組み合わせて，最適な酵素候補を予測する．

§ 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 番号予測の研究例を述べる．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と実験結果，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

有機合成分野と情報分野の関わり

§ 2.1　有機合成と情報技術

有機合成では人工的に有機化合物を作り出すことを目的としている．古くから病の治療として天然の有機化合物が用いられており，薬として有効な成分のみを取り出すことが近年行われてきた．1805 年，F.W.A.Serürner がアヘンから強い麻酔作用を持つ morphine を取り出すに成功したことを皮切りに, 薬効成分が次々に抽出されるようになっていき，その発展とともに有機化学が発展していった [3]．また，1828 年には，Wähler が有機化合物として初となる尿素の合成に成功し，その後 Liebig を筆頭として，有機化合物の扱い方，構造式での構造理解が明確化されていった [4].

現在まで, 比較的簡単に入手できる化合物から，天然に存在する薬の成分などを人工的に生成する全合成によって，様々なものが合成されてきた．計算機が発達するようになると，実験結果で得られた情報がデータベースに蓄積されていき，化学の現象を計算機上で上手く表現することで，高速なデータ処理が可能となった．そして，情報技術によって，データベースからデータを分類・予測する分野としてケモインフォマティクスは現在まで発展してきた.

ケモインフォマティクスの研究分野として以下のものが挙げられている [5].

1. ケモインフォマティクス情報検索，データベース，グラフ理論，反応設計など
2. マテリアルズ・インフォマティクス構造物性相関など
3. バイオインフォマティクス
4. 計算機科学
5. 理論・計算科学 (量子科学，分子軌道法，分子科学)
6. コンビナトリアルケミストリー
7. 通信・システム (コンピュータネットワーク，並列化，専用機，コンピュータグラフィックスなど)
8. ラボラトリーオートメーション
9. 関連する化学教育・学習システム

化学分野では情報学に適用できる問題が多く存在する．例えば，化合物の構造に着目したとき，原子の部分を頂点，結合部分を辺とみなすことでグラフ理論の問題になる．合成時の反応経路の設計においては，どの化合物から出発し，いかにステップ数やコストなどを抑え，かつ効率的に目的物を生成していくかという最適化問題に帰着できる．例として，化

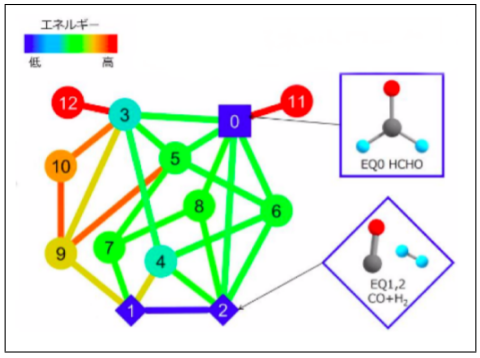


図 2.1: HCHO の化学反応ネットワーク [6]

学反応ネットワーク中の最適な反応経路に関する研究がある [6]．ここでは，化学反応における安定平衡構造を頂点，遷移状態構造を辺とした化学反応ネットワークにおける最短経路候補について検討している．莫大なパターンの経路を調べる代わりに，反応が経由するそれぞれの状態における最大・最小エネルギー差が小さい経路に絞り，合成経路設計，反応予測，逆合成解析の 3 つの場合において，K 最短経路問題などを適用して，計算機実験による探索の性能評価を行っている．図 2.1 は原子 H,C,H,O で構成される化合物の化学反応ネットワークを表している．

化合物を合成する過程においては，化学合成経路設計と，化学反応予測の 2 つのアプローチがある [7]．化学合成経路設計では，目的とする最終的な生成物を設定し，何の物質から出発して，どのような合成経路をたどって合成していくか，という逆合成的な手順を用いた手法である．化学反応予測は，出発物質を決め，目的の生成物を得るための反応は実際に起こるのか，副産物は何が生成されるのかといった手順をたどる．効率的に合成を進めるため，これらの手法を計算機上で行うためのシステム開発が行われてきた．化学合成経路を設計するシステムは，1970 年頃から多数開発されてきた一方で，化学反応の予測を行うシステムはあまり開発されてこなかった．それは，化学反応は様々な要因が複雑に絡み合うことで発生するため，反応の予測が困難であるためである．しかし，近年では計算機上での化学反応特性の表現方法や機械学習の発展によって，予測のハードルが下がりつつある．化学反応時に関わってくる要因を計算機側で上手く表現し，一部の要因に重点を置いて予測を行うことで，困難性を解消している．

研究例として，化学反応時の電子移動に関する，極性反応とラジカル反応について予測したものがある [8]．ここでは，1110 個の極性反応，103 個のラジカル反応からデータベースを作成し，機械学習の分類を行っている．分子の反応部位や構造情報，原子の複数の性質などを，量子科学計算によって数値化および特徴ベクトルとして表現し，10 分割交差検証によって，分類精度を評価している．

このように，化学反応や化合物における特定の特徴を数値化し，計算機上で扱いやすく，かつ機械学習に組み込みやすい形式を開発にすることによって，精度の良い反応予測を可能にしている．

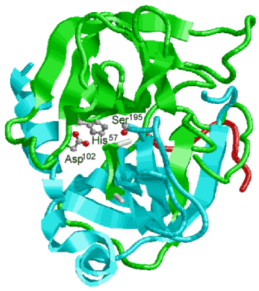


図 2.2: α-キモトリプシンの立体構造 [10]

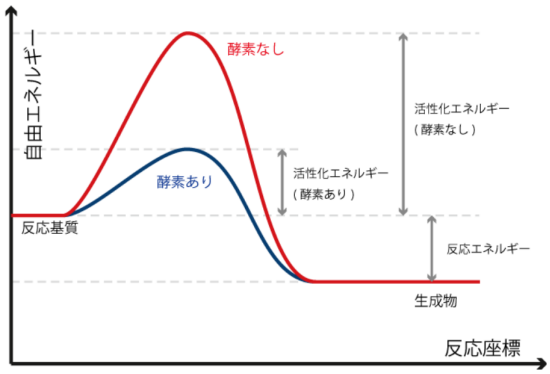


図 2.3: 反応の進行と必要なエネルギー [11]

§ 2.2　酵素と EC 番号

酵素は生体内で必要な化学反応を触媒するタンパク質で，生物が生きていくためには必要不可欠なものである．酵素には基質特異性という，特定の反応物 (基質) のみに触媒反応を示す特性を持っている．これは，Emil.H.Fischer が唱えた「鍵と鍵穴説」と呼ばれる，基質を鍵，酵素を鍵穴とする考え方が用いられている．基質が酵素に結合することで反応が始まり，基質が生成物へと変化すると結合が外れる．このとき，酵素自体は変化することなく元の状態に戻るため，触媒として繰り返し利用できる．酵素の構造イメージを図 2.2 に示す．

より多くの基質と結びついて作用する用途の広い触媒とするために，基質特異性を広げるタンパク質工学と呼ばれる分野がある．ここでは，アミノ酸配列の一部を置き換えることで酵素の性質を改変したり，ランダムに変異させた変異体ライブラリを作成し，スクリーニングによって所望の触媒機能をもったものを選択するといったことが行われている [9].

酵素を用いることのメリットとして以下のことが挙げられる．

反応速度の増加

酵素は生体触媒として，基質の化学反応をより早く，安定して進めることができる．化学反応において, 反応が進むにつれてエネルギーが増加し，遷移状態をピークに減少していく．この反応開始から遷移状態になるために，必要なエネルギーを活性化エネルギーと呼び，大きいほど反応が進みにくくなる．しかし，酵素を用いることで必要な活性化エネルギーが低下し，反応を速く進めることができる．酵素を用いた場合と用いなかった場合のエネルギー遷移の様子を図 2.3 に示す．

環境への影響の低減

通常，化学触媒は高温や高圧といった条件下で使用することが適している場合が多い．一方，生体触媒は常温，常圧で使用することができ，これらの条件下での反応であれば，高温・高圧にするためにかかるエネルギーの削減につながる．

高選択性

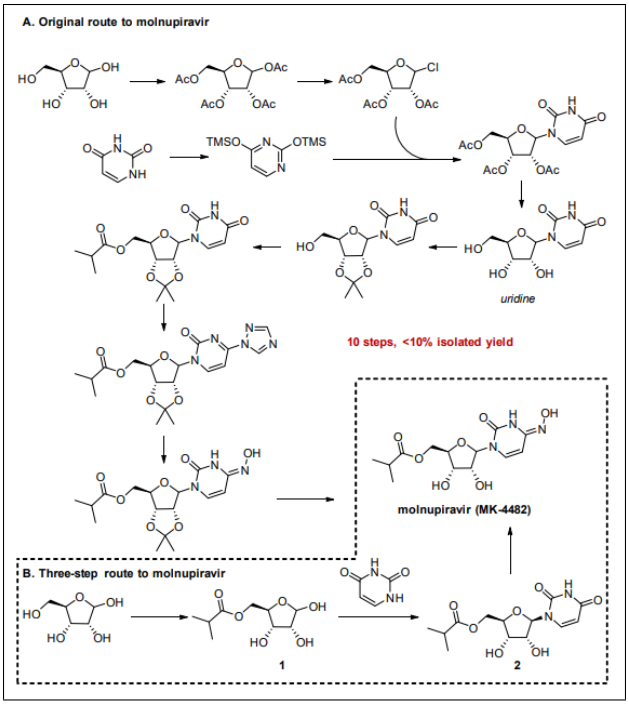


図 2.4: モルヌピラビルの合成

選択性とは，反応が起こりうる化合物の部位が複数ある中で，どれだけ特定の部位のみに反応するかの度合いを表す．選択性が高いほど，特定の構造を持つ部位のみを選んで，反応させることができるため，効率的な合成につながる．

上記の理由から，生体触媒を使って，医薬品を生成する事例が増えている．例として新型コロナウイルスの治療薬として，治験が進められているモルヌピラビル (MK-4482) の合成がある [2]．ここでは従来 10 ステップで行っていたものを，生体触媒を取り入れることによって 3 ステップまで短縮している．図 2.4 にモルヌピラビルの従来における合成方法と，提案された方法の比較を示す．合成ステップを短縮することは，使用する試薬などのコストが減り，結果として環境にも優しい．

酵素番号 (Enzyme Commission numbers：EC 番号)

酵素は EC 番号という，4 組の数字の組み合わせからなる番号で管理されており，酵素の性質ごとに分類されている．EC ○. ○. ○. ○というように番号が振られ，1 番目の数字はどの反応を触媒するかによって，1(酸化還元酵素),2(転移酵素),3(加水分解酵素),4(離脱酵素),5(異性化酵素),6(合成酵素),7(輸送酵素) の 7 つに分類されている [12] [13]．2 番目の数字では，どの結合に作用するか，3 番目の数字ではどの基質 (化合物) に反応するかや，必要とする補酵素情報というように分類され，4 番目の数字で 1 から 3 番目までの組み合わせ番号 (EC ○. ○. ○) に属する酵素の名前 (登録順) を表している．図 2.5 に EC 番号分類のイメージを示す．EC3(加水分解酵素) を例に見ると，エステル結合に作用する EC 3.1，グリコシド結合に作用する EC 3.2，・・・と分類されている [14]．さらに，EC 3.1 の下層に注目すると，カルボン酸エステルに作用する EC 3.1.1，チオエステルに作用する EC 3.1.2，・・・

