P04-051

Proposal of next-generation system in big data era based on chemical data science

--- Integrated toxicity research support system adapted to the new era ---

Kohtaro Yuta

In Silico Data, Ltd. (http://www.insilicodata.com)

Introduction: In recent years, information-related environments have changed greatly, and various new technologies such as big data and data science are rapidly developing. In this poster, we will consider the development of future-type systems that can cope with such major changes in IT-related technologies. In particular, we will discuss the problems that occur when building systems that handle compounds.

Discussion: The core technology of the future system is big data. The difficulty of collecting a large number of samples in the field dealing with compounds is much more difficult than in other fields. For this reason, many systems with a small number of chemical compound samples are constructed, and it is extremely difficult to construct big data which is the core of future systems. This problem is considered from the viewpoint of chemical and compound information.

System configuration and technical contents of Big Data & Data Science era

Big Data
Several hundreds of thousand of data
Several millions to billions of data

→ **Data Science**
Chemometrics

→ **Artificial Intelligence**
Deep learning, CNN, RNN

→ **Various Application Fields**

Current system configuration

General features in compound-related systems:
1. Extremely difficult to collect samples
2. Compound-related notation methods are not unified
(compound notation and operation differ for each system)
3. There are many toxic items, and the protocol is almost uniform

☐**Independent system construction (single-function system)**

Compounds Database 1 — System 1
Compounds Database 2 — System 2
Compounds Database ・ — System ・
Compounds Database N — System Z

Compound safety system
1. COSMOS
2. electronic Toxicity: eTox
3. QSARToolbox
4. RepDose
5. HESS
6. KATE
2. ECOTOX
3. eChem Portal
4. ACToR
6. SIDER (Side Effect Resource)
7. ISSCAN (Chemical carcinogens)
8. Many others

＊Each systems are constructed according to individual research objects
・System linkage is quite difficult ・Small scale due to independent system

Future system configuration

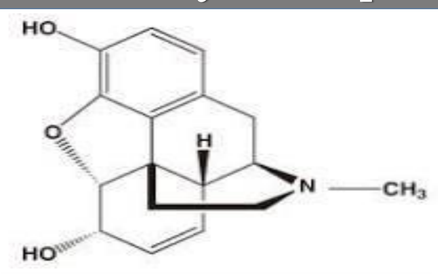
☐**Integrate individual systems into big data**

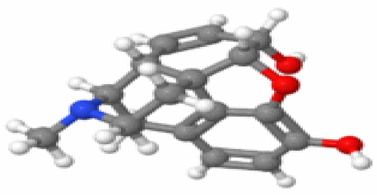
Future Compounds Database 1 — New System 1
Future Compounds Database 2 — New System 2
Future Compounds Database ・ — New System ・
Future Compounds Database N — New System Z

→ **Virtual Big Data** → **Data Science** / **Artificial Intelligence**

Problems related to compound manipulation / storage

◇ Diversity of compound notation: No unified information





■**Chemical ID Number**
CAS number:57-27-2
ATC code:N02AA01 (WHO)
PubChem:CID: 5288826
DrugBank:APRD00215
ChemSpider:4450907
KEGG:D08233

■**compound properties**
Chemical formula:C17H19NO3

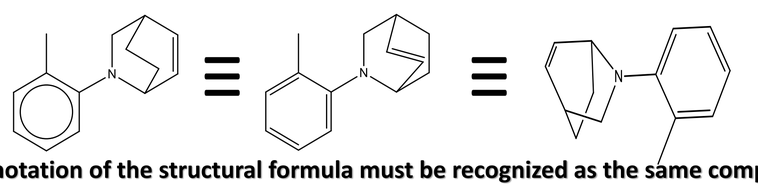
■**Reproducibility of chemical compounds: Linear notation of compounds**
Compound name：Morphine
IUPAC: (5α,6α)-7,8-didehydro-4,5-epoxy-17-methylmorphinan-3,6-diol
SMILES: OC(C=CC1CC2N3C)=C(OC4C(O)C=5)C1C4(CC3)C2C5
InChIKey: InChI=1S/C17H19NO3/c1-18-7-6-17-10-3-5-13(20)16(17)21-15-12(19)4-2-9(14(15)17)8-11(10)18/h2-5,10-11,13,16,19-20H,6-8H2,1H3/t10-,11+,13-,16-,17-/m0/s1

