Plumbing for Chemistry

Stephen Heller, PhD

BS Chemistry (1 of 5) Class of ‘63

The main site for the NIST mas spectral database is:

https://www.nist.gov/srd/nist-standard-reference-database-1a-v17

The main web sites for the IUPAC InChI project are:

http://www.iupac.org/inchi

And

http://www.inchi-trust.org

10/19/2018

Slides are available at http://www.hellers.com/steve/SUNY-SB-10-18.pdf

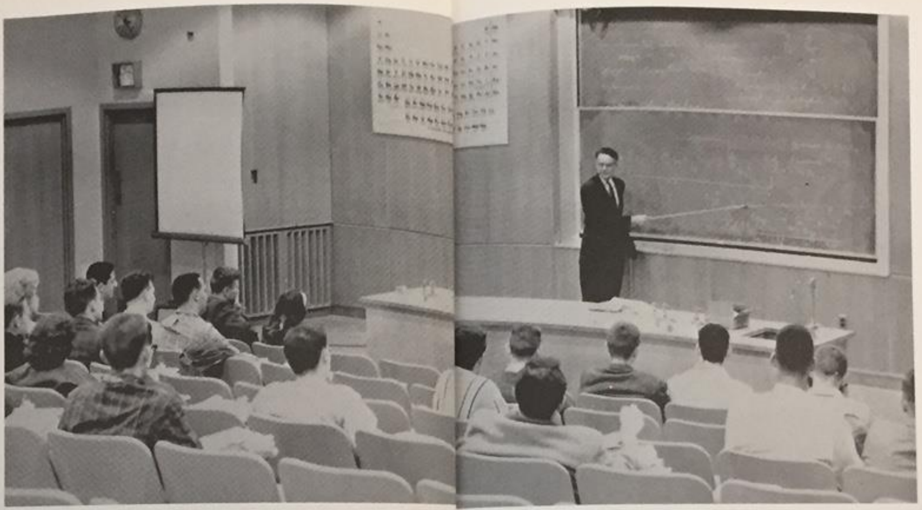
Outline

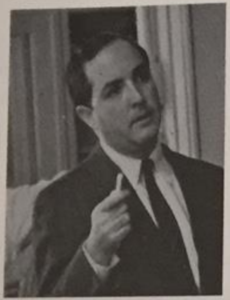
- The value of being in the right place at the right time
- The start in the Chemistry Department at Stony Brook
- Two major projects
    - The NIH/EPA/NIST Mass Spectral Data Base
    - IUPAC InChI Chemical Structure Standard

This is a green talk

These slides were made from 100% recycled electrons

FRANCIS T. BONNER
Chairman





FAUSTO RAMIREZ

SEI SUJISHI

CHEMISTRY

EDWARD M. KOSOWER

PAUL C. LAUTERBUR

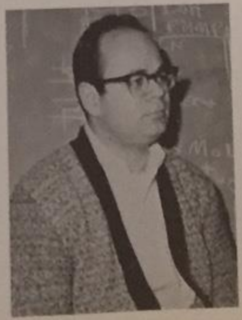
THEODORE GOLDFARB

BARRY M. GORDON

WILLIAM J. LeNOBLE

ARTHUR R. LEPLEY

ROBERT SCHNEIDER



RICHARD SOLO

Chemistry, Stony Brook, and the Early 1960s

* No graduate students the first few years

* I was the first summer NMR technician for the future Nobel Laureate Paul Lauterbur

* Professor Fausto Ramirez and the search for the correct phosphate structure

F. Ramirez, A. V. Patwardhan, and S. R. Heller, The Reaction of Trialkyl Phosphites with Aliphatic Aldehydes. P-31 and H-1 Nuclear Magnetic Resonance Spectra of Tetraoxyalkyl Phosphoranes, J. Am. Chem. Soc., 86, 514-516(1964)