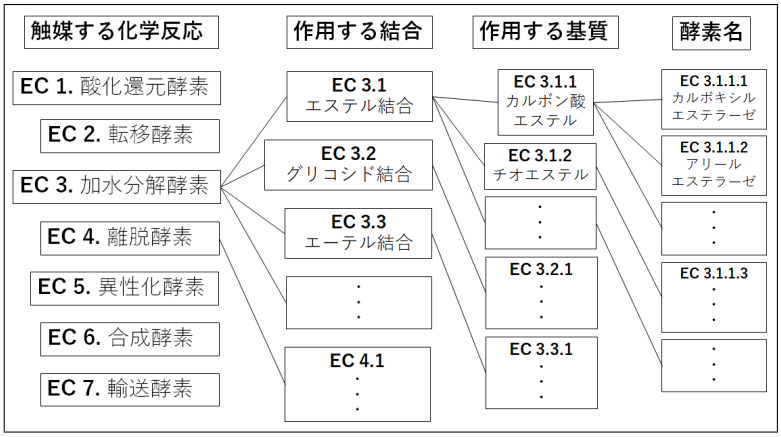


図 2.5: EC 番号による酵素の分類

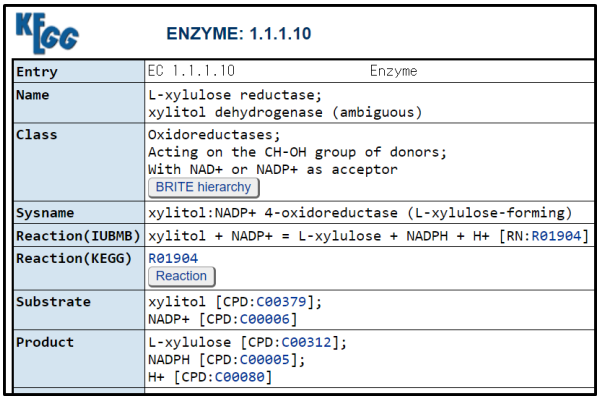


図 2.6: KEGG ENZYME データベース

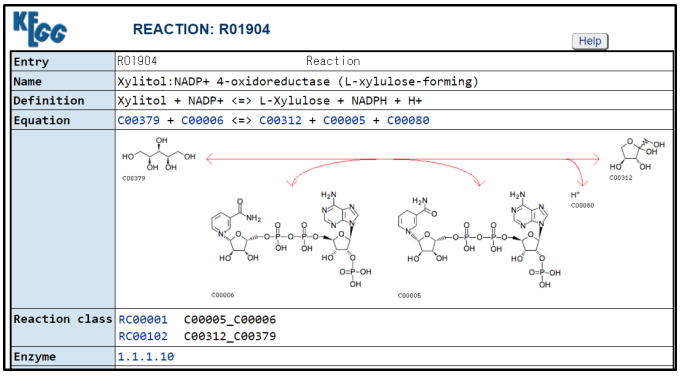


図 2.7: KEGG REACTION データベース

と分かれ，EC 3.1.1 からは EC 3.1.1.1 のカルボキシルエステラーゼ，EC 3.1.1.2 のアリールエステラーゼ，・・・と分類されている．

§ 2.3　化学・酵素データベース

化学・生物分野において用いられているデータベースについて，いくつか説明する．

Kyoto Encyclopedia of Genes and Genomes(KEGG) [15]

遺伝子・タンパク質情報，タンパク質相互作用を可視化した KEGG PATHWAY，酵素情報を表した KEGG ENZYME，主に酵素反応の反応式について記した KEGG REACTION，生態系に関連する化合物を集めた KEGG COMPOUND 等のデータからなるデータベースである．KEGG ENZYME では各酵素の情報を該当する EC 番号から検索して得ることができ，酵素の別名，その酵素を用いた生体内の反応式，基質・生成物情報，遺伝情報，文献情報等について書かれている．KEGG REACTION には酵素を用いて起こる化学反応についての情報を記している．それぞれの反応は R から始まる 5 桁の数字で管理されており，反応に用いられる酵素と EC 番号，化合物名・C 番号・構造式でそれぞれ表した反応式等が書かれている．KEGG COMPOUND では C から始まる 5 桁の C 番号で化合物を管理して

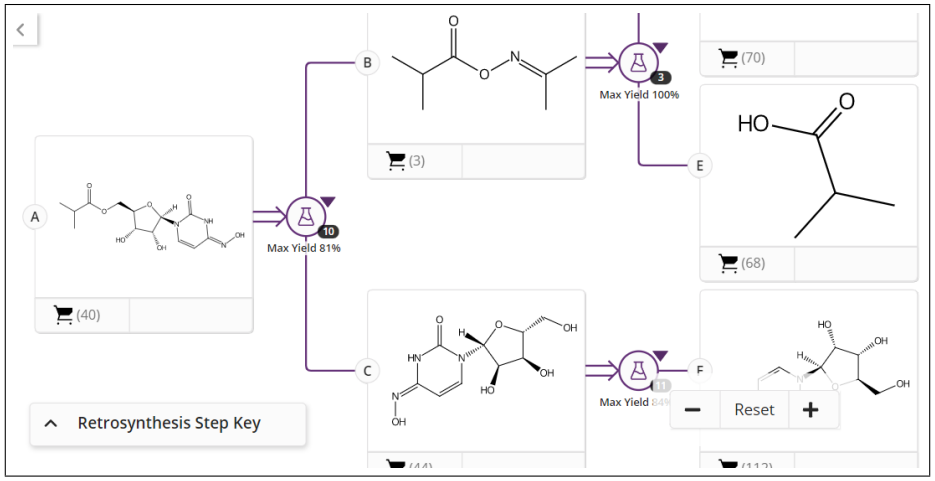


図 2.8: SciFinder^n による逆合成設計予測

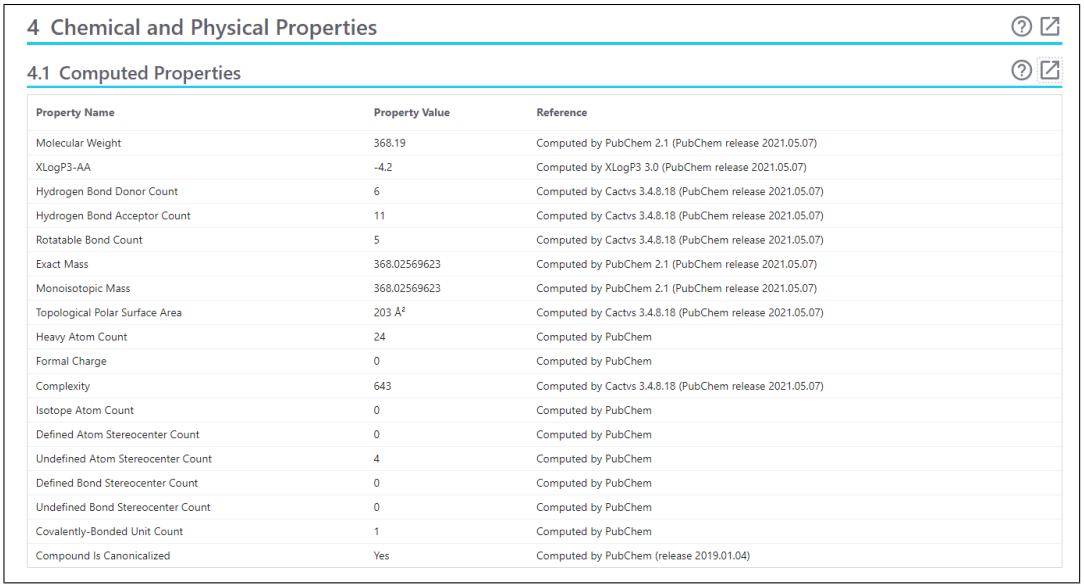
おり，主に KEGG PATHWAY 中や KEGG REACTION 中に現れる化合物を扱っている．また，C 番号，名前，分子式で検索することができ，そのリンク先には，別名，分子式，分子量，構造式，登場する R 番号，PATHWAY MAP の MAP 番号，EC 番号のリンク先，他のデータベースへのリンク先などが掲載されている．サイト内のリンクのつながりによって，EC 番号から R 番号，R 番号から C 番号とたどることができる．図 2.6 および図 2.7 に KEGG データベースの例を示す．

SciFinder^n [16]

Chemical Abstracts Service(CAS) が提供する，データベース．主に，「Substances(化学物質情報)」,「Reactions(反応情報)」,「References(文献情報)」,「Suppliers(カタログ情報)」,「Biosequences(配列情報)」の項目から検索することができる．

「Substances」では化学物質の名前，CAS 登録番号，分子式やスペクトル，物性値などで検索できる [17].「Reactions(反応情報)」では化学物質名，構造式などから検索され，その化合物が反応式・生成物として用いている反応式を調べることができる．また，生成物の収率，反応に用いる試薬や溶媒，文献情報などを条件に入れてフィルター検索もできる．「References」ではキーワード，著者名，文献番号，雑誌情報，機関名などで検索される．「Suppliers(カタログ情報)」では，検索した化合物を取り扱う企業などのカタログ情報を取り扱っている．検索結果には取扱業者名と純度情報，化合物の購入サイトへのリンクと取り扱い分量等が表示される．「Biosequences」では DNA,RNA, タンパク質の配列情報や類似する配列などで検索される．

SciFinder^n では，構造式をユーザ自信が描画・編集して検索することが可能である．化合物の構造が一致するもの，または構造の類似度に基づいて検索できる他，関連する反応式，文献情報，提供元の情報も参照できる．さらに，作成した構造式を生成物として，逆合成ルートを設計・予測する「Retrosynthesis Planner」というツールが存在する．ここでは合成ステップ数やコストなどを設定し，複数パターンの合成プランが設計される．各合成ルートは既知の反応または，予測された反応で構成され，最大の収率が表示される．図 2.8 にモルヌピラビルをターゲットとして，逆合成ルートの設計・予測をした様子を示す．



4 Chemical and Physical Properties

4.1 Computed Properties

Property Name	Property Value	Reference
Molecular Weight	368.19	Computed by PubChem 2.1 (PubChem release 2021.05.07)
XLogP3-AA	-4.2	Computed by XLogP3 3.0 (PubChem release 2021.05.07)
Hydrogen Bond Donor Count	6	Computed by Cactvs 3.4.8.18 (PubChem release 2021.05.07)
Hydrogen Bond Acceptor Count	11	Computed by Cactvs 3.4.8.18 (PubChem release 2021.05.07)
Rotatable Bond Count	5	Computed by Cactvs 3.4.8.18 (PubChem release 2021.05.07)
Exact Mass	368.02569623	Computed by PubChem 2.1 (PubChem release 2021.05.07)
Monoisotopic Mass	368.02569623	Computed by PubChem 2.1 (PubChem release 2021.05.07)
Topological Polar Surface Area	203 Å²	Computed by Cactvs 3.4.8.18 (PubChem release 2021.05.07)
Heavy Atom Count	24	Computed by PubChem
Formal Charge	0	Computed by PubChem
Complexity	643	Computed by Cactvs 3.4.8.18 (PubChem release 2021.05.07)
Isotope Atom Count	0	Computed by PubChem
Defined Atom Stereocenter Count	0	Computed by PubChem
Undefined Atom Stereocenter Count	4	Computed by PubChem
Defined Bond Stereocenter Count	0	Computed by PubChem
Undefined Bond Stereocenter Count	0	Computed by PubChem
Covalently-Bonded Unit Count	1	Computed by PubChem
Compound Is Canonicalized	Yes	Computed by PubChem (release 2019.01.04)

図 2.9: PubChem の例

PubChem [18]

化合物名，分子式，化合物の 2D(もしくは 3D) 形式の構造イメージ，化学・物理特性，生物学的活性情報，毒性情報，文献情報等のデータを収録している．データ提供者からアップロードされた，約 2.8 億種の化学物質情報や約 140 万種の生物学的実験データ，標準化された約 1.1 億種の化学構造情報，また，約 10 万種の遺伝子データなどから構成される [19]．さらに，PubChem Compound，PubChem Substance，PumChem BioAssay の 3 つのデータベースがある．

PubChem Substance では，研究者がアップロードしたデータを管理している．複数の提供者から重複するデータがアップロードされることがあるため，標準化によって，同様の情報を集約し，PubChem Compound に格納される [20]．また，PumChem BioAssay では，データ提供者の実験環境によってばらつきが生じる生物活性データ等を，実験に用いられた化合物と，実験結果ごとに紐づけを行うことで管理している．それぞれのデータベース中のデータには，SID(SubstanceID)，CID(CompoundID), AID(AssayID) が割り振られている．特に SID は KEGG のほとんどの C 番号と対応している．図 2.9 に PubChem のデータベースの例を示す．

