WorldFAIR Chemistry Doc-a-thon: Chemical representation best practices for humans AND machines

Evan Bolton, Ph.D.
evan.bolton@nih.gov

ACS National Meeting, 27 March 2023

National Library of Medicine
National Center for Biotechnology Information

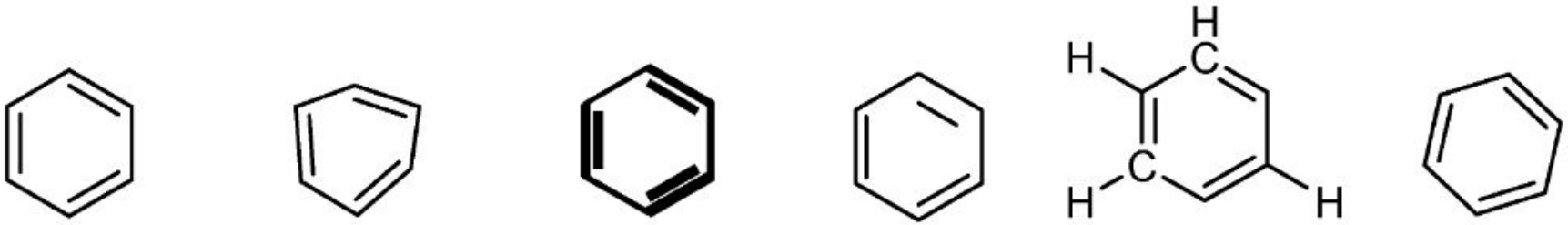
https://doi.org/10.1351/pac200880020277

GR-0. INTRODUCTION

Although chemical structures have been called "the language of chemistry" [1], few documents have attempted to provide any sort of guidelines for the production of chemical structure diagrams [2–5]. The

Graphical representation of chemical structures by humans, for humans. Best Practices 101, circa 2005

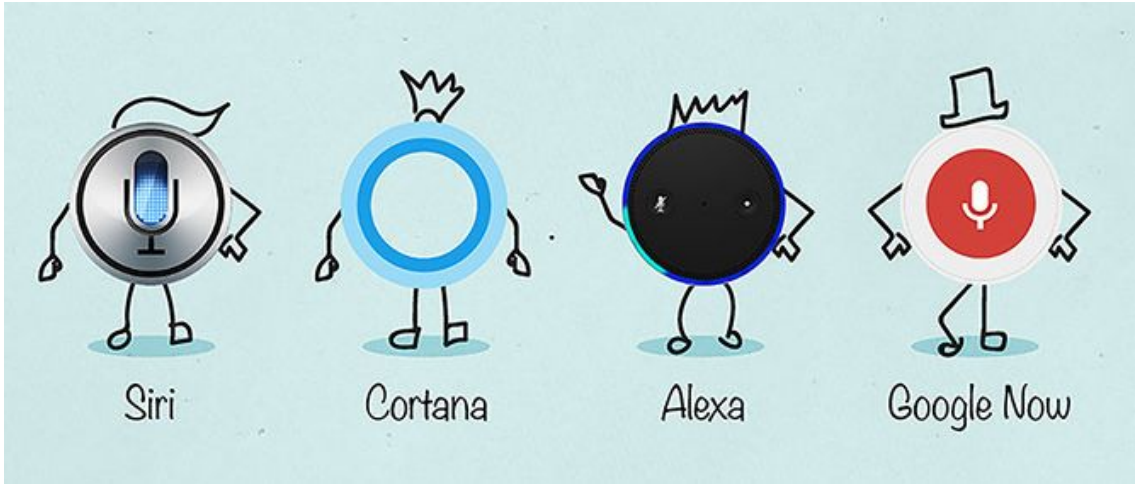
[8]. And yet chemists have strong feelings for how chemical structure diagrams should look, even in the absence of formal guidelines. Show most chemists a series of diagrams of something as simple as benzene, and there will be near-unanimity about which ones are "good" diagrams and which ones are "bad". As Robert Pirsig writes in *Zen and the Art of Motorcycle Maintenance*, "But even though Quality cannot be defined, *you know what Quality is*!" [9].



