

PubChem data issues: Chemistry, bioactivities, etc.

Evan Bolton, Ph.D.

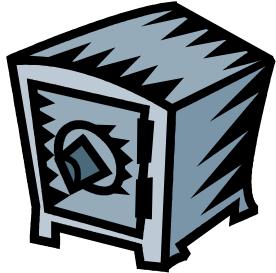
NCBI/NLM/NIH

Open PHACTS / GEN2PHEN Workshop

September 19, 2011

bolton@ncbi.nlm.nih.gov





What is PubChem?

- An open archive
 - anyone can contribute
 - chemical structures
 - synonyms
 - comments
 - biological experiments
 - cross references
 - records versioned
 - URLs
 - links external resources
 - voluntary data push
 - automated updates
- A public resource
 - anyone can access
 - data downloadable
 - search, browse, retrieve
 - integrated
 - literature
 - sequences, protein 3-D
 - analysis capabilities
 - programmatic layers
 - PUG, PUG/SOAP
 - Entrez Utilities
 - URL-based interfaces



What is PubChem?

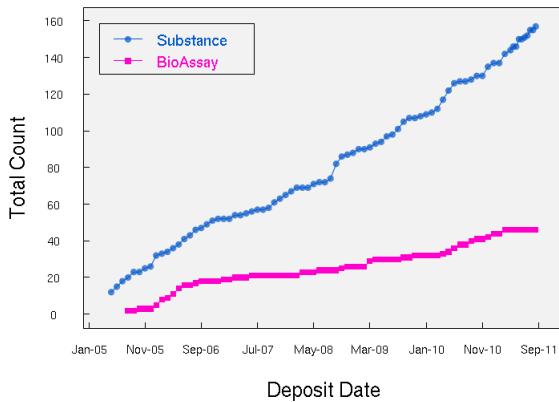


- An open archive
 - anyone can contribute
 - chemical structures
 - synonyms
 - comments
 - biological experiments
 - cross references
 - records versioned
 - URLs
 - links external resources
 - voluntary data push
 - automated updates
- A public resource
 - anyone can access
 - data downloadable
 - search, browse, retrieve
 - integrated
 - literature
 - sequences, protein 3-D
 - analysis capabilities
 - programmatic layers
 - PUG, PUG/SOAP
 - Entrez Utilities
 - URL-based interfaces

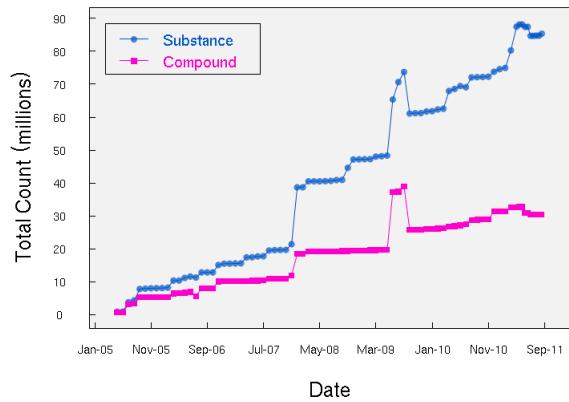


PubChem has a lot of data...

