Chemoinformatics Hackathon in YOKOHAMA

2025

Hiroaki Gotoh

Hackathonの目標

ハッカソンは、「ハック（Hack）」と「マラソン（Marathon）」を組み合わせた言葉で、限られた時間内にチームや個人でアイデアを出し合い、プログラムやサービスを一気に作り上げるイベントです。スピード感と創造性が重視されます！

スキルを学び競い合う

分子構造データから **記述子を作成し、予測する**

- **機械学習モデル**を構築して予測
- **可視化・評価**を通じて結果を理解

協力して何かを形にする

普段一緒に研究していない人と
共同作業を行う

- **役割分担して成果物を作成**する経験
- チームでの発見や工夫を共有する

Hackathon の本質：
「スキルを学ぶ」ことと「協力して何かを形にする」ことの両方を体験する

全体の流れとスケジュール

2025年8月28(木)	
13:15 ～	受付開始
13:30 ～ 13:40	開会式
13:40 ～ 14:40	ショートプレゼン
14:40 ～ 15:00	テーマ発表！
15:00 ～ 17:45	ハッカソン
18:00 ～ 20:00	夕食（BBQ）・暫定順位発表
20:15 ～ 21:00	レクレーション
21:00 ～	フリータイム

2025年8月29(金)	
8:15 ～	朝食
9:00 ～ 11:30	ハッカソン
11:30 ～ 12:00	成果報告会
12:00 ～ 12:30	表彰式　記念撮影・閉会式
12:30 ～	解散

