Describing chemical substances and chemical structures: we have a long way to go

Evan Bolton, Ph.D.

Jun 22, 2017

U.S. National Center for Biotechnology Information (NCBI)

PubChem resource

https://pubchem.ncbi.nlm.nih.gov

PubChem is an ope blic resource with the p ide information on the of chemical substances

(simple question .. not a simple answer)

WHAT IS A CHEMICAL SUBSTANCE?

Image credit: https://dj1hlxw0wr920.cloudfront.net/userfiles/wyzfiles/3c8febc6-a9af-473c-a29b-313a2ffde9eb.png

What is a chemical substance? [scientific definition]

aka “pure substance”

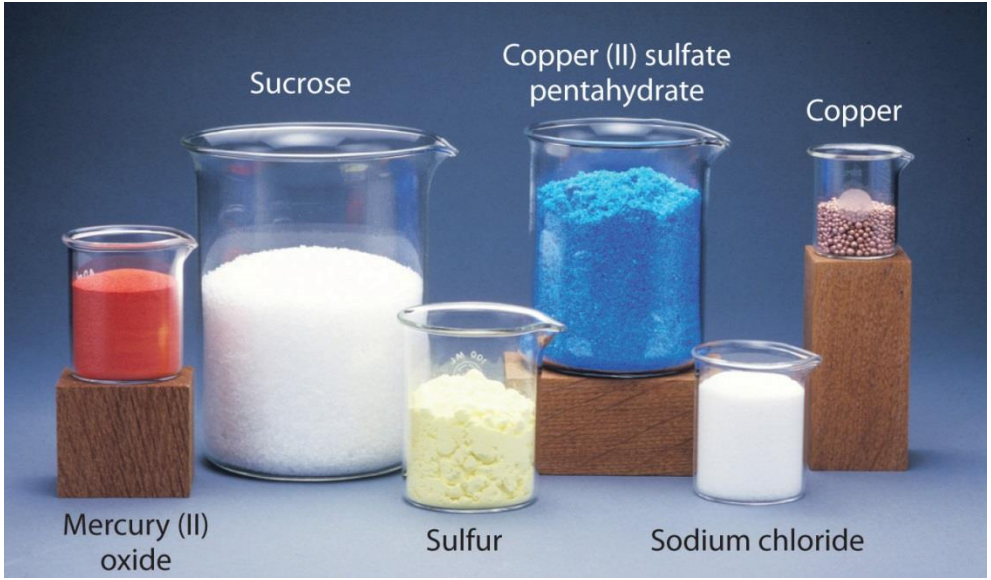


Image credit:
http://pindex.com/uploads/post_images/original/image_1410.jpg

Form of matter that has constant chemical composition and characteristic properties.

<u>Properties of chemical substances</u>

- Cannot be separated by physical separation
- Can be elements, compounds, ions or alloys
- Chemical reactions convert one into another
- Exist in different phases of matter such as solids, liquids, gases or plasma, and may change between them

Scientific definition of chemical substance breaks down

- Non-stoichiometric compounds (chemical composition is not a constant!)
  - Usually alloys or crystalline materials (inorganic solids) where some atoms are missing or extra atoms are present in an otherwise perfect lattice
  - Tungsten oxides [$W_nO_{(3n-2)}$, where n=20,24,25,40]
  - Yttrium barium copper oxide [$Y_xBa_2Cu_3O_{7-x}$, where x=0 to 1] ←important class of superconductors

ferrous oxide
[$Fe_{0.95}O$]

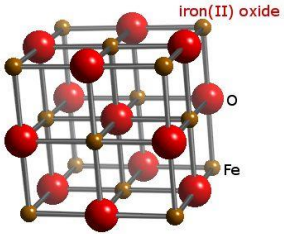


Image credit:
http://herb02.weebly.com/uploads/1/2/4/2/12420435/2087029_orig.jpg

pyrrhotite
[$Fe_{(1-x)}S$, where x=0 to 0.2]

Image credit:
http://www.dakotamatrix.com/images/products/pyrrhotite31943a.jpg

palladium hydrides
[PdH_x, where x=0.02 to 0.58]

Reversibly absorbs hydrogen →

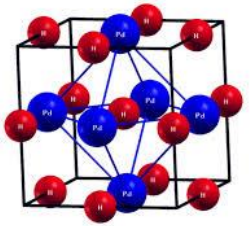


Image credit:
https://encrypted-tbn1.gstatic.com/images?q=tbn:ANd9GcS5baYlUSWs7sLQXPzt8mHLAQnH6GiYCsl8qXRvD-r3cm9guFag

Scientific definition of chemical substance breaks down

- In geology, substances of uniform composition are called minerals
- Many minerals mutually dissolve into solid solutions!
  - rocks that are a uniform substance despite being a mixture of minerals
  - e.g., feldspars such as anorthoclase, an alkali aluminum silicate where the alkali metal can be interchangeably sodium or potassium

(Note: aggregates of minerals are called rocks)

Anorthoclase typically consists of 10-36% of $KAlSi_3O_8$ and 64-90% of $NaAlSi_3O_8$

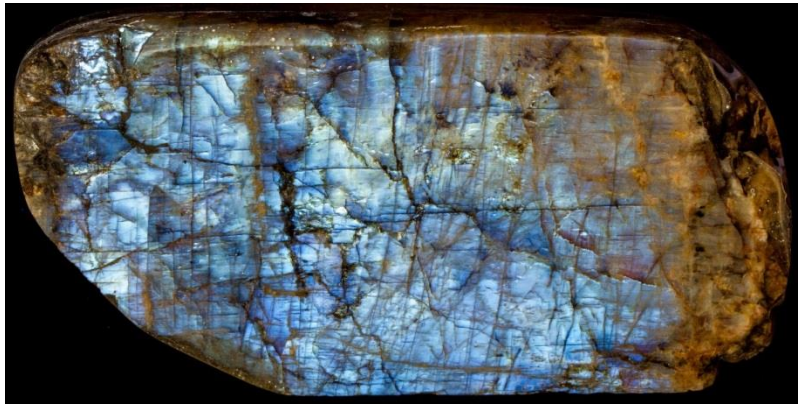


Image credit:
https://wgnhs.uwex.edu/wp-content/uploads/2013/01/anorthoclase-moonstone-1.jpg

Scientific definition of chemical substance breaks down

- Monomers (the repeated subunits) are discrete chemical substance entities
- Polymers are:
  - made up of one or more types of ‘monomers’
  - a large molecule (macromolecule) composed of repeated subunits
  - synthetic or natural (DNA, proteins, starches, etc.)
  - described by monomers, reactions, molar mass distribution (molecular weight range), properties, …

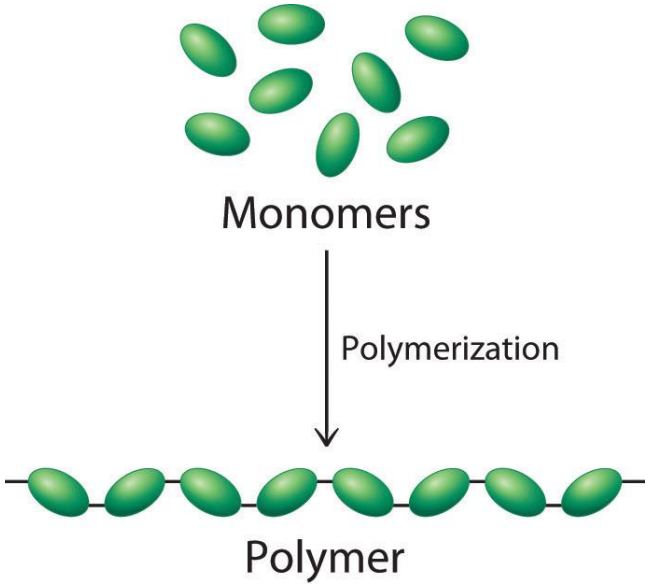


Image credit: http://images.flatworldknowledge.com/averillfwk/averillfwk-fig12_031.jpg

Polymer example

- Polyethylene is made up of ethylene (a simple gaseous unsaturated hydrocarbon) into long chains and classified by its density and branching
- Commonly known forms of polyethylene include LDPE (recycling code 4) and HDPE (recycling code 2)

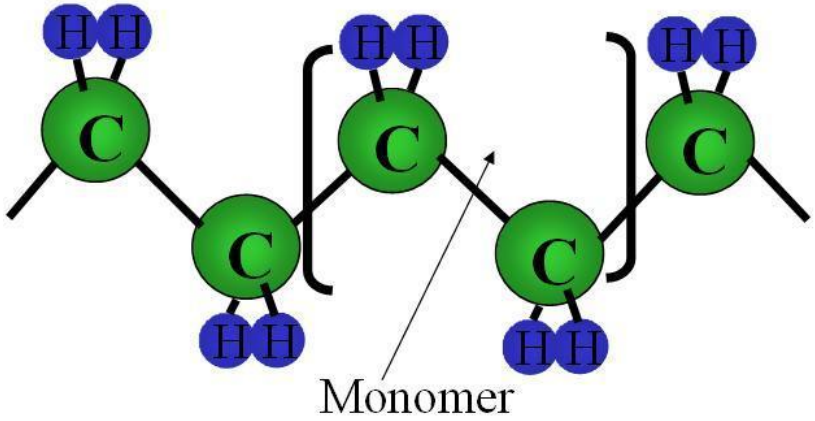


Image credit:
http://polymerkali.weebly.com/uploads/2/3/8/7/23873288/249832_orig.jpg

Image credit: http://www.recyclingdepotadelaide.com.au/wp-content/uploads/2014/10/plastic-types.gif

Chemical substance can also have a legal definition

- Can include mixtures with a defined composition or manufacturing process
- European Union (EU) regulation REACH defines ‘mono-constituent substances’, ‘multi-constituent substances’ and ‘substances of unknown or variable composition’ .. all of which are referred to as ‘chemical substances’
- Identity can be established either by direct chemical analysis or reference to a manufacturing process

Image credit:
http://www.vista-industrial.com/blog/wp-content/uploads/2013/07/reach.jpg

Legal definition example: Charcoal

- is an extremely complex, partially polymeric mixture that can be defined by its manufacturing process
- while exact chemical identity may be unknown, charcoal can be ‘identified’ to a sufficient accuracy

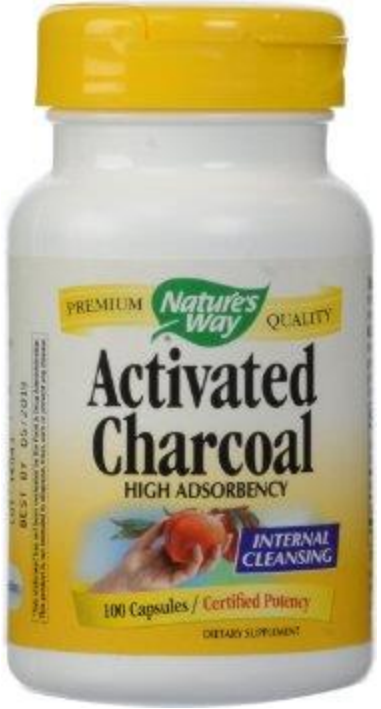


Image credit:
http://ecx.images-amazon.com/images/I/819fxj83xnL._SY355_.jpg

Practical real world use mirrors legal use

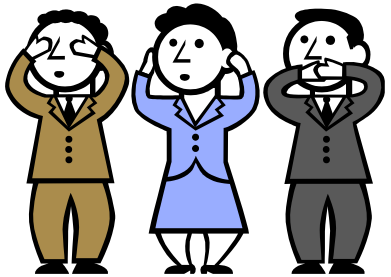
- anything you can assign a name and identify, isolate, or make is considered a ‘substance’
- if a chemical is associated to it .. one typically calls it a chemical substance or chemical or compound
- may be ‘pure’ or a mixture (e.g., stereoisomers, salts)
- often a ‘pure substance’ in a solvent (where solvent is ignored)
- convolution to the extreme