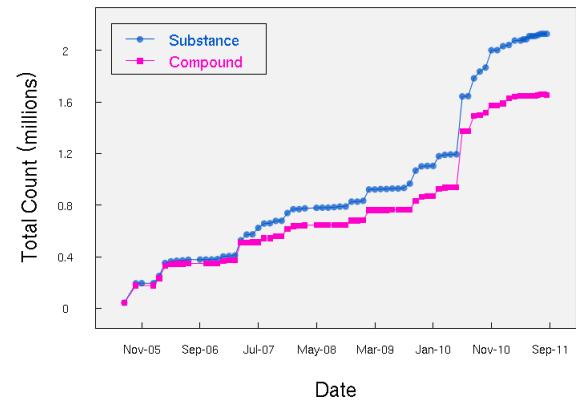
Depositors



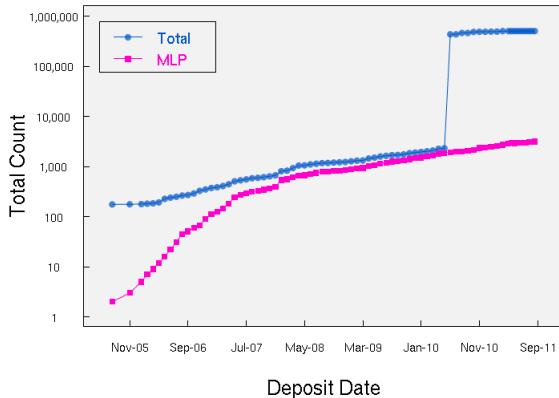
Chemicals



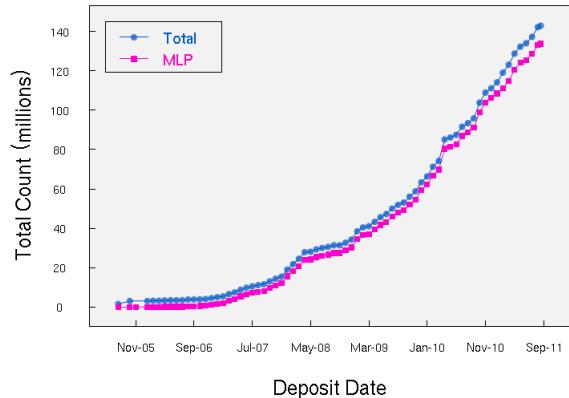
Tested Chemicals



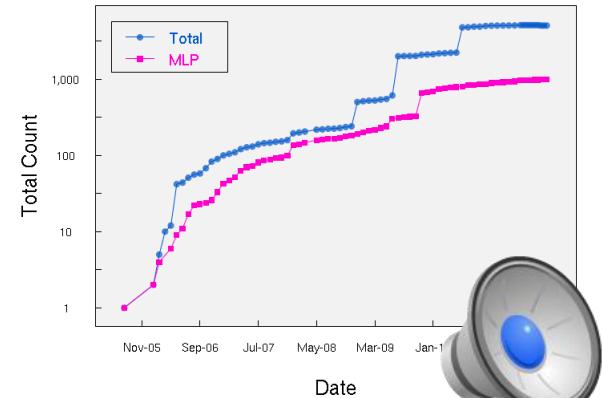
Biological Assays



Bioactivities



Protein Targets



PubChem is one resource amongst others...

NCBI

Entrez, The Life Sciences Search Engine

HOME | SEARCH | SITE MAP PubMed All Databases Human Genome GenBank Map Viewer BLAST

Search across databases Help

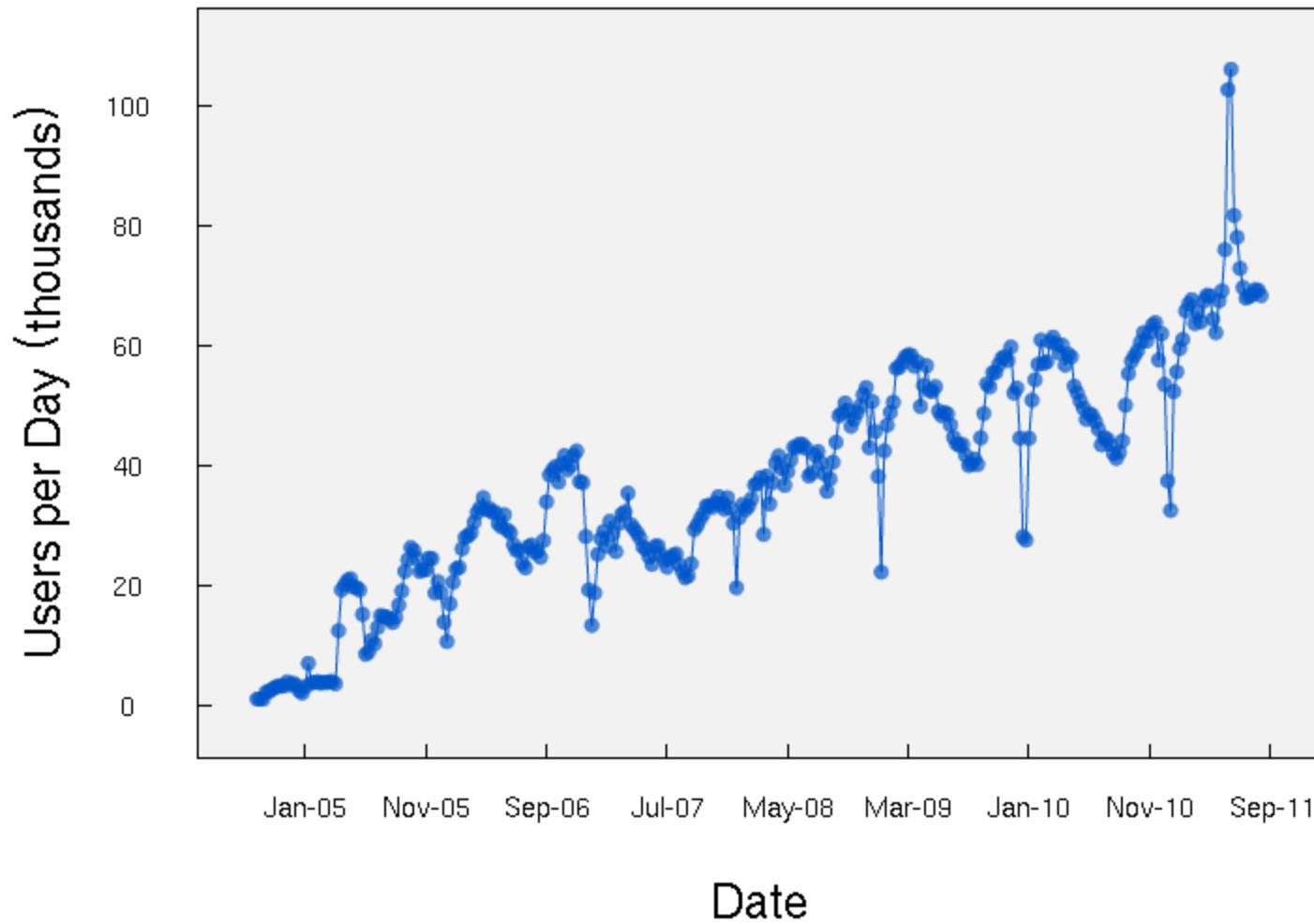
- Result counts displayed in gray indicate one or more terms not found