We live in a rapidly changing world

Digital assistants

Siri, Cortana, Alexa, …



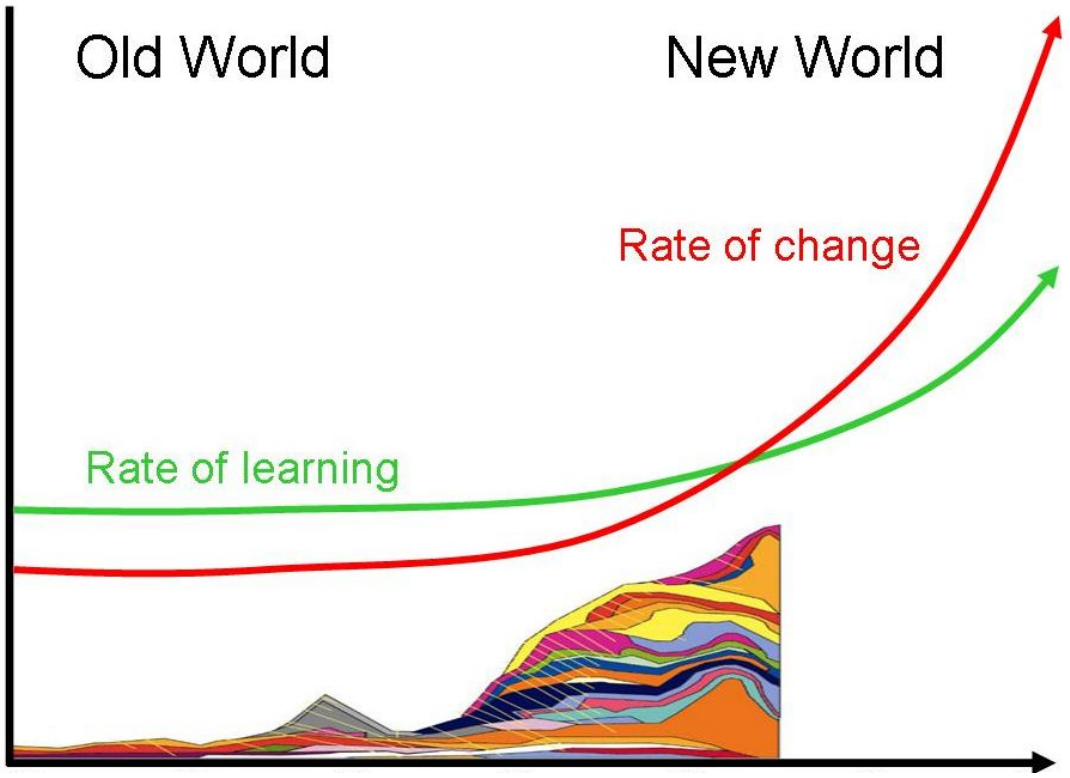
We live in a rapidly changing world

Digital assistants

Siri, Cortana, Alexa, …

Machine learning, deep learning,

data science, …



We live in a rapidly changing world

Digital assistants

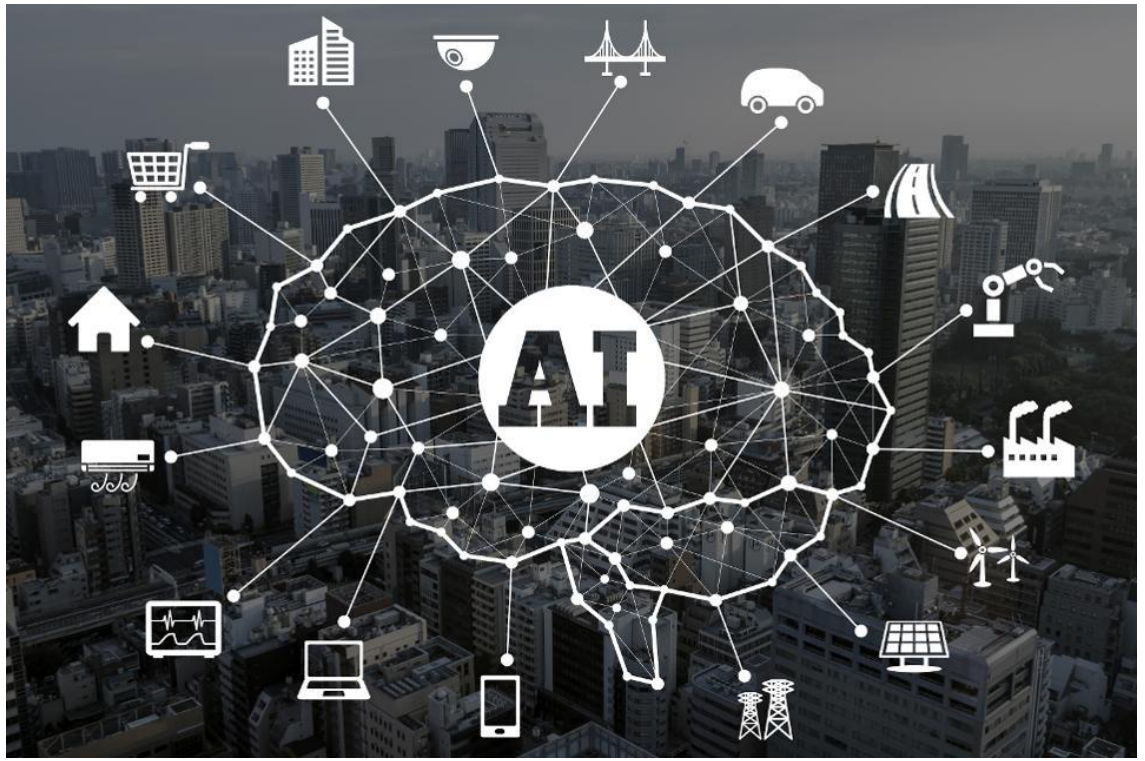
Siri, Cortana, Alexa, …

Machine learning, deep learning,

data science, …

Dawning of the age of

Artificial Intelligence (AI)



We live in a rapidly changing world

Do Large Language Models Understand Chemistry? A Conversation with ChatGPT

Cayque Monteiro Castro Nascimento and André Silva Pimentel*

Abstract

Large language models (LLMs) have promised a revolution in answering complex questions using the ChatGPT model. Its application in chemistry is still in its infancy. This viewpoint addresses the question of how well ChatGPT understands chemistry by posing five simple tasks in different subareas of chemistry.

J. Chem. Inf. Model. 2023, 63, 6, 1649–1655
Publication Date:March 16, 2023
https://doi.org/10.1021/acs.jcim.3c00285

understanding

The chemical information divide

Human understanding

Depictions
Schemes
Table
Text

Computer understanding

Explicit
Complete
Annotated
Interpreted

Image credit:
https://hcldr.files.wordpress.com/2016/12/1-rr-0116-industry-divided.jpg

This is what it means to be useful for many use-cases

Imagine you wanted to make a modern scientific resource, what do you need to focus on?

2020s Cloud-first, Mobile-first, Machine-first, FAIR-first, Open-first

Use of persistent research identifiers

Data use cases explicitly considered

Standards-based approaches

Explicit data licensing (e.g., CC-BY 4.0, CC0)

