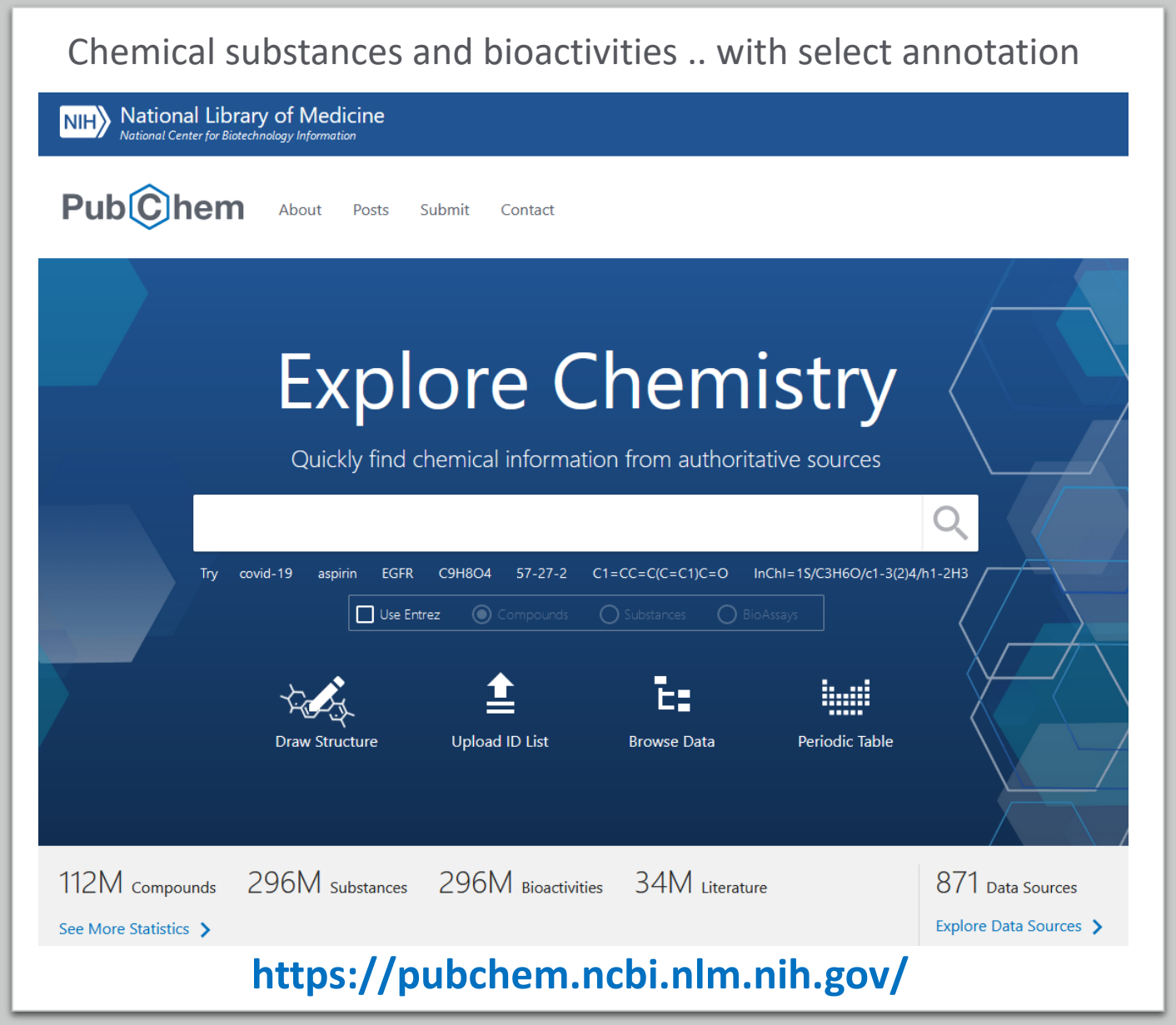
PubChem: Advancing chemical information through InChI

Evan Bolton, Ph.D.

National Library of Medicine
National Center for Biotechnology Information

PubChem is a data repository

- World's largest collection of freely accessible chemical information
- Helps researchers make sense of the biological roles and health effects of chemicals on human health and the environment



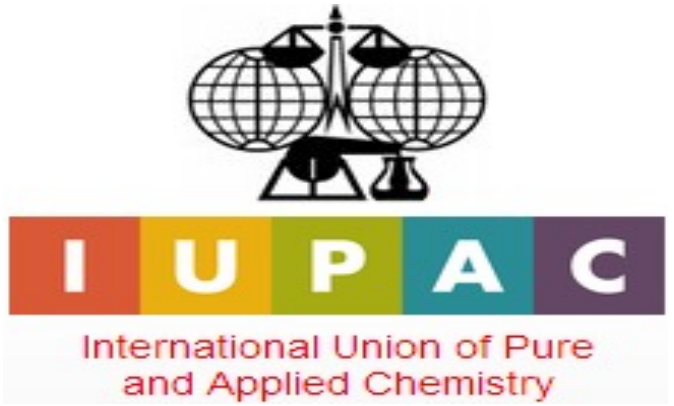
A lot of data .. enabling many use cases

PubChem Data Counts

Data Collection	Live Count	Description
Compounds	111,665,090	Unique chemical structures extracted from co
Substances	295,791,648	Information about chemical entities provided
BioAssays	1,466,013	Biological experiments provided by PubChem
Bioactivities	296,237,788	Biological activity data points reported in Pub
Genes	103,622	Genes tested in PubChem BioAssays and tho
Proteins	185,291	Proteins tested in PubChem BioAssays and th
Taxonomy	112,547	Organisms of proteins/genes tested in PubCh
Pathways	239,173	Interactions between chemicals, genes, and p
Cell Lines	1,964	Cell Lines tested in PubChem BioAssays
Literature	34,473,384	Scientific publications with links in PubChem
Patents	42,395,312	Patents with links in PubChem
Data Sources	871	Organizations contributing data to PubChem

What is an InChI?

