◇第一部　計算機化学（Computer Chemistry）関連
Part1. Computer Chemistry

株式会社　インシリコデータ
湯田　浩太郎

Contents:

◇第一部
計算機化学（Computer Chemistry）関連

□化合物保存形式（Compound storage format）
□化合物命名法（Compound nomenclature）
□一元一項対応（Canonicalization）
□化合物検索手法；（Compound search method）
（完全一致、部分構造、２・３次元構造検索、他）
(Complete match, partial structure, 2 / 3-dimensional structure search, etc.)
□データベース連携（ビッグデータ化）；串刺し検索
（Database linkage (big data); skewered search）
□化合物の扱い；縮合多環、互変異性、立体／幾何異性、塩、他
（Handling of compounds; condensed polycycles, tautomerism, stereo / geometric isomerism, salts, etc.）
□化合物表記；ケトエノール、ニトロ／ニトロソ、他
（Compound notation; ketoenol, nitro / nitroso, etc.）

□化学分野の基本情報

◇化学データ（アナログ）のデジタル化：Digitization of chemical data（analog）

アナログ（英: analog、英語発音: [ˈænəˌlɔːg] 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: [ˈdɪdʒətl]。ディジタル）
　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　Basic information in the chemical field

◇化学データ(アナログ)のデジタル化:Digitization of chemical data（analog）

＊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)

Compound structural formula　(analog data, image data, topological data)

⇒　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

Cannot be handled by a computer based on digital information

＊二次元及び三次元構造式　2D and 3D structural formulas

⇒　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)

The computer can handle only 0/1 data in one dimension
(2/3 dimensions are not expected)

＊コンピューターサイエンスによる検索及びデータ解析

Computer science search and data analysis

⇒デジタル情報を用いて展開されている⇒Developed using digital information

＊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

Technology and science of computer application to the chemical field exist

⇒コンピューター化学(Computer Chemistry)が展開されてきた

Computer Chemistry has been deployed

□化学分野の基本情報：Basic information on the chemical field

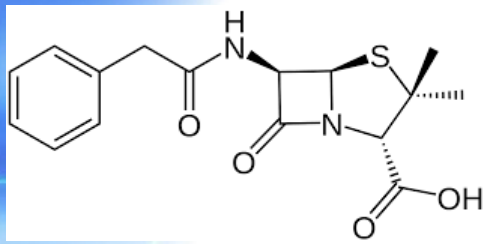
◇化学データ（アナログ）のデジタル化　；　Digitization of chemical data（analog）

＊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）

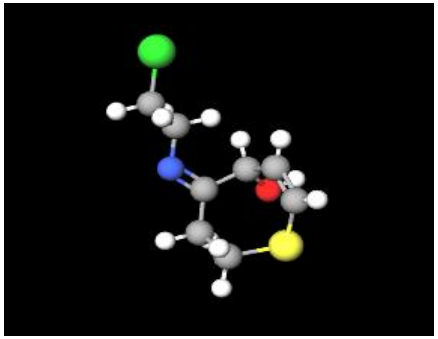
Compound structural formula　(analog data, image data, topological data)

⇒　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

Cannot be handled by a computer based on digital information



二次元化合物構造表示
2D compound structure display



三次元化合物の
ボール＆スティック表示
Three-dimensional compound
Ball and stick display

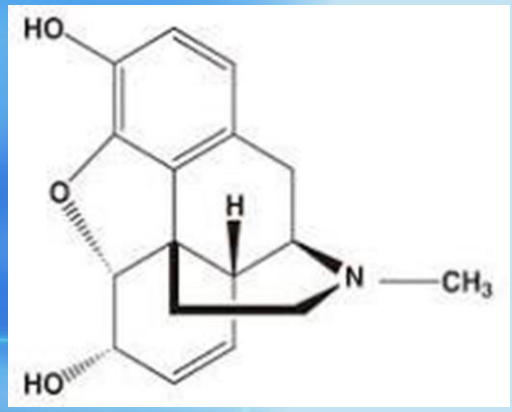
Graph Convolutional Networks

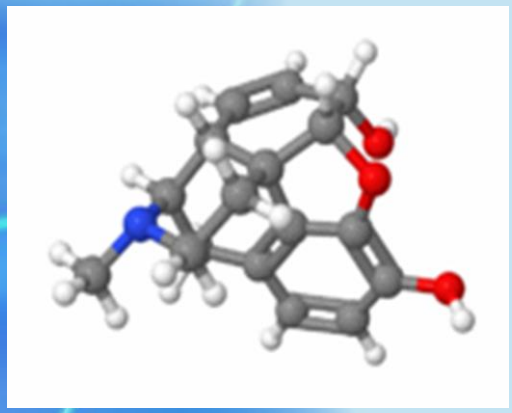
0010011011011000010100001101010111100101101001011101001001111111001101 00

□化合物命名法およびID番号

Compound nomenclature and ID numbers

■Reproducibility of chemical compounds: Linear notation and Chemical ID number of compounds





■Compound Name; Morphine

IUPAC: (5α,6α)-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：Basic information on the chemical field

■化合物表記法の種類と違い（分子式、化合物名、WLN、Smiles）

Types and differences in compound notation
(molecular formula, compound name, WLN, Smiles)

線形による化合物表記：Linear notation of compound

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

CAS (Chemical Abstracts Service) number Number of the latest registered compounds: 144 million compounds

- 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。

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

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

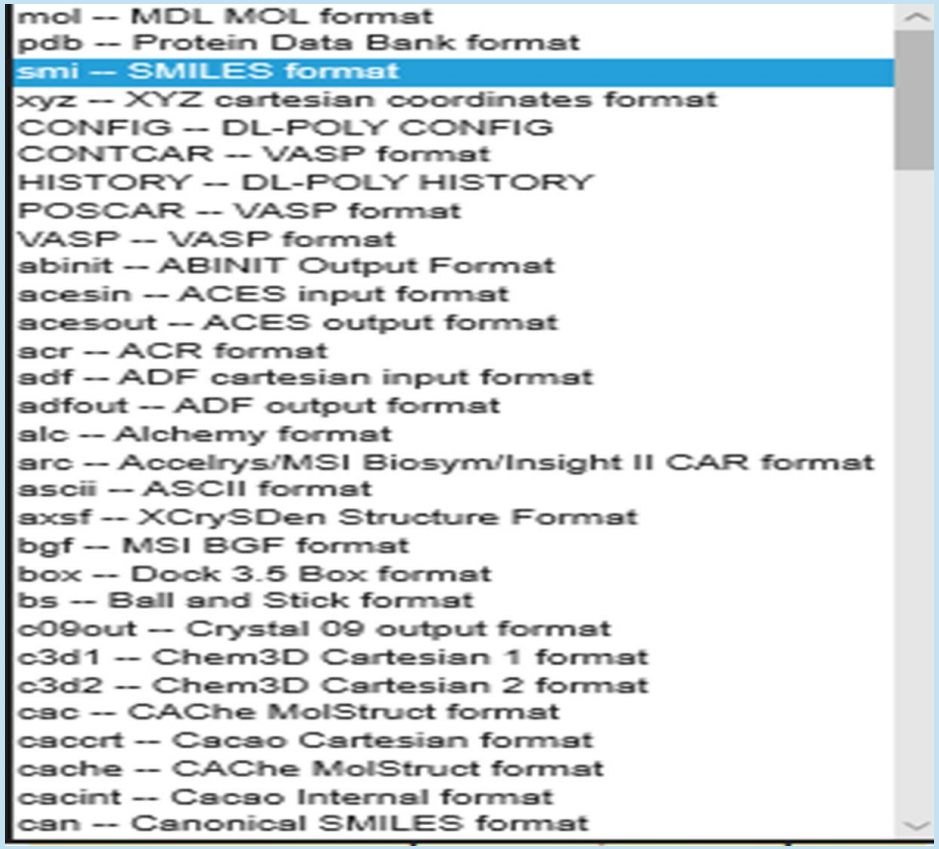
https://ja.wikipedia.org/wiki/CAS登録番号

□多種多様の化合物ファイル形式
A wide variety of compound file formats

In Silico Data
Miracles by the KY-methods

■Reproducibility of chemical compounds: Notation by connection table

List of file formats handled by the "OpenBabel system"