The Mass Spectral Database & Search System - MSSS

* Early 1970s

* NIH Division of Computer Research and Technology

Scotty Pratt, Hank Fales

* Showing off what computers could do

Timesharing computers

No internet, 110 baud modems

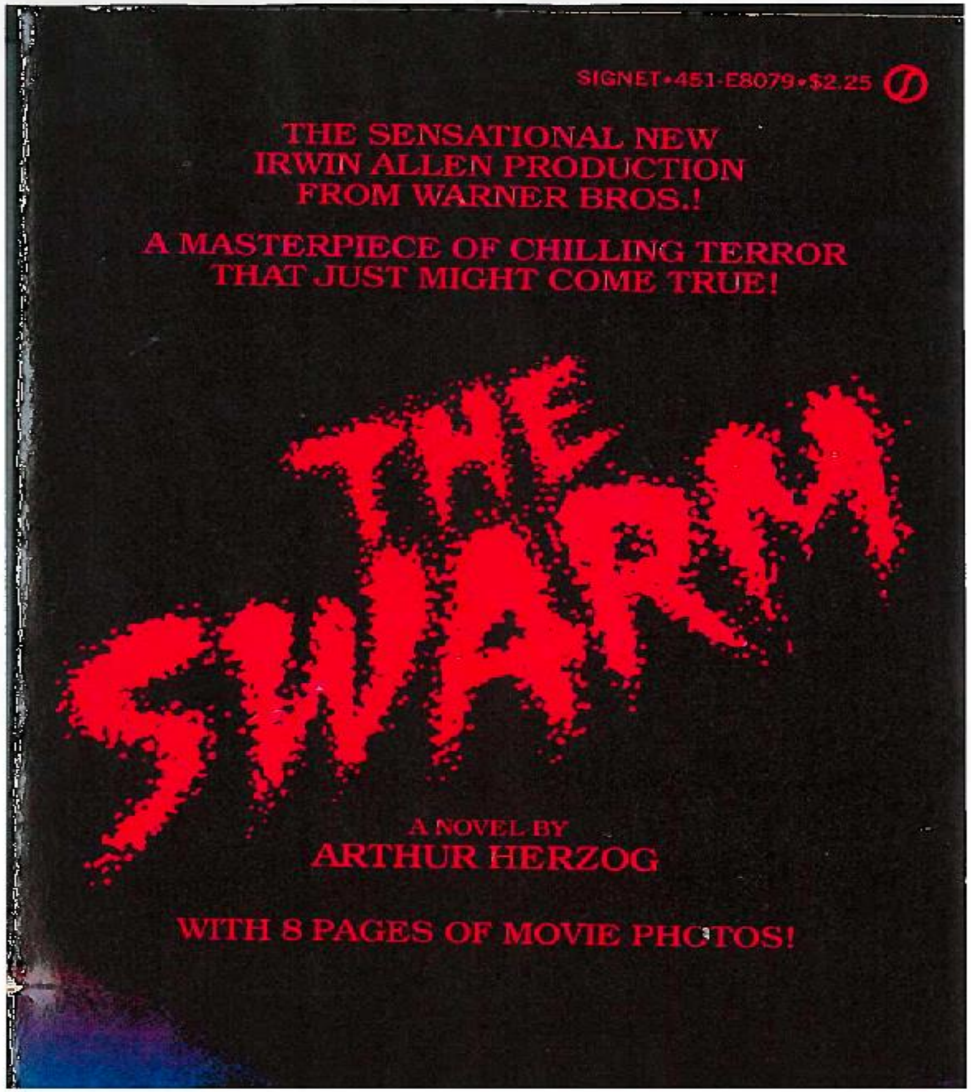
* Public was invited to experiment

* 30,000 El spectra in 1975; 10 times that today

