



CBI学会2019年大会、2019年10月21日(月)、13:00-17:00

〈チュートリアル〉
計算毒性学と
化学データサイエンスの基本

株式会社 インシリコデータ
湯田 浩太郎

本日のプログラム:

1. 13:00-13:05: (5分) 挨拶:株式会社 インシリコデータ 湯田浩太郎

2. 13:05-13:20(15分) ◆導入 計算毒性学と「化学データサイエンス」

計算毒性学でのコンピューター導入原理、二大毒性評価関連技術(化学多変量解析/パターン認識アプローチ、人工知能アプローチ)、データサイエンスから「化学データサイエンス」へ

3. 13:20-13:50(30分) ◇第一部 計算機化学(Computer Chemistry)関連

化合物保存形式、化合物命名法、化合物検索(完全一致、部分構造、2・3次元構造検索、他)手法、一元一項対応串刺し検索、化合物の扱い(縮合多環、互変異性、立体/幾何異性)、化合物表記(ケトエノール、ニトロニトロソ、他)

4. 13:50-15:20(90分) ◇第二部 化学多変量解析/パターン認識(ケモトリックス(Chemometrics))関連

化学パラメーター、2/3次元パラメーター、種々データ解析手法、過剰適合、偶然相関、線形/非線形性、特徴抽出、最少サンプル数、最少パラメーター数、クラスポピュレーション、次元変換/圧縮/縮小、分類率/予測率、要因解析、オートスケーリング、アウトライヤー/インライヤー、解析信頼性指標(サンプル数/パラメーター数)、KY(K-step Yard sampling)法、パーセプトロン、バックプロパゲーション、遺伝的アルゴリズム、ファジー理論、内挿/外挿問題、他

<15:20-15:40 休憩 20分>

5. 15:40-16:20(40分) ◇第三部 人工知能(Artificial Intelligence)関連

人工知能の歴史、ルールベース型人工知能、ニューラルネットワーク型人工知能、深層学習、サンプル数問題、要因説明問題、ルールのコンピューターへの組み込み、ネットワーク構造、LISP、FORTRAN、PYTHON、

6. 16:50-17:00(30分) ◇第四部 計算機科学(Computer Science)関連

データベース理論、プログラミング言語、クラスター、クラウド、スーパーコンピューター、ネットワーク、WEB、他

7. 16:50-17:00(10分) ◇討論および名刺交換会

◇第一部

計算機化学(Computer Chemistry)関連

- 化合物保存形式
- 化合物命名法
- 一元一項対応(Canonicalization)
- 化合物検索手法;
(完全一致、部分構造、2・3次元構造検索、他)
- データベース連携(ビッグデータ化);串刺し検索
- 化合物の扱い;縮合多環、互変異性、立体/幾何異性、塩、他
- 化合物表記;ケトエノール、ニトロ/ニトロソ、他

□化学分野の基本情報

◇化学データ(アナログ)のデジタル化

アナログ(英: analog、英語発音: [ˈænəˌlɒɡ] アナローグ):

連続した量(例えば時間)を他の連続した量(例えば角度)で表示すること。デジタルが連続量をとびとびな値(離散的な数値)として表現(標本化・量子化)することと対比される。時計や温度計などがその例である。

<https://ja.wikipedia.org/wiki/アナログ>

□化合物の世界はアナログである

デジタル(英語: digital、英語発音: [ˈdɪdʒəl]。デジタル)


離散量(とびとびの値しかない量)のこと。連続量を表すアナログと反対の概念である。工業的には、状態を示す量を量子化・離散化して処理(取得、蓄積、加工、伝送など)を行う方式のことである。

<https://ja.wikipedia.org/wiki/デジタル>

■コンピューターの世界はデジタルである

□化学分野の基本情報

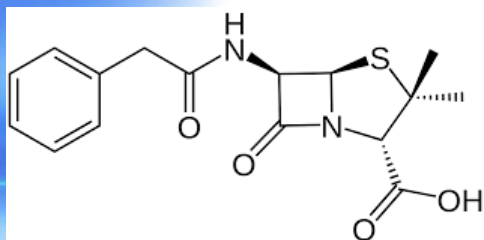
◇化学データ(アナログ)のデジタル化

- 
- * 化合物構造式(アナログデータ、イメージデータ、トポロジカルデータ)
⇒ デジタル情報を基本とするコンピューターでは扱えない
 - * 二次元及び三次元構造式
⇒ コンピューターは一次元で0/1のデータしか扱えない(2/3次元は想定外)
 - * コンピューターサイエンスによる検索及びデータ解析
⇒ デジタル情報を用いて展開されている
 - * 化学分野へのコンピューター適用の技術やサイエンスが存在
⇒ コンピューター化学(Computer Chemistry)が展開されてきた

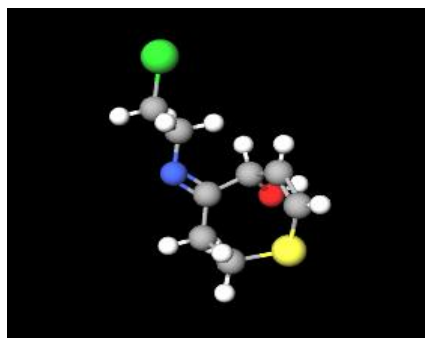
□化学分野の基本情報

◇化学データ(アナログ)のデジタル化

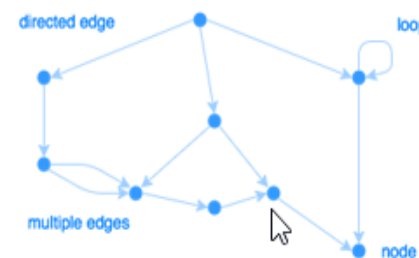
- * 化合物構造式(アナログデータ、イメージデータ、トポロジカルデータ)
⇒ デジタル情報を基本とするコンピューターでは扱えない



二次元化合物構造表示



三次元化合物の
ボール&スティック表示




Graph Convolutional
Networks



001001101101100001010000110101011110010110100101110100
10011111100110100

□化学分野の基本情報

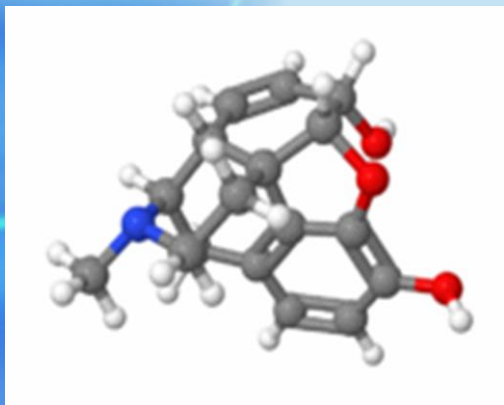
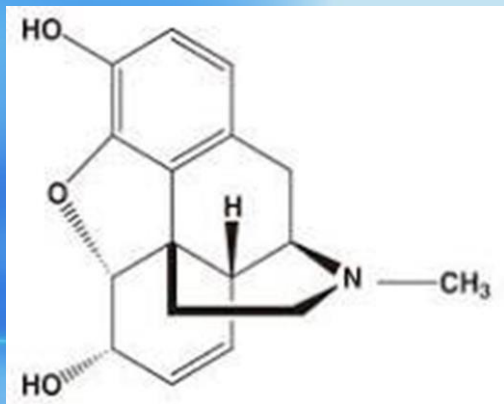
◇化学データ(アナログ)のデジタル化