48172 PubMed: biomedical literature citations and abstracts	2321 Books: online books
24538 PubMed Central: free, full text journal articles	52 OMIM: online Mendelian Inheritance in Man
20 Site Search: NCBI web and FTP sites	
143 Nucleotide: Core subset of nucleotide sequence records	520 dbGaP: genotype and phenotype
none EST: Expressed Sequence Tag records	none UniGene: gene-oriented clusters of transcript sequences
none GSS: Genome Survey Sequence records	4 CDD: conserved protein domain database
143 Protein: sequence database	none UniSTS: markers and mapping data
none Genome: whole genome sequences	none PopSet: population study data sets
60 Structure: three-dimensional macromolecular structures	8799 GEO Profiles: expression and molecular abundance profiles
none Taxonomy: organisms in GenBank	536 GEO DataSets: experimental sets of GEO data
none SNP: single nucleotide polymorphism	none Epigenomics: Epigenetic maps and data sets
none dbVar: Genomic structural variation	none Cancer Chromosomes: cytogenetic databases
85 Gene: gene-centered information	1906 PubChem BioAssay: Bioactivity screens of chemical substances
none SRA: Sequence Read Archive	69 PubChem Compound: unique small molecule chemical structures
15 BioSystems: Pathways and systems of interacting molecules	none PubChem Substance: deposited chemical substance records
2 HomoloGene: eukaryotic homology groups	none Protein Clusters: a collection of related protein sequences
none GENSAT: gene expression atlas of mouse central nervous system	1 OMIA: online Mendelian Inheritance in Animals
none Probe: sequence-specific reagents	none BioSample: biological material descriptions
1 BioProject: aggregated biological research project data	
330 NLM Catalog: catalog of books, journals, and audiovisuals in the NLM collections	33 MeSH: detailed information about NLM's controlled vocabulary