* 6,000 copies sold per year

* Generates in excess of $5 million for NIST

https://www.nist.gov/srd/nist-standard-reference-database-1a-v17



SIGNET • 451-E8079 • $2.25

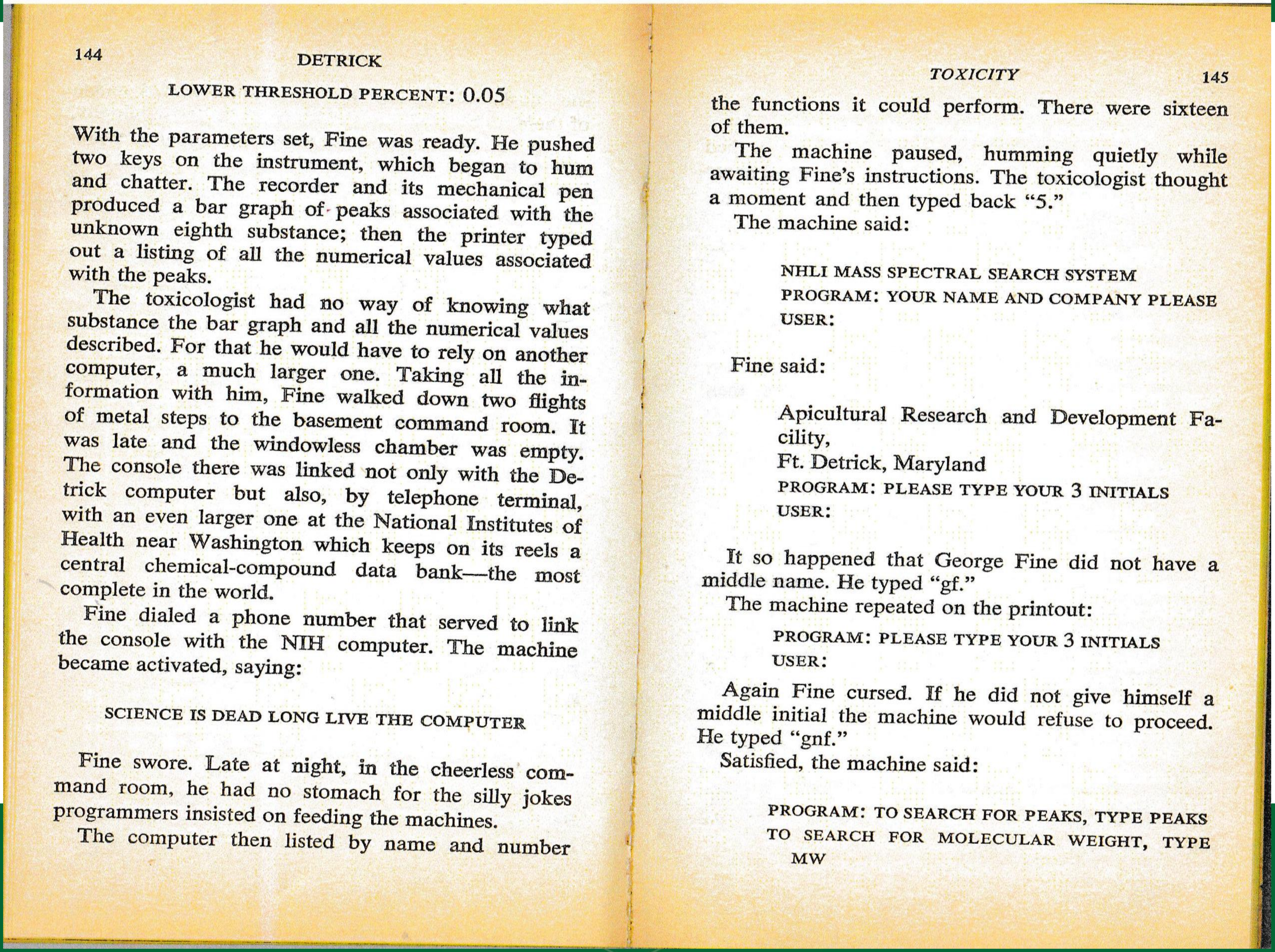
THE SENSATIONAL NEW
IRWIN ALLEN PRODUCTION
FROM WARNER BROS.!

A MASTERPIECE OF CHILLING TERROR
THAT JUST MIGHT COME TRUE!