- 
- * 化合物構造式(アナログデータ、イメージデータ、トポロジカルデータ)
⇒ デジタル情報を基本とするコンピューターでは扱えない
 - * 二次元及び三次元構造式
⇒ コンピューターは一次元で0/1のデータしか扱えない(2/3次元は想定外)
 - * コンピューターサイエンスによる検索及びデータ解析
⇒ デジタル情報を用いて展開されている
 - * 化学分野へのコンピューター適用の技術やサイエンスが存在
⇒ コンピューター化学(Computer Chemistry)が展開されてきた

□ 化合物命名法およびID番号

■ Reproducibility of chemical compounds:

Linear notation and Chemical ID number of compounds



■ Compound Name; Morphine

IUPAC: (5 α ,6 α)-7,8-didehydro-4,5-epoxy-17-methylmorphinan-3,6-diol

SMILES: OC(C=CC1CC2N3C)=C(OC4C(O)C=5)C1C4(CC3)C2C5

InChIKey: InChI=1S/C17H19NO3/c1-18-7-6-17-10-3-5-13(20)16(17)21-15-12(19)4-2-9(14(15)17)8-11(10)18/h2-5,10-11,13,16,19-20H,6-8H2,1H3/t10-,11+,13-,16-,17-/m0/s1

■ Compound Properties

Chemical formula: C₁₇H₁₉NO₃

■ Chemical ID Number

CAS number: 57-27-2

ATC code: N02AA01 (WHO)

PubChem:CID: 5288826

DrugBank: APRD00215

ChemSpider: 4450907

KEGG: D08233

□化学分野の基本情報

■化合物表記法の種類と違い(分子式、化合物名、WLN、Smiles) 線形による化合物表記

①CAS(Chemical Abstracts Service)番号 最新の登録化合物数: 1億4千4百万化合物

- 番号は基本的に登録順で、左の数値、中央の数値を用いた通し番号がつけられる
- 構造や物性などとは関連付けることなく割り当てられ、番号に化学的な意味は持たせていない

異性体は異なる物質なので、CAS登録番号の割り当ても異なる。例えばD-グルコースは50-99-7、L-グルコースは921-60-8である。まれに、分子の種類全体に対して1つのCAS登録番号が割り当てられることもある(全てのアルコール脱水素酵素は9031-72-5である)。

チェックディジットの計算式は次のとおりである。

CAS登録番号が $N_8N_7N_6N_5N_4N_3-N_2N_1-R$ (R, N_i は各桁の0~9の数字、桁が存在しない場合は0とみなす) の場合、

$$R = (8 \times N_8 + 7 \times N_7 + 6 \times N_6 + 5 \times N_5 + 4 \times N_4 + 3 \times N_3 + 2 \times N_2 + N_1) \bmod 10$$

たとえば、水のCAS登録番号は 7732-18-5 なので、以下の通り5になる。

$$(6 \times 7 + 5 \times 7 + 4 \times 3 + 3 \times 2 + 2 \times 1 + 1 \times 8) = 105$$


$$105 \bmod 10 = 5 \quad (105 = 10 \times 10 + 5)$$

<https://ja.wikipedia.org/wiki/CAS登録番号>

□ 多種多様の化合物ファイル形式

■ Reproducibility of chemical compounds: Notation by connection table

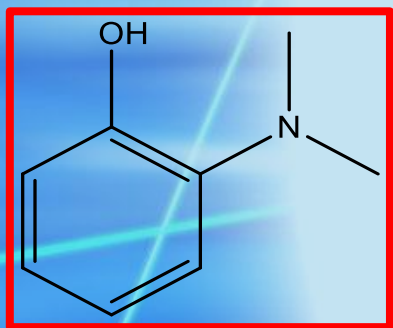
List of file formats handled
by the
“OpenBabel system”



```
mol -- MDL MOL format
pdb -- Protein Data Bank format
smi -- SMILES format
xyz -- XYZ cartesian coordinates format
CONFIG -- DL-POLY CONFIG
CONTCAR -- VASP format
HISTORY -- DL-POLY HISTORY
POSCAR -- VASP format
VASP -- VASP format
abinit -- ABINIT Output Format
acesin -- ACES input format
acesout -- ACES output format
acr -- ACR format
adf -- ADF cartesian input format
adfout -- ADF output format
alc -- Alchemy format
arc -- Accelrys/MSI Biosym/Insight II CAR format
ascii -- ASCII format
axsf -- XCrySDen Structure Format
bfg -- MSI BGF format
box -- Dock 3.5 Box format
bs -- Ball and Stick format
c09out -- Crystal 09 output format
c3d1 -- Chem3D Cartesian 1 format
c3d2 -- Chem3D Cartesian 2 format
cac -- CAChe MolStruct format
cacrt -- Cacao Cartesian format
cache -- CAChe MolStruct format
cacint -- Cacao Internal format
can -- Canonical SMILES format
```

□化合物特定上での問題点

- ◆コンピューター上では全く同じ化合物と認識されるか？
化合物構造式入力に異なる化合物エディターを用いると
同じSmilesでストアしても、全く異なるSmilesとなる
- ◆化合物データベース上で問題が生じる
 - ①化合物の重複登録が発生する
 - ②化合物検索でヒットしなくなる