口化合物特定上での問題点 :Problems in compound identification

◆コンピューター上では全く同じ化合物と認識されるか？

Is it recognized as exactly the same compound on the computer?

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

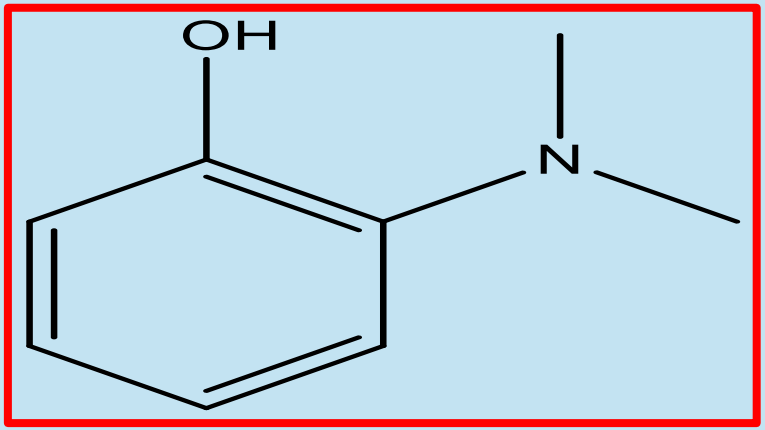
If you use a different compound editor to enter the compound structure
Even if you store with the same Smiles, it will be a completely different Smiles

◆化合物データベース上で問題が生じる

Problems occur in the compound database

①化合物の**重複登録**が発生するCompound duplicate registration occurs
②化合物検索で**ヒットしなくなる**No hit in compound search

口化合物特定上での問題点 :Problems in compound identification



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
Variability problem of two-dimensional compound structural formula

◆全く同じ化合物が作画状態の違いで異なる図となる
Exactly the same compound is different depending on the drawing state

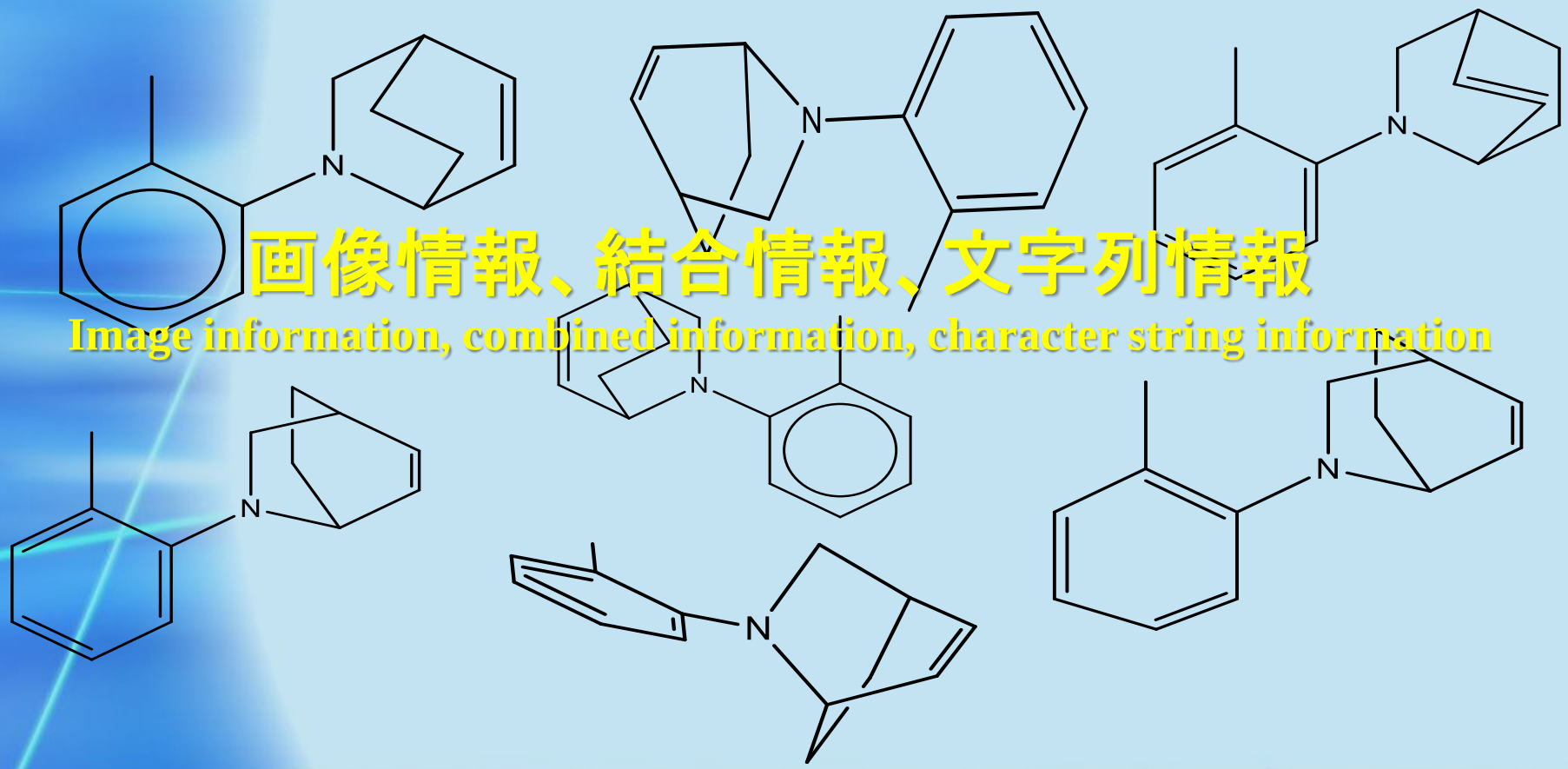
1．化合物の方向性の違い（上下／左右／表裏）
Difference in directionality of compounds (up / down / left / right / front / back)

2．表記の違い（芳香族結合、ブリッジ構造、他）
Difference in notation (aromatic bond, bridge structure, etc.)

◆コンピューター上では全く同じ化合物と認識されるか？
Is it recognized as exactly the same compound on the computer?