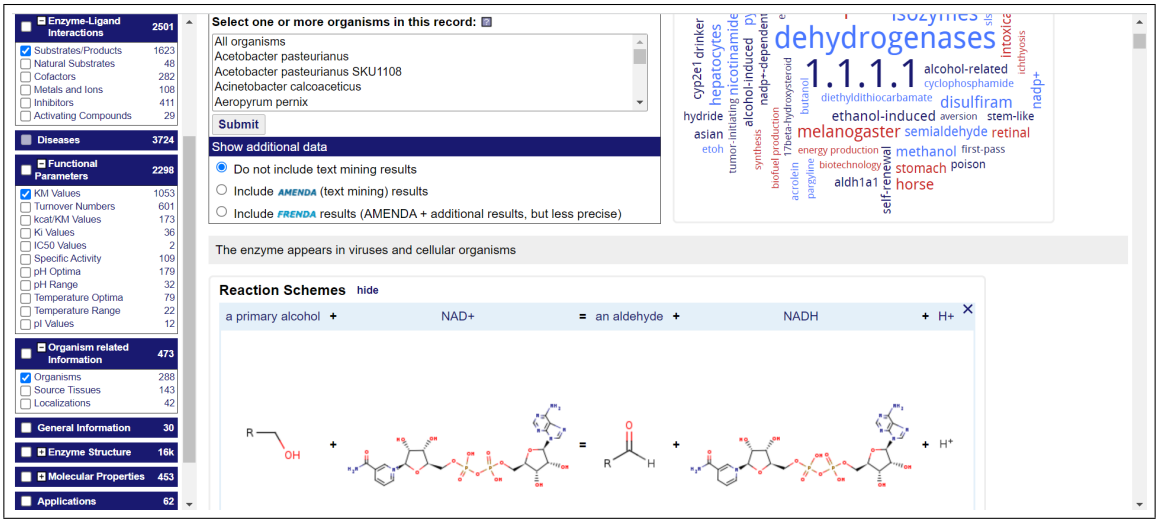


図 2.10: BRENDA 上の EC1.1.1.1 に関する情報

BRENDA [21]

酵素に関するデータを，文献の情報をもとに網羅したデータベース．酵素名，生物種，CAS 登録番号，EC 番号，特性値などで検索することができる．例として，EC 番号で検索した様子を図 2.10 に示す．検索した EC 番号のページに行くと，その酵素に関係している単語のワードマップや用いられている反応式が書かれている．図 2.10 の画面左にある画面から目的とする詳細情報を表示できる．例えば，Substrates/Products では，検索した EC 番号の酵素を使った反応の基質・生成物のペアが記されている．Organisms では，酵素を作る由来となった生物種のリストが表示されている．また，「Functional Parameters」ボックス内の KM Values では，酵素の由来となった生物種・基質ごとの Km 値 (基質と酵素の親和性を表す指標) を見ることができる．

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 [22], PubChem では POWER USER GATEWAY(PUG) [19] 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，図 3.1 のようになる [22]．<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，図 3.2 のフォーマットに従う．

ここでは，<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 3.3 にその対応リストを示す．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 [23]．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

```
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
```

図 3.3: マップ番号と Pathway 名の対応リスト

```
<PCT-Data>
  <PCT-Data_input>
    <PCT-InputData>
      <PCT-InputData_download>
        <PCT-Download>
          <PCT-Download_uids>
            <PCT-QueryUids>
              <PCT-QueryUids_ids>
                <PCT-ID-List>
                  <PCT-ID-List_db>pccompound</PCT-ID-List_db>
                  <PCT-ID-List_uids>
                    <PCT-ID-List_uids_E>1</PCT-ID-List_uids_E>
                    <PCT-ID-List_uids_E>99</PCT-ID-List_uids_E>
                  </PCT-ID-List_uids>
                </PCT-ID-List>
              </PCT-QueryUids_ids>
            </PCT-QueryUids>
          </PCT-Download_uids>
          <PCT-Download_format value="sdf"/>
          <PCT-Download_compression value="gzip"/>
        </PCT-Download>
      </PCT-InputData_download>
    </PCT-InputData>
  </PCT-Data_input>
</PCT-Data>
```

図 3.4: PUG における XML 応答の例 [19]