- Acronym
  - **In**ternational **Ch**emical **I**dentifier
- Standard that is Software
  - Initially created by NIST
  - Under auspices of IUPAC (and maintained by the InChI Trust)
  - Open source, non-proprietary
- Algorithm
  - Normalizes chemical representation
  - Includes a ‘hashed’ form (InChIKey)



https://iupac.org/inchi

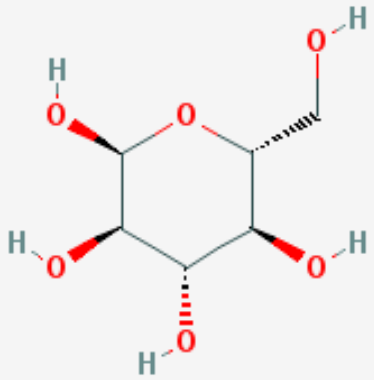
https://www.inchi-trust.org/

InChI is a string

InChI=1S/C6H12O6/c7-1-2-3(8)4(9)5(10)6(11)12-2/h2-11H,1H2/t2-,3-,4+,5-,6+/m1/s1

Version Type
Chemical formula
Connectivity
Charge&Proton
Stereochemical
Other (e.g., Isotopic)

"layered" line notation



alpha-D-Glucose

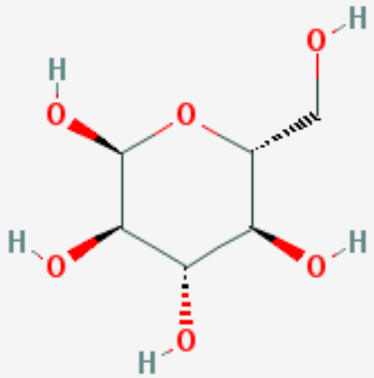
InChIKey is a “hashed” InChI

- Fixed length and search engine friendly InChI
- May allow for ‘secure’ lookup of a chemical

Chemical formula
& Connectivity
Stereochemical &
Proton &
Other (e.g., Isotopic)
Type Version
Charge

WQZGKKKJIJFFOK-DVKNGEFBSA-N

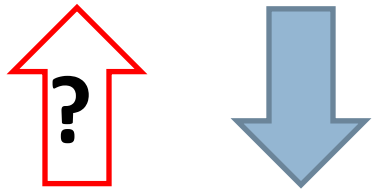
“layered” line notation



alpha-D-Glucose

InChIKey can be a ‘secret’

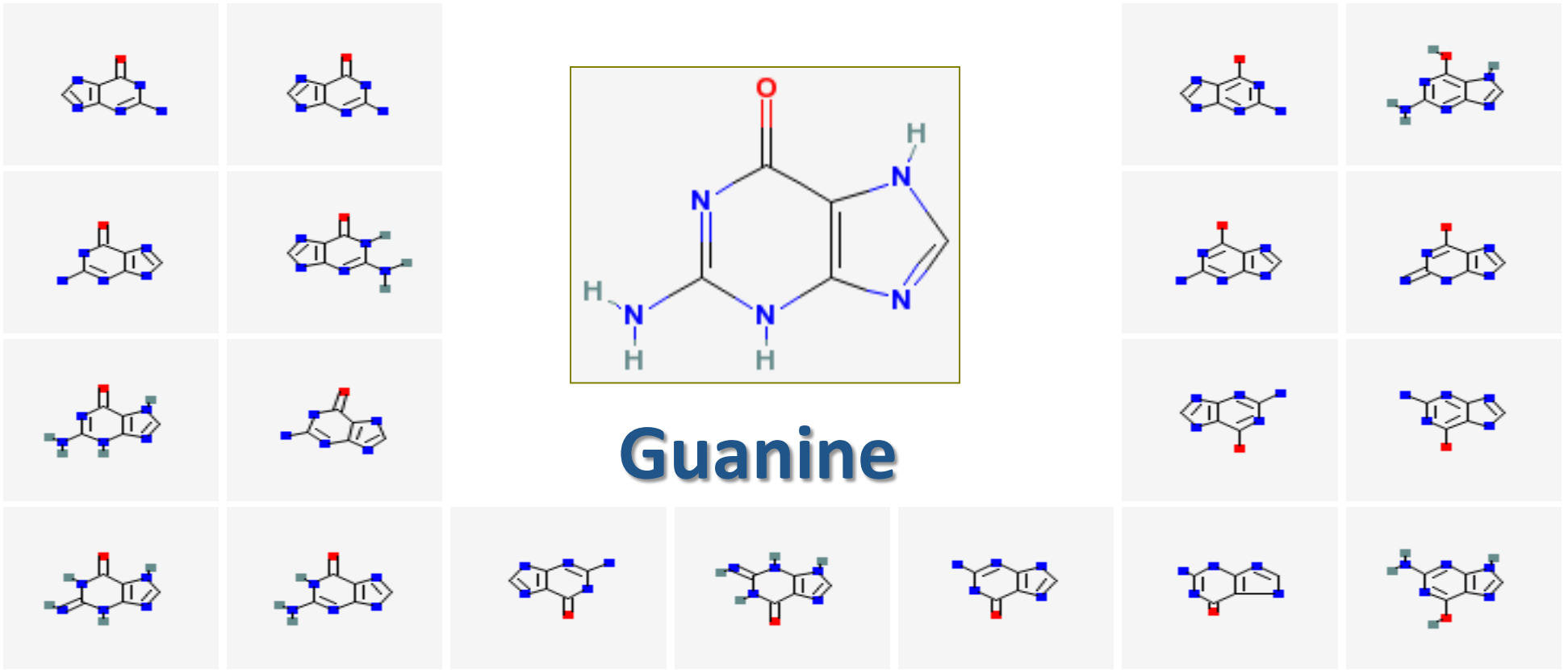
InChI=1S/C6H12O6/c7-1-2-3(8)4(9)5(10)6(11)12-2/h2-11H,1H2/t2-,3-,4+,5-,6+/m1/s1