THE SWARM

A NOVEL BY
ARTHUR HERZOG

WITH 8 PAGES OF MOVIE PHOTOS!



```
LOWER THRESHOLD PERCENT: 0.05
```

With the parameters set, Fine was ready. He pushed two keys on the instrument, which began to hum and chatter. The recorder and its mechanical pen produced a bar graph of peaks associated with the unknown eighth substance; then the printer typed out a listing of all the numerical values associated with the peaks.

The toxicologist had no way of knowing what substance the bar graph and all the numerical values described. For that he would have to rely on another computer, a much larger one. Taking all the information with him, Fine walked down two flights of metal steps to the basement command room. It was late and the windowless chamber was empty. The console there was linked not only with the Detrick computer but also, by telephone terminal, with an even larger one at the National Institutes of Health near Washington which keeps on its reels a central chemical-compound data bank—the most complete in the world.

Fine dialed a phone number that served to link the console with the NIH computer. The machine became activated, saying:

```
SCIENCE IS DEAD LONG LIVE THE COMPUTER
```

Fine swore. Late at night, in the cheerless command room, he had no stomach for the silly jokes programmers insisted on feeding the machines.

The computer then listed by name and number

the functions it could perform. There were sixteen of them.

The machine paused, humming quietly while awaiting Fine's instructions. The toxicologist thought a moment and then typed back "5."

The machine said:

```
NHLI MASS SPECTRAL SEARCH SYSTEM
PROGRAM: YOUR NAME AND COMPANY PLEASE
USER:
```

Fine said:

```
Apicultural Research and Development Facility,
Ft. Detrick, Maryland
PROGRAM: PLEASE TYPE YOUR 3 INITIALS
USER:
```

It so happened that George Fine did not have a middle name. He typed "gf."

The machine repeated on the printout:

```
PROGRAM: PLEASE TYPE YOUR 3 INITIALS
USER:
```

Again Fine cursed. If he did not give himself a middle initial the machine would refuse to proceed. He typed "gnf."

Satisfied, the machine said:

```
PROGRAM: TO SEARCH FOR PEAKS, TYPE PEAKS
TO SEARCH FOR MOLECULAR WEIGHT, TYPE
  MW
```

The Birth of InChI

*   Issues relating to CAS Registry Numbers in the late 1990’s

*   NIST mass spec compound registration software for finding replicate spectra for the same compound).

*   Chemical names - NOT the future for chemical structures

IUPAC and the Birth of InChI

* IUPAC convened a meeting in March 2000 in Washington.

Ted Becker (NIH) and Alan McNaught (RSC)

* NIST offered to provide staff to create and program this chemical identifier standard for IUPAC.

InChI Project Goal

* But before you can share and use data and information do you need to find it.

* Link everything (data and information) about a chemical

  * from many and varied sources

  * purpose of creating new information and perhaps/hopefully knowledge.

Unique InChI Features

* Only IUPAC International structure standard

* Open Source structure standard

* Only structure standard support by
  majority of publishers
  database producers
  chemistry software companies

InChI Characteristics

1. Easy to generate

2. Expressive (it will contain structural information)

3. Unambiguous/Unique

4. Does not require a centralized operation (it can be generated anywhere – can use crowdsourcing/free labor)

5. Easy to search for structure via Internet search engines (Google, Yahoo, Bing, etc.) using the InChI (hash) Key.

What “*is*” the InChI standard?

The InChI standard programmed into the algorithm is an arbitrary decision as to how structures are handled. In most cases there is total agreement (e.g., methane).

In cases of more complex molecules where there is not agreement among chemists, one representation is chosen. As long as the arbitrarily chosen representation is properly programmed, one will always get the SAME result using it – which is what a standard is!