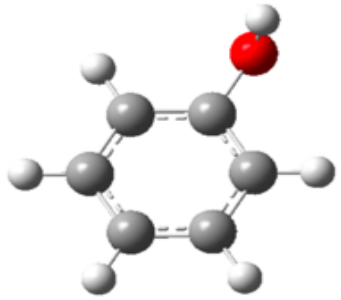
データセット(test_private)で評価
＋ プレゼン資料を提出

データセット(train, test_public)で評価

データは、3つに分割してお渡しします。その中でモデルを作成して、評価してください

ケモインフォマティクスの概要

化合物の情報



直接変換困難

変換, 抽出, 選定

分子記述子

原子数: **C 6, H 6, O 1**
フラグメント: Ar**OH** , **1**
分子量: **94.1**

分子の特性などの特徴量を数字に変換

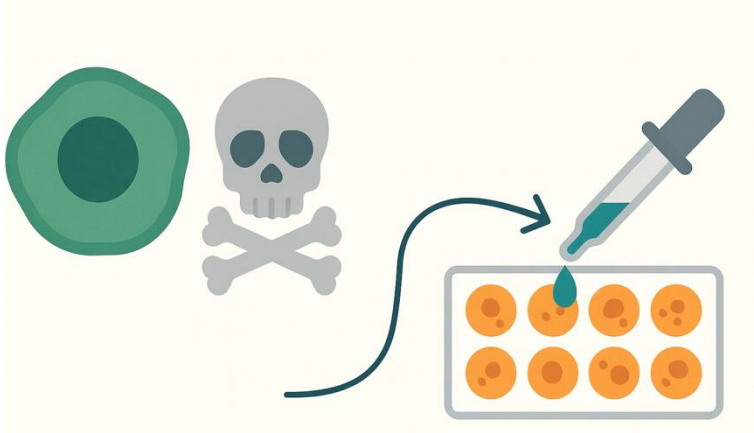
アルゴリズム、解釈

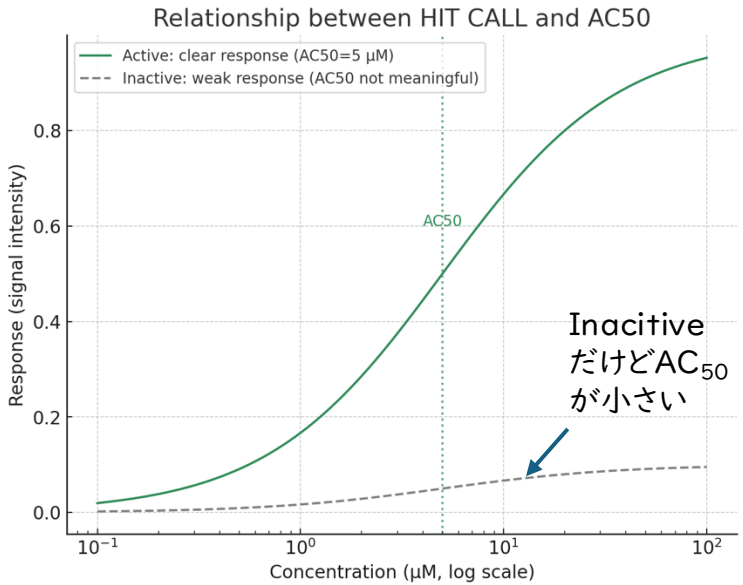
物性や選択性

選択性: 位置、ジアステレオ、エナンチオ
反応性: 進行可否、要因解明
食品科学: 抗酸化能、キレート能
機能性物質: 光重合開始剤

ケモインフォマティクスは、分子の特徴を数値化し、機械学習や統計的手法を用いて、物性予測・機能解明・材料設計へと応用する枠組みを提供する

今回のテーマは、細胞の毒性予測





1. HIT CALL

- アッセイデータで用いられる指標。
- 化合物が **活性（Active）か非活性（Inactive）か** を二値的に判定した結果。
- 判定基準はアッセイごとに異なりますが、一般的には **濃度反応曲線（dose-response curve）** をフィットして、
  - **有意なシグナル変化があるかどうか**
  - **閾値を超えるかどうか**

  によって決まります。

2. AC_{50}

- **Activity Concentration at 50%** の略。
- 化合物がアッセイ応答の **半分の効果を示す濃度**。
- 濃度応答曲線をロジスティック関数などでフィッティングして求めます。
- 単位は通常 **μM**。

3. 両者(HIT CALLとAC50)の関係

- **HIT CALL = Active** であれば、原則としてその化合物には **AC50値が定義されます**。
- **HIT CALL = Inactive** の場合、応答が基準を超えないため **AC50は算出されない**（欠損値や大きな上限値が入ることが多い）。
- つまり：
  - HIT CALL = Active → 有意な応答あり、AC50が算出可能。
  - HIT CALL = Inactive → 応答なし、AC50はNaNや ＞ 最大濃度(20μMや30μM)。

4. HIT CALL は 有意な応答強度を伴わないと Active にならない

HIT CALL は単純に AC50 だけで判定されるのではなく、

濃度応答曲線の形状が妥当か
応答の大きさ（TOP）が一定の閾値以上か
統計的に有意か
などを含めた総合的な基準で決まる

Active だと毒性がある。Activeとされた化合物の中でAC_{50}が小さいと、毒性がより強い。

データセットの紹介

同じ細胞に対する毒性試験。 データ元を探して正解データを導くことは、お控えください。

データセット(train, test_public)

	CASRN	Name	IUPAC_NAME	INCHI_STRING	SMILES	MOLECULAR_FORMULA	HIT	AC50	LOGAC50
0	2588/4/7	Phorate sulfone	S-[(Ethanesulfonyl)methyl] O,O-diethyl phospho...	InChI=1S/C7H17O4PS3/c1-4-10-12(13,11-5-2)14-7-...	CCOP(=S)(OCC)SCS(=O)(=O)CC	C7H17O4PS3	Inactive	30.000000	1.477121
1	96-45-7	4,5-Dihydro-2-mercaptoimidazole	Imidazolidine-2-thione	InChI=1S/C3H6N2S/c6-3-4-1-2-5-3/h1-2H2,(H2,4,5,6)	S=C1NCCN1	C3H6N2S	Active	0.940777	-0.026513
2	110-40-7	Diethyl decanedioate	Diethyl decanedioate	InChI=1S/C14H26O4/c1-3-17-13(15)11-9-7-5-6-8-1...	CCOC(=O)CCCCCCCCC(=O)OCC	C14H26O4	Inactive	20.000000	1.301030
3	104-76-7	2-Ethyl-1-hexanol	2-Ethylhexan-1-ol	InChI=1/C8H18O/c1-3-5-6-8(4-2)7-9/h8-9H,3-7H2,...	CCCCC(CC)CO	C8H18O	Inactive	20.000000	1.301030
4	74051-80-2	Sethoxydim	2-(N-Ethoxybutanimidoyl)-5-[2-(ethylsulfanyl)p...	InChI=1/C17H29NO3S/c1-5-8-14(18-21-6-2)17-15(1...	CCCC(=NOCC)C1=C(O)CC(CC(C)SCC)CC1=O	C17H29NO3S	Inactive	7.720201	0.887629