□二次元化合物構造式の変化性問題

Variability problem of two-dimensional compound structural formula

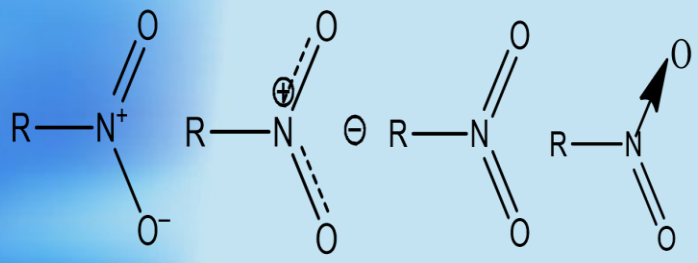


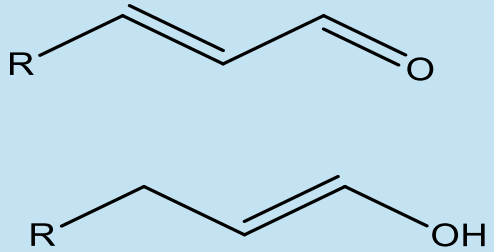
□化合物構造表記の多様性に関する問題

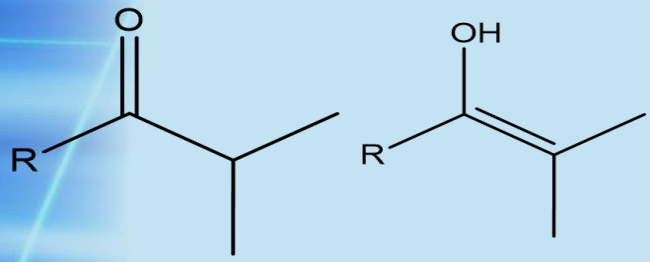
Problems related to the diversity of compound structure notation

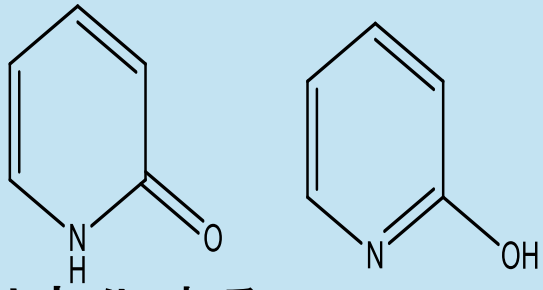
◇ Problem in compound structure:

・tautomer ・nitro ・aromatic ・salt









・作画者の考えや習慣、用途等により変化する

Changes depending on the creator's thoughts, customs, and usage

・表記を間違えたわけではなく、すべて正しい

It is not a mistake in the notation, everything is correct

□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

In system development and use of compounds Problems in compound operation

◆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

The following content is required depending on the purpose of system use.

1. 化合物表記手法の変化性:Variability in compound notation
   - ・多種多様の表記法が存在・Various notations exist
   - ・同じ表記法であっても内容が異なる

     Even if the same notation is used, the contents are different.
2. 化合物入力時の変化性:Variability when entering compounds
   - ・二次元構造式の変化性・Change in two-dimensional structural formula
   - ・三次元構造式の変化性(ローカル／グローバル)

     Change in 3D structural formula (local / global)
3. 化合物構造式の多様性: Diversity of compound structural formulas
   - ・同じ化合物に正しい表記が複数存在

     Multiple correct notations exist for the same compound

□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
In system development and use of compounds Problems in compound operation

◆システム開発や利用上での留意点
Notes on system development and use

1．データベースの連携:Database linkage
- 組み合わせ解析が困難で、不安定
  Combination analysis is difficult and unstable
- 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
  There is a problem in compound registration even with a single database
- 複数データベースの串刺し検索等が不可能
  Multi-database skewer search is not possible
- データを集積してのビッグデータ化が出来ない
  Big data cannot be created by collecting data