Image credit:
http://cliparts.co/cliparts/qTB/oqe/qTBoqeKAc.jpg

Chemical structure data is **rather** troublesome…

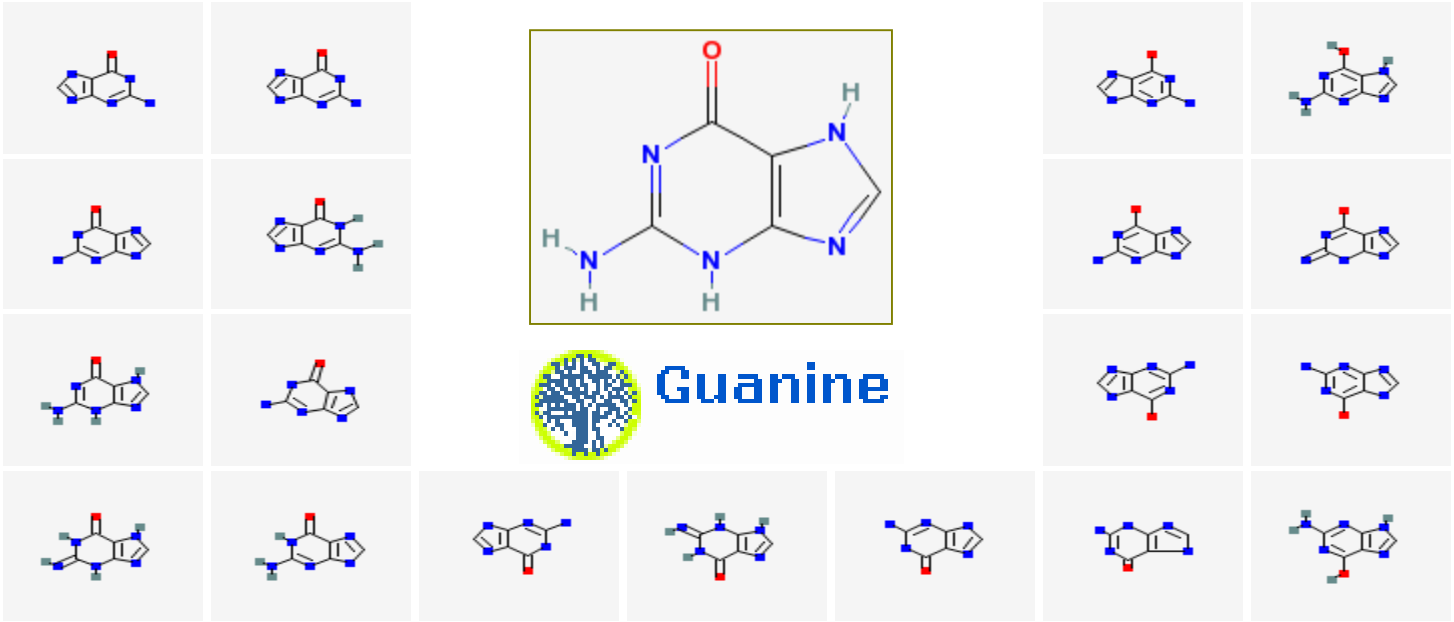
Let’s talk chemical information…

- **No “Global” rules or standards**
  - based on individual organizational needs
  - often based on individual preferences
  - “depictions” of chemical structures



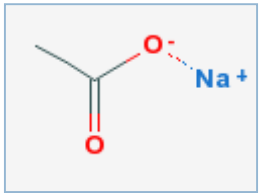
- **Data from many organizations**
  - conflicting “business rules”
  - previously unseen data representation schemes
  - combinatorial ways of drawing the same structure

A chemical structure may be represented in many different ways

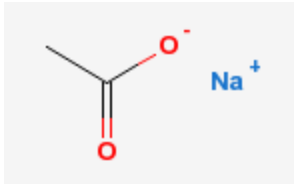


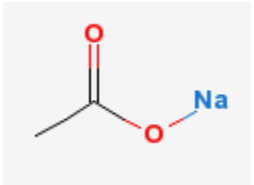
Tautomers and resonance forms of same chemical structure are prolific

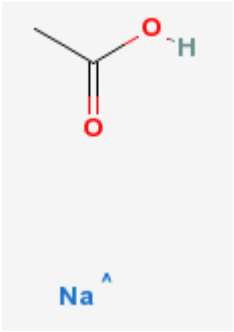
A chemical structure may be represented in many different ways



Sodium Acetate

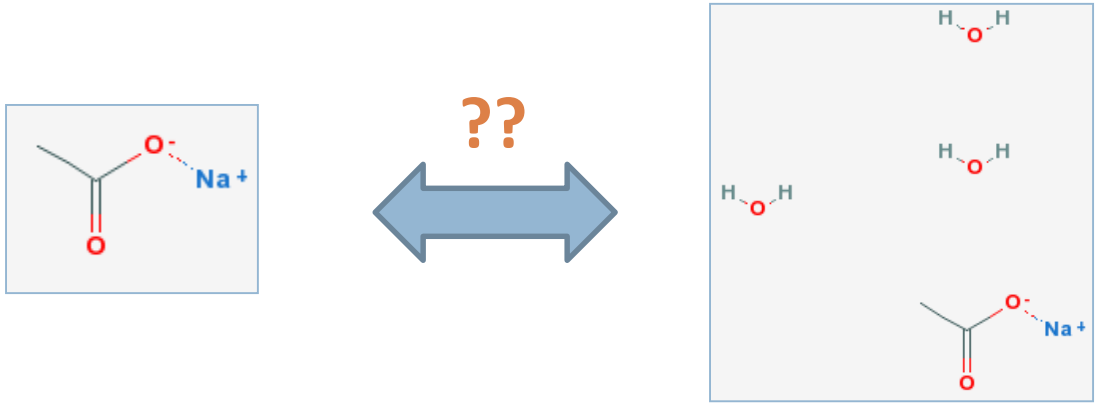






Salt-form drawing variations are common

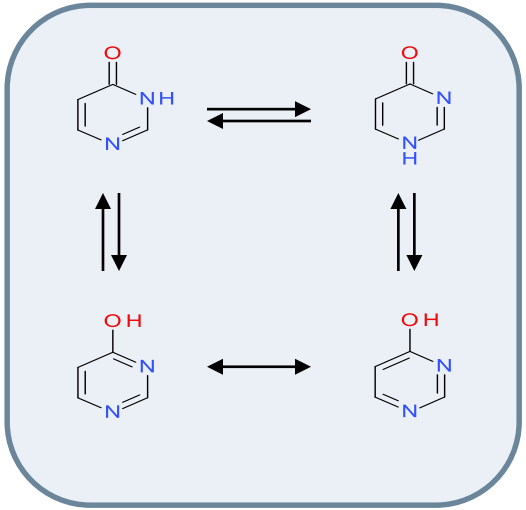
What do you mean by “sodium acetate”?

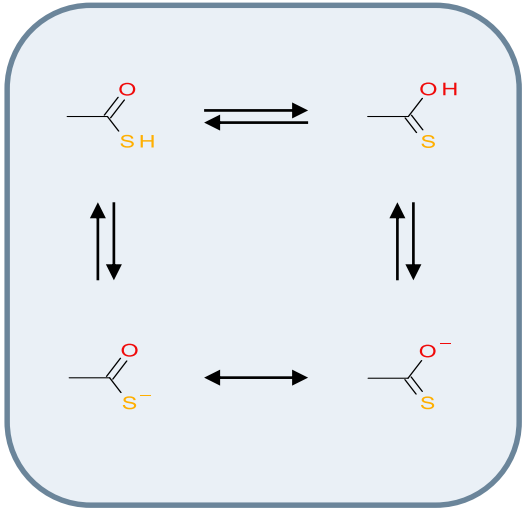


Sodium Acetate
The trihydrate sodium salt of acetic acid, which is used as a source of sodium ions in solutions for dialysis and as a systemic and urinary alkalizer, diuretic, and expectorant.

Chemical meaning of a substance may change upon context

Resonance, Tautomerism, Ionization, and Mesomerism





Which form does your software prefer?

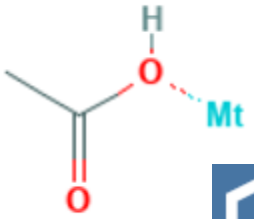
Which form do you prefer?

Special thanks to Roger Sayle and Volker Hähnke

Even elements can be troublesome

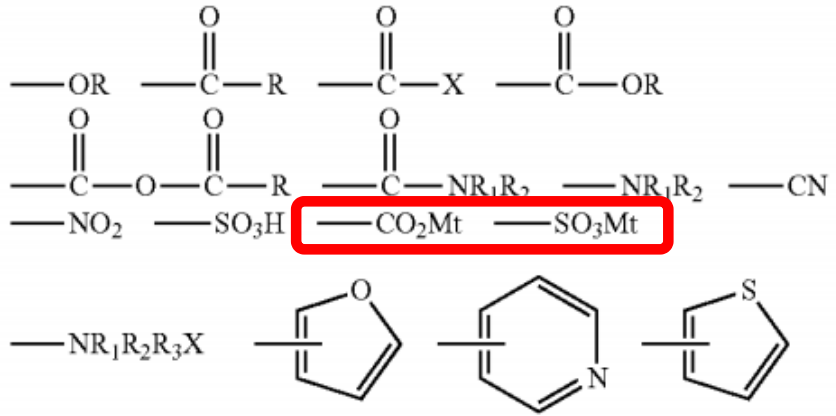
- Chemical diagrams often use abbreviations that can be mistaken for something else

Mt = Meitnerium



CID 60016057

Mt indicates Li, Na or K metal



Patent ID	Patent Title
US7871572	Chemical sensors based on metal nanoparticle encapsulated by mixed ligand and sensor array

Even elements can be troublesome

- ^{99m}Tc is a metastable form of ^{99}Tc used to image cancer

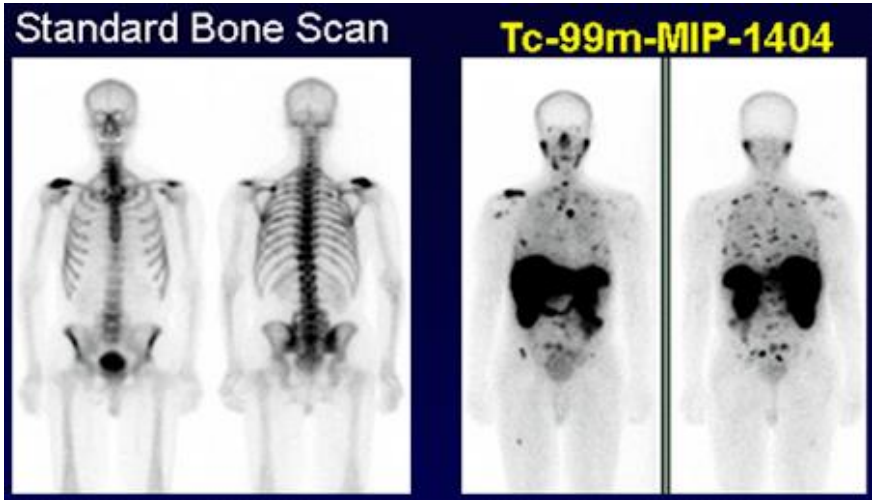


Image credit:
http://www.auntminnie.com/user/images/content_images/sup_mol/2013_06_11_17_05_49_101_snmmi_prostate_mip1404_450.jpg

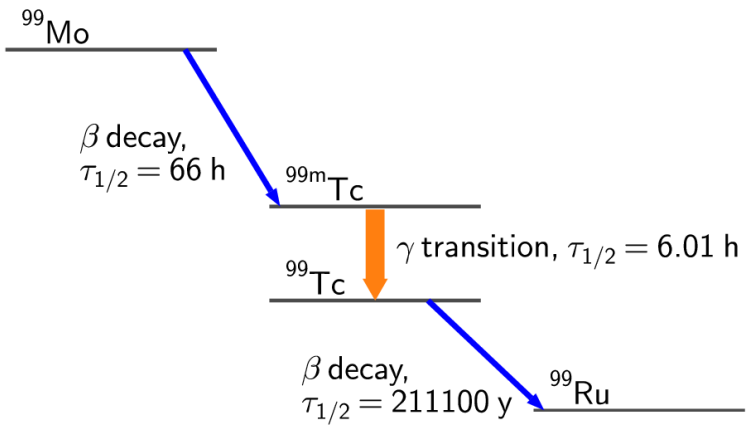
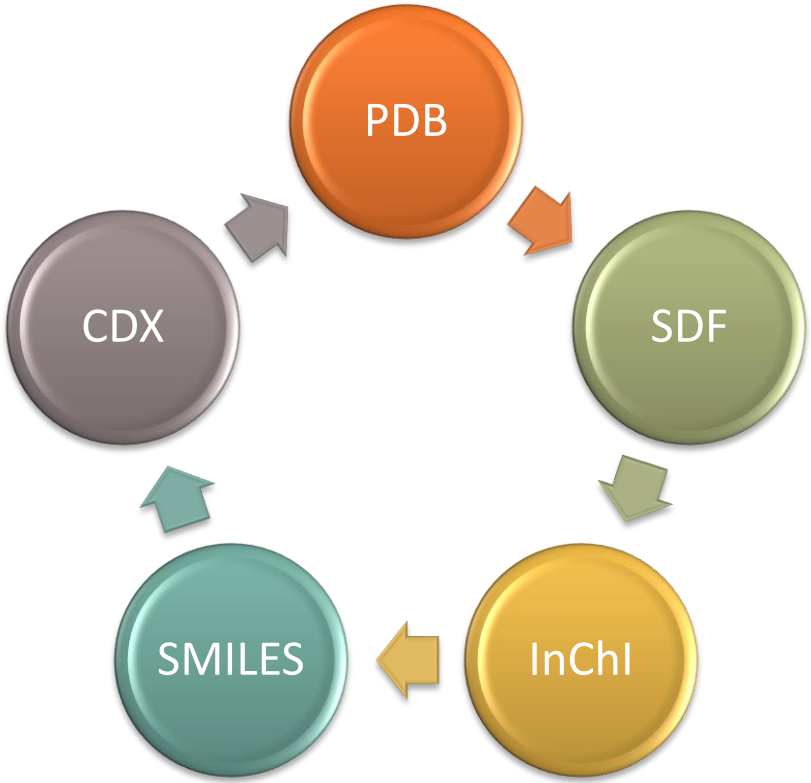


Image credit:
http://th.physik.uni-frankfurt.de/~scherer/Blogging/Tc99/decays_scheme.png

Where has this structure been?