HIT 部分を予測するための訓練や評価に使用する。

test_publicとtest_privateのデータは、学習には使わないでください。
使われていたら、点数を0点にします。

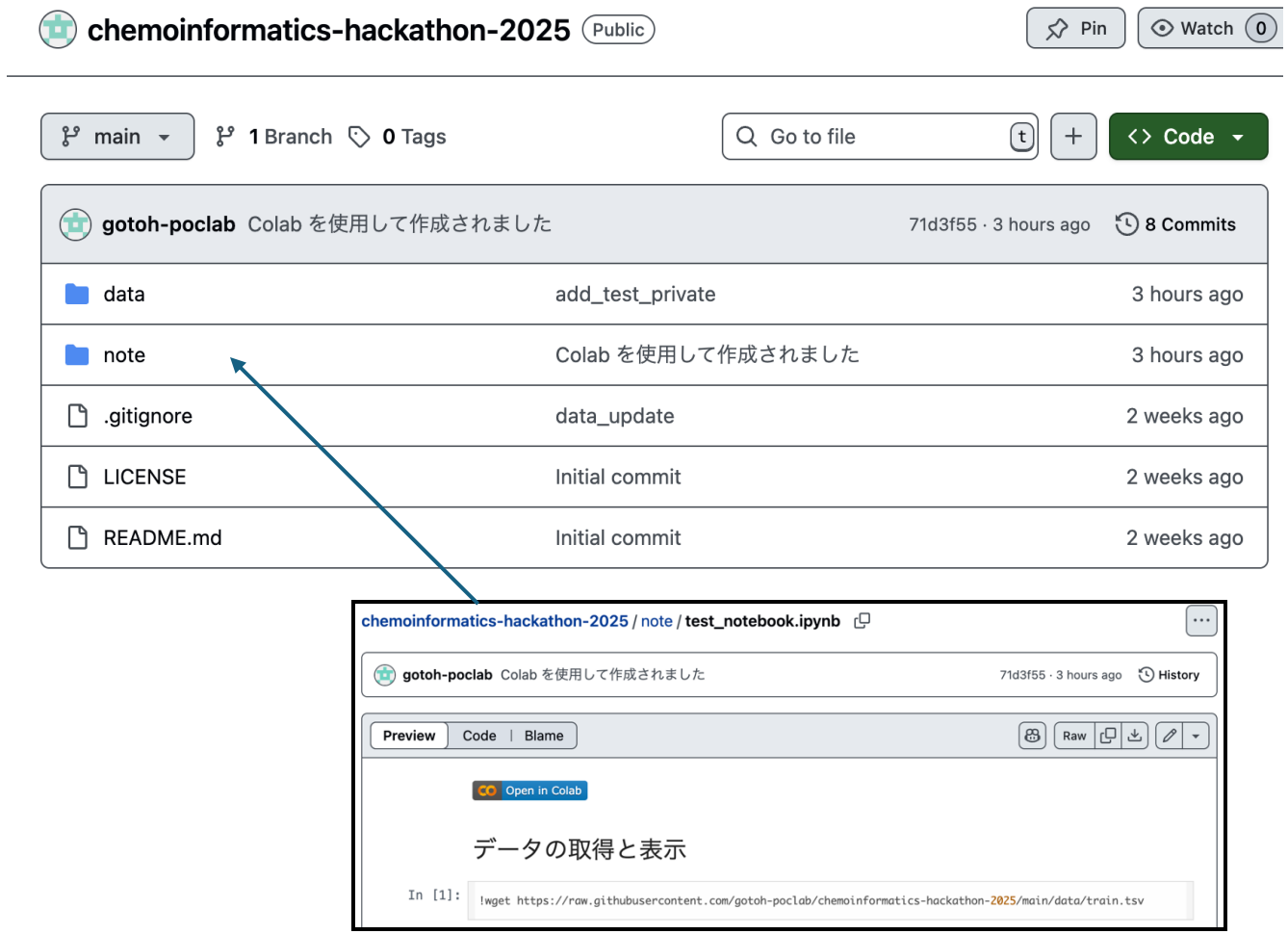
データセット(test_private)