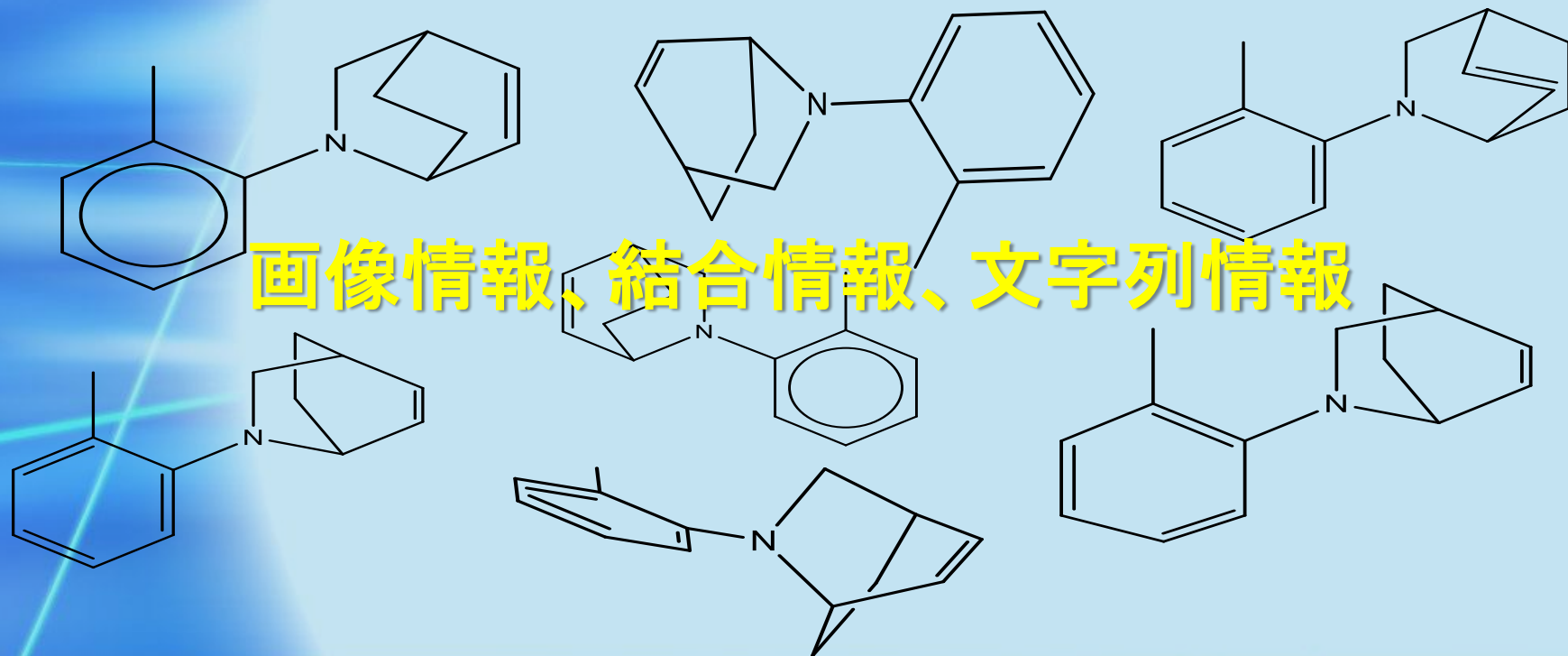
SMILES 1: OC1=C(N(C)C)C=CC=C1 ;by ChemDraw
2: c1(O)c(N(C)C)cccc1 ;by Ecosar
3: C1=CC(=C(C=C1)N(C)C)O ;by QSAR Toolbox
4: CN(C)c1ccccc1O ;by OpenBabel
5: C1=CC(O)=C(N(C)C)C=C1 ;Manual Input by Yuta
6: C1(O)=C(N(C)C)C=CC=C1 ;Manual Input by Yuta

□ 二次元化合物構造式の変化性問題

◆ 全く同じ化合物が作画状態の違いで異なる図となる

1. 化合物の**方向性**の違い(上下/左右/表裏)
2. **表記**の違い(芳香族結合、ブリッジ構造、他)

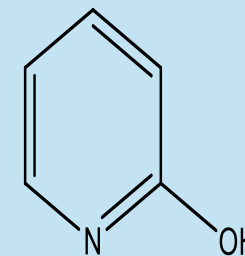
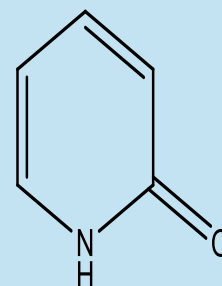
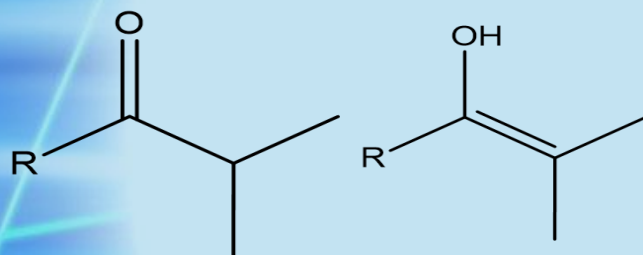
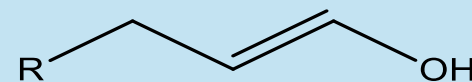
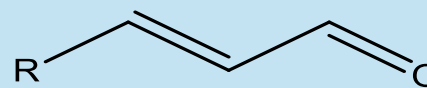
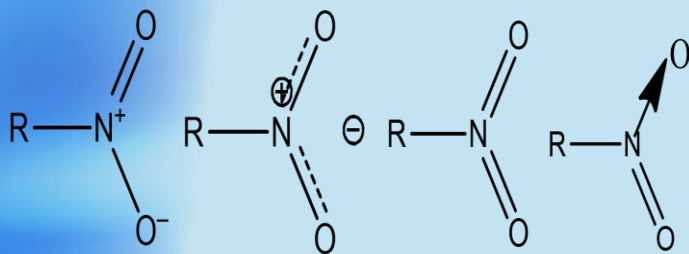
◆ コンピューター上では全く同じ化合物と認識されるか？



□化合物構造表記の多様性に関する問題

◇ Problem in compound structure:

- tautomer
- nitro
- aromatic
- salt



- 作画者の考えや習慣、用途等により変化する
- 表記を間違えたわけではなく、すべて正しい

□化合物のシステム開発や利用上での 化合物操作上での問題点

◆以下の内容に関して、システム利用目的に応じて対応必要

1. 化合物表記手法の変化性
 - ・多種多様の表記法が存在
 - ・同じ表記法であっても内容が異なる
2. 化合物入力時の変化性
 - ・二次元構造式の変化性
 - ・三次元構造式の変化性(ローカル／グローバル)
3. 化合物構造式の多様性
 - ・同じ化合物に正しい表記が複数存在

□化合物のシステム開発や利用上での 化合物操作上での問題点

◆システム開発や利用上での留意点

1. データベースの連携

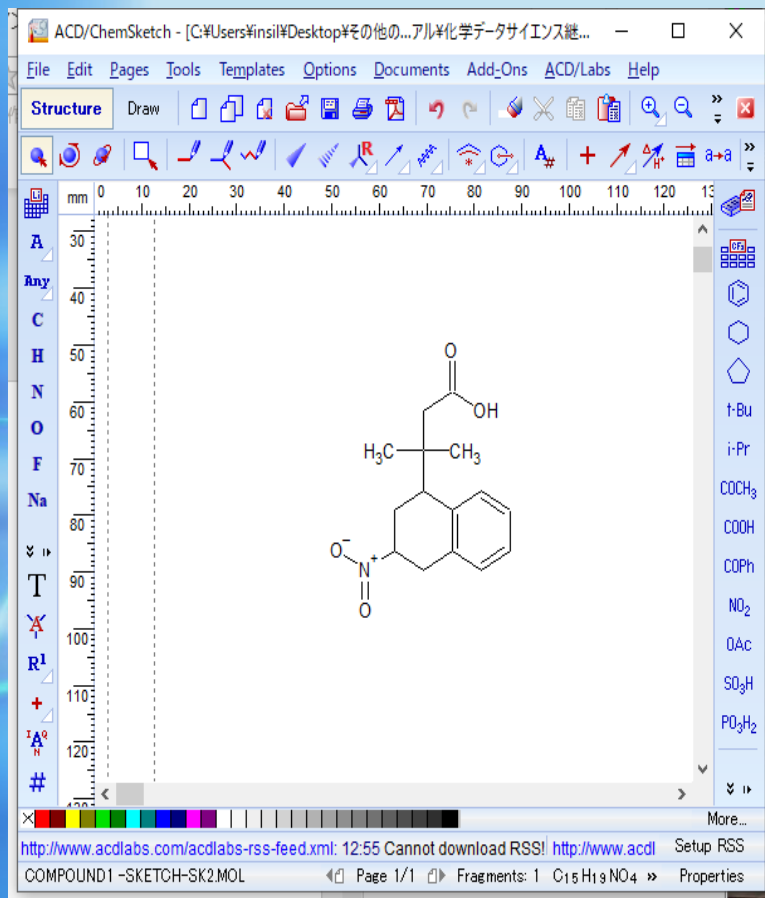
- ・組み合わせ解析が困難で、不安定
- ・単体のデータベースであっても、化合物登録上問題が発生
- ・複数データベースの串刺し検索等が不可能
- ・データを集積してのビッグデータ化が出来ない