Receive cloud-based data

Make data accessible within cloud

UI/UX from a device screen size agnostic perspective

Use of controlled vocabulary and machine interpretable statements

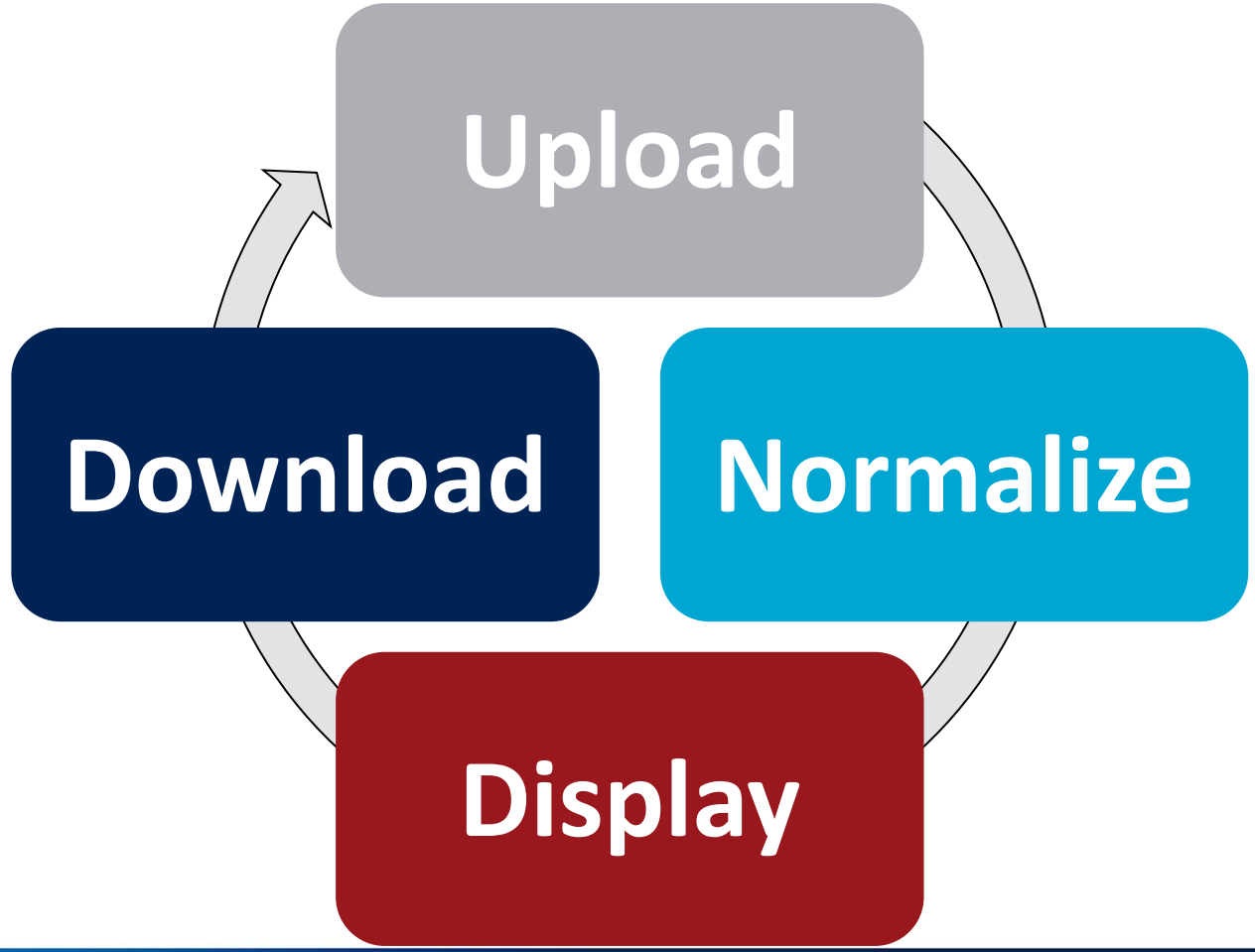
Sufficient meta data to reproduce the science

FAIR means "**F**ully **AI** **R**eady" .. also means "**F**indable", "**A**ccessible", "**I**nteroperable", "**R**eusable"

