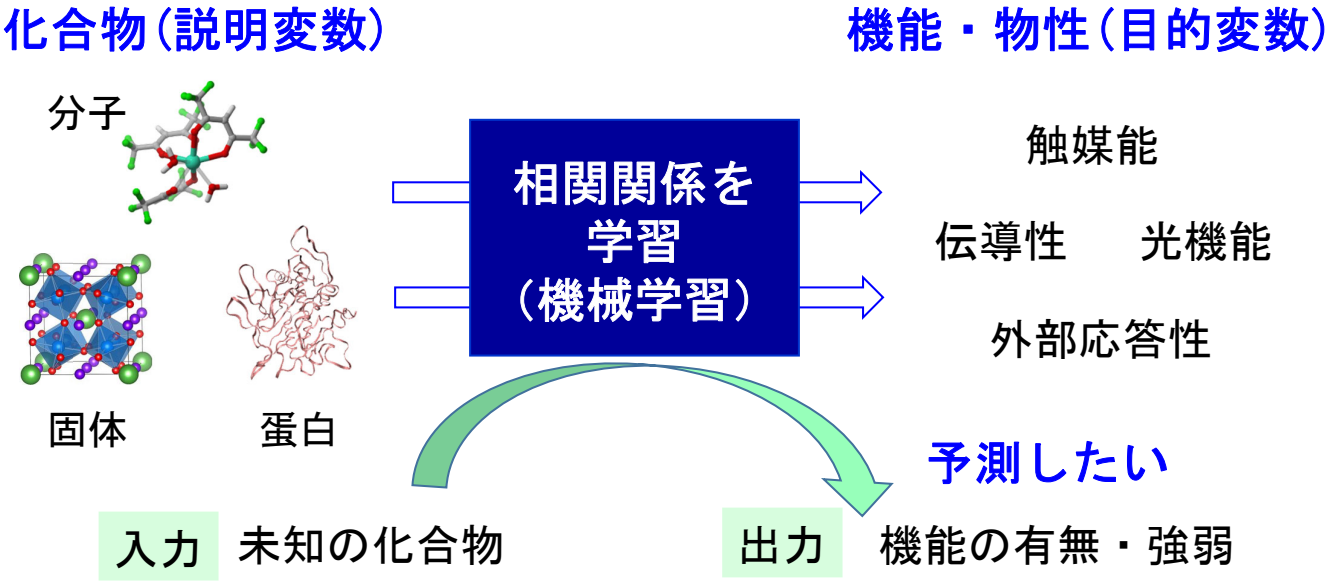
基礎科学チュートリアル

すぐできるマテリアルズ・インフォマティクス

～材料×機械学習の融合～

慶應義塾大学理工学部化学科
畑中美穂

Materials Informaticsの基本概念



考えるべき３項目

- Q1. 相関関係をどう学習するか？（機械学習）
- Q2. 化合物をどう表現するか？（記述子・特徴量）
- Q3. データをどう集めるか？（ﾃﾞｰﾀﾍﾞｰｽの扱い）

データを数値で記述する (記述子, 特徴量)

例①　白黒画像

28×28 pixelの
手書き数字の画像

各pixelの値を
要素に持つ
28×28 の行列で
記述可

例②　カラー画像

RGB
分割

3×3000 pixel
×4000 pixelの
テンソルで
記述可

例③　言語

I have a pen.
I have an apple.

予め定義した単語の
出現回数を数える

I	2
want	0
have	2
go	0
a / an	2
the	0

apple	1
banana	0
pineapple	0
pen	1
cup	0
scarf	0

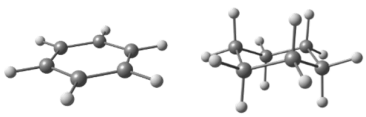
(2, 0, 2, 0, 2, 0, 1, 0, 0, 1, 0, 0)
ベクトルで記述可