	CASRN	IUPAC_NAME	INCHI_STRING	SMILES	MOLECULAR_FORMULA
0	825643-57-0	2-[3,5-Bis(trifluoromethyl)phenyl]-N-{4-(4-flu...	InChI=1S/C30H30F7N3O3/c1-16-9-20(31)5-6-21(16)...	Cc1cc(F)ccc1-c1cc(N2CC[C@@H](O)[C@H]2CO)ncc1N(...	C30H30F7N3O3
1	138261-41-3	N-{(2E)-1-[(6-Chloropyridin-3-yl)methyl]imidaz...	InChI=1S/C9H10ClN5O2/c10-8-2-1-7(5-12-8)6-14-4...	O=[N+]([O-])/N=C1\NCCN1Cc1ccc(Cl)nc1	C9H10ClN5O2
2	104206-82-8	2-[4-(Methanesulfonyl)-2-nitrobenzoyl]cyclohex...	InChI=1S/C14H13NO7S/c1-23(21,22)8-5-6-9(10(7-8...	CS(=O)(=O)c1ccc(C(=O)C2C(=O)CCCC2=O)c([N+](=O)...	C14H13NO7S
3	464930-42-5	Nalpha-{(3R)-3-(2H-1,3-Benzodioxol-5-yl)-3-[(6...	InChI=1S/C42H52N4O7S.ClH/c1-27(2)45(5)42(48)38...	COc1ccc2cc(S(=O)(=O)N[C@H](CC(=O)N[C@H](Cc3ccc...	C42H53ClN4O7S
4	173584-44-6	Methyl (4aS)-7-chloro-2-{(methoxycarbonyl)[4-(...	InChI=1S/C22H17ClF3N3O7/c1-33-18(30)21-10-12-9...	COC(=O)N(C(=O)N1CO[C@@]2(C(=O)OC)Cc3cc(Cl)ccc3...	C22H17ClF3N3O7

最終的には、[Inactive]か[Active]のラベルをつける。
test_privateはtest_publicに化合物が100個追加されています。

予測するのは、[Inactive]か[Active]の2値分類です。

チュートリアルコードと使用したライブラリー



https://github.com/gotoh-poclab/chemoinformatics-hackathon-2025

データの可視化

- **Pandas**：表形式データ処理（CSV/Excelの読み書き、集計）
- **Matplotlib**：汎用的なグラフ描画
- **Seaborn**：美しい統計可視化（Matplotlib拡張）

機械学習

- **scikit-learn**：基本的な機械学習（分類・回帰・評価）
- **XGBoost**：高精度な勾配ブースティング
- **LightGBM**：高速・軽量な勾配ブースティング

ケモインフォマティクス

- **RDKit**：分子構造の扱い（読み込み、可視化、操作）
- **Mordred**：分子記述子の計算

後で、全員、チュートリアルコードは動くことを確認いたします。

全体像と使用ライブラリー

データ収集

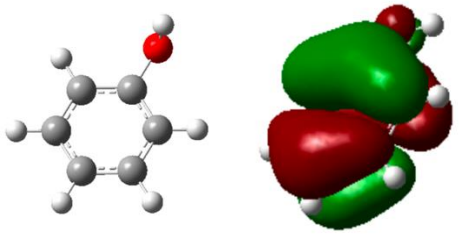
データベース
実験
文献

化合物と問題に関する情報を収集

化合物の情報

Phenol

C1=CC=C(C=C1)O

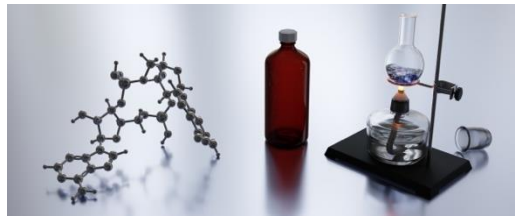


2次元，3次元表記，反応表記

RDKit

直接変換困難

物性や選択性



物理化学: 融点、水溶性など
スペクトル: NMR, IR など
食品科学: 匂い、味、毒性 など
選択性: オルト、パラの位置で反応など

検索
(完全、部分、類似)

数値化

解釈

評価

データの分割
(訓練，検証，テスト)