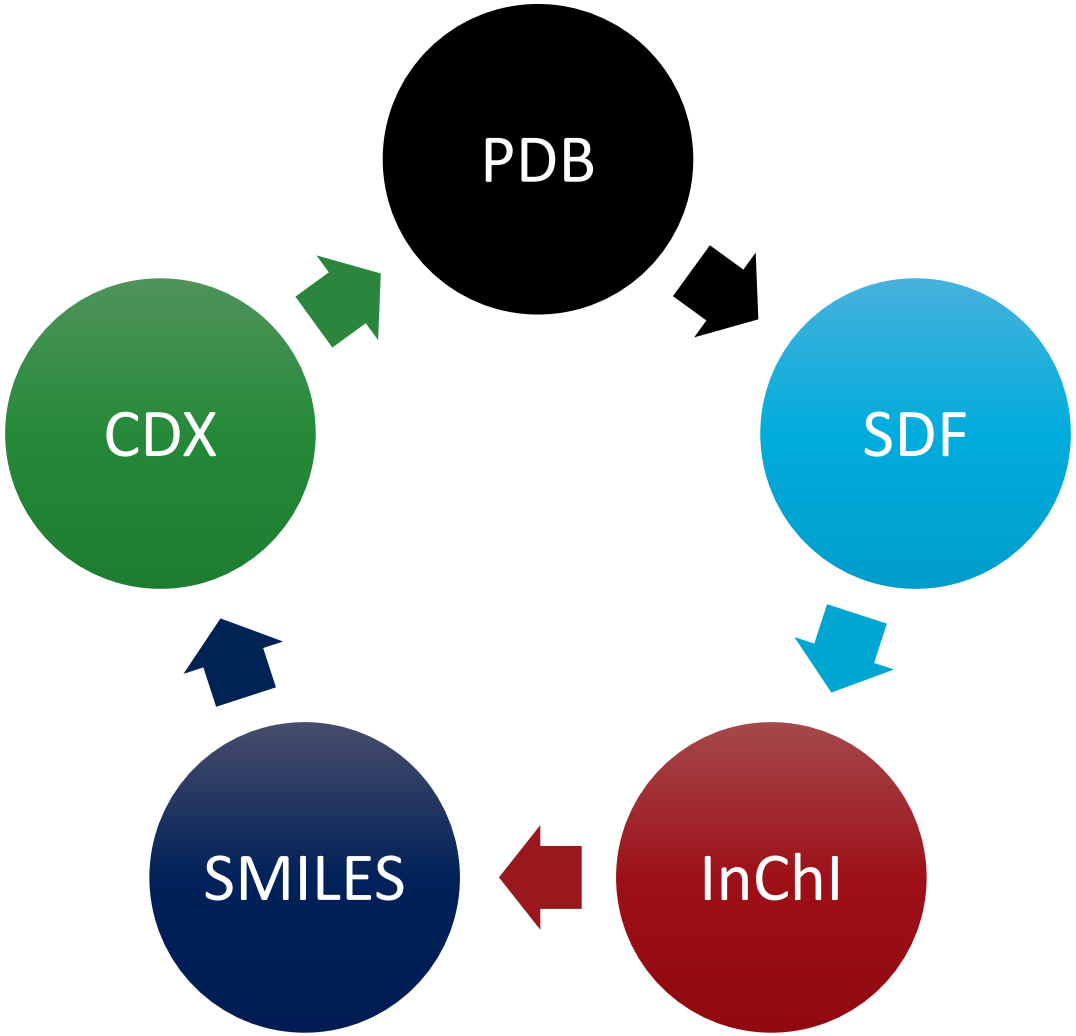
Variability in interpretation by software can lead to confusion

- Algorithms normalize
- Sometimes they go wrong
- Can introduce ambiguity
- Later processing results in error