InChI layered structure design

The current InChI layers are:

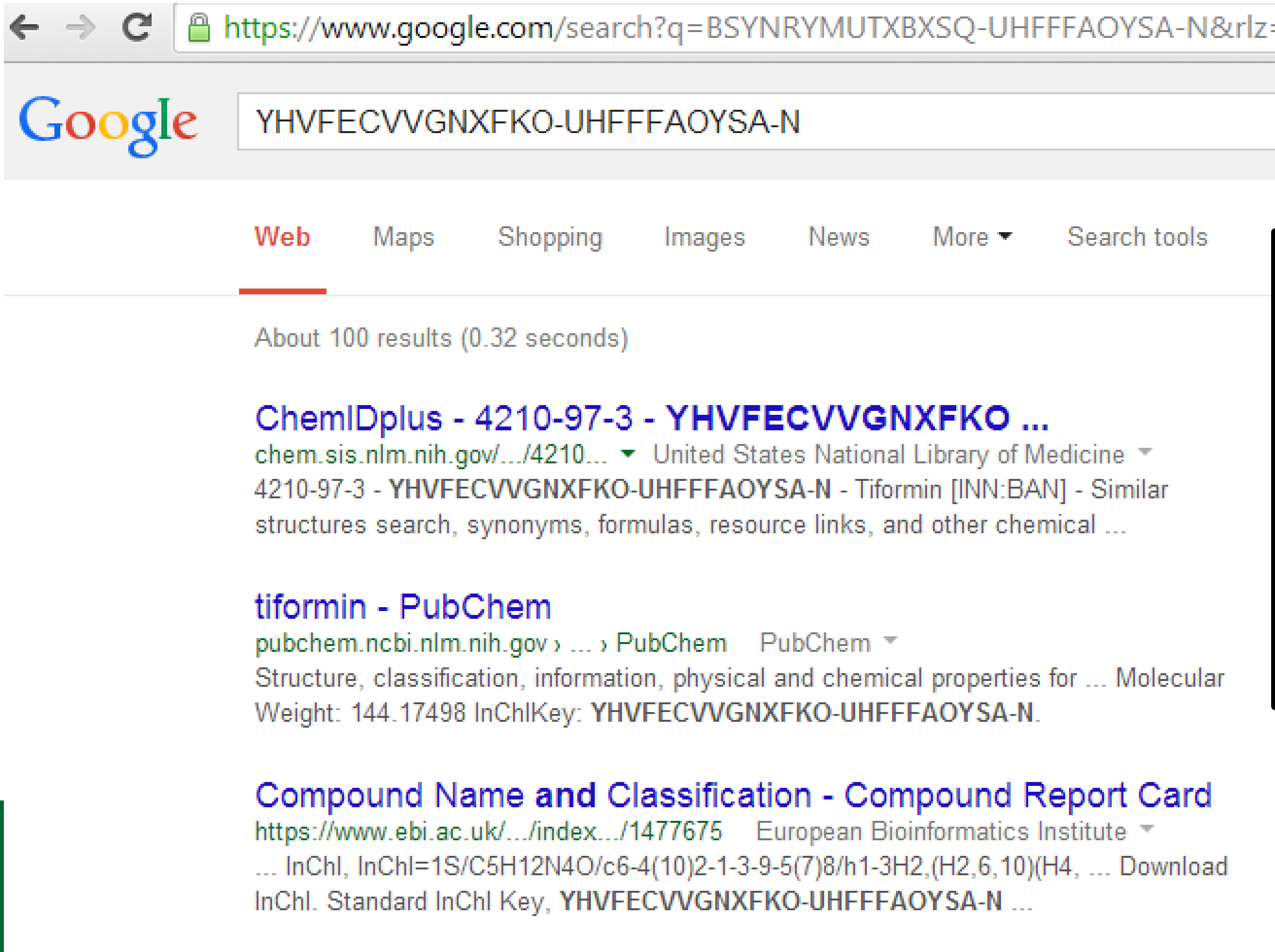
1. Formula
2. Connectivity (no formal bond orders)
    a. disconnected metals
    b. connected metals
3. Isotopes
4. Stereochemistry
    a. double bond *(Z/E)*
    b. tetrahedral (sp3)
5. Tautomers (on or off)