□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

In system development and use of compounds Problems in compound operation

◆システム開発や利用上での留意点

Notes on system development and use

2. データサイエンスでの解析:Data science analysis

- 入力構造依存のパラメーター値が変化する

  Parameter value dependent on input structure changes

- 予測モデルの信頼性が低下する

  Reliability of prediction model decreases

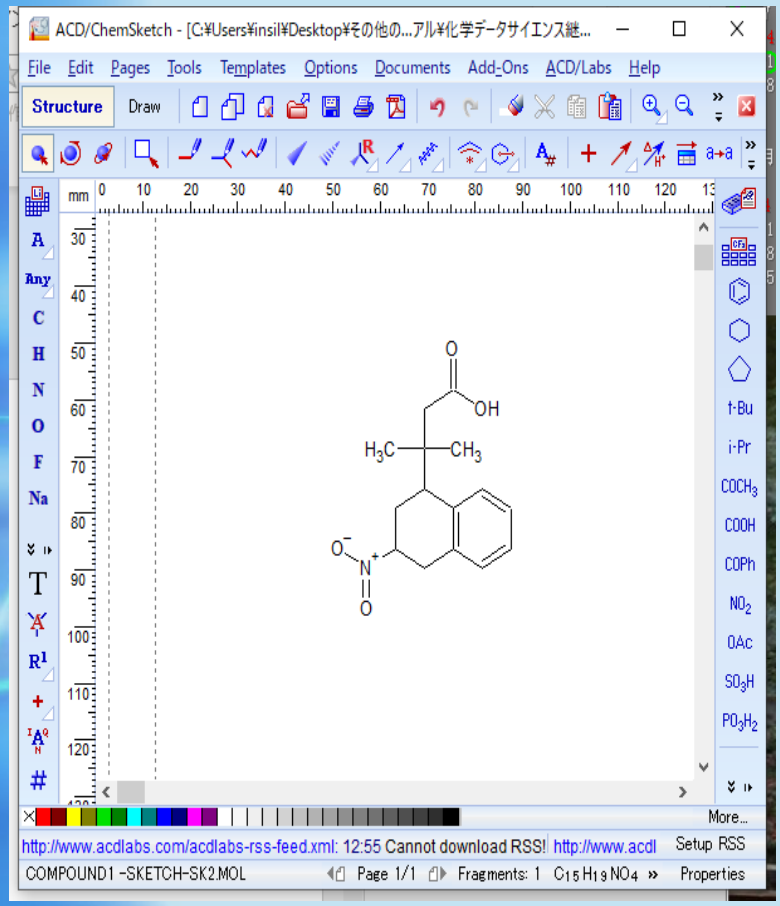
- 予測結果が入力構造依存となる

  The prediction result depends on the input structure

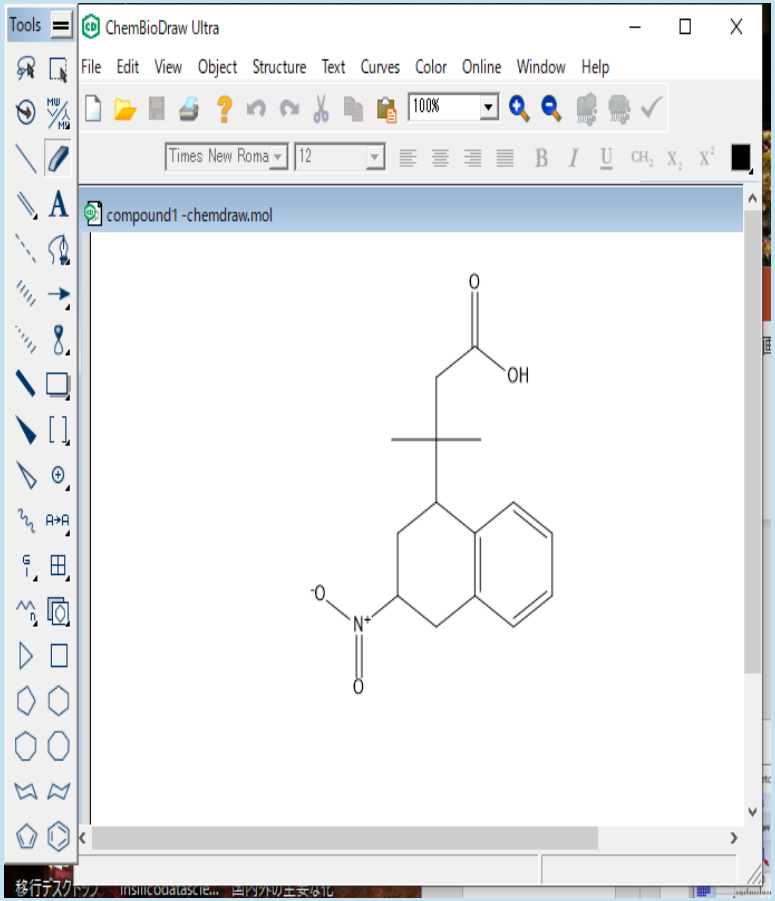
二次元化合物構造式の作画（異なるソフト）

Drawing of two-dimensional compound structural formula (different software)

ACD/ChemSketch



ChemBioDraw



□二次元化合物構造式の同一Molファイル

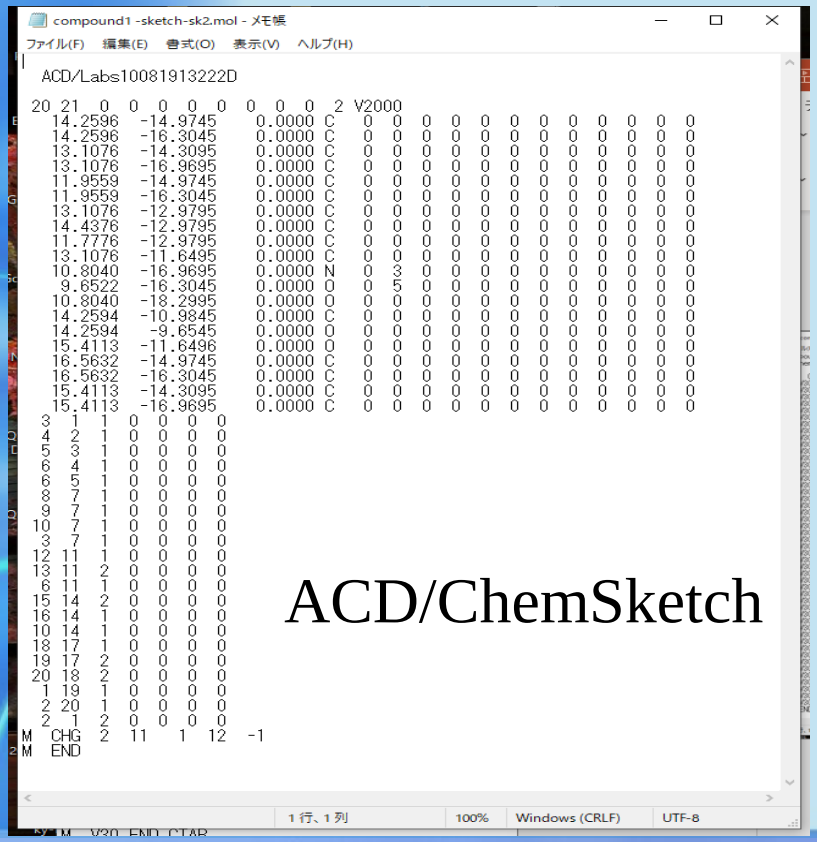
Same Mol file of 2D compound structural formula

◆同一ファイル形式でも、互換性に留意が必要

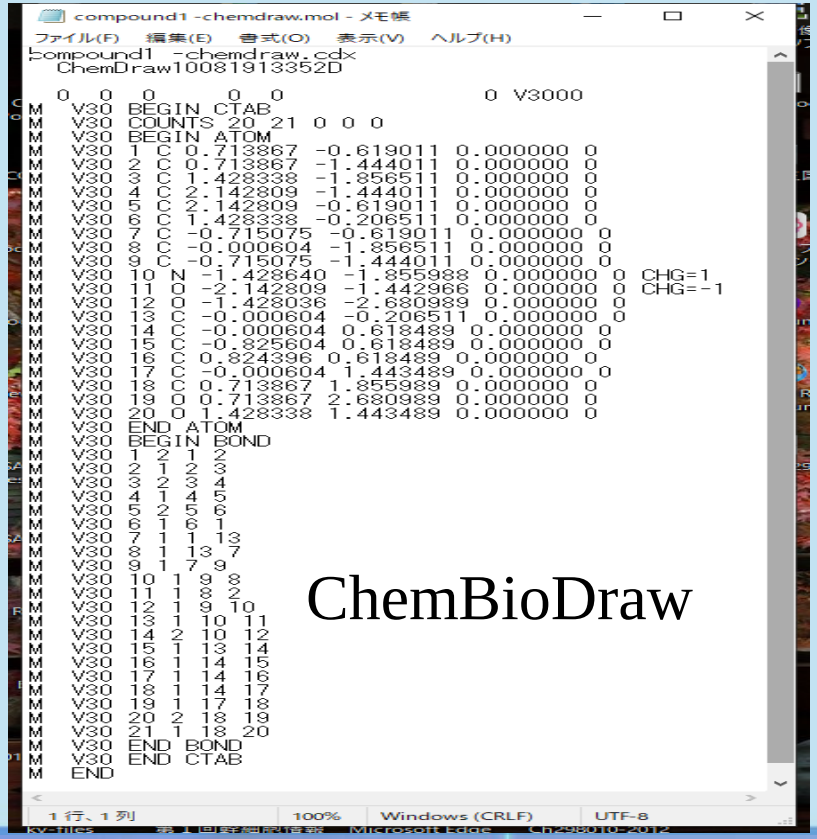
Even in the same file format, attention must be paid to compatibility

・システム依存性、 ・バージョンの変化、記述桁数の違い、他

System dependency, Version change, difference in number of description digits, etc.



