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

富山県立大学大学院 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: 令和6年 (2024年) 2月

目 次

図一覧 ii

表一覧 iii

記号一覧 iv

第 1 章　はじめに 1
§ 1.1 本研究の背景 1
§ 1.2 本研究の目的 2
§ 1.3 本論文の概要 3

第 2 章　有機合成と酵素 4
§ 2.1 有機合成分野の概要 4
§ 2.2 酵素と EC 番号 6
§ 2.3 酵素の探索 8

第 3 章　機械学習による EC 番号予測 15
§ 3.1 計算機上における化学反応の表現法 15
§ 3.2 EC 番号予測手法 17
§ 3.3 機械学習と特徴量エンジニアリング 19

第 4 章　提案手法 23
§ 4.1 特徴ベクトルと特徴エンジニアリングによる EC 番号予測 23
§ 4.2 EC 番号予測モデルの構築と予測 25
§ 4.3 提案手法の実装と流れ 27

第 5 章　実験結果並びに考察 32
§ 5.1 数値実験の概要 32
§ 5.2 実験結果と考察 33

第 6 章　おわりに 41

謝辞 42

参考文献 43

図一覧

2.1 HCHO の化学反応ネットワーク [9] 5
2.2 α-キモトリプシンの立体構造 [13] 6
2.3 反応の進行と必要なエネルギー [14] 6
2.4 モルヌピラビルの合成 [1] 7
2.5 EC 番号による酵素の分類 8
2.6 EC3.1.1.2 の代表的な反応式 9
2.7 KEGG ENZYME データベース 10
2.8 KEGG REACTION データベース 11
2.9 BLAST 検索画面 12
2.10 BLAST 検索結果 13
2.11 5-isobutyryl ribose 合成における酵素選択 (一部抜粋) [1] 14

3.1 KEGG COMPOUND で取得できる構造式と MOL ファイルの例 16
3.2 rdkit を用いた化合物の情報 17
3.3 SMOTE によるオーバーサンプリングのイメージ 20
3.4 決定木におけるクラス分類の様子 21

4.1 特徴ベクトルのイメージ 24
4.2 構造式オブジェクトに変換された酵素反応 25
4.3 モデル構築と予測の流れ 26
4.4 データ加工の流れ 28

5.1 実験 2 の流れ 34
5.2 予備実験 1 の結果 35
5.3 EC 1.X.X に対する分類 35
5.4 EC 2.X.X, 3.X.X に対する分類 37
5.5 EC 4.X.X, 5.X.X に対する分類 37
5.6 EC 6.X.X, 1 桁目に対する分類 37

表一覧

4.1 実際の特徴ベクトル 24
4.4 除外リスト 28
4.5 SMILE 反応式 (左辺) 28
4.2 SMILE 形式のデータセット (一部抜粋および改変) [5] 30
4.3 学習データの内訳と SMOTE 増加数 31

5.1 EC 3 の 2,3 桁目クラスのデータの内訳 33
5.2 予備実験 2 で選ばれた記述子リスト 36
5.3 EC 番号 1 桁〜3 桁目までの分類結果 1 38
5.4 EC 番号 1 桁〜3 桁目までの分類結果 2 39
5.5 EC 番号各 2,3 桁目の分類精度 40

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		
正例データを正しく正と予測した数	TP	負例データを誤って正と予測した数	FP
負例データを正しく負と予測した数	TN	正例データ誤って負と予測した数	FN
正に分類されたデータのうち，実際に正であったデータの割合	$Precision$	正に属するデータを正しく正に分類した割合	$Recall$
全体の予測精度	$Accuracy$	Precision と Recall の調和平均	$F1$
近傍データ数	K	着目するデータ点	x_o
選択されたデータ点	x_k	ランダムな生成値	r
D_p を含むノードにおける情報利得	$IG(D_P, f)$	ノード分割に用いる特徴量	f
決定木上位ノード内のデータ	D_p	決定木上位ノードのデータ数	N_p
上位ノードに対する下位ノード左 (右)	D_{left}, D_{right}	上位ノードに対する下位ノード左 (右) 数	N_{left}，N_{right}
ノード t におけるジニ不純度	$I_G(t)$	ノード t におけるクラス i のデータの割合	$p(i\|t)$
ノード t 内のクラス数	c		

第1章

はじめに

§ 1.1　本研究の背景

有機合成化学分野では，化学反応の予測や設計に酵素を生体触媒として利用する機会が増加している．生体触媒となる酵素は，特定の反応物 (基質) に対して触媒反応を示し，化学触媒に比べてしばしば化学反応をより効率的に進め，かつ環境に優しい条件で利用できる．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事例がある [1]．そのような背景から，基質から目的の生成物を生み出す化学反応に対して，最適な酵素を探索することが重要視されている．

酵素には 4 桁の Enzyme Commission number (EC 番号) が割り振られており，どの反応を触媒するか，どの結合や基質に作用するかによって酵素が分類されている [2]．EC 番号は代謝経路の解析で得られる酵素の性質の特定や触媒する化学反応 (酵素反応) の探索に用いられる．代謝経路の解析では遺伝子解析によって得られたタンパク質配列から性質を特定し，酵素反応の探索では，化合物の合成を通して反応効率のよい組み合わせを探索する必要があるが，これらは実験によって正確が解析が行われる．近年，実験コストや時間を短縮する目的で，EC 番号を機械学習を用いて予測する研究が行われている．EC 番号を予測することで対象となる酵素の大まかな性質や触媒する反応物 (基質) の種類などを知ることができるため，実験による解析や探索の効率化が期待できる．EC 番号が登録されている酵素には，酵素反応やタンパク質配列といった情報がデータベース上で付与されており，これらの情報から既存のデータに対して正しい EC 番号を予測する機械学習モデルの開発が行われてきた．既存のデータに対する予測精度を高め，将来的に未知の酵素に対して自動的な EC 番号の割り当てが行われることが目指されている．

EC 番号予測が活発的になる以前から情報技術を用いた解析技術が数多く存在しており，化学合成経路の解析や未知酵素と既存酵素のタンパク質の類似性探索を行うデータベースなどの技術が土台となって様々な手法が開発されてきた．EC 番号の予測手法として，タンパク質配列に着目した遺伝子ベースの手法や化合物の構造的特徴や物理・化学的な特性値の変化に着目した化学的手法がある．タンパク質配列に着目した手法では，配列の類似性や画像処理的方法によって，与えられた配列がどの EC 番号に属するのかを予測する [3]．化学的手法では化合物が持っている特徴的な分子の有無に関するバイナリ値 (フィンガープリント) や，化合物に対する物理・化学的指標を複数用いた特徴ベクトルを用い，EC 番号が割り当てられた酵素反応を特徴ベクトルとして学習器に学習させることで EC 番号を予測する [5] [4]．

§ 1.2　本研究の目的

本研究では，有機合成を行うことを主目的として，基質から生成物に変化する際の物理・化学的な特性値の変化に着目し，機械学習を用いて特定の化学反応に最適な酵素の候補をEC番号として予測する．物理・化学特性値はPythonライブラリのRDKitで得られる特徴量 (記述子) を用いる．これらの記述子は分子量や親油性，電荷分布などの125種の物理・化学的特性値と，特徴的な構造 (分子断片) を持っているかを判定する85種のバイナリ値で構成されている．基質から生成物の変化に着目した従来手法として，物理・化学的特性値を用いた手法 [4] と差分フィンガープリントを用いたものがあるが [5]，本研究はこの2つの手法の一部を組み合わせたものとなる．化合物の構造情報だけでなく，物理，化学的な指標を加えることで，化学合成時の特徴をより詳細に捉えることができると考えられる．

本研究の下準備として，初めに酵素を用いた化学反応 (酵素反応) を収録した4つのデータベースからなるEC番号が割り当てられている既存の酵素反応のデータセットを提案手法に合わせて加工する．このデータセットは左辺と右辺からなる化学反応式中の化合物の化学構造を文字列で表現したもので構成されている．それによって，RDKitの210種類の記述子を用いた，基質から生成物に変化する際の特性値変化量からなる210次元の特徴ベクトルのデータセットを作成した．次に記述子間の相関係数が1の記述子の片方を削除し，欠損値が含まれる記述子を削除することで200次元の特徴ベクトルとした．提案するEC番号予測モデルは従来と同様にEC番号の1桁目から3桁目を予測するモデルとする，既存のEC番号が割り当てられている酵素反応に対して，正しいEC番号に予測ができるかの多クラス分類を行う．

初めに，データセットの訓練データに対して1桁目と2,3桁目に対する分類の交差検証を行い，200種類の中から分類精度に貢献する記述子を選択する．1桁目の分類，およびEC 1からEC 6クラスの2,3桁目分類のそれぞれで選択された記述子をマージ，および重複を削除した組合せを最終的に用いる．分類器としてはRandom Forest (RF) を用い，記述子はForward法で1つずつ選択する．また，2,3桁分類による記述子選択ではクラス間のデータ分布に偏りが生じる懸念があることから，検証用訓練データにSMOTEを適用しオーバサンプリングを行う．次に，選択された記述子組合せと訓練データを用いて，EC番号の1,2,3桁目を同時に予測する分類モデルを作成する．RFのハイパーパラメータである最大深さと，決定木数に対して，グリッドサーチでパラメータ調整を行ったのち，テストデータに対する分類精度を評価した．この際，それぞれのクラス分類における特徴選択と，グリッドサーチでは5分割交差検証を用いた．各分割で検証用訓練データにSMOTEを適用し，作成したモデルで検証用テストデータを分類した．このとき，Macro F1-Scoreが最も高い記述子組合せ・パラメータを選択された．

数値実験では2つの予備実験と本実験を行った．予備実験1では，従来行われたEC 3クラスの2,3桁目分類に対してSMOTEを適用し，適用前のものとF1-macroスコアで比較することでSMOTEの有効性を確認する．予備実験2では，EC番号1桁目とEC 1からEC 6の2,3桁目クラスの記述子選択後に作成される分類モデルに対して，モデル単体での予測精度を評価する．分類スコアが最も高い記述子組合せを用いてSMOTEとグリッドサーチを適用し，Macro F1-Scoreで評価した．本実験では予備実験2で得られた記述子をマージ・重複削除した組合せを用いてRFのグリッドサーチを行ったのち，EC番号1〜3桁目がラ

ベリングされたテストデータを用いて，多クラス分類を行う．

§ 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．本研究の背景では，有機合成における酵素の利用について，生体触媒を用いることのメリットと EC 番号の利用について述べた．目的では，有機合成に最適な酵素候補を予測する機械学習モデルを提案し，データの加工やモデル構築，数値実験の大まかな流れを説明した．

第 2 章 有機合成で行われている内容や情報分野との関係，酵素の性質と EC 番号による分類体系，および酵素探索を行うための EC 番号予測の重要性について述べる．

第 3 章 化学反応の機械学習を行うにあたっての計算機上での化学反応の表現法や機械学習による EC 番号予測手法の先行研究を述べる．また，本研究で用いる機械学習や特徴エンジニアリング関する手法について述べる．

第 4 章 提案手法についての説明，および提案システムの実装と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

有機合成と酵素

§ 2.1　有機合成分野の概要

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，アヘンから強い麻酔作用を持つ morphine の抽出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 [6]．また，1828 年には有機化合物として初となる，尿素の合成が成功し，その後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 [7].

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，情報技術を用いて化学分野の研究を行う分野である，ケモインフォマティクスが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 [8].

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 [9]．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

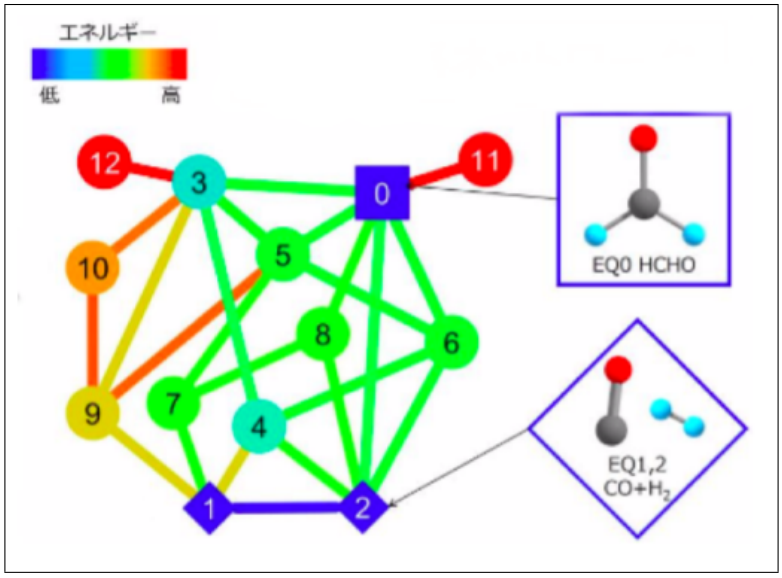


図 2.1: HCHO の化学反応ネットワーク [9]

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 [10]．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 [11]．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．

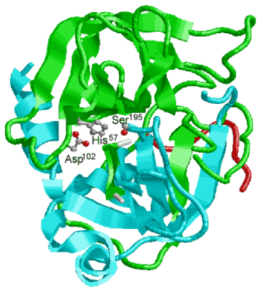


図 2.2: α-キモトリプシンの立体構造 [13]

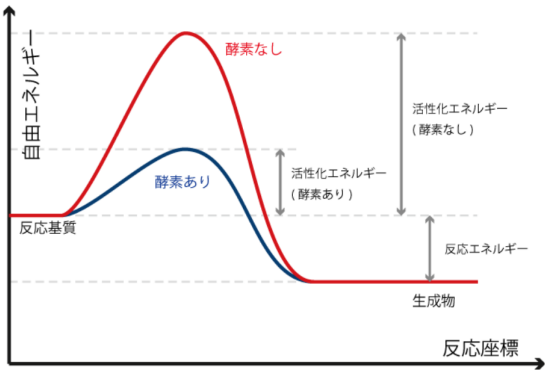


図 2.3: 反応の進行と必要なエネルギー [14]

§ 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「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 [12].

