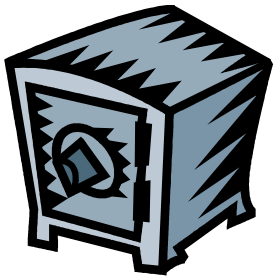
PubChem: small molecules and their biological activities

Evan Bolton, Ph.D.
NCBI/NLM/NIH
Feb. 5, 2013

PubChem home page…

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



What is PubChem?

- An open archive
  - anyone can contribute
    - biological experiments
    - chemical structures
    - substance names
    - cross-reference IDs
    - link-back URLs
    - textual commentaries
    - hierarchical annotation
  - records are versioned
  - voluntary data push
  - interconnects chemical biology resources
- A public resource
  - anyone can access
    - search, subset, select, analyze, download
  - integrated
    - biomedical literature, sequences, pathways, patents, ontologies, etc.
  - programmatic layers
    - URL-based interfaces
    - NCBI Entrez Utilities
    - PUG [XML, SOAP, REST]
    - JavaScript display ‘widgets’ [display and data]

PubChem original primary goal…

… to be an on-line resource providing comprehensive information on the biological activities of small molecules

where “small” means less than 1,000 atoms or bonds

PubChem primary goal…

… to be an on-line resource providing comprehensive information on the biological activities of *substances*

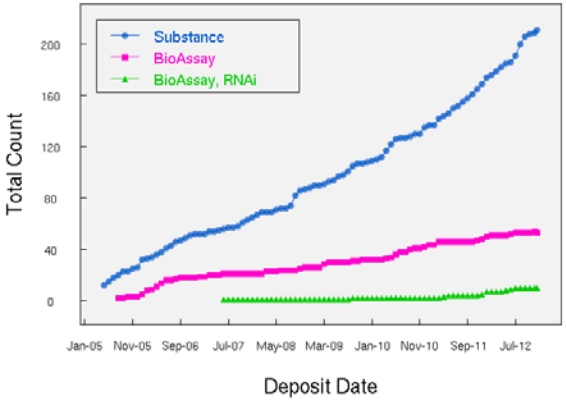
where "substance" means any biologically testable entity

Small molecules, RNAi's, etc.

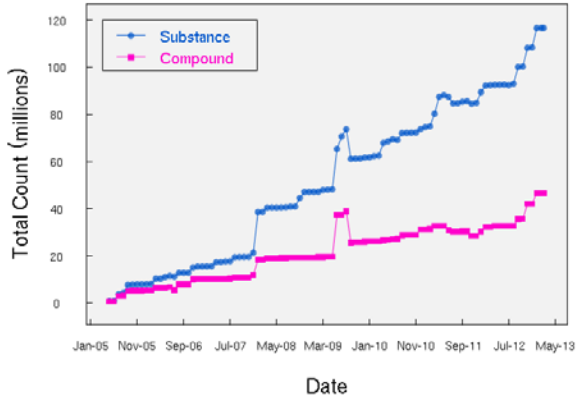
PubChem is an information hub…

PubChem contains a lot of data…

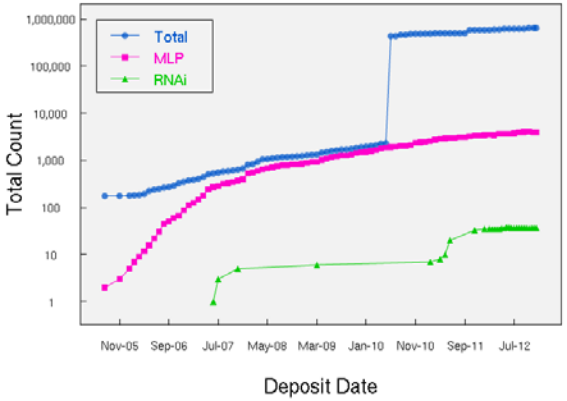
Contributors



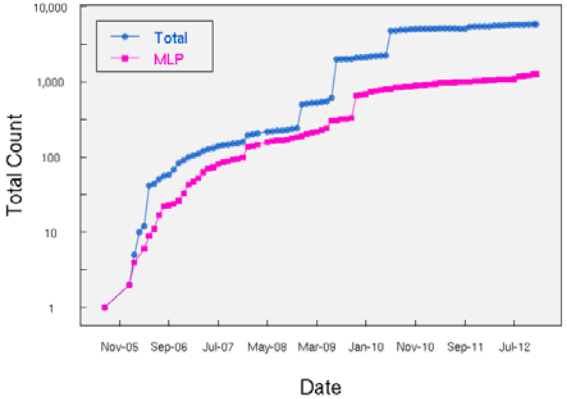
Chemicals



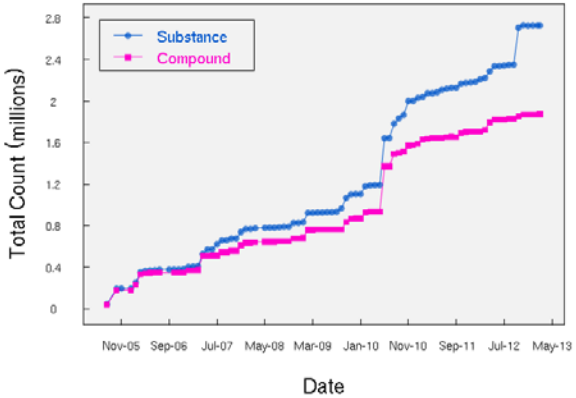
Biological Assays



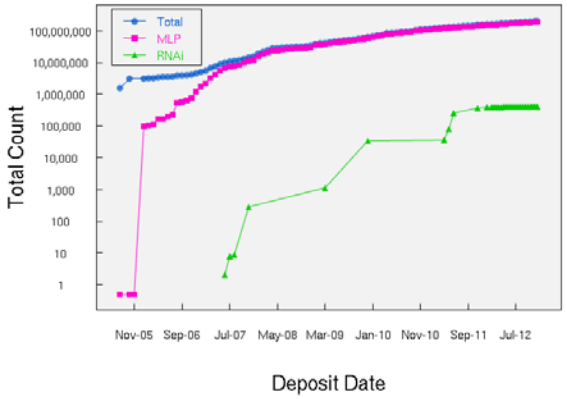
Protein Targets



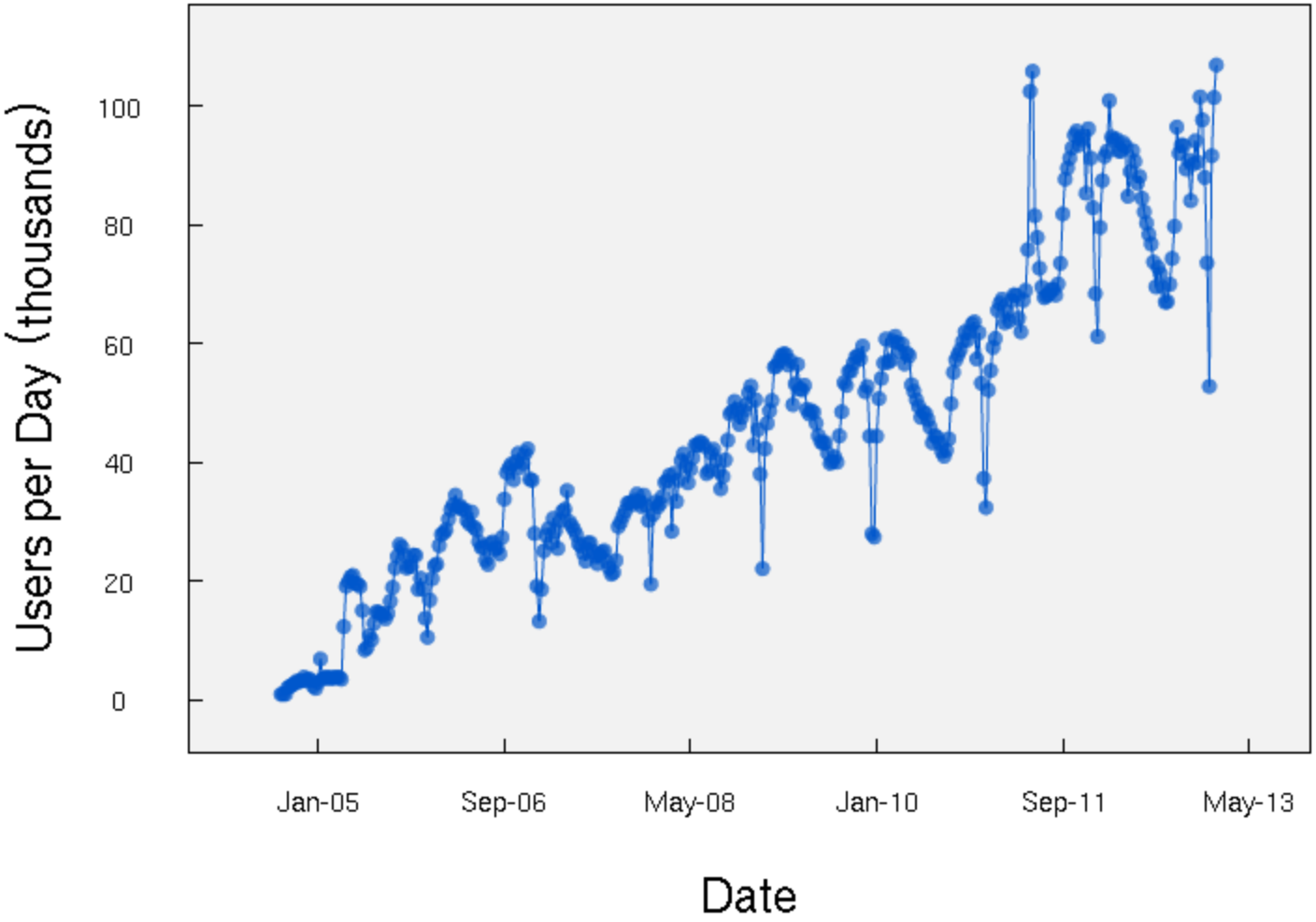
Tested Chemicals



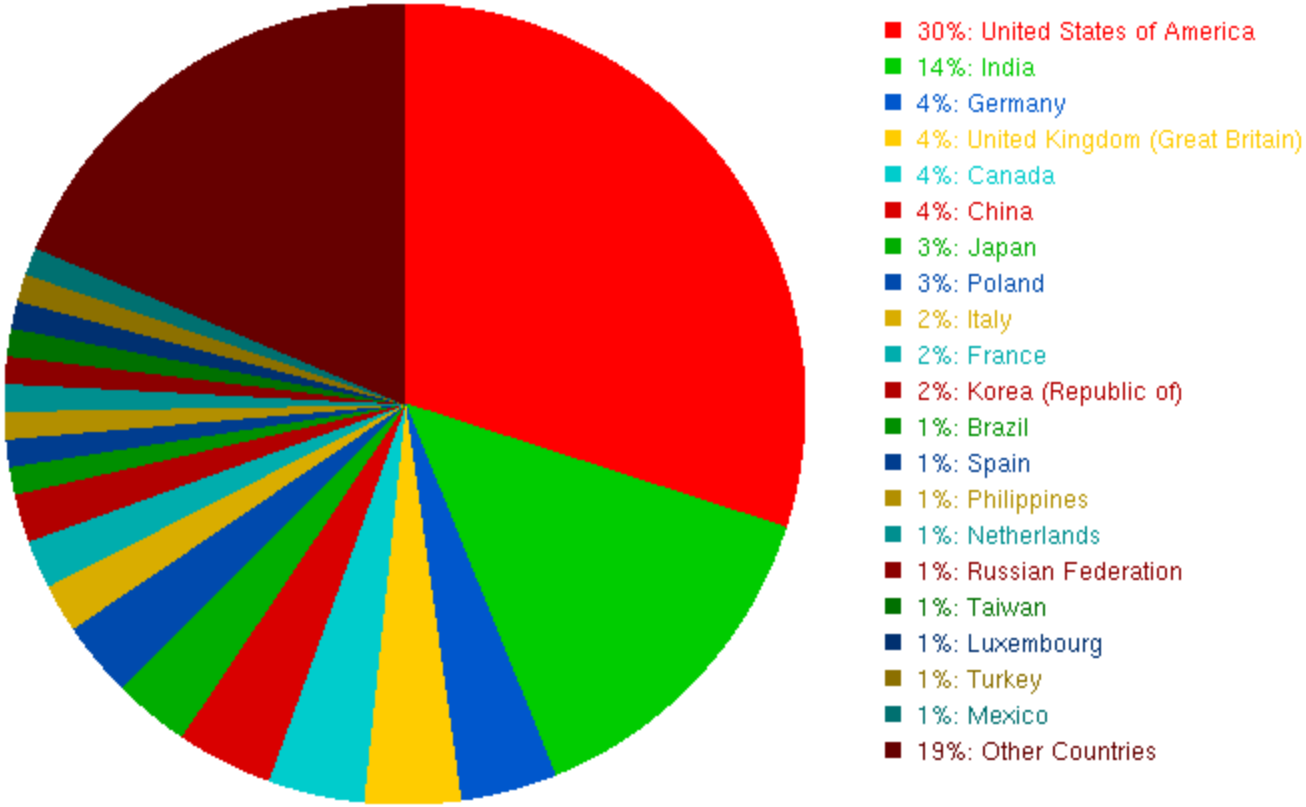
Bioactivity Results



PubChem is heavily used…



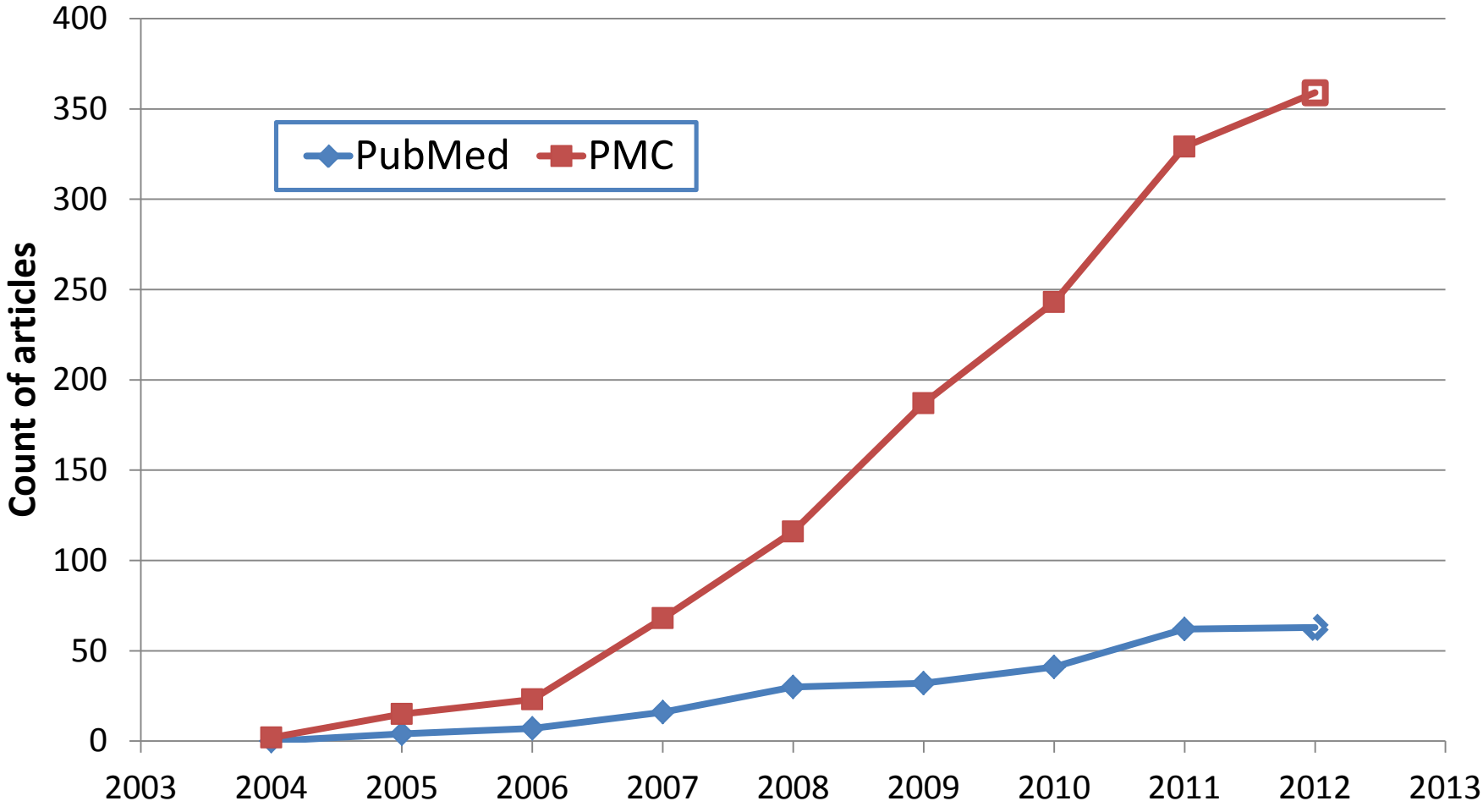
PubChem is a global resource…



Interactive usage by country in the last four weeks
(Oct 2012 – Nov 2012)

Publications mentioning PubChem…

(per year results … +1,100 in total)



PubMed abstracts and full text PubMedCentral (PMC) articles

[Note: 2012 PMC results are incomplete due to rolling 1 year article embargo]

PubChem contributed data…

Primary accession **SID**

PubChem Substance

Substances can be anything in a test tube:
- a discrete small molecule (e.g., aspirin)
- a complex mixture (e.g., plant extract)
- a biomolecule (e.g., RNAi)

Each depositor controls their own version of a substance records independent of others (i.e., can be highly redundant)

PubChem BioAssay

Assays test substances and are essentially a table of readouts

Primary accession **AID**

The nature of a test can be highly variable:
- phenotypic, whole cell, defined target, high throughput, dose-response, validation, physical property measurement, etc.