ACD/ChemSketch



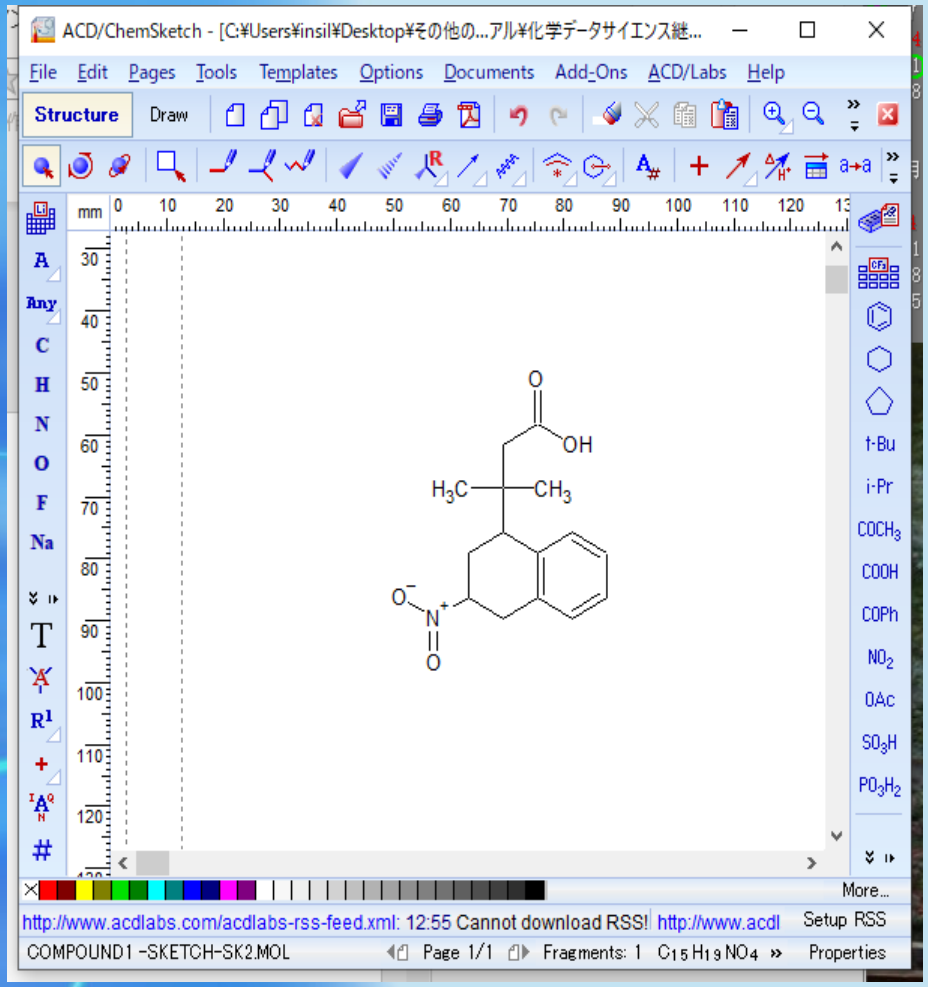
ChemBioDraw

口二次元化合物構造式の作画（同一のソフト）

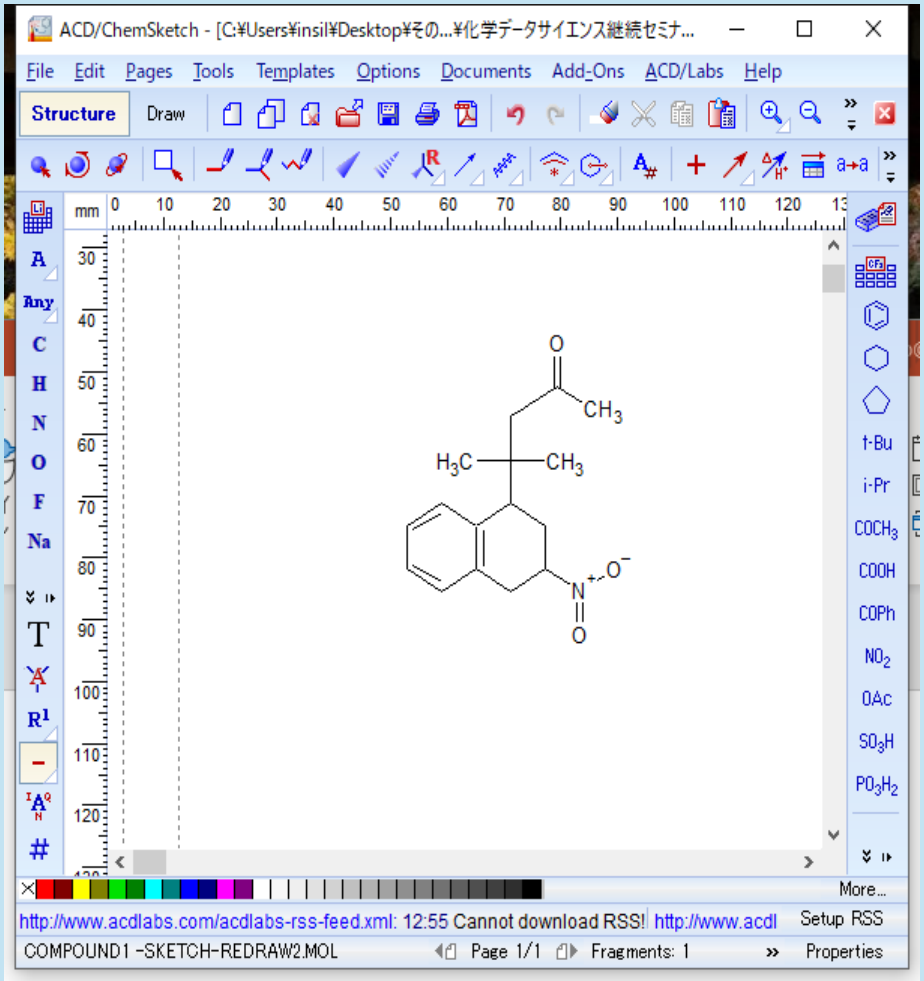
In Silico Data
Miracles by the KY-methods

Drawing of two-dimensional compound structural formula (same software)

ACD/ChemSketch



ACD/ChemSketch



□二次元化合物構造式の同一Molファイル

Same Mol file of 2D compound structural formula

◆原子の番号付けが異なる Different atom numbering

・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**

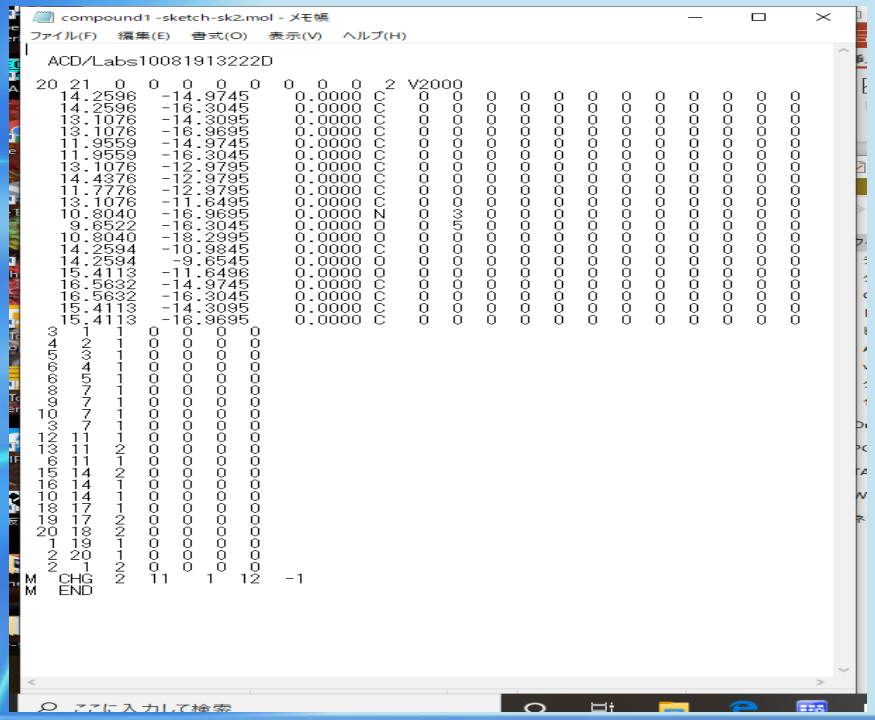
Even the same compound created by the same system has different atomic numbers