WQZGKKKJIJFFOK-DVKNGEFBSA-N

There is no chemical information in an InChIKey … if you do not know the InChI, you cannot convert the InChIKey back into a chemical structure

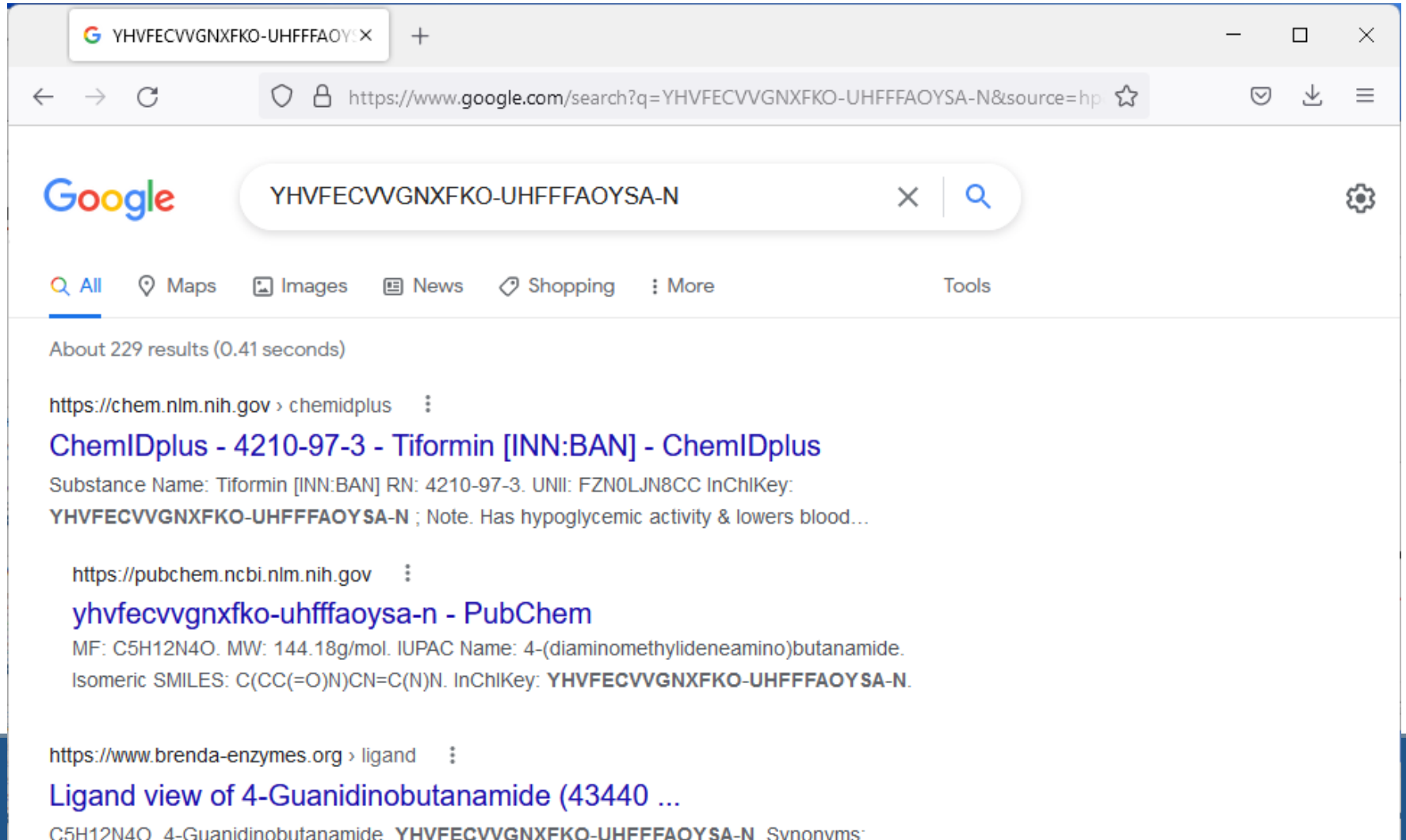
Different equivalent (tautomer) forms have different SMILES but same InChI