Charges are added to end of the string

The InChI Algorithm normalizes chemical representation and includes a “standardized” InChI, and the ‘hashed’ form called the InChIKey

Search Engines can use InChIKey

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



https://www.google.com/search?q=BSYNRYMUTXBXSQ-UHFFFAOYSA-N&rlz=1C1CHFX_enUS4

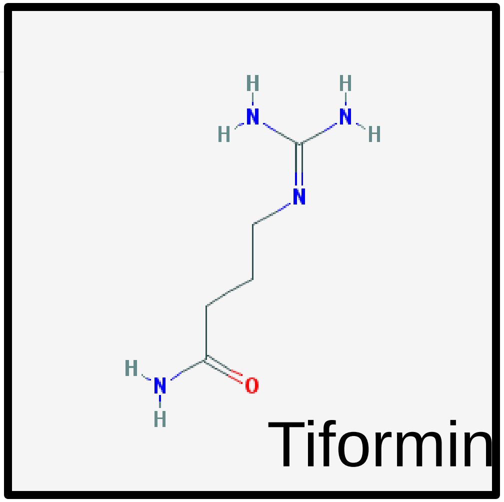
Google YHVFECVVGNXFKO-UHFFFAOYSA-N

About 100 results (0.32 seconds)

ChemIDplus - 4210-97-3 - YHVFECVVGNXFKO ...
chem.sis.nlm.nih.gov/.../4210... United States National Library of Medicine
4210-97-3 - **YHVFECVVGNXFKO-UHFFFAOYSA-N** - Tiformin [INN:BAN] - Similar structures search, synonyms, formulas, resource links, and other chemical ...

tiformin - PubChem
pubchem.ncbi.nlm.nih.gov › ... › PubChem PubChem
Structure, classification, information, physical and chemical properties for ... Molecular Weight: 144.17498 InChIKey: **YHVFECVVGNXFKO-UHFFFAOYSA-N**.

Compound Name and Classification - Compound Report Card
https://www.ebi.ac.uk/.../index.../1477675 European Bioinformatics Institute
... InChI, InChI=1S/C5H12N4O/c6-4(10)2-1-3-9-5(7)8/h1-3H2,(H2,6,10)(H4, ... Download InChI. Standard InChI Key, **YHVFECVVGNXFKO-UHFFFAOYSA-N** ...



Tiformin

InChI Standard

Developing a standard and having it accepted are two very different things.

InChI Standard

"Standards are like toothbrushes – everyone has one but no one wants to use someone else's."

Phil Bourne,
Former Associate Director for Data Science (Big Data), NIH

The Kasson Metric System Act of the US Congress back in ’66 is an example of slower acceptance of a standard.

How difficult is it to create an InChI?

Today, all the major structure drawing programs have incorporated the InChI algorithm in their products, with usually an “InChI” button for generating the InChI.

Current InChI Status

*	InChI can handle simple organic molecules

*	99%+ of what scientists use every day

InChI is the worst computer readable structure representation except for all those other forms that have been tried from time to time.

With apologies to Sir Winston Churchill (House of Commons speech on November 11, 1947)

Large Databases with InChIs/InChIKeys

EBI UniChem –157 million
NIH/NCI – 110 million
NIH/PubChem - 97 million active
RSC/ChemSpider – 67 million
Elsevier/Reaxys – 30 million
IUPAC – 0 million

InChI Videos

1. What on Earth is InChI?

http://www.youtube.com/watch?v=rAnJ5toz26c

2. The Birth of the InChI

http://www.youtube.com/watch?v=X9c0PHXPfso

3. The Googlable InChIKey

http://www.youtube.com/watch?v=UxSNOtv8Rjw

4. InChI and the Islands

http://www.youtube.com/watch?v=qrCqJ0o4jGs

Success is uncoerced adoption

If you are not part of the solution; you are part of the precipitate

steve@hellers.com

Slides are available at http://www.hellers.com/steve/SUNY-SB-10-18.pdf