An Open Standard for Chemical Structure Representation

The IUPAC Chemical Identifier

S. E. Stein, S. R. Heller, D. V. Tchekhovskoi
Physical & Chemical Properties Division
NIST

"Classification and Categorization"
International Chemical Information Conference
Nimes, France
October 22, 2003

IUPAC was formed in 1919 .. for international standardization in chemistry. The standardization of weights, measures, names and symbols is essential to .. scientific enterprise and the growth of international trade and commerce

NIST

United States Department of Commerce
National Institute of Standards and Technology

I U P A C

Current Project

Chemical Nomenclature and Structure Representation Division (VIII)

Number: 2000-025-1-800

Title: IUPAC Chemical Identifier (IChI)

Task Group Chairman: A. McNaught

Members: S. Heller and S. Stein

Completion Date: 2003

> 1 July 2002 <
CAS/IUPAC Conference on Chemical Identifiers and XML for Chemistry

> *clipping*

Communication of Chemical Identity

- Human
  - Verbal – Common name
  - Text – Systematic/Common name
  - Pictorial – Structure diagram
- Computers
  - Electronic - Precise standards

Digital ‘Naming’ of Chemicals

- Chemical structure is the true ‘identifier’
- But, structure representations are not unique or convenient for computers.
- So, convert structure to a unique ‘name’ by fixed algorithms
  - The Iupac CHemical Identifier (IChI)

Customer Needs

- “Authors”
  - Precise
  - Convention-free
  - Wide coverage
- “Readers”
  - Robust
  - Variable specificity
  - Long life
- “Publishers” (Software)
  - Ready access

Two Problems

- Chemicals
  - Fast isomerization (tautomerization)
  - Ill-defined connectivity
- Chemists
  - Differing conventions
    - Depends on discipline, education and convenience
  - Imprecision/uncertainty

3 Steps to IChI

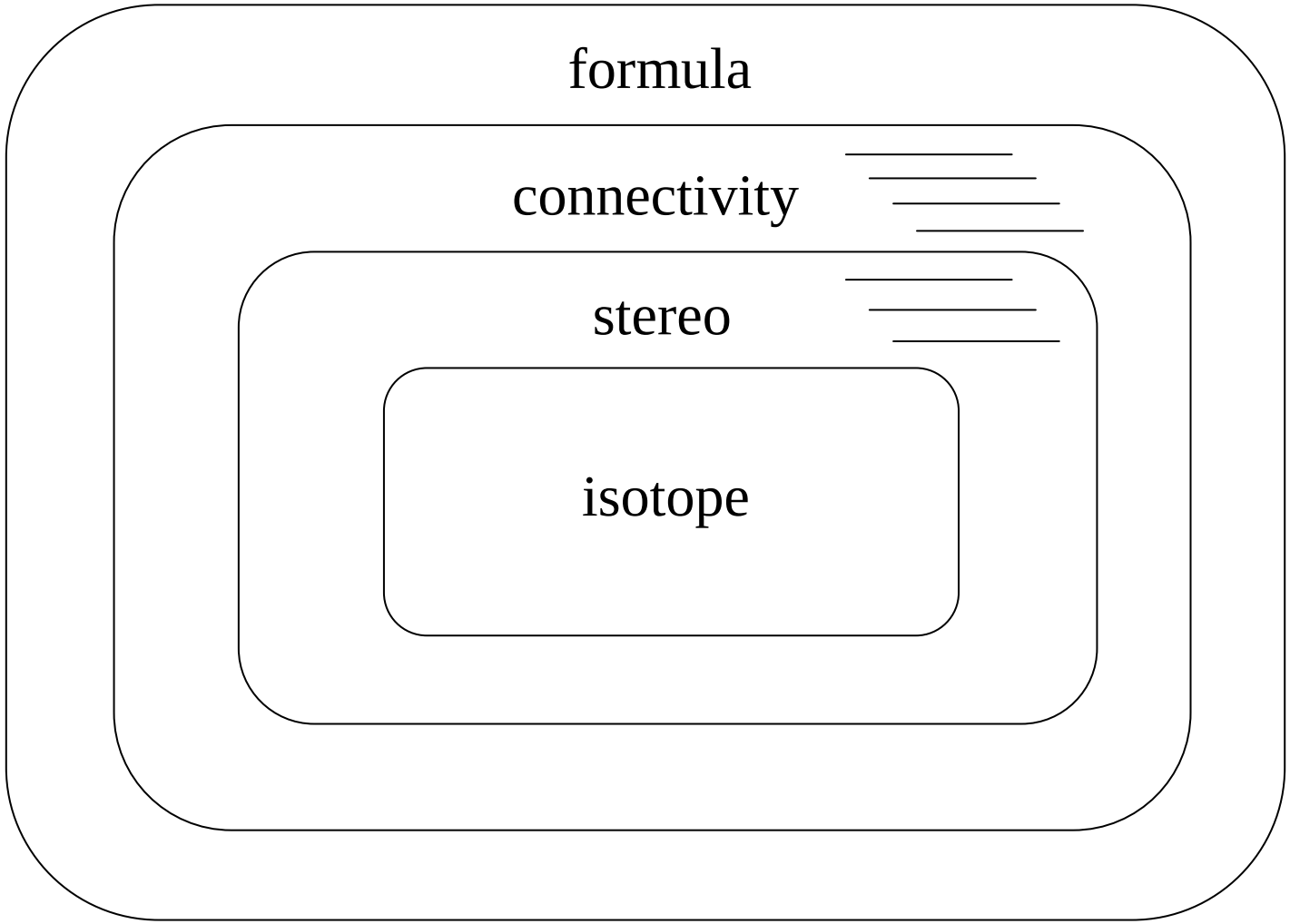
- Chemistry
  - ‘Normalize’ Input Structure
    - Implement chemical rules
- Math
  - ‘Canonicalize’ (label the atoms)
    - Equivalent atoms get the same label
- Format
  - ‘Serialize’ Labeled Structure
    - Output as character string (‘name’)

Normalize
Simplify

- Divide structure into ‘layers’
  - Each layer ‘refines’ structure
- Ignore ‘Electron Density’
  - Use simple ‘connectivity’ only
  - Ignore bond type and electron location
- Stereochemistry
  - sp^2 and sp^3 only
  - Free rotation around single bonds
  - No Z/E stereo for small rings (default)

“Layers”

Chemical Substances

