Open for Innovation®
KNIME

Open file formats for chemical information

IUPAC InChI Workshop
August 2017

Greg Landrum, Ph.D.
KNIME.com AG

This presentation in two bullets

- **What:** Define open standards for an important subset of widely used formats used for interchanging small molecule chemical information.

- **Why:** Make sure that we can correctly interpret the information (metadata) about chemical compounds in our repositories (open or otherwise)

Motivation

- Go to ChEMBL, PubChem, Reaxys (or any other large data source), find some data you're interested in, and download the data.
- What if you want to know about the chemical structures that the data was generated for?

Motivation

- Go to ChEMBL, PubChem, Reaxys (or any other large data source), find some data you're interested in, and download the data.
- What if you want to know about the chemical structures that the data was generated for?
- Small molecule structures typically come as SMILES or Mol blocks.

Motivation

- Go to ChEMBL, PubChem, Reaxys (or any other large data source), find some data you're interested in, and download the data.
- What if you want to know about the chemical structures that the data was generated for?
- Small molecule structures typically come as SMILES or Mol blocks.
- Question: What do these actually mean?
- Note: this is **not** Roger's cutting edge

Why?

- Where are our most common small molecule file/interchange formats actually defined? How do we know what they mean?
  - SMILES:
    http://www.daylight.com/dayhtml/doc/theory/theory.smiles.html
    http://www.opensmiles.org/opensmiles.htm
    https://github.com/opensmiles
  - SMARTS:
    http://www.daylight.com/dayhtml/doc/theory/theory.smarts.html
    https://github.com/timvdm/OpenSMARTS (very preliminary)
  - CTAB/MOL/SDF:
    ctfile.pdf (somewhat publicly available)
    Various MDL/Symyx/Accelrys/Biovia manuals (not publicly available)

Why?

- Many/all of the formats have been around a fairly long time and have been “embraced and extended” by various groups, which (if any) of these should be standard?

What?

- Define open standards for an important subset of widely used formats used for interchanging small molecule chemical information.
- Start with the basics:
  - SMILES
  - SMARTS
  - CTAB/Mol/SDF
- After these are done we can move on to other things like SMIRKS, RXN, etc.

But we have InChI! Why bother?

- InChI is an identifier. As part of the canonicalization process it standardizes the input structure

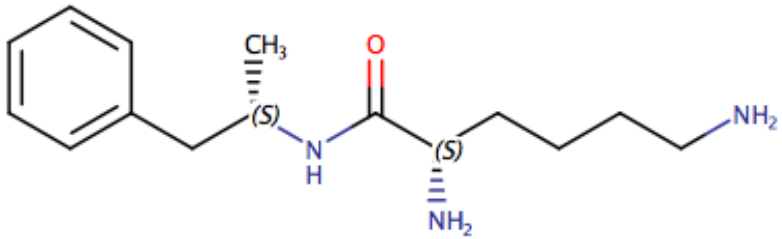
- Often you want to keep track of what the input structure actually was.[1] You need another format for that.

- It'd be great if that format was also well defined

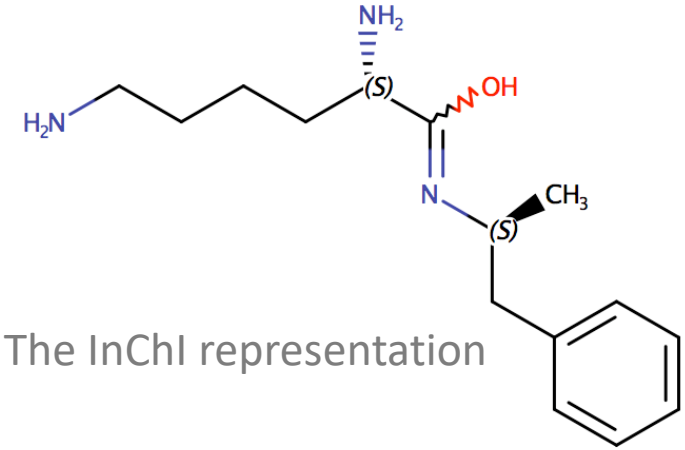
[1] This is a good thing to do in most every case

A problem of tautomers

Vyvanse
(Lisdexamfetamine)