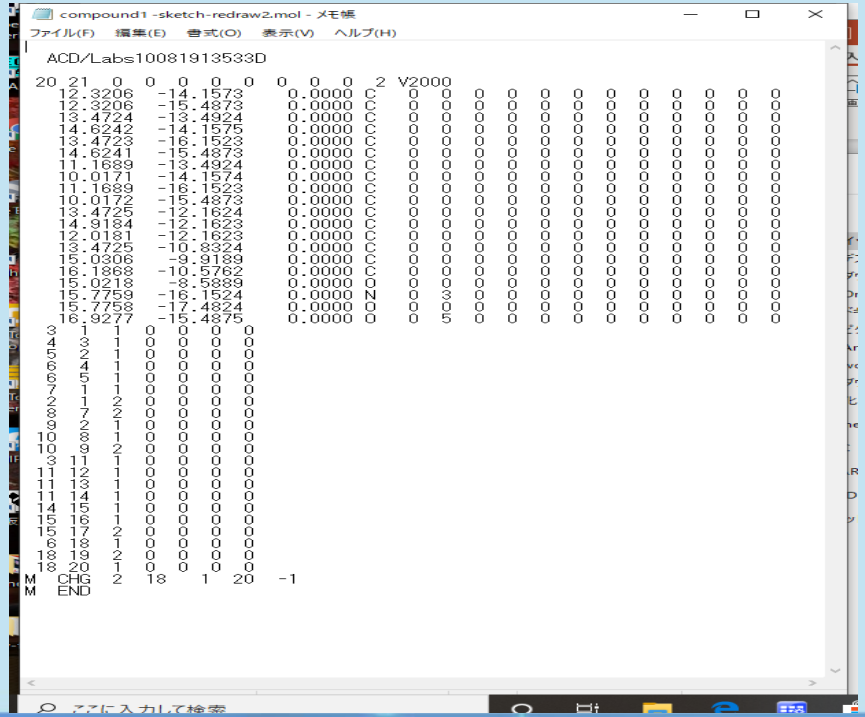
・原子番号の一元化が必要（Morganアルゴリズム、SEMA等）

Need to unify atomic numbers (Morgan algorithm, SEMA, etc.)

◆化合物の種々検索での問題解決が必要

Need to solve problems in various compound searches





口化合物原子の番号付け　Compound atom numbering

■完全一致検索　Exact search

1．変換コードを用いたアプローチ　Conversion code approach

①MORGAN 名：立体情報を持たない化合物　Compound without 3D information

＊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 name is a compound name assigned to a compound in a one-to-one term

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：Unique numbering

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

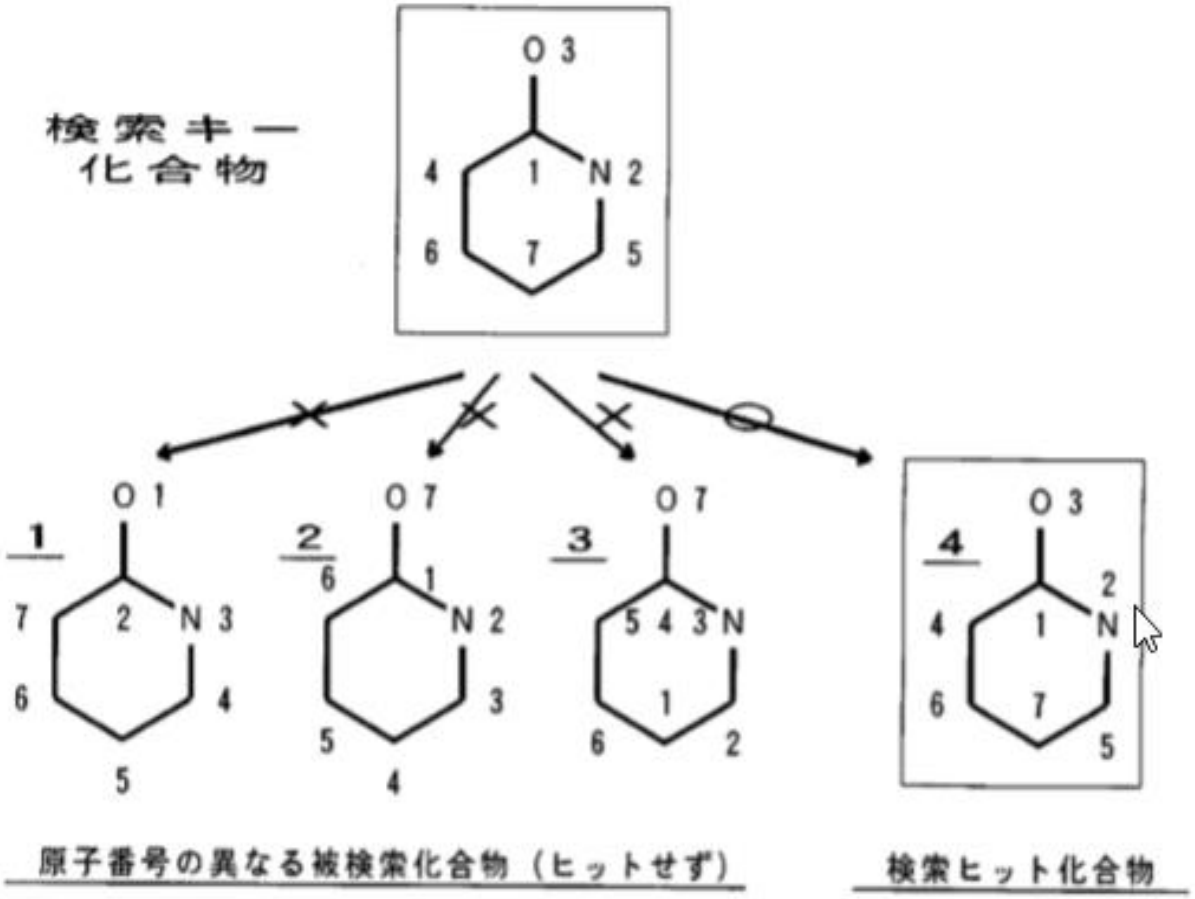
Decide the number assigned to each atom that constitutes a compound for each compound.

口化合物原子の番号付け　Compound atom numbering

■MORGAN名の実施形態：MORGAN name embodiment

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。If the atomic number assigned to a compound is different, the same compound is recognized as a different compound inside the computer as shown in the right figure.