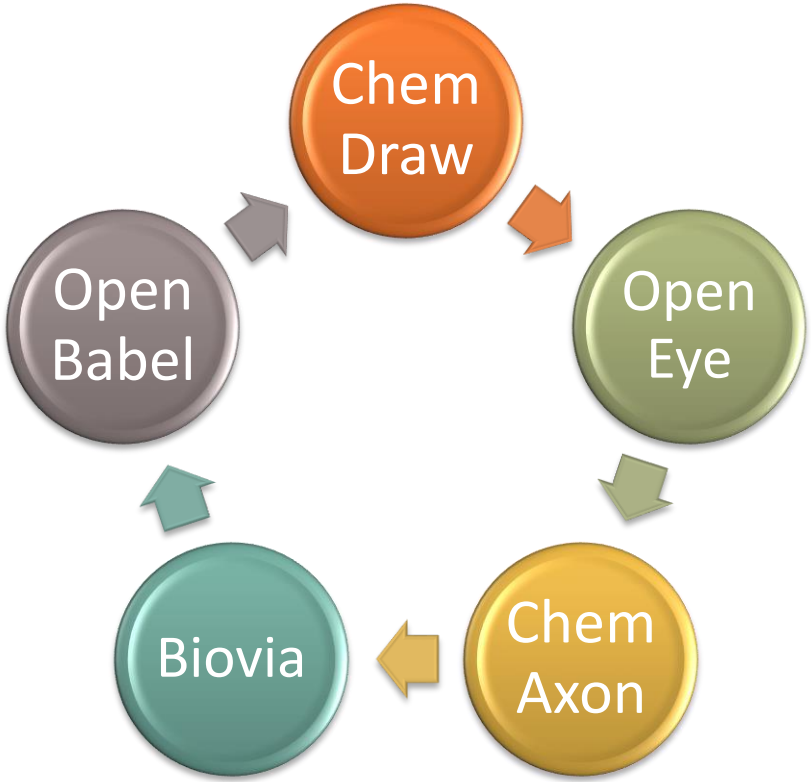
Export between file format flavors can result in reinterpretation of structure (2-D coordinates, atom types, bond types, etc.)



Chemical file format interconversion can be lossy

With whom has this structure been?

Different structure packages support structural information content to different extents



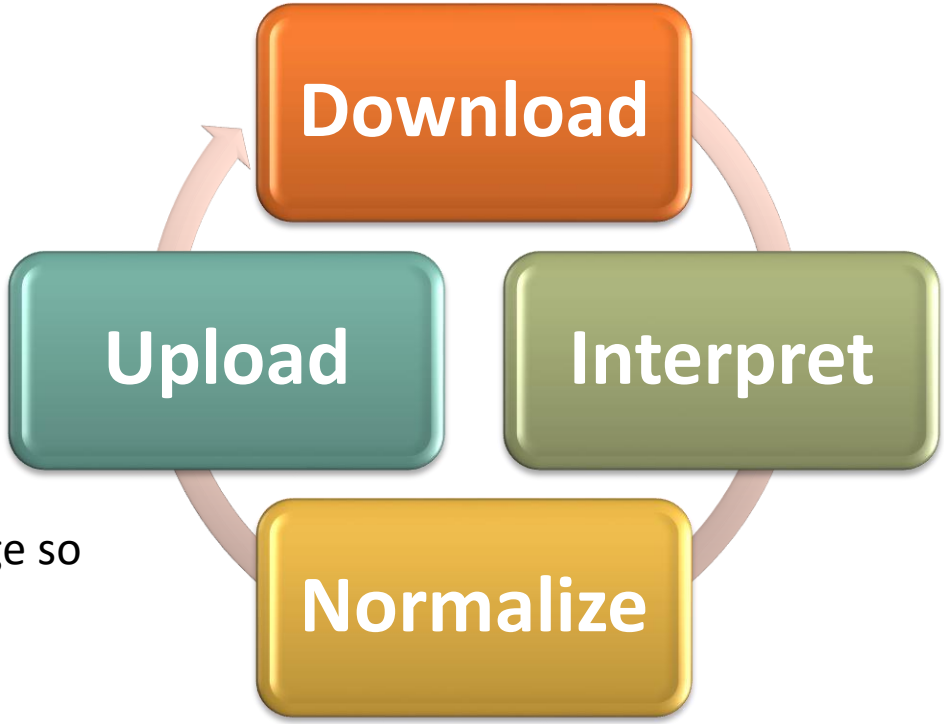
Different software packages may ‘normalize’ your chemical structure in different ways

Free flow of information

- Algorithms interpret information
- Can introduce ambiguity, errors
- Later use can mislead

- Now cycle this many times

- Need ways to lock down knowledge so not corrupted



The rise of the machines…
"RoboChemistry"…

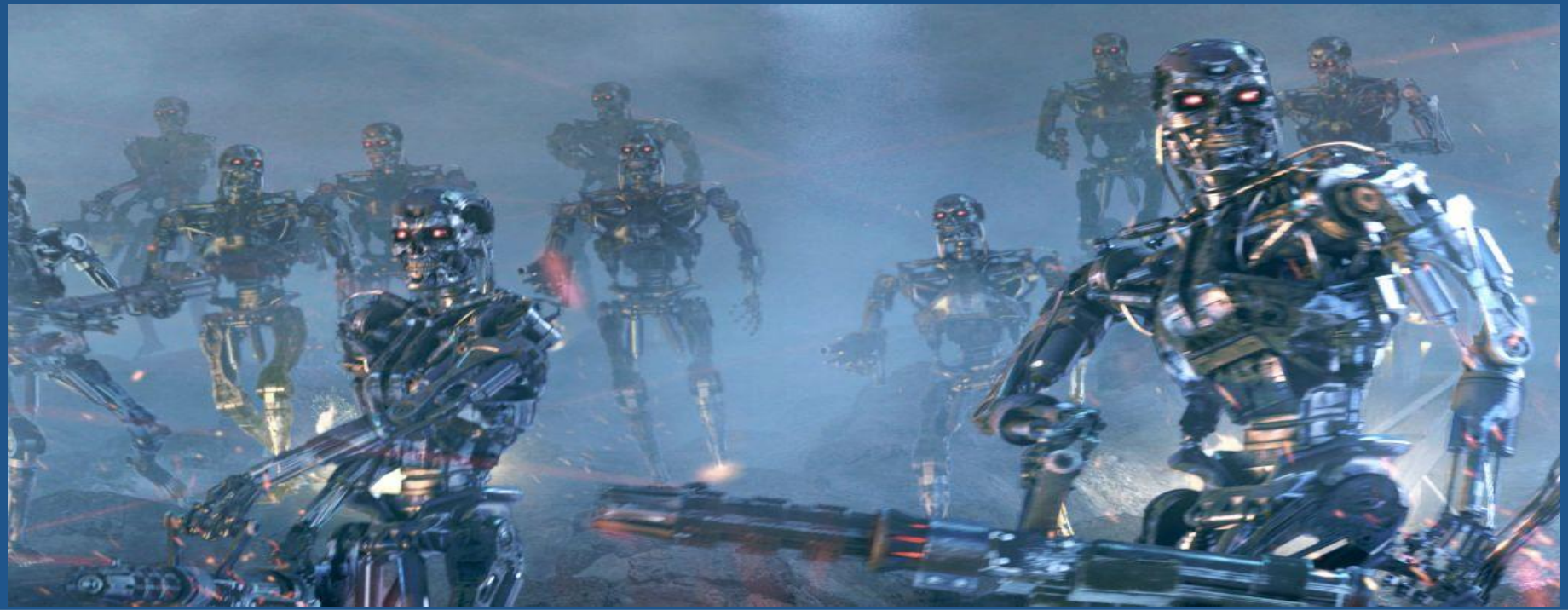
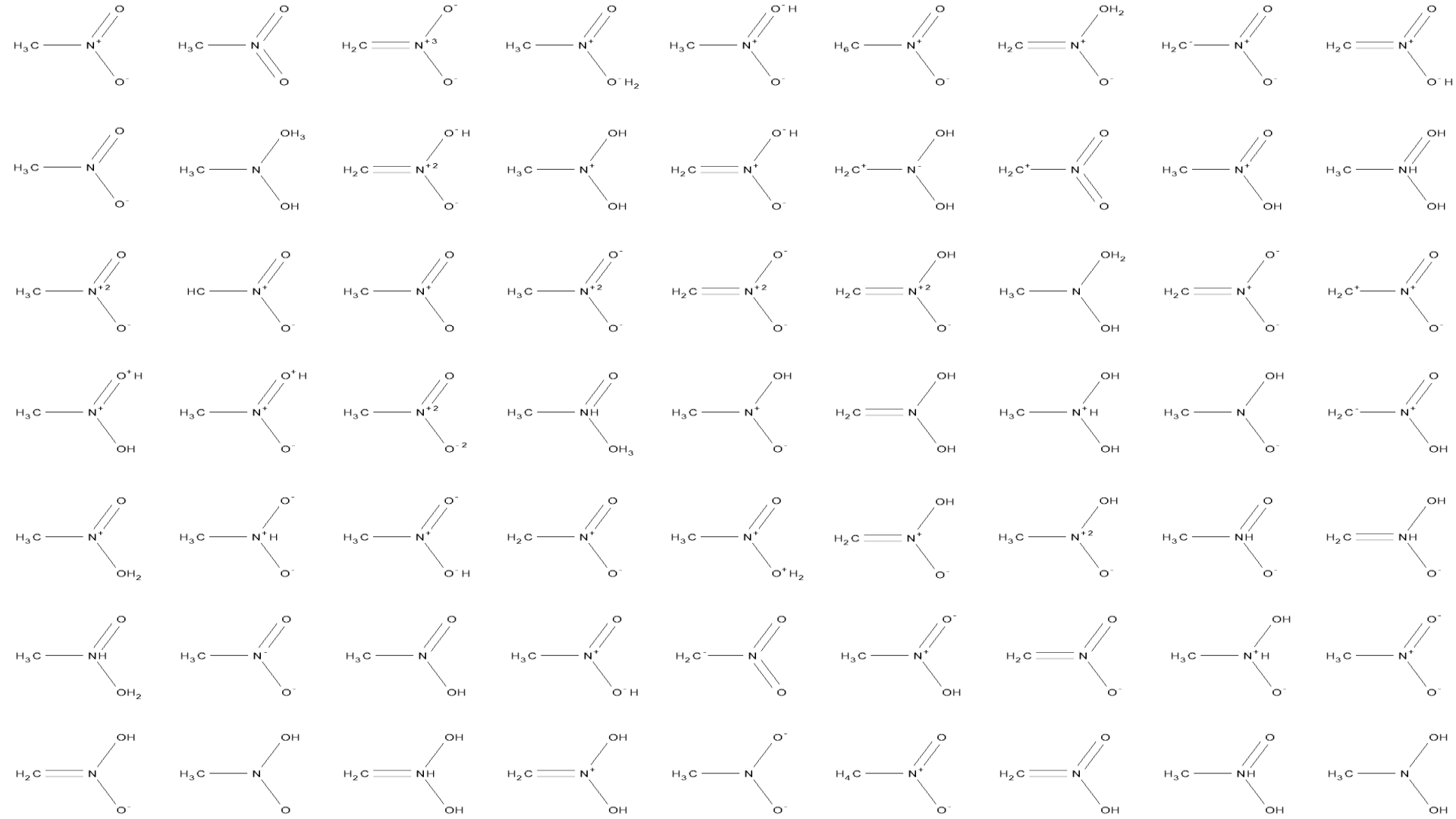
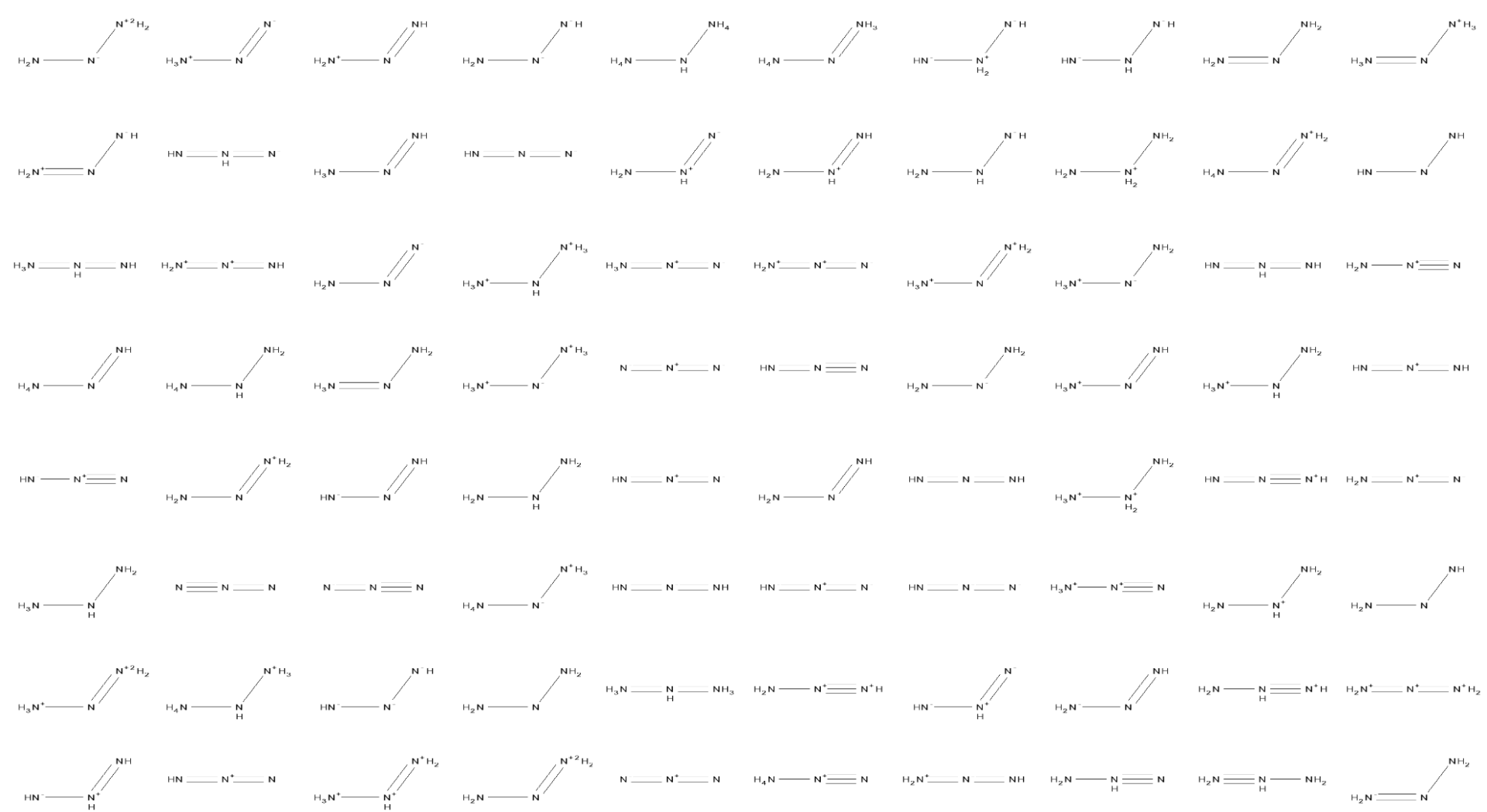