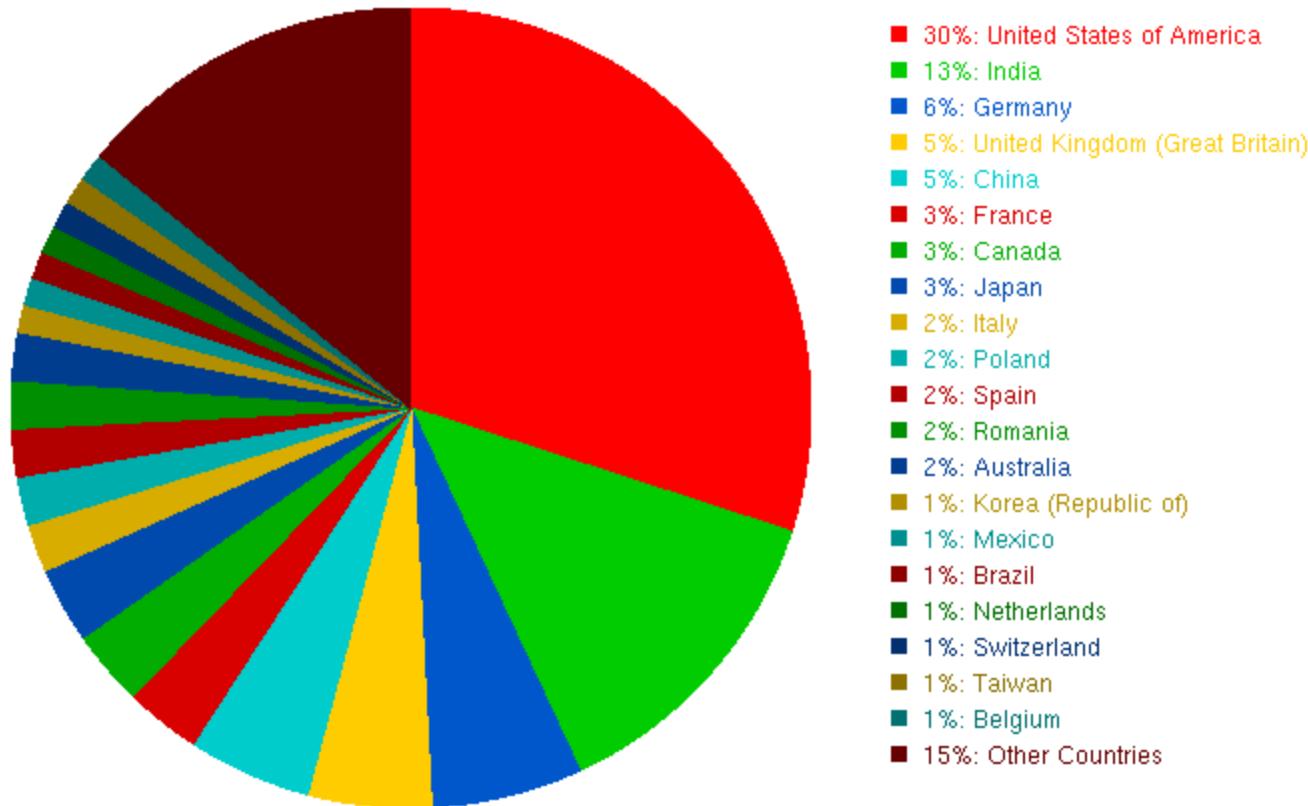
A red box highlights the row for PubChem BioAssay: Bioactivity screens of chemical substances.



PubChem is heavily used...



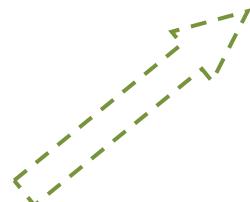
PubChem is a global resource...



Interactive usage by country
(Jul 15 2010 – Aug 15 2010)



PubChem data relationships...



Depositor provided
Primary accession **AID**

Unique chemical structure
content of PubChem

Mixture
Salt
Parent
Components

“Identity groups”

Exactly Same
Same Isotope
Same Stereo
Same Connectivity
Tautomers



Automated structure processing...



- **Verification**

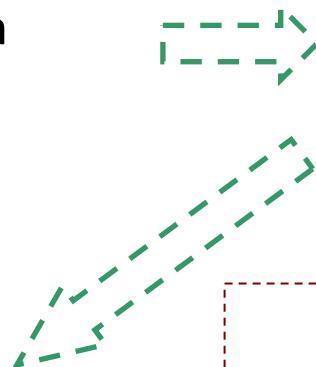
- Atom element
- Implicit hydrogen
- Functional group
- Valence

- **Standardization**

- Tautomer invariance
- Aromaticity detection
- Stereochemistry
- Explicit hydrogen

- **Calculation**

- Coordinates
- Properties
- Descriptors



- **Components**

- Isolate covalent units
- Neutralize (+/- proton)
- **Reprocess**
- Detect unique





PubChem data access...

- Interfaces
 - text/numeric search
 - fielded/range search
 - precomputed similarities
 - 2-D, 3-D, identity groups
 - inter-database links
 - biomedical literature
 - MeSH ontology
 - biological roles
 - protein 3-D
 - pathways
 - external resource links
- Tools
 - bioactivity analysis
 - chemical clustering
 - chemical structure search
 - data download
 - FTP site
 - heatmap analysis
 - integrated 3-D layer
 - similarity computation
 - source summary
 - structure normalization



PubChem data access...