■**Reproducibility of chemical compounds: Notation by connection table**

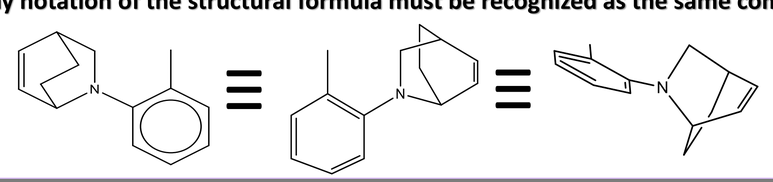
List of file formats handled by the “OpenBabel system” →

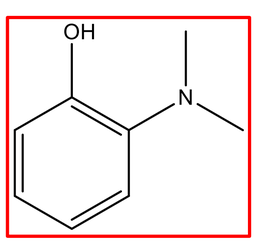
```
mol -- MDL MOL format
pdb -- Protein Data Bank format
smi -- SMILES format
xyz -- XYZ cartesian coordinates format
CONFIG -- DL-POLY CONFIG
CONTCAR -- VASP format
HISTORY -- DL-POLY HISTORY
POSCAR -- VASP format
VASP -- VASP format
abinit -- ABINIT Output Format
acesin -- ACES input format
acesout -- 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
bgf -- 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
caccrt -- Cacao Cartesian format
cache -- CAChe MolStruct format
cacint -- Cacao Internal format
can -- Canonical SMILES format
```

◇Necessity of canonicalization of compounds: Response to compound diversity



Any notation of the structural formula must be recognized as the same compound



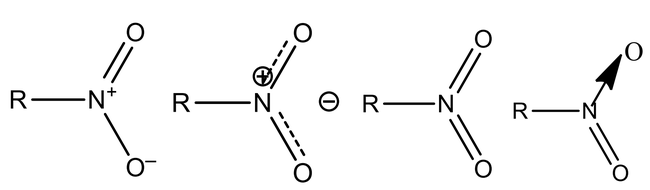


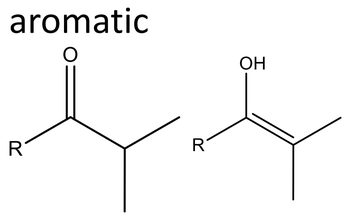
☐**Canonicalization is required to correctly perform compound searches**
There are many structural patterns in one compound. Compound does not hit in search.

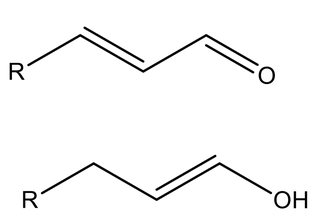
SMILES 1: OC1=C(N(C)C)C=CC=C1 ;by **ChemDraw**
2:c1(O)c(N(C)C)cccc ;by **Ecosar**
3:C1=CC(=C(C=C1)N(C)C)O ;by **QSAR Toolbox**
4:CN(C)c1ccccc1O ;by **OpenBabel**
5:C1=CC(O)=C(N(C)C)C=C1 ;Manual Input by Yuta
6:C1(O)=C(N(C)C)C=CC=C1 ;Manual Input by Yuta

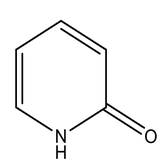
◇Compound-specific problems in the compound structure: Response to compound diversity is required

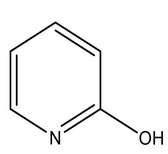
◇ Problem in compound structure: tautomer, nitro, aromatic











Many others