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-REST は，PUG や PUG-SOAP で用いられている XML 形式の記述を必要とせず，簡単な記述法でデータを取得することができる API である．PUG-REST のリクエストは図 3.5 のような URL で表記される [19]．<input specification>はさらに<domain>/<namespace>/<identifiers>で構成されており，何のデータを取得するのかを定める．<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>において，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

```
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 3.5: PUG-REST のリクエスト

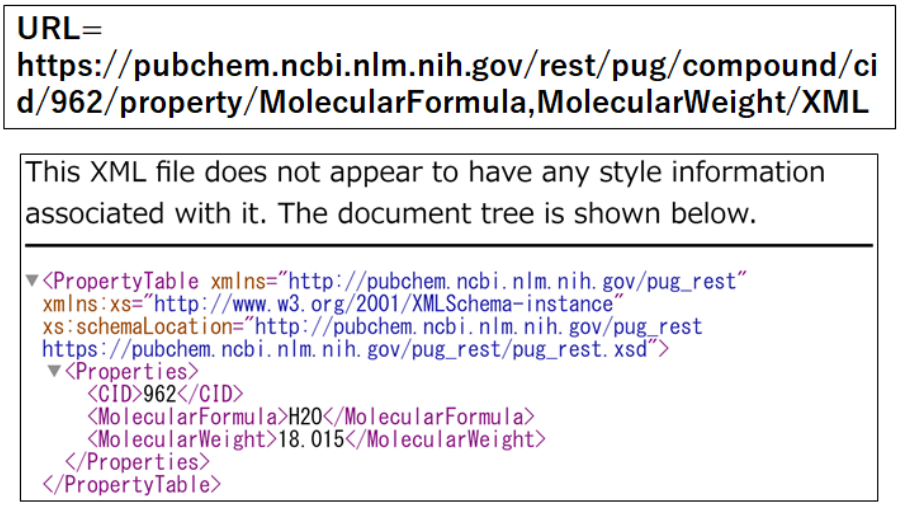


図 3.6: 水 (H_2O) の情報を取得するリクエスト URL とデータ取得結果

取得することができる．

<output 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 3.6 に示す．

§ 3.2　化合物の構造表現法と EC 番号予測手法

ケモインフォマティクスでは，化合物の構造を計算機上において扱いやすい形式で表現し，化学反応の予測や分類などを行う．ここでは，その表現法や用いられているライブラリの一部について説明する．また，EC 番号予測に関する研究についても述べる．

MOL ファイル・SDF 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

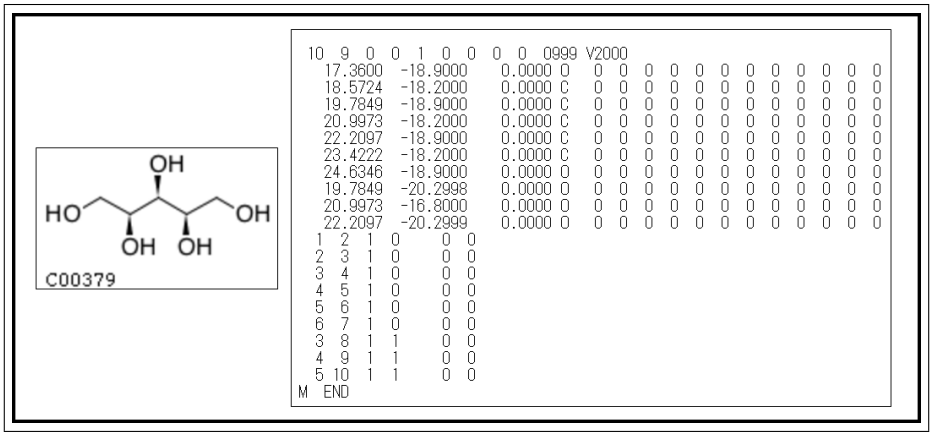


図 3.7: KEGG COMPOUND で取得できる構造式と MOL ファイルの例

のを図 3.7 に示す．複数の化合物 MOL ファイルを統合したものは，拡張子「.sdf」からなる SDF ファイルとなる．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，SDF ファイルとなることが多い．

SMILES

化合物の構造を文字列で表したもの．以下の規則に従い，化学構造を文字列に変換していく [24].

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．

これらに加えて，さらに以下の規則を加えた isomeric SMILES を本研究では用いる．

8. 同位体 (例えば炭素) がある場合 [13C] という表記にする
9. 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
10. 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で，化合物同士の類似性比較などに用いられる．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 [25] 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．

化合物の数値化 (特徴ベクトル)

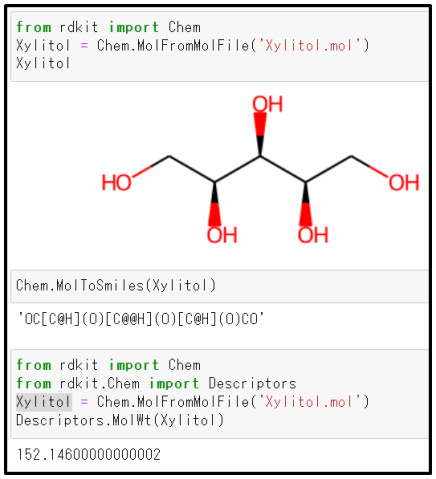


図 3.8: rdkit を用いた化合物の情報

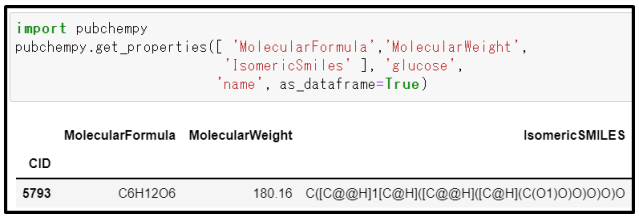


図 3.9: PubChempy でグルコースの情報を取得した結果

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・化学特性値を用いることもある．一般的に複数の特性値が用いられ，多次元の特徴ベクトルとして化合物の特徴を表現する．これらの化合物を数値で表現する特徴を，記述子と呼ぶことがある．

RDKit を用いた化合物のデータ化

RDKit [26] 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 ファイルを読み込み，構造式を描画する構造式オブジェクトを出力したり，SMILES やフィンガープリントに変換することができる．また，構造式オブジェクトから，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，化合物の分子量を表す記述子 MolWt を知りたい場合，Descriptors クラスにある MolWt メソッドに生成した構造式オブジェクトを渡すことで，MolWt が計算され出力される．図 3.8 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 [27]

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 3.9 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 R 番号として表記されており，図 3.10 のような化学反応となる．これらの反応式に対して

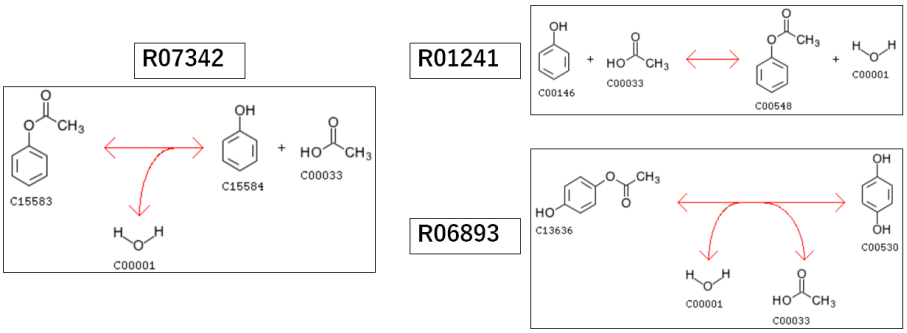


図 3.10: EC3.1.1.2 の代表的な反応式

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し，EC 番号を予測する研究が行われている．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 [28]，構造として特徴的な部分の化学変化に注目したもの [29] 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) と，その近傍の原子で異なっている領域 (D atom)，および一致している領域 (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．クラスタリングは教師なし学習の一つで，特定の基準に従って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 [30].

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．

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

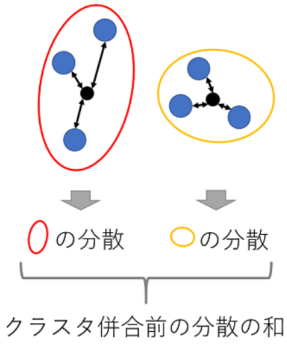
ウォード法

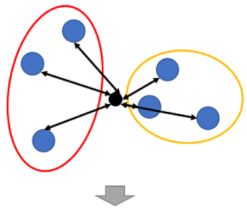
あらかじめ 2 つのクラスタを結合し，結合したクラスタ内の重心に対するデータの分散 $E(C_1 \cup C_2)$ に対して，結合前の各クラスタ内のデータの分散 $E(C_i)$ を引いた差が，最小となるクラスタのペアを結合する方法．計算量は多くなるものの，分類感度が良いとされ，階層的クラスタリングで最も用いられている．ウォード方のイメージを図 3.11 に示す．クラスタ C_1 の重心を $\boldsymbol{c_i}$ として，以下のように表される．

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

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

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





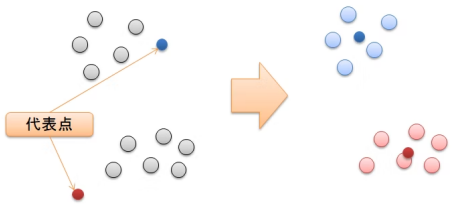


図 3.12: k-means 法のイメージ [32]

図 3.11: ウォード法のイメージ [31]

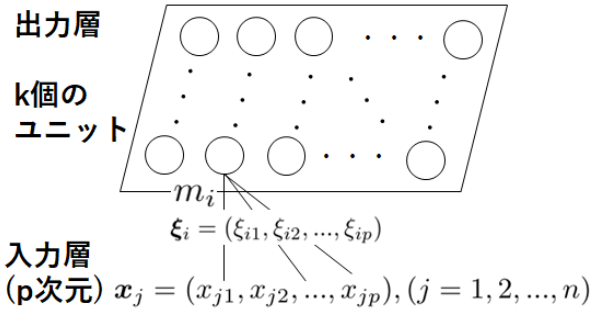


図 3.13: SOM における入力データのマッピング

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 個のクラスタに，データをランダムに割り振る
2. 各クラスタ内のデータに対する重心を計算し，それぞれの重心に対して最短距離にあるデータが，全て同じクラスタに属するように，再クラスタリングする．
3. クラスタ内のデー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) [33]

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 [34]．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 3.13 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)

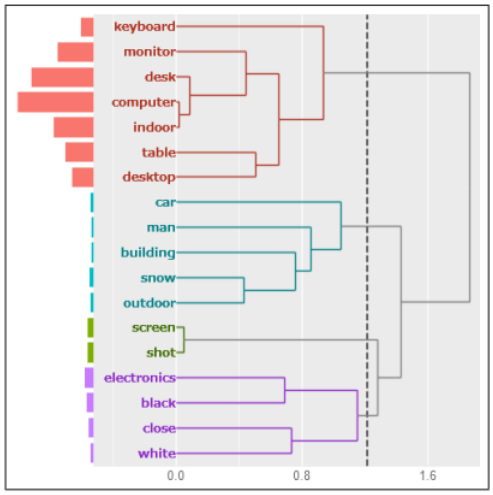


図 3.14: 凝集型クラスタリングの行動識別

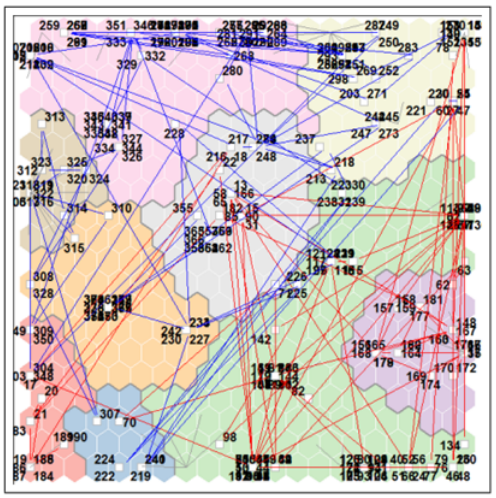


図 3.15: SOM を用いた行動時系列分析

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

クラスタリングを応用した研究として，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 [35]．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 [36] を用いて，物体名の同時出現頻度に関して，凝集型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．凝集型クラスタリングと SOM を用いた行動分析の様子を図 3.14 および図 3.15 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．酵素を分類しているEC番号には，その酵素を用いた自然界でみられる代表的な反応が，反応式として記載されている．また，EC番号に登録されている酵素には様々な生物由来のものが存在し，製品として開発されている．これらのことから，EC番号を予測することでスクリーニングの酵素候補を，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，逆合成解析で，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 [37] に基づいている．

特性値変化量の導入

反応物から生成物への構造変化の指標として，物性値・化学特性値の変化量 (特性値変化量) を用いる．3.2説で述べたように，構造変化に着目してEC反応式を分類する研究は多く行われている．その1つにフィンガープリントを用いたものがある [28]．ここでは反応物・生成物の部分構造に関するフィンガープリントを求め，差分を取ることで部分構造の変化

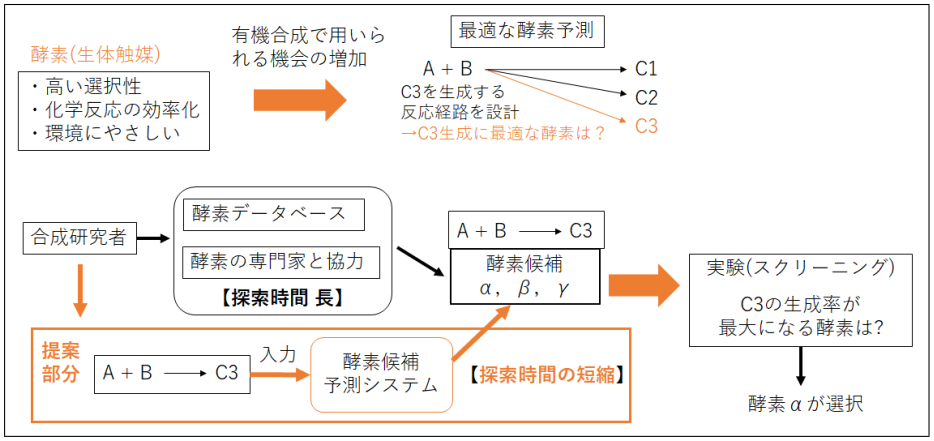


図 4.1: 従来の酵素探索と提案する酵素探索の比較

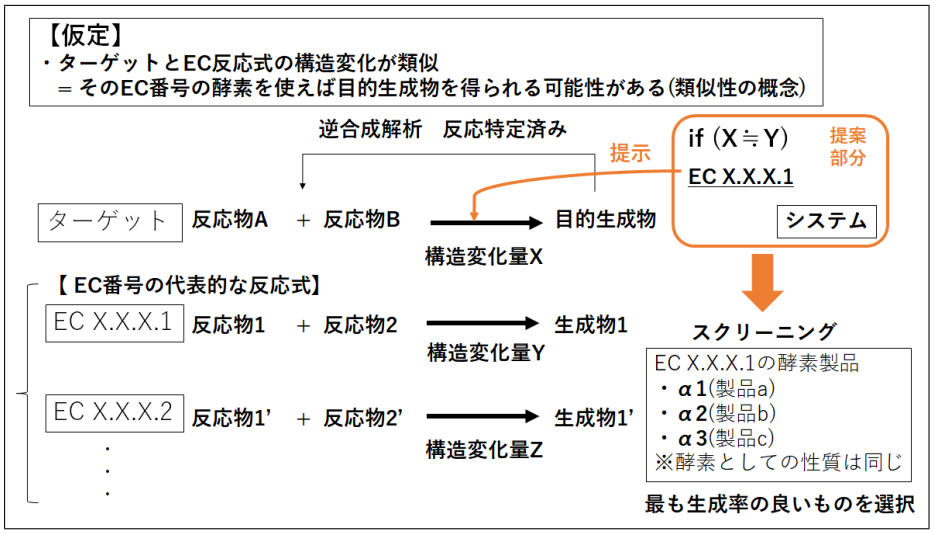


図 4.2: 反応式の類似性比較

を表現している．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，化合物の構造を表現する指標として，用いることができると考えられる．また，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．このことから，RDKit 記述子によって求めることができる多数の特性値変化量を用いて，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．

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 $cv_j(j = 1, 2, ..., n)$ を以下のように定義する．

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