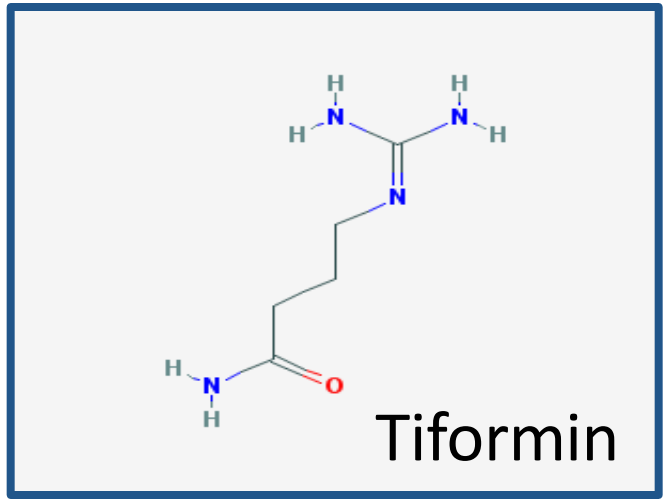


Tautomers and resonance forms of the same chemical structure are prolific

Internet Search Engines use InChIKey

They can use InChI too! .. but your mileage may vary

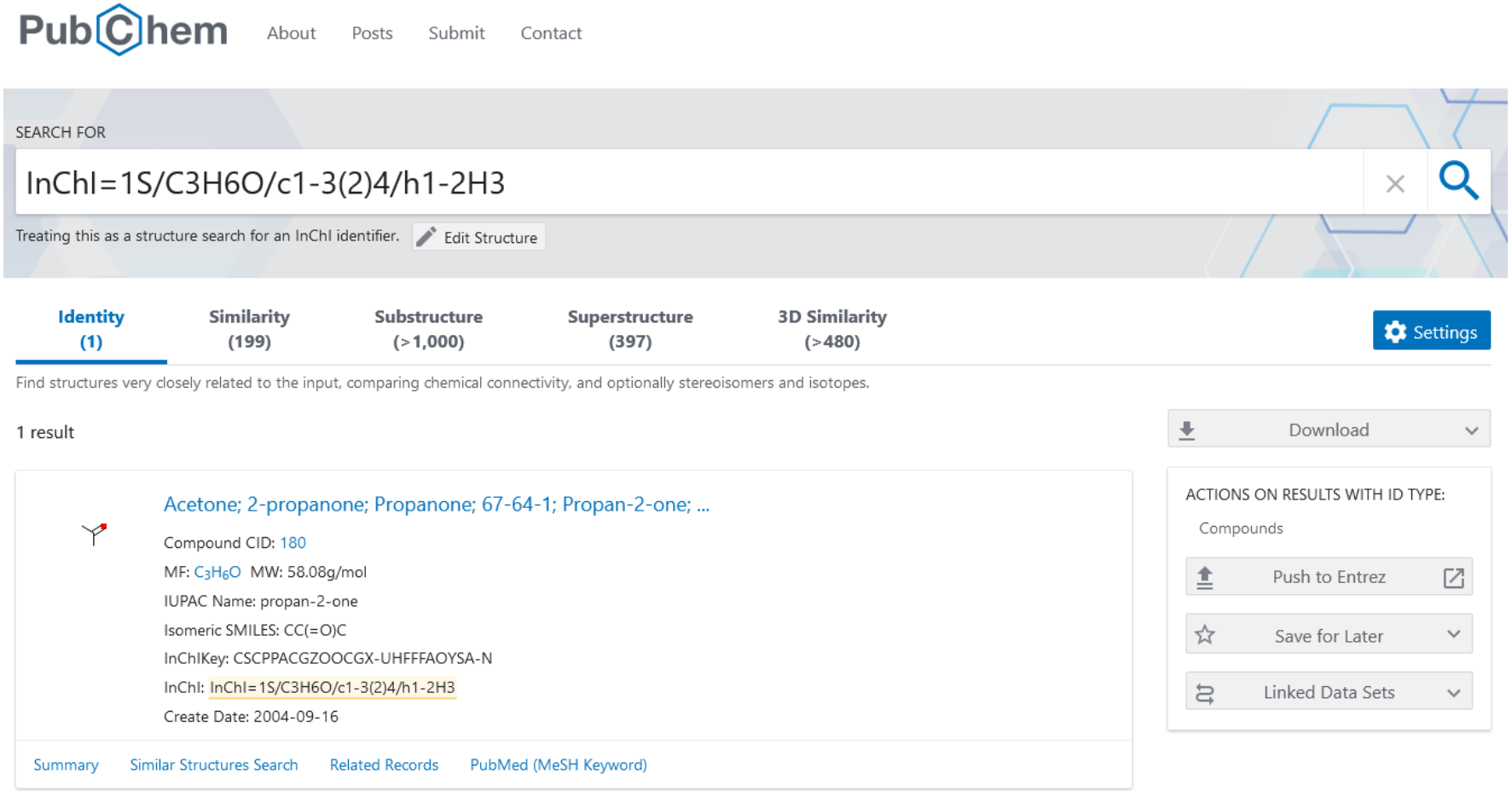




The many uses of InChI in PubChem

InChI is very heavily integrated into PubChem

Query by InChI



Query by InChIKey

PubChem About Posts Submit Contact

SEARCH FOR

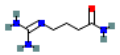
YHVFECVVGNXFKO-UHFFFAOYSA-N

Treating this as a text search.