2. データサイエンスでの解析

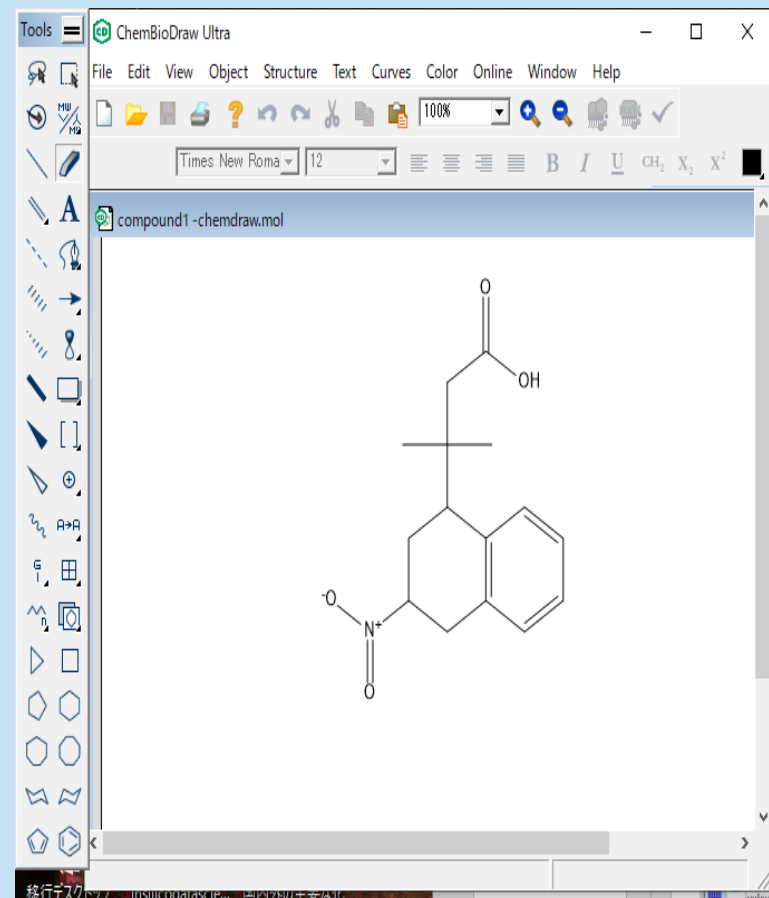
- ・入力構造依存のパラメーター値が変化する
- ・予測モデルの信頼性が低下する
- ・予測結果が入力構造依存となる

□二次元化合物構造式の作画(異なるソフト)

ACD/ChemSketch



ChemBioDraw



□二次元化合物構造式の同一Molファイル

◆同一ファイル形式でも、互換性に留意が必要

・システム依存性、 ・バージョンの変化、記述桁数の違い、他

ACD/ChemSketch

ChemBioDraw

```

compound1 -sketch-sk2.mol -メモ帳
ファイル(F) 編集(E) 書式(O) 表示(V) ヘルプ(H)
ACD/Labs10081913222D
20 21 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 2 V2000
14.2596 -14.9745 0.0000 C 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
14.2596 -16.3045 0.0000 C 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
13.1076 -14.3095 0.0000 C 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
13.1076 -16.9695 0.0000 C 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
11.9559 -14.9745 0.0000 C 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
11.9559 -16.3045 0.0000 C 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
13.1076 -12.9795 0.0000 C 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
14.4376 -12.9795 0.0000 C 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
11.7776 -12.9795 0.0000 C 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
13.1076 -11.6495 0.0000 C 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
10.8040 -16.9695 0.0000 N 0 0 3 0 0 0 0 0 0 0 0 0 0 0 0 0
9.6522 -16.3045 0.0000 O 0 0 3 0 0 0 0 0 0 0 0 0 0 0 0 0
10.8040 -18.2995 0.0000 O 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
14.2594 -10.9845 0.0000 C 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
14.2594 -9.6545 0.0000 O 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
15.4113 -11.6496 0.0000 O 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
16.5632 -14.9745 0.0000 C 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
16.5632 -16.3045 0.0000 C 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
15.4113 -14.3095 0.0000 C 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
15.4113 -16.9695 0.0000 C 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
3 1 1 0 0 0 0 0
4 2 1 0 0 0 0 0
5 3 1 0 0 0 0 0
6 4 1 0 0 0 0 0
8 5 1 0 0 0 0 0
9 7 1 0 0 0 0 0
10 7 1 0 0 0 0 0
3 7 1 0 0 0 0 0
12 11 1 0 0 0 0 0
13 11 2 0 0 0 0 0
6 11 1 0 0 0 0 0
15 14 2 0 0 0 0 0
16 14 1 0 0 0 0 0
10 14 1 0 0 0 0 0
18 17 1 0 0 0 0 0
19 17 2 0 0 0 0 0
20 18 2 0 0 0 0 0
1 19 1 0 0 0 0 0
2 20 1 0 0 0 0 0
2 1 2 0 0 0 0 0
M CHG 2 11 1 12 -1
M END