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

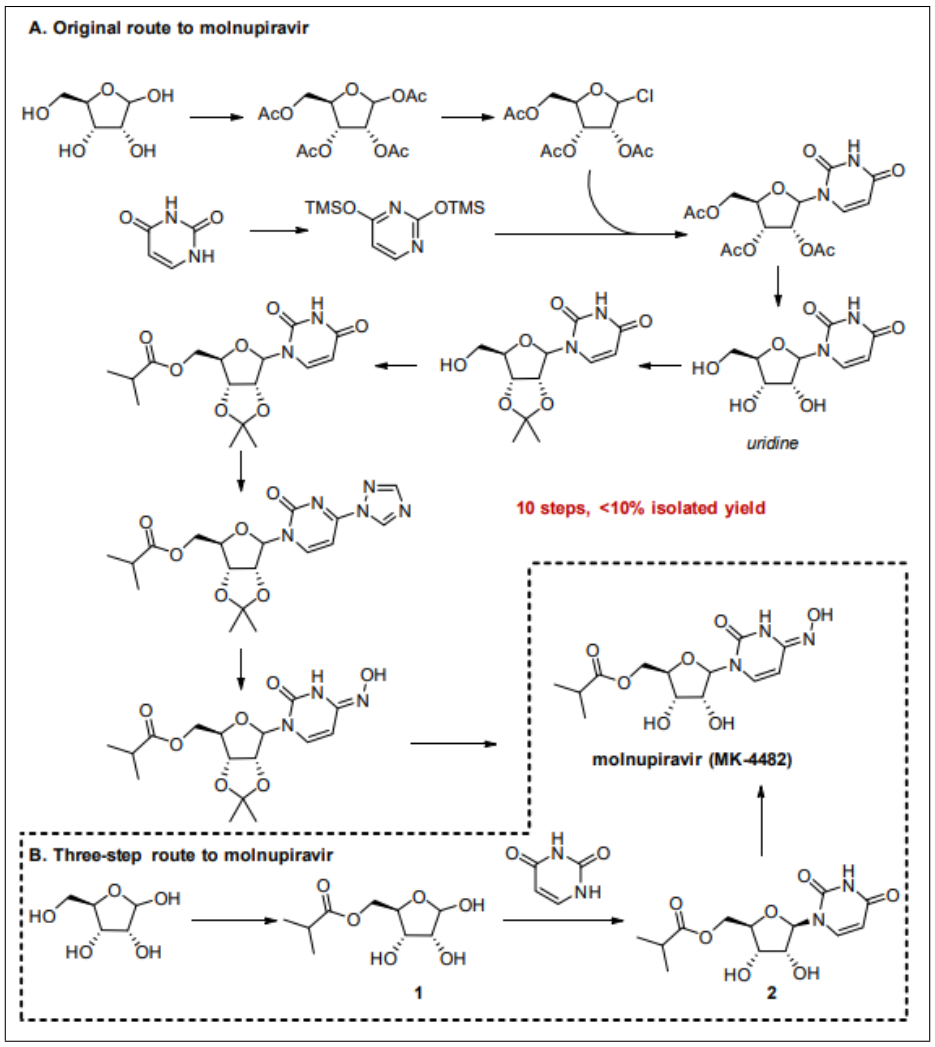


図 2.4: モルヌピラビルの合成 [1]

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 [1]．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 桁目の数字 (class)

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 [15] [16]．2 桁目の数字では，どの結合に作用するか，3 桁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 桁目の数字で 1 から 3 桁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 [2]．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,・・・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．EC 番号を扱うデータベースでは，4 桁の EC 番号に分類されている酵素の名前と触媒する反応式が記載されている．例えば，Kyoto Encyclopedia of Genes and Genomes (KEGG) [17] では，図 2.6 のように自然界で起こる酵素を用いた代表的な化学反応 (酵素反応) が 1 つまたは複数登録されている．このように，既存の酵素反応データを機械学習に用いることで，EC 番号が分かっている酵素反応を入力した際，正しい EC 番号に分類できるかの予測を行うことができる．

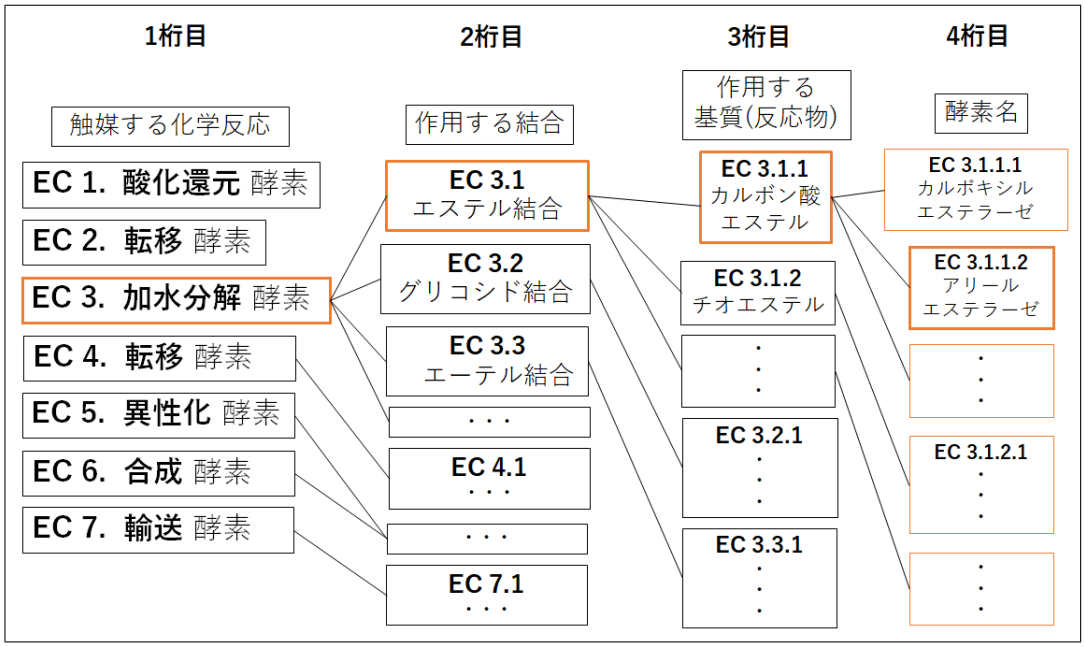


図 2.5: EC 番号による酵素の分類

§ 2.3 酵素の探索

ここでは，代謝経路の解析と有機合成における酵素探索の概要に触れ，機械学習を用いて EC 番号を予測する背景を述べる．

データベースによる EC 番号と化学情報との関連付け

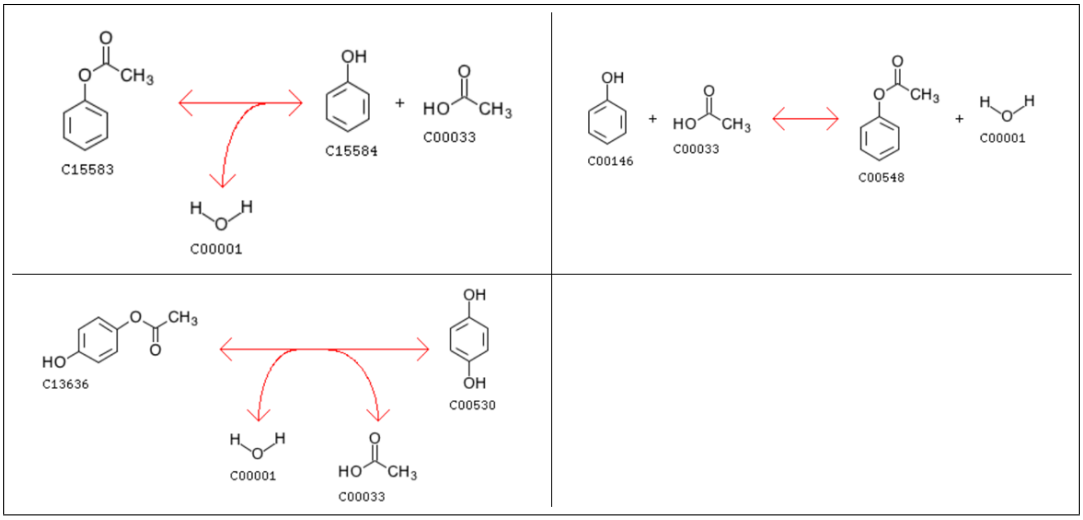
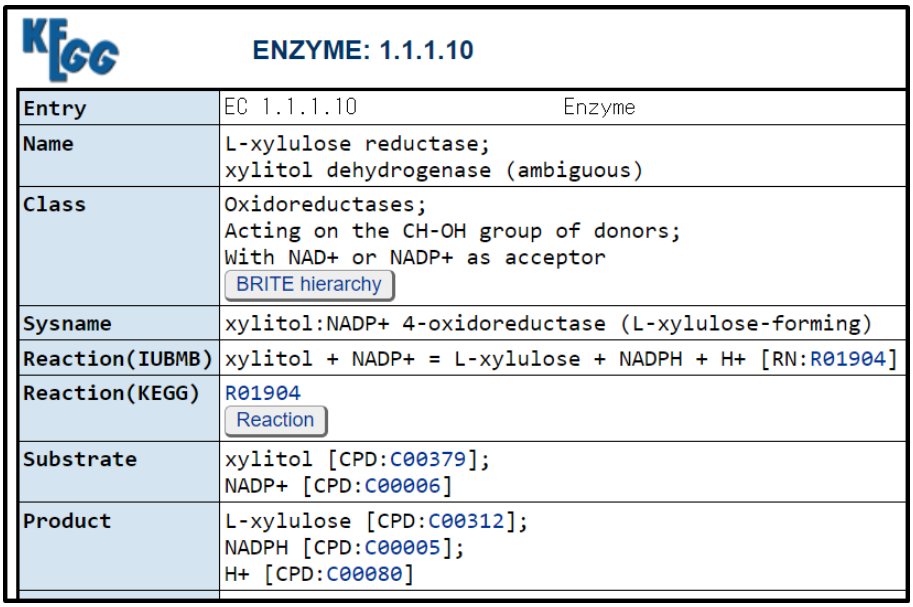


図 2.6: EC3.1.1.2 の代表的な反応式

EC 番号を扱うデータベースの 1 つである KEGG を例に挙げる．KEGG は遺伝子・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 2.7 および図 2.8 に KEGG データベースの例を示す．

類似タンパク質配列の検索

タンパク質配列どうしが類似する場合，類似の機能を有するという類似性の概念から，配列が特定されたタンパク質に対して類似する配列を探索することで，タンパク質の性質解析につながる．探索方法の一つとして Basic Local Alignment Search Tool（BLAST）[18] を用いた方法がある．BLAST では配列同士を並べ，部分的に等しくなる箇所を探していくことで配列間の類似性を計算するシステムで，データベース内の配列と比較することで，未知タンパク質配列の性質の分析を行うことができる．例として，KEGG に登録されている酵素 EC 3.1.1.1 のタンパク質配列を図 2.9 の検索窓に入力した場合，図 2.10 のようにデータベース内で類似する配列が検索結果として表示される．ここで，図 2.10 内の Max Score, Total Score は類似性の評価スコア，Query Cover は配列の長さに対する一致した箇所の割



ENZYME: 1.1.1.10

Entry	EC 1.1.1.10 Enzyme
Name	L-xylulose reductase; xylitol dehydrogenase (ambiguous)
Class	Oxidoreductases; Acting on the CH-OH group of donors; With NAD+ or NADP+ as acceptor BRITE hierarchy
Sysname	xylitol:NADP+ 4-oxidoreductase (L-xylulose-forming)
Reaction(IUBMB)	xylitol + NADP+ = L-xylulose + NADPH + H+ [RN:R01904]
Reaction(KEGG)	R01904 Reaction
Substrate	xylitol [CPD:C00379]; NADP+ [CPD:C00006]
Product	L-xylulose [CPD:C00312]; NADPH [CPD:C00005]; H+ [CPD:C00080]

図 2.7: KEGG ENZYME データベース

合，E value はその類似性スコアが偶然えられるものかどうかの期待値，Per. Ident は配列の一致度合いを表す [19]．Description 内のリンクを参照することで，調べたいタンパク質配列とデータベース内の配列との一致箇所や検索結果配列の情報を調べることができる．

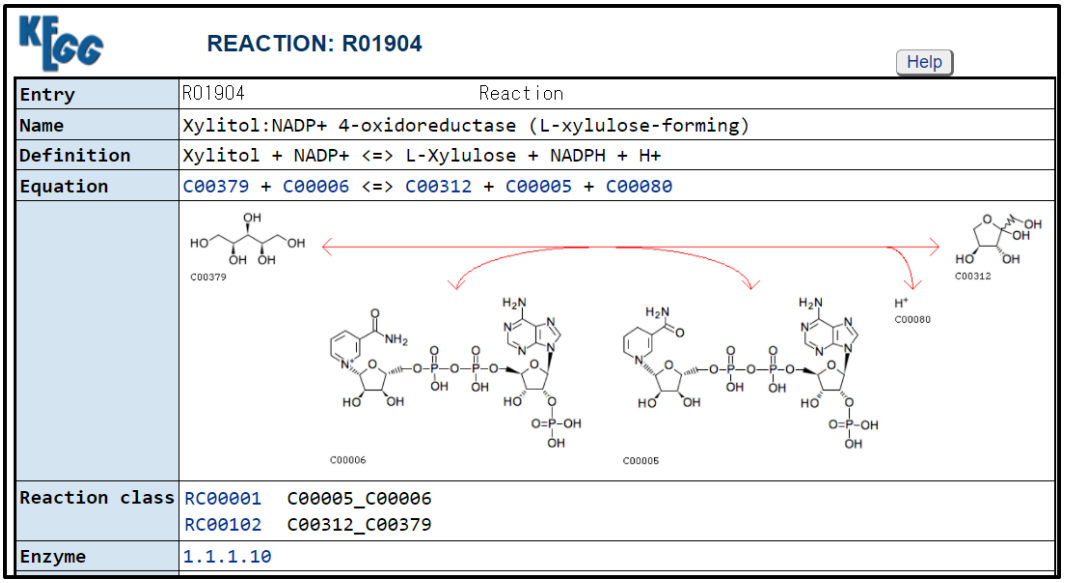
代謝経路解析における EC 番号予測手法の利用

酵素が触媒することで起こる生体内の反応が別の反応を次々に引き起こす流れを代謝経路と呼ぶ．代謝経路内で起こる酵素反応を実験的に解析し，性質を特定 (アノテーション) していくことで，生体機能の解明につながる．また，酵素がどの化学反応を触媒するのかを解析することで，材料や新薬の開発につながる．遺伝子配列の解析技術の発展によって，タンパク質配列を特定する速度が向上している一方で，実験等によるアノテーションが追いついておらず，機能が未解明の配列が蓄積されている [5] [20]．そのため，機械的な機能予測や自動的なアノテーション法を開発することが重視されている．

従来行われてきた割り当て方法として，アノテーションを必要とするタンパク質配列を BLAST などで検索する方法がある．しかし，検索結果に表示される配列が類似度の低いもののみの場合，信頼性のあるアノテーションが行えないという問題がある．機械学習による予測はそのような問題を解決するために用いられ，類似性検索の改善や類似性を超えた自然の法則をデータによって推定することで，より高性能な割り当てを可能にすることが期待されている．具体的には，EC 番号によって分類されている酵素のタンパク質配列を数値化し，ベクトルを学習器に入力することで，EC 番号の分類体系に沿った分類モデルを作成できる．

有機合成における酵素探索

2.1 節で述べたように有機合成では逆合成的なアプローチで化学合成を行うことがある．最終的に得たい化合物を設定し，切断が容易な部位を探し，2 つの化合物に分解する．同様



REACTION: R01904

Help

Entry	R01904 Reaction
Name	Xylitol:NADP+ 4-oxidoreductase (L-xylulose-forming)
Definition	Xylitol + NADP+ <=> L-Xylulose + NADPH + H+
Equation	C00379 + C00006 <=> C00312 + C00005 + C00080
Reaction class	RC00001 C00005_C00006 RC00102 C00312_C00379
Enzyme	1.1.1.10

図 2.8: KEGG REACTION データベース

に分解した化合物からさらに切断部位を設定する手順を繰り返し，低コストで入手しやすい化合物が現れるまで分解する．このように合成経路と用いる化合物を決定した後，化合物を合成していく．スムーズに目的の化合物を得るため，様々な試薬や溶媒，触媒，温度，気圧を設定し，最も収率の良い組み合わせを実験によって選択する．このとき，反応を効率良く進める生体触媒として酵素が用いられることがある．一般的に酵素反応は自然界で発生する反応であり，特定の基質のみに作用するため，非天然の化合物には反応しない．そのため，非天然物から目的化合物を作る場合，酵素の働きを活性化させたり，作用する基質の種類を増やすような試薬・溶媒の探索が行われる．