Compounds (1) **Substances (2)**

Searching chemical names and synonyms including IUPAC names and InChIKeys across the compound collection. Note that annotations text from compound summary pages is not searched. Read More...

1 result



Tiformin; 4-carbamimidamidobutanamide; 4-guanidinobutyramide; 4-Guanidinobutanamide; 4210-97-3; ...

Compound CID: 123

MF: $C_5H_{12}N_4O$ MW: 144.18g/mol

IUPAC Name: 4-(diaminomethylideneamino)butanamide

Isomeric SMILES: C(CC(=O)N)CN=C(N)N

InChIKey: YHVFECVVGNXFKO-UHFFFAOYSA-N

InChI: InChI=1S/C5H12N4O/c6-4(10)2-1-3-9-5(7)8/h1-3H2,(H2,6,10)(H4,7,8,9)

Create Date: 2004-09-16

Summary Similar Structures Search Related Records

Download

Search in Entrez

ACTIONS ON RESULTS WITH ID TYPE:

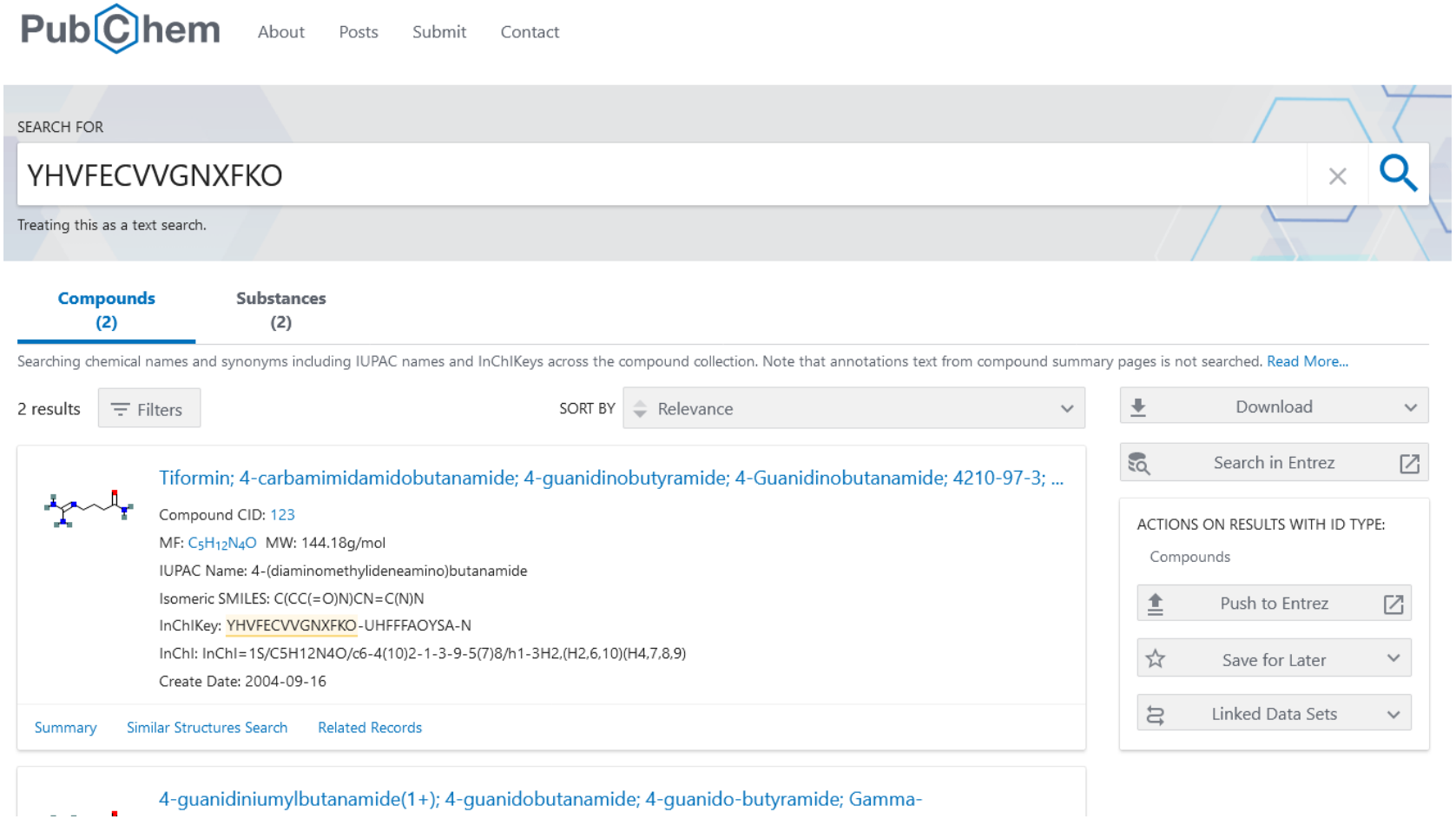
Compounds

Push to Entrez

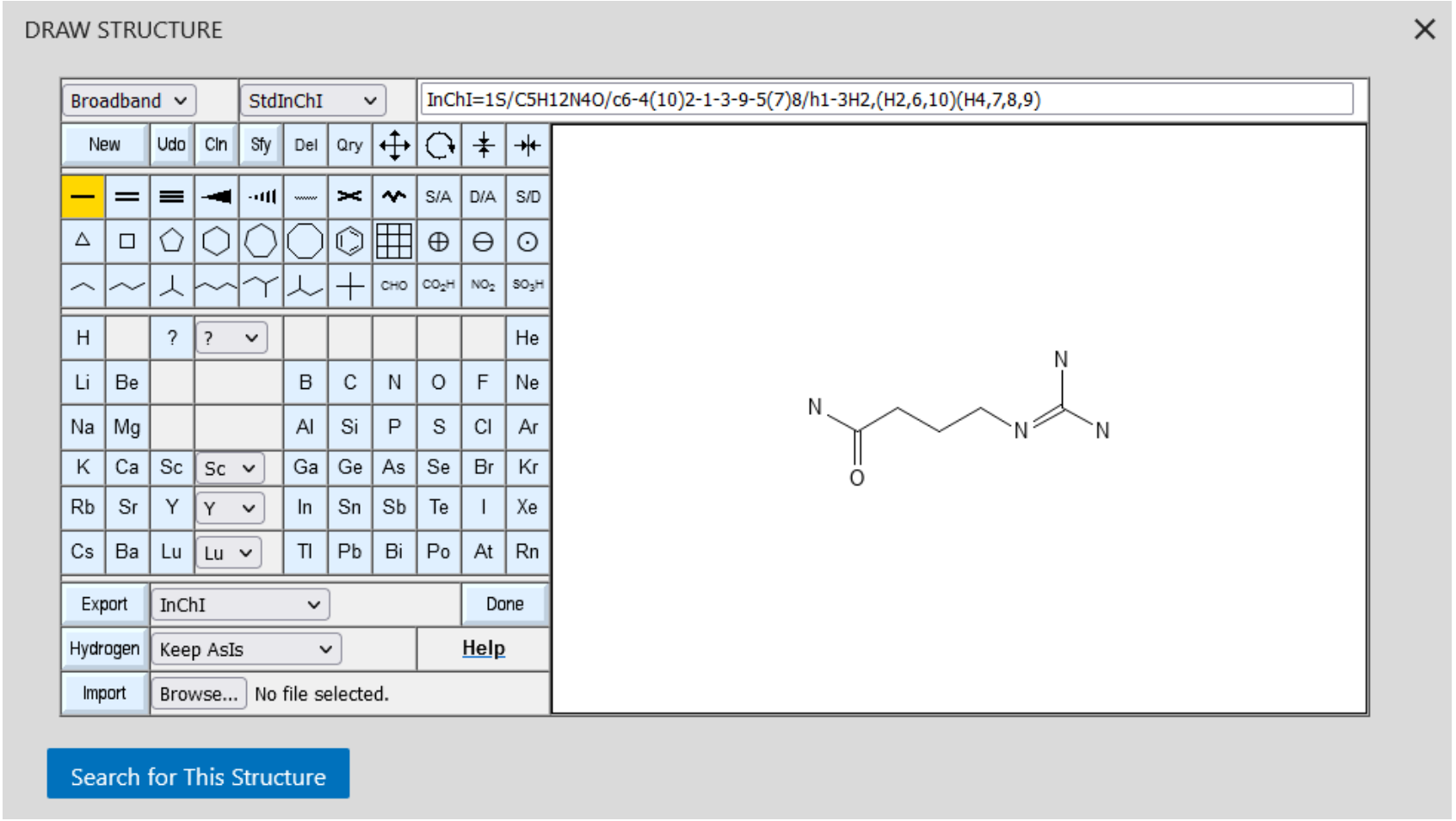
Save for Later

Linked Data Sets

Query by partial InChIKey



Computed in the PubChem Sketcher



Prominently displayed

PubChem Tiformin (Compound)

2 Names and Identifiers

2.1 Computed Descriptors

2.1.1 IUPAC Name

4-(diaminomethylideneamino)butanamide

Computed by Lexichem TK 2.7.0 (PubChem release 2021.05.07)

▸ PubChem

2.1.2 InChI

InChI=1S/C5H12N4O/c6-4(10)2-1-3-9-5(7)8/h1-3H2,(H2,6,10)(H4,7,8,9)

Computed by InChI 1.0.6 (PubChem release 2021.05.07)