```

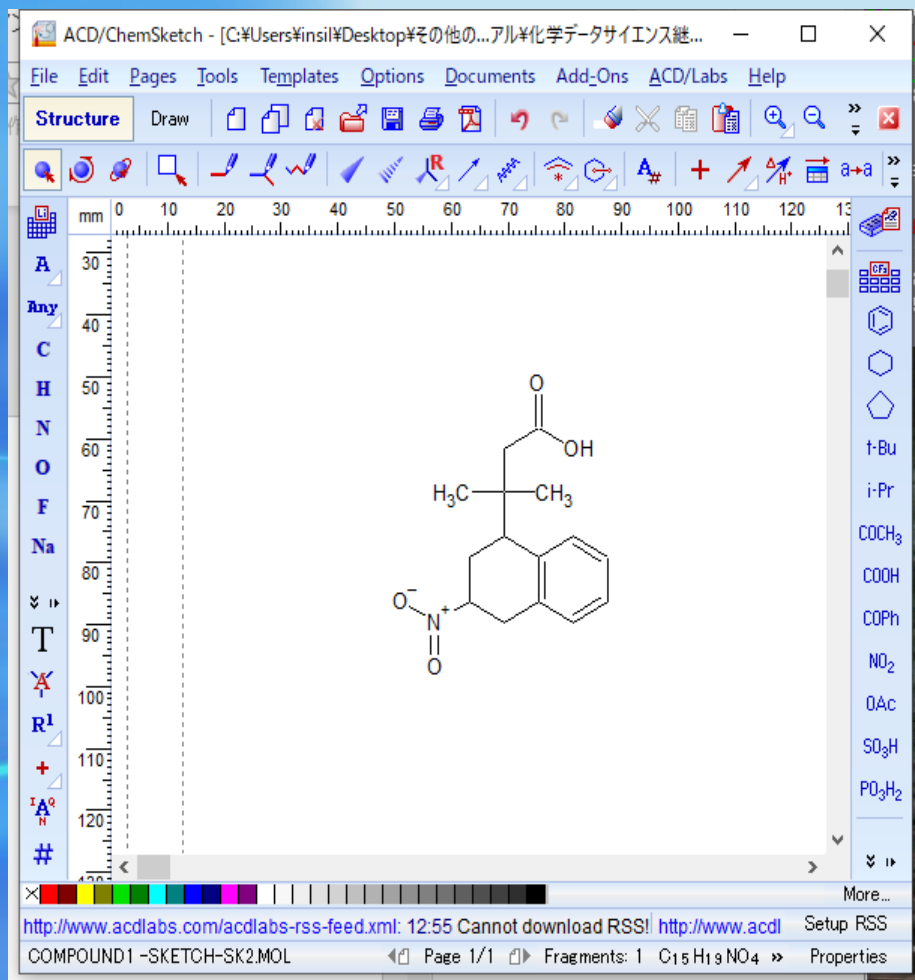
```

compound1 -chemdraw.mol -メモ帳
ファイル(F) 編集(E) 書式(O) 表示(V) ヘルプ(H)
compound1 -chemdraw.cdx
ChemDraw10081913352D
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 V3000
M V30 BEGIN CTAB
M V30 COUNTS 20 21 0 0 0
M V30 BEGIN ATOM
M V30 1 C 0.713867 -0.619011 0.000000 0
M V30 2 C 0.713867 -1.444011 0.000000 0
M V30 3 C 1.428338 -1.856511 0.000000 0
M V30 4 C 2.142809 -1.444011 0.000000 0
M V30 5 C 2.142809 -0.619011 0.000000 0
M V30 6 C 1.428338 -0.206511 0.000000 0
M V30 7 C -0.715075 -0.619011 0.000000 0
M V30 8 C -0.000804 -1.856511 0.000000 0
M V30 9 C -0.715075 -1.444011 0.000000 0
M V30 10 N -1.428640 -1.855988 0.000000 0 CHG=1
M V30 11 O -2.142809 -1.442966 0.000000 0 CHG=-1
M V30 12 O -1.428036 -2.680989 0.000000 0
M V30 13 O -0.000804 -0.206511 0.000000 0
M V30 14 C -0.825804 0.618489 0.000000 0
M V30 16 C 0.824396 0.618489 0.000000 0
M V30 17 C -0.000804 1.443489 0.000000 0
M V30 18 C 0.713867 1.855989 0.000000 0
M V30 19 O 0.713867 2.680989 0.000000 0
M V30 20 O 1.428338 1.443489 0.000000 0
M V30 END ATOM
M V30 BEGIN BOND
M V30 1 2 1 2
M V30 2 1 2 3
M V30 3 2 3 4
M V30 4 1 4 5
M V30 5 5 6 1
M V30 6 6 1 13
M V30 8 1 13 7
M V30 9 7 9 8
M V30 10 1 9 8
M V30 11 1 8 2
M V30 12 1 9 10
M V30 13 1 10 11
M V30 14 2 10 12
M V30 15 1 13 14
M V30 16 1 14 15
M V30 17 1 14 16
M V30 18 1 14 17
M V30 19 1 17 18
M V30 20 2 18 19
M V30 2 1 13 20
M V30 END BOND
M V30 END CTAB
M END

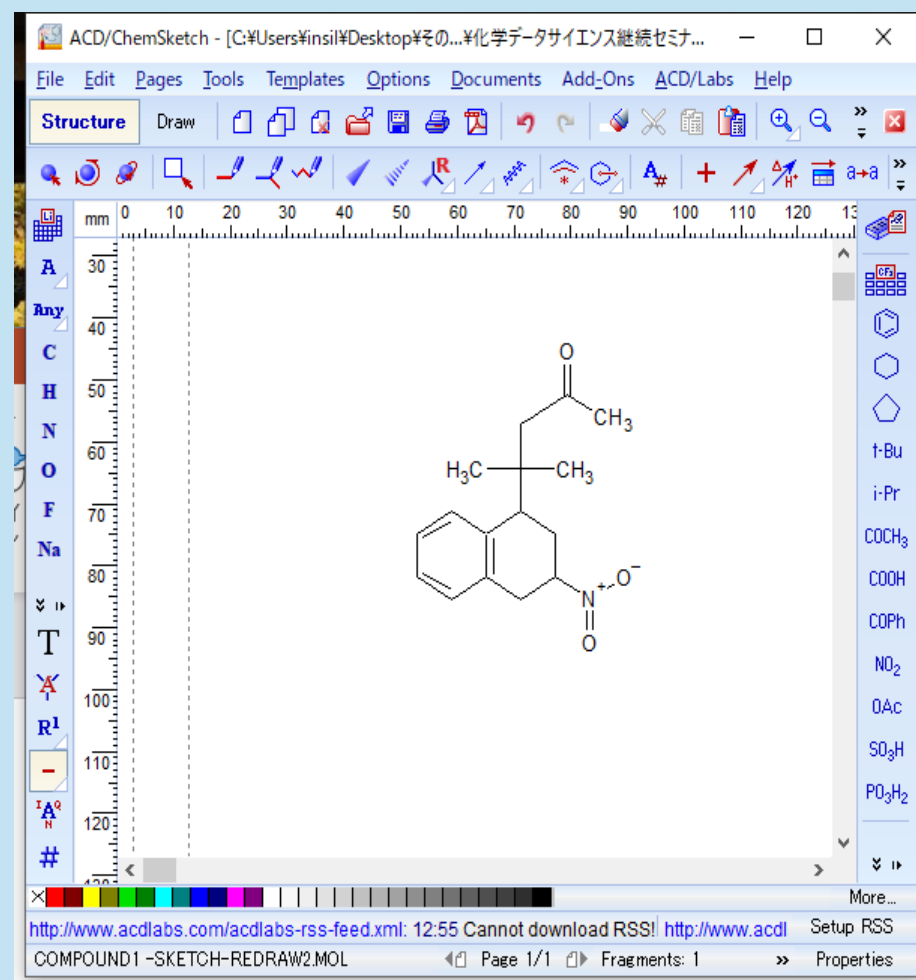
```

□ 二次元化合物構造式の作画 (同一のソフト)

ACD/ChemSketch



ACD/ChemSketch



□二次元化合物構造式の同一Molファイル

◆原子の番号付けが異なる

同一システム作成の同一化合物であっても、原子番号が異なる
原子番号の一元化が必要 (Morganアルゴリズム、SEMA等)