回帰
YYPLOT

分類
混同行列

分子記述子

計算化学で算出

分子力学
分子軌道法
など

原子数: C 6, H 6, O 1
フラグメント: ArOH , 1
分子量: 94.1

Mordred

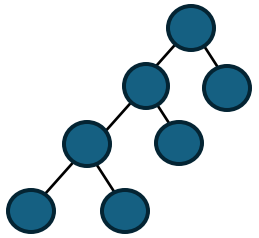
Pandas

前処理

記述子選択

アルゴリズムと予測

モデルの種類の検討



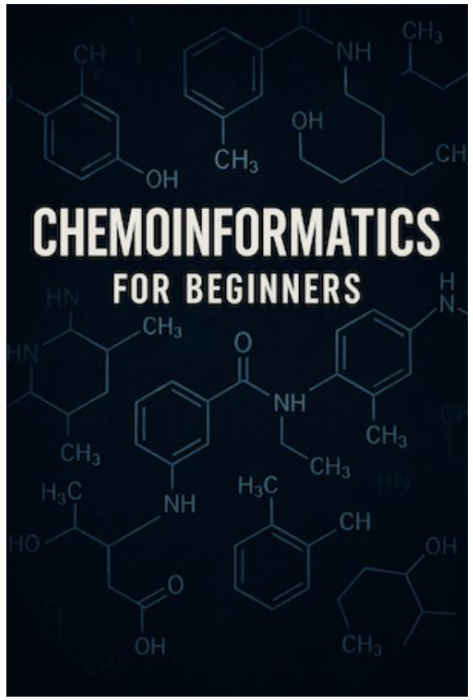
線形モデルと非線形モデル

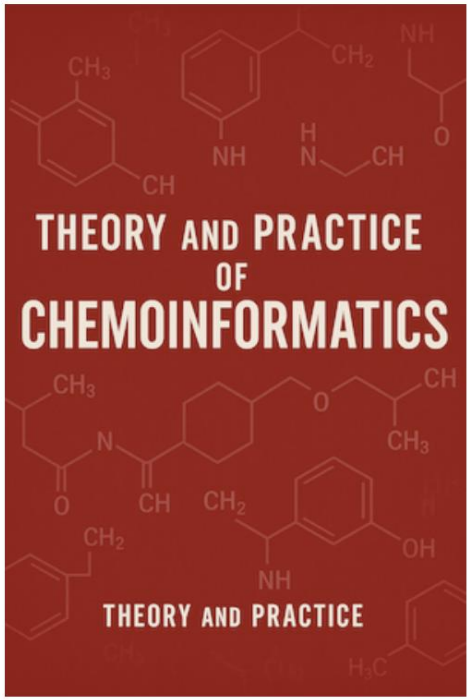
KEYWORDS

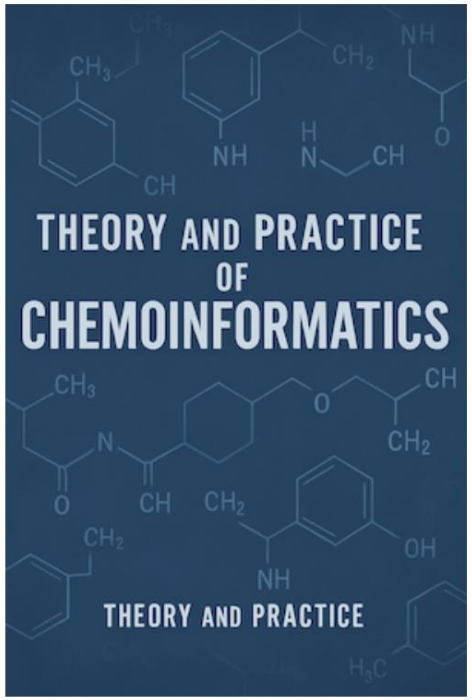
- 記述子特性
- 逆解析
- 化学空間
- 内挿と外挿

Scikit-learn

学習や情報提供







https://zenn.dev/poclabweb?tab=books

作成中ですが途中経過を**無料公開中**です。
インターネットにアクセスできる人ならば、1から学べます。

さらに、学びたい人は、「**ケモインフォマティクス 入門書**」で検索

記述子作成

分子記述子とは?