- Cycle

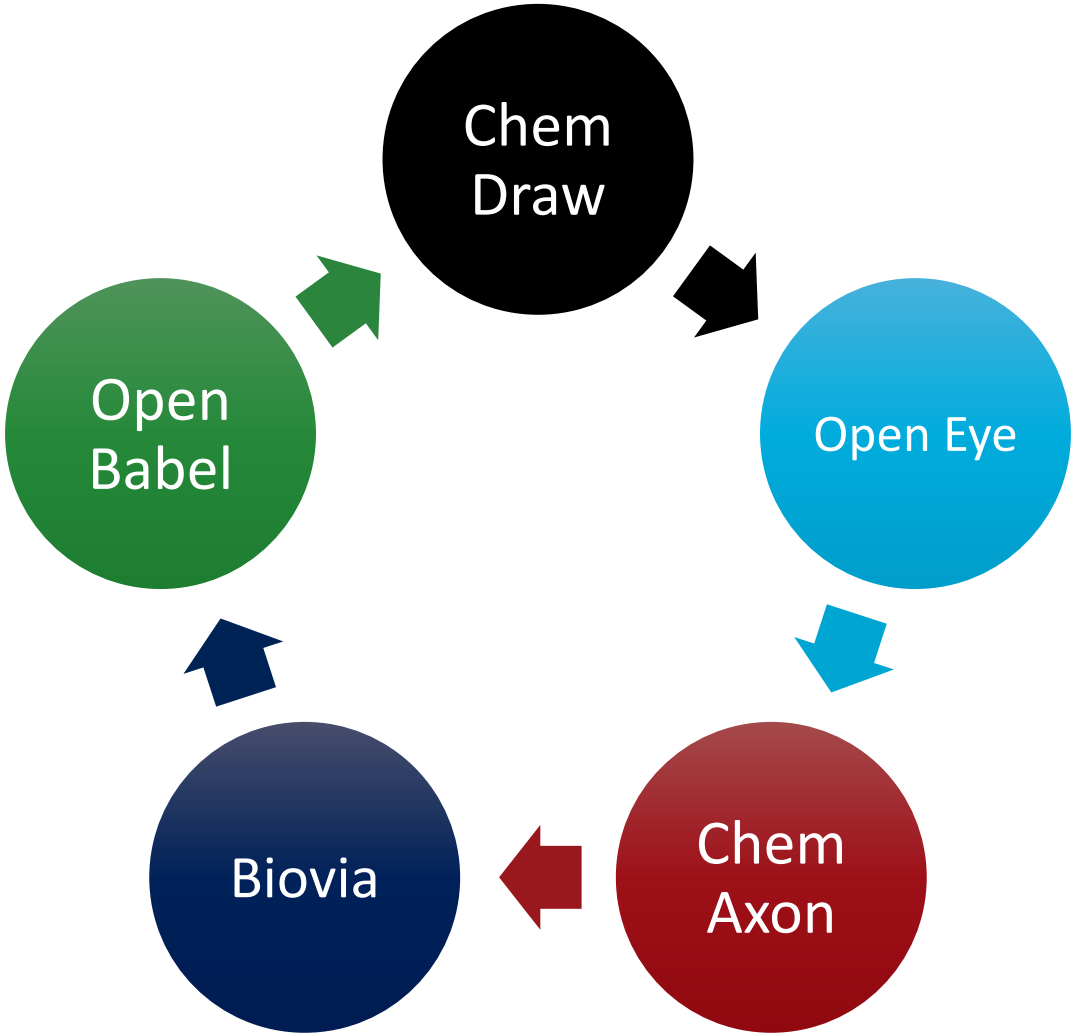


Where has this structure been?