▸ PubChem

2.1.3 InChIKey

YHVFECVVGNXFKO-UHFFFAOYSA-N

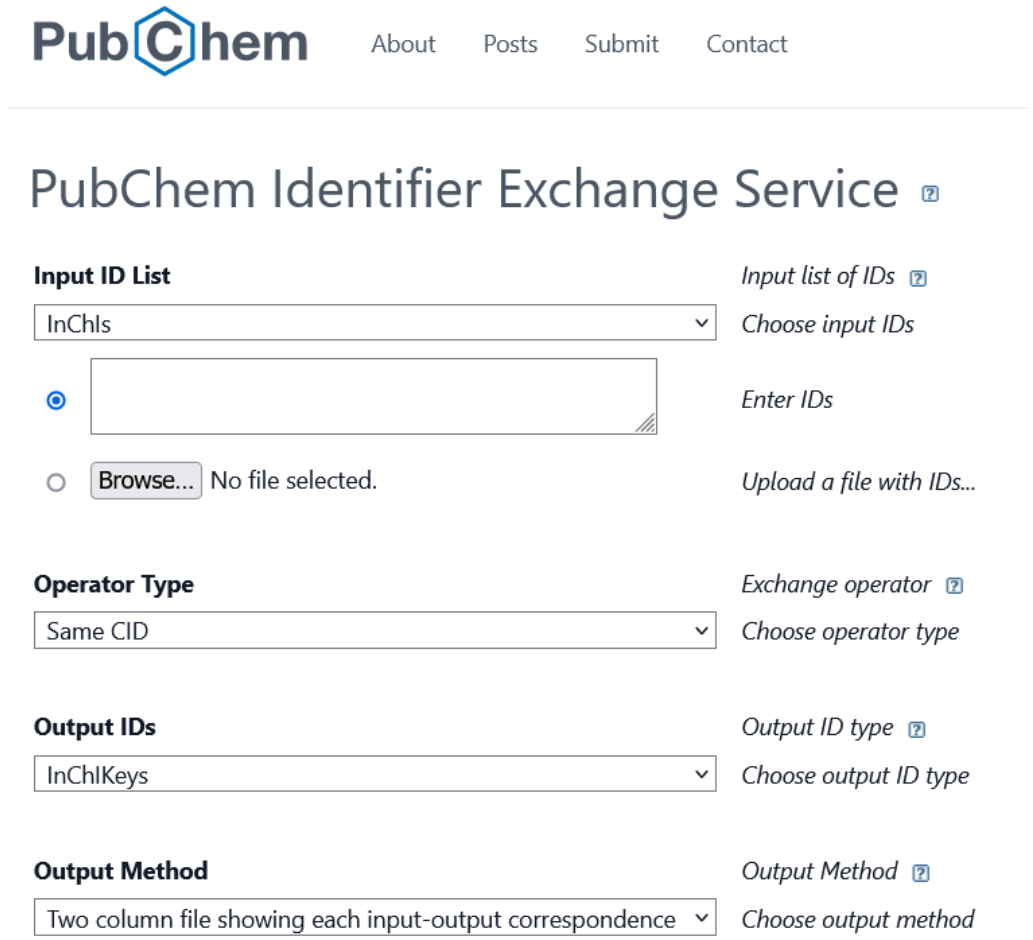
Computed by InChI 1.0.6 (PubChem release 2021.05.07)

▸ PubChem

Provided in downloads

```
> <PUBCHEM_IUPAC_NAME>
4-(diaminomethylideneamino)butanamide

> <PUBCHEM_IUPAC_SYSTEMATIC_NAME>
4-[bis(azanyl)methylideneamino]butanamide

> <PUBCHEM_IUPAC_TRADITIONAL_NAME>
4-guanidinobutyramide

> <PUBCHEM_IUPAC_INCHI>
InChI=1S/C5H12N4O/c6-4(10)2-1-3-9-5(7)8/h1-3H2,(H2,6,10)(H4,7,8,9)

> <PUBCHEM_IUPAC_INCHIKEY>
YHVFECVVGNXFKO-UHFFFAOYSA-N
```

Used as input/output in services



InChIKey used in PubChemRDF

4.4 PubChem InChIKey

PubChem InChIKey RDF triples expose the type, value and the link to the corresponding compound(s) for a given InChIKey. For example, to resolve the URI for the InChIKey with a value of "BSYNRYMUTXBXSQ-UHFFFAOYSA-N" (case-insensitive):
http://rdf.ncbi.nlm.nih.gov/pubchem/inchikey/BSYNRYMUTXBXSQ-UHFFFAOYSA-N

Link Type	Example RDF Triple
type	inchikey:BSYNRYMUTXBXSQ-UHFFFAOYSA-N rdf:type cheminf:CHEMINF_000399 .
value	inchikey:BSYNRYMUTXBXSQ-UHFFFAOYSA-N sio:has-value "BSYNRYMUTXBXSQ-UHFFFAOYSA-N"@en .
compound	inchikey:BSYNRYMUTXBXSQ-UHFFFAOYSA-N sio:is-attribute-of compound:CID2244 .

If the InChKey represents a chemical structure in the FDA UNII database , and the UNII code is incorporated as registry number in a MeSH concept, the annotation of InChIKey using MeSH concept is provided:

```
inchikey:ADUKCCWBEDSMEB-NSHDSACASA-N dcterms:subject <http://id.nlm.nih.gov/mesh/M0017537> .
```

Full set of InChI/Key are downloadable

Index of /pubchem/Compound/Extras

Name	Last modified	Size
Parent Directory		-
CID-Component.gz	2022-08-20 12:34	110M
CID-Component.gz.md5	2022-08-20 12:34	51
CID-Date.gz	2022-08-20 12:57	282M
CID-Date.gz.md5	2022-08-20 12:57	46
CID-IUPAC.gz	2022-08-20 15:03	1.5G
CID-IUPAC.gz.md5	2022-08-20 15:03	47
CID-InChI-Key.gz	2022-08-20 14:57	6.0G
CID-InChI-Key.gz.md5	2022-08-20 14:58	51
CID-LCSS.xml.gz	2022-08-20 13:43	224M
CID-LCSS.xml.gz.md5	2022-08-20 13:43	50
CID-Mass.gz	2022-08-20 14:15	1.1G
CID-Mass.gz.md5	2022-08-20 14:15	46
CID-MeSH	2022-08-20 12:26	4.0M
CID-MeSH.md5	2022-08-20 12:26	43
CID-PMID.gz	2022-08-20 14:19	359M
CID-PMID.gz.md5	2022-08-20 14:19	46
CID-Parent.gz	2022-08-20 13:01	466M
CID-Parent.gz.md5	2022-08-20 13:01	48
CID-Patent.gz	2022-08-20 15:03	7.6G
CID-Patent.gz.md5	2022-08-20 15:04	48

CID-Parent.gz:

This is a listing of all CIDs with parents. It is a gzipped text file with CID, tab, and parent on each line. Note that a CID may be a parent of itself, or may have no parent.