- 分子の構造を数値ベクトルに変換したもの
- 機械学習モデルに入力できるようにするための「特徴量」

代表的な分子記述子

- **物理化学的特徴**
  - 分子量、LogP（水溶性/疎水性）、極性表面積（TPSA）
- **構造的特徴**
  - 原子数、結合数、環の数、芳香環の有無
- **部分構造フラグメント**
  - 官能基（–OH, –NH_2, –Cl など）の有無
- **分子フィンガープリント**
  - 分子をビット列で表現（類似性検索や分類でよく使う）

記述子作成に使うツール

- **RDKit**：分子の基本的な記述子やフィンガープリント
- **Mordred**：1800種類以上の記述子を自動計算可能

入力例

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

# 分子の準備(SMILES表記)
mol = Chem.MolFromSmiles("CC(=O)Oc1ccccc1C(=O)O")  # アスピリン

# 代表的な記述子の計算
print("分子量:", Descriptors.MolWt(mol))
print("LogP:", Descriptors.MolLogP(mol))
print("TPSA:", Descriptors.TPSA(mol))
print("水素結合ドナー数:", Descriptors.NumHDonors(mol))
print("水素結合アクセプター数:", Descriptors.NumHAcceptors(mol))
```

出力例

```
分子量: 180.16
LogP: 1.19
TPSA: 63.60
水素結合ドナー数: 1
水素結合アクセプター数: 4
```

チュートリアルコードでは、RDKitのDescriptorsから数個選んで使用しているが
自作や他のライブラリーなど、何を使用しても問題ありません。

機械学習モデル

機械学習モデルの役割

•分子記述子（数値ベクトル）を入力として活性（Active / Inactive）を予測する分類器を構築

代表的なモデル

- ランダムフォレスト (RandomForest)
  - 多数の決定木を組み合わせたモデル
  - 初学者向け、解釈しやすい、安定した精度
- XGBoost / LightGBM
  - 勾配ブースティング（Gradient Boosting）系
  - 高精度・高速・チューニングしやすい
  - 実務やコンペでよく使われる
- ロジスティック回帰 / SVM / KNN
  - シンプルだが基礎が理解しやすい
  - 少数の特徴量なら有効
  - SVMやKNNでは、特徴量の正規化が必要

チュートリアルコードでは、RandomForestを使用していますが、自作や他のモデルを検討するなども問題ありません。

入力例

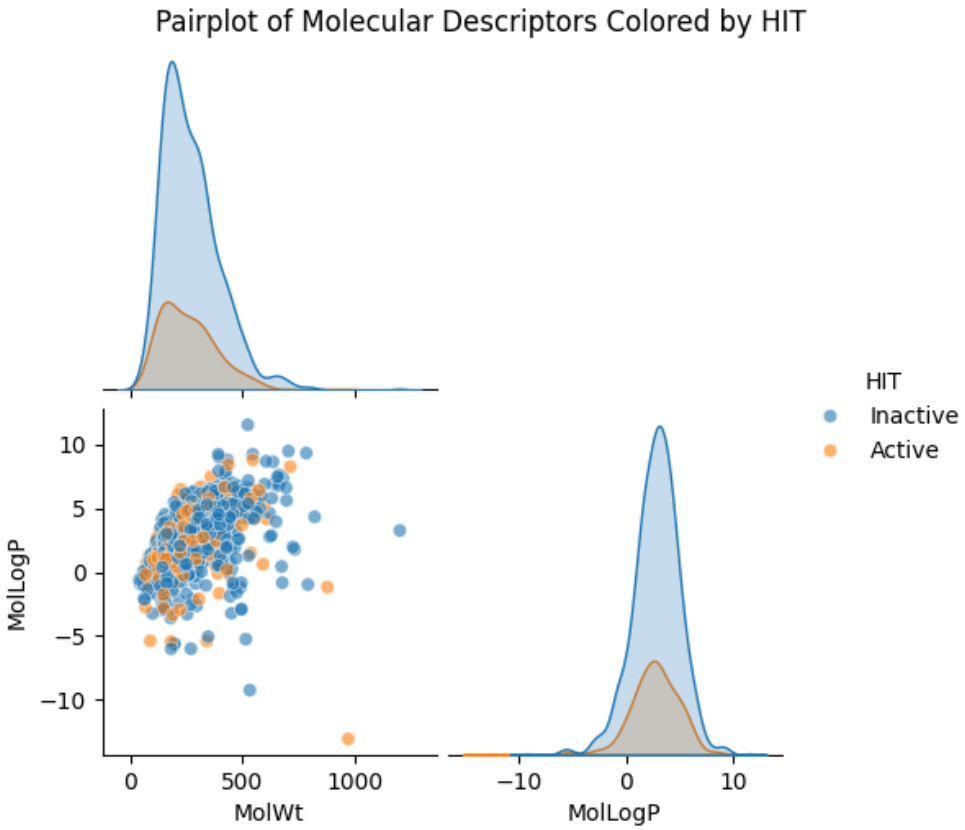
```python
import numpy as np
from sklearn.ensemble import RandomForestClassifier
from xgboost import XGBClassifier