各反応式に対して，n 種の特性値変化量を要素に持つ n 次元特徴ベクトルを，反応式の構造変化の特徴として用いる．m 個の反応式に対する特徴ベクトル $\boldsymbol{DF}_i(i = 1, 2, ..., m)$ を以下

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}

のように表す.

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

これらをもとに，表4.1のような $m \times n$ 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「target」と EC 反応式の EC 番号，列ラベルは記述子名となる．

特徴ベクトルの取得方法

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を構成する各化合物の ID を取得する [44].

PubChem Compound に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．

PubChem では，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．その後，SMILES 形式の化合物を RDKit で読み込み，化合物の構造式オブジェクトに変換する．同様に RDKit で実装されている，記述子 208 種を用いて特性値変化量を計算し，各反応式に対して 208 次元の特徴ベクトルを取得する．

2 つの EC 番号予測手法

今回は EC 3 クラス内の予測を目的とし，2 つの手法を用いる．1 つは，EC 3 クラスの 4 桁目の予測手法，もう 1 つは，2,3 桁目の予測手法である．予測手法を 2 つに分けた理由と

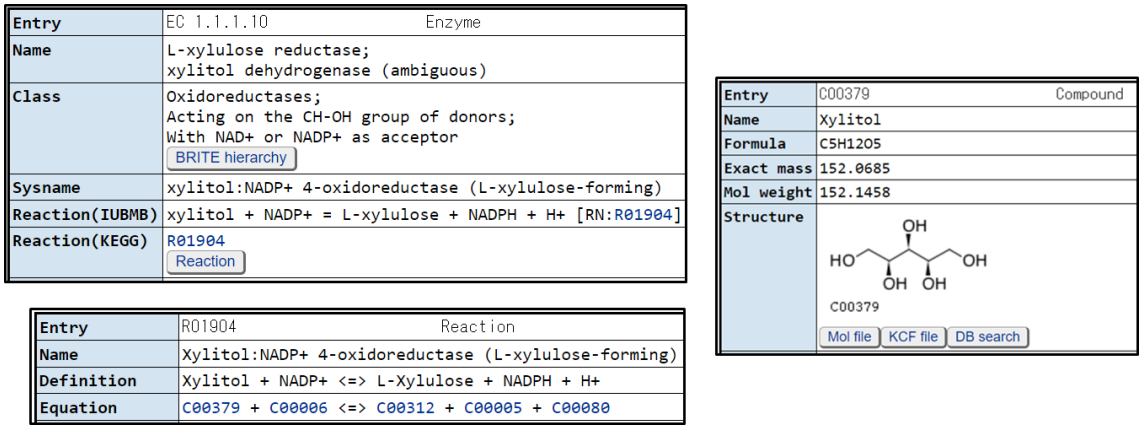


図 4.3: EC 番号，R 番号，C 番号の参照 [15](一部抜粋)

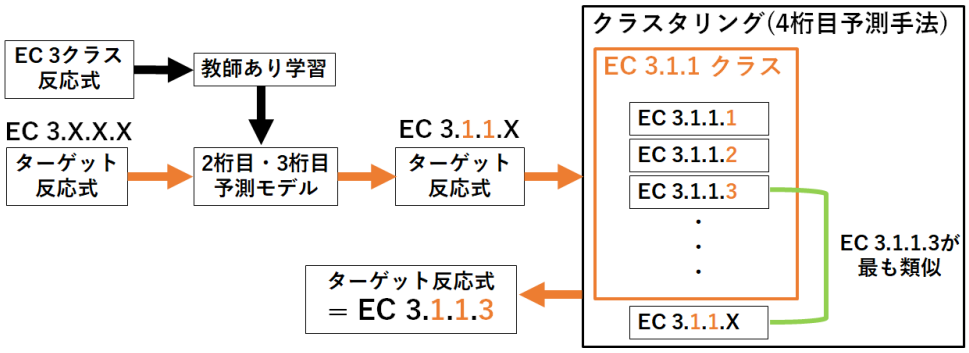


図 4.4: EC 3 クラスの予測手順

して，4 桁目の予測に用いることができる学習データが少ないことが挙げられる．KEGG ENZYME では，1 つの EC 番号に複数の代表的な反応式が登録されている場合もあるが，反応式が 1 つのみの EC 番号が多く，学習データとして用いることができない．一方で，2,3 桁目においては，各クラスに含まれるデータ数が増加するため，学習データを十分に準備することができる．そのため，教師あり学習を用いて 2,3 桁目の予測モデルを作成し，4 桁目の予測では教師なし学習を用いた類似性によるクラスタリングを行う．これら 2 つの手法を用いて，ターゲット反応式に最適な EC 3 クラスの 3 桁目までを予測したのち，4 桁目の予測を行う．図 4.4 にその流れを示す．

本研究では，初めに EC 番号 4 桁目の予測についての数値実験を行ったのち，2,3 桁目の予測を行う．2 節，および 3 節で各予測手法について述べる．

§ 4.2　EC 番号 4 桁目に対する予測手法

EC 3 クラスの 4 桁目の予測においては，ターゲット反応式と各 EC 番号反応式の特徴ベクトルの類似性を評価し，評価が高い EC 番号反応式の EC 番号を最適な酵素候補として提示する．類似性の評価方法としては SOM を用いるが，前処理として凝集型クラスタリングによる次元削減を行う．

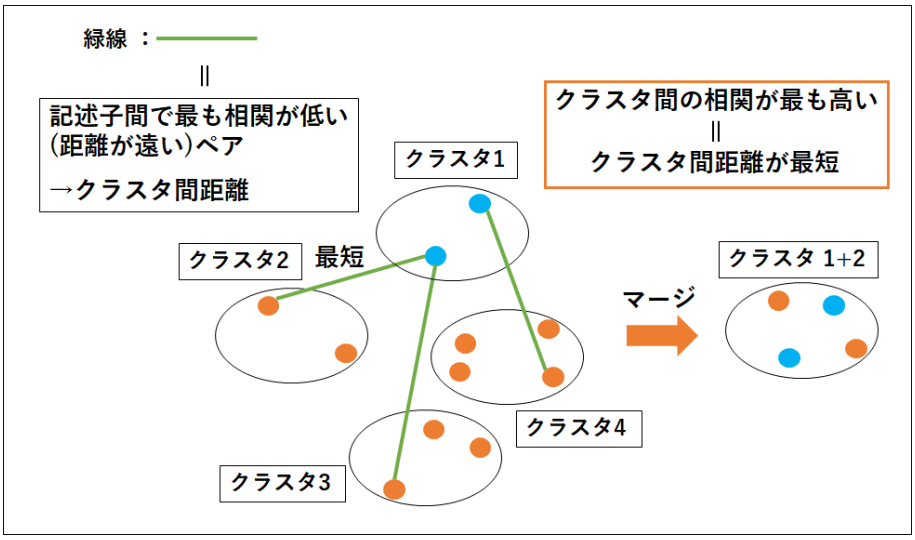


図 4.5: 最長距離法による記述子のクラスタリング

次元削減のための凝集型クラスタリング

多次元の特徴ベクトルを用いる際，次元サイズの大きさ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あり，次元の呪いは，用いる特徴量が多い場合に起こる．いずれも，汎化性能や分類精度の低下の原因とされている．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，他の記述子に比べて，それらの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，記述子間で凝集型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 [38]．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を行うため，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．少なくともその中で最も相関の高いペアを持つクラスタどうしは，クラスタ間での相関が最も高い関係と考えられる．よって，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．その後，次元削減として，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．それにより，相関の高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型クラスタリングである，AgglomerativeClustering ライブラリ [39] 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．

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	$\cdots$	記述子 n
記述子 1	0	$1/s_{12}$	$\cdots$	$1/s_{1n}$
記述子 2	$1/s_{21}$	0	$\cdots$	$1/s_{2n}$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
記述子 n	$1/s_{n1}$	$1/s_{n2}$	$\cdots$	0

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

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 $s_{uv} \geqq 0.9$ すなわち，$1/s_{uv} \leqq 1/0.9 \approx 1.11$ 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，最長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．

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

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. $1/s_{uv} \leqq 1.11$ 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. $1/s_{uv} \leqq 1.11$ 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における，特性値変化量の列に対し標準化，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．

SOM による反応式のクラスタリング

次元削減した特徴ベクトルの類似性に基づいて，ターゲット反応式，および EC 反応式を SOM によってクラスタリングする．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

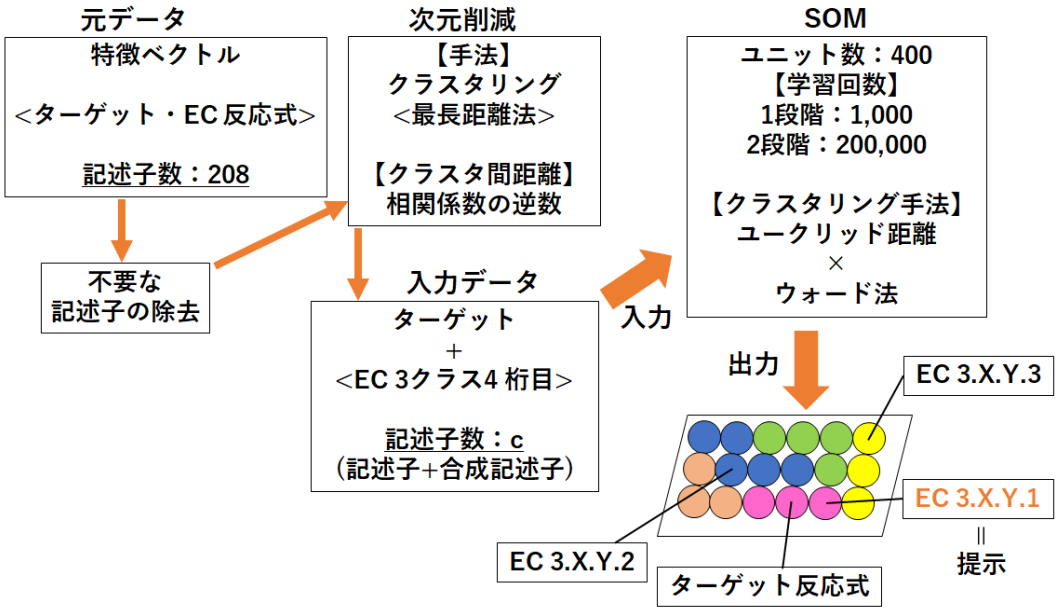


図 4.6: 4 桁目予測手法の流れ

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 EC 反応式どうしの類似性も見ることができるようになる．

SOM のプログラムは，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 言語のソースコードを用いる [35]．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 [40]．プロット点のラベルはターゲットを表す「target」と 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．

予測の流れ 1

EC 3 クラスの 4 桁目の予測は，次に述べる手順で行う．初めに，ターゲット反応式と EC 反応式の特徴ベクトルを，合わせて表 4.1 の形式で求める．行ラベルには，ターゲットを表す「target」と，ターゲットが所属すると予測される，EC 3.X.Y クラスにおける EC 反応式の「4 桁目番号」，列ラベルには，208 種の記述子名が入る．次に，不要な記述子を除去したうえで，凝集型クラスタリングによる次元削減を行い，合成された記述子をまとめて 1 つのラベル「cluster X」で表す．各反応式の合成記述子における，特性値変化量は，クラ

スタ内の記述子の特性値変化量を標準化・平均化したものとなる．その後，SOM によって反応式をクラスタリングし，ターゲットと同クラスタ内であり，かつ最も近くに位置する EC 反応式の EC 番号を，最適な酵素候補として提示する．図 4.6 に上記手順のイメージを示す．

§ 4.3　EC 番号 2 桁目・3 桁目に対する予測手法

EC3 クラスの 2 桁目・3 桁目の予測においては，EC 3.X.Y における X，Y の組み合わせからなる，各クラスの EC 反応式を分類する予測モデルを作成し，ターゲット反応式が分類される EC 3.X.Y を最適な EC 番号として提示する．EC 3.X.Y を予測したのち，4 桁目の予測に移るような流れとなる．ここでは，ランダムフォレスト (Random Forests) をモデル作成に用いるが，特徴選択も同時に行い，必要となる記述子を選択する．

ラッパー法を用いた記述子選択

必要以上に記述子を用いることは，構造変化を分類する際の汎化性能低下につながるため，適切な種類の記述子のみに限定することが望ましい．ラッパー法 (Wrapper Method) では分類・回帰モデルの予測精度を評価し，最も評価の高い特徴の組み合わせを選択するため，今回の記述子選択に適している．

ラッパー法には，指定した特徴数まで特徴を 1 つずつ追加し，追加するたびに最も評価の高い組み合わせを選択する Step Forward 法，全ての特徴を選択した状態で，評価が最も高い組み合わせとなるよう，指定した特徴数まで特徴を 1 つずつ削減していく Step Backwards 法，全ての組み合わせを探索し，最高評価の組み合わせを選択する Exhaustive 法がある．今回は，Step Forward 法を適用し，Python の mlxtend ライブラリ内にある，SequentialFeatureSelector に実装されている，Sequential Forward Selection(SFS) を用いる [41]．特徴選択の手順は以下のようになる．

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，記述子を選択する．
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する．
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する．
5. 指定した特徴数になるまで 3 と 4 を繰り返す．

モデルの評価基準としては，各 EC 番号クラス分類時の F1 スコア平均を用いる．

ランダムフォレストによる EC 番号分類

ランダムフォレスト (Random Forests) は複数の決定木を用いる機械学習モデルであり，モデルの表現力は高いが，過学習に陥りやすい決定木を組み合わせることで，汎化性能を高めることができる．決定木のイメージを図 4.7 に示す．決定木ではまず根のノードに特徴量を割り当て，特徴量の閾値に応じてデータを下の 2 つのノード内に分割する．それ以降