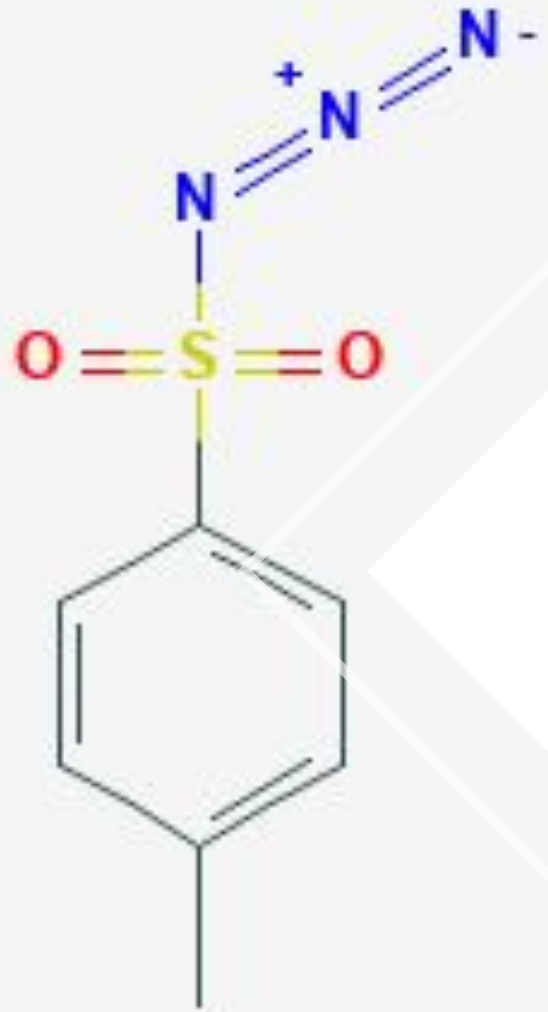


Chemical file format interconversion can be lossy

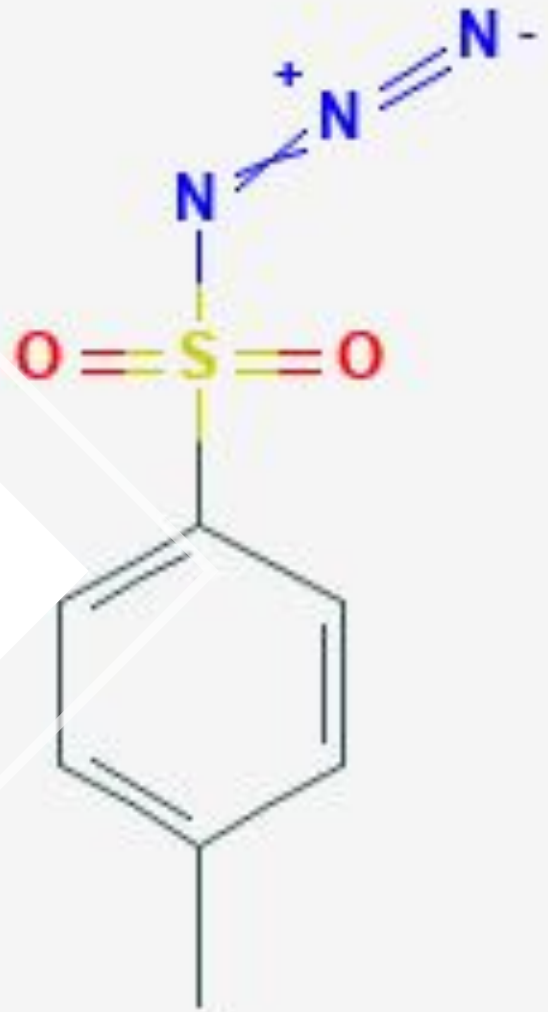
With whom has this structure been?