PubChem data relationships…

Depositor provided
Primary accession **SID**

PubChem Substance

Derived content
Primary accession **CID**

PubChem Compound

Unique chemical structure content of PubChem

Substances not well-defined are not a “compound”

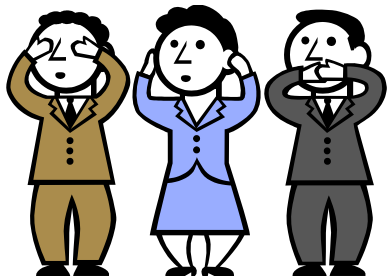
Inconsistent chemical representations across depositors are a challenge

PubChem BioAssay

Depositor provided
Primary accession **AID**

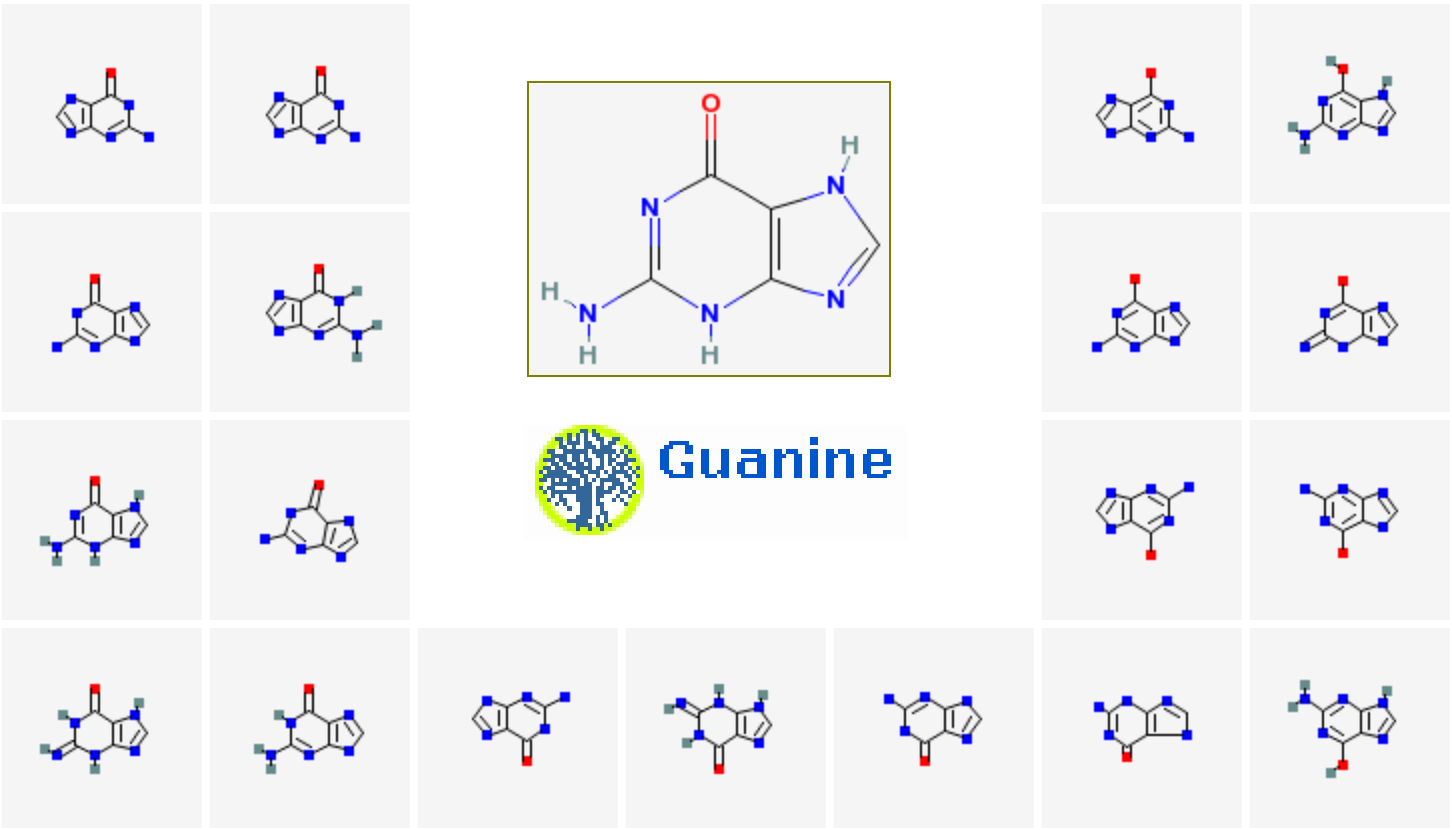
Let's talk chemical information…

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

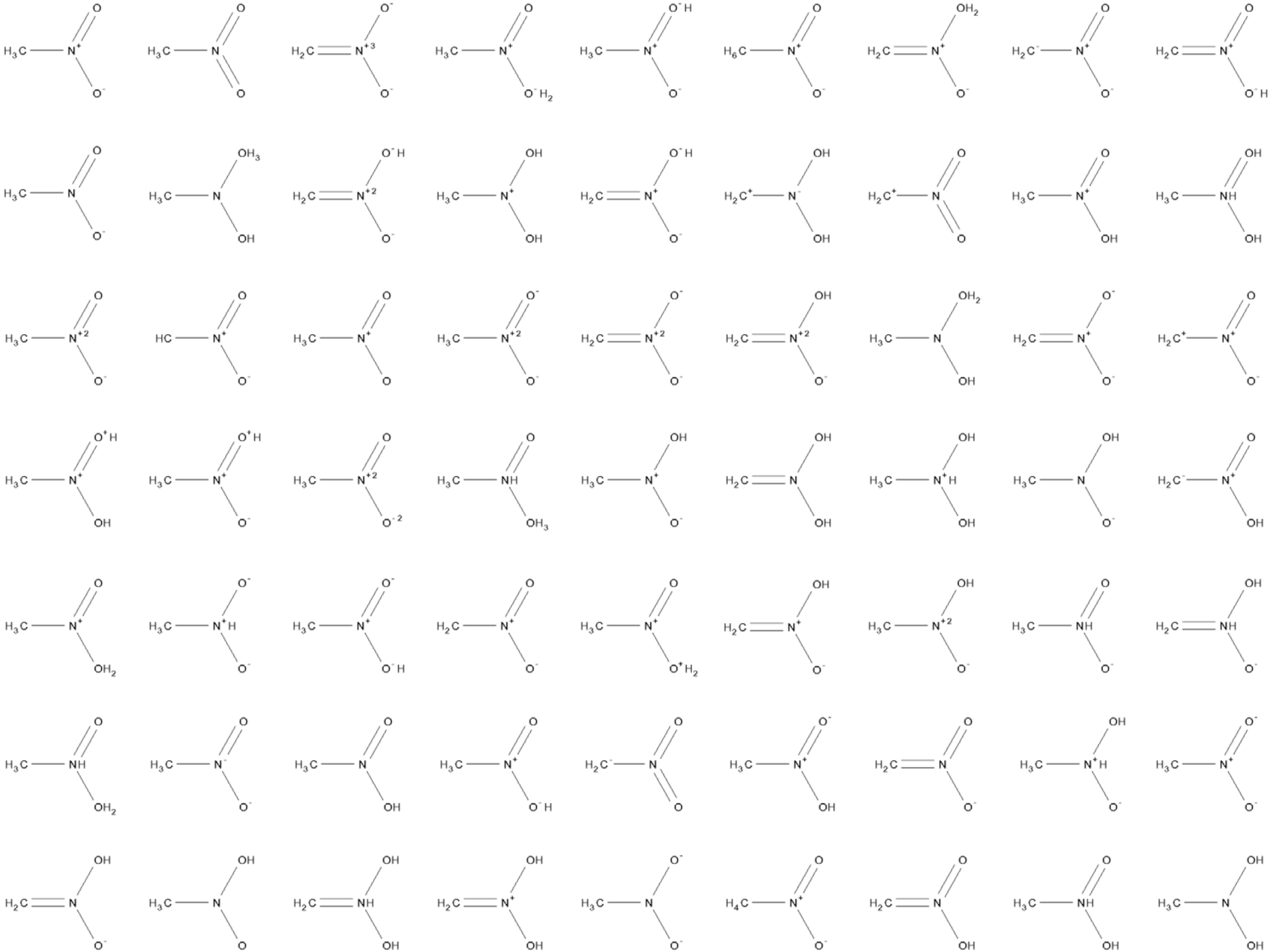


- **PubChem accepts data from many organizations**
  - conflicting "business rules"
  - previously unseen data representation schemes
  - combinatorial ways of drawing the same structure

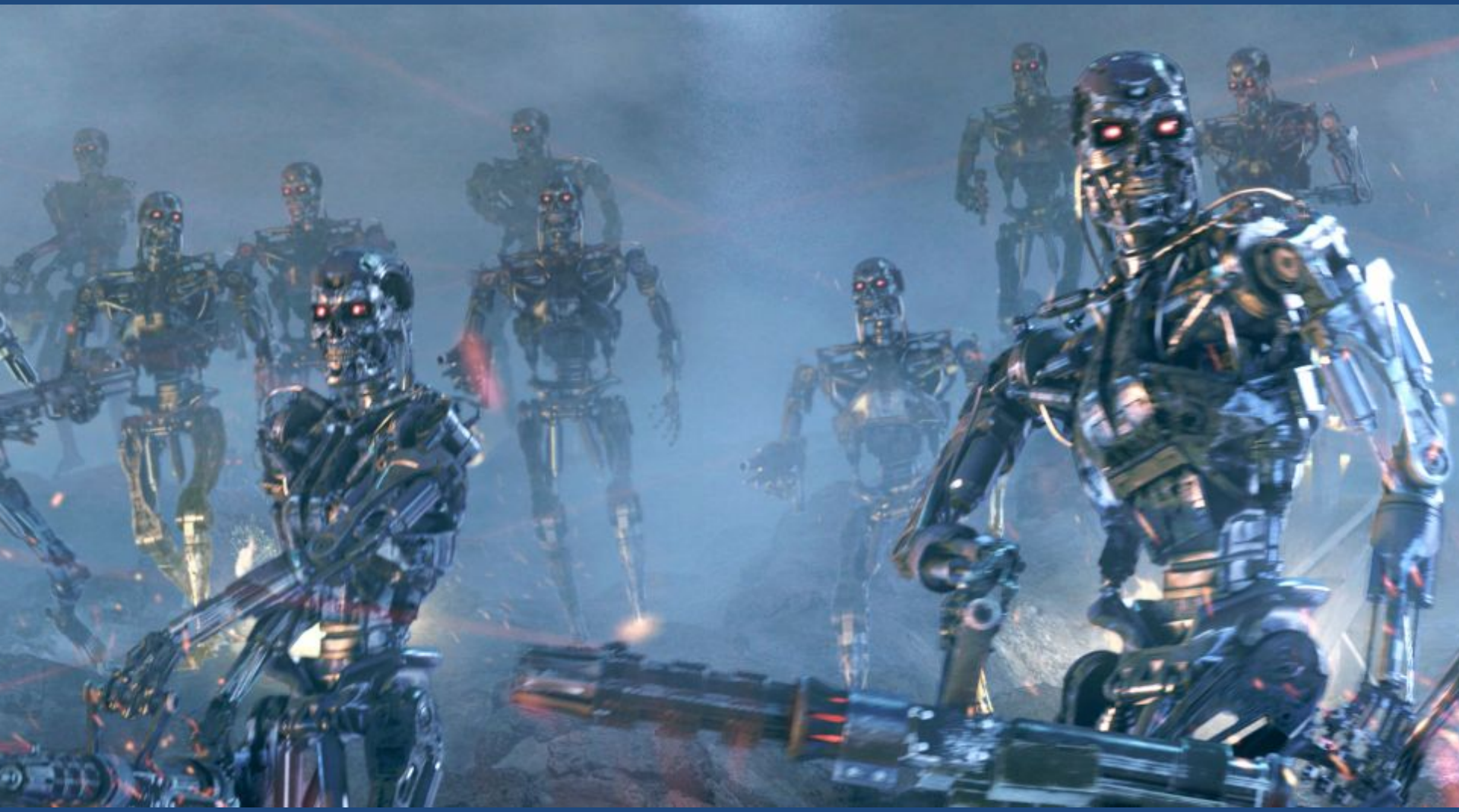
A chemical structure may be represented in many different ways



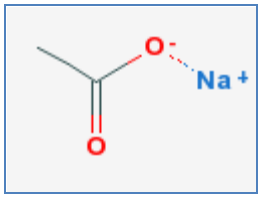
Tautomers and resonance forms of same chemical structure are prolific



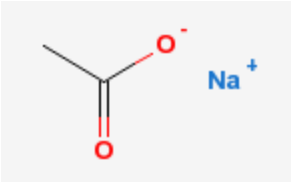
We made them… and now they are everywhere… “RoboChemistry”…

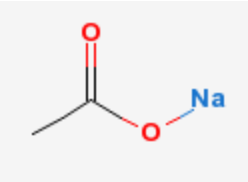


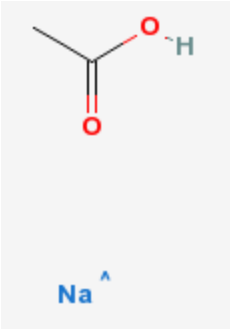
A chemical structure may be represented in many different ways



Sodium Acetate







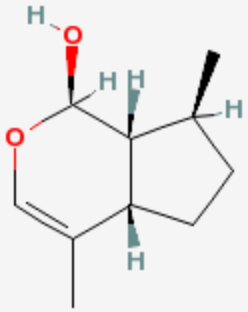
Salt-form drawing variations are common

What did you mean by that?

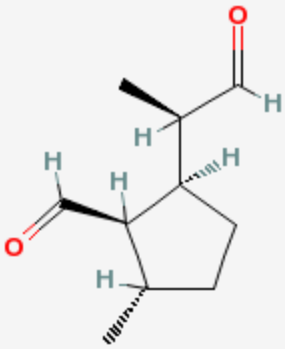
- Case Study:

(+)-Iridodial

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



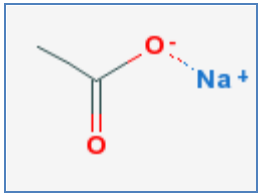
Ring Closed



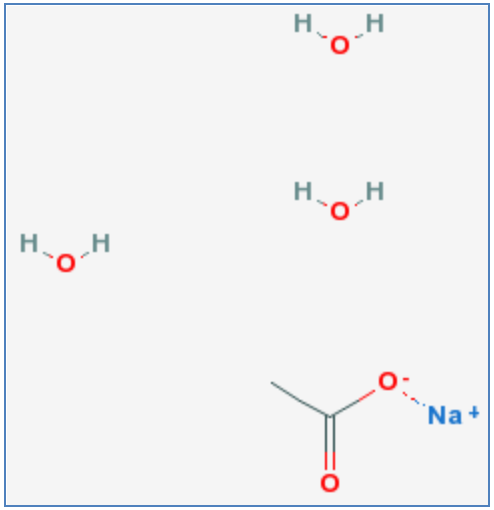
Ring Open

Some chemical species naturally have multiple forms

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.