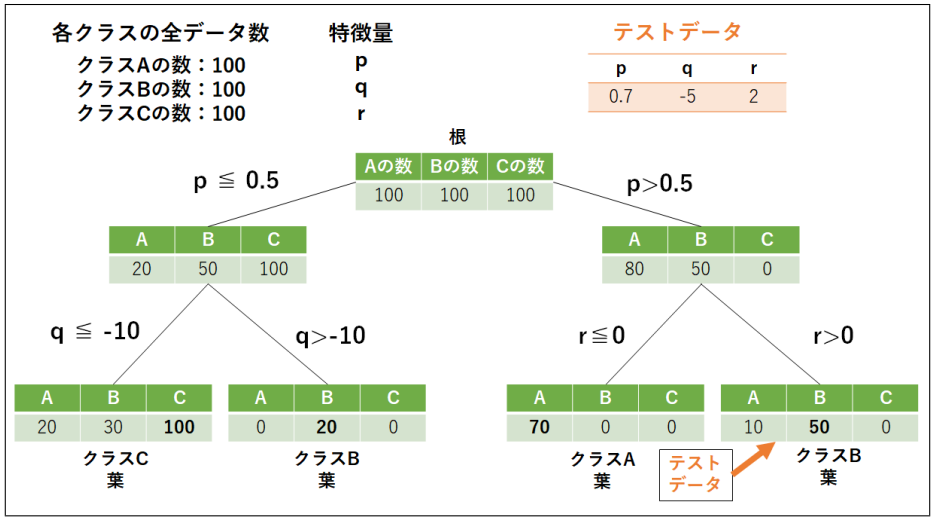


図 4.7: 決定木におけるクラス分類の様子

は，根の下にある各ノードに対しても特徴量を割り当て，あるノードに到達したデータを特徴量の閾値に応じて，下位ノードに 2 分割する工程を繰り返していく．

ノードを分割する際の分割基準と用いる特徴量は，以下で定義される情報利得 IG によって決まる [42]．

$$IG(D_P, f) = I(D_p) - \frac{N_{left}}{N_p} I(D_{left}) - \frac{N_{right}}{N_p} I(D_{right}) \tag{4.5}$$

f は分割時に用いられる特徴量，D_p は上位ノード内のデータ，D_{left}，D_{right} はそれぞれ分割先の下位ノード内のデータを表し，N_p，および N_{left}，N_{right} は上位ノード，下位ノードのデータ数を表す．I は不純度を表し，ジニ不純度，エントロピーなどが用いられる．今回はジニ不純度を用い，ノード t に対するジニ不純度 $I_G(t)$ は以下のようになる．

$$I_G(t) = \sum_{i=1}^{c} p(i|t)(1 - p(i|t)) = 1 - \sum_{i=1}^{c} p(i|t)^2 \tag{4.6}$$

$p(i|t)$ はノード t におけるクラス i のデータの割合，c はノード t 内のクラス数を表す．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が決定される．

全ての葉ノードで $I_G(t) = 0$，すなわちデータの属するクラスが 1 種類となるまで，分割が続行されるが，過学習を避けるため，決定木の最大深さを設定して途中で打ち切ることが多い．分割が停止したとき，葉に含まれるデータのクラス割合が，最も大きいクラスを葉ノードのクラスとして決定される．テストデータは葉ノードのいずれかに分類され，葉ノードのクラスとして予測される．

ランダムフォレストでそれぞれの決定木を作成する際には，全データから一部のデータを復元抽出し，決定木に入力される．また，特徴量も全てではなく，一部のみ選択されて決定木に用いられる．それによって，異なる決定木モデルを表現する．

Python の sklearn ではパラメータとして，決定木の数 n_estimators や，各決定木の最大深さ max_depth，用いる特徴量の数 max_features などを設定する．また，テストデータの予測クラスは，それぞれの決定木で分類された，葉ノードの各クラス確率の推定値に対す

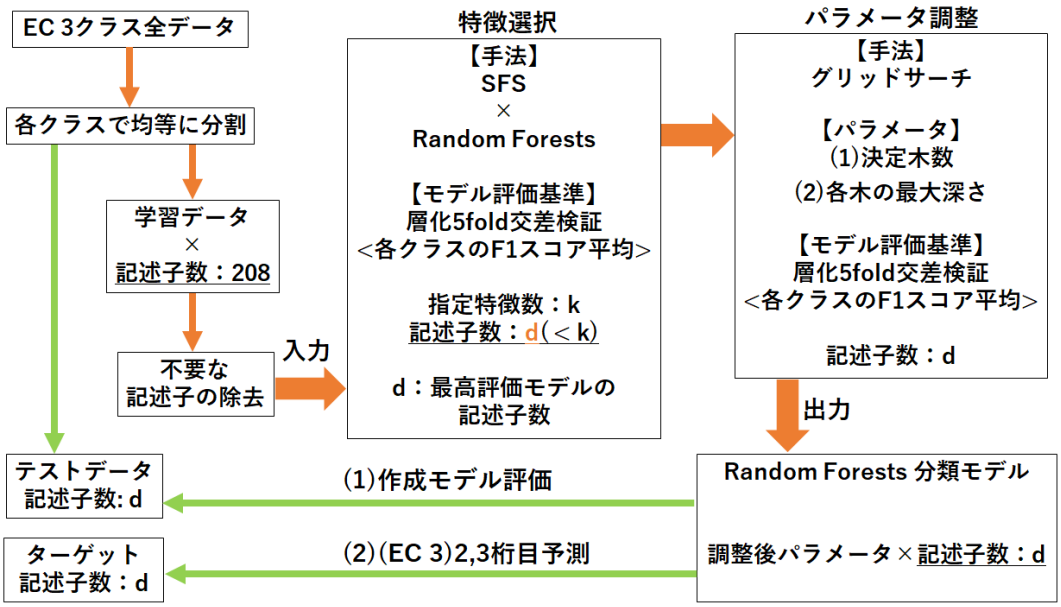


図 4.8: 2,3 桁目予測手法の流れ

る平均値が最も高いクラスとなる [43]．例えば，ある 1 つのテストデータが 3 つの決定木において，クラス A,B,C の確率がそれぞれ (0.10, 0,40, 0.50)，(0.00, 0.10, 0.90), (0.20, 0.50, 0.30) となる葉ノードに分類された場合，平均値は (0.10, 0.33, 0.57) となり，最終的にクラス C に分類される．

予測の流れ 2

上記で述べた内容を EC 番号の分類に適用する．各データはラベルに EC 3.X.Y が割り当てられている，208 次元の特徴ベクトルとし，特徴量はその 208 種の記述子とする．初めに，不要な記述子を削除したのち，ランダムフォレストに SFS を適用し，記述子選択によって次元を削減する．その後グリッドサーチによって決定木数，および各決定木の最大深さを設定する．さらに，ランダムフォレストによる EC 反応式分類モデルを作成し，ターゲット反応式の EC 番号を予測する．図 4.8 に，その流れを示す．

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．まず準備として，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，4 章で述べた 2 つの予測を行う．

EC 3 クラス 4 桁目の予測では，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，新たな合成記述子を作成することで，次元削減を行う．そして，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最適な EC 番号 4 桁目を提示する．

EC 3 クラス 2 桁目・3 桁目の予測では，SFS による記述子選択，およびグリッドサーチによるパラメータ調整を経て，ランダムフォレストの EC 反応式分類モデルを作成する．そして，分類精度の評価を行い，ターゲット反応式の EC 番号を予測する．

具体的なデータ整理，前処理，および分類に用いるデータについて，以下で説明する．

データの対応表取得および整理

まず，KEGG の ID や反応式を取得するソースコードを用いて [44]，EC 番号と反応式情報を取得し，表 5.1 のような対応表を取得した．ここでは，行ラベルは EC 番号，列ラベルは EC 反応式の各項となっている．同様に表 5.2 のような C 番号と PubChem CID・SID の対応表も取得した．次に，表 5.1 と表 5.2 を参照し，PubChemPy によって，EC 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生じた．そこで，PUG-REST における，SID から化合物の SDF ファイルを取得する URL を用い，自動取得した SDF ファイルを RDKit で読み込むことで，SMILES に変換した．表 5.1 を置き換え，表 5.3 のような EC 番号と SMILES 化合物の対応表を作成した．また，ターゲット反応式における化合物の 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．

対応表の前処理

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．そのため，SMILES が空白，または「N」と表記される部分が発生するため，その項を含む

表 5.1: EC 番号と化合物 C 番号の対応表

ENZYME	left1	left2	left3	right1	right2	right3	right4
3.6.1.10	C00404	C00001	N	C02174	N	N	N
3.6.1.1	C00013	C00001	N	C00009	N	N	N
3.5.1.54	C01010	C00001	N	C00011	C00014	N	N
3.2.1.28	C01083	C00001	N	C00031	N	N	N
3.2.1.52	C01674	C00001	N	C00140	N	N	N
...	...	...	...	...	...	...	...
3.1.6.21	C02000	C00001	N	C00069	C00059	N	N
3.1.6.22	C22403	C00001	N	C22404	C00059	N	N
3.4.14.14	C22365	C00001	N	C22407	C22330	N	N
3.7.1.28	C22278	C00001	N	C00160	C00197	N	N
3.6.1.-	C00235	C00001	N	C21214	C00009	N	N

表 5.2: 各化合物の ID 対応表

cid	pubchem_SID	pubchem_CID
C00001	3303	962
C00002	3304	5957
C00003	3305	5893
C00004	3306	439153
C00005	3307	5884
...	...	...
C22269	405226444	6365572
C22272	405226445	11788398
C22273	405226446	11411510
C22274	405226447	135567131
C22275	405226448	44468216

表 5.3: EC 番号と化合物 SMILES の対応表

ENZYME	left1	left2	left3	right1	right2	right3	right4
3.6.1.10	O=P(O)(O)OP(=O)(O)OP(=O)(O)O	[H]O[H]	N	O=P(O)(O)OP(=O)(O)OP(=O)(O)O	N	N	N
3.6.1.1	O=P(O)(O)OP(=O)(O)O	[H]O[H]	N	O=P(O)(O)O	N	N	N
3.5.1.54	NC(=O)NC(=O)O	[H]O[H]	N	O=C=O	[H]N([H])[H]	N	N
3.2.1.28	OC[C@H]1O[C@H](O[C@H]2O[C@H](CO)[C@@H](O)[C@H]...	[H]O[H]	N	OC[C@H]1OC(O)[C@H](O)[C@@H](O)[C@@H]1O	N	N	N
3.2.1.52	CC(=O)N[C@H]1C(O)O[C@H](CO)[C@@H](O[C@@H]2O[C@...	[H]O[H]	N	CC(=O)N[C@H]1C(O)O[C@H](CO)[C@@H](O)[C@@H]1O	N	N	N
...	...	...	...	...	...	...	...
3.1.4.61	O=C(O)C1COP(=O)(O)OP(=O)(O)O1	[H]O[H]	N	O=C(O)C(COP(=O)(O)O)OP(=O)(O)O	N	N	N
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	N	*C(=O)O	N	N
3.1.1.118	*C(=O)OCC(COP(=O)(O)O)OC(*)=O	[H]O[H]	N	*C(=O)O[C@H](CO)COP(=O)(O)O	*C(=O)O	N	N
3.1.6.21	*OS(=O)(=O)O	[H]O[H]	N	*O	O=S(=O)(O)O	N	N

EC 反応式は除外した．また，EC 反応式は反応物・生成物がそれぞれ 2 つずつのものを採用した．

特徴ベクトルの作成

表 5.3 の各 SMILES を，RDKit の構造式オブジェクトに変換し，rdkit.chem.descriptor から 208 種の特性値を計算した．その後，特性値変化量を求め，各反応式において 208 次元特徴ベクトルを作成した．表 5.4 に EC 反応式の特徴ベクトルを示す．ここでは，行に EC 番号，列には EC 反応式の特性値変化量算出に用いられた記述子名が記されている．

さらに，表 5.4 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，特性値変化量が 98%以上等しくなる記述子を除外した．

EC3 クラス 4 桁目予測の準備

ターゲット反応式と 4 桁目予測手法の評価方法

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

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ているが，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反応である．初めにターゲット 1 のような反応を試したのち，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ターゲット 2 の化合物に置き換えたと考えられる．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，ターゲット 2 に比べて反応しやすいと推測されることから，ターゲット 1 も比較対象とする．

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 4 桁目予測手法を評価する．

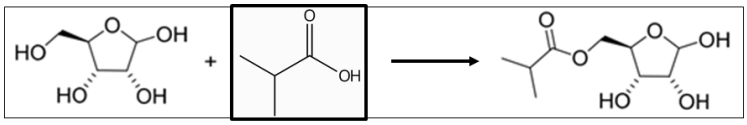
比較対象となる EC 反応式

SOM によるクラスタリングでは，比較する EC 反応式を EC 3.1.1 クラスの約 100 種類に制限して行う．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 のカルボン酸エステル加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．