Image credit:
http://www.sequential.com.au/wp-content/uploads/2015/07/robots.jpg



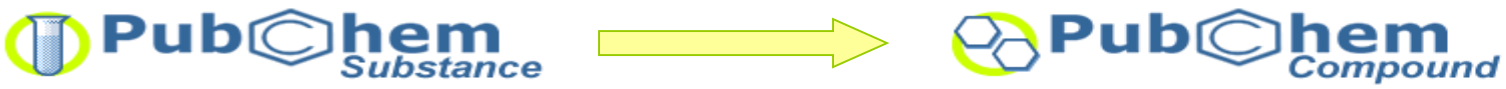


“RoboChemistry” = apocalypse?

(wastelands of chemical structures)

Image credit:
http://vignette2.wikia.nocookie.net/terminator/images/3/3e/Judgementday.jpg/revision/latest?cb=20090419205037

Automated structure processing...

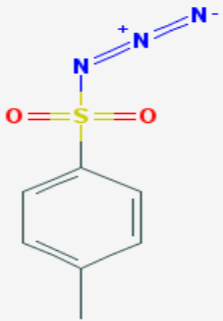


- **Verify chemical content**
  - Atoms defined/real
  - Implicit hydrogen
  - Functional group
  - Atom valence sanity
- **Calculate**
  - Coordinates
  - Properties
  - Descriptors

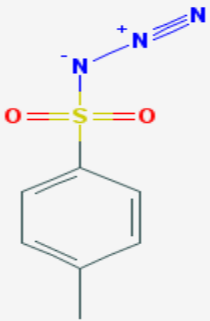
- **Normalize representation**
  - Tautomer invariance
  - Aromaticity detection
  - Stereochemistry
  - Explicit hydrogen
- **Detect components**
  - Isolate covalent units
  - Neutralize (+/- proton)
  - **Reprocess**
  - Detect unique

Sadly, we are always fighting the last war…

new contributor = new chemical representation

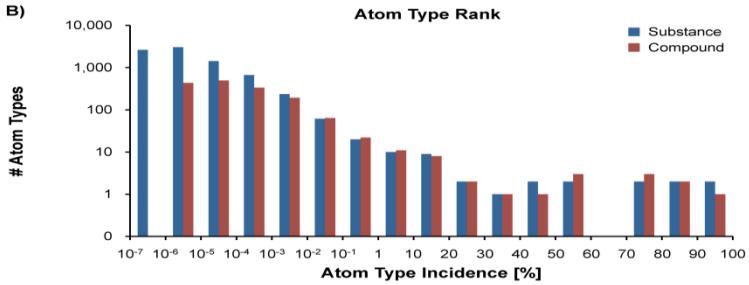


Tosyl azide
CID 13661



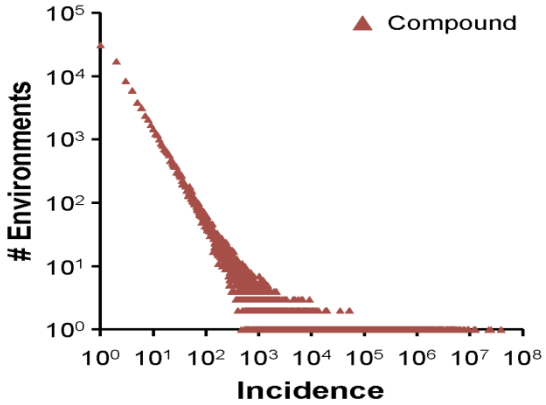
Tosyl nitride
CID 5359526

Mapping chemistry space using molecular (fragment) descriptors

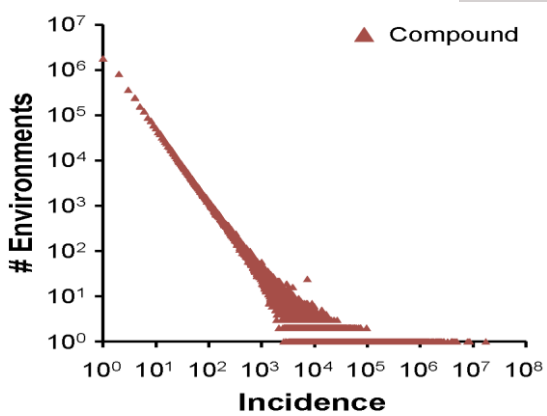


The +46M PubChem Compound records (as of Jan. 2013) contained unique atom environments:

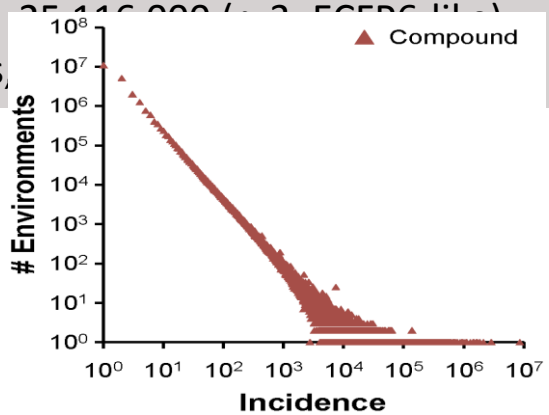
1 583 (r=0, atom types)
109 308 (r=1, ECFP2-like)
4 559 753 (r=2, ECFP4-like)



r=1 (ECFP2-like)



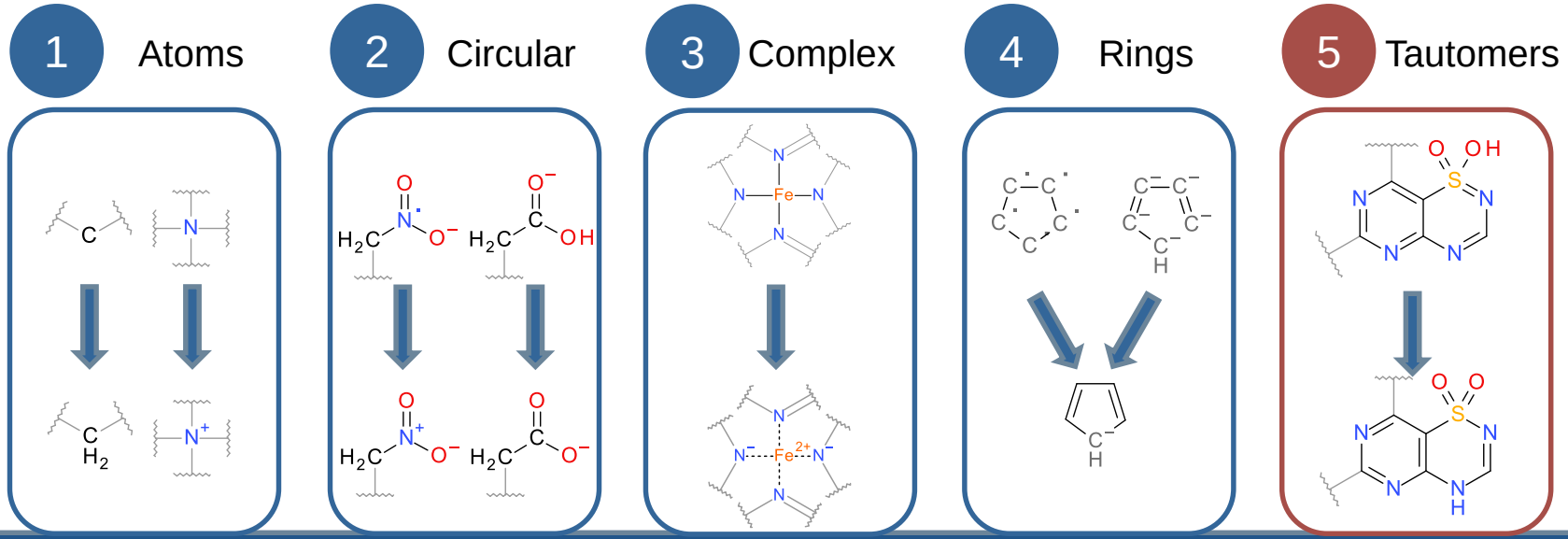
r=2 (ECFP4-like)



r=3 (ECFP6-like)

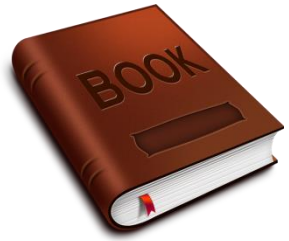
Atom Environment Standardization

- **Atom Environment Types**



WAIT .. IT GETS BETTER

Image credit:
https://image.freepik.com/free-icon/evil-emoticon_318-40171.png



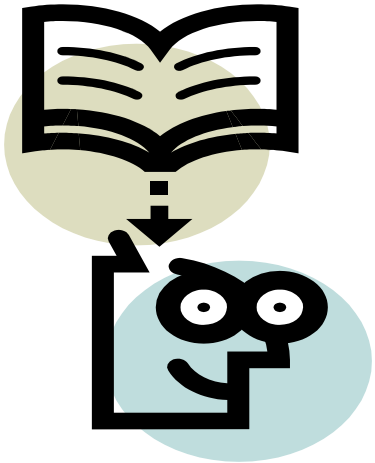
Premise

MOST CHEMISTRY KNOWLEDGE IS LOCKED UP IN TEXT.

Image credits:
https://play.google.com/store/apps/details?id=com.uc.addon.web2pdf
http://ideasuccessnetwork.com/idea-discovery-article-how-good-idea-mushroomed/
http://xk2.ahu.cn/
http://libraryschool.libguidescms.com/content.php?pid=682172
https://www.freshrelevance.com/blog/real-time-marketing-report-for-may-2014

Scientific Information

Presentation



Data

Human understanding

Scientific Information

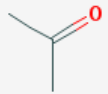
Presentation

Data

What about computer understanding?

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

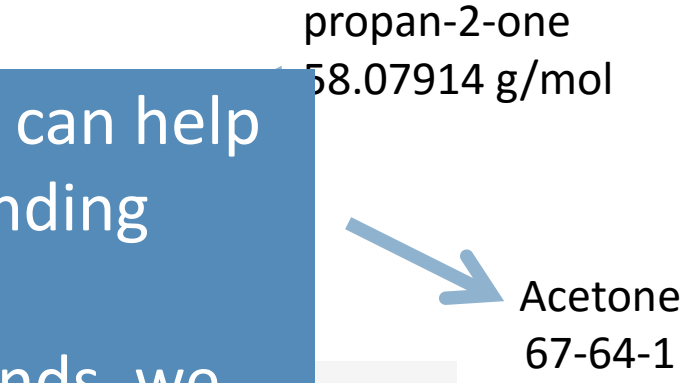
Almost all chemical information is geared towards human understanding in the form of text and images