具体例として，図 2.4 下部の 1 と示された，ribose と o-isobutyrl oxime から 5-isobutyryl ribose を作成する過程を挙げる [1]．初めに 8 種類の酵素製品を用いて最も効率的に 5-isobutyryl ribose が得られる酵素を選択する実験が行われている．ここでは, EC 3.1.1.3 である Novozym 435 が選択された．図 2.11 にその結果を示す．次に Novozym 435 を用いて 5-isobutyryl ribose の収率が最も高くなる，溶媒や Novozym 435 の濃度が選択され，最終的な合成経路として決定される．

将来的に EC 番号を予測するモデルを有機合成の場で用いることができれば，合成に用いるべき酵素の候補を同じ EC 番号内の酵素製品に絞り込むことができる．それによって酵素候補を選出する過程を短縮し，次の過程である，酵素で目的物を効率よく得るための，実験条件の探索にスムーズに進むことができる．また，酵素に関する知識が十分でない場合でも，実験に用いる酵素候補が自動的に選ばれるため，有機合成をより容易に行えることが期待できる．

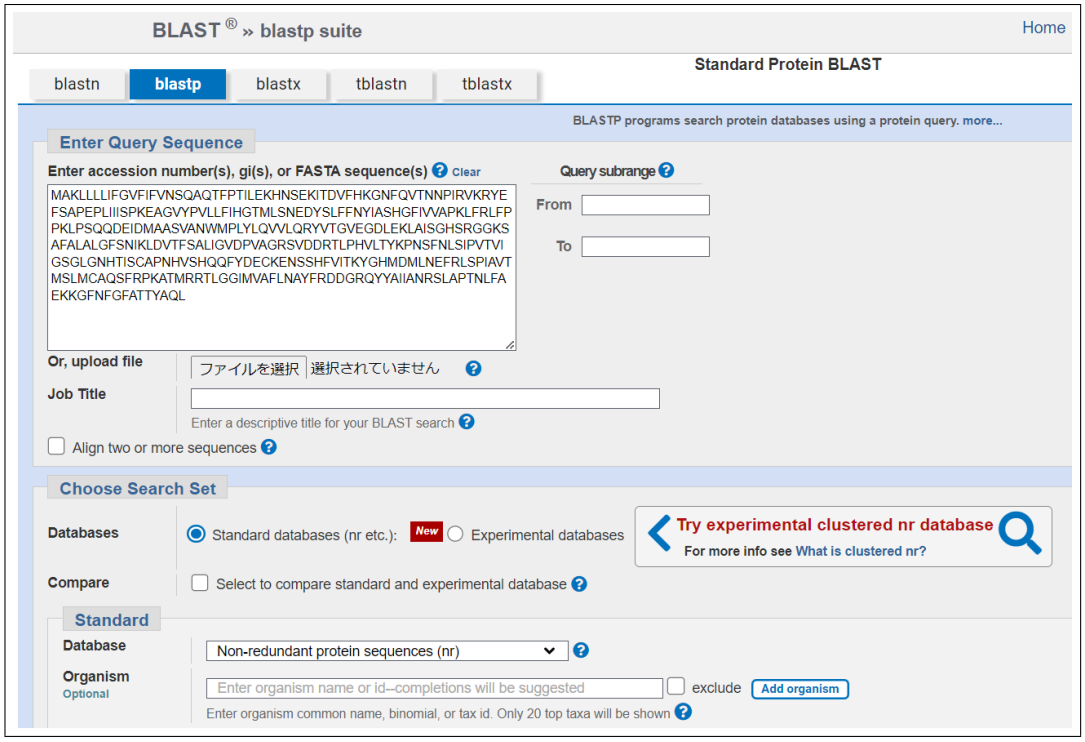


図 2.9: BLAST 検索画面

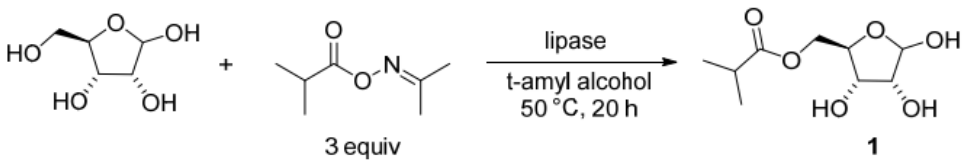
Descriptions | Graphic Summary | Alignments | Taxonomy

Sequences producing significant alignments Download Select columns Show 100

select all *100 sequences selected* GenPept Graphics Distance tree of results Multiple alignment MSA Viewer

Description	Scientific Name	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession
RecName: Full=Chlorophyllase type 0; AltName: Full=CaCLH0; AltName: Full=Chlorophyll-chlorophyllido hydr…	Chenopodium al…	718	718	100%	0.0	100.00%	347	Q9LE89.1
LOW QUALITY PROTEIN: chlorophyllase type 0-like [Chenopodium quinoa]	Chenopodium q…	686	686	100%	0.0	95.97%	347	XP_021735971.1
chlorophyllase type 0-like [Chenopodium quinoa]	Chenopodium q…	638	638	100%	0.0	89.34%	345	XP_021754029.1
chlorophyllase type 0-like [Spinacia oleracea]	Spinacia oleracea	441	441	98%	2e-151	65.71%	354	XP_056688230.1
chlorophyllase type 0-like [Chenopodium quinoa]	Chenopodium q…	416	416	100%	9e-142	61.67%	355	XP_021735970.1
chlorophyllase type 0-like [Spinacia oleracea]	Spinacia oleracea	404	404	99%	8e-137	61.00%	359	XP_021844444.1
hypothetical protein BVRB_2g033410 [Beta vulgaris subsp. vulgaris]	Beta vulgaris su…	390	390	98%	4e-131	59.28%	365	KMT17996.1
chlorophyllase type 0-like [Spinacia oleracea]	Spinacia oleracea	387	387	97%	2e-130	58.96%	344	XP_021844517.2
chlorophyllase type 0 [Beta vulgaris subsp. vulgaris]	Beta vulgaris su…	387	387	91%	2e-130	62.58%	365	XP_010669257.2
chlorophyllase type 0-like [Amaranthus tricolor]	Amaranthus tric…	380	380	97%	2e-127	56.58%	351	XP_057524082.1
chlorophyllase type 0-like [Chenopodium quinoa]	Chenopodium q…	334	334	99%	1e-109	52.53%	335	XP_021753754.1
hypothetical protein Cgig2_020904 [Carnegiea gigantea]	Carnegiea gigan…	299	299	88%	8e-96	51.71%	349	KAJ8421835.1

図 2.10: BLAST 検索結果



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 2.11: 5-isobutyryl ribose 合成における酵素選択 (一部抜粋) [1]

第3章

機械学習によるEC番号予測

§ 3.1　計算機上における化学反応の表現法

化学反応を学習器に入力する場合，化合物の構造などを計算機上において扱いやすい形式で表現する必要がある．化学構造を文字列表現したり，化学反応時の物理的・化学的な指標から得られる数値を用いることによって，化学反応の予測や分類を行う．ここでは，計算機上での各種表現法や用いれれるライブラリについて説明する．

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のを図 3.1 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，主に SDF ファイル形式となる．

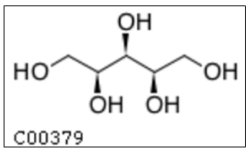
SMILES

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 [21].

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




```
 10  9  0  0  1  0  0  0  0  0999 V2000
   17.3600  -18.9000    0.0000 O   0  0  0  0  0  0  0  0  0  0  0  0
   18.5724  -18.2000    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
   19.7849  -18.9000    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
   20.9973  -18.2000    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
   22.2097  -18.9000    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
   23.4222  -18.2000    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
   24.6346  -18.9000    0.0000 O   0  0  0  0  0  0  0  0  0  0  0  0
   19.7849  -20.2998    0.0000 O   0  0  0  0  0  0  0  0  0  0  0  0
   20.9973  -16.8000    0.0000 O   0  0  0  0  0  0  0  0  0  0  0  0
   22.2097  -20.2999    0.0000 O   0  0  0  0  0  0  0  0  0  0  0  0
  1  2  1  0     0  0
  2  3  1  0     0  0
  3  4  1  0     0  0
  4  5  1  0     0  0
  5  6  1  0     0  0
  6  7  1  0     0  0
  3  8  1  1     0  0
  4  9  1  1     0  0
  5 10  1  1     0  0
M  END
```

図 3.1: KEGG COMPOUND で取得できる構造式と MOL ファイルの例

フィンガープリント

化合物の構造や特徴をバイナリベクトルで表現したもので，化合物どうしの分子の類似性などを比較する場合に用いられる．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．例として 2 つの表現法を挙げる．

Molecular Access System for Screening and Selection (MACCS) Keys [22]

化合物の分子に対して，166 種類の部分構造を持っているかどうかを参照する．持っている場合は 1，持っていない場合は 0 と判定して，166 次元のベクトルで表現したもの．

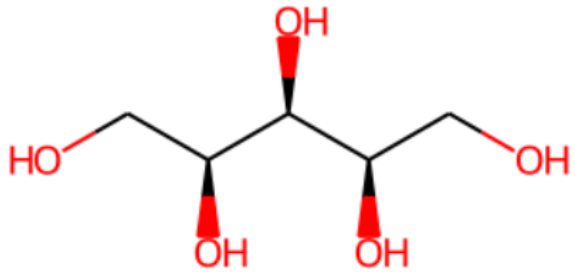
Extended Connectivity Fingerprint (ECFP) [26]

隣接する分子の繋がりをバイナリベクトルで表現したもの．ベクトル化の手順として，初めに，フィンガープリントの長さ L と隣接原子の探索数 F とする．また各原子の特徴を表す値を計算し，初期化を行う．具体的には，結合内の各原子に対して，水素を除く重原子数，原子価から水素数を引いたもの，原子の番号・質量・電荷，結合水素数，環構造に含まれる原子かどうかのバイナリ値を元に数値化を行い，ハッシュ関数によって 32 ビットの整数変換する．次に，隣接する原子の整数値を結合し，ハッシュ関数に入力することで整数値を更新する．各更新において，L で割った余りをフィンガーベクトルのビットとみなし，対応するビットを 1 にする．それぞれの原子に対して，同様の作業を F の数だけ離れた，隣接原子を超えるまで繰り返し行う．最終的に各原子の整数値に該当する箇所が 1 として表されたフィンガープリントが生成される．

RDKit [23]