- Interfaces
 - text/numeric search
 - fielded/range search
 - precomputed similarities
 - 2-D, 3-D, identity groups
 - inter-database links
 - biomedical literature
 - MeSH ontology
 - biological roles
 - protein 3-D
 - pathways
 - external resource links
- Tools
 - bioactivity analysis
 - chemical clustering
 - chemical structure search
 - data download
 - FTP site
 - heatmap analysis
 - integrated 3-D layer
 - similarity computation
 - source summary
 - structure normalization



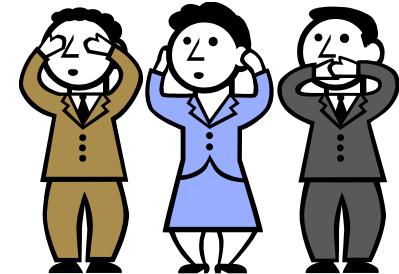


The **sad** state of
chemical information

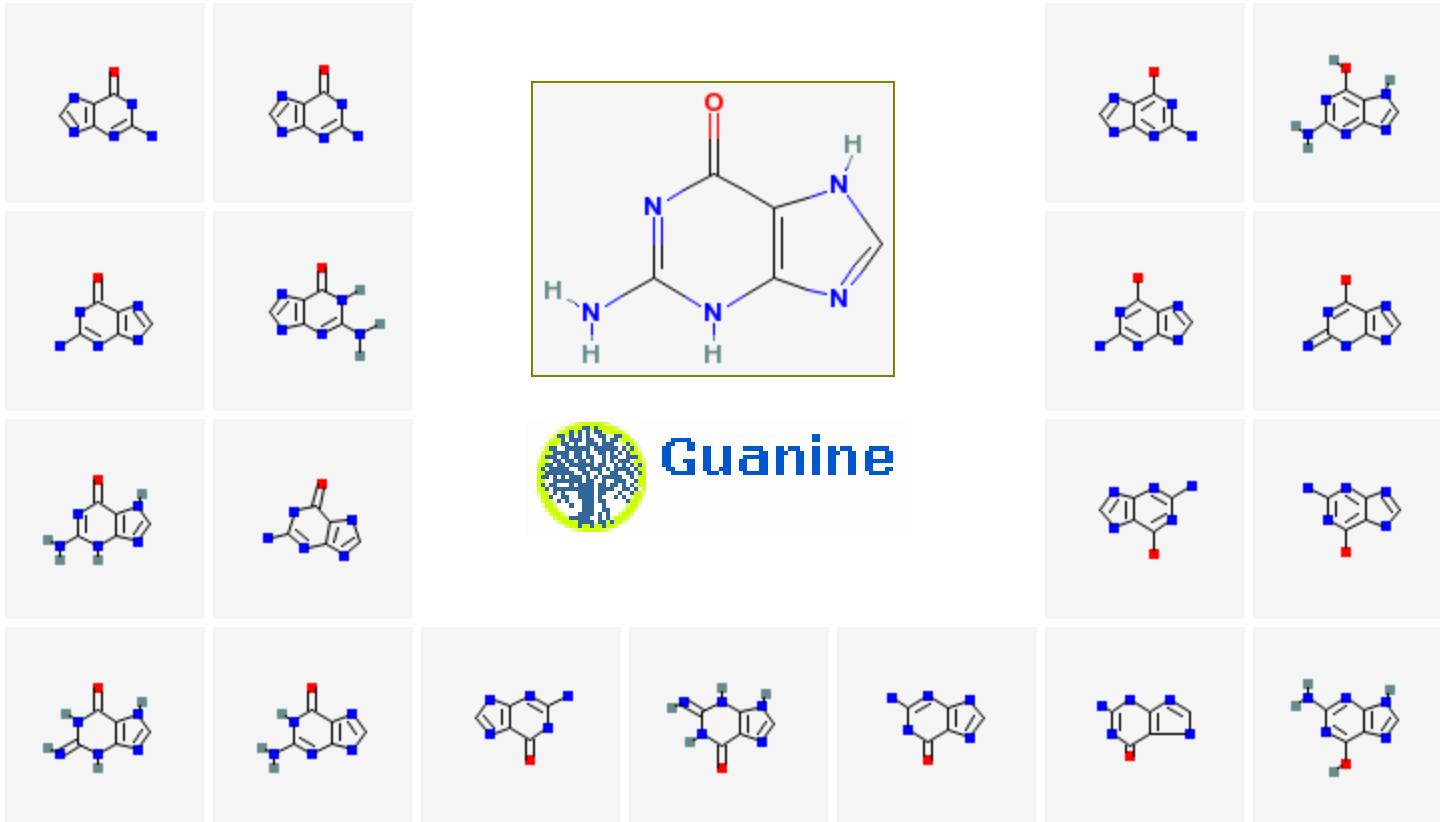


Let's talk chemical information...