Computer understanding of chemical information

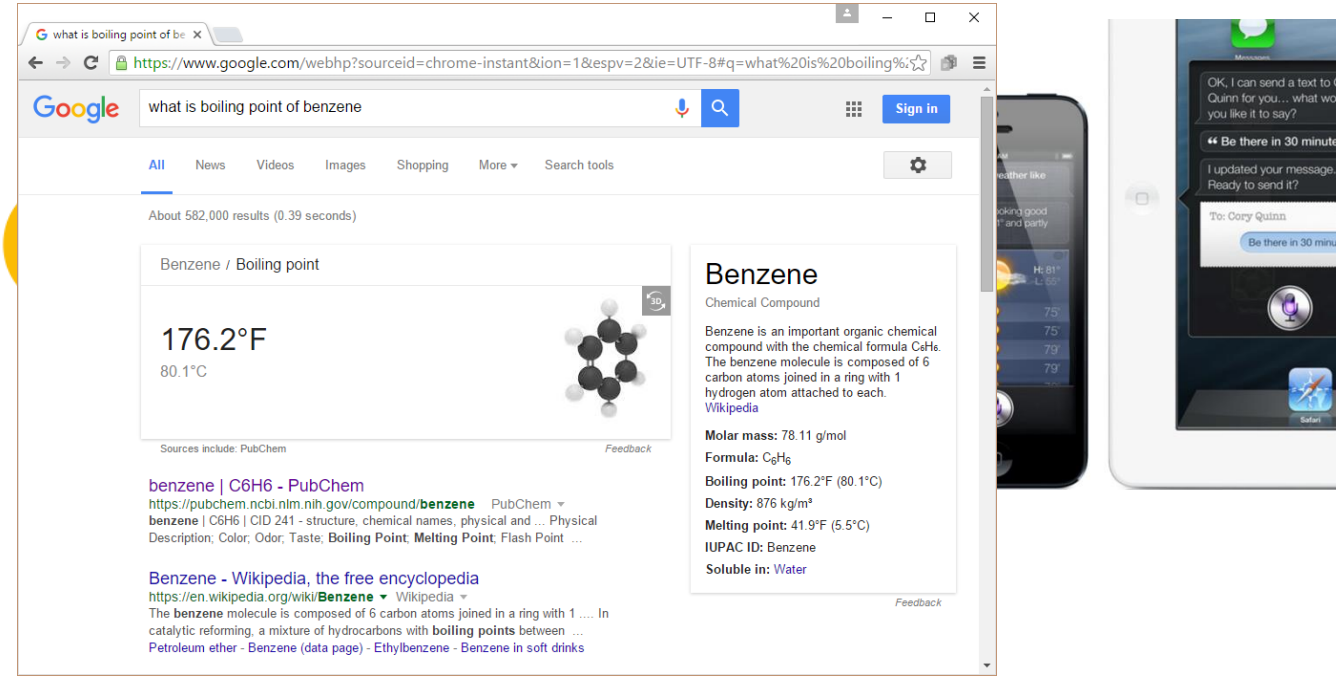
- Give a computer a chemical structure a
cannot und
- A computer
from the st
- A computer
information
- A computer
key information from structure

Computer understanding can help provide human understanding

If the computer understands, we can leverage it for search, analysis, and more



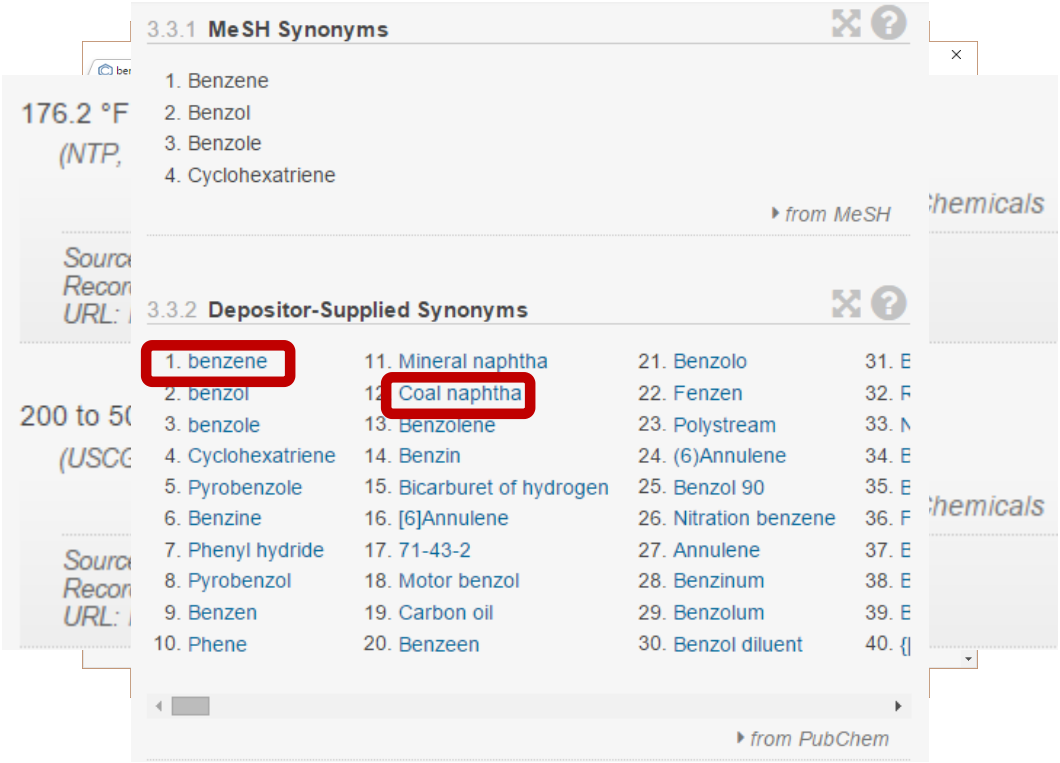
Effects of computer understanding you can experience



Benzene boiling point case study

Benzene

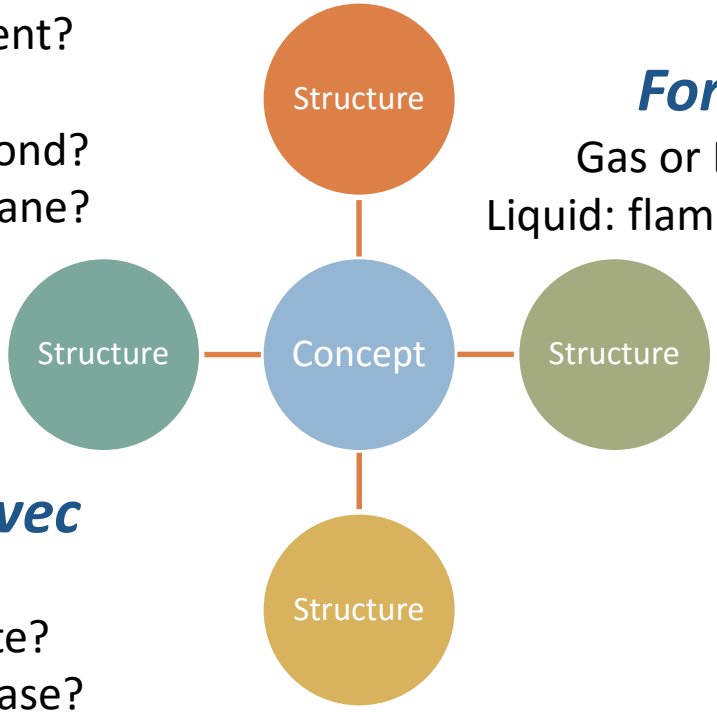
Coal tar



Many to many relationships

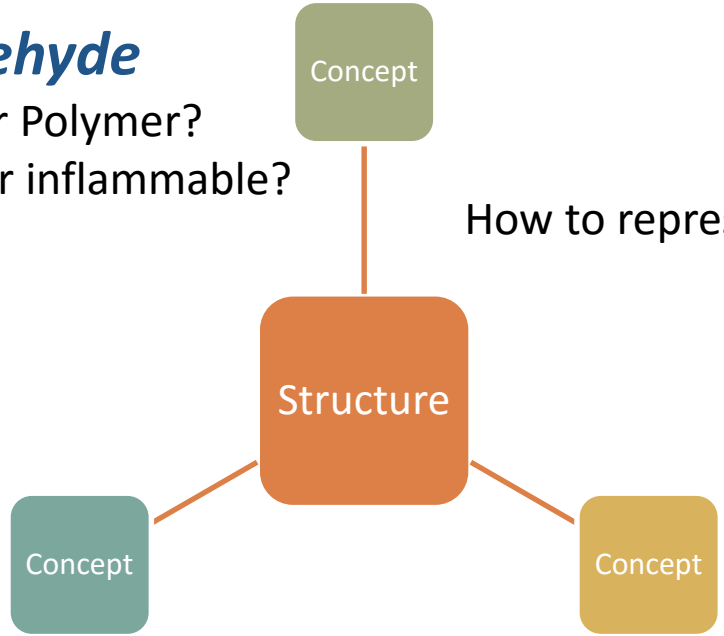
Carbon

Element?
Coal?
Diamond?
Methane?



Formaldehyde

Gas or Liquid or Polymer?
Liquid: flammable or inflammable?



How to represent?

Gleevec

Salt?
Hydrate?
Free base?

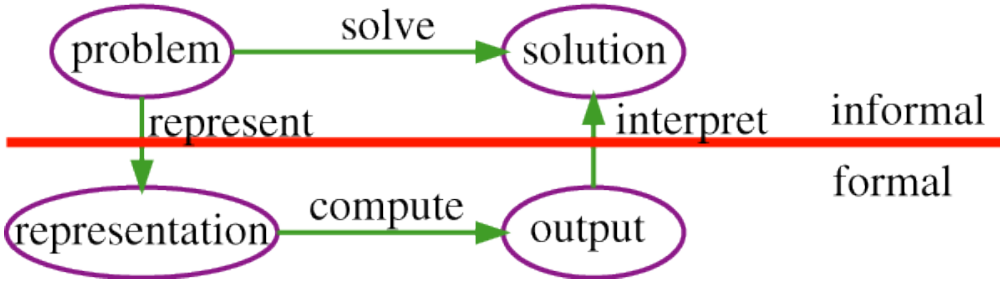
Same chemical structure can represent many things, most of which are use case dependent

structure is not enough

- Chemical information is a bit of a mess and can be rather nuanced
  - Names, names, and more names (+300M in PubChem)
    - Some standard names are not open and cannot be used/verified without $$$
  - Name/structure associations vary by use case (many overlapping)
    - Acetic acid vs. Acetic acid tri-hydrate
    - Formaldehyde: (gas) vs. Formalin (liquid, 40% formaldehyde w/ water)
    - Sulfuric acid: SO_3 (gas) vs. H_2SO_4 (liquid w/ water)
    - Glucose: L/D, ring open/closed (f/p), alpha/beta/both vs. Glucose monohydrate
    - Large corpus in the 'wild' .. data source dependent nuances
- Verify with primary source(s) prior to information use
  - i.e., is this the form of the chemical I care about?

Go below 32% formaldehyde in water and it is non-flammable

Users are not happy to see overlapping data on different forms of same chemical

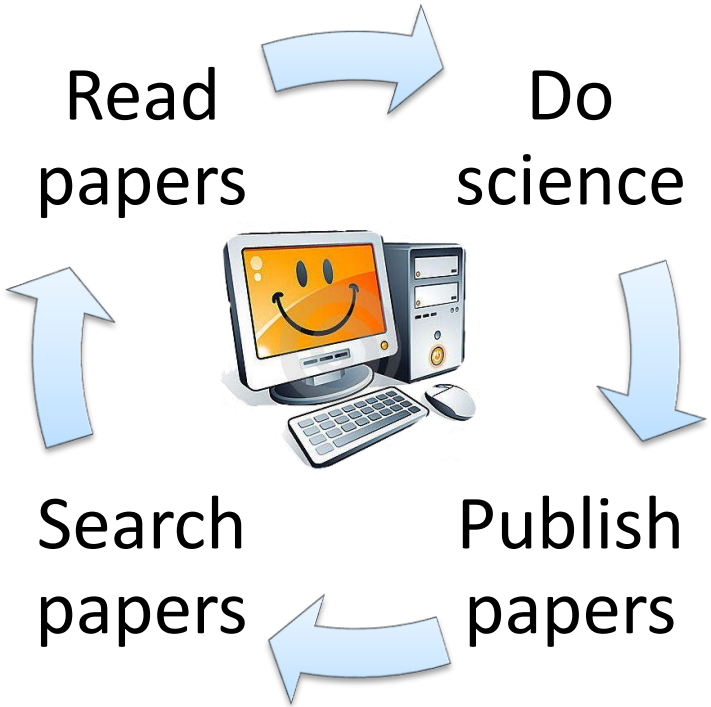


Knowledge representation helps to provide “information about the world in a form that a computer system can utilize to solve complex tasks”…

HOW DO WE BRING IT ALL TOGETHER?

Image credit: http://artint.info/html/ArtInt_8.html

Simplistic science workflow



Computers help yet we rely on humans who abstract out keywords, article gist, data, etc

CAS, Medline, ChEMBL, etc.