RDKi は Python のライブラリとして提供されており，化学分野に関わる様々な情報処理を行える．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，125 種類の物理・化学的特性値と，85 種類の分子断片の有無を判定するバイナリ値からなる 210 種類の記述子が用意されており，化学合成時の材料探索 [24] や生体内で起こる化学反応の機械学習

```
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 3.2: rdkit を用いた化合物の情報

予測 [25] などに用いられている．物理・化学的特性値として，化合物の分子量や油 (水) への溶けやすさを表す MolLogP [28]，分子の静電的な相互作用を表す 14 種の PEOE_VSA [29]，分子断片の有無に関する指標として，芳香族アミンの数を表す fr_Ar_NH，チオール基の数を表す fr_SH などがある．例えば，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 3.2 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．

§ 3.2　EC 番号予測手法

EC 番号は代謝経路の解析や酵素反応の探索において，おおまかな酵素の性質や探索範囲を特定し，実験コストや時間を短縮するために用いられてきた．機械学習による自動的な EC 番号の予測が行われており，データベース上にある既存のタンパク質配列や酵素反応データを用いて予測モデルを構築し，精度をより高める予測手法を開発することで，将来的に未知データに対して最適な EC 番号を提示することが目指されている．既に EC 番号が割り当てられているデータを用い，EC 番号の 1 桁から 3 桁または 4 桁目までのクラスに対する多クラス分類を行うことで，データが正しい EC 番号のクラスに予測されるかを評価する．EC 番号の予測手法として，タンパク質配列に着目した手法や化合物の物理・化学的

特性値，構造的特徴を用いた手法があり，以下でそれぞれの手法を述べる．

タンパク質配列を用いた予測

タンパク質配列は文字列で構成されているが，2 次元上に配列をマッピングし画像の機械学習で用いられる手法などを用いることで，タンパク質配列を正しい EC 番号に割り振ることを目指す．EnzymeNet [3] では，酵素機能を有する配列約 1,000,000 と有していない配列約 140,000 を用い，EC7 を除く EC 番号の 4 桁全ての予測を行っている．ここではディープラーニングによって文字列を変換し，配列の長さを 1024 に調整後，分類器の中で 1024 × 1024 の配列特徴を表すマップを出力している．Convolutional Neural Network (CNN) を用いて配列のマップを深層学習し，配列が酵素機能を有する (非酵素) か否か，および EC 番号 1 桁目クラス (EC1〜EC6) に対してどのクラスに属するかのクラス分類が行われた．さらに，転移学習を用いて，1 桁目が予測された配列の 2〜4 桁目を予測する手法を開発し，Macro F1 スコア 0.85 の従来の配列ベースの予測で最も高い精度を出力した．

化合物の物理・化学的特性値を用いた予測

MOLecular Mapping of Atom-level Properties (MOLMAP) 反応記述子 [4] を用いた手法では，68 種類の構造・物理・化学的指標を用い，EC 番号が判明している化学反応の分類が行われた．手順として，初めに，68 種類の化合物の特性値情報を Self-Organizing Map (SOM) [30] によって 2 次元平面上に凝縮マッピングし，7 × 7 から 29 × 29 個程度のセルで構成される数値の MOLMAP を生成した．次に基質と生成物の MOLMAP の差を取ることで酵素反応における化合物の化学変化を表現した．最終的に 55 種類の特性値によって生成される 25×25 の MOLMAP の差分が用いられ，反応式中全ての反応物と生成物を用いた反応データ約 7,500 に対し，Random Forests による 1〜3 桁目の分類が行われた．結果として，1 桁目，2 桁目，3 桁目それぞれの予測において，正解率 0.95, 0.90, 0.85 を得ている．

フィンガープリントの差分を用いた予測

反応物 (基質と試薬) から生成物に変化する際のフィンガープリントの変化に着目した手法として Differential Reaction Fingerprint (DRFP) [5] がある．具体的な実装手順を以下に示す．

1. 化合物内の各分子に着目し，分子単体，分子と 1〜2 つ隣に隣接する分子が結び付いた部分構造の SMILES を抽出する．
2. 抽出した全ての部分構造を反応物の集合と生成物の集合に分割し，対称差と取ることで，2 つの集合に重複する部分構造を除外する．
3. 各部分構造をを ECFP に基づいて 32 ビットの整数に変換し，10240 で割った余りをフラグ立てるビットとみなし，10240 ビットのフィンガープリントを作成する．

多層パーセプトロンを用いて EC 番号の 3 桁 (EC 7 を除く) まで予測し，F1-Score として Rhea [31] の反応データ (7010 データ) のみで 0.87±0.02, BRENDA [32], PathBank [33], MetaNetX [?] を含めたデータ (81205 データ) では 0.77±0.01 を得ている．

§ 3.3　機械学習と特徴量エンジニアリング

ここでは，本研究で用いる機械学習手法について述べる．

多クラス分類による機械学習予測

2 種類のいずれかの正解ラベルが割り振られたデータを分類する 2 クラス分類に対して，3 種類以上の正解ラベルのいずれかが付与されたデータを分類する問題は多クラス分類と呼ばれる．多クラス分類では，学習データを分類器に入力することでデータの分類体系を近似した分類モデルを構築し，テストデータを入力することで分類モデルがテストデータを正しく分類できるかを評価する．多クラス分類を評価する指標を以下に示す．なお簡便化のため，2 クラス分類を元に説明する．

Precition, Recall および Accuracy

D 個のデータが 2 つのクラス正, 負のいずれかに割り当てられているとする．また，分類モデルが正に分類されるデータを正しく正と予測した数を True Positive (TP), 負に分類されるデータを誤って正と予測した数を False Positive (FP) とする．さらに，負に属するデータを負と予測した数を True Negative (TN)，正のデータを誤って負と予測した数を False Negative (FN) とする．このとき，分類器が正に分類したデータのうち，実際に正に所属するデータであった割合を Presicion，正に所属するデータに対して分類器が正しく正に分類した割合を Recall と呼び，以下の式で表される．

$$Precition = \frac{TP}{TP + FP} \tag{3.1}$$

$$Recall = \frac{TP}{TP + FN} \tag{3.2}$$

Presicion は分類モデルの正確性を測る指標に対して，Recall はデータを正しいクラスであると判断する感度を表す指標であり，トレードオフの関係となっている [27]．また，全体の予測精度を表す指標である Accuracy は以下の式で表される．

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{3.3}$$

$$\tag{3.4}$$

多クラス分類の場合，1 つのクラスに着目し，着目したクラスを正，それ以外のクラスを負のクラスとみなして Precition，Recall が計算される．そして，各クラスの平均を取った，Macro Average Precition，Macro Average Recall が用いられる．

F1-Score

Presicion と Recall のバランスを測る指標で，それぞれの調和平均を取った指標となる．F1-Score (F1) を表す式は以下のようになる．

$$F1 = \frac{2Precidion \cdot Recall}{Precision + Recall} \tag{3.5}$$

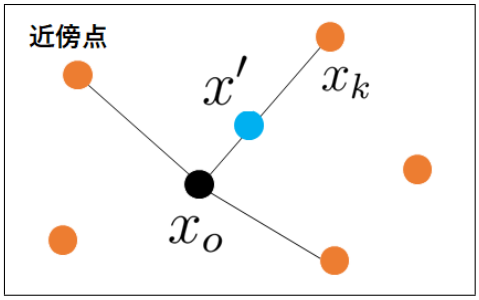


図 3.3: SMOTE によるオーバーサンプリングのイメージ

(3.6)

多クラス分類では，Macro Average Precition，Macro Average Recall の調和平均である，Macro F1-Score が用いられる．クラス間でデータに偏りがある場合，Accuracy ではデータ数の多い多数クラスに影響され，少数クラスに対する分類精度が低い場合でも，多数クラスの分類精度が高ければ，精度が高い分類モデルと評価される．一方で，Macro F1-Score は各クラスに対する分類精度を評価し，平均を取っているため，各クラスの重要度が等しい場合の評価に適している．

オーバーサンプリング

各クラスのデータ数に偏りが生じているデータを不均衡データと呼ぶ．多クラス分類において，少数クラスデータが誤って多数クラスに分類される可能性が高い，もしくは多数クラスのデータが少数クラスに誤分類される可能性が低くくなるのは必然的である．その対処法として，少数クラスのデータを仮想的に増やすオーバーサンプリングや多数クラスのデータを減らすアンダーサンプリングなどの手法がある．ここでは，Synthetic Minority Over-sampling Technique (SMOTE) [35] を用いたオーバーサンプリングについて述べる．SMOTE はデータが多次元の特徴量を持った特徴空間において，少数クラスに属するデータ間で線分を取り，線分上に同じ少数クラスのデータを新たに生成することでオーバーサンプリングを行う手法である．具体的なアルゴリズムは次のようになる．

1. ある少数クラスの 1 つのデータに着目し，図 3.3 のようにデータの近傍にある K 個の同じ少数クラスのデータを抽出する．
2. K 個の近傍データから 1 つを選択し，以下の式のように，データ点 x_o と選択された近傍データ点 x_k の間に新たなデータ点 x' が生成される．
   ただし，r はランダムに生成される $[0, 1]$ の値である．

$$x' = x_o + r(x_k - x_o) \tag{3.7}$$

Random Forests による多クラス分類

Random Forests (RF) [36] 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

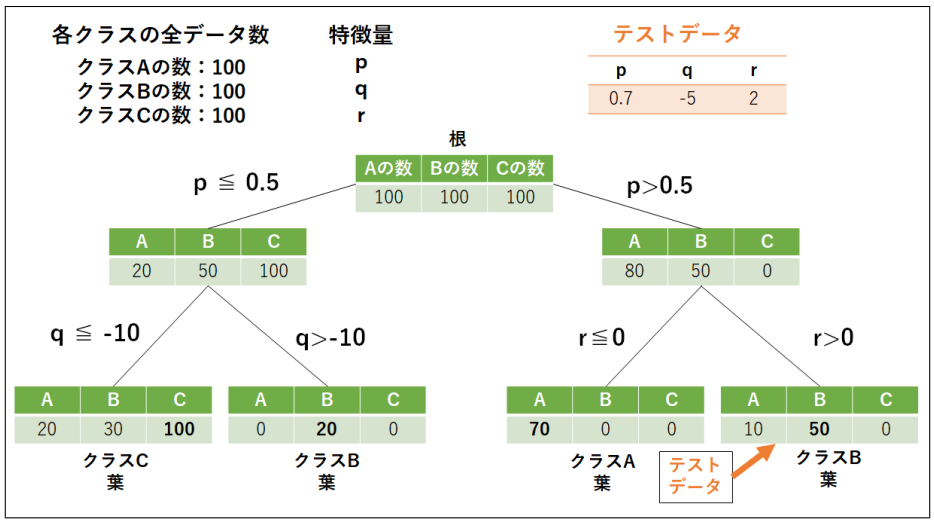


図 3.4: 決定木におけるクラス分類の様子

ができる．決定木のイメージを図 3.4 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 [37]．

$$IG(D_P, f) = I(D_p) - \frac{N_{left}}{N_p} I(D_{left}) - \frac{N_{right}}{N_p} I(D_{right}) \tag{3.8}$$

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．ジニ不純度を用いた場合，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．

$$I_G(t) = \sum_{i=1}^{c} p(i|t)(1 - p(i|t)) = 1 - \sum_{i=1}^{c} p(i|t)^2 \tag{3.9}$$

$p(i|t)$ 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．

基本的に全ての葉ノードで $I_G(t) = 0$，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 (分割されるデータの最大の階層数) を設定することで分割を打ち切ることもある．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．

RF 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，特徴量も一部 (または全て) 選択されて決定木に用いられる．それによって，異なるデータと特徴量からなる決定木モデルを作成し，組み合わせることで汎化性能を高める．

RF ライブラリである Python の sklearn.ensemble.RandomForestClassifier [38] 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る平均値が最も高いクラスとなる．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．

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, 基質から生成物の化学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 [39]．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．

第4章

提案手法

§ 4.1 特徴ベクトルと特徴エンジニアリングによるEC番号予測

有機合成では，反応の効率性や環境などの観点で目的の化合物を生成する際に酵素を生体触媒として用いられる機会が増加している．酵素にはEC番号が割り当てられており，酵素の性質の特定や触媒する化学反応の探索効率化のため，機械学習で入力された酵素に対するEC番号を予測する研究が行われてきた．手法としてタンパク質配列や化合物の構造的特徴，物理・化学的特性値などが用いられてきたが，これらのEC番号予測手法を有機合成に適用することで，探索する酵素の種類に対して絞り込みができ，実験によって最適な酵素を探索する時間やコストの削減が期待できる．

本研究では，有機合成における酵素探索に焦点を当て，化学反応に用いるべき最適な酵素の候補をEC番号として予測する機械学習モデルを開発する．提案手法として，従来化学反応の探索や予測の場面で多く用いられてきた210種類のRDKit記述子を用い，Random Forest (RF) に特徴選択とオーバーサンプリングを組み合わせた予測モデルを作成する．

有機合成の観点から見れば，従来手法の物理・化学特性値やフィンガープリントを用いた予測手法が適用可能であり，最近の研究ではフィンガープリントの差分を用いた手法 (DRFP) で優れた予測精度が得られている．主に，フィンガープリントは化合物の構造や化学反応時の構造変化を捉える指標であるため，物理・化学的な情報を加えることで，酵素を用いた化学反応の特徴をより詳細に捉えることが可能になると考えられる．しかし，フィンガープリントは次元が1000や2000を超えるものがあり，物理・化学的特性値と組み合わせた場合，学習コストが増大する懸念がある．そこで，RDKitで実装されている，125種類の物理・化学的特性値とMACCS Keysの表現法に類似する85種類の分子断片のバイナリ値を用いることで，学習コストを抑えつつ，各EC番号における化学反応の特徴をより説明できるようなモデルの作成が期待できる．

特徴ベクトルの作成

RFに入力するデータとして，基質から生成物に変化する際のRDKit記述子の変化を用いる．EC番号が割り当てられた酵素反応式において，各化合物の210種類の物理・化学的特性値を計算し，基質の特性値の和と生成物の特性値の和に関する差を取った特性値変化量を計算する．例えば，図4.1において，基質A, Bと生成物C, Dからなる反応式を考えた場合，A, B, C, Dはそれぞれ210次元の特性値を持ち，(A + B) - (C + D) を計算することで1つの反応式が210次元の特性値変化量を持つことになり，これを特徴ベクトルと定

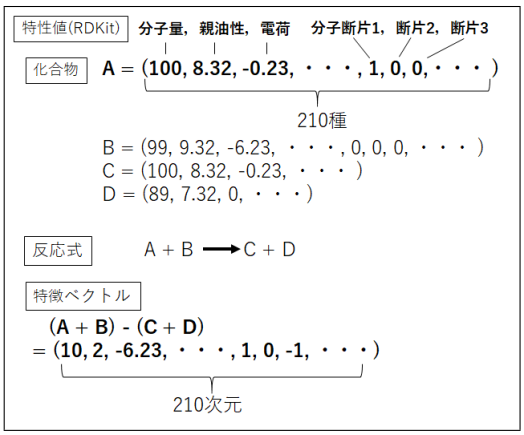


図 4.1: 特徴ベクトルのイメージ

EC	MaxAbsEState Index	MinAbsEState Index	MinEState Index	qed	...
4.1.1.74	-7.449074	-0.064815	-2.62963	-0.360209	...
1.2.1.8	-0.197403	0.405116	1.30787	-0.079826	...
2.5.1.85	0.593569	-2.312624	0.196239	0.488055	...
1.4.1.4	0.234718	0.418052	-0.413194	0.389325	...
1.1.1.3	-0.15593	0.059255	-0.317778	0.016389	...
...	...	...	...	...	...
4.4.1.13	-7.622896	-1.049639	0.2744	-0.411722	...
4.4.1.13	-7.478685	-1.049639	0.25852	-0.440793	...
2.3.1.-	-0.286151	0.454936	0.039395	-0.386083	...
2.3.1.-	-0.306521	0.11352	0.039395	-0.375732	...
2.3.1.-	-0.254718	0.523159	0.028495	-0.386848	...

表 4.1: 実際の特徴ベクトル

義する．最終的に表 4.1 のような，行に EC 番号，列に記述名が記載された反応データを用いる．

EC 番号予測における記述子選択