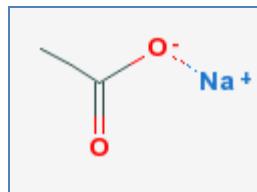
- **No “Global” rules or standards**
 - based on individual organizational needs
 - often based on individual preferences
 - depictions of chemical structures
- **PubChem accepts data from many organizations**
 - conflicting “business rules”
 - previously unseen data representation schemes
 - combinatorial ways of drawing the same structure



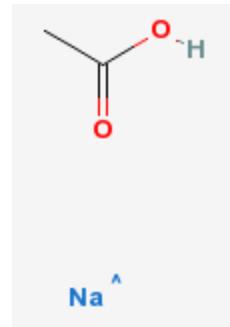
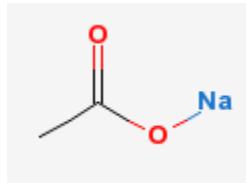
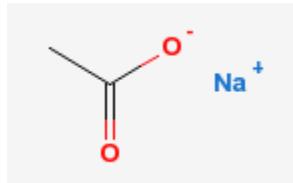
A chemical structure may be represented in many different ways



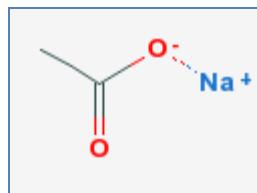
A chemical structure may be represented in many different ways



 Sodium Acetate

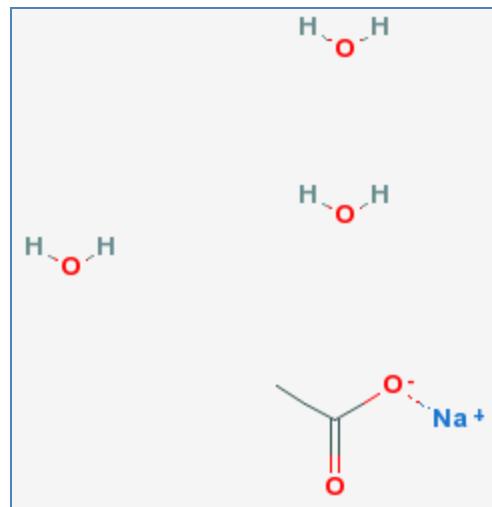


What do you mean by “sodium acetate”?



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.

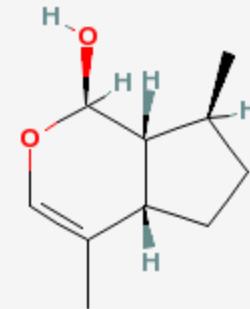


What did you mean by that?

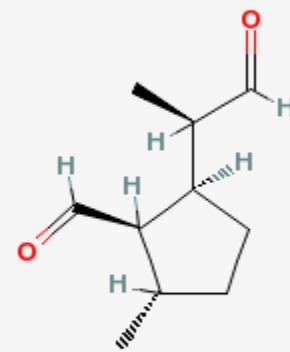
- Case Study:

(+)-Iridodial

Defense chemicals from
abdominal glands of 13
rove beetle species of
subtribe Staphylinina



Ring Closed



Ring Open



What do you mean by that?

- “C” means?
 - form of carbon?
 - which one?
 - diamond?
 - graphite?
 - coal?
 - graphene?
 - charcoal?
 - carbon black?
 - nanotube?
 - methane?