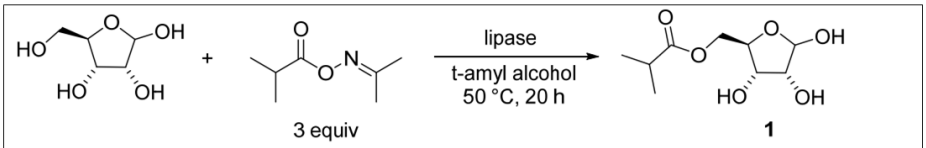
EC 3 クラス 2,3 桁目予測の概要

モデルの分類クラスとターゲット反応式

target1



target2



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 ([45] より引用および一部改変)

表 5.5: EC 番号クラスと全データ数

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

EC 3 クラス 2,3 桁目のモデル作成に用いるクラスは，2,3 桁目の組み合わせ 34 クラスのうち，EC 反応式が 6 個以上取得できた 20 クラスとなる．表 5.5 に，用いるクラスを示す．なお，4 桁目が「-」となっている EC 反応式も用いており，EC 番号が重複している反応式は，先頭の EC 番号を所属クラスとした．各クラスで学習データとテストデータの割合が 4:1 となるように均等に分割し，769 個の EC 反応式からランダムフォレストモデルを作成する．

ターゲット反応式は，4 桁目予測手法に使用する 2 種類と，BRENDA で取得した EC 3.1.1 に属する酵素の文献反応 2 種類，他の EC 3 クラスにおける複数の文献反応を用いる．いずれも正解クラスが分かっている反応となる．

§ 5.2　実験結果と考察

EC3 クラス 4 桁目予測の結果および考察

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	...	Kappa3	EState_VSA8
target1	-0.811847	0.083828	4.450476	0.487355	0.249143	-0.157514	0.192650	-0.269549	0.670170	-0.277871	...	-105.319683	0.0
33	-0.811847	0.083828	-0.101680	0.194684	0.249143	-0.157514	0.210124	-0.042677	0.616525	-0.277871	...	-103.916364	0.0
6	-0.811847	0.083828	-0.101680	0.058794	0.249143	-0.157514	0.142875	-0.042677	-0.312148	1.964942	...	-1.005737	5.106527
1	1.141110	0.083828	-0.101680	0.061725	0.249143	1.630786	0.195868	-0.122933	-0.448792	1.964942	...	-1.278912	5.106527
7_8	-0.811847	0.083828	-0.101680	0.192381	0.249143	-0.157514	0.209894	-0.042677	0.643709	-0.277871	...	-103.686009	0.0
...	...	...	...	...	...	...	...	...	...	...	...	...	...
111	1.141110	0.083828	-0.101680	0.137586	0.249143	-0.157514	0.205520	-0.042677	-0.922301	-0.277871	...	-105.038909	4.736863
118	1.141110	0.083828	-0.101680	0.135367	0.249143	-0.157514	0.214994	-0.042677	-1.328664	-0.277871	...	-105.080776	0.0
2	-0.811847	0.083828	-0.101680	0.205582	-3.984159	-0.157514	0.206484	-0.042677	0.808984	1.964942	...	-104.229693	0.0
5.1	1.141110	0.083828	-0.101680	0.163351	0.249143	1.630786	0.218918	-0.122933	-0.693718	-0.277871	...	-106.172572	-4.523747
26.1	1.141110	0.083828	-0.101680	0.586779	0.249143	-1.945814	0.397498	0.852081	-0.713532	-2.520683	...	-104.746194	0.0

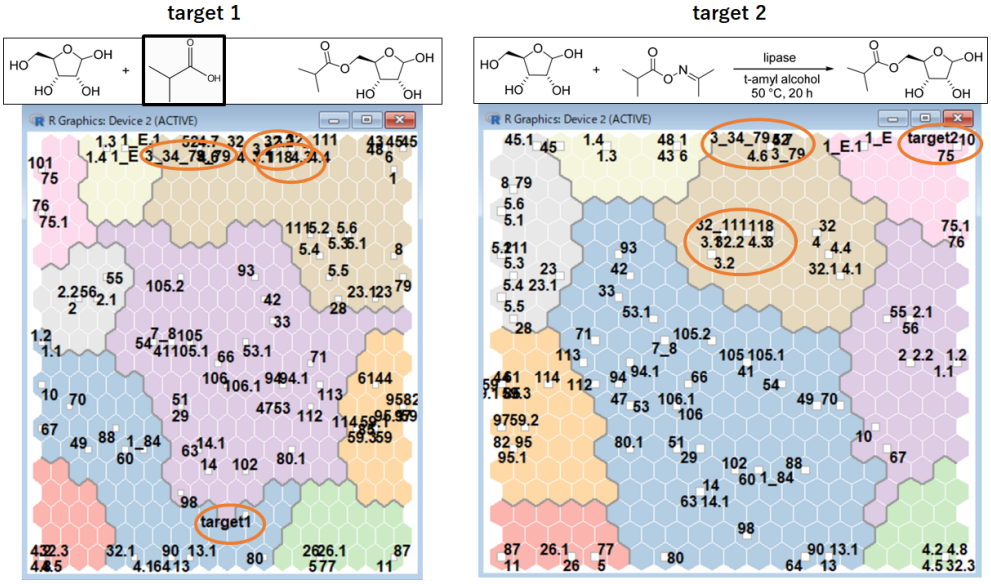


図 5.2: SOM による反応式のクラスタリング結果

特徴ベクトルの次元削減

不要な記述子を除外した，113 次元の特徴ベクトルに対して，次元削減を行った．記述子間における相関係数の逆数である距離行列を入力として，最長距離法の凝集型クラスタリング行った．ターゲット 1 では 16 個のクラスタが形成された 73 次元の特徴ベクトル，ターゲット 2 では 15 個のクラスタが形成され，76 次元の特徴ベクトルとなった．

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，数字は EC 3.1.1 の 4 桁目を表し，ピリオド以下の数字は，同じクラスの EC 反応式が複数ある場合の区別に用いられている．そして，アンダーバー以下の数字は，その反応式が複数のクラス間で重複している場合の区別となっている．

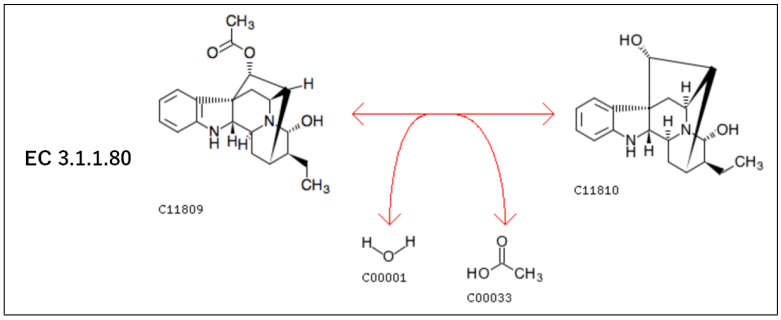


図 5.3: ターゲット 1 の付近に位置している反応式

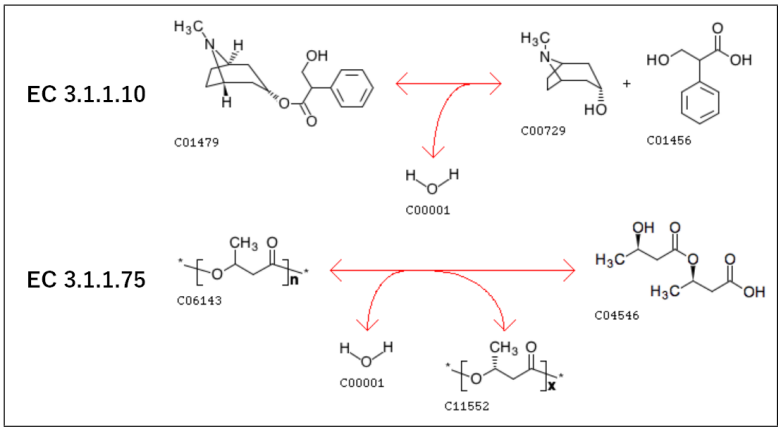


図 5.4: ターゲット 2 の付近に位置している反応式

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 EC 3.1.1 以外の反応式を表し，楕円で囲まれている部分は，ターゲット反応式と EC 3.1.1.3 反応式を示す．ターゲット 1 と同じクラスタに属し，かつ付近に位置する EC 反応式は EC3.1.1.80 となった．KEGG に記載されている反応式を図 5.3 上部に示す．これはノルアジマリンのエステル化反応である．また，ターゲット 2 においては，図 5.4 に示すような，EC 反応式となった．EC 3.1.1.10 はアトロピンの加水分解反応，EC 3.1.1.75 はポリ - β - ヒドロキシ酪酸の加水分解反応となっている．

EC 3.1.1.3 の代表反応式は，他 EC 番号の重複を含めて 5 種類用いられているが，いずれもターゲットに対して離れており，異なるクラスタに属する結果となった．

考察 1

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，相関係数に基づくクラスタリングで合成記述子を作成した際に，構造変化を説明するのに重要な記述子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，重要な方を誤って削除するのを避けるため，相関の高い記述子クラスタ内

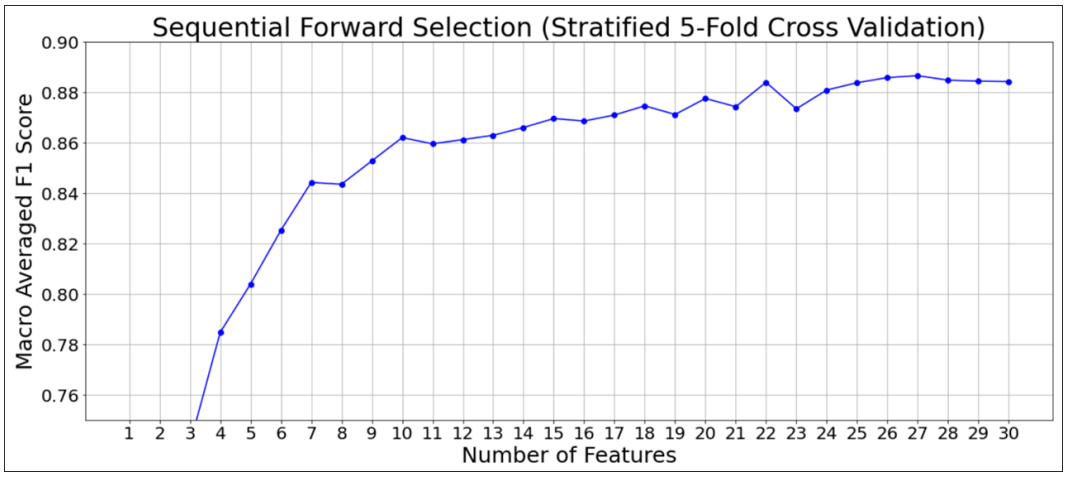


図 5.5: SFS におけるスコアが最も高い記述子の組み合わせ

で標準化・平均化を行い，合成記述子を作成した．しかし，クラスタ内で重要な記述子と重要でない記述子の区別を行わなかったため，平均化によって，重要な記述子の特徴性を薄めてしまった可能性がある．クラスタ内で，各記述子の重要度を認識して重みづけする手法を提案できれば，適切な記述子のみを用いて反応式どうしを比較できると考えられる．

一方で，各ターゲットの付近に位置していた反応式について，その EC 番号の酵素がターゲットの反応において，EC 3.1.1.3 よりも優れた性質を示す可能性が考えられる．BRENDA において，各ターゲットで予測された EC 番号酵素に関する文献数は，EC 3.1.1.3 が約 370 件あるのに対して，EC 3.1.1.80 は 2 件，EC 3.1.1.10 は 12 件，EC 3.1.1.75 は 81 件と少ない傾向にある．EC 番号内の酵素製品の種類や実験数も少ないと考えられ，より優れた反応が見つかる可能性がある．そのためにも，ターゲットが上記反応式に類似していると判断された特徴を特定し，詳しく調べる必要がある．

EC 3 クラス 2,3 桁目予測の結果および考察

分類モデルの作成と評価

初めに，769 個の学習データに対して，ランダムフォレストモデルによる SFS を用いた特徴選択を行った．選択された記述子の組み合わせに対して，層化 5-fold 交差検証を適用してモデルを生成し，各クラスにおける F1 スコアの平均値によって評価した．図 5.5 は，評価の最も良い組み合わせにおける，交差検証の平均スコアである．モデルに用いられる記述子数が 30 になるまで評価を行い，最も高くなったときの 27 種類の記述子を選択した．さらに，同様のモデル評価手法でグリッドサーチを行い，決定木数 100，各決定木の最大深さ 15 が設定された．

最終的に生成されたモデルに用いられている決定木の一つを図 5.6 に示す．ここでは，深さ 3 の部分までを表示し，各ノードには上から，分割に用いられた記述子，ジニ不純度 (gini)，ノード内反応式数 (samples)，各クラスのデータ数 (value)，ノード内の最多クラス (class) が記されている．

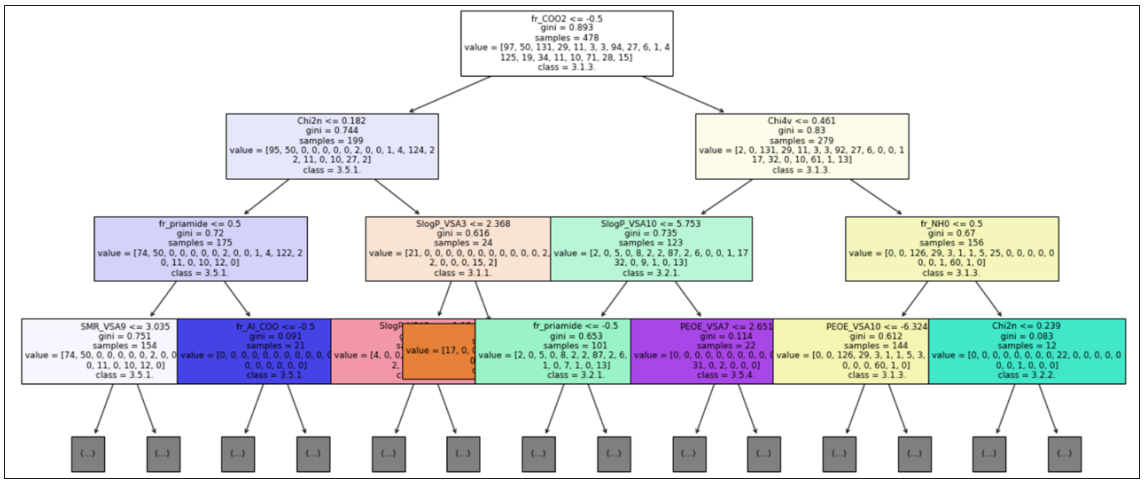


図 5.6: 作成モデルに用いられている決定木の一部

次に，設定した記述子，およびパラメータで生成されたランダムフォレストモデルを，193 個のテストデータを用いて評価した．表 5.7 には，各クラスの要素数と評価値が書かれている．適合率，再現率，f1 スコアの平均値はそれぞれ，0.85, 0.80, 0.81, となり，データ数の多いクラスは高精度で分類できているが，少ないクラスの一部は適合率・再現率がともに 0 という結果になった．

ターゲット反応式に対する予測

最後に，ターゲット反応式の EC 番号予測を行った．正解クラスが EC 3.1.1 であるターゲット反応式の分類結果を図 5.7 に示す．ここでは，各決定木でターゲットが分類された葉ノードにおける，クラス確率の平均値を描画している．target1, target2 は 4 桁目予測で用いたターゲットであり，残り 2 つは BRENDA に登録されている文献の EC 3.1.1 クラスの酵素反応である．結果として，3 つのターゲットは EC 3.1.1 が最も高い割合で予測され，ターゲット 2 のみ EC 3.2.1 と予測された．

同様に BRENDA 内の EC 3.5.3，EC 3.7.1 クラスにおける文献の酵素反応を複数抽出し，ターゲット反応式として予測を行ったものをそれぞれ表 5.8 に示す．ここでは，予測されたクラス確率が高い上位 5 クラスを掲載している．ターゲット EC 3.5.3 では優れた予測性能を示したが，ターゲット EC 3.7.1 では，3 つの反応式が異なる EC 番号として予測される結果となった．

考察 2

正解が EC 3.1.1 であるターゲット 2 が，EC 3.2.1 クラスと予測されたことについて，特性値変化量以外の要因が多く影響したことが考えられる．ターゲット 2 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，今回は 2 つの反応物に対して tert-アミルアルコールも反応物として加えた，反応物 3, 生成物 1 の特性値変化量だったため，異なるクラスに予測された可能性が高い．実験に用いられた試薬や化合物の配合比率など

表 5.7: 作成モデルの各 EC クラスにおける分類精度

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

も考慮した特徴量を作成できれば，複雑な反応に対してもより精度の高い予測ができると考えられる．