RDKit 記述子を全て用いて RF のクラス分類を行った場合，RF は各ノードで 1 つの記述子の閾値によって特徴ベクトルを分類する．それによって，ある EC 番号の特徴ベクトルを説明するのに不要な記述子やデータに同様の特徴を付与する記述子が複数存在することで，異なる EC 番号の酵素反応の区別があいまいになる可能性がある．そのため，記述子選択によって必要な記述子のみを分類に用いる．本研究では Python の mlxtend.feature_selection にある SequentialFeatureSelector を適用し，Step Forward 法を用いる．また，記述子数が 30 になるまで記述子選択を継続し，評価が最も高い記述子の組合せを用いる．最終的に 210 種類から RF の分類精度が最も高くなる記述子 10〜30 程度の組み合わせを用いる．

オーバーサンプリングによる少数クラスのデータ増大

EC 番号が割り当てられた酵素反応式もクラスごとにばらつきがあり，不均衡データである．背景として酵素の汎用性と機械学習の対象となるデータの形式が関わっている．例えば，EC3.1.1.3 に分類されるリパーゼは様々な反応に作用するため，産業分野での汎用性が高く，多くの酵素製品が生産されている [40]．そのため，提供されている酵素反応のデータが多い．一方で，その他の EC 番号は酵素反応データが 5 種以下のクラスなども多くあり，新たな酵素に EC 番号を割り当てる作業にも時間を要する．よって，汎用性が高い酵素の EC 番号は急増するが，応用機会が少ないものは年々ほとんど増加しないため，多数クラスと少数クラスのデータ数の隔たりは拡大していく一方となる．

本研究ではオーバーサンプリング手法の 1 種である SMOTE を適用する．多数クラスのデータ数を減らすアンダーサンプリングも考えられるが，多数クラスの重要なデータを削除してしまう可能性が挙げられており [41]，多数クラスを特徴づける説明力が減少する可能性を考慮し，今回は用いないこととする．代わりに少数クラスのデータを仮想的に増やすことで，少数クラスに対する分類精度の向上を図る．数値実験では Python の imblearn.over_sampling [42] に搭載されている SMOTE を用い，近傍データ数 K(k_neighbors) を $K = 3$ に設定した．また，デフォルトでは少数クラスのデータ数を最多クラスのデータ数に合わせてオーバーサンプリングされるが，最少クラスに対する最多クラスの比率が 2000

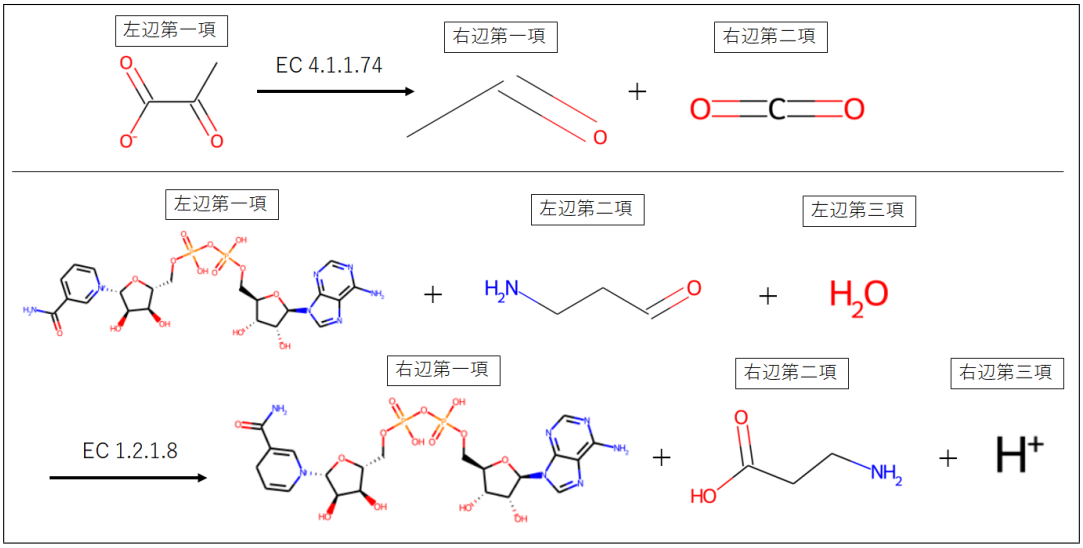


図 4.2: 構造式オブジェクトに変換された酵素反応

を超えるデータを用いるため，学習時間が元のデータ数の場合とに比べて膨大になることが懸念される．そのため，オーバーサンプルデータ数 (sampling_strategy) を合計データ数に合わせて調整するものとする．

§ 4.2　EC 番号予測モデルの構築と予測

本研究では，酵素反応ベースで EC 番号予測を行った従来手法 [4] [5] と同様に EC 番号の 1 桁目から 3 桁目を予測するモデルとする．データセットとして Rhea, BRENDA, PathBank, MetaNetX の 4 つのデータベースからなる，図 4.2 のような SMILES 形式で記載された酵素反応 [5] を用いる．初めにデータセットを反応式の左辺と右辺に分解し，各化合物を図 4.2 のような RDKit の構造式オブジェクトに変換する．次に，210 種類の記述子を用いて各化合物の特性値変化量を計算し，EC 番号と 210 次元の酵素反応式の特徴ベクトルを取得する．次に，適切な EC 番号が割り当てられていないものや SMILE から特性値変化量に変換した際に無効値となる要素を含む特徴ベクトルを削除した．また，記述子間とデータ間で相関係数が 1 のものや記述子選択に用いることが不可能な大きな値を要素に持つ特徴ベクトルを削除した．さらに，5 分割交差検証を実施するため，特徴ベクトルが 5 以下のクラスを除外した．最終的に，47090×200 記述子のデータを用いて予測モデルの構築と予測を行う．

データ数 47090 のうち 80%を学習データ，20%をテストデータとし，各クラス均等な割合で行き渡るように分割する．EC 番号予測は記述子選択と学習，予測の 3 段階で構成される．なお，それぞれの段階で RF を用い，決定木における各ノードの分割と用いられる記述子の基準としてジニ不純度を用いる．また，決定木の各ノード分割において，RF で指定する記述子数のみでは適切に特徴ベクトルを分割できない可能性を考慮し，記述子選択された全ての記述子を用いる．実装は Python の sklearn.ensemble.RandomForestClassifier を用い，記述子選択の段階ではデフォルトのパラメータ，モデル作成と予測には設定されたパラメータを用いる．図 4.3 にモデル構築と予測の流れを示す．

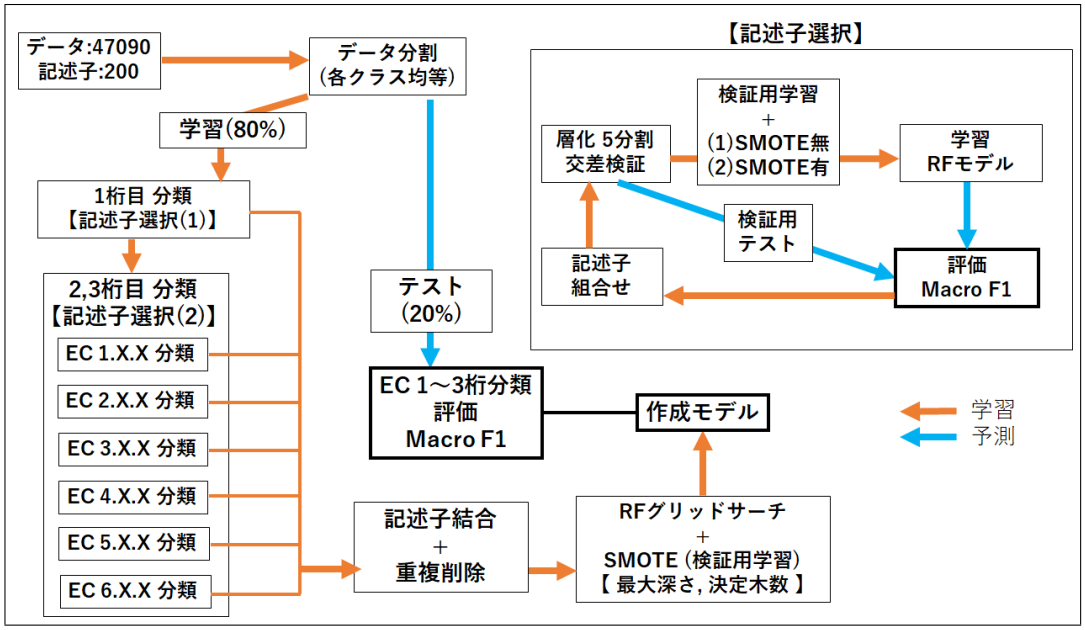


図 4.3: モデル構築と予測の流れ

記述子選択では学習データに対して RF を用い，EC 番号 1 桁目 (EC 1, EC 2, EC 3, EC 4, EC 5, EC 6) の多クラス分類と EC 番号 1 桁目に属する 2,3 桁目 (EC 1.X.X, EC 2.X.X, EC 3.X.X, EC 4.X.X, EC 5.X.X, EC 6.X.X) クラスの多クラス分類に分けて行う．なお EC 7 クラスは特徴ベクトル数が少ないため，本研究の予測対象外とする．このように分割した理由として，2 つ挙げられる．1 つ目は学習に要する時間が関係する．数値実験ではクラス数 148，データ数 37,672 の学習データを用いるため，学習に膨大な時間を要するためである．2 つ目は，EC 番号の分類法によるものである．EC 番号 1 桁目の分類で酵素の性質が大きく決まり，2 桁目 3 桁目と進むにつれて，性質が似通った酵素が同じクラスに属する構造となっているためである．そのため，1 桁目を分類する記述子と 2,3 桁目分類する記述子を分割して選択することで，分類粒度が異なる記述子を適切に選択できると考えられる．

7 回の記述子選択それぞれに 5 分割交差検証を実施し，学習データをさらに検証用学習データと検証用テストデータに分割する．そして，各分割の検証用学習データに対して SMOTE を適用する．なお，1 桁目分類では SMOTE を適用せず，素のデータで記述子選択を行っている．表 4.3 に各クラス分類における学習データの内訳と SMOTE ので増加させるデータ数を示す．少数クラスは表内の SMOTE 増加数以下のクラスとし，この閾値以下のクラス数を閾値の数になるようにオーバーサンプリングを行う．EC 2.X.X は最もデータが多く，記述子選択の学習コストを考慮して閾値をデータ数合計の 2%とした．EC 1.X.X は EC 2.X.X の 1/4 程度のデータだが，クラス数が多いため 3%とした．また，EC 3.X.X と EC 4.X.X は合計の減少に応じて割合を増やし，EC 5.X.X と EC 6.X.X は最多クラス数に合わせてオーバーサンプリングした．実際は，検証用学習データに SMOTE が適用され，学習データの分割率が 4:1 の場合を想定して，SMOTE 増加数 ×0.8 としたものを閾値に用いている．

記述子の組合せに対して検証用テストデータを適用し，Macro F1-Score で分類精度を評価する．新しい記述子が 1 つ追加されたときの組合せで，5 分割分の平均値が最も高いものを選択し，次の記述子選択を行う．30 種までの選択で，平均スコアの最高値が 4 回更新さ

れなくなる直前における記述子の組み合わせを，そのクラス分類で選ばれた記述子とする．最終的に 7 回分の記述子集合をマージし，重複を削除したものを学習に用いる．

学習では選択された記述子数の次元からなる学習データに対して，RF のパラメータ選択を行う．ここでは，EC 番号 1〜3 桁目の多クラス分類を実施する．方法としてグリッドサーチ (sklearn.model_selection.GridSearchCV [43]) を適用し，5 分割交差検証を実施する．記述子選択と同様に各分割の検証用学習データに SMOTE を適用し，Macro F1-Score で最適なパラメータの組み合わせを選択する．このとき，オーバーサンプリングの閾値は (最少クラスのデータ数)×100 とし，調整するパラメータとしては，決定木の数 (n_estimators) と最大深さ (max_depth) とする．

予測ではグリッドサーチで得られたパラメータと選択された記述子で作成されたモデルに対して，データ数 9,418 のテストデータを入力し，1〜3 桁目の 148 クラスに対する多クラス分類を行い，予測精度を評価する．評価指標として各クラスに対する Precidion, Recall, Macro F1-Score を用い，全体の平均値と Accuracy を出力する．

§ 4.3　提案手法の実装と流れ

SMILES 形式のデータの加工から特徴ベクトルの作成，機械学習モデルの構築から予測までの流れは，Jupyter Notebook 上で実装を行った．本節では，前半で提案システムの実装，後半で提案システムによる流れについて述べる．

SMILES 反応式の加工

図 4.2 の SMILES 形式の酵素反応式は rxn の項目に記載されている．中央の EC 番号を含んだ文字列で反応式の右辺左辺が分割されており，“.” で第一項，第二項などを区別されている．初めに，正規表現を用いて右辺と左辺の SMILES に分解した．このとき，MetaNetX のデータに含まれている，EC 2.A.3.1 などのデータを削除した．さらに文字列の “.” をもとにして右辺第一項，第二項，$\cdots$，左辺第一項，第二項のように，化合物単体の SMILE に切り分け，分割された SMILES の酵素反応式を取得した．次に SMILES 化合物を変換した構造オブジェクトで特性値を計算する際，無効な値を出力する SMILE 化合物の削除を行う．[Co+] や [Ca+2] などの電荷を含む SMILES の一部は記述子の正確な計算が行えない問題がある．そのため，右辺左辺から重複のない化合物 SMILES リストを作成し，特性値計算時に無効値が出る SMILES の除外リストを出力した．そして，先ほど取得した酵素反応式内で除外リストに該当する SMILES を検索し，右辺左辺いずれかに含まれる，反応式を削除した．除外リスト，および削除後の SMILES 反応式を表 4.4，4.5 に示す．

特徴ベクトルの作成

表 4.5 の右辺と左辺の応式を構造オブジェクトの反応式にそれぞれ変換し，RDKit 記述子を呼び出すことで，全ての化合物に 210 次元の特性値が付与される．左辺の特性値の和と右辺の和に対する差を取ることで，表 4.1 のような 210 次元の特徴ベクトルが得られる．ここで，全てのデータで値が等しくなる記述子の片方 (10 種類) と，同様に全ての記述子で値が等しい特徴ベクトルの片方を削除した．また，記述子選択で用いることができない float32 を超える要素や全て 0 の要素を持つ特徴ベクトルも削除した．

機械学習モデルの構築

Removal List
OP(=O)([O-])OC[C@H]1O[C@@H](n2cnc3c(=O)n4c(CCC([NH3+])C(=O)OC)c(C)nc4n(C)c32)[C@H](O)[C@@H]1OP(=O)([O-])O
O[As](O)c1ccccc1
*OP(=O)([O-])OCC(C)(C)C(O)C(=O)NCCC(=O)NCCSC(=O)CC(=O)CCCCCCCCCCCCCCC
*O[C@@H]1O[C@H](CO)[C@H](O)[C@H](O)[C@H]1NC(C)=O
*N[C@@H](CCCCNC(=O)CCCC[C@H]1CCSS1)C(*)=O
. . .
*OP(=O)([O-])O[C@H]1[C@@H](O)[C@H](*)O[C@@H]1COP(=O)([O-])O[C@@H]1[C@@H](COP(=O)([O-])OP(=O)([O-])OP(=O)([O-])[O-])O[C@@H](n2cnc3c(=O)[nH]c(N)nc32)[C@@H]1O

表 4.4: 除外リスト

左辺第一項	左辺第二項	左辺第三項	左辺第四項
CC(=O)C(=O)[O-]			
NC(=O)c1ccc[n+]([C@@H]2O[C@H](COP(=O)(O)OP(=O)(O)OC[C@H]3O[C@@H](n4cnc5c(N)ncnc54)[C@H](O)[C@@H]3O)[C@@H](O)[C@H]2O)c1	NCCC=O	O	
C=C(C)CCOP(=O)([O-])OP(=O)([O-])[O-]	CC(C)=CCOP(=O)(O)OP(=O)(O)O		
N	NC(=O)C1=CN([C@@H]2O[C@H](COP(=O)(O)OP(=O)(O)OC[C@H]3O[C@@H](n4cnc5c(N)ncnc54)[C@H](OP(=O)(O)O)[C@@H]3O)[C@@H](O)[C@H]2O)C=CC1	O=C([O-])CCC(=O)C(=O)[O-]	[H+]
NC(=O)C1=CN([C@@H]2O[C@H](COP(=O)(O)OP(=O)(O)OC[C@H]3O[C@@H](n4cnc5c(N)ncnc54)[C@H](OP(=O)(O)O)[C@@H]3O)[C@@H](O)[C@H]2O)C=CC1	NC(CC=O)C(=O)[O-]		
. . .	. . .	. . .	
CC(=O)SCCNC(=O)CCNC(=O)[C@H](O)C(C)(C)COP(=O)([O-])OP(=O)([O-])OC[C@H]1O[C@@H](n2cnc3c(N)ncnc32)[C@H](O)[C@@H]1OP(=O)([O-])[O-]	[NH3+]CCSCCC([NH3+])C(=O)[O-]		

表 4.5: SMILE 反応式 (左辺)

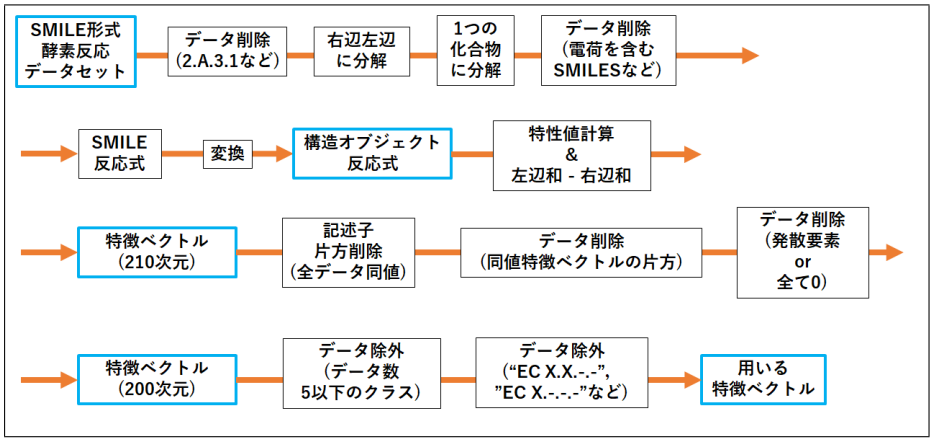


図 4.4: データ加工の流れ

初めに，特徴ベクトル数が 5 以下のクラスを削除した．これは学習データにおいて最少クラスのデータ数が 5 で 5 分割交差検証を行えるようにし，さらに検証用学習データに分割した際に，最少クラスのデータ数が 4 で近傍データ数 $K = 3$ による SMOTE を実施できるようにするためである．次に正規表現を用い，3 桁まで表示されている (EC X.X.-.-や EC X.-.-.-を除く)EC 番号を選択し，EC 番号 3 桁までの分類クラスを作成した．ここまでの流れを図 4.4 に示す．このデータに対して，各クラスで 4:1 の割合になるように学習データとテストデータに分割する．

次に，学習データから EC 番号 1 桁目のラベルのみ抽出し，1 桁目分類に対する記述子選択を行った．ここでは， mlxtend.feature_selection (SFS) に RF 分類器を入力し，Macro F1-Score の交差検証平均スコアで 30 種まで選択した．また，EC T.X.X (T= $1, 2, \cdots, 6$) における記述子選択では，EC T.X.X のラベルを抽出し，SFS に SMOTE と RF を組み合わせたパイプラインを入力するように設定した．これにより，検証用学習データにオーバーサンプリングを適用し，同様の記述子選択を実施した．選択記述子による評価値においてスコアが 4 回更新されなくなった時点に注目し，その直前の記述子の組み合わせを選出する．

4 回目を選んだ理由として，3 回目ではスコアの低い状態で打ち止めとなり，5 回目付近ではスコアの増加速度が低下し，データに対して類似の記述子が増加する懸念があったためである．

全ての EC T.X.X 学習データに対し，7 回の記述子選択で選ばれた記述子をマージし，重複を削除した記述子リストの要素のみを取り出すことで，表 4.1 のようなデータから次元削減された特徴ベクトルを取得する．最後に sklearn.model_selection の GridSearchCV を呼び出し，上記と同様に SMOTE と RF を組み合わせたパイプラインを入力し，検証用学習データをオーバサンプリングしたうえでパラメータが選択された．ここで，n_estimators は 10〜1000，max_depth は 1〜30 の 182 通りの組合せでグリッドサーチが行われた．

最後に Macro F1-Score の最も高い，パラメータ組合せで作成されたモデルにテストデータを入力し，分類精度を評価した．sklearn.metrics の classification_report を用いることで，Precision, Recall, F1-Score, Accuracy を出力した．

提案システムの流れ

以下の手順で提案手法を実施する．

1. 200 次元，特徴ベクトル数 47090 のデータを入力し，4:1 の割合で学習データと，テストデータに分割する．
2. 5 分割交差検証を実施し，それぞれの分割で学習データを検証用学習データと検証用テストデータに分割する．
3. EC 番号 1 桁目の分類に対する記述子選択を行う．
4. 検証用データに SMOTE を適用し，EC 番号 T.X.X (T$= 1, 2, \cdots, 6$) の分類に対する記述子選択を行う
5. 各記述子選択で選ばれた記述子をマージし重複を削除した記述子の組み合わせ (最終記述子) を作成する．
6. 最終記述子次元の学習データに対してグリッドサーチを実行する．(5 分割交差検証で検証データに SMOTE を実行)
7. 決定したパラメータで作成した RF モデルにテストデータを適用し，Macro Average Precision・Recall，Macro F1-Score，Accuracy で EC 番号の 1〜3 桁目分類の分類精度を評価する．

表 4.2: SMILE 形式のデータセット（一部抜粋および改変）[5]

rxn	ec	source	rxn	ec	source