Image credits:
http://www.how-to-draw-funny-cartoons.com/cartoon-scientist.htm
http://computertutorinc.net/computer-maintenance-safety-tips/

Computer aided abstraction

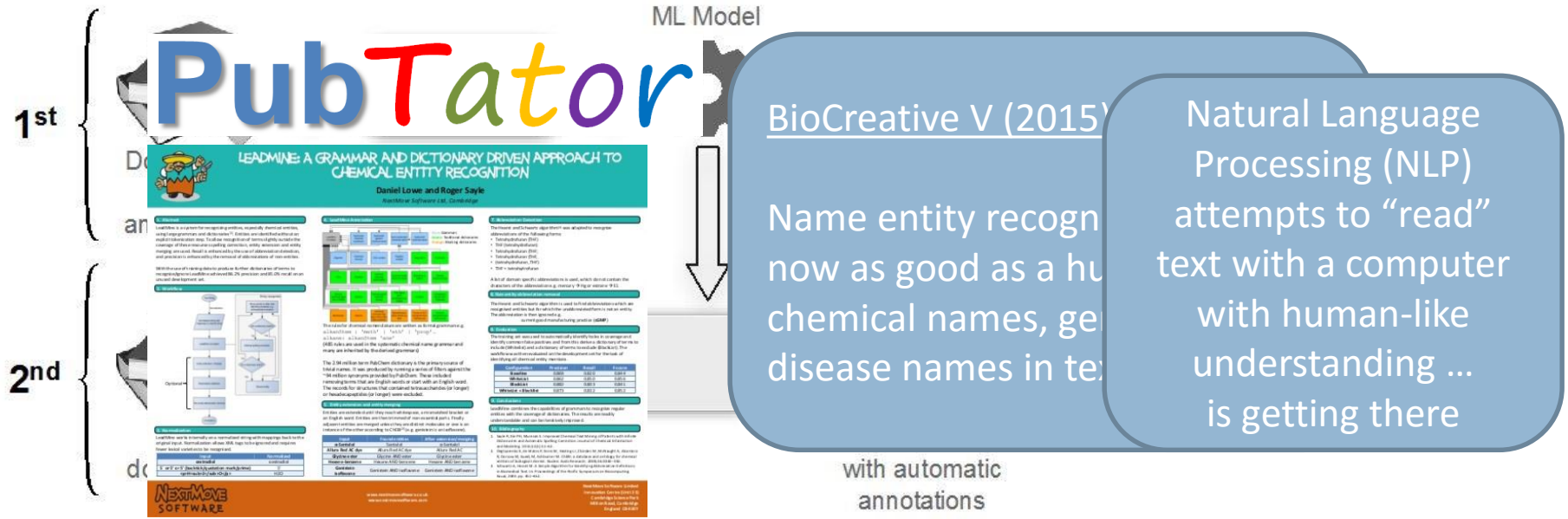


Image credits:
http://www.ncbi.nlm.nih.gov/CBBresearch/Lu/ner/
http://www.ncbi.nlm.nih.gov/CBBresearch/Lu/Demo/PubTator/
http://www.intechopen.com/books/theory-and-applications-for-advanced-text-mining/biomedical-named-entity-recognition-a-survey-of-machine-learning-tools
http://www.slideshare.net/NextMoveSoftware/leadmine-a-grammar-and-dictionary-driven-approach-to-chemical-entity-recognition

Concept-based information

"subject-*predicate*-object"

"atorvastatin *may treat* hypercholesterolemia"

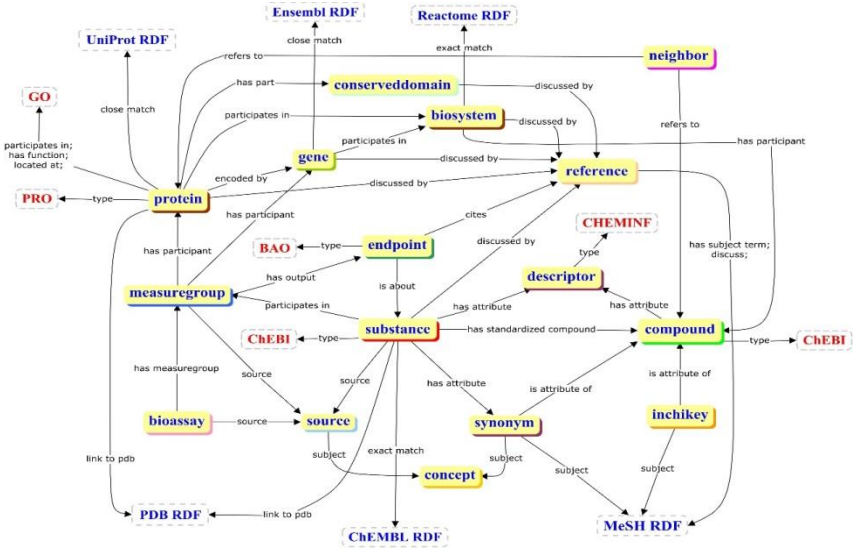
Provenance information

Evidence citation (PMID)

From whom? (Data Source)

PubChem data complexity

- Many links between large record collections
  - ~230M Substances <-> ~90M Compounds
  - ~90M Compounds <-> ~90M Compounds
  - ~230M Bioactivities <-> ~3M Substances
  - ~230M Bioactivities <-> ~2M Compounds
  - ~230M Bioactivities <-> ~1M BioAssays
  - ~10M PMIDs <-> ~100K Compounds
  - ~3M Patents <-> ~30M Substances
  - ~3M Patents <-> ~15M Compounds



Community drive towards knowledge representation

Dear Colleagues,
We would like to thank you for taking the time to participate in our first meeting to address chemical ontologies (CO). Below please find some notes / minutes from the **meeting held in Basel, Switzerland, October 2, 2015**.

The purpose was to explore using computers to ingest machine-readable forms of molecules and to generate molecular attributes (descriptors). For example, ingesting a SMILES string and producing a set of triples that describe the molecule [molecule X "is a " ketone ; " is a " amino acid ; "is a " steroid etc.]. The output of which would provide the basis of a chemical ontology to be used for classification purposes as well as for input for downstream operations such as knowledge graphs, data mining, chemical text mining and cognitive computing experiments. Historically these operations were performed manually or semi-automated; however, it is desirable to have a computer process for large scale processing to meet current day demands resulting from computer curation of the scientific literature. To date, two programs have been developed to accomplish this objective: one at OntoChem, a German informatics company, and another (ClassyFire) at the University of Alberta. While both programs produce reasonable output, there are differences that could lead to non-conformity in the resulting ontologies. **One motivation for the workshop was anticipation that all parties will benefit from *common* standards for a computer-derived chemical ontology.** Overall we believe that a Chemical Ontology can make contributions when it comes to answering scientific relevant questions.

Workshop hosted by Fatma Oezdemir-Zaech, Novartis Pharma AG

Computational vs. Manual Classification

- Humans make mistakes
  - This includes programming mistakes, classification mistakes, encoding mistakes
- Manual classifications only handle a small number of chemical structures
  - Automated classification can help to extend classifiers to all known chemicals
- Harmonization of terminology?
  - Can they speak a common language?

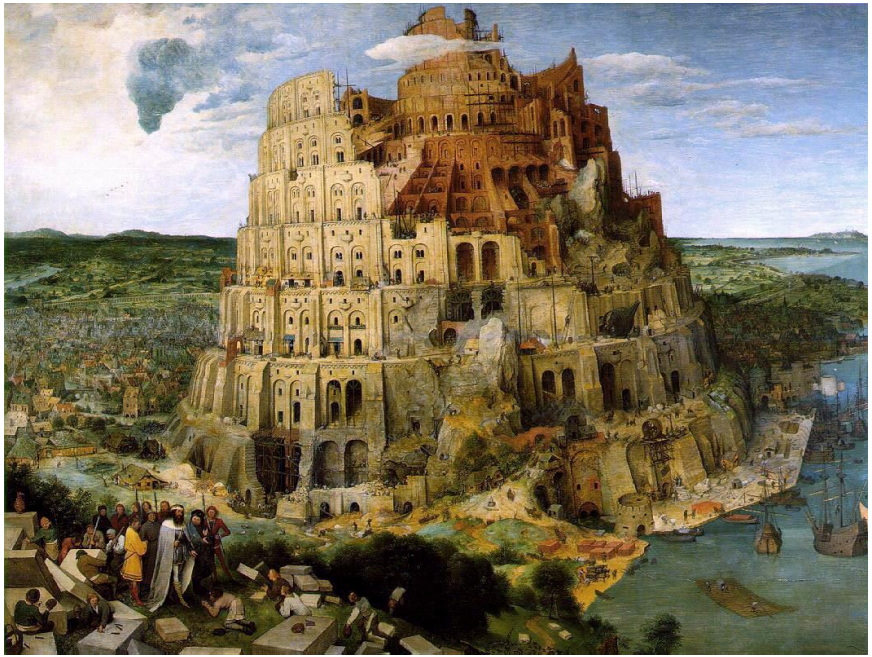
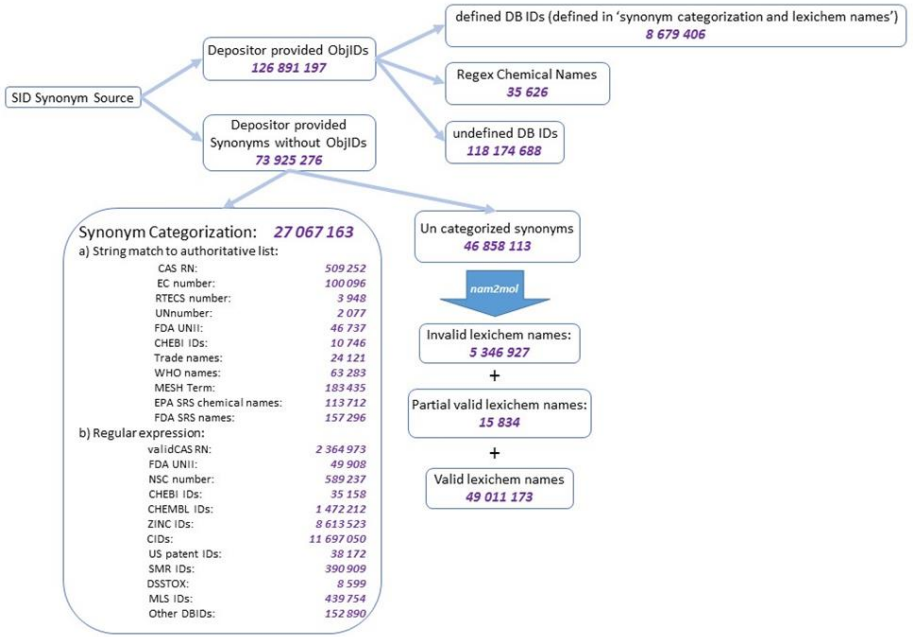


Image credit: https://commons.wikimedia.org/wiki/File:Brueghel-tower-of-babel.jpg

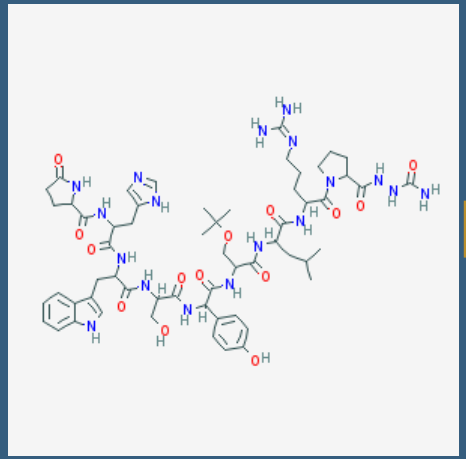
PubChem Synonym Classification

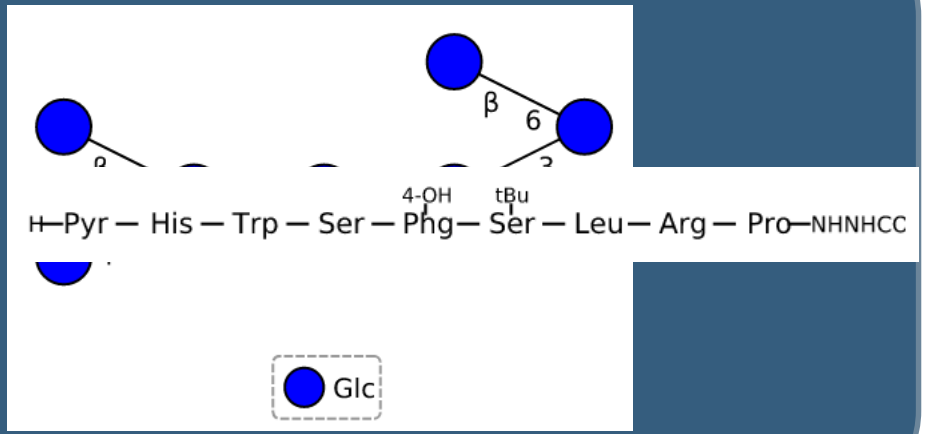
- available in RDF first
- indicates chemical name type
- allows grouping of names
- can involve guess work
- adding more name sources
- most non-classified names are not helpful (chemical name corruption/fragments?)