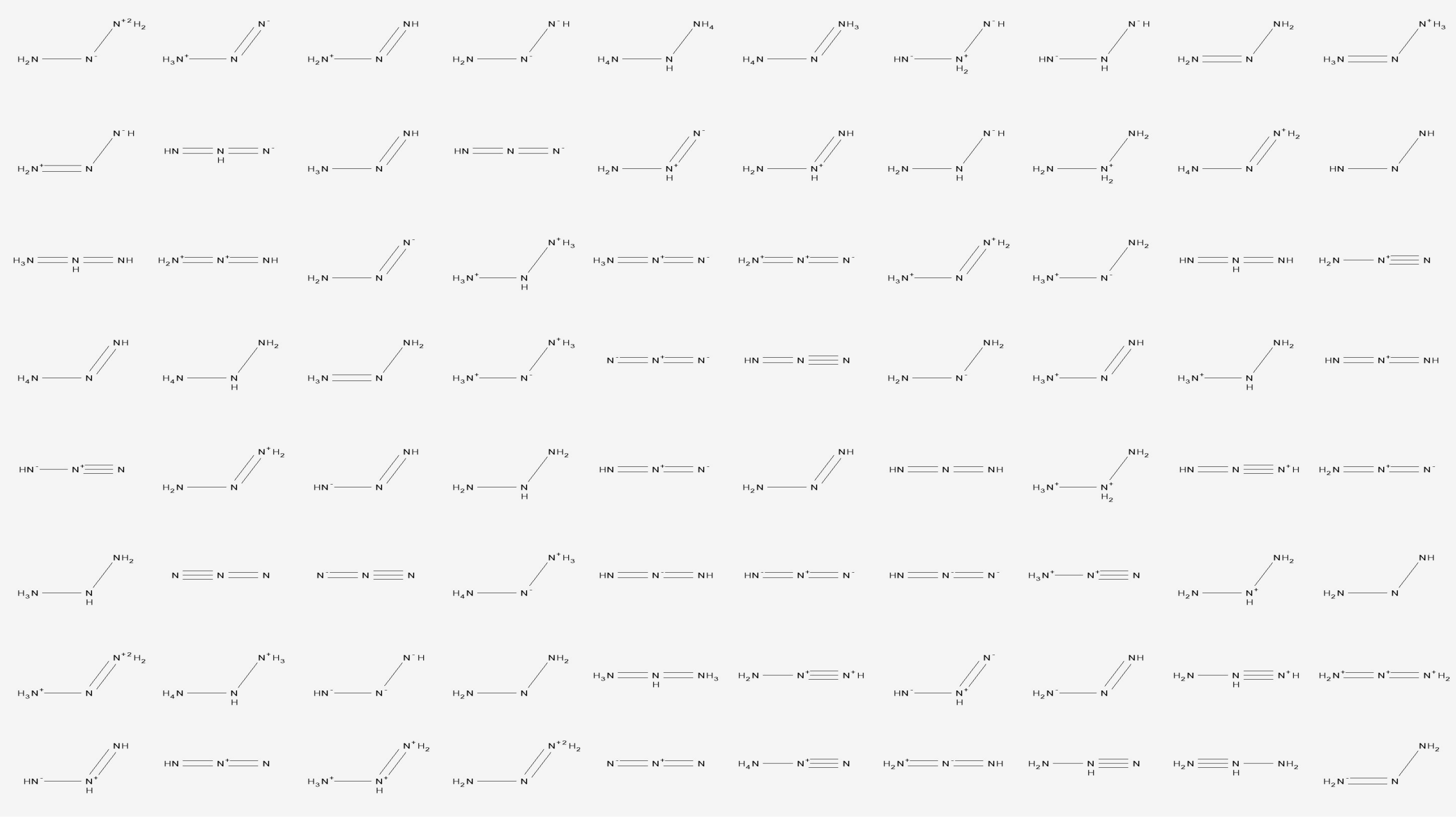


Different software packages may ‘normalize’ your chemical structure in different ways

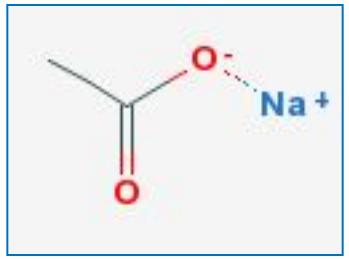


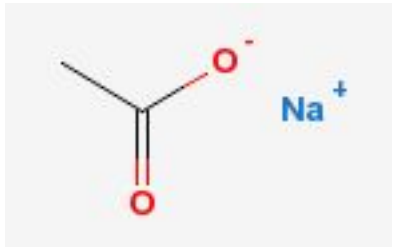
Examples of the need for standards and best practices abound with chemical structure representation

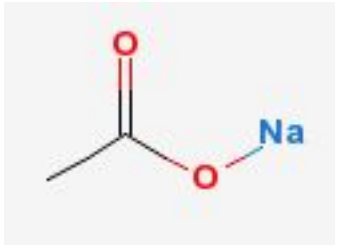


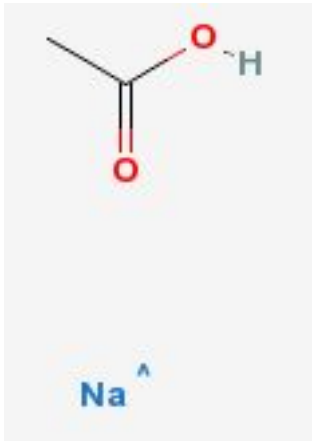


A chemical structure may be represented in many different ways





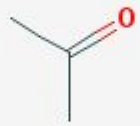




Salt-form drawing variations are common

Chemical information is not designed for computers

- As a chemist, you can understand and recognize that this picture is the chemical acetone
- You can put a chemical name or registry identifier next to it
- Is this not good enough?



Acetone
67-64-1

- Many names for structure
- The computer ‘sees’ a binary image not a structure

Almost all chemical information is geared towards human understanding in the form of text and images

Image credit:
https://i.pinimg.com/originals/51/91/68/5191682657c2a8650e68ad906b062e51.jpg

Specifically, what do we (chemists) need?

- Incentives for researchers to share their data
  - Funders/societies/publishers requirement
- Make data sharing easy and machine ready
  - LIMS/ELN software that can export data to be shared direct to repositories or for publication supplementary materials
- Improved interoperability for machine understanding
  - ***Standards and best practices for sharing chemical information (structures)***
  - Standards for structure formats
  - Appropriate terms and terminology

NCBI