◆化合物の種々検索での問題解決が必要

compound1-sketch-sk2.mol - メモ帳

ファイル(F) 編集(E) 書式(O) 表示(V) ヘルプ(H)

```

ACD/Labs10081913222D
20 21 0 0 0 0 0 0 0 0 0 0 2 V2000
14.2596 -14.9745 0.0000 C 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
14.2596 -16.3045 0.0000 C 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
13.1076 -14.3095 0.0000 C 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
13.1076 -16.9695 0.0000 C 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
11.9559 -14.9745 0.0000 C 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
11.9559 -16.3045 0.0000 C 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
13.1076 -12.9795 0.0000 C 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
14.4376 -12.9795 0.0000 C 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
11.7776 -12.9795 0.0000 C 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
13.1076 -11.6495 0.0000 C 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
10.8040 -16.9695 0.0000 N 0 0 3 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
9.8522 -16.3045 0.0000 O 0 0 5 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
10.8040 -16.2995 0.0000 O 0 0 5 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
14.2594 -10.9845 0.0000 C 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
14.2594 -9.8545 0.0000 O 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
15.4113 -11.6496 0.0000 O 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
16.5632 -14.9745 0.0000 C 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
16.5632 -16.3045 0.0000 C 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
15.4113 -14.3095 0.0000 C 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
15.4113 -16.9695 0.0000 C 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
3 1 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
4 2 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
5 3 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
6 4 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
7 5 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
8 6 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
9 7 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
10 8 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
11 9 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
12 10 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
13 11 2 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
14 12 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
15 13 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
16 14 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
17 15 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
18 16 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
19 17 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
20 18 2 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
21 19 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
22 20 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
23 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
24 2 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
M CHG 11 1 12 -1
M END

```

compound1-sketch-redraw2.mol - メモ帳

ファイル(F) 編集(E) 書式(O) 表示(V) ヘルプ(H)

```

ACD/Labs10081913533D
20 21 0 0 0 0 0 0 0 0 0 0 2 V2000
12.3206 -14.1573 0.0000 C 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
12.3206 -15.4873 0.0000 C 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
13.4724 -13.4924 0.0000 C 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
14.6242 -14.1575 0.0000 C 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
13.4723 -16.1523 0.0000 C 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
14.6241 -15.4873 0.0000 C 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
11.1689 -13.4924 0.0000 C 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
10.0171 -14.1574 0.0000 C 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
11.1689 -16.1523 0.0000 C 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
10.0172 -15.4873 0.0000 C 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
13.4725 -12.1624 0.0000 C 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
14.9184 -12.1623 0.0000 C 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
12.0181 -12.1623 0.0000 C 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
13.4725 -10.3324 0.0000 C 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
15.0306 -9.9189 0.0000 C 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
16.1868 -10.5762 0.0000 C 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
15.0218 -8.5889 0.0000 O 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
15.7759 -16.1524 0.0000 N 0 0 3 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
15.7758 -17.4824 0.0000 O 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
16.9277 -15.4875 0.0000 O 0 0 5 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
3 1 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
4 2 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
5 3 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
6 4 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
7 5 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
8 6 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
9 7 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
10 8 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
11 9 2 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
12 10 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
13 11 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
14 12 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
15 13 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
16 14 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
17 15 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
18 16 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
19 17 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
20 18 2 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
21 19 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
22 20 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
23 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
24 2 18 1 20 -1
M CHG 18 1 20 -1
M END

```

□化合物原子の番号付け

■完全一致検索

1. 変換コードを用いたアプローチ

①MORGAN名: 立体情報を持たない化合物

MORGAN名とは化合物に一元一項対応で付けられた化合物名

MORGAN名 = ユニークナンバリング + 原子／結合情報

ユニークナンバリング:

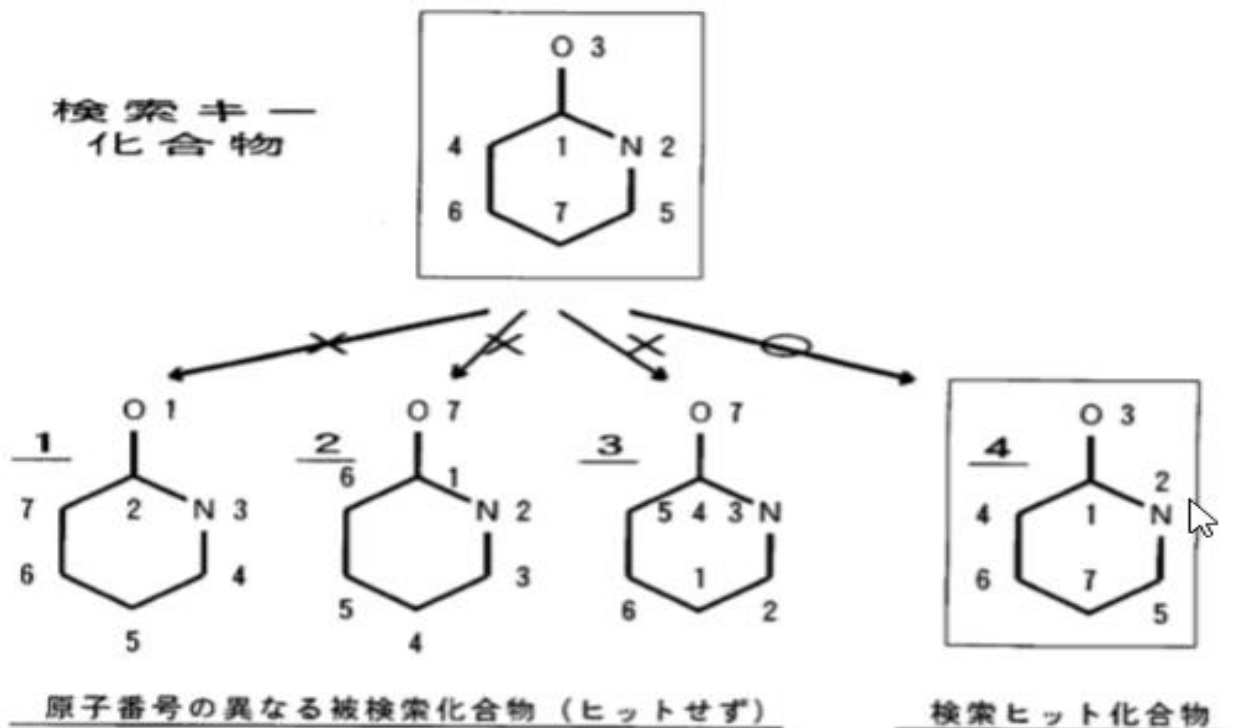
化合物を構成する原子につけられる番号を、1化合物1通りに決定すること

□化合物原子の番号付け

■MORGAN名の実施形態

化合物に付けられた原子番号が異なると右図のように、まったく同じ化合物であってもコンピューター内部では異なる化合物として認識する。

この故、化合物の原子に付けられる番号はユニーク（一元一項対応）でなければならない



□一元一項対応の重要性

化合物情報をコンピューターで扱う上での
重要な基本事項

◆一元一項対応 (Canonicalization: 規範化) とは
一つの化合物は同じ記述方式であれば1:1に定義される

◆一元多項対応とは
一つの化合物が複数の記述(1:多)で定義される

□化学分野の基本情報

■一元一項対応関連の考え

一元一項対応 ⇒ 化合物(1) ⇒ 化合物名(1)