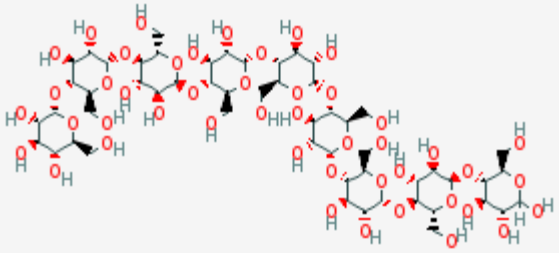
Towards adapting chemical biology approaches for structure representation

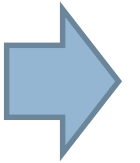
- Substances that are (chemically modified) carbohydrate, protein, nucleotide
- Need not be fully defined (positional isomers, motif, etc.)
- May include lipids, nanoparticles
- Limits on size

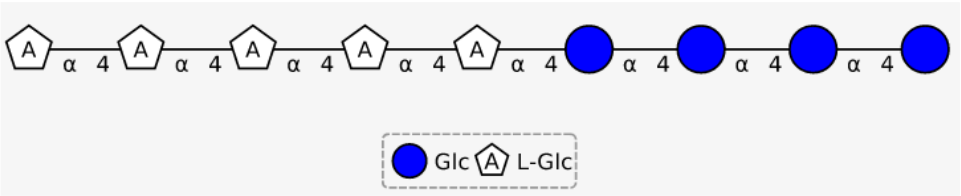




Maltononaose

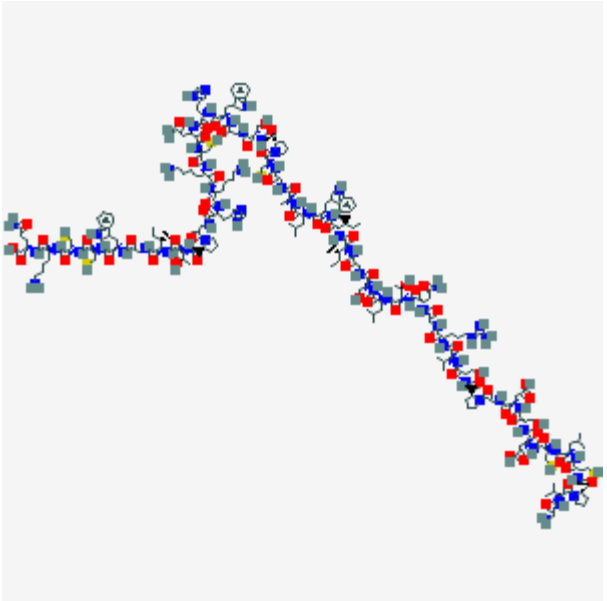


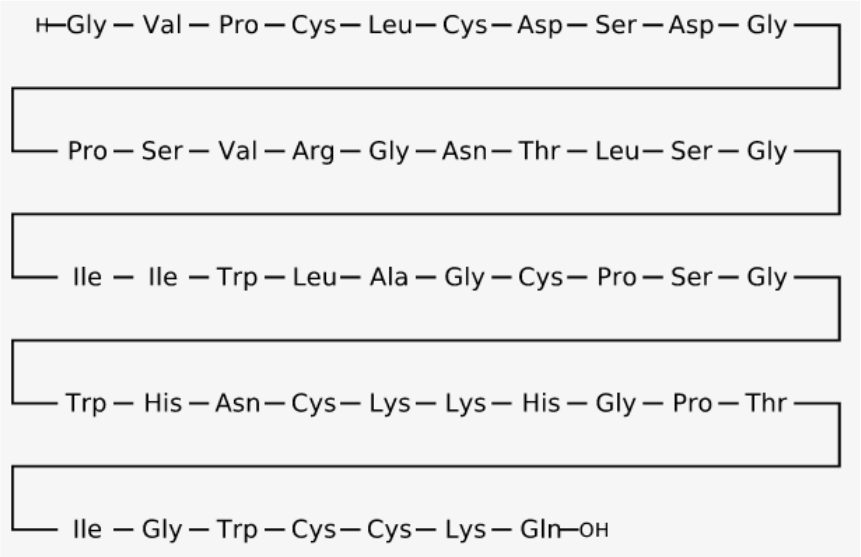




Sea Anemone Toxin II

Cardiotonic Agent





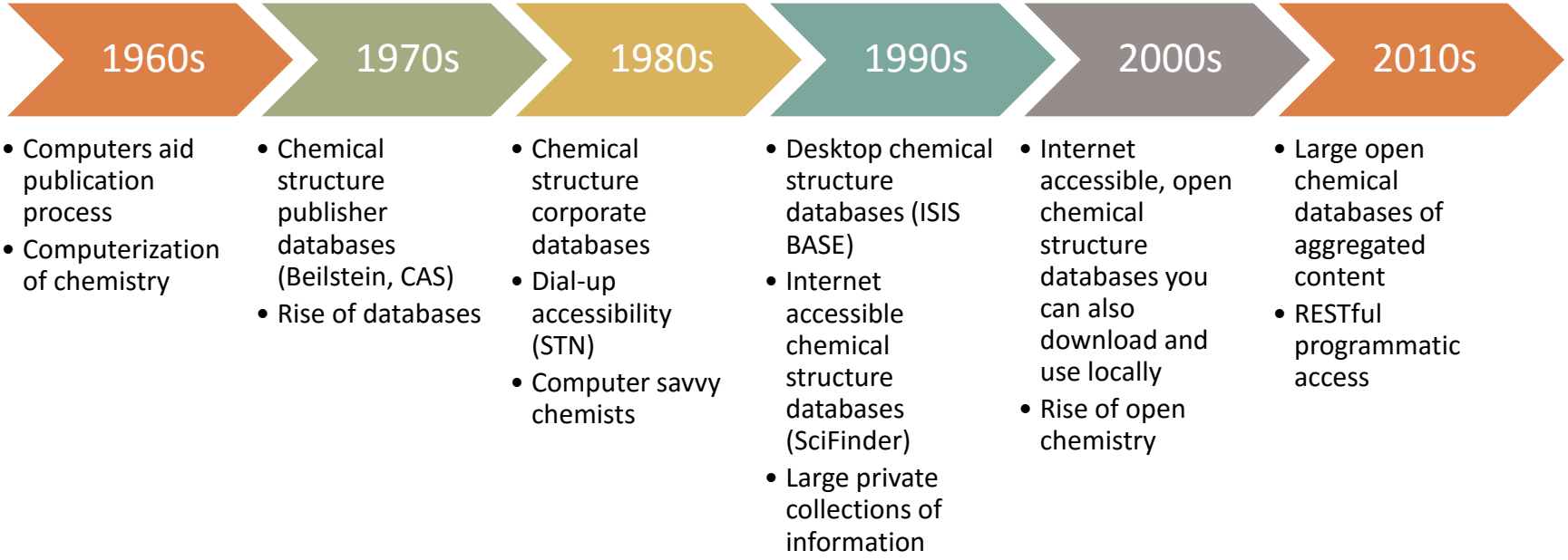
Sea anemone toxin II (ATX II) is a useful tool for investigation of sodium channels in excitable membranes.
http://www.ncbi.nlm.nih.gov/pubmed/2880413

Clean up approaches

- Structure standardization
- Consistency filtering
  - Name/Structure – what structure for a name?
    - Minor variation means new determination
  - Authoritative sources
  - Black lists, Grey lists, White lists
  - Concepts (group of synonyms for same ‘thing’)
    - Preferred concept for a structure
    - Preferred structure for a concept
- Cross validation via text mining – can we find evidence for association/link between gene-chemical?

Image credit:
http://3.bp.blogspot.com/_KG_4as681ns/S9ihvrpm-uI/AAAAAAAAAEc/yuJhJ4-gcuU/S1600-R/small+title.jpg

An evolution of chemical structure databases



Special thanks to discussions with Dr. John Rumble and Dr. Evan Hepler-Smith

Image credit:
http://www.worldsviewacademy.com/wp-content/uploads/2015/11/Where-are-we-now-header.png

Community opportunities

- 2010s – large open collections of chemical structures (10s of millions) – non-curation combined with open exchange of data leads to error proliferation – need methods to lock down the data without curation
- Digital standards to improve data exchange and prevent error proliferation
- Close attention to provenance – where did this data come from?
- Set of clear definitions for chemical concepts

Entering a new era …

Image credit:
https://fortunedotcom.files.wordpress.com/2014/12/mac_mshadow1.jpg?w=840

2020s – large, extensively machine curated, open collections – clear provenance, standard approaches to file formats and normalization, errors do not proliferate, links cross-validated, emergence of open knowledge bases that contain all open scientific knowledge that is computable (inference-able, natural language questions, …)

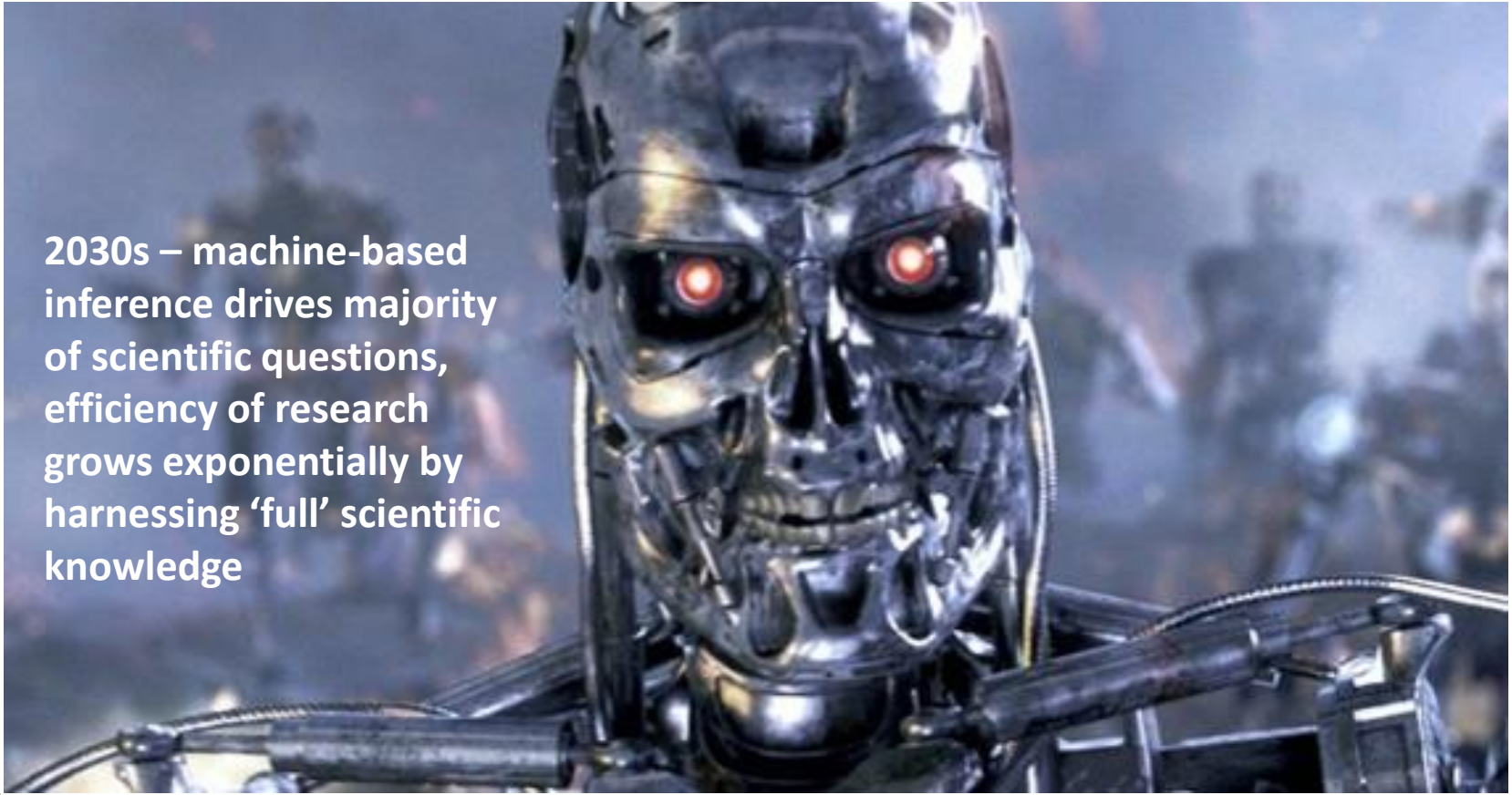


Image credit:
http://static.srcdn.com/slir/w570-h300-q90-c570:300/wp-content/uploads/terminator-5-release-date-new-trilogy.jpg

HOW DO WE GET THERE?


WE NEED
YOU

You can help improve chemical informatics

Image credit:
http://www.idreamcareer.com/img/blog/1449649713-make-a-difference.jpg

Image credit:
https://media.licdn.com/mpr/mpr/shrinknp_400_400/AAEAAQAAAAAAAAcPAAAAJDFjOTk4YzRiLWJiZTItNDBkNi1hYTYyLTJkOTFiZTBlNTMzMQ.jpg

Feel free to email me with questions and thoughts .. **evan_bolton@nih.gov**

What is this all about?

- An attempt to identify opportunities to advance chemical
- A ‘call to arms’ on areas in need of focus
- An assessment of the landscape we work
- A search for interested parties and champions

Image credit:
http://www.riseresearchproject.com/wp-content/uploads/2015/10/Assessment.jpg

We know there are issues

- We struggle with chemical structure representation (e.g., human implicit depictions) that result in loss of scientist intention
- Different flavors/extensions of file formats (e.g., v2000 CTAB/MOL/SDF, SMILES) can conflict
- Different implementations/variations of common algorithms (e.g., aromaticity, depiction) create ambiguity or corruption

Chemical structure information can (irreversibly) change when exchanging between file formats and software packages

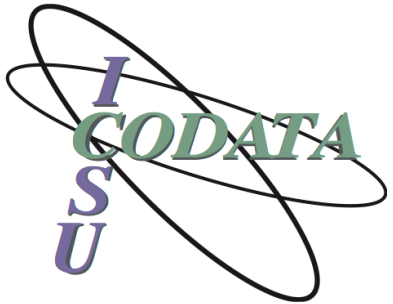
- Computed IUPAC InChI (and IUPAC systematic name) may differ after data exchange
- Scientists often unaware of what can occur during data exchange
  - Implicit (presentation) vs. explicit (machine interpretable) information
  - Loss of information (e.g., coordinates, relative stereo) when using different formats
- Software can help correct or warn the scientist of issues
  - Ambiguous stereo centers or missing explicit parity information
  - Tautomeric/resonance systems containing stereo centers/bonds
- Lack of agreed processing rules between software packages and publicly accessible databases
  - Same input can produce different output (still the same chemical structure?)
  - Proliferation of different structure variations of the same chemical with different InChI
- Free flow of chemical information makes establishment of ***best practices***, ***adherence to standards***, and ***scientist education*** of utmost importance