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 Therefore, the number assigned to the atom of the compound must be unique (corresponding to one unit and one term)



口一元一項対応の重要性
Importance of one-on-one correspondence

In Silico Data
Miracles by the KY-methods

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
When handling compound information on a computer

◆一元一項対応（Canonicalization:規範化）とは
What is one-to-one correspondence (Canonicalization)

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
One compound is defined as 1: 1 if it has the same description system

◆一元多項対応とは What is one-way multiplet correspondence

一つの化合物が複数の記述（1：多）で定義される
One compound is defined by multiple descriptions (1: many)

□化学分野の基本情報　Basic information in the chemical field

■一元一項対応関連の考え　Thoughts related to one-on-one correspondence

一元一項対応　⇒　化合物（1）　⇒　化合物名（1）
One-on-one correspondence ⇒ Compound (1) ⇒ Compound name (1)

多元一項対応　⇒　化合物（N）　⇒　化合物名（1）
Plural terms ⇒ Compound (N) ⇒ Compound name (1)

一元多項対応　⇒　化合物（1）　⇒　化合物名（N）
One-way multiple terms ⇒ Compound (1) ⇒ Compound name (N)

多元多項対応　⇒　化合物（N）　⇒　化合物名（N）
Multifactorial correspondence ⇒ Compound (N) ⇒ Compound name (N)

但し、化合物名＝化合物表記　However, compound name = compound notation
（　）内は数を示す　（　）Indicates number

□一元一項対応の重要性
Importance of one-on-one correspondence

◆一元多項対応の時に発生する問題点
Problems that occur when dealing with one-way multinomials

1. 化合物データベース関連上での問題
Problems related to compound databases

①多重登録が頻発する　Multiple registrations occur frequently

②化合物検索の精度が保たれない
Compound search accuracy is not maintained

③複数の化合物データベース間の連携が困難
Collaboration between multiple compound databases is difficult

2. 化合物データ解析関連上での問題
Problems related to compound data analysis

①重複データの存在可能性　Possibility of duplicate data

②パラメーターの不安定性　Parameter instability

③予測モデルの汎用性減少　Reduced versatility of prediction model

□化学分野の基本情報
Basic information in the chemical field

■なぜ一元一項対応が重要となるのか
Why the one-on-one correspondence is important

	実在の化合物 Real compounds	化合物表記 Compound notation	
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 ○**
One-on-one correspondence	⇒ Compound (1)	⇒ Compound name (1)	
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 ×**
Plural terms	⇒ Compound (N)	⇒ Compound name (1)	
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 ×**
One-way multiple terms	⇒ Compound (1)	⇒ Compound name (N)	
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 ×**
Multifactorial correspondence	⇒ Compound (N)	⇒ Compound name (N)	

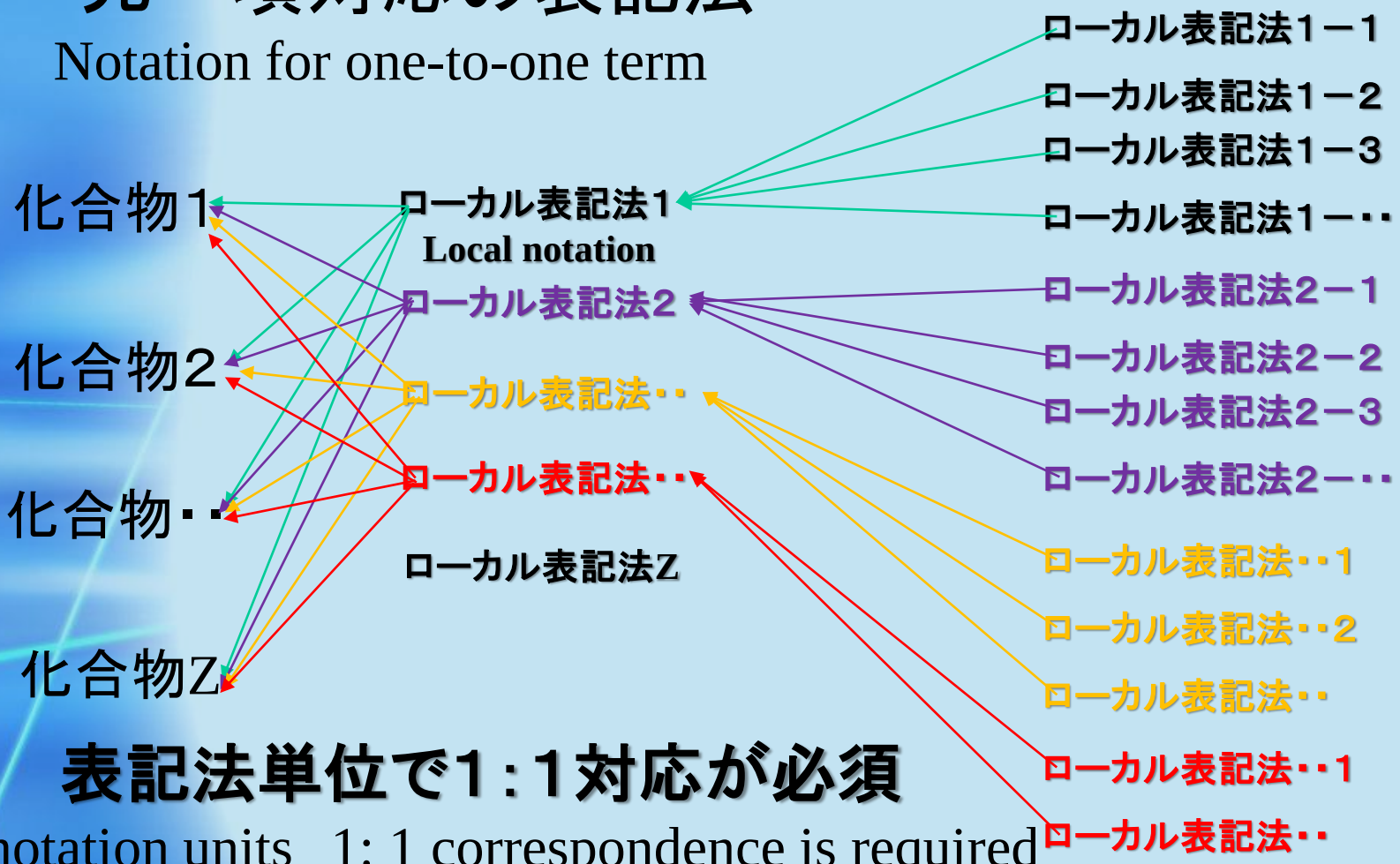
◇化合物に関する検索や化合物データ解析を想定すると**一元一項対応**が必須
About compounds Search and compound data analysis Assuming One-on-one correspondence is required

口一元一項対応の概念図

Conceptual diagram corresponding to one item

一元一項対応の表記法

Notation for one-to-one term



表記法単位で1：1対応が必須

In notation units　1: 1 correspondence is required

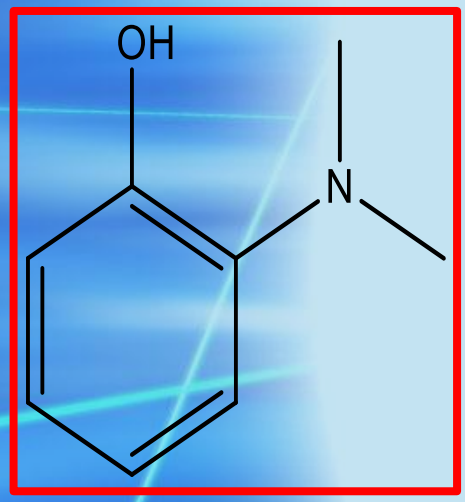
口一元一項対応の重要性（Smilesを例に）

Importance of dealing with one item and one term (Smiles as an example)