あるクラスにおいて，ほとんどの反応式が他クラスに分類されるというケースがいくつか見られたことについて，反応物の構造で反応が生じる部分のみに着目し，各クラスにおける構造変化としての特徴をより強調する．または，EC 番号の酵素分類法に加えて，別の分類規則を追加して予測する必要があると考えられる．

全体の考察

2,3 桁目，および 4 桁目いずれの予測手法においても，予測精度に関して問題点がいくつか見られた．共通となる原因の 1 つに，EC 番号が重複している EC 反応式の正解ラベルを，1 つの EC 番号に限定したことが挙げられる．重複する EC 番号間で同様の特徴ベクトルを学習したことで，EC 番号の反応式としての特徴があいまいになったことが考えられる．今後 EC 3 以外のクラス分類に拡張するためには，重複する反応式を除外した予測手法や，EC 番号だけでなく，特徴ベクトルの類似度に基づくクラスタリングで一度分類し，クラスタ内で性質が似ている EC 反応式を見つけ出す手法などの提案が必要である．

また今回は，反応式中における各項の係数を考慮しなかった．今後，化合物の比率を考慮した特性値変化量を作成することで，より化学的な構造変化を特徴抽出できると考えられる．

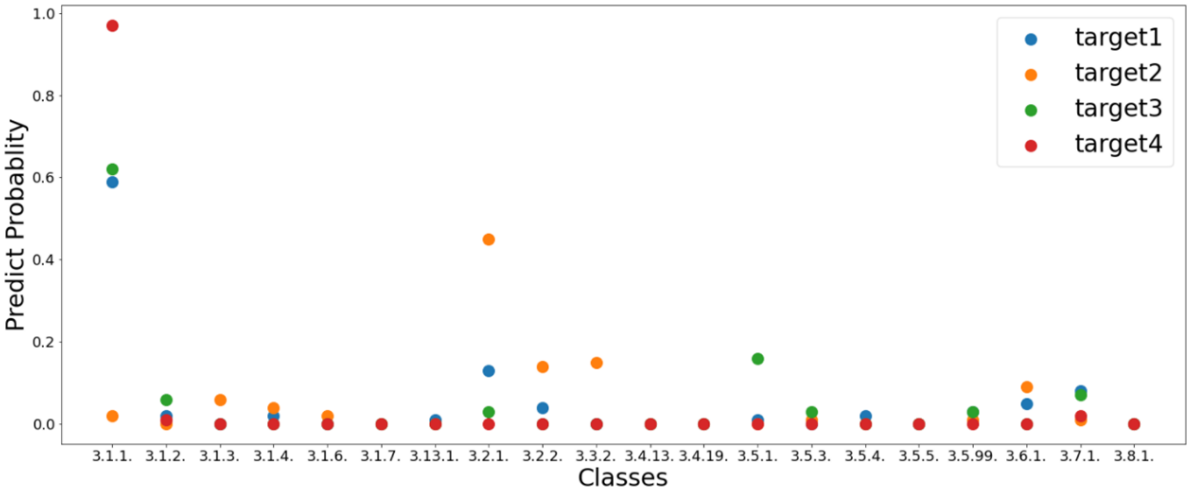


図 5.7: ターゲット反応式 (正解 EC 3.1.1) の EC 番号予測結果

表 5.8: ターゲット反応式 (正解 EC 3.5.3，EC 3.7.1) の予測結果

ターゲット(正解ラベルEC 3.5.3)

	第1位	第2位	第3位	第4位	第5位
target1	3.5.3.	3.1.6.	3.5.99.	3.5.4.	3.1.1.
〃	0.96	0.02	0.01	0.01	0.0
target2	3.5.3.	3.5.4.	3.5.99.	3.1.1.	3.1.2.
〃	0.95	0.03	0.02	0.0	0.0
target3	3.5.3.	3.5.4.	3.5.99.	3.2.2.	3.1.1.
〃	0.89	0.06	0.03	0.02	0.0
target4	3.5.3.	3.2.1.	3.1.1.	3.4.19.	3.7.1.
〃	0.99	0.01	0.0	0.0	0.0

ターゲット(正解ラベルEC 3.7.1)

	第1位	第2位	第3位	第4位	第5位
target1	3.7.1.	3.1.2.	3.5.1.	3.1.1.	3.5.5.
〃	0.31	0.17	0.152	0.1	0.08
target2	3.2.2.	3.13.1.	3.5.4.	3.5.1.	3.1.2.
〃	0.3495	0.210526	0.060526	0.052526	0.05
target3	3.1.1.	3.7.1.	3.6.1.	3.1.2.	3.5.1.
〃	0.29	0.22	0.11	0.106	0.1
target4	3.1.1.	3.5.1.	3.7.1.	3.1.2.	3.5.5.
〃	0.28	0.24	0.13	0.12	0.1

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の面から，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，酵素研究の専門家と協力する，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．本研究では，ターゲット反応式と EC 反応式を比較し，類似する EC 反応式の EC 番号を，最適な酵素候補として予測する方法を提案した．化合物の物理・化学的な特性値を計算し，反応物から生成物への特性値変化量を要素とする特徴ベクトルをもとに，クラス分類やクラスタリングを行い，ターゲット反応式に最適な EC 3 クラスの酵素候補の予測を行った．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 種の特性値変化量からなる特徴ベクトルを作成した．

EC 3 クラス 4 桁目の予測では，特徴ベクトルの次元を約 75 次元に削減したのち，SOM によって反応式のクラスタリングを行った．結果として，ターゲット 1，ターゲット 2 反応式の付近には，正解ラベルである EC 3.1.1.3 反応式は存在せず．他の EC 反応式が位置していた．EC 3 クラス 2 桁目・3 桁目予測手法では，SFS とグリッドサーチを用いたランダムフォレストモデルを作成し，モデルの精度検証，およびターゲット反応式の EC 番号を予測した．高い精度で分類できているクラスがある一方で，一部のクラスでは，ほとんどの反応式が別のクラスに分類されていた．

以上の結果から，ターゲット反応式に最適な酵素候補を提示するシステムとしての性能を高めるためには，構造変化の特徴抽出をより詳細にする，あるいは EC 番号に対して，さらに別の分類規則を加えるなどの工夫が必要であることが明らかになった．また，最適な酵素候補と予測された EC 番号に対して，その酵素が実際に優れた反応を示すかどうか，検証する必要があると考えられる．

今後の課題として，EC 3 以外のクラスに対する予測への拡張や，2・3 桁目予測手法から 4 桁目予測手法への接続，特徴ベクトルのクラスタリングなどの，EC 番号以外の分類法の追加などが挙げられる．

謝辞

本研究を遂行するにあたり，多大なご指導とご鞭撻を賜りました，富山県立大学工学部電子・情報工学科情報基盤工学講座 奥原浩之教授，António Oliveira Nzinga René 講師に深く感謝の意を表します．また，同大学工学部生物工学科酵素化学工学講座 浅野泰久教授，同大学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 岩﨑源司博士には，有機化学・酵素分野の立場から大変貴重なご意見，および当該分野に関して，一からご指導を賜りました．心よりお礼申し上げます．さらに、ケモインフォマティクスにおける機械学習や化学構造の表現法についてご助言を賜りました、国立研究開発法人医薬基盤・健康・栄養研究所上級研究員・プロジェクトリーダー 荒木通啓博士，神戸大学大学院工学研究科応用化学専攻 渡邉直暉氏に深く感謝申し上げます。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 3 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.

[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.

[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] “特異なタンパク質進化 Circular permutation による酵素の機能改変”, https://www.amano-enzyme.co.jp/corporate/foundation/pdf/19/pg09.pdf, 閲覧日 2022.1.25.

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

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

[12] “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.

[13] 白兼孝雄, “酵素の分類と命名法”, JAS 情報, 2017.

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

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

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

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

[18] “PubChem”, https://pubchem.ncbi.nlm.nih.gov/, 閲覧日 2022.1.17.

[19] “About PubChem”, https://pubchemdocs.ncbi.nlm.nih.gov/about, 閲覧日 2022.2.6.

[20] 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.

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

[22] “KEGG API”, https://www.kegg.jp/kegg/rest/keggapi.html, 閲覧日 2022.2.1.

[23] 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.

[24] “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.

[25] 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.

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

[27] “PubChemPy documentation”, https://pubchempy.readthedocs.io/en/latest/#, 閲覧日 2022.2.6.

[28] 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”, *PLoS ONE*, Vol. 7, No. 12, 2012.

[29] 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.

[30] “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.

[31] “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.

[32] “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.

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

[34] 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.

[35] 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.

[36] “KH Coder”, http://khcoder.net/, 閲覧日 2022.1.30.

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

[38] “[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.

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

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

[41] “Mlxtend.feature selection ”, http://rasbt.github.io/mlxtend/api_subpackages/mlxtend.feature_selection/#sequentialfeatureselector, 閲覧日 2022.3.8.

[42] Sebastian Raschka, Vahid Mirjalili 著, 株式会社クイープ訳, 福島真太朗監訳, “[第 3 版] Python 機械学習プログラミング 達人データサイエンティストによる理論と実践”, 株式会社インプレス, 2020.

[43] “sklearn.ensemble.RandomForestClassifier”, https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html, 閲覧日 2022.3.15.

[44] “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.

[45] “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.