多元一項対応 ⇒ 化合物(N) ⇒ 化合物名(1)

一元多項対応 ⇒ 化合物(1) ⇒ 化合物名(N)

多元多項対応 ⇒ 化合物(N) ⇒ 化合物名(N)

但し、化合物名＝化合物表記
()内は数を示す

□一元一項対応の重要性

◆一元多項対応の時に発生する問題点

1. 化合物データベース関連上での問題

- ① 多重登録が頻発する
- ② 化合物検索の精度が保たれない
- ③ 複数の化合物データベース間の連携が困難

2. 化合物データ解析関連上での問題

- ① 重複データの存在可能性
- ② パラメーターの不安定性
- ③ 予測モデルの汎用性減少

□化学分野の基本情報

■なぜ一元一項対応が重要となるのか

◇化合物に関する
検索や化合物データ解析を
想定すると

一元一項対応が必須

実在の化合物

化合物表記

一元一項対応	⇒	化合物(1)	⇒	化合物名(1)
化合物検索	○	データ解析	○	

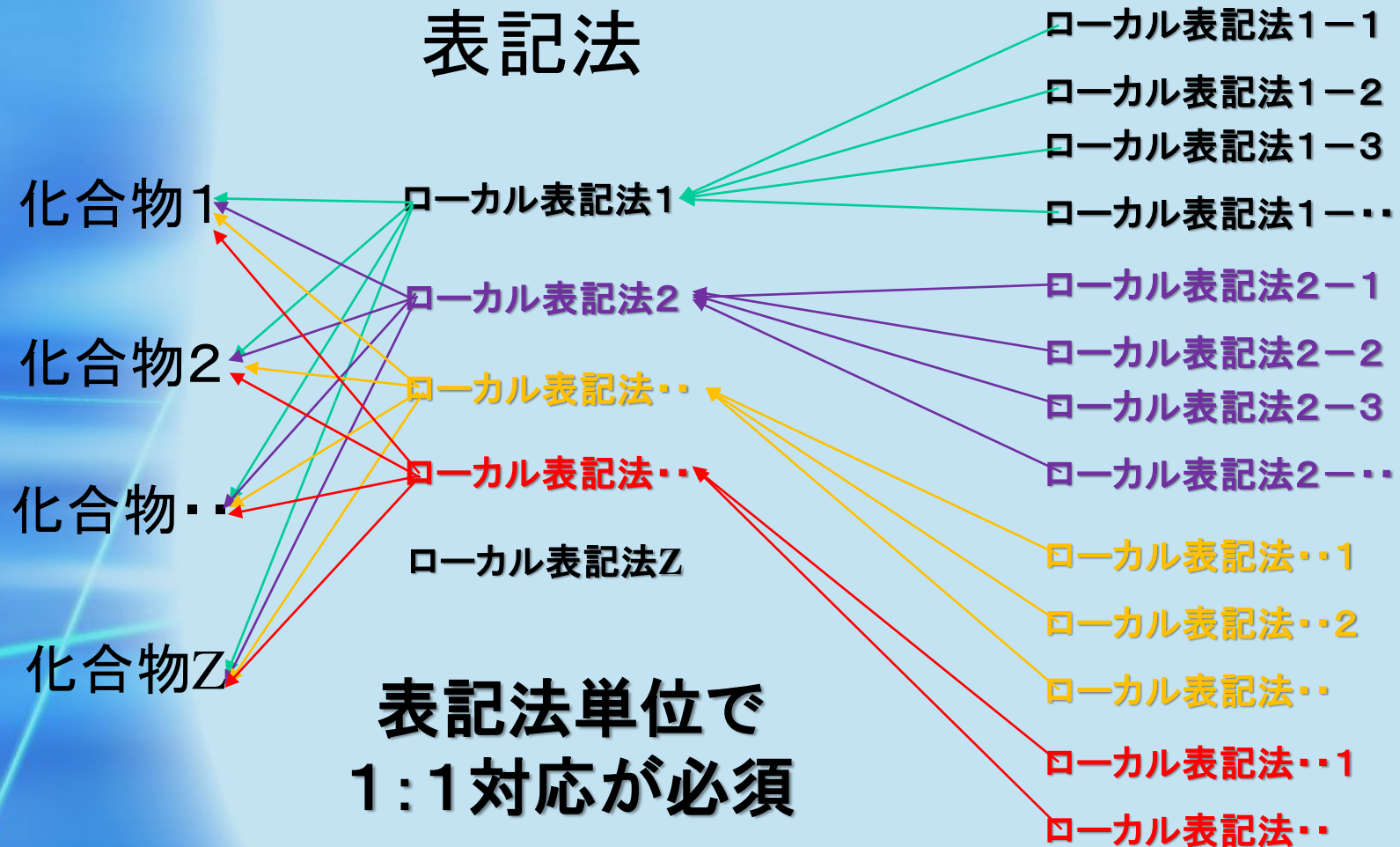
多元一項対応	⇒	化合物(N)	⇒	化合物名(1)
化合物検索	×	データ解析	×	

一元多項対応	⇒	化合物(1)	⇒	化合物名(N)
化合物検索	×	データ解析	×	

多元多項対応	⇒	化合物(N)	⇒	化合物名(N)
化合物検索	×	データ解析	×	

□一元一項対応の概念図

一元一項対応の 表記法

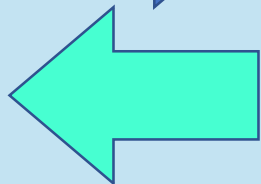
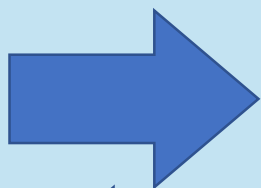
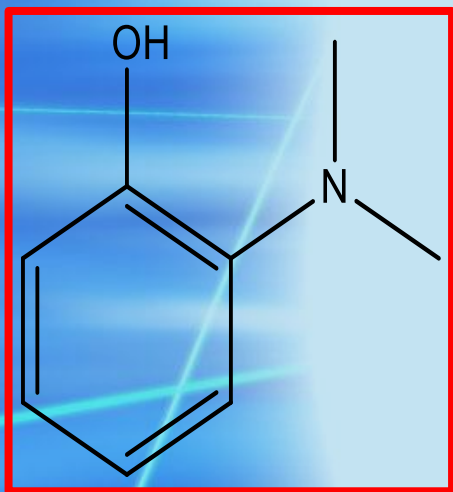


□一元一項対応の重要性 (Smilesを例に)