化合物の特徴量

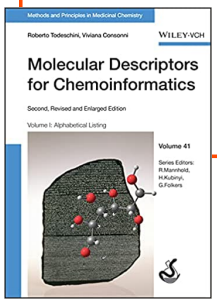
無機塩　**NaCl**　**$MgCl_2$**　**$FeCl_2$**　**$EuCl_3$**

- 各原子固有の情報：　原子番号・分子量・イオン半径
- 塩の情報：　各原子の形式電荷 (Na^+, Mg^{2+}, Fe^{2+}, Eu^{3+})
  結晶構造　(面心立方格子 etc)
  融点・密度・水への溶解度

分子



- 測定から得られる情報：　スペクトル
- DFT計算から得られる情報：軌道のエネルギー・電荷の偏り
- 構造式から得られる情報：　分子量・部分構造の有無・
  水素結合donor/acceptor数
  極性表面積 (TPSA)
  回転可能な結合数　など



**ケモインフォマティクス分野で蓄積されてきた
分子の特徴量に広く用いられている**

分子の表記方法① SMILES記法

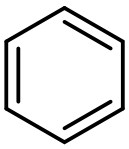
SMILES (simplified molecular-input line-entry system)

- 各原子のつながりを線形で表記・分子構造のコンパクトな表現
- 文法
  - 原子名：C, H, Nなど
    （芳香環に含まれる原子のみ小文字で書く場合も）
  - 結合： 二重結合は「=」， 三重結合は「#」
    （共有結合以外の場合は、「.」でつなぐ）
  - 分岐：分岐した枝の先を全てカッコ内()に書く
  - 環　：環のはじめ・終わりの原子に同じ数字をつける

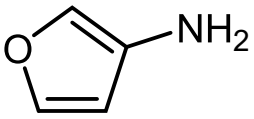
例① CO

H_3C-OH

例② C1=CC=CC=C1



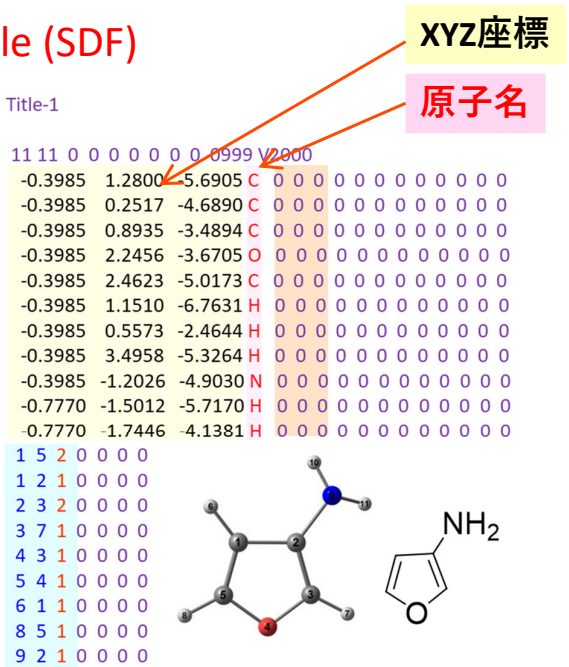
例③ O1C=C(N)C=C1



分子の表記方法② MDLフォーマット

MDL (Molecular Design Ltd)

- MLDを含むファイル：Structure-Data file (SDF)
- 原子名・結合情報で分子を表す
  (XYZ座標を含む場合も)
- 化合物の性質も書き込める
- 原子名の右側の列
  同位体/電荷/立体の情報