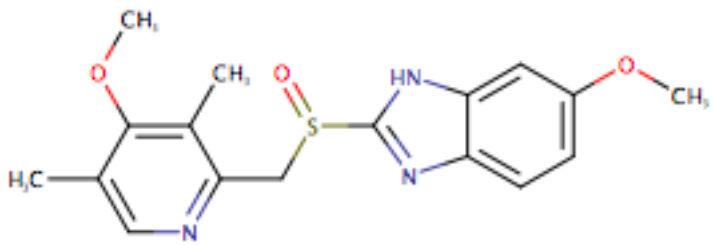
How the chemist (likely) thinks of it

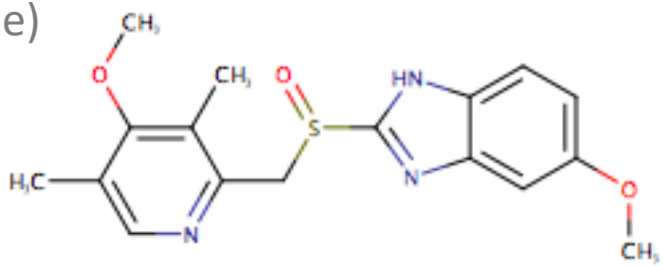


The InChI representation

A problem of tautomers

Prilosec
(Omeprazole)



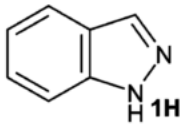
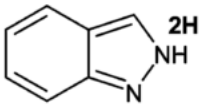
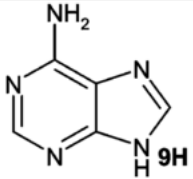
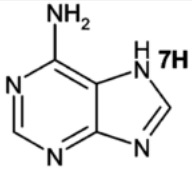
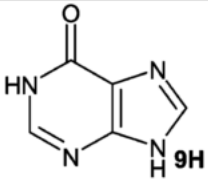
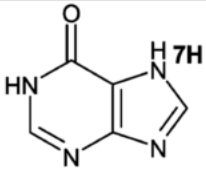


- “Tautomeric polymorphism”: Both tautomers are observed in crystal structures

A problem of tautomers

- Prilosec is **not** an isolated example. Many, many examples of different tautomers observed in crystal structures and protein–ligand complexes

a)

Name	ΔG (kcal mol^{-1}) / medium / relevant form(s)	Major form in Water (M)			Minor form in Water (m)			PDB with identical HB score
		Structure/Name	CSD	PDB	Structure/Name	CSD	PDB	
1. Indazole	2.3/water/M; ND/gas (calc.)/M.		45	26		0	0	7
2. Adenine	0.8/water/M; 12/gas (calc.)/M; 0.8/DMSO/M.		17	10		3	23	6
3. Hypoxanthine	0/water/M and m; ND/gas (calc.)/M; ND/solid/M.		2	27		2	21	14

Milletti, F. & Vulpetti, A. Tautomer preference in PDB complexes and its impact on structure-based drug discovery. *J. Chem. Inf. Model.* **50,** 1062–1074 (2010).

What?

- Define open standards for an important subset of widely used formats used for interchanging small molecule chemical information.
- Start with the basics:
  - SMILES
  - SMARTS
  - CTAB/Mol/SDF
- After these are done we can move on to other things like SMIRKS, RXN, etc.

An Open Standard?

- Documents are freely accessible and redistributable
  - Need some kind of explicit license. A suggestion would be to use something well known like Creative Commons
- Clear, publicly visible process for suggesting and discussing improvements and corrections
- Clear stewardship

https://docs.google.com/document/d/1DgDYE1Jt-VxtqevSQcBfJSmCy57irdrUtBi3oIhkzD4/edit?usp=sharing

How?

- Review the status of the existing documentation and identify the gaps.
- Review the various extensions to/dialects in use and decide on which, if any, of these will be incorporated in v1 of the new reference document.
- Develop and publish the new reference and tutorial documents along with a good collection of examples
- *Document and recommend a portable subset of the format that can be relied upon in legacy/existing files*
- Seek adoption by tool and toolkit producers as well as data providers. Note: this needs to happen with a core subset very early in the process so that support for the new format is available in at least some tools/toolkits/data sources when the documentation is published.
- *Develop and deploy an open syntax checker and depiction service and webpage*

Italics: optional but very useful

How? (the main points)

- Review the status of the existing documentation and identify the gaps, if any.
- Develop and publish the new reference and tutorial documents along with a good collection of examples.
- Seek adoption by tool and toolkit producers as well as data providers.

The problem

- The person supposedly organizing this seems highly motivated, but hasn't managed to do any significant work on this effort aside from talking about it.

- If this is going to go anywhere, it needs a more effective ~~cat herder~~ steward

Want to help?

- Volunteer to help directly
- Point us to important pathologies
- Point us to important extensions for SMILES / SMARTS / CTAB
- Follow progress and make suggestions

Thanks!

This presentation in two bullets

- **What:** Define open standards for an important subset of widely used formats used for interchanging small molecule chemical information.

- **Why:** Make sure that we can correctly interpret the information (metadata) about chemical compounds in our repositories (open or otherwise)

Backups

ToDos

- A place to work: https://github.com/OpenChemistryFileFormats
- Identify participants
  - Core: actually doing work, responsible for decisions and delivery
  - Consulted: providing input and comments
- Concrete project plan, with target dates for at least the next deliverables
- Get started!