# データ(X: 特徴量, y: ラベル 0=Inactive, 1=Active)
X = np.array([
    [0.1, 1.0],
    [0.2, 0.9],
    [0.8, 0.2],
    [0.9, 0.1],
    [0.5, 0.5]
])
y = np.array([0, 0, 1, 1, 0])

# モデル学習と予測
models = {
    "RandomForest": RandomForestClassifier(),
    "XGBoost": XGBClassifier()
}

for name, clf in models.items():
    clf.fit(X, y)
    pred = clf.predict(X)
    print(f"{name}:", pred)
```

出力例

```
RandomForest: [0 0 1 1 0]
XGBoost     : [0 0 1 1 1]
```

解釈や評価

可視化による解釈



記述子選びやプレゼン資料では、みやすい

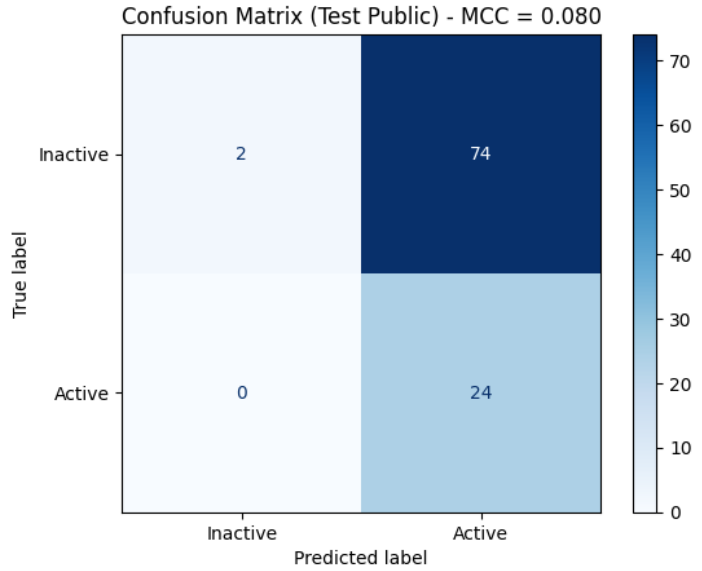
評価指標 (MCC)

		予測値 Positive	予測値 Negative
実験値	Positive	TP	FN
	Negative	FP	TN

正しい分類　　間違った分類

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}$$

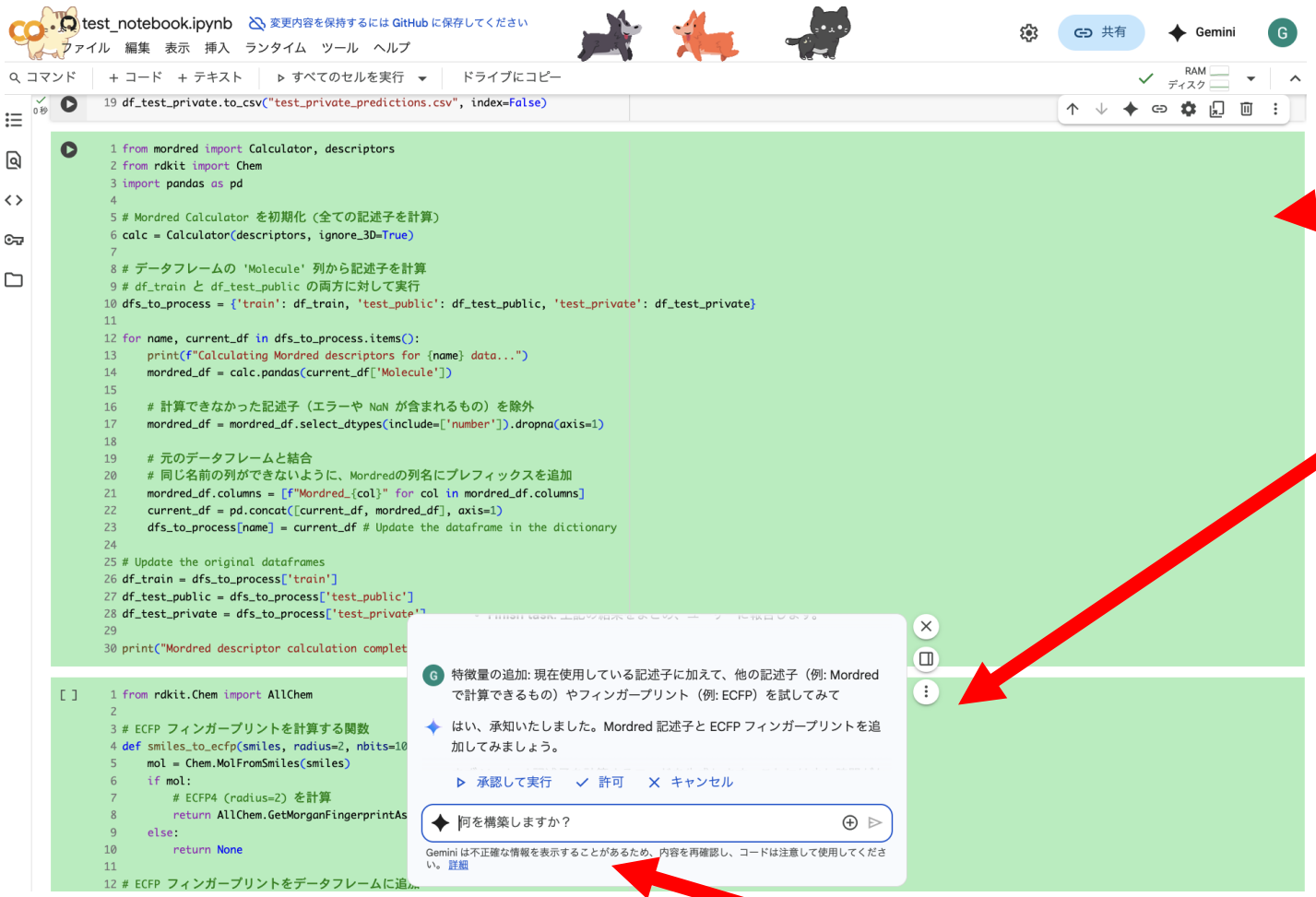
全体のスケールを調整をして、範囲が-1から1になるようにする。



左の例だとモデルがActive
に予測しすぎているので
MCCの値が低くなっている。

今回は、評価指標としてMCCを採用し、MCCの値がより1に近いものを、良いモデルとして評価

Geminiの利用など



Geminiを呼び出す

生成された部分が緑色

指令文章（プロンプト）
を書けば、コードを書いてくれる

生成AIに聞くときは、**明確な指示**と**適切な文脈**を与えると精度が上がる。