CC(=O)C(=O)[O-] |4.1.1.74>>| CC=O.O=C=O	4.1.1.74	brenda_ reaction_smiles	N[C@@H](CCC(=O)O)C(=O)O.O=C(O)C(=O)Cc1ccccc1 |2.6.1.57>>| N[C@@H](Cc1ccccc1)C(=O)O.O=C(O)CCC(=O)C(=O)O	2.6.1.57	pathbank_ reaction_smiles
NC(=O)c1ccc[n+]([C@@H]2O[C@H](COP(=O)(O)OP(=O)(O)OC[C@H]3O[C@@H](n4cnc5c(N)ncnc54)[C@H](O)[C@@H]3O)[C@@H](O)[C@H]2O)c1.NCCC=O.O |1.2.1.8>>| NC(=O)C1=CN([C@@H]2O[C@H](COP(=O)(O)OP(=O)(O)OC[C@H]3O[C@@H](n4cnc5c(N)ncnc54)[C@H](O)[C@@H]3O)[C@@H](O)[C@H]2O)C=CC1.NCCC(=O)O.[H+]	1.2.1.8	brenda_ reaction_smiles	O=CC(O)COP(=O)(O)O |5.3.1.1>>| O=C(CO)COP(=O)(O)O	5.3.1.1	pathbank_ reaction_smiles
・・・	・・・	・・・	・・・	・・・	・・・
NC(=O)CC[C@H]([NH3+])C(=O)[O-].Nc1ncnc2c1ncn2[C@@H]1O[C@H](COP(=O)([O-])OP(=O)([O-])[O-])[C@@H](O)[C@H]1O.O=P([O-])([O-])O.[H+] |6.3.1.2>>| Nc1ncnc2c1ncn2[C@@H]1O[C@H](COP(=O)([O-])OP(=O)([O-])OP(=O)([O-])[O-])[C@@H](O)[C@H]1O.[NH3+][C@@H](CCC(=O)[O-])C(=O)[O-].[NH4+]	6.3.1.2	metanetx_ reaction_smiles	C[NH+]1CCC[C@@H]1c1ccc(O)nc1.O.O=O |1.5.3.6>>| C[NH2+]CCCC(=O)c1ccc(O)nc1.OO	1.5.3.6	rhea_ reaction_smiles
NC(=O)CC[C@H]([NH3+])C(=O)[O-].O |1.4.1.13>>| [NH3+][C@@H](CCC(=O)[O-])C(=O)[O-].[NH4+]	1.4.1.13	metanetx_ reaction_smiles	COc1cc(/C=C/C(=O)OCC[N+](C)(C)C)cc(OC)c1O.O |3.1.1.49>>| COc1cc(/C=C/C(=O)[O-])cc(OC)c1O.C[N+](C)(C)CCO.[H+]	3.1.1.49	rhea_ reaction_smiles
・・・	・・・	・・・	・・・	・・・	・・・

表 4.3: 学習データの内訳と SMOTE 増加数

EC 1.X.X計	クラス数	SMOTE 増加数	1.1.1	1.14.13	1.2.1	1.14.14	1.3.1	1.13.11	1.14.11	1.4.3	・・・	1.23.5
6380	64	3%(191)	1745	761	666	408	348	242	173	156	・・・	5
EC 2.X.X計	クラス数	SMOTE 増加数	2.7.8	2.3.1	2.1.1	2.4.1	2.7.1	2.7.7	2.5.1	2.6.1	・・・	2.7.3
23160	24	2%(463)	10074	7309	2797	686	602	426	303	280	・・・	5
EC 3.X.X計	クラス数	SMOTE 増加数	3.1.1	3.1.3	3.6.3	3.2.1	3.5.1	3.6.1	3.1.2	3.1.4	・・・	3.3.1
5377	27	10%(538)	2277	589	508	448	347	237	169	104	・・・	5
EC 4.X.X計	クラス数	SMOTE 増加数	4.1.1	4.2.1	4.1.2	4.4.1	4.2.3	4.1.99	4.3.1	4.1.3	・・・	4.6.1
1878	14	20%(376)	1037	361	106	106	101	44	39	32	・・・	5
EC 5.X.X計	クラス数	SMOTE 増加数	5.5.1	5.3.1	5.3.3	5.4.99	5.3.2	5.1.3	5.4.2	5.4.3	・・・	5.1.1
273	12	最多(80)	80	46	44	20	15	13	13	12	・・・	5
EC 6.X.X計	クラス数	SMOTE 増加数	6.2.1	6.3.2	6.3.4	6.3.5	6.3.1	6.4.1	6.1.2			
604	7	最多(266)	266	233	30	29	22	18	6			

第5章

実験結果並びに考察

§ 5.1　数値実験の概要

本研究では提案手法の数値実験 (本実験) の前に予備実験を 2 つ行う．予備実験 1 では，従来行われた EC 3 の 2,3 桁目のクラス分類に対して SMOTE を適用し，SMOTE 未適用と適用した場合の分類精度を比較することで，オーバーサンプリングの有効性を示す．従来のクラス分類では，KEGG から取得した基質に水を加えることで基質が 2 つの生成物に分解する，左 2 辺，右 2 辺の加水分解反応の EC 番号分類が行われていた．962 種 20 クラスの特徴ベクトルに対して RF に特徴選択ととグリッドサーチを適用し，769 種のテストデータを用いて分類モデルを構築した．そして，193 種のテストデータに対して予測を行い，全体の Macro F1-Score，Accuracy としてそれぞれ，0.81, 0.92 の精度を得ている．しかし，クラスのデータ分布が 6〜155 となっており，少数クラスに対する分類精度が問題視されていた．表 5.1 にデータの内訳を示す．5 分割交差検証の検証用学習データに対して SFS に SMOTE と RF を組み込んだパイプラインを適用し，閾値を最多クラスとしてオーバサンプリングを行った．そして，グリッドサーチにもパイプラインを入力することでモデルを構築し，テストデータに対する予測精度の比較を行った．

予備実験 2 では，本実験のための記述子選択を行うが，記述子選択後に各クラス分類で，グリッドサーチを適用し，EC 番号 1 桁目分類，EC T.X.X (T$= 1, 2, \cdots, 6$) クラス単体でのテストデータに対する分類を行う．図 5.1 に実験手順を示す．本実験には，30 種までの記述子組合せの中で Macro F1-Score が 4 回未更新状態になる直前の組合せが用いられるが，この実験では，最も高くなる記述子の組合せを用いる．また，いずれも SMOTE と RF のパイプラインを適用し，7 回それぞれのグリッドサーチで異なるパラメータが決定されることで，特定のクラスに特化した分類モデルが作成される．

本実験では，記述子組合せを全てマージし，重複削除したものを選択記述子として用いる．EC 番号 3 桁分 (EC X.X.X) の多クラス分類に対して，データ数 37672 の学習データに対して，グリッドサーチに SMOTE と RF のパイプラインを適用した 5 分割交差検証によるパラメータ調整を行う．ここで，検証用学習データの少数クラスの増量閾値を，最少クラス $\times 100 \times 0.8 = 400$ とし，n_estimators と max_depth の調整範囲を，それぞれ 100〜1000, 10〜100 の 100 通りでグリッドサーチが行われた．最後に最適パラメータで作成されたモデルに，データ数 9418 のテストデータを入力し，平均の Precision, Recall, F1-Score, および Accuracy で分類精度を評価した．

表 5.1: EC 3 の 2,3 桁目クラスのデータの内訳

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

§ 5.2 実験結果と考察

2 つの予備実験の結果，および本実験の結果と考察を述べる．

予備実験 1 の結果

従来の EC 3 の 2,3 桁目に対する RF×SFS 分類モデルと，新たに SMOTE を適用した分類モデルの比較結果を図 5.2 に示す．Macro F1-Score が従来に比べて 0.06 向上し，SMOTE の有効性が示された．

予備実験 2 の結果

EC 番号 1 桁目，EC T.X.X (T= $1, 2, \cdots, 6$) に対する各クラス分類で得られた記述子リスト，およびテストデータに対するクラス分類結果を表 5.2, および図 5.3, 5.4, 5.5, 5.6, に示す．得られた記述子をマージし，重複を削除すると，93 種の記述子が得られた．

本実験の結果

選ばれた 93 種の記述子を用いてグリッドサーチを行うと, パラメータとして n_estimators=300, max_depth=90 が得られた．テストデータに対する分類結果は表 5.3, 5.4 のようになった．

考察

図 5.5 に本実験での EC1〜EC6 それぞれの分類精度の平均を示す．EC 5 が最も分類精度が低かった原因として，データ数が他のクラスに比べて少ないことの他に，EC 5 と他クラスの酵素反応が類似している点が挙げられる．DRFP を用いた酵素反応のクラスタリング結果において，EC 5 の酵素が他クラスのクラスタに多く混在していることが示されており [5]，類似性の高い他クラスの酵素として誤分類された可能性が考えられる．

同様に，データ数が類似している EC 1.X.X と EC 3.X.X の Macro F1-Score に着目すると，データ数が多い EC 3.X.X の方が精度が低くなっている．このことから，データ数の多さだけでなく，クラス数の多さも分類精度に影響してくることがうかがえる．

最終的な分類精度が Macro F1-Score 0.79 となり，DRFP を用いた手法 [5] の EC 番号 1〜3 桁目における分類精度 0.77± 0.01 とほぼ同等の性能となった．しかし，先行研究と比べ

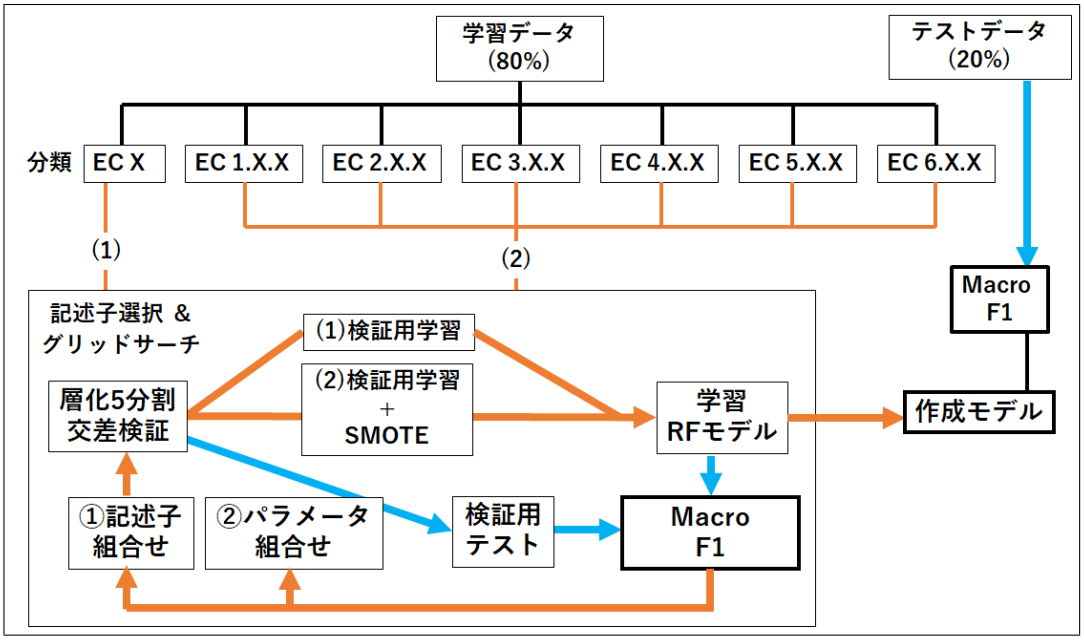


図 5.1: 実験 2 の流れ

る記述子選択とパラメータ調整に多くの時間がかかる点．また，SMILES を RDKit 構造オブジェクトに変換した際，特性値を計算できない化合物を含む反応式を削除したため，先行研究よりもデータ数が少ない点などから，予測精度の改善が望まれる．

EC 番号 1 桁目，EC T.X.X (T= $1, 2, \cdots, 6$) 分類で選択された記述子の共通の特徴や傾向を分析することで，EC 番号を振り分けるより詳細な特徴を抽出できると考えられる．また，選択された記述子間で類似の傾向を持つ記述子どうしをまとめることで次元の圧縮を行い，さらに RDKit 以外の記述子を用いる．特徴選択時，分類に大きく貢献する記述子を評価し，重みづけを行うなども分類精度向上の一種として期待できる．

フィンガープリント手法と物理・化学的特性値を比較した場合，フィンガープリントはバイナリ値のため，学習時間が短いが化学反応を捉える特徴は主に構造変化という点．物理・化学的特性値は化学反応に関する情報を多く取り扱うが，学習時間が長い点と，トレードオフの関係が考えられる．一方で分類精度をより向上させるためには，物理・化学的記述子とフィンガープリントを最適な割合で用いることが必要である考えられる．例えば，特性値を 10 個以下まで絞り込むことができれば，長さ 1000 以上のフィンガープリントの次元に対して学習コストの増加が少ない．そのため，従来のフィンガープリント手法の学習時間に影響を与えず，酵素反応をより詳細に捉え，高精度に分類できるモデル作成が期待できる．

従来モデル

	precision	recall	f1-score	support
3.1.1.	0.96	0.96	0.96	25
3.1.2.	0.92	1.00	0.96	12
3.1.3.	0.91	0.94	0.92	31
3.1.4.	0.86	1.00	0.92	6
3.1.6.	1.00	1.00	1.00	3
3.1.7.	0.00	0.00	0.00	2
3.13.1.	1.00	0.50	0.67	2
3.2.1.	0.96	0.96	0.96	26
3.2.2.	0.83	1.00	0.91	5
3.3.2.	1.00	1.00	1.00	1
3.4.13.	0.00	0.00	0.00	1
3.4.19.	1.00	1.00	1.00	1
3.5.1.	0.94	0.97	0.95	31
3.5.3.	0.83	1.00	0.91	5
3.5.4.	0.89	0.89	0.89	9
3.5.5.	1.00	1.00	1.00	2
3.5.99.	1.00	0.50	0.67	2
3.6.1.	0.86	0.95	0.90	19
3.7.1.	1.00	0.71	0.83	7
3.8.1.	1.00	0.67	0.80	3
accuracy			0.92	193
macro avg	0.85	0.80	0.81	193
weighted avg	0.91	0.92	0.91	193

SMOTE適用後

	precision	recall	f1-score	support
3.1.1.	1.00	0.96	0.98	25
3.1.2.	1.00	1.00	1.00	12
3.1.3.	0.97	0.94	0.95	31
3.1.4.	1.00	1.00	1.00	6
3.1.6.	0.75	1.00	0.86	3
3.1.7.	1.00	1.00	1.00	2
3.13.1.	0.50	0.50	0.50	2
3.2.1.	1.00	0.96	0.98	26
3.2.2.	0.71	1.00	0.83	5
3.3.2.	1.00	1.00	1.00	1
3.4.13.	0.00	0.00	0.00	1
3.4.19.	1.00	1.00	1.00	1
3.5.1.	0.94	0.97	0.95	31
3.5.3.	1.00	1.00	1.00	5
3.5.4.	0.88	0.78	0.82	9
3.5.5.	1.00	1.00	1.00	2
3.5.99.	0.67	1.00	0.80	2
3.6.1.	0.89	0.89	0.89	19
3.7.1.	0.88	1.00	0.93	7
3.8.1.	1.00	0.67	0.80	3
accuracy			0.94	193
macro avg	0.86	0.88	0.87	193
weighted avg	0.94	0.94	0.94	193

図 5.2: 予備実験 1 の結果

	precision	recall	f1-score	support
1.1.1	0.93	0.96	0.94	440
1.1.3	0.86	0.76	0.81	33
1.1.5	0.79	0.94	0.86	16
1.1.98	0.67	0.67	0.67	3
1.1.99	0.64	0.78	0.70	18
1.10.3	0.86	0.86	0.86	7
1.10.5	0.00	0.00	0.00	1
1.10.99	0.50	0.50	0.50	2
1.11.1	0.81	0.85	0.83	20
1.11.2	0.64	0.70	0.67	10
1.12.98	1.00	1.00	1.00	1
1.13.11	0.88	0.95	0.91	61
1.13.12	0.69	0.60	0.64	15
1.13.99	0.00	0.00	0.00	1
1.14.11	0.82	0.84	0.83	44
1.14.12	0.96	0.77	0.85	30
1.14.13	0.80	0.90	0.85	192
1.14.14	0.91	0.71	0.80	102
1.14.15	0.50	0.25	0.33	4
1.14.16	1.00	0.60	0.75	5
1.14.17	0.75	1.00	0.86	3
1.14.18	0.73	0.67	0.70	12
1.14.19	0.91	1.00	0.95	10
1.14.20	0.78	0.82	0.80	17
1.14.21	1.00	0.50	0.67	4
1.14.99	0.95	0.86	0.90	22
1.17.1	1.00	0.50	0.67	8
1.17.3	0.50	0.67	0.57	6
1.17.5	1.00	0.50	0.67	2
1.18.1	0.00	0.00	0.00	2
1.2.1	0.96	0.96	0.96	170
1.2.3	0.69	0.69	0.69	16
1.2.4	1.00	0.89	0.94	9
1.2.5	0.50	1.00	0.67	1
1.2.99	0.67	1.00	0.80	2
1.20.1	1.00	1.00	1.00	1
1.21.1	1.00	1.00	1.00	2
1.21.3	0.50	0.67	0.57	3
1.23.5	1.00	1.00	1.00	1
1.3.1	0.92	0.92	0.92	85
1.3.3	0.87	0.81	0.84	16
1.3.5	1.00	0.92	0.96	13
1.3.8	1.00	0.92	0.96	13
1.3.98	0.33	1.00	0.50	1
1.3.99	1.00	0.33	0.50	3
1.4.1	0.90	1.00	0.95	18
1.4.3	0.92	0.85	0.88	39
1.4.5	1.00	1.00	1.00	2
1.4.99	1.00	1.00	1.00	5
1.5.1	0.92	0.85	0.88	27
1.5.3	0.67	0.57	0.62	7
1.5.5	0.33	0.50	0.40	2
1.5.99	0.86	1.00	0.92	6
1.6.3	0.67	1.00	0.80	2
1.6.5	0.85	0.81	0.83	27
1.7.1	1.00	0.85	0.92	13
1.7.2	1.00	1.00	1.00	1
1.7.3	1.00	1.00	1.00	4
1.7.5	1.00	0.50	0.67	2
1.7.99	0.50	1.00	0.67	1
1.8.1	1.00	0.57	0.73	7
1.8.3	1.00	1.00	1.00	2
1.8.5	0.88	1.00	0.93	7
1.97.1	1.00	1.00	1.00	2
accuracy			0.88	1601