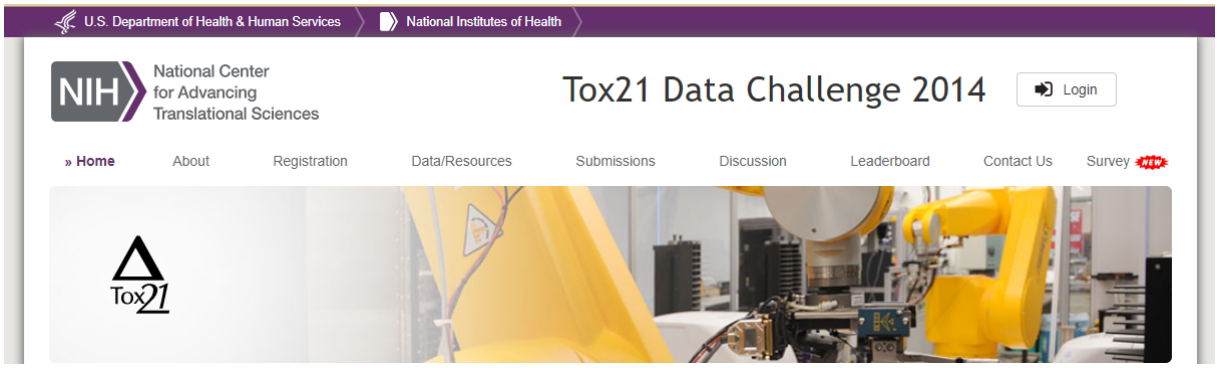


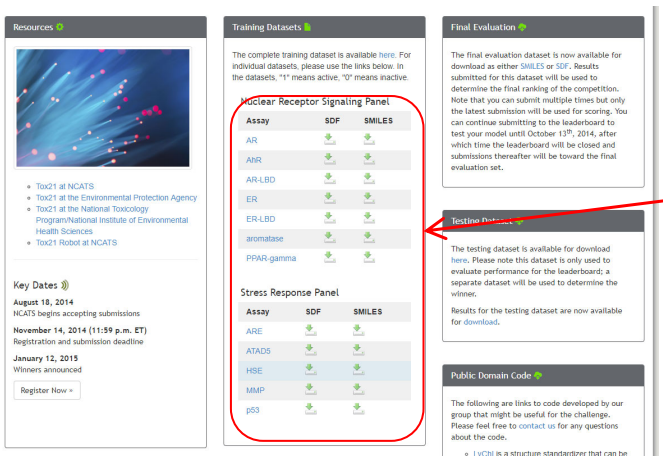
演習

①化合物の構造から毒性を
予測できるか検証しよう

②毒性に大きく関わる重要な
パラメタが存在するか
議論する

化学物質の毒性データ

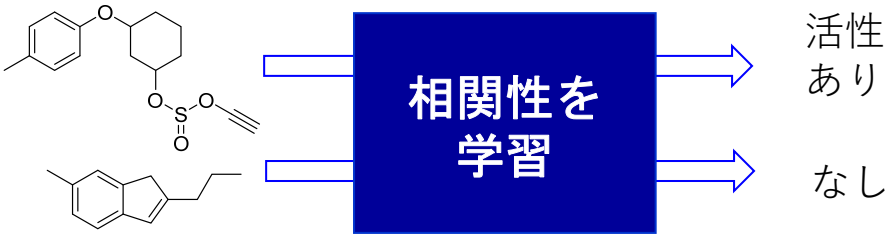




GitHub掲載
データは
ここから取得

機械学習モデルを作るには…

- SDFの情報



- 機械学習モデル構築に向けて考えるべきこと

1）化合物をどう記述するか？
RDKit@Pythonを利用

2）どのように学習するか？
教師あり学習・分類問題
Scikit-learnを利用

- 事前準備(要インストール)

```
!pip install rdkit
import rdkit
from rdkit import rdBase, Chem, DataStructs
 (…以下略…)
```


nr-ar.sdfファイルの中身を見てみよう

```
$$$$
NCGC00181091-01
  Marvin  07111412562D

 29 29  0  0  0  0            999 V2000
    2.1489   -1.6117    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
    2.1489   -2.4303    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
4.3937   -3.4024    0.0000 Cl  0  0  0  0  0  0  0  0  0  0  0  0
  1  2  1  0  0  0  0
  1  6  2  0  0  0  0
  中略
24 25  1  0  0  0  0
M  END
> <Formula>
C25H39ClN2O

> <FW>
419.0430 (382.5820+36.4609)

> <DSSTox_CID>
26837

> <Active>
0
```

毒性の有無が **0/1**で
記録されている