CID-InChI-Key.gz:

This is a listing of all CIDs with their full InChI strings and InChI keys. It is a gzipped text file with CID, tab, InChI, tab, InChI Key on each line.

CID-SMILES.gz:

This is a listing of all CIDs with their isomeric SMILES. It is a gzipped text file with CID, tab, SMILES on each line.

https://ftp.ncbi.nlm.nih.gov/pubchem/Compound/Extras/

InChI enables use cases…

- PubChem Compound TOC ? 49,493,641
  - Agrochemical Information ? 3,045
  - Associated Disorders and Diseases ? 20,847
  - Biologic Description ? 2,056,521
  - Biological Test Results ? 3,622,920
  - Biomolecular Interactions and Pathways ? 125,253
  - Chemical and Physical Properties ? 263,015
  - Classification ? 1,454,824
  - Drug and Medication Information ? 17,922
  - Food Additives and Ingredients ? 8,414
  - Identification ? 4,968
  - Information Sources ? 20,271,277
  - Literature ? 1,833,941
  - Names and Identifiers ? 1,275,170
  - Patents ? 36,351,418
  - Pharmacology and Biochemistry ? 110,628
  - Related Records ? 9,224,590
  - Safety and Hazards ? 149,319
  - Spectral Information ? 480,730
  - Structures ? 9,117,635
  - Toxicity ? 114,012
  - Use and Manufacturing ? 115,321
- Chemical Safety ? 147,023

J Cheminform. 2021 Mar 8;13(1):19. doi: 10.1186/s13321-021-00489-0.

Empowering large chemical knowledge bases for exposomics: PubChemLite meets MetFrag

Emma L Schymanski [1], Todor Kondić [2], Steffen Neumann [3 4], Paul A Thiessen [5], Jian Zhang [5], Evan E Bolton [6]

Affiliations + expand

PMID: 33685519 PMCID: PMC7938590 DOI: 10.1186/s13321-021-00489-0

Free PMC article

PubChemLite EXPOSOMICS

~400,000 entries “small”

Breakdown of PubChem for exposomics use by annotation availability using unique InChIKey first block to group and gather different stereo forms, charge states, and salt forms

Many thanks to Dr. Emma Schymanski for portions of this slide

PubChem-InChI limitations/future

- Organometallics
- Tautomers
- Inorganics
- Mixtures
- Reactions
- Polymers

- PubChem normalization
  - Many CIDs to one InChI
  - Many InChI to one CID
- InChI is a descriptor
- InChI is used as a structural format

Igor Pletnev

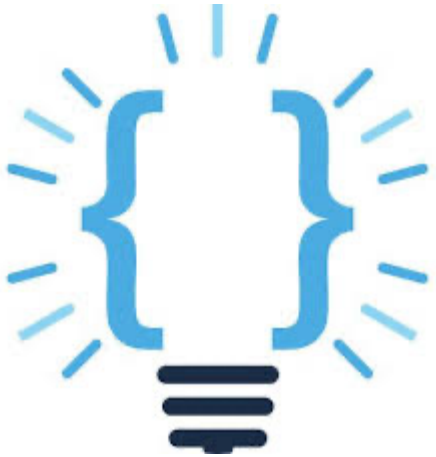
Many words to describe Igor

Professional thoughts on Igor

- Enjoyed working with him
- Took great care in technical documentation
- Translated aspirations into reality
- Put blood sweat and tears into his efforts
- Directly enabled InChI to be successful

Image credit:
https://leadg2.thecenterforsalesstrategy.com/hubfs/C-Suite%20Thought%20Leadership.png

Igor enabled InChI



- Interpreter in chief
  - code needs details not aspirations
- Coordinator in chief
  - many people involved with InChI, each with ideas
- Implementor in chief
  - Igor advanced InChI with each line he wrote

Image credit:
https://encrypted-tbn0.gstatic.com/images?q=tbn:ANd9GcTBzxQE6d4CT8OBYIJCwv3IiqrQe3vxX8r6Fw&usqp=CAU

PubChem Crew …

Evan Bolton

Jie Chen

Tiejun Cheng

Asta Gindulyte

Jane He

Siqian He

Sunghwan Kim

Qingliang Li

Ben Shoemaker

Paul Thiessen

Bo Yu

Leonid Zaslavsky

Jian Zhang

Special thanks to the NCBI Help Desk, the NCBI Systems team, and past PubChem group members

Special thanks

- Exposomics collaborators (NORMAN SLE, Emma Schymanski)
- All PubChem Contributors and Collaborators
- Software collaborators
  - NextMove Software (Roger Sayle, John May) and Daniel Lowe / Noel O'Boyle
  - Xemistry GmbH (Wolf D. Ihlenfeldt)
  - OpenEye Scientific Software

- This research was supported by the National Center for Biotechnology Information of the National Library of Medicine (NLM), National Institutes of Health