macro avg	0.79	0.78	0.77	1601
weighted avg	0.88	0.88	0.88	1601

図 5.3: EC 1.X.X に対する分類

表 5.2: 予備実験 2 で選ばれた記述子リスト

	1	2	3	4	5	6	7	8	9	10	11	12	13	14
EC X	MinEState Index	MolWt	NumValence Electrons	FpDensity Morgan1	BCUT2D _MWHI	BCUT2D _MWLOW	BCUT2D _CHGHI	HallKierAlpha	Ipc	Kappa3	PEOE_VSA4	SlogP_VSA2	SlogP_VSA3	FractionCSP3
EC 1.X.X	MinEState Index	BCUT2D _MRLOW	BalabanJ	Chi4v	Kappa1	Kappa3	PEOE_VSA4	SMR_VSA1	SMR_VSA10	SMR_VSA3	SMR_VSA7	SlogP_VSA2	EState_VSA1	EState_VSA3
EC 2.X.X	AvgIpc	Chi1	Ipc	Kappa1	PEOE_VSA2	SlogP_VSA12	SlogP_VSA3	EState_VSA4	VSA_EState1	VSA_EState6	VSA_EState7	VSA_EState8	VSA_EState9	NumAliphatic Carbocycles
EC 3.X.X	BCUT2D _MWHI	BCUT2D _MWLOW	BCUT2D _LOGPLOW	Chi1	Chi2n	Chi2v	Chi3v	Chi4v	Kappa1	Kappa3	PEOE_VSA12	PEOE_VSA5	PEOE_VSA9	SMR_VSA10
EC 4.X.X	NumValence Electrons	BCUT2D _MWHI	BCUT2D _MWLOW	BCUT2D _LOGPHI	LabuteASA	SMR_VSA5	SMR_VSA6	VSA_EState2	VSA_EState6	VSA_EState7	RingCount	fr_Al_OH _noTert	fr_COO2	fr_C_O_ noCOO
EC 5.X.X	AvgIpc	BalabanJ	Chi1n	Chi2n	Chi3v	Chi4v	HallKierAlpha	VSA_EState5	Num Heteroatoms	fr_C_O	fr_aldehyde	fr_ester	fr_ketone _Topliss	**end13**
EC 6.X.X	BCUT2D _MWLOW	AvgIpc	HallKierAlpha	Ipc	EState_VSA3	VSA_EState6	Fraction CSP3	NumAromatic Heterocycles	fr_COO	fr_benzene	fr_ether	fr_halogen	fr_piperzine	fr_sulfide
	15	**16**	**17**	**18**	**19**	**20**	**21**	**22**	**23**	**24**	**25**	**26**	**27**	**28**
EC X	fr_Ar_COO	fr_COO	fr_COO2	fr_allylic_oxid	fr_lactone	**end 19**								
EC 1.X.X	EState_VSA7	EState_VSA8	VSA_EState6	VSA_EState8	VSA_EState9	NumAliphatic Heterocycles	NumSaturated Carbocycles	fr_Ar_COO	fr_Ar_OH	fr_NH2	fr_SH	fr_aniline	fr_guanido	**end27**
EC 2.X.X	fr_Al_OH	fr_C_S	fr_epoxide	fr_ether	fr_lactone	fr_phos_ester	**end20**							
EC 3.X.X	SlogP_VSA8	VSA_EState6	NumAliphatic Rings	NumAromatic Heterocycles	NumRotata- bleBonds	NumSaturated Carbocycles	fr_C_O	fr_NH0	fr_guanido	fr_hdrzone	fr_ketone	fr_ketone _Topliss	fr_pyridine	fr_sulfide
EC 4.X.X	fr_NH1	fr_aldehyde	fr_benzene	fr_halogen	fr_methoxy	fr_phenol_no OrthoHbond	fr_sulfone	**end 21**						**↑ end28**
EC 5.X.X	**end13**													
EC 6.X.X	fr_unbrch _alkane	**end 15**												

	precision	recall	f1-score	support
2.1.1	1.00	1.00	1.00	694
2.1.2	1.00	1.00	1.00	4
2.1.3	1.00	0.60	0.75	5
2.1.4	0.00	0.00	0.00	2
2.2.1	0.80	0.89	0.84	18
2.3.1	1.00	1.00	1.00	1827
2.3.2	0.75	1.00	0.86	3
2.3.3	1.00	1.00	1.00	5
2.4.1	0.97	0.98	0.97	171
2.4.2	0.97	0.88	0.92	41
2.4.99	1.00	0.50	0.67	2
2.5.1	0.97	0.92	0.95	78
2.6.1	0.86	0.92	0.88	71
2.7.1	0.96	0.97	0.96	153
2.7.2	1.00	0.62	0.77	8
2.7.3	0.50	1.00	0.67	1
2.7.4	0.75	0.87	0.81	31
2.7.6	1.00	0.57	0.73	7
2.7.7	0.96	0.99	0.98	107
2.7.8	1.00	1.00	1.00	2518
2.7.9	1.00	0.50	0.67	2
2.8.1	1.00	0.50	0.67	4
2.8.2	0.95	0.95	0.95	21
2.8.3	0.93	0.88	0.90	16
accuracy			0.99	5789
macro avg	0.89	0.81	0.83	5789
weighted avg	0.99	0.99	0.99	5789

	precision	recall	f1-score	support
3.1.1	0.97	0.98	0.97	570
3.1.2	0.93	0.91	0.92	43
3.1.3	0.87	0.91	0.89	148
3.1.4	0.90	0.69	0.78	26
3.1.6	1.00	1.00	1.00	2
3.1.8	1.00	0.83	0.91	6
3.13.1	0.00	0.00	0.00	2
3.2.1	0.93	0.98	0.95	113
3.2.2	0.92	1.00	0.96	12
3.3.1	0.50	1.00	0.67	1
3.3.2	1.00	0.88	0.93	16
3.4.11	0.78	0.90	0.84	20
3.4.13	0.82	0.78	0.80	18
3.4.17	1.00	0.83	0.91	6
3.4.19	1.00	1.00	1.00	6
3.4.21	1.00	1.00	1.00	2
3.5.1	0.89	0.84	0.86	87
3.5.2	0.71	0.71	0.71	7
3.5.3	1.00	1.00	1.00	4
3.5.4	0.86	0.80	0.83	15
3.5.5	0.95	1.00	0.97	18
3.5.99	1.00	0.33	0.50	3
3.6.1	0.73	0.62	0.67	58
3.6.3	1.00	1.00	1.00	126
3.6.4	0.33	1.00	0.50	1
3.7.1	0.84	0.94	0.89	17
3.8.1	0.94	0.94	0.94	18
accuracy			0.93	1345
macro avg	0.85	0.85	0.83	1345
weighted avg	0.93	0.93	0.93	1345

図 5.4: EC 2.X.X, 3.X.X に対する分類

	precision	recall	f1-score	support
4.1.1	0.98	0.98	0.98	260
4.1.2	0.93	0.96	0.95	28
4.1.3	0.44	0.50	0.47	8
4.1.99	0.78	0.70	0.74	10
4.2.1	0.93	0.95	0.94	88
4.2.2	1.00	1.00	1.00	1
4.2.3	0.95	0.87	0.91	23
4.3.1	1.00	0.89	0.94	9
4.3.2	0.50	0.50	0.50	2
4.3.3	1.00	0.33	0.50	3
4.4.1	0.92	0.92	0.92	24
4.5.1	1.00	0.75	0.86	4
4.6.1	1.00	1.00	1.00	1
4.99.1	0.33	1.00	0.50	1
accuracy			0.94	462
macro avg	0.84	0.81	0.80	462
weighted avg	0.95	0.94	0.94	462

	precision	recall	f1-score	support
5.1.1	0.00	0.00	0.00	1
5.1.3	0.75	0.75	0.75	4
5.2.1	0.00	0.00	0.00	1
5.3.1	0.92	1.00	0.96	11
5.3.2	1.00	1.00	1.00	4
5.3.3	0.92	1.00	0.96	11
5.3.99	1.00	0.67	0.80	3
5.4.2	0.60	1.00	0.75	3
5.4.3	1.00	0.67	0.80	3
5.4.4	0.67	1.00	0.80	2
5.4.99	0.83	1.00	0.91	5
5.5.1	1.00	0.89	0.94	19
accuracy			0.90	67
macro avg	0.72	0.75	0.72	67
weighted avg	0.89	0.90	0.88	67

図 5.5: EC 4.X.X, 5.X.X に対する分類

	precision	recall	f1-score	support
6.1.2	0.50	0.50	0.50	2
6.2.1	0.99	0.99	0.99	67
6.3.1	0.86	1.00	0.92	6
6.3.2	0.98	0.98	0.98	59
6.3.4	0.86	0.75	0.80	8
6.3.5	0.71	0.71	0.71	7
6.4.1	1.00	1.00	1.00	5
accuracy			0.95	154
macro avg	0.84	0.85	0.84	154
weighted avg	0.95	0.95	0.95	154

	precision	recall	f1-score	support
1	0.98	0.98	0.98	1601
2	0.99	0.99	0.99	5789
3	0.97	0.98	0.98	1345
4	0.95	0.94	0.94	462
5	0.83	0.72	0.77	67
6	0.99	0.92	0.95	154
accuracy			0.98	9418
macro avg	0.95	0.92	0.94	9418
weighted avg	0.98	0.98	0.98	9418

図 5.6: EC 6.X.X, 1 桁目に対する分類

表 5.3: EC 番号 1 桁～3 桁目までの分類結果 1

EC Class	Num of Data	precision	recall	f1-score	EC Class	Num of Data	precision	recall	f1-score
1.1.1	440	0.95	0.95	0.95	**1.3.99**	3	1.00	0.33	0.50
1.1.3	33	0.89	0.76	0.82	**1.4.1**	18	0.89	0.94	0.92
1.1.5	16	0.84	1.00	0.91	**1.4.3**	39	0.95	0.95	0.95
1.1.98	3	1.00	0.67	0.80	**1.4.5**	2	1.00	1.00	1.00
1.1.99	18	0.76	0.72	0.74	**1.4.99**	5	1.00	1.00	1.00
1.10.3	7	0.57	0.57	0.57	**1.5.1**	27	0.86	0.93	0.89
1.10.5	1	0.00	0.00	0.00	**1.5.3**	7	1.00	0.57	0.73
1.10.99	2	1.00	0.50	0.67	**1.5.5**	2	0.50	0.50	0.50
1.11.1	20	0.80	0.80	0.80	**1.5.99**	6	0.86	1.00	0.92
1.11.2	10	0.70	0.70	0.70	**1.6.3**	2	0.67	1.00	0.80
1.12.98	1	0.50	1.00	0.67	**1.6.5**	27	0.79	0.85	0.82
1.13.11	61	0.92	0.93	0.93	**1.7.1**	13	1.00	0.85	0.92
1.13.12	15	0.71	0.67	0.69	**1.7.2**	1	1.00	1.00	1.00
1.13.99	1	1.00	1.00	1.00	**1.7.3**	4	1.00	1.00	1.00
1.14.11	44	0.90	0.84	0.87	**1.7.5**	2	0.50	0.50	0.50
1.14.12	30	0.96	0.80	0.87	**1.7.99**	1	0.00	0.00	0.00
1.14.13	192	0.83	0.87	0.85	**1.8.1**	7	0.80	0.57	0.67
1.14.14	102	0.88	0.71	0.78	**1.8.3**	2	1.00	1.00	1.00
1.14.15	4	0.33	0.25	0.29	**1.8.5**	7	0.78	1.00	0.88
1.14.16	5	1.00	0.60	0.75	**1.97.1**	2	1.00	1.00	1.00
1.14.17	3	0.75	1.00	0.86	**2.1.1**	694	1.00	1.00	1.00
1.14.18	12	0.56	0.75	0.64	**2.1.2**	4	1.00	1.00	1.00
1.14.19	10	0.83	1.00	0.91	**2.1.3**	5	1.00	0.80	0.89
1.14.20	17	0.74	0.82	0.78	**2.1.4**	2	0.00	0.00	0.00
1.14.21	4	0.67	0.50	0.57	**2.2.1**	18	0.88	0.78	0.82
1.14.99	22	0.90	0.86	0.88	**2.3.1**	1827	0.99	0.99	0.99
1.17.1	8	1.00	0.63	0.77	**2.3.2**	3	0.00	0.00	0.00
1.17.3	6	0.56	0.83	0.67	**2.3.3**	5	0.83	1.00	0.91
1.17.5	2	1.00	0.50	0.67	**2.4.1**	171	0.97	0.97	0.97
1.18.1	2	0.00	0.00	0.00	**2.4.2**	41	0.94	0.83	0.88
1.2.1	170	0.97	0.96	0.97	**2.4.99**	2	1.00	1.00	1.00
1.2.3	16	0.68	0.81	0.74	**2.5.1**	78	0.92	0.86	0.89
1.2.4	9	0.88	0.78	0.82	**2.6.1**	71	0.84	0.86	0.85
1.2.5	1	1.00	1.00	1.00	**2.7.1**	153	0.95	0.95	0.95
1.2.99	2	1.00	1.00	1.00	**2.7.2**	8	0.83	0.63	0.71
1.20.1	1	1.00	1.00	1.00	**2.7.3**	1	0.50	1.00	0.67
1.21.1	2	1.00	1.00	1.00	**2.7.4**	31	0.76	0.90	0.82
1.21.3	3	0.67	0.67	0.67	**2.7.6**	7	0.80	0.57	0.67
1.23.5	1	1.00	1.00	1.00	**2.7.7**	107	0.97	0.99	0.98
1.3.1	85	0.93	0.92	0.92	**2.7.8**	2518	1.00	1.00	1.00
1.3.3	16	0.93	0.81	0.87	**2.7.9**	2	1.00	0.50	0.67
1.3.5	13	1.00	0.85	0.92	**2.8.1**	4	1.00	1.00	1.00
1.3.8	13	0.92	0.92	0.92	**2.8.2**	21	1.00	1.00	1.00
1.3.98	1	0.25	1.00	0.40	**2.8.3**	16	0.79	0.94	0.86

表 5.4: EC 番号 1 桁〜3 桁目までの分類結果 2

EC Class	Num of Data	precision	recall	f1-score	EC Class	Num of Data	precision	recall	f1-score
3.1.1	570	0.97	0.96	0.96	**4.1.99**	10	1.00	0.70	0.82
3.1.2	43	0.91	0.91	0.91	**4.2.1**	88	0.93	0.93	0.93
3.1.3	148	0.89	0.90	0.89	**4.2.2**	1	0.50	1.00	0.67
3.1.4	26	0.95	0.81	0.88	**4.2.3**	23	0.95	0.78	0.86
3.1.6	2	1.00	1.00	1.00	**4.3.1**	9	1.00	0.89	0.94
3.1.8	6	1.00	1.00	1.00	**4.3.2**	2	1.00	1.00	1.00
3.13.1	2	0.00	0.00	0.00	**4.3.3**	3	0.67	0.67	0.67
3.2.1	113	0.91	0.96	0.94	**4.4.1**	24	0.77	1.00	0.87