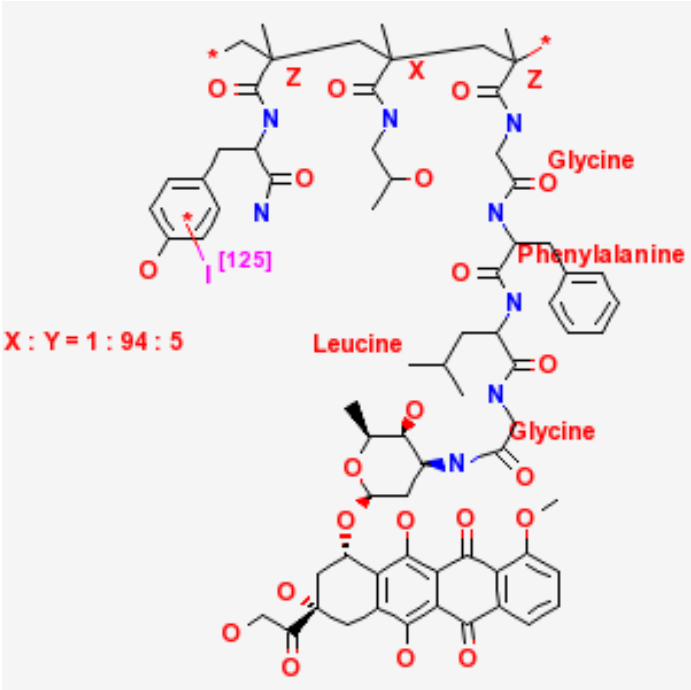
Chemical meaning of a substance may change upon context

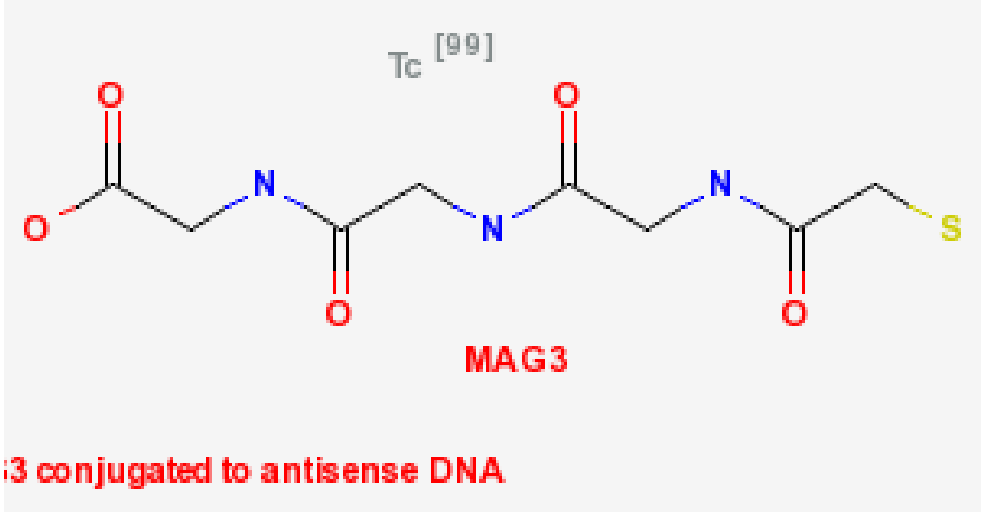
Difficulty in representing some substances using a chemical structure…

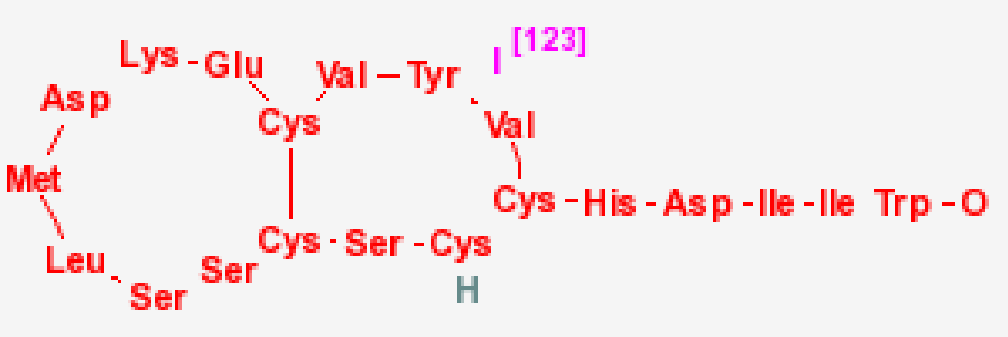
- "C" means?
  - form of carbon?
    - which one?
      - diamond?
      - graphite?
      - coal?
      - charcoal?
      - graphene?
      - carbon black?
      - nanotube?
  - methane?

Even seemingly trivial cases can be ambiguous or chemical representation singularities

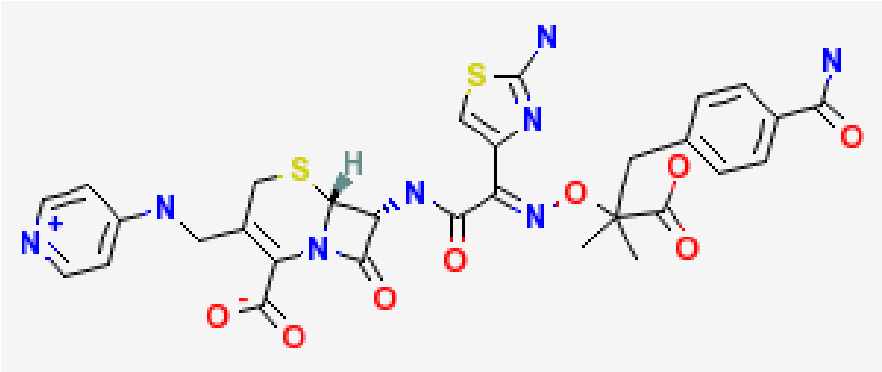
Do we have a “well defined” structure?

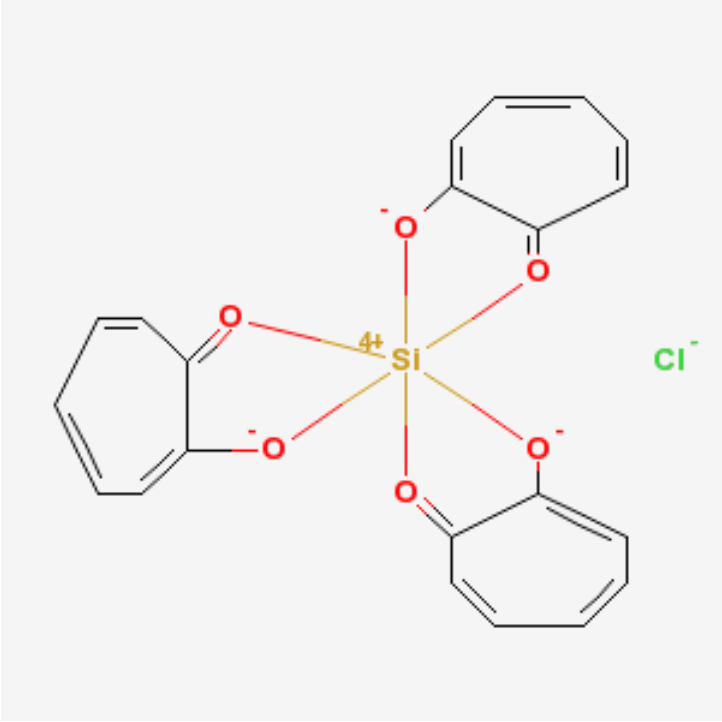


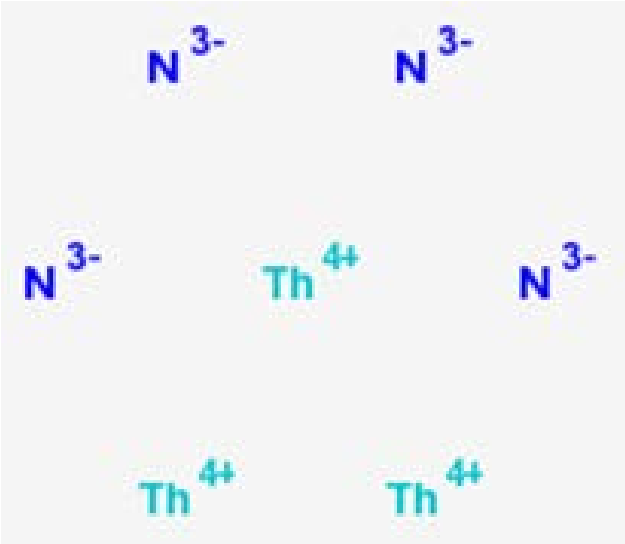




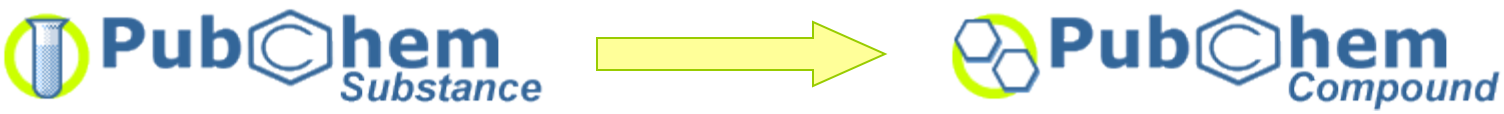
Is the structure reasonable?







Automated structure processing...



- **Verify chemical content**
  - Atoms defined/real
  - Implicit hydrogen
  - Functional group
  - Atom valence sanity
- **Normalize representation**
  - Tautomer invariance
  - Aromaticity detection
  - Stereochemistry
  - Explicit hydrogen
- **Calculate**
  - Coordinates
  - Properties
  - Descriptors
- **Detect components**
  - Isolate covalent units
  - Neutralize (+/- proton)
  - **Reprocess**
  - Detect unique

Structure as "key" between chemical biology resources

Primary accession **SID**

PubChem Substance

Primary accession **CID**

PubChem Compound

Unique chemical structure content of PubChem

- Each CID groups SIDs with same normalized structure
- Various chemical concepts help to group CIDs irrespective of salt-form, charge-state, stereo-form, isotopic-form, or mixture-type
- Original deposited structure is retained but not structure searchable

Mixture
Salt
Parent
Components

"Identity groups"
Exactly Same
Same Isotope
Same Stereo
Same Connectivity
Tautomers

Entrez interface…

- Primary (text-based) search engine

User query

Google-like approach… most likely answer is at the top…

Result record summaries

Rapid result subsets

aspirin - PubChem Compou

www.ncbi.nlm.nih.gov/pccompound?term=aspirin

NCBI Resources How To My NCBI Sign In

PubChem Compound | PubChem Compound | aspirin | Search

Save search Limits Advanced Help

Display Settings: Summary, 20 per page, Sorted by Default order Send to:

Results: 1 to 20 of 64 << First < Prev Page 1 of 4 Next > Last >>

1. aspirin; ACETYLSALICYLIC ACID; Acylpyrin ...
MW: 180.157420 g/mol MF: $C_9H_8O_4$
IUPAC: 2-acetyloxybenzoic acid
Active in 380 BioAssays Tested in 1420 BioAssays
CID: 2244
Similar Compounds Same Parent, Connectivity Mixture/Component Compounds BioAssays, activity ≤ 1 µM PubMed (MeSH Keyword)

2. Calcascorbin; Calcium aspirin; Calscorbate ...
MW: 398.376960 g/mol MF: $C_{18}H_{14}CaO_8$
IUPAC: calcium 2-acetyloxybenzoate
CID: 6247
Similar Compounds Same Parent, Connectivity Mixture/Component Compounds PubMed (MeSH Keyword)

3. acetaminophen; Paracetamol; APAP ...
MW: 151.162560 g/mol MF: $C_8H_9NO_2$
IUPAC: N-(4-hydroxyphenyl)acetamide
Active in 207 BioAssays Tested in 822 BioAssays
CID: 1983
Similar Compounds Same Parent, Connectivity Mixture/Component Compounds BioAssays, activity ≤ 1 µM PubMed (MeSH Keyword)

4. Axotal; BUTALBITAL ASPIRIN AND CAFFEINE; BUTAL COMPOUND ...
MW: 598.604360 g/mol MF: $C_{28}H_{34}N_6O_9$
IUPAC: 2-acetyloxybenzoic acid; 5-(2-methylpropyl)-5-prop-2-enyl-1,3-diazinane-2,4,6-trione; 1,

Filters: Manage Filters

Actions on your results
- BioActivity Analysis — Analyze the BioActivities of the compounds
- Structure Clustering — Cluster structures based on structural similarity
- Structure Download — Download the structures in various formats
- Pathways — Analyze pathways containing the compounds

Refine your results
• What's this?

Chemical Properties
Rule of 5 (15)

BioActivity Experiments
BioAssays, Active (10)
BioAssays, Tested (14)
Protein 3D Structures (3)
Bromodomain Of Human Crebbp With N-(4-Hydroxyphenyl) Acetamide (1)

Why so many results for “aspirin”?

User query

Google-like approach… most likely answer is at the top…

Result record summaries

Rapid result subsets

aspirin - PubChem Compou

www.ncbi.nlm.nih.gov/pccompound?term=aspirin

NCBI Resources How To My NCBI Sign In

PubChem Compound | PubChem Compound | aspirin | Search

Save search Limits Advanced Help

Display Settings: Summary, 20 per page, Sorted by Default order Send to:

Results: 1 to 20 of 64 << First < Prev Page 1 of 4 Next > Last >>

1. aspirin; ACETYLSALICYLIC ACID; Acylpyrin ...
MW: 180.157420 g/mol MF: $C_9H_8O_4$
IUPAC: 2-acetyloxybenzoic acid
Active in 380 BioAssays Tested in 1420 BioAssays
CID: 2244
Similar Compounds Same Parent, Connectivity Mixture/Component Compounds BioAssays, activity ≤ 1 µM PubMed (MeSH Keyword)

2. Calcascorbin; Calcium aspirin; Calscorbate ...
MW: 398.376960 g/mol MF: $C_{18}H_{14}CaO_8$
IUPAC: calcium 2-acetyloxybenzoate
CID: 6247
Similar Compounds Same Parent, Connectivity Mixture/Component Compounds PubMed (MeSH Keyword)

3. acetaminophen; Paracetamol; APAP ...
MW: 151.162560 g/mol MF: $C_8H_9NO_2$
IUPAC: N-(4-hydroxyphenyl)acetamide
Active in 207 BioAssays Tested in 822 BioAssays
CID: 1983
Similar Compounds Same Parent, Connectivity Mixture/Component Compounds BioAssays, activity ≤ 1 µM PubMed (MeSH Keyword)

4. Axotal; BUTALBITAL ASPIRIN AND CAFFEINE; BUTAL COMPOUND ...
MW: 598.604360 g/mol MF: $C_{28}H_{34}N_6O_9$
IUPAC: 2-acetyloxybenzoic acid; 5-(2-methylpropyl)-5-prop-2-enyl-1,3-diazinane-2,4,6-trione; 1,

Filters: Manage Filters

Actions on your results
- BioActivity Analysis — Analyze the BioActivities of the compounds
- Structure Clustering — Cluster structures based on structural similarity
- Structure Download — Download the structures in various formats
- Pathways — Analyze pathways containing the compounds

Refine your results
• What's this?

Chemical Properties
Rule of 5 (15)

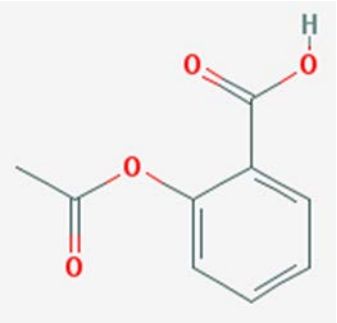
BioActivity Experiments
BioAssays, Active (10)
BioAssays, Tested (14)
Protein 3D Structures (3)
Bromodomain Of Human Crebbp With N-(4-Hydroxyphenyl) Acetamide (1)

What is “the” structure for “aspirin”?