Digital standard organizations

https://rd-alliance.org/

INTERNATIONAL UNION OF PURE AND APPLIED CHEMISTRY

https://iupac.org/



http://www.codata.org/

https://www.w3.org/

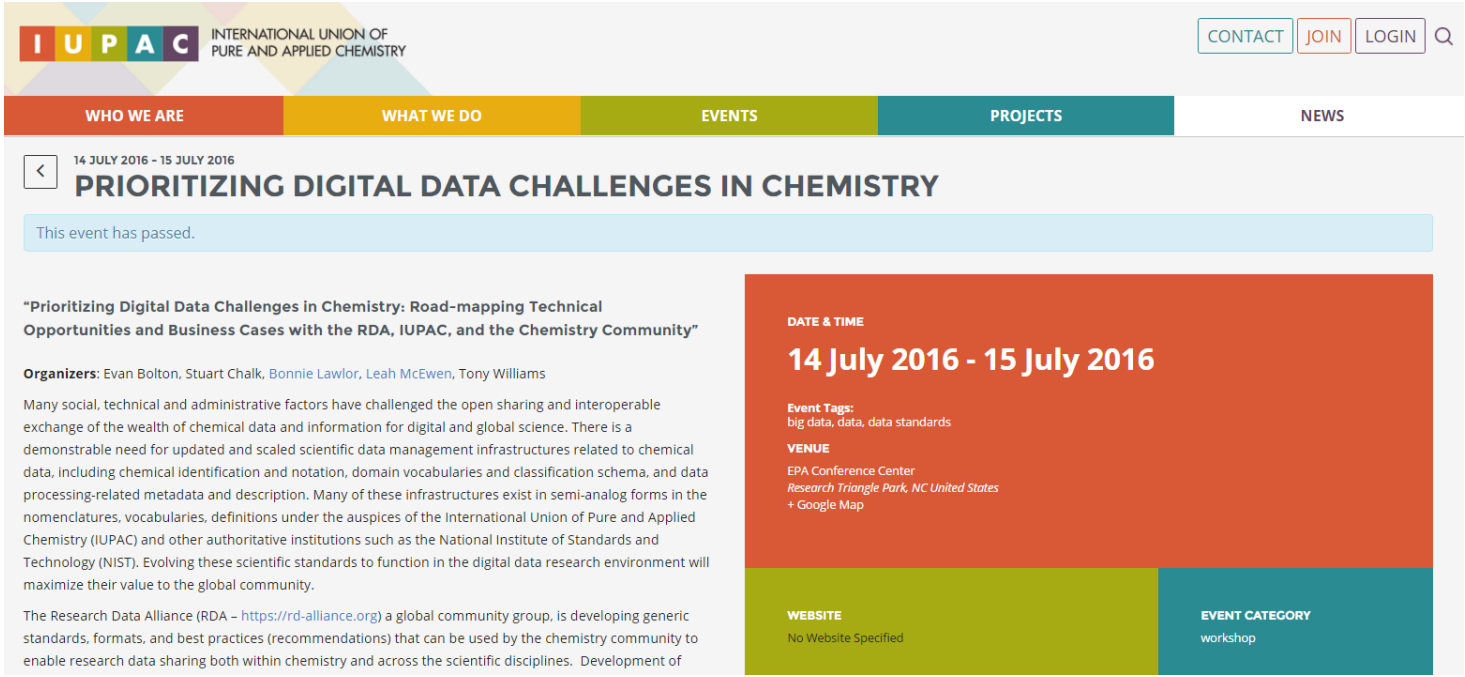
https://www.cdisc.org

Seven thematic areas of pain points

1. **Access** – *problems finding and/or getting access to (meta)data*
2. **Audience** – *who is looking at the data, how do they perceive it, perspectives, language of discipline*
3. **Chemical structure representation** – *inorganic, organometallic, large molecules, mixtures, chirality*
4. **Community** – *policies, procedures, and best practices needed to be adopted to move things forward*
5. **(Meta)Data** – *standardization, interoperability, metadata, gaps, scale, sharing, dark data*
6. **Ontology/Vocabulary** – *consensus on terms, maintenance, versions, optimal vocabularies, areas where needed*
7. **Tools to help (meta)data capture** – *adding metadata, feedback, consistency, synchronization*

From CINF Data Summit Mar 2016

RDA/IUPAC CPCDS Workshop at EPA



IUPAC INTERNATIONAL UNION OF PURE AND APPLIED CHEMISTRY

CONTACT JOIN LOGIN

WHO WE ARE | WHAT WE DO | EVENTS | PROJECTS | NEWS

14 JULY 2016 – 15 JULY 2016

PRIORITIZING DIGITAL DATA CHALLENGES IN CHEMISTRY

This event has passed.

"Prioritizing Digital Data Challenges in Chemistry: Road-mapping Technical Opportunities and Business Cases with the RDA, IUPAC, and the Chemistry Community"

Organizers: Evan Bolton, Stuart Chalk, Bonnie Lawlor, Leah McEwen, Tony Williams

Many social, technical and administrative factors have challenged the open sharing and interoperable exchange of the wealth of chemical data and information for digital and global science. There is a demonstrable need for updated and scaled scientific data management infrastructures related to chemical data, including chemical identification and notation, domain vocabularies and classification schema, and data processing-related metadata and description. Many of these infrastructures exist in semi-analog forms in the nomenclatures, vocabularies, definitions under the auspices of the International Union of Pure and Applied Chemistry (IUPAC) and other authoritative institutions such as the National Institute of Standards and Technology (NIST). Evolving these scientific standards to function in the digital data research environment will maximize their value to the global community.

The Research Data Alliance (RDA – https://rd-alliance.org) a global community group, is developing generic standards, formats, and best practices (recommendations) that can be used by the chemistry community to enable research data sharing both within chemistry and across the scientific disciplines. Development of

DATE & TIME

14 July 2016 - 15 July 2016

Event Tags:
big data, data, data standards

VENUE

EPA Conference Center
Research Triangle Park, NC United States
+ Google Map

WEBSITE
No Website Specified

EVENT CATEGORY
workshop

RDA/IUPAC Workshop at EPA

Chemical Structure Standardization Education and Outreach

(https://drive.google.com/open?id=14n8JMsXgN92-zUy1rBmwxiLDf42DL96BCz2pto80tCI)

Help chemists and other stakeholders to understand the issues of chemical structure standardization, its importance for chemical data exchange among humans and machines, and how these issues relate to their own work. Foundational activities will focus on identifying examples of how lack of standardization hinders research, articulating benefits to authors, readers, publishers, reviewers and educators and better understanding why chemists draw molecules the way they do and where the critical points exist in communicating chemistry among humans and machines.

Graphical Representation Guidelines Update

(https://drive.google.com/open?id=1SD_w5xbx6WpYYGRy7iJD1zlvgTr5o756qkfxj1f3F6Y)

Update the IUPAC Chemical Structure Drawing standards to consider machine interpretation of chemical depictions into chemical structures to prevent corruption of chemist intention (often apparent in chemical depiction). A primary goal is to harmonize the existing guidelines with machine interpretation, structure standardization, and nomenclature considerations in mind. These guidelines need to be taught to new chemists by educators to ensure the next generation of informatics savvy chemists will be minted.

RDA/IUPAC Workshop at EPA

Open Chemical Structure File Formats

(https://drive.google.com/open?id=11wNDCsdTDG_LUxwGpDHWkqWwuMQL3JxhIIvarwlzzsE)

Recommending and standardizing use of a small number of open chemical file formats/representations will improve interoperability and reduce error for chemistry data exchange. This strawman covers three commonly used *de facto* community based file formats, including SMILES, SMARTS, and CTAB. Each project proposal is of fairly limited scope and likely able to be completed in a comparatively short period of time with some degree of overlap in processes and concerns.

Normalizing Chemical Structures Best Practices

(https://drive.google.com/open?id=1S9wTAapSTLSABZ8TLZlxi9e_HThTefWrIxDf3zHFNjw)

Each organization and chemist has their own way of representing chemical structures. This makes chemical structure standardization necessary for data transfer. Standardization is performed in different ways, often with different goals in mind. These can be incompatible such that renormalization between disparate approaches can result in different standard InChI. For example, there are different approaches to SMILES aromaticity such that encoding in one and decoding in another can result in the loss of aromaticity and/or addition/subtraction of H_2. In addition, some tautomer 'normalizers' can change structure identity by making assumptions, such as the presence of acid or base or heating to 40°C, which transform the chemical substance into a related, separately isolatable structure. There is great potential to facilitate the sharing of chemical data if all standardizing software could be directed to operate in "sanctioned" fashion using best practices established for standardization of structures in chemical databases.

RDA/IUPAC Workshop at EPA

IUPAC Orange Book Ontology

(https://drive.google.com/open?id=1jRiJM048EyFfhE2u3ikI37wxIsG5rAaKZFirklNpA0g)

Develop a small scale ontology of chemical terms based on terms in IUPAC Orange Book as a case study. Foundational activities will look for example terminologies that have been converted to ontologies, identify where terms are currently being used and in what contexts, and look at relationships of those terms to others and potential differences in definitions. Terms will be transferred to a formal ontology in a plain bibliographic format, and a framework will be developed for augmenting the definition of terms to clarify the semantic meaning and context.

IUPAC Gold Book Data Structure

(https://drive.google.com/open?id=1hJdM7h90MBVLLUWBPtHe6cM-URJGi4zSlwYn8rXWNb8)

The IUPAC Gold Book is a valued compendium of terms sourcing from IUPAC published recommendations, including other Color Books and Pure and Applied Chemistry. The content is electronically accessible and linkable but not easily machine readable. This project is related to a current effort to extract the content data and term identifiers and migrate them into a more accessible and machine digestable format for increased usability.

Use Cases for Semantic Chemical Terminology Applications

(https://drive.google.com/open?id=1Ss5-qsIrgzSMTkcvEd52Iq-lwEYN2qCCN-BxN1ogGll)

This scoping project will focus on researching the current chemical data transfer and communication landscape for potential applications of semantic terminology. Example use cases might include text books, patents, article and data indexing, standard protocols, experimental literature, published ontologies and thesauri with chemical terms, dictionaries for text mining, etc. Initial activities will analyze citations to terminology in the IUPAC Color Books (including the Gold Book) and Pure and Applied Chemistry.

Workshop on future of InChI

AUG
16

Status and Future of the IUPAC InChI: context and use cases

by InChI Trust / IUPAC

Free

REGISTER

http://www.inchi-trust.org/
Aug. 16-18, 2017

DESCRIPTION

This conference / workshop builds on recent events to expand and prioritise future developments for the IUPAC InChI standard for chemical structure representation.

The event will include a mix of themed talks and working breakout sessions, finishing with summaries and actions.

Working groups will cover InChI extensions for organometallics, tautomerism, reactions, mixtures and large molecules amongst others. Applications and challenges for InChI implementation in QR codes and cross database resolvers will also be covered, as well as the challenges in chemical structure standardization and for open molecular file formats.

See the InChI Trust or IUPAC InChI site for background on the InChI

DATE AND TIME

Wed, Aug 16, 2017, 8:30 AM –
Fri, Aug 18, 2017, 12:30 PM EDT
Add to Calendar

LOCATION

National Institutes of Health
9000 Rockville Pike
Bethesda, MD 20892
View Map

TAKE AWAY SUMMARY

Chemical information is not easy

- How do you describe a chemical substance?
  - No standards, meta data associated with chemical representation (e.g., purity)
- How do you describe a chemical mixture?
  - No standards (InChI?), often free-form text
- How do you describe a bioactivity?
  - Emerging standards, not widely adopted
    - Minimum Information About a Bioactive Entity (MIABE)
- How do you draw a chemical structure?
  - IUPAC Graphical Representation Standards for Chemical Structure Diagrams
    - Not widely adopted, large pre-existing corpus
    - Not tuned for computer understanding

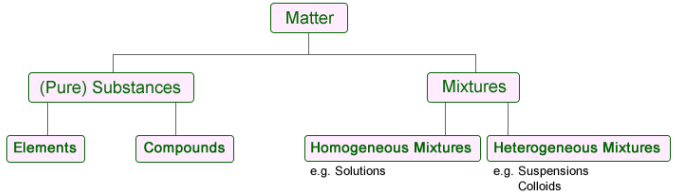


Image credit: http://www.ivyroses.com/Chemistry/GCSE/What-is-a-substance.php

Chemical data issues are fundamental

- Many answers to “How do describe my substance and its data?”
- Information is geared towards humans
- Need computer understanding of information
- Chemical structure representation is insufficient
- Publically available chemical information is heavily fragmented
- Lots of data links (use case dependent, relevancy, etc.)

Summary

- There are many opportunities to improve the quality, quantity, variety, relevancy, and integration of (open) chemical data and chemical knowledge
- Knowledge representation in chemistry helps to enable computer understanding of the domain
- Community efforts are underway to help build and harmonize chemical ontologies

PubChem Crew …

Evan Bolton

Jie Chen

Tiejun Cheng

QingLiang Li

Asta Gindulyte

Jane He

Siqian He

Sunghwan Kim

Ben Shoemaker

Paul Thiessen

Bo Yu

Leonid Zaslavsky

Jian Zhang

Special thanks to the NCBI Help Desk, especially Rana Morris, and past PubChem group members.

Special thanks

- Software collaborators
  - NextMove Software (Roger Sayle, Daniel Lowe, Noel O'Boyle, John May)
  - Xemistry GmbH (Wolf D. Ihlenfeldt)
  - OpenEye Scientific Software
- Chemical Health and Safety collaborators
  - Especially: Leah McEwen (Cornell U.), Ralph Stuart (Keene State College)
- Chemical Ontology Collaborators
  - Especially: Stephen Boyer, Yannick Djoumbou, Lutz Weber
- All PubChem Contributors and Collaborators
- Intramural Research Program of the NIH, National Library of Medicine

Have any questions?