3.2.2	12	1.00	1.00	1.00	**4.5.1**	4	0.50	0.50	0.50
3.3.1	1	1.00	1.00	1.00	**4.6.1**	1	1.00	1.00	1.00
3.3.2	16	0.88	0.94	0.91	**4.99.1**	1	1.00	1.00	1.00
3.4.11	20	0.76	0.80	0.78	**5.1.1**	1	1.00	1.00	1.00
3.4.13	18	0.76	0.89	0.82	**5.1.3**	4	0.50	0.50	0.50
3.4.17	6	0.71	0.83	0.77	**5.2.1**	1	0.00	0.00	0.00
3.4.19	6	1.00	1.00	1.00	**5.3.1**	11	0.75	0.82	0.78
3.4.21	2	1.00	1.00	1.00	**5.3.2**	4	0.60	0.75	0.67
3.5.1	87	0.84	0.91	0.87	**5.3.3**	11	0.82	0.82	0.82
3.5.2	7	1.00	1.00	1.00	**5.3.99**	3	0.67	0.67	0.67
3.5.3	4	0.67	1.00	0.80	**5.4.2**	3	0.75	1.00	0.86
3.5.4	15	0.87	0.87	0.87	**5.4.3**	3	1.00	0.67	0.80
3.5.5	18	0.94	0.94	0.94	**5.4.4**	2	0.33	0.50	0.40
3.5.99	3	0.25	0.33	0.29	**5.4.99**	5	0.71	1.00	0.83
3.6.1	58	0.68	0.69	0.68	**5.5.1**	19	0.83	0.79	0.81
3.6.3	126	1.00	1.00	1.00	**6.1.2**	2	1.00	0.50	0.67
3.6.4	1	0.33	1.00	0.50	**6.2.1**	67	0.99	0.99	0.99
3.7.1	17	0.81	1.00	0.89	**6.3.1**	6	1.00	0.83	0.91
3.8.1	18	1.00	0.89	0.94	**6.3.2**	59	0.95	0.97	0.96
4.1.1	260	0.97	0.99	0.98	**6.3.4**	8	0.78	0.88	0.82
4.1.2	28	0.75	0.96	0.84	**6.3.5**	7	1.00	0.29	0.44
4.1.3	8	1.00	0.50	0.67	**6.4.1**	5	1.00	0.80	0.89
					Total	9418			
					macro average		0.81	0.80	0.79
					weighted average		0.96	0.95	0.95
					accuracy		0.95		

表 5.5: EC 番号各 2,3 桁目の分類精度

ECクラス	学習データ数	テストデータ数	クラス数	Precision	Recall	Macro F1-Score
EC 1.X.X	6380	1601	64	0.80	0.78	0.78
EC 2.X.X	23160	5789	24	0.83	0.81	0.81
EC 3.X.X	5377	1345	27	0.81	0.87	0.83
EC 4.X.X	1878	462	14	0.86	0.85	0.84
EC 5.X.X	273	67	12	0.66	0.71	0.68
EC 6.X.X	604	154	7	0.96	0.75	0.81

第6章

おわりに

有機合成化学分野では，酵素の効率性や環境面から，酵素を生体触媒として利用される機会が増加している．酵素は性質に応じて4桁のEC番号に割り振られており，酵素の性質特定や探索による実験コスト短縮のため，機械学習を用いたEC番号予測が行われている．

本研究では，化学反応の特徴をより詳細に捉えるため，物理・化学的特性値と構造的特徴であるフィンガープリントを組み合わせ，有機合成において用いるべき最適な酵素候補をEC番号として予測する機械学習モデルを開発した．210種類のRDKit記述子を用い，EC番号が割り当てられた酵素反応に対して，基質から生成物の変化を表現した210次元特徴ベクトルを作成し，予測モデルを構築した．テストデータの特徴ベクトルに対して，EC番号の1〜3桁までを予測することを目的とし，Random Forests (RF) とStep Forward法による記述子選択，SMOTEによるオーバーサンプリングを用いてモデルを構築した．

数値実験において，予備実験1では従来行われたクラス分類にSMOTEを適用し，適用前と比較することでSMOTEの有効性を示した．予備実験2ではEC番号1桁目分類，EC T.X.X (T= $1, 2, \cdots, 6$) のクラス分類を行い，分類精度 (Macro F1-Score) を高める記述子組合せを選択した．7回の記述子選択で選ばれた記述子をマージし，重複を削除することで最終的な記述子組合せを決定した．本実験では，選択された記述子93種からなる学習データに対してグリッドサーチを行い，RFの最適パラメータを決定した．そして，テストデータに対してEC番号1〜3桁目の多クラス分類を行い，Macro F1-Score 0.79が得られた．

結果として，先行研究であるフィンガープリントの差分を用いた手法と同程度の予測精度となったが．モデル構築に時間がかかる点，利用可能なデータが限られる点などがら，用いるデータと機械学習手法に改善が望まれる．7回の記述子選択で選ばれた記述子の特徴を分析し，類似する記述子をまとめて次元削減を行う，または記述子ごとに重要度を評価し，重みづけを行うことが分類精度向上に有効であると考えられる．また，フィンガープリント手法と物理・化学的特性値を用いる手法は学習コストと化学反応表現の面でトレードオフの関係と考えられ，化学反応情報をより詳細に捉えるため，2つの特徴量を最適な割合で用いることが必要である．

今後の課題として，2つ挙げられる．1つ目は，EC番号の1〜4桁目を同時に予測する手法の開発である．4桁目の分類は3桁目の分類よりもさらに細かく，酵素同士の違いを区別するのが難しい．より詳細に酵素反応を分類できる機械学習手法や記述子組合せを用い，実装することで，さらなる酵素候補の絞り込みが行える．2つ目は，実際の有機合成で提案予測モデルを用いることである．有機合成の研究者に利用してもらい，現実的な制約や実験条件，予測誤差などをモデルに反映することができれば，より応用性のあるモデル構築が期待できる．

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 浅野泰久名誉教授，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、生物・化学分野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 AI 健康・医薬研究センター 副センター長 荒木通啓先生，国立研究開発法人医薬基盤・健康・栄養研究所 渡邉直暉様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．

2024 年 2 月

武藤 克弥

参考文献

[1] 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.

[2] "Enzyme Nomenclature", `https://iubmb.qmul.ac.uk/enzyme/`, 閲覧日 2024.1.21.

[3] Naoki Watanabe, Masaki Yamamoto, Masahiro Murata, Yuki Kuriya, Michihiro Araki, "EnzymeNet: residual neural networks model for Enzyme Commission number prediction", *Bioinformatics Advances*, Vol. 3, No. 1, 2023.

[4] Latino Diogo ARS and Aires-de-Sousa Joao "Assignment of EC Numbers to Enzymatic Reactions with MOLMAP Reaction Descriptors and Random Forests", *Journal of chemical information and modeling*, Vol. 49, No. 7, pp. 1839-1846, 2009.

[5] Probst Daniel, "An explainability framework for deep learning on chemical reactions exemplified by enzyme-catalysed reaction classification", *Journal of Cheminformatics*, Vol. 15, No. 1, pp. 113, 2023.

[6] 北川勲，磯部稔, "天然物化学・生物有機化学 I", 朝倉書店, 2008.

[7] 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.

[8] "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.

[9] 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.

[10] 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.

[11] 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.

[12] "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.

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

[14] "酵素基質とは", `https://bizcomjapan.co.jp/iris-biotech/knowledge/substrate/`, 閲覧日 2022.1.31.

[15] “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.

[16] 白兼孝雄, “酵素の分類と命名法”, JAS 情報, 2017.

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

[18] “BLAST: Basic Local Alignment Search Tool”https://blast.ncbi.nlm.nih.gov/Blast.cgi, 閲覧日 2024.1.31

[19] “バイオインフォマティクスの基礎実習（ホモロジー検索）”https://lecture.ecc.u-tokyo.ac.jp/~ashimizu/database/lec2.html, 閲覧日 2024.1.31

[20] Yu Chenggang, Zavaljevski Nela, Desai Valmik and Reifman Jaques, “Genome-wide enzyme annotation with precision control: Catalytic families (CatFam) databases”, *Proteins: Structure, Function, and Bioinformatics*, Vol. 74, No. 2, pp. 449-460, 2009.

[21] “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.

[22] 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.

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

[24] Wenbo Sun et al., “Machine learning–assisted molecular design and efficiency prediction for high-performance organic photovoltaic materials”, *Science advances*, Vol. 5, No. 1, eaay4275, 2019.

[25] Sakiyama, Hiroshi, Motohisa Fukuda, and Takashi Okuno., “Prediction of blood-brain barrier penetration (bbbp) based on molecular descriptors of the free-form and in-blood-form datasets”, *Molecules*, Vol. 26, No. 24, 7428, 2021.

[26] Rogers David and Hahn Mathew, “Extended-connectivity fingerprints”, *ournal of chemical information and modeling*, Vol. 50, No. 5, pp. 742-754, 2021.

[27] “Python: 機械学習で分類問題のモデルを評価する指標について”, https://blog.amedama.jp/entry/2017/12/18/005311, 閲覧日 2022.1.27.

[28] Scott A. Wildman and Gordon M. Crippen., “Prediction of Physicochemical Parameters by Atomic Contributions”, *Journal of chemical information and computer sciences*, Vol. 39, pp. 868-873, 1999.

[29] 小林誠一., “化合物活性予測におけるベイジアンモデルの利用”, *CICSJ Bulletin* , Vol. 27, No. 4, 2009.

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

[31] “Rhea- Annotated reactions database”, `https://www.rhea-db.org/`, 閲覧日 2024.2.3.

[32] “BRENDA Enzyme Database”, `https://www.brenda-enzymes.org/`, 閲覧日 2024.2.3.

[33] “PathBank”, `https://pathbank.org/`, 閲覧日 2024.2.3.

[34] “MetaNetX: Summary”, `https://www.metanetx.org/`, 閲覧日 2024.2.3.

[35] Chawla Nitesh V., Bowyer Kevin W., Hall Lawrence O., Kegelmeyer W. Philip, “SMOTE: synthetic minority over-sampling technique”, *Journal of artificial intelligence research*, Vol 16, pp. 321-357, 2002.

[36] Leo Breiman., “Random Forests”, *Machine Learning*, Vol 45, pp. 5-32, 2001.

[37] 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.

[38] “sklearn.ensemble.RandomForestClassifier — scikit-learn 1.4.0 documentation”, `https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html`, 閲覧日 2024.1.31.

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

[40] Chandra, Prem and Enespa and Singh, Ranjan and Arora, Pankaj Kumar “Microbial lipases and their industrial applications: a comprehensive review”, *Microbial cell factories*, Vol 19, pp. 1-42, 2020.

[41] Fernández Alberto, Garcia Salvador, Herrera Francisco, Chawla, Nitesh V, “SMOTE for learning from imbalanced data: progress and challenges, marking the 15-year anniversary”, *Journal of artificial intelligence research*, Vol 61, pp. 863-905, 2018.

[42] “SMOTE — Version 0.12.0”, `https://imbalanced-learn.org/stable/references/generated/imblearn.over_sampling.SMOTE.html`, 閲覧日 2024.1.31.

[43] “sklearn.model_selection.GridSearchCV — scikit-learn 1.4.0 documentation”, `https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.GridSearchCV.html`, 閲覧日 2024.1.31.