化合物

表記法およびローカル表記

重複登録
検索ヒットせず

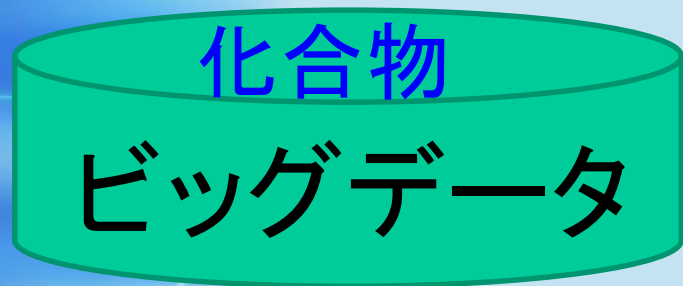


SMILES

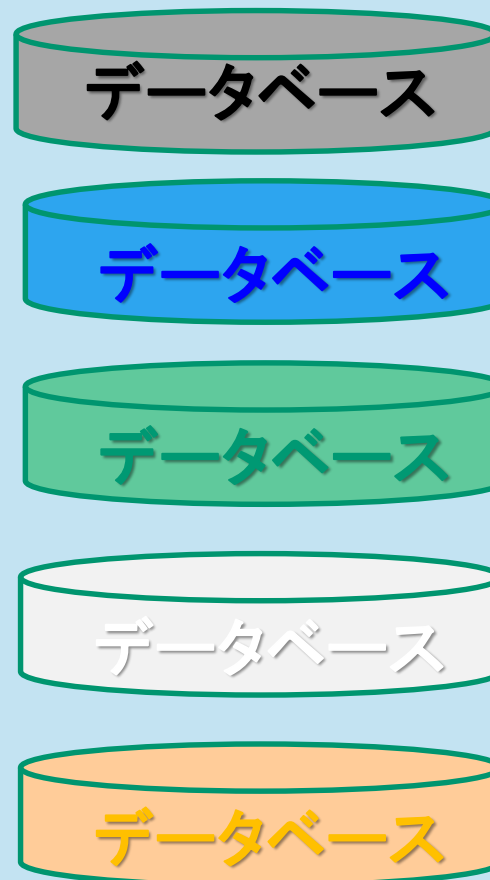
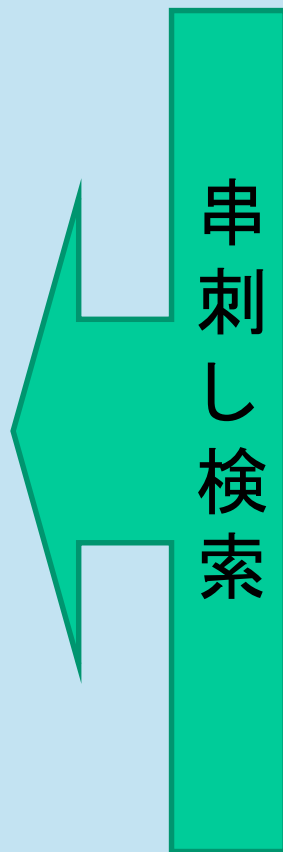
- 1: OC1=C(N(C)C)C=CC=C1 ;by ChemDraw
- 2: c1(O)c(N(C)C)cccc1 ;by Ecosar
- 3: C1=CC(=C(C=C1)N(C)C)O ;by QSAR Toolbox
- 4: CN(C)c1ccccc1O ;by OpenBabel
- 5: C1=CC(O)=C(N(C)C)C=C1 ;Manual Input by Yuta
- 6: C1(O)=C(N(C)C)C=CC=C1 ;Manual Input by Yuta

データベース連携や統合によるビッグデータ化

データベース統合による
化合物ビッグデータ化



極めて大きな
化合物数



◇ビッグデータの5Vとは

データベース構築では達成の最適化目標として、右に示される5種類のVの実現が理想とされる

データの価値
(Value)

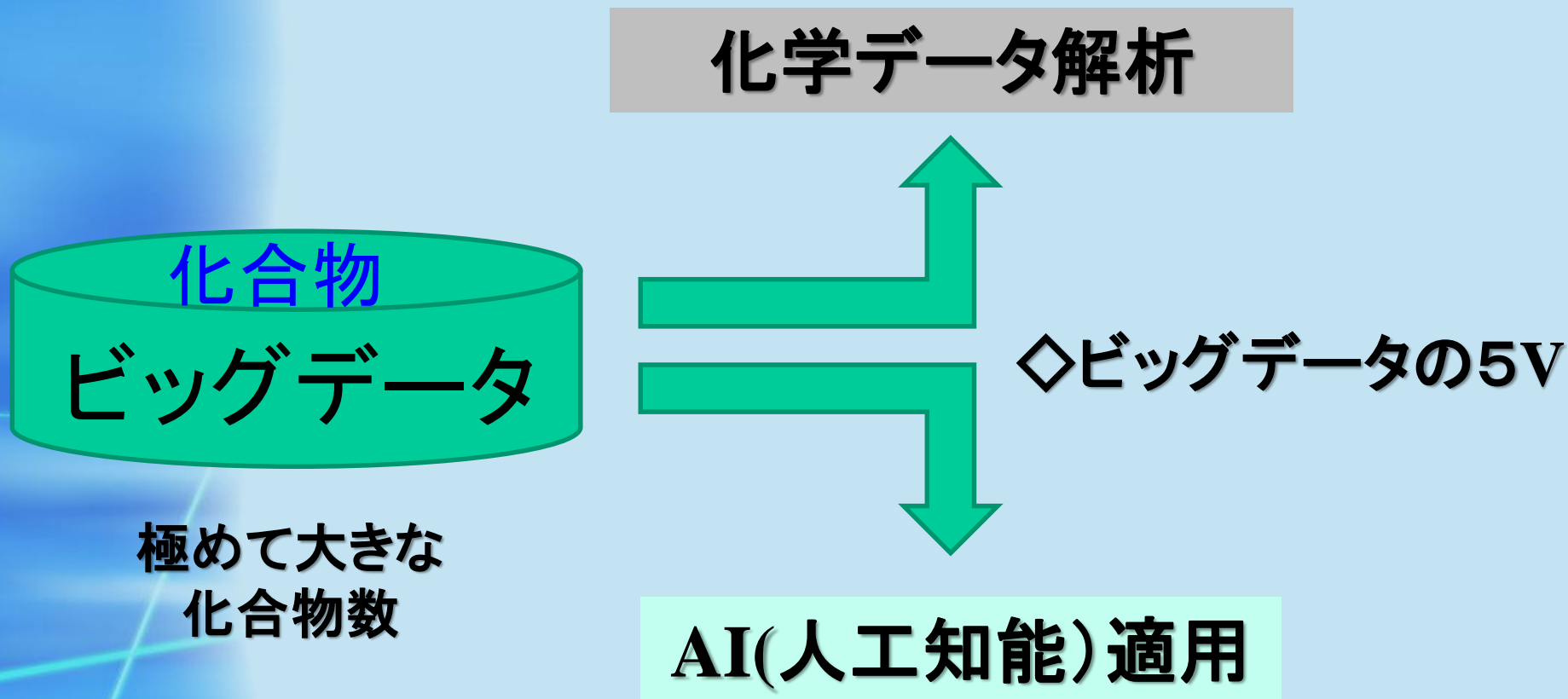
データ量 (Volume)

データの収集の速さ (Velocity)

データの種類 (Variety)

データの正確さ (Veracity)

□ビッグデータ構築上での留意点



□ 化学関連システム間連携上での留意点