- Fielded queries to the rescue!

Search just chemical name indexes

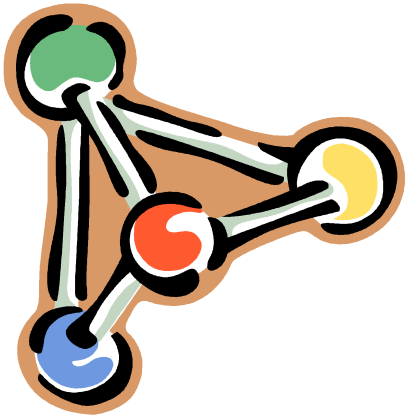
- “aspirin”
  - global keyword search - 67 hits
- “aspirin”[Synonym]
  - keyword search - 48 hits - many derivatives, mixtures, salts
- “aspirin”[CompleteSynonym]
  - exactly matches name - 1 hit
- “aspirin”[DepositorCompleteSynonym]
  - All exact matches without filtering - 16 hits



Chemical name/structure cleanup

Ensure name-chemical associations are consistent at some level of structural "sameness"

- **Same structure**
- **Same stereo isomer**
  - varies by isotope
- **Same parent structure**
  - varies by charge/salt
- **Same parent stereo isomer**
  - varies by charge/salt/isotope
- **Same connectivity**
  - varies by isotope/stereo
- **Same parent connectivity**
  - varies by charge/salt/isotope/stereo



Affect of cleanup on chemical names

- 4.65M unique chemical names used by multiple PubChem contributors

- Assign Synonym to a single "CID" ← **Sliding quality scale**
  - 4.61M (99.1%) names with "consistent" structure

Observation:

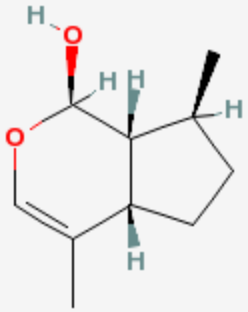
Very few cases where inconsistency is found!

What did you mean by that?

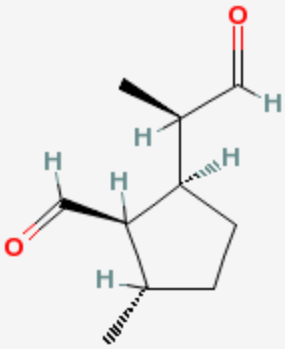
- Case Study:

(+)-Iridodial

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



Ring Closed



Ring Open

Some chemical species naturally have multiple forms

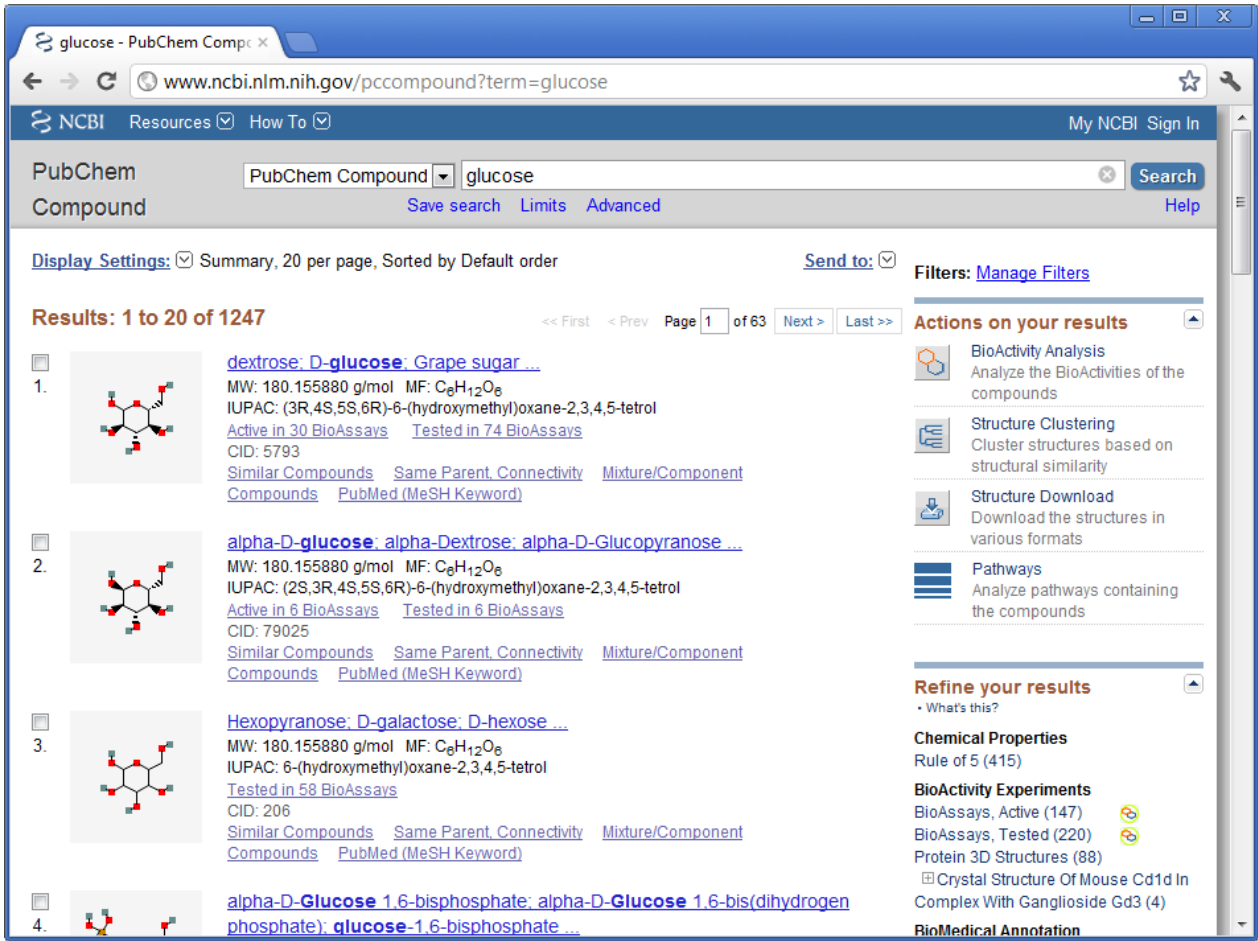
Depositors agree… but stereo an issue

Consistency type	Chemical name count	% of Total
CID	3,671,623	79.7%
STE	4,591	0.1%
PCID	40,209	0.9%
PSTE	6	0.0%
CON	887,314	19.3%
PCON	4,643	0.1%

- CID – same exact structure
  - no variation
- STE – same structure stereo form
  - variable isotopic form
- CON – same structure connectivity
  - variable stereo/isotopic form
- PCID – same exact parent structure
  - variable salt/charge state form
- PSTE – same parent structure stereo form
  - variable salt/charge state/isotopic form
- PCON – same parent structure connectivity
  - variable salt/charge state/isotopic/stereo form

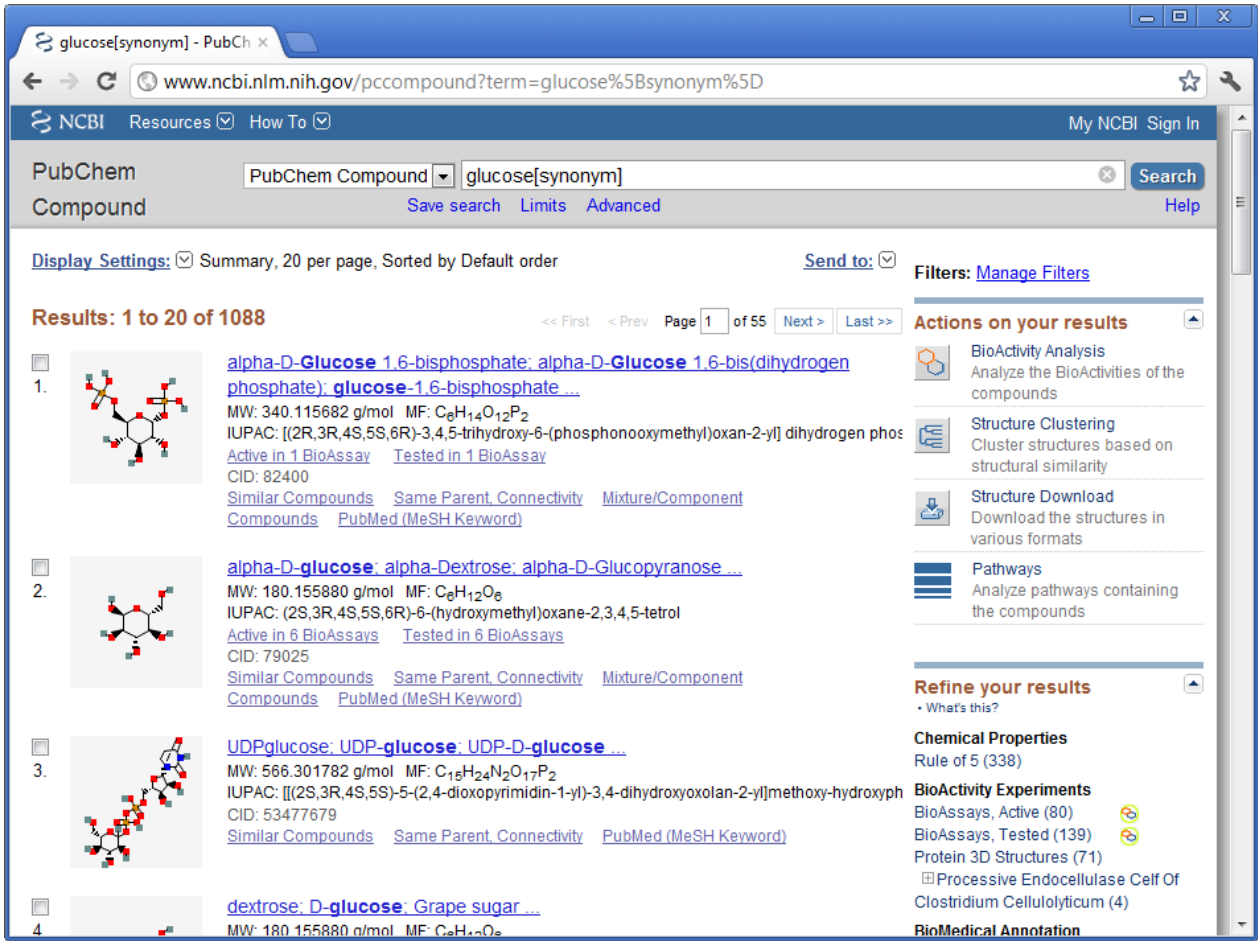
Case study… “glucose”

- Search by global keyword … 1,247 hits!



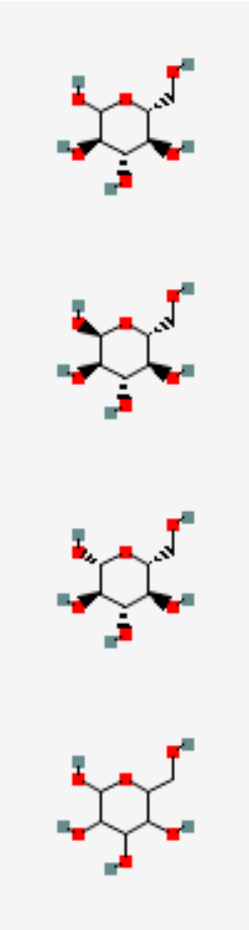
Case study… “glucose”

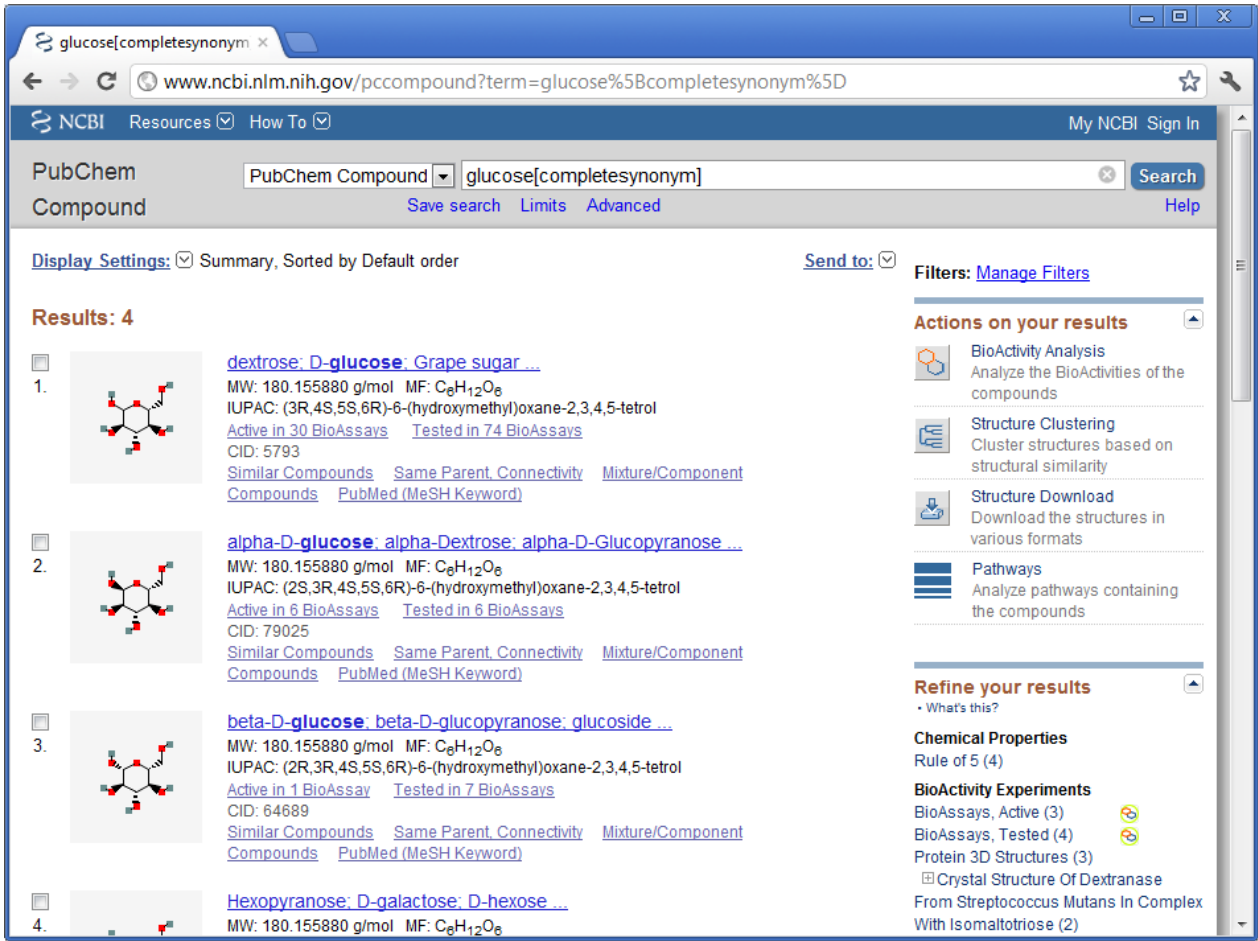
- Search by “glucose[Synonym]”… 1,088 hits!



Case study… “glucose”

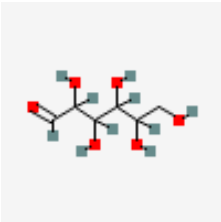
- “glucose[CompleteSynonym]”… 4 hits!

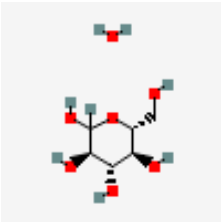


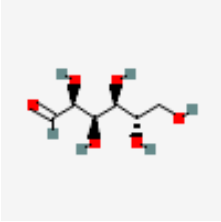


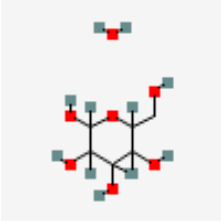
Case study… “glucose”

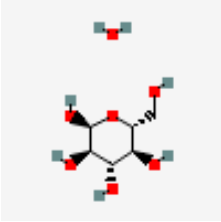
- “glucose[DepositorCompleteSynonym]”… 10 hits!