化合物
Compound

表記法およびローカル表記
Notation and local notation

重複登録 Duplicate registration
検索ヒットせず No search hit

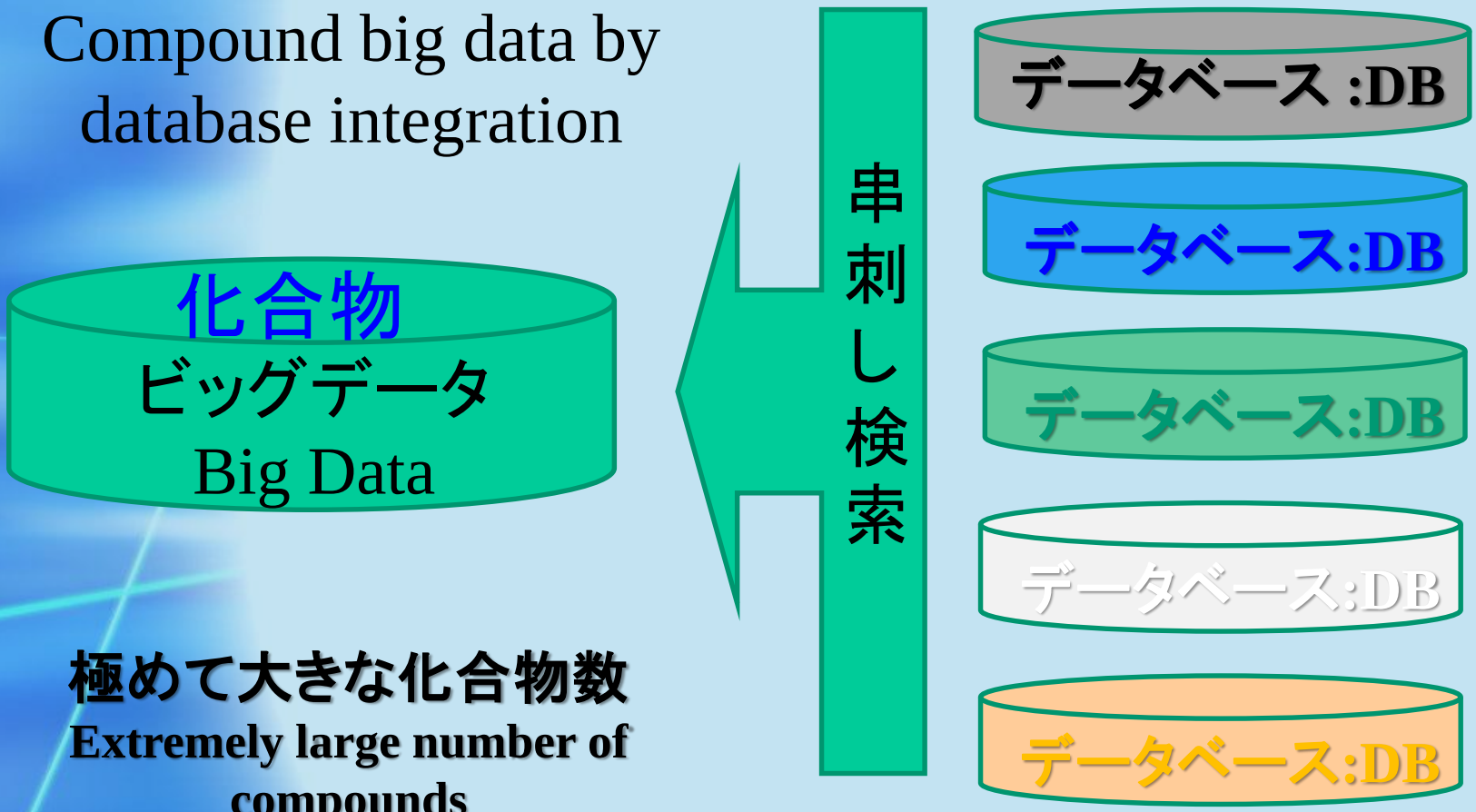


SMILES

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
Big data by database linkage and integration

データベース統合による
化合物ビッグデータ化
Compound big data by
database integration

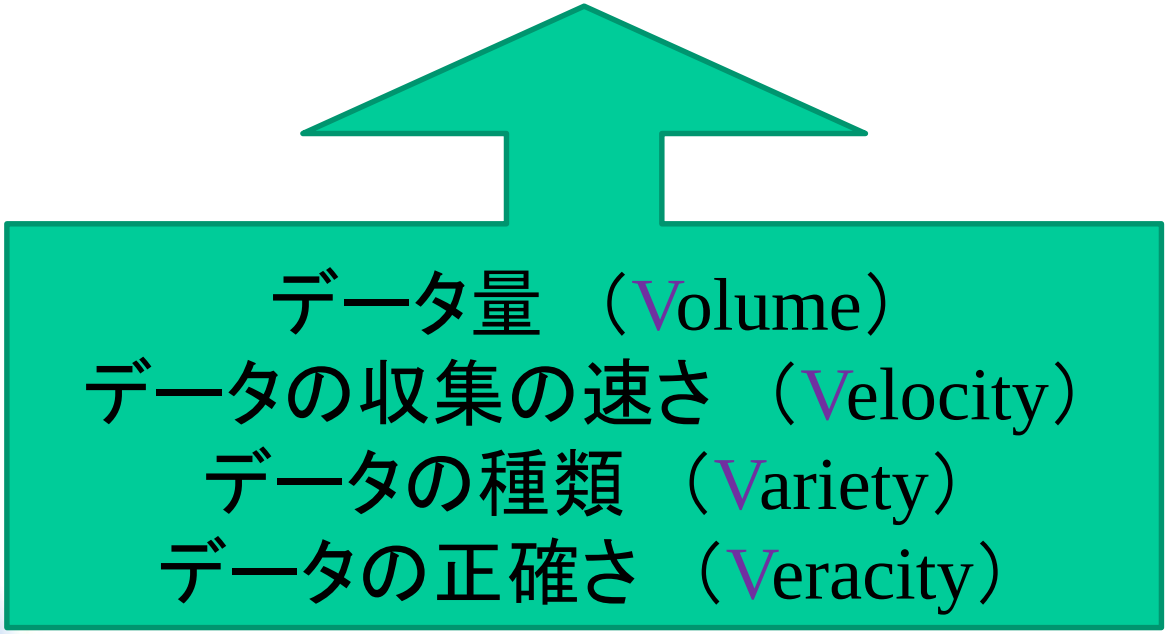


極めて大きな化合物数
Extremely large number of compounds

口化合物関連データベース(ビッグデータ)
Compound-related database (big data)

◇ビッグデータの5Vとは:Big data 5V

データベース構築では達成の最適化目標として、5種類のVの実現が理想とされる
In the database construction, the realization of the five types of V shown on the right is ideal as an optimization goal for achievement.



□ビッグデータ構築上での留意点
Points to remember when building big data

化学データ解析
Chemical data analysis

◇ビッグデータの5V
Big data 5V

極めて大きな化合物数
Extremely large number of compounds

AI(人工知能）適用
Application of AI (artificial intelligence)

□化学関連システム間連携上での留意点
Points to keep in mind when linking chemical-related systems