C

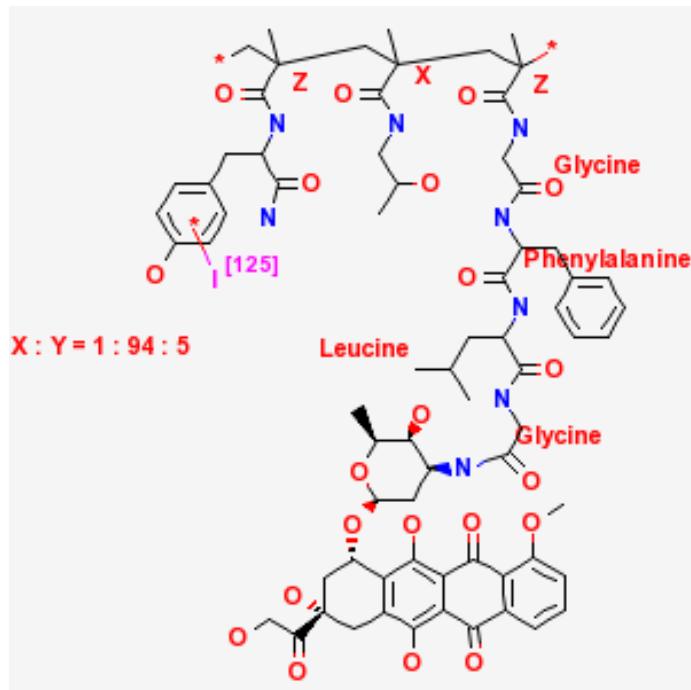


Stereochemistry ← **Big** problem

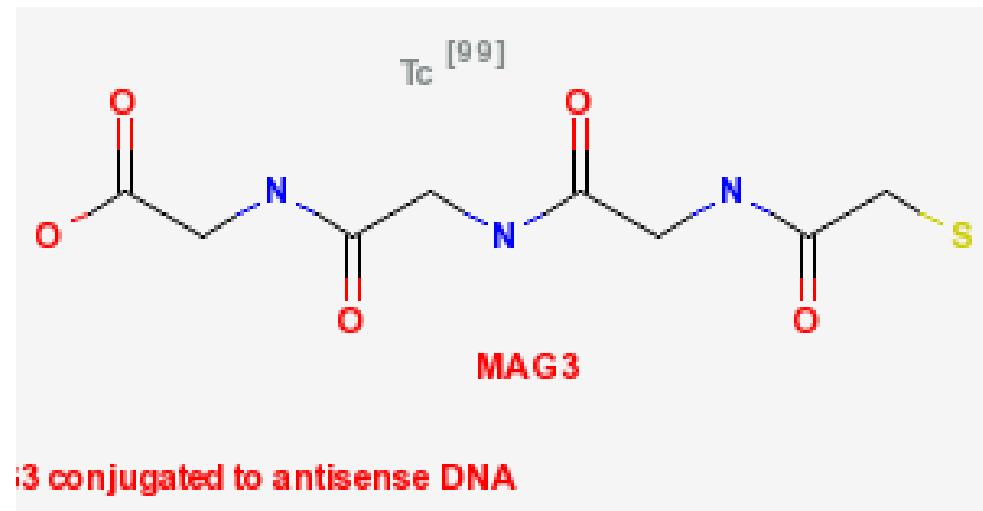
- Import issues
 - Often obtained by perception of atom coordinates
 - Coordinates or stereo wedges may be ambiguous
 - Inconsistency between software packages for same file
- Export issues
 - Improper/inconsistent use of file format
- Format conversion adds/removes/changes stereo
- Relative stereochemistry improperly treated
- Depiction vs. machine readable
- Curated data may become corrupted!



Do we have a “defined” structure?

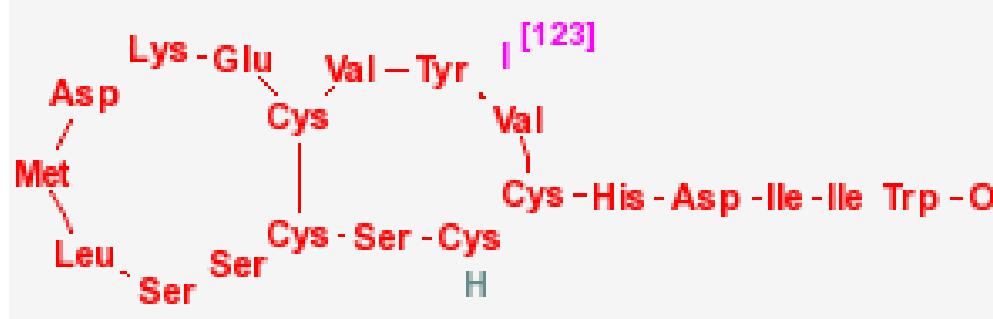


SID: 5



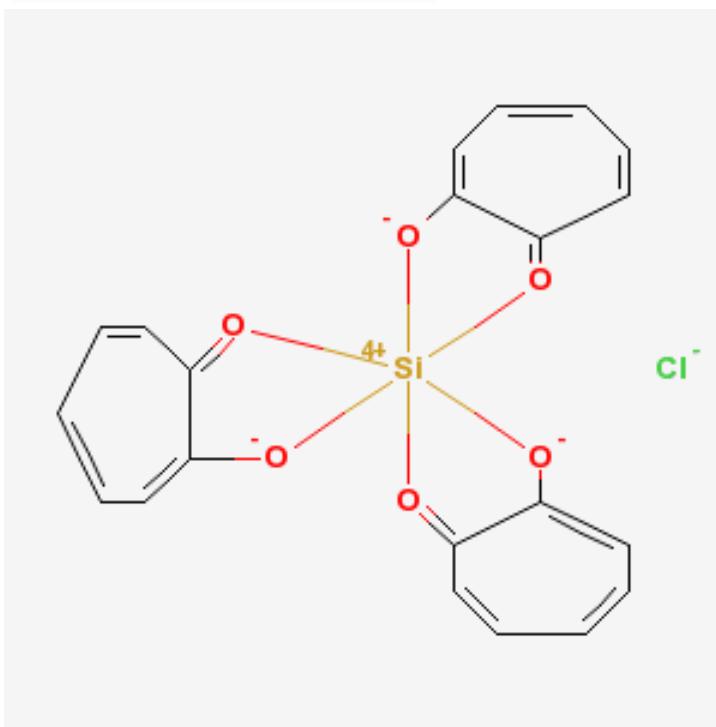
MAG3

i3 conjugated to antisense DNA

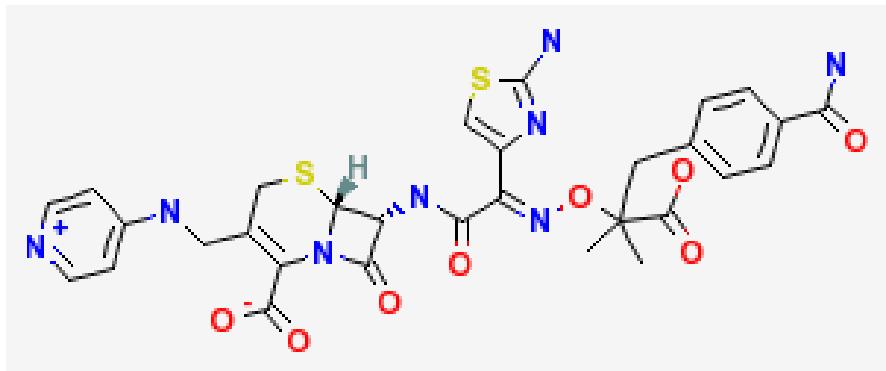


SID: 8

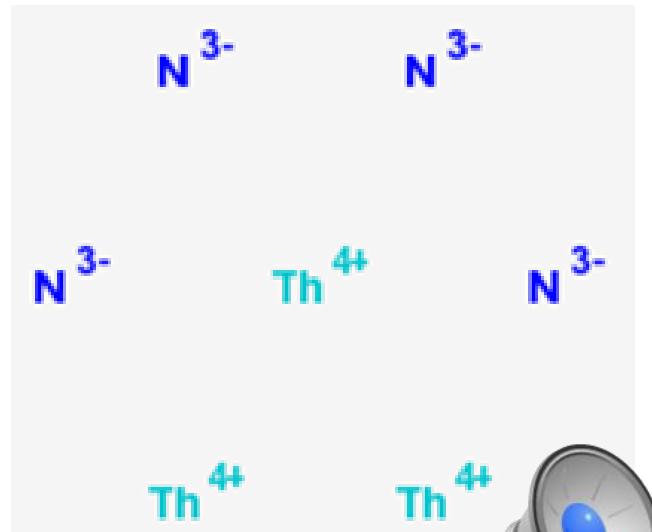
Is the structure reasonable?



SID: 772267



SID: 743503



What is data quality?

Ideal

- Validated
- Available
- Complete
- Succinct
- Useful
- Facile
- Seamless
- Happy user



Usually found

- Best guess
- Something close
- Fragmented
- Verbose
- Might help
- Lots of work
- Issues
- Frustrated user



What is data quality?

Ideal

- Validated
- Available
- Complete
- Succinct
- Useful
- Facile
- Seamless
- Happy user



Usually found

- Best guess
- Something close
- Fragmented
- Verbose
- Might help
- Lots of work
- Issues
- Frustrated user



PubChem Crew ...

Steve Bryant

Jie Chen

Tiejun Chen

Lewis Geer

Asta Gindulyte

Volker Hahnke

Lianyi Han

Jane He

Siqian He

Kenneth Karapetian

Sunghwan Kim

Qingliang Li

Ben Shoemaker

Tugba Suzek

Paul Thiessen

Jiyao Wang

Yanli Wang

Jewen Xiao

Bo Yu

Jian Zhang

Jun Zhang