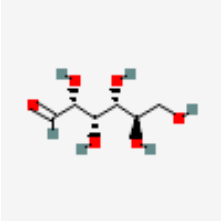


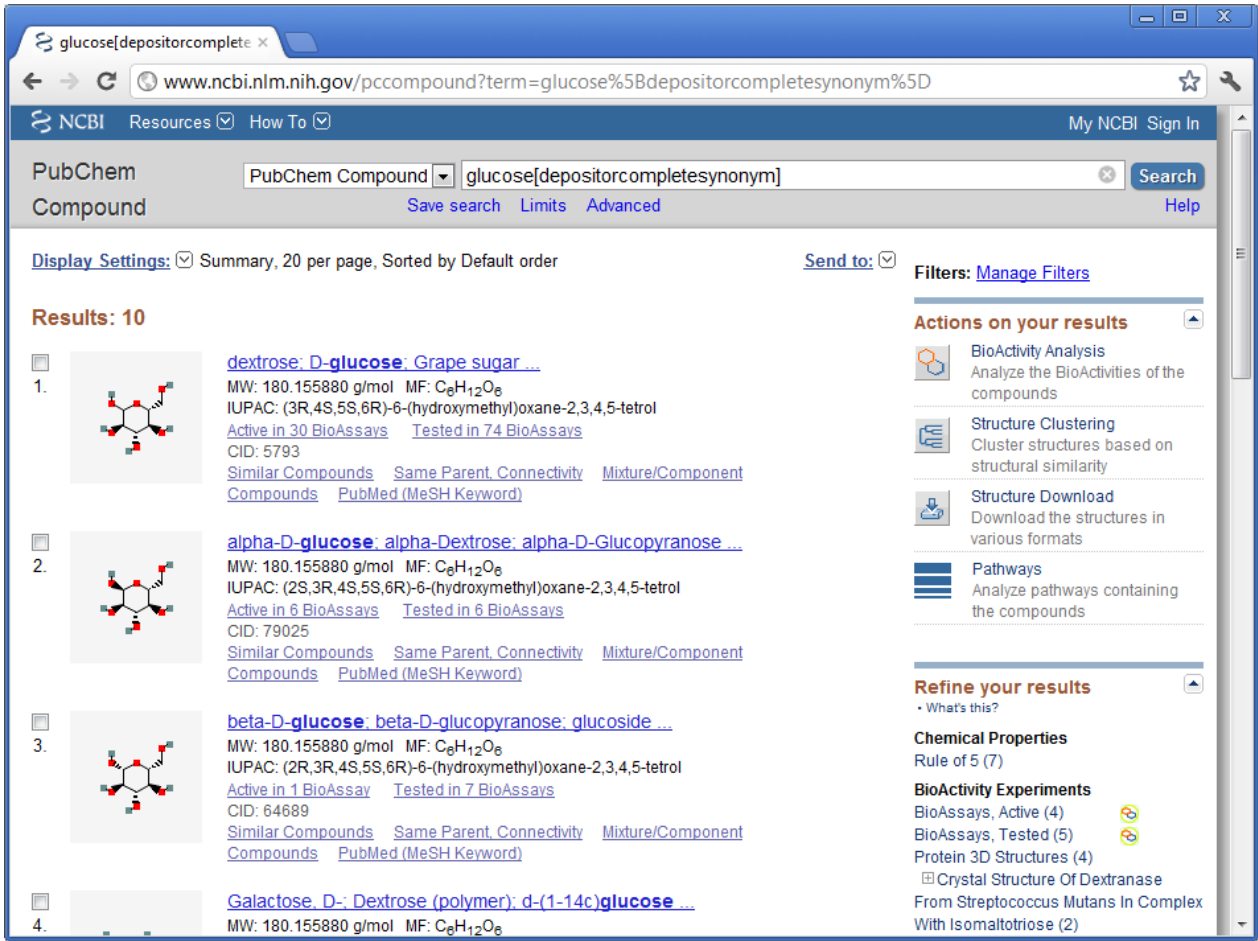




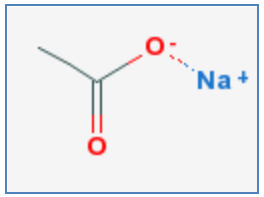




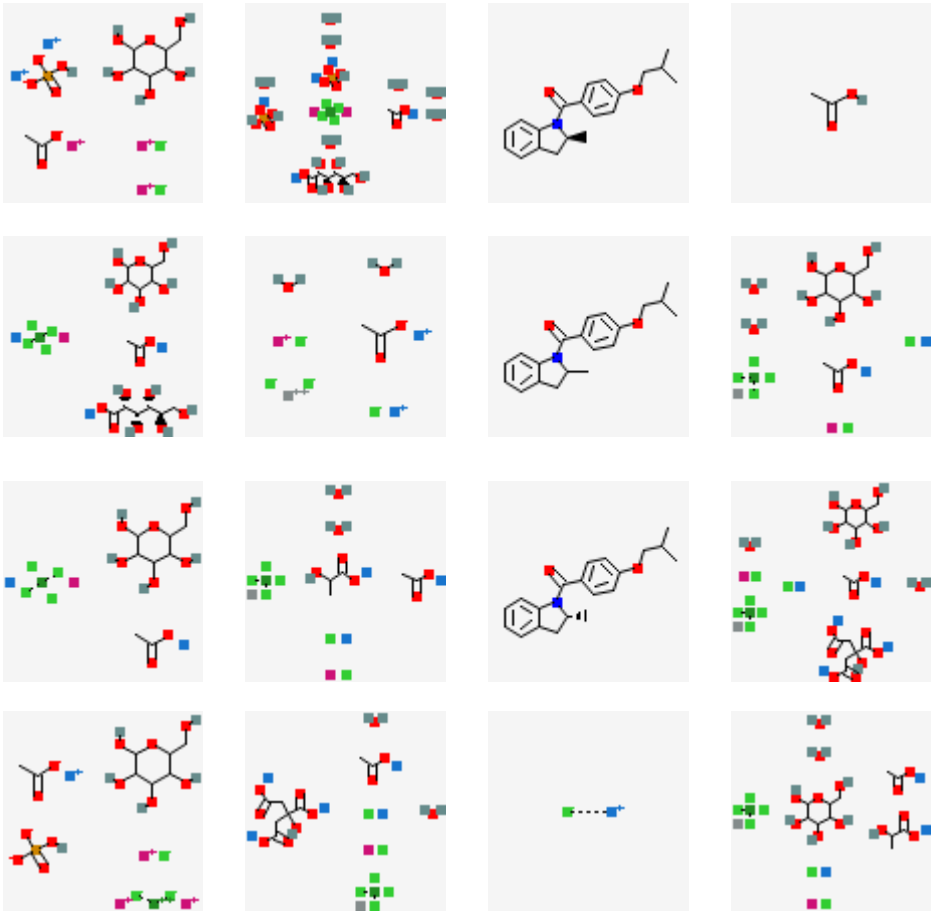


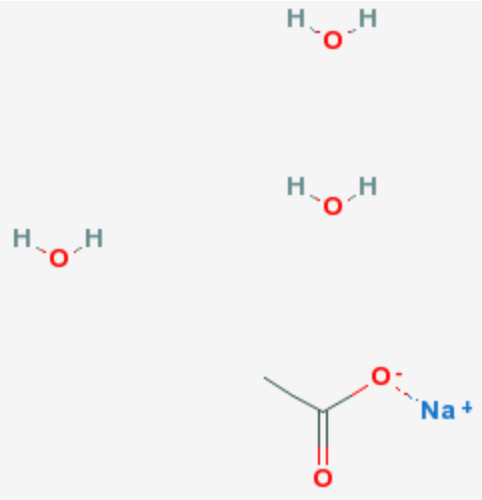


What do you mean by “sodium acetate”?



"sodium acetate"[depositorcompletesynonym]
23 hits





Chemical record for “aspirin”

Aspirin - PubChem

pubchem.ncbi.nlm.nih.gov/summary/summary.cgi?cid=2244&loc=ec_rcs

NCBI Resources How To

PubChem Compound

PubChem Compound Search

Limits Advanced search Help

Aspirin - Compound Summary (CID 2244)

ASN.1 XML SDF

Also known as: ACETYLSALICYLIC ACID, Acylpyrin, Ecotrin, 2-Acetoxybenzoic acid, Acenterine, Acetosal, Colfarit, Polopiryna, Aceticyl

Molecular Formula: $C_9H_8O_4$ **Molecular Weight:** 180.15742

The prototypical analgesic used in the treatment of mild to moderate pain. It has anti-inflammatory and antipyretic properties and acts as an inhibitor of cyclooxygenase which results in the inhibition of the biosynthesis of prostaglandins. Aspirin also inhibits platelet aggregation and is used in the prevention of arterial and venous thrombosis. (From Martindale, The Extra Pharmacopoeia, 30th ed, p5) *From: MeSH*

Table of Contents Show subcontent titles

Identification and Related Records

Use and Manufacturing

Pharmacology

Biomedical Effects and Toxicity

Safety and Handling

Environmental Fate and Exposure Potential

Exposure Standards and Regulations

Monitoring and Analysis Methods

Literature

Biomolecular Interactions and Pathways

Biological Test Results

Classification

Chemical and Physical Properties

Expand all sub-sections

2D Structure 3D Conformer

» Links and Related Information

Properties

Compound ID: 2244

Molecular Weight: 180.15742 [g/mol]

Molecular Formula: $C_9H_8O_4$

XLogP3: 1.2

H-Bond Donor: 1

H-Bond Acceptor: 4

BioActivity Data Links

with Similar Conformers

Related Compounds

Same, Connectivity (7)

Similar Compounds (1796)

Similar Conformers (2469)

Related Substances

All (985)

Same Structure (167)

Biomedical annotation from MeSH, ChEBI, KEGG, DrugBank, DailyMed, LipidMaps, HSDB, ChemIDplus, and other depositors

Compound Summary

Aspirin - PubChem

pubchem.ncbi.nlm.nih.gov/summary/summary.cgi?cid=2244&loc=ec_rcs

NCBI Resources How To

PubChem Compound

PubChem Compound Search

Limits Advanced search Help

Aspirin - Compound Summary (CID 2244)

ASN.1 XML SDF

Also known as: ACETYLSALICYLIC ACID, Acylpyrin, Ecotrin, 2-Acetoxybenzoic acid, Acenterine, Acetosal, Colfarit, Polopiryna, Aceticyl

Molecular Formula: $C_9H_8O_4$ **Molecular Weight:** 180.15742

The prototypical analgesic used in the treatment of mild to moderate pain. It has anti-inflammatory and antipyretic properties and acts as an inhibitor of cyclooxygenase which results in the inhibition of the biosynthesis of prostaglandins. Aspirin also inhibits platelet aggregation and is used in the prevention of arterial and venous thrombosis. (From Martindale, The Extra Pharmacopoeia, 30th ed, p5) *From: MeSH*

Table of Contents Show subcontent titles

- Identification and Related Records
- Use and Manufacturing
- Pharmacology
- Biomedical Effects and Toxicity
- Safety and Handling
- Environmental Fate and Exposure Potential
- Exposure Standards and Regulations
- Monitoring and Analysis Methods
- Literature
- Biomolecular Interactions and Pathways
- Biological Test Results
- Classification
- Chemical and Physical Properties

Expand all sub-sections

2D Structure 3D Conformer

» Links and Related Information

Properties

Compound ID: 2244

Molecular Weight: 180.15742 [g/mol]

Mo

XL

H-Bond Donor: 1

H-Bond Acceptor: 4

BioActivity Data Links

This Compound

with Similar Compounds

with Similar Conformers

Related Compounds

Same, Connectivity (7)

Similar Compounds (1796)

Similar Conformers (2469)

Related Substances

All (985)

Same Structure (167)

Structure write-up from MeSH or DrugBank

Compound Summary

Aspirin - PubChem

pubchem.ncbi.nlm.nih.gov/summary/summary.cgi?cid=2244&loc=ec_rcs

NCBI Resources How To

PubChem Compound

PubChem Compound Search

Limits Advanced search Help

Aspirin - Compound Summary (CID 2244)

ASN.1 XML SDF

Also known as: ACETYLSALICYLIC ACID, Acylpyrin, Ecotrin, 2-Acetoxybenzoic acid, Acenterine, Acetosal, Colfarit, Polopiryna, Aceticyl

Molecular Formula: $C_9H_8O_4$ **Molecular Weight:** 180.15742

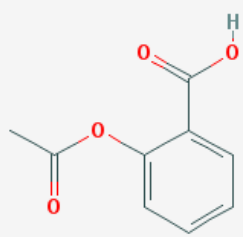
The prototypical analgesic used in the treatment of mild to moderate pain. It has anti-inflammatory and antipyretic properties and acts as an inhibitor of cyclooxygenase which results in the inhibition of the biosynthesis of prostaglandins. Aspirin also inhibits platelet aggregation and is used in the prevention of arterial and venous thrombosis. (From Martindale, The Extra Pharmacopoeia, 30th ed, p5) *From: MeSH*

Table of Contents Show subcontent titles

- Identification and Related Records
- Use and Manufacturing
- Pharmacology
- Biomedical Effects and Toxicity
- Safety and Handling
- Environmental Fate and Exposure Potential
- Exposure Standards and Regulations
- Monitoring and Analysis Methods
- Literature
- Biomolecular Interactions and Pathways
- Biological Test Results
- Classification
- Chemical and Physical Properties

Expand all sub-sections

2D Structure 3D Conformer



» Links and Related Information

Properties

Compound ID: 2244
Molecular Weight: 180.15742 [g/mol]
Molecular Formula: $C_9H_8O_4$
XLogP3: 1.2
H-Bond Donor: 1
H-Bond Acceptor: 4

BioActivity Data Links

This Compound
with Similar Compounds
with Similar Conformers

Related Compounds

Same, Connectivity (7)
Similar Compounds (1796)
Similar Conformers (2469)

Related Substances

All (985)
Same Structure (167)

Compound Summary

Aspirin - PubChem

pubchem.ncbi.nlm.nih.gov/summary/summary.cgi?cid=2244&loc=ec_rcs

NCBI Resources How To

PubChem Compound

PubChem Compound Search

Limits Advanced search Help

Aspirin - Compound Summary (CID 2244)

ASN.1 XML SDF

Also known as: ACETYLSALICYLIC ACID, Acylpyrin, Ecotrin, 2-Acetoxybenzoic acid, Acenterine, Acetosal, Colfarit, Polopiryna, Aceticyl

Molecular Formula: $C_9H_8O_4$ **Molecular Weight:** 180.15742

The prototypical analgesic used in the treatment of mild to moderate pain. It has anti-inflammatory and antipyretic properties and acts as an inhibitor of cyclooxygenase which results in the inhibition of the biosynthesis of prostaglandins. Aspirin also inhibits platelet aggregation and is used in the prevention of arterial and venous thrombosis. (From Martindale, The Extra Pharmacopoeia, 30th ed, p5) *From: MeSH*

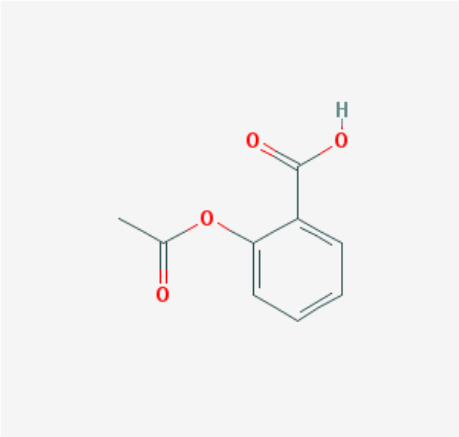
Click!