1. **目的・入力・出力の形式** を指定
2. **「なぜ知りたいか」「どう使うか」**を伝える
3. **制約条件**をつける
4. **具体例**を伝える

承認して実行を行えば、実行される

積極的に使って、いただいても構いません。

提出物(1日目の17:45、2日目11:30 締切)

1日目 17:45

1. 学習を行った結果が記載した
.ipynbのファイル

- train, test_publicの2つで評価されていることが必須
- test_publicは学習には使用しないようにモデルを作成してください。
- 提出された.pynbのファイルは全員で共有する。

2日目 11:30

1. test_privateにpredictionのラベル(active or inactive)をつけたcsvファイル

2. 成果発表会に使用するPDFファイル
- 形式や枚数は自由ですが、各チーム5分以内で、発表を行なってください。
- 自分のチーム以外で発表が良かったチームに全員投票してもらいます。

3. 最終的な学習結果をまとめた.ipynbのファイル

提出は、Slackの全員が参加しているチャンネルに行ってもらいます。

まとめ

- 今回のHackathonを通じて、
  「分子構造 → 記述子 → モデル → 評価」の一連の流れを体験する。
- **技術的なゴール：**
  ケモインフォマティクスの基礎を学び、機械学習モデルで予測を行う力を養う
- **体験的なゴール：**
  普段一緒に研究していない人と協力し、限られた時間で成果物を作る経験を得る
- 限られた時間の中で工夫したアイデアや、チームでの協力は大きな成果です。
  今後の研究や学びにつなげてください。

制限時間内にできるだけ良いものを作成したチームには表彰がありますが、、、

「できたこと」よりも「何を学んだか」「どう協力したか」が大切です。

Q&A

今までのところで何か質問がある人は、いますか？

チュートリアルのコードを動かしながら、実演して、その後、各チームごとに別れて進めますので、その時に質問していただいても構いません。