Table of Contents Show

- Identification and Related Records
- Use and Manufacturing
- Pharmacology
- Biomedical Effects and Toxicity
- Safety and Handling
- Environmental Fate and Exposure Potential
- Exposure Standards and Regulations
- Monitoring and Analysis Methods
- Literature
- Biomolecular Interactions and Pathways
- Biological Test Results
- Classification
- Chemical and Physical Properties

Expand all sub-sections

2D Structure 3D Conformer



» Links and Related Information

Properties

Compound ID: 2244

Molecular Weight: 180.15742 [g/mol]

Molecular Formula: $C_9H_8O_4$

XLogP3: 1.2

H-Bond Donor: 1

H-Bond Acceptor: 4

BioActivity Data Links

This Compound

with Similar Compounds

with Similar Conformers

Related Compounds

Same, Connectivity (7)

Similar Compounds (1796)

Similar Conformers (2469)

Related Substances

All (985)

Same Structure (167)

Aspirin: Table of Contents

Identification and Related Records
- Depositor-Supplied Synonyms
- Compound Information
- Compound Descriptors
- Substance Information

Use and Manufacturing
- Medication Information
- Methods of Manufacturing
- Impurities
- Formulations/Preparations
- Consumption Patterns
- U.S. Production
- U.S. Imports
- U.S. Exports

Pharmacology
- Pharmacological Action
- Therapeutic Uses
- Drug Warning
- Drug Idiosyncracies
- Minimum/Potential Fatal Human Dose
- Maximum Drug Dose

Biomedical Effects and Toxicity
- Absorption, Distribution and Excretion
- Metabolism/Metabolites
- Biological Half-Life
- Mechanism of Action
- Interactions
- Toxicity Summary
- Antidote and Emergency Treatment
- Human Toxicity Excerpts
- Non-Human Toxicity Excerpts
- Human Toxicity Values
- Non-Human Toxicity Values
- Ecotoxicity Values
- Populations at Special Risk
- Toxicology Links

Safety and Handling
- Safety Links
- Fire Potential
- Reactivities and Incompatibilities
- Decomposition
- Skin, Eye, and Respiratory Irritations
- Protective Equipment and Clothing
- Other Preventative Measures
- Stability/Shelf Life
- Storage Conditions
- Disposal Methods

Environmental Fate and Exposure Potential
- Environmental Fate/Exposure Summary
- Artificial Sources
- Environmental Fate
- Biodegredation
- Abiotic Degredation
- Bioconcentration
- Soil Adsorption/Mobility
- Volatilization from Water/Soil
- Water Concentrations
- Effluents Concentrations
- Milk Concentrations
- Probable Routes of Human Exposure
- Body Burdens

Exposure Standards and Regulations
- OSHA Standards
- NIOSH Recommendations
- Threshold Limit Values

Monitoring and Analysis Methods
- Analytic Laboratory Methods
- Clinical Laboratory Methods

Literature
- Depositor Provided PubMed Citations
- NLM Curated PubMed Citations

Biomolecular Interactions and Pathways
- Protein Bound 3-D Structures
- Biosystems and Pathways

Biological Test Results
- BioAssay Results

Classification
- Ontologies
- Substance Categorization Classification

Chemical and Physical Properties
- Color/Form
- Odor
- Melting Point
- Density/Specific Gravity
- Dissociation Constants
- Octanol/Water Partition Coefficient
- Solubilities
- Spectral Properties
- Vapor Pressure
- Other Chemical/Physical Properties
- Computed Properties

Classifications in PubChem

Aspirin - PubChem

pubchem.ncbi.nlm.nih.gov/summary/summary.cgi?cid=2244#x393

Classification

MeSH (Medical Subject Headings) is the NLM controlled vocabulary thes
PubMed.

- Chemical Actions and Uses
  - Pharmacologic Actions
    - Physiological Effects of Drugs
      - Peripheral Nervous System Agents
        - Sensory System Agents
          - Analgesics
            - Analgesics, Non-Narcotic
              - Anti-Inflammatory Agents, Non-Steroidal
                - Cyclooxygenase Inhibitors
      - Antipyretics
    - Therapeutic Uses
      - Anti-Inflammatory Agents
        - Anti-Inflammatory Agents, Non-Steroidal
          - Cyclooxygenase Inhibitors
      - Antirheumatic Agents
        - Anti-Inflammatory Agents, Non-Steroidal
      - Cardiovascular Agents
        - Fibrinolytic Agents
      - Hematologic Agents
        - Fibrinolytic Agents
        - Platelet Aggregation Inhibitors
    - Molecular Mechanisms of Pharmacological Action
      - Fibrin Modulating Agents
        - Fibrinolytic Agents
      - Enzyme Inhibitors
        - Cyclooxygenase Inhibitors

AID 466 - PubChem BioAssay

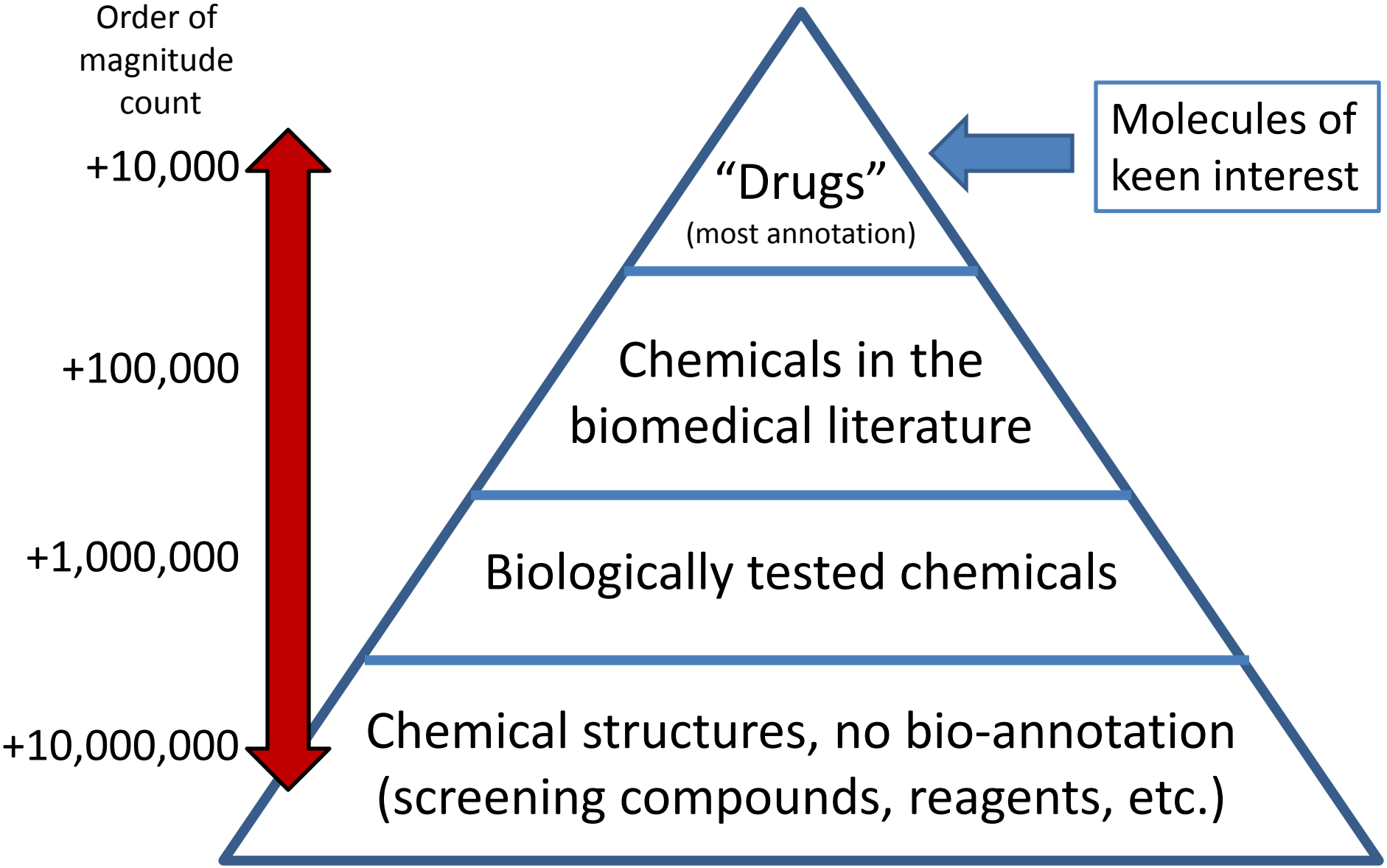
pubchem.ncbi.nlm.nih.gov/assay/assay.cgi?aid=466#aClass

Classification:

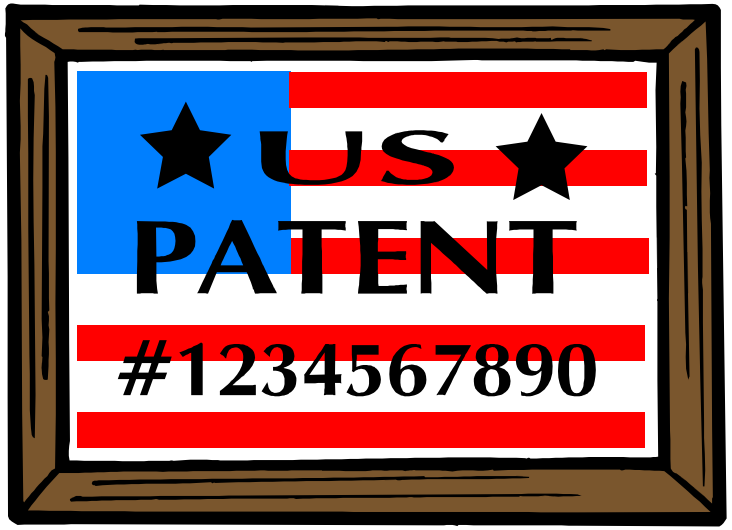
The Gene Ontology (GO) project provides a controlled vocabulary of terms for describing the fun
products, and is divided into three domains. Each term in the biological processes domain, shown
represents recognized series of events, or a collection of molecular events with a defined begin
Mutant phenotypes often reflect disruptions in biological processes. The terms below apply to the
gene/protein target(s) tested by the BioAssay.

- biological_process
  - biological regulation
    - regulation of biological process
      - regulation of cellular process
        - positive regulation of cellular process
          - positive regulation of cell proliferation
            - positive regulation of smooth muscle cell proliferation
              - sphingosine 1-phosphate receptor 1 [Homo sapiens]
      - regulation of cellular component biogenesis
        - regulation of actin filament bundle assembly
          - negative regulation of actin filament bundle assembly
            - negative regulation of stress fiber assembly
              - sphingosine 1-phosphate receptor 1 [Homo sapiens]
  - cellular process
    - cellular response to stimulus
      - signal transduction
        - intracellular signal transduction
          - second-messenger-mediated signaling
            - inositol phosphate-mediated signaling

PubChem biological annotation pyramid



2012
the year of the patent



Patent Information in PubChem

- Dec 2011 – IBM added 2.4M substances (SIDs)
  - Derived by automated text processing
  - Patent data from years 1800 through 2000
    - 783,175 USPTO patents (23.1M SID-Patent links)
    - 399,276 EPO patents (14.8M SID-Patent links)
    - 186,899 WIPO patents (5.5M SID-Patent links)

  - 5 MLP chemical "probes"
  - 3,829 found in PDB
  - 92,760 biologically tested (~4%)

Patent Information in PubChem

- Aug 2012 – ScripDB added 6.1M SIDs
  - Structures from USPTO complex work units (i.e., figures)
  - Patent data from Jan 2001 through Feb 2012
    - 15,895,209 SID-Patent links
      - 202,775 USPTO applications
      - 108,171 USPTO patents

  - 16 MLP chemical "probes"
  - 4,594 found in PDB
  - 122,668 biologically tested (~2%)

See ScripDB paper: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3245107/

Old and New Patent intersects

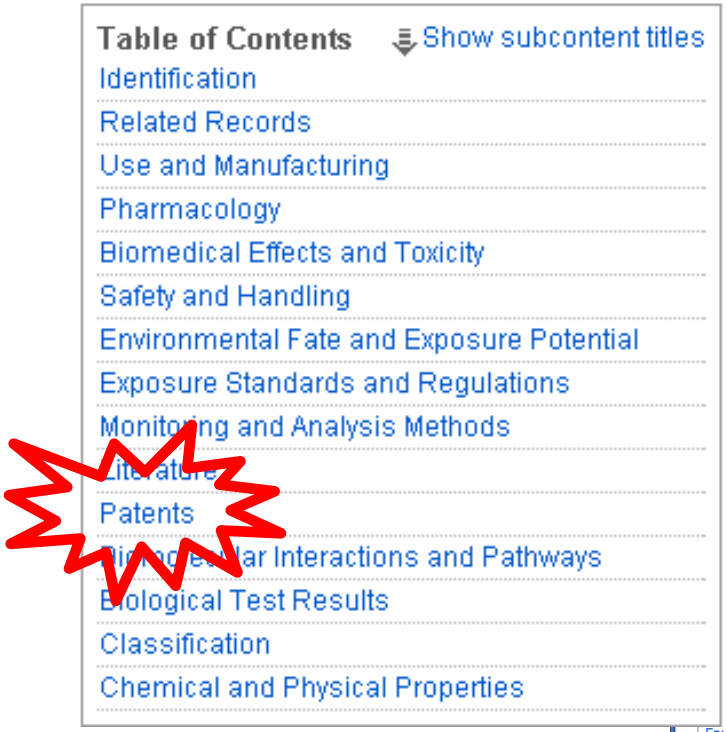
- Unique structure counts
  - 2.36M from IBM and 3.99M from ScripDB
  - Only 206,863 structures in common
  - 6.14M total unique between IBM and ScripDB
    - 5.97M with patent data (59M SID-patent links)
      - IBM also gave PMID/structures associations

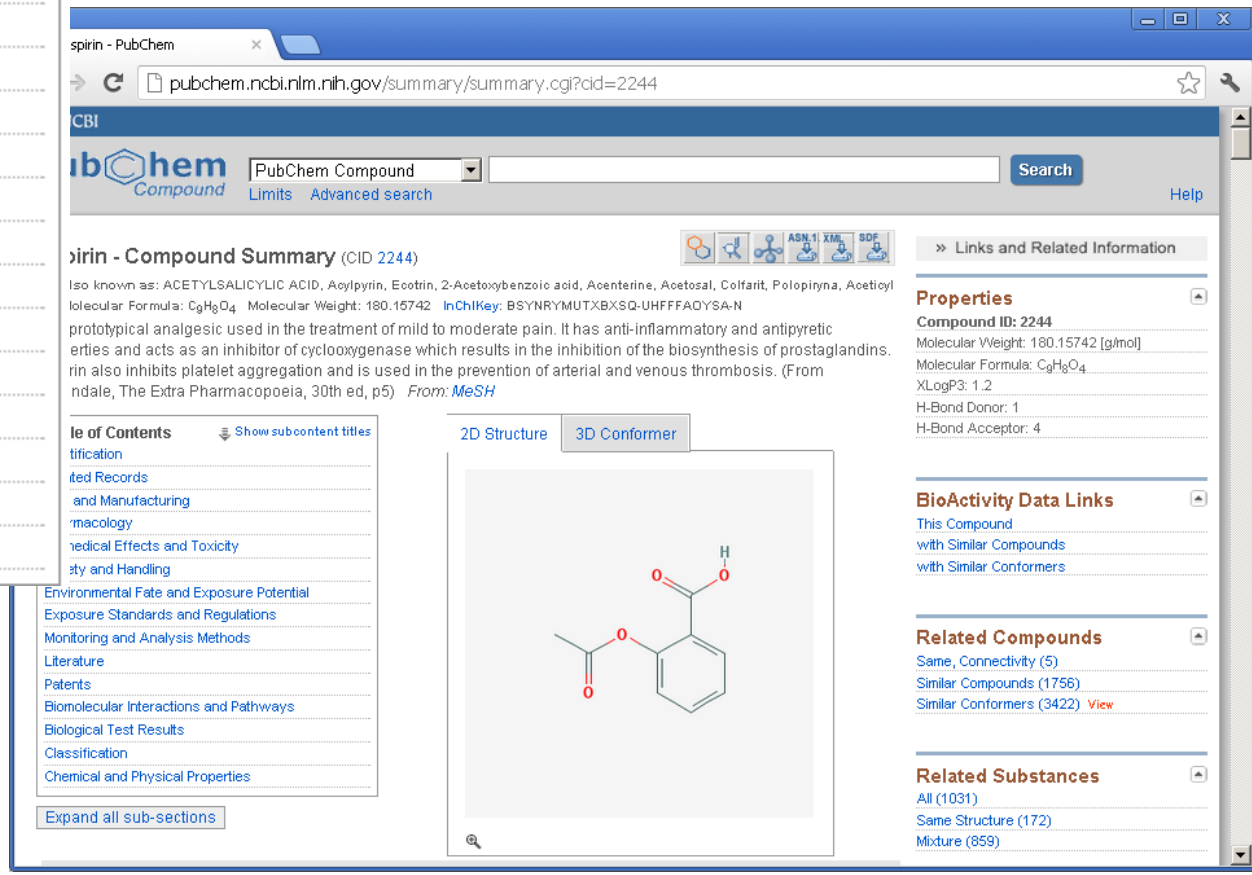
  - 16 MLP chemical "probes"
  - 5,678 links to PDB
  - 180,870 biologically tested (10% of all tested)
  - 106,133 biologically active (13% of all actives)

New Patent Interfaces (things to come)

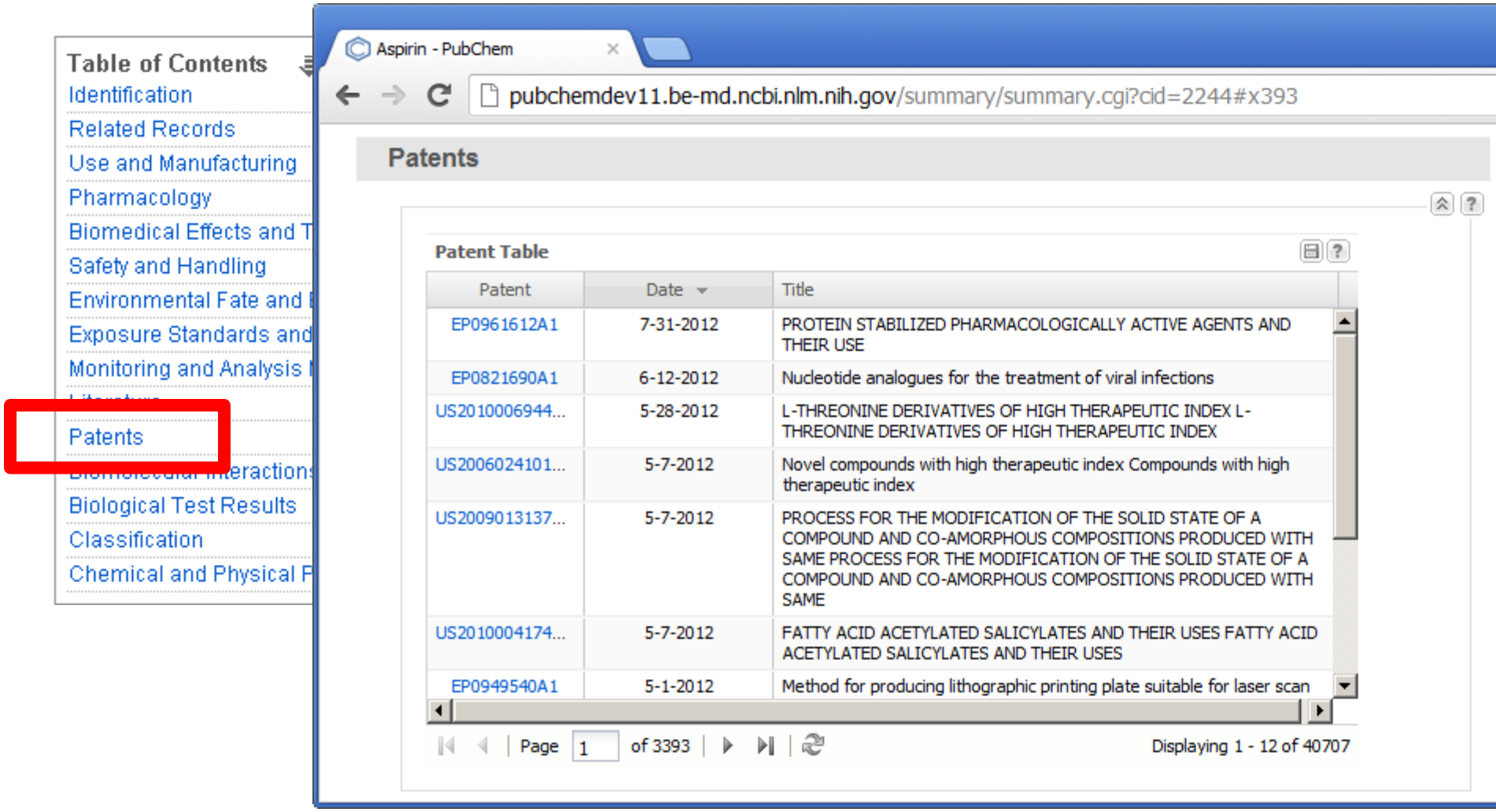
- Patent “family” associations
- Per patent summary pages
  - Summarize what is know about structures mentioned in a patent (bioactivity, PMIDs, etc.)
  - Cross-link information/access
- Navigation/Filtering of patents/structures by…
  - Patent classification tree
  - Location in document (title, abstract, claims,…)
  - Structure type (Markush, fragment, whole,…)
  - Date, originator, …

PubChem Widgets provide interactive content

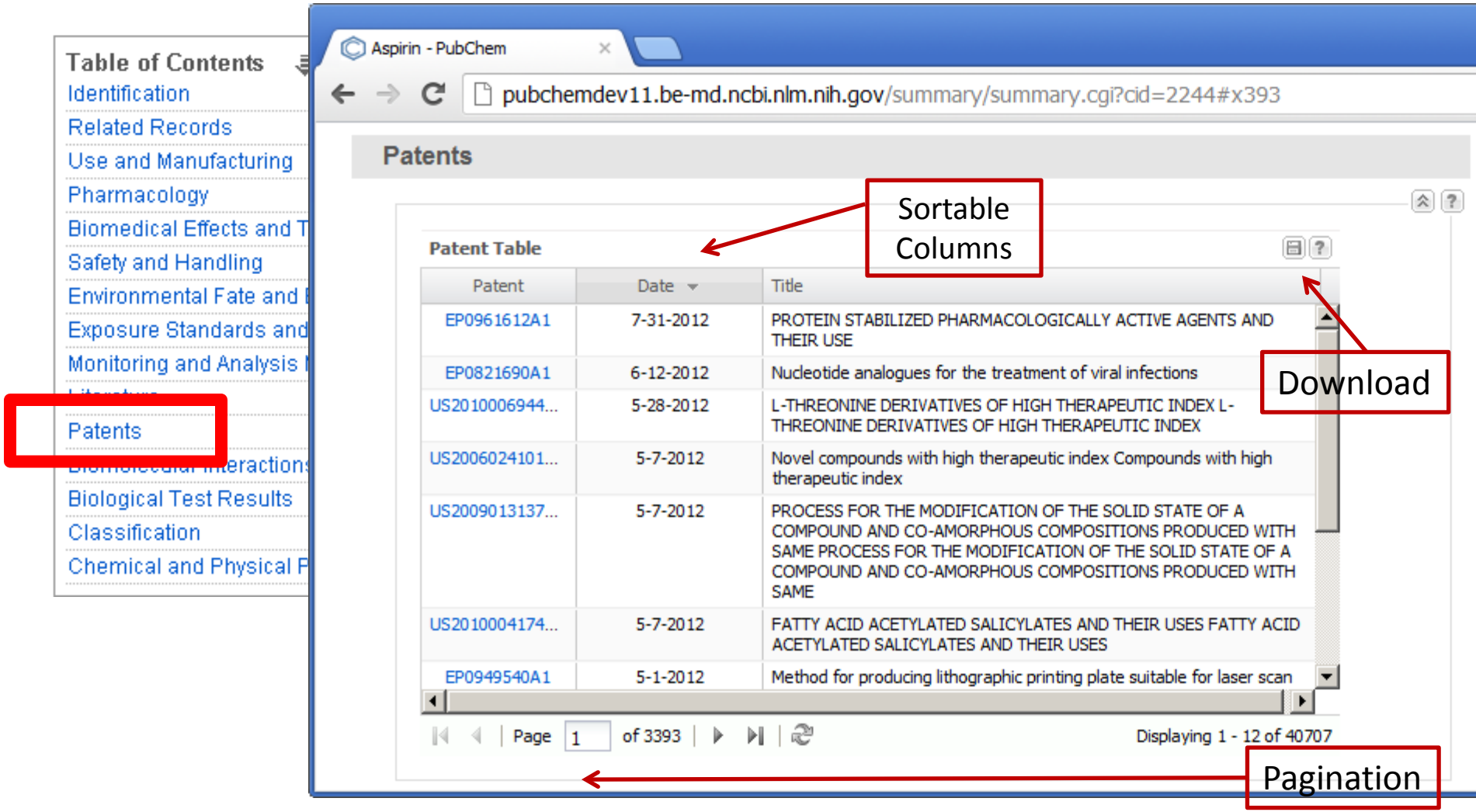




Provide interactive content for PubChem Web pages



Provide interactive content for PubChem Web pages



Modular entities you can add into your own web pages…

CID 2244 - PubChem Result

pubchem.ncbi.nlm.nih.gov/data/cidpatentdata?cid=2244

PubChem Patent Data - CID 2244
Aspirin

Patent Table

Patent	Date	Title
EP0961612A1	7-31-2012	PROTEIN STABILIZED PHARMACOLOGICALLY ACTIVE AGENTS AND THEIR USE
EP0821690A1	6-12-2012	Nucleotide analogues for the treatment of viral infections
US2010006944...	5-28-2012	L-THREONINE DERIVATIVES OF HIGH THERAPEUTIC INDEX L-THREONINE DERIVATIVES OF HIGH THERAPEUTIC INDEX
US2006024101...	5-7-2012	Novel compounds with high therapeutic index Compounds with high therapeutic index
US2009013137...	5-7-2012	PROCESS FOR THE MODIFICATION OF THE SOLID STATE OF A COMPOUND AND CO-AMORPHOUS COMPOSITIONS PRODUCED WITH SAME PROCESS FOR THE MODIFICATION OF THE SOLID STATE OF A COMPOUND AND CO-AMORPHOUS COMPOSITIONS PRODUCED WITH SAME
US2010004174...	5-7-2012	FATTY ACID ACETYLATED SALICYLATES AND THEIR USES FATTY ACID ACETYLATED SALICYLATES AND THEIR USES
EP0949540A1	5-1-2012	Method for producing lithographic printing plate suitable for laser scan exposure, and photopolymerizable composition
EP1058538A1	4-10-2012	FAST DISINTEGRATING TABLETS
US2010003365...	4-2-2012	LIGHT ABSORPTION LAYER FOR A DISPLAY DEVICE LIGHT ABSORPTION LAYER FOR A DISPLAY DEVICE
US2009003646...	2-27-2012	ALKYLSULFONYL-SUBSTITUTED THIAZOLIDE COMPOUNDS ALKYLSULFONYL-SUBSTITUTED THIAZOLIDE

Page 1 of 3393

Displaying 1 - 12 of 40710

Not just limited to patents…

AID 1234 - PubChem Result

pubchemdev11.be-md.ncbi.nlm.nih.gov/data/data?aid=1234

NCBI Help

PubChem PubChem Compound Search

PubChem BioAssay Data - AID 1234

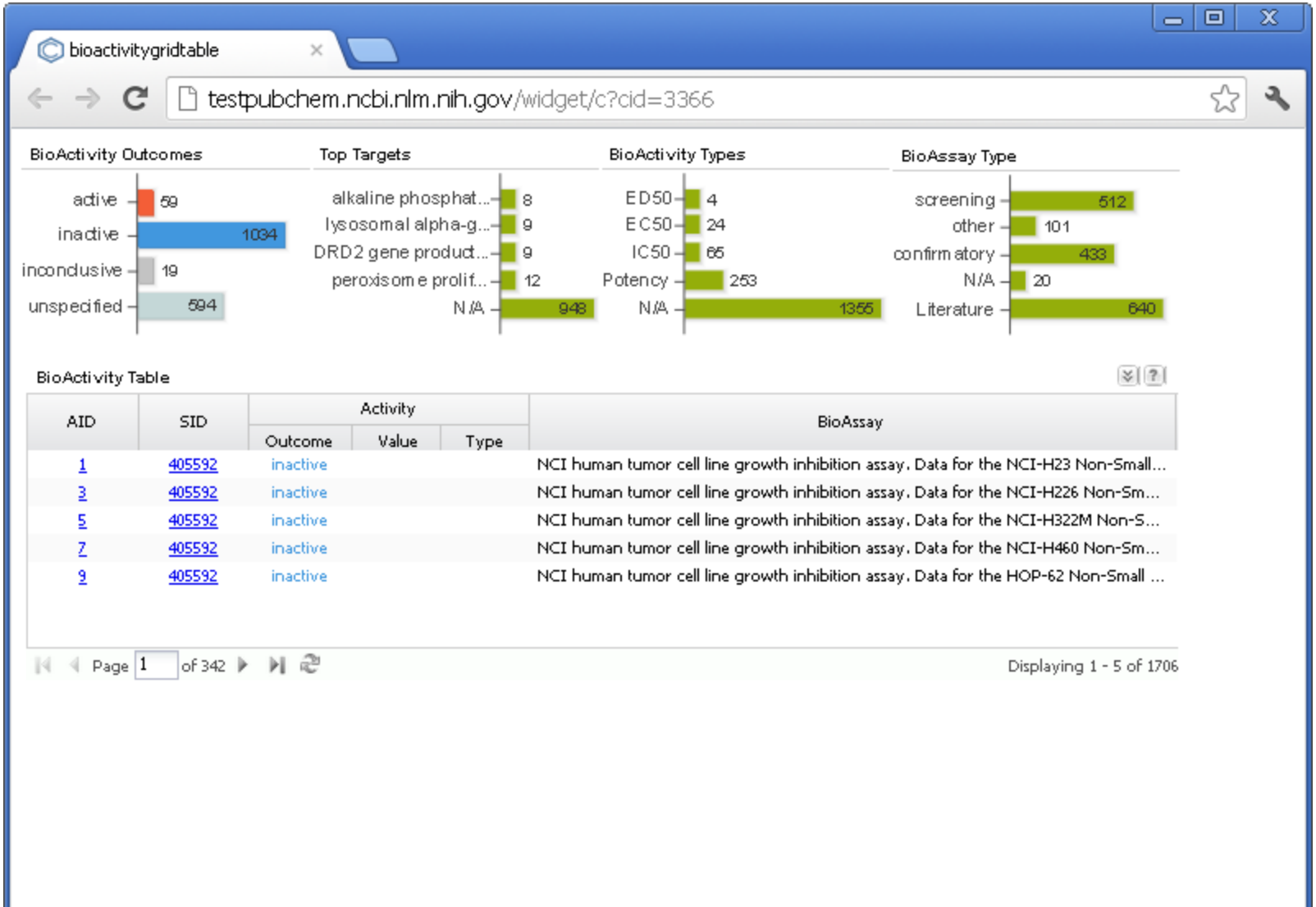
BioActivity Table

Structure	Activity			Name	Substance SID	Compound CID
	Outcome	Value[uM]	Type			
	active	0.04	DR EC50 uM	ethyl 8-bromo-2,3-dimethyl-4,5-dioxo-4,5-dihydro-3H-benzo[e]indole-1-carboxylate	14727734	1736381
	active	0.43	DR EC50 uM	ethyl 2-methyl-4,5-dioxo-4,5-dihydro-3H-benzo[e]indole-1-carboxylate	22404392	1736294
	active	0.71	DR EC50 uM	Phenanthrenequinone	16952991	6763
	active	0.89	DR EC50 uM	STK397048	14723875	460749
	active	0.9	DR EC50 uM	MLS000027832	845167	647501
	active	0.91	DR EC50	SMR000115581	14719117	9549141

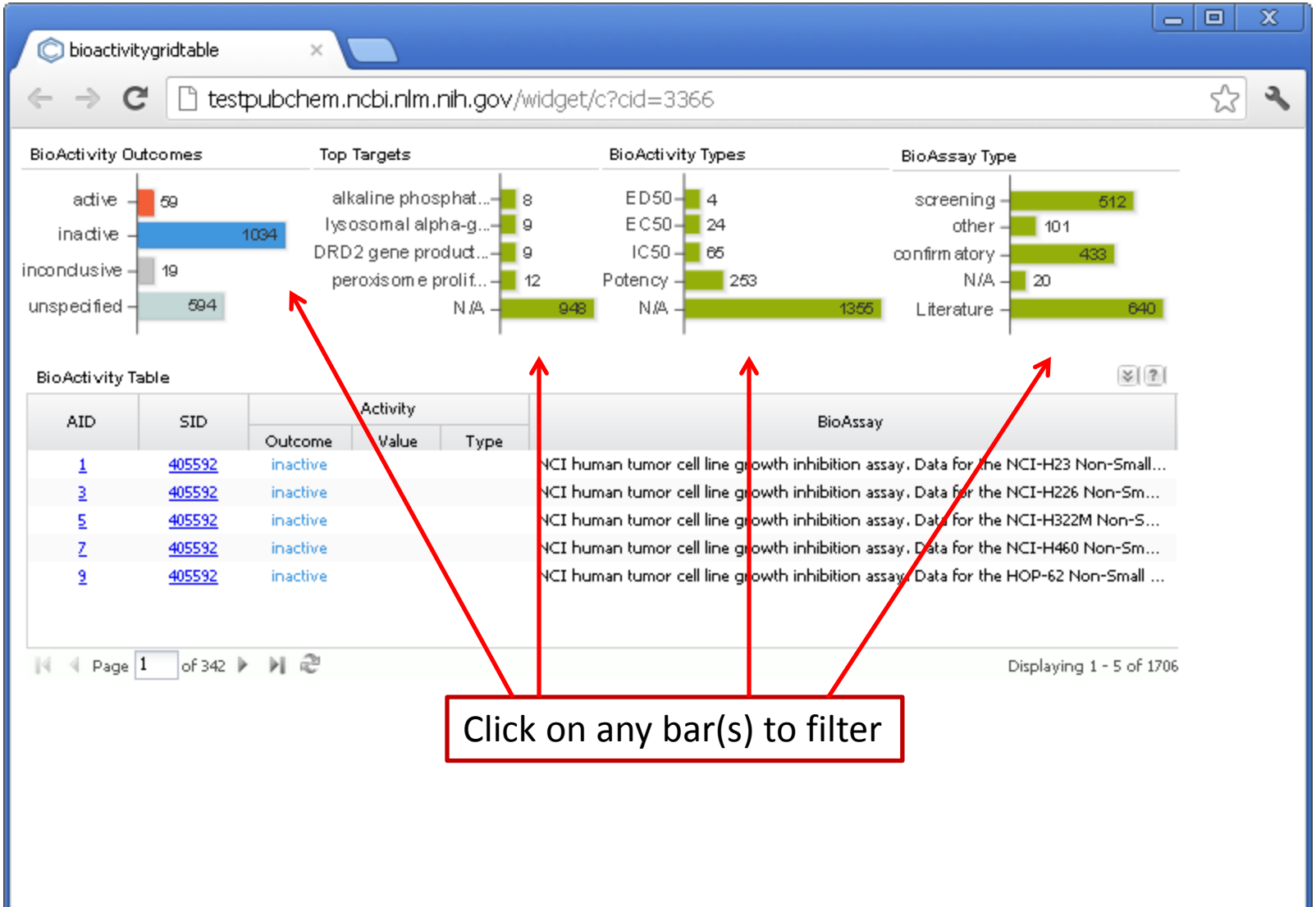
Page 1 of 6

Displaying 1 - 12 of 61

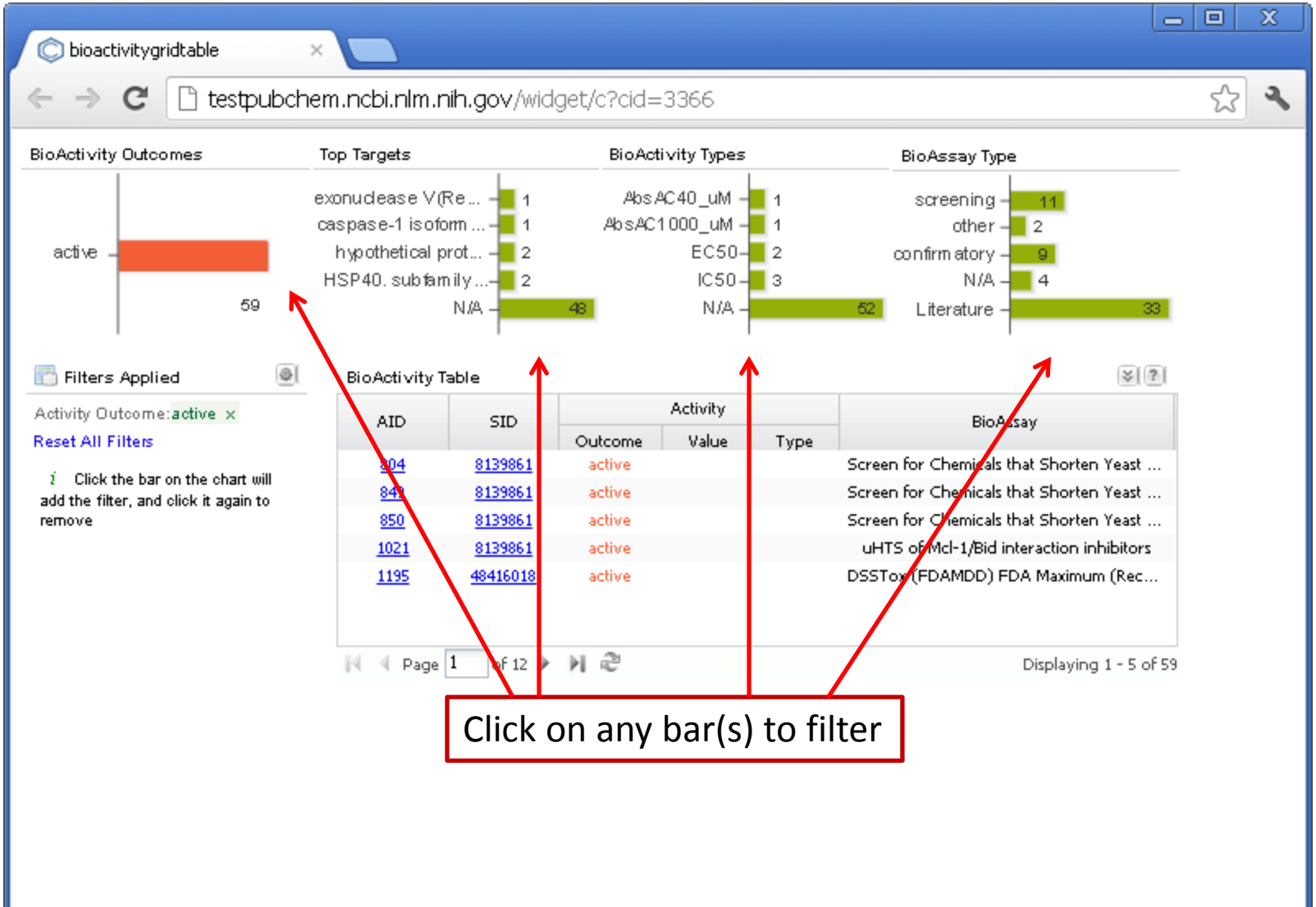
Not just limited to tables… summarization/filtering capabilities



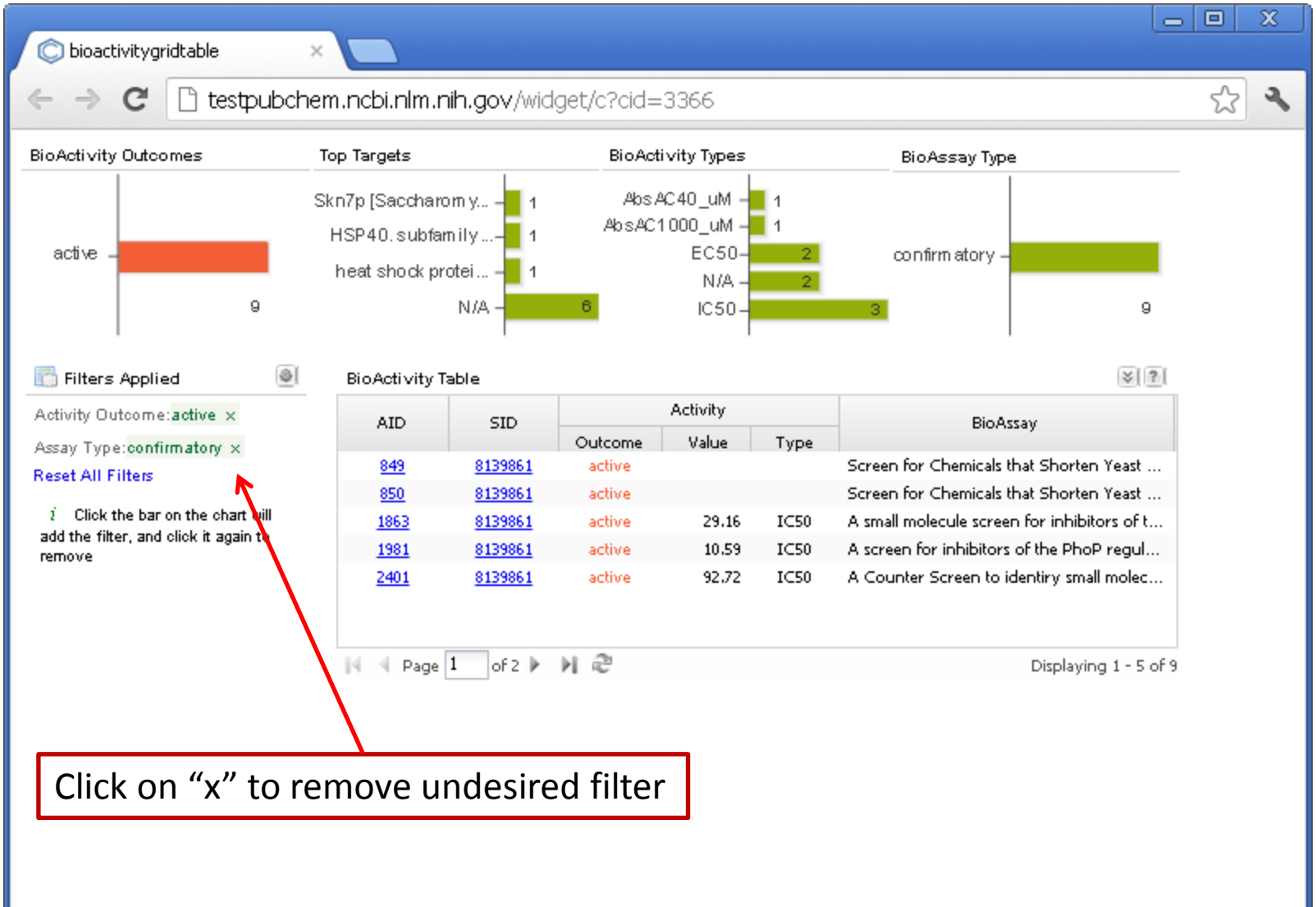
Not just limited to tables… summarization/filtering capabilities



Not just limited to tables… summarization/filtering capabilities



Not just limited to tables… summarization/filtering capabilities



HTTPS available NCBI-wide

- PubChem is HTTPS compliant for all interfaces and services
  - Including Entrez, and programmatic access
  - If you see anything not compliant, please let us know when you see them…

    pubchem-help@ncbi.nlm.nih.gov

- Yes it took a while… but it required a lot of arm twisting to make it happen…

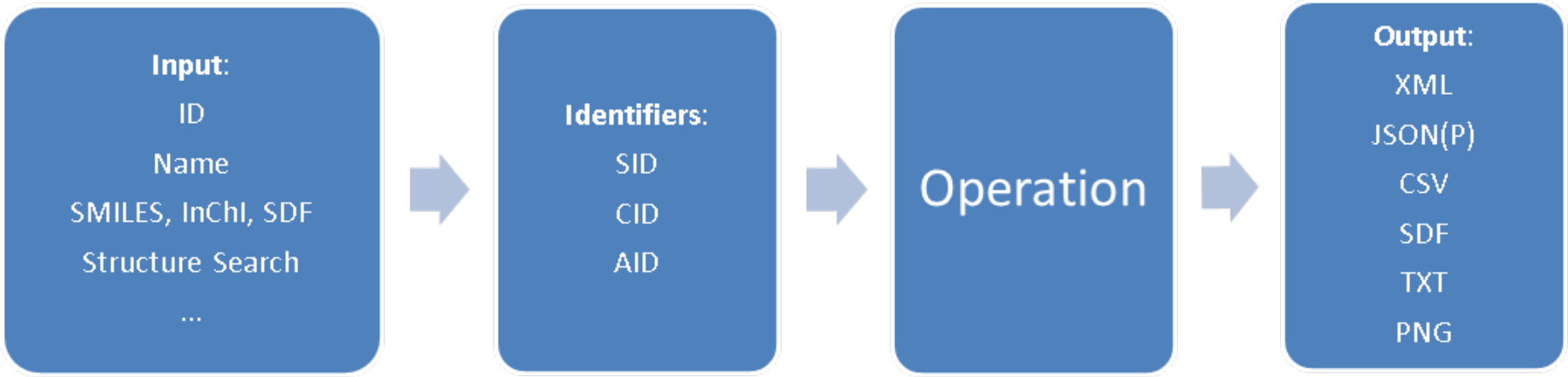
https://pubchem.ncbi.nlm.nih.gov

PUG/REST Interface

- PUG = Power User Gateway
  - Family of interfaces (ticket based)
    - PUG (deep XML, queries but also for saving state)
    - PUG SOAP (simplified XML)
- RESTful (not REST)
  - URL-based approach
- Gives access to AID/CID/SID-based data
  - Also patent IDs and …
- Synchronous data access

Tutorial provided at http://pubchem.ncbi.nlm.nih.gov/pug_rest/

Simplified Data Access



http://pubchem.ncbi.nlm.nih.gov/rest/pug
prolog

/compound/name/vioxx
input

/property/InChI
operation

/TXT
output

Output Format	Description
XML	standard XML, for which a schema is available
JSON	JSON, JavaScript Object Notation
JSONP	JSONP, like JSON but wrapped in a callback function
ASNB	standard binary ASN.1, NCBI's native format in many cases
ASNT	NCBI's human-readable text flavor of ASN.1
SDF	chemical structure data
CSV	comma-separated values, spreadsheet compatible
PNG	standard PNG image data
TXT	plain text

PubChem Crew …

Steve Bryant

Tiejun Chen

Gang Fu

Lewis Geer

Asta Gindulyte

Volker Hahnke

Lianyi Han

Jane He

Siqian He

Sunghwan Kim

Qingliang Li

Ben Shoemaker

Tugba Suzek

Paul Thiessen

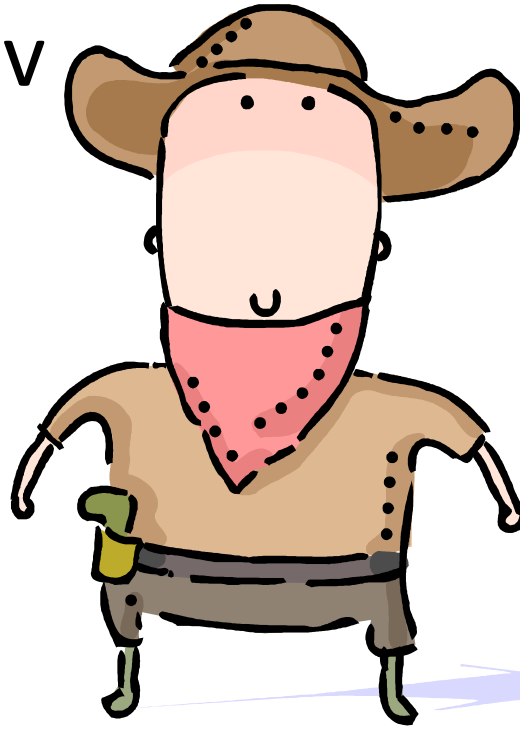
Jiyao Wang

Yanli Wang

Bo Yu

Jian Zhang

Have any questions?

If you think of a question later, email me:

bolton@ncbi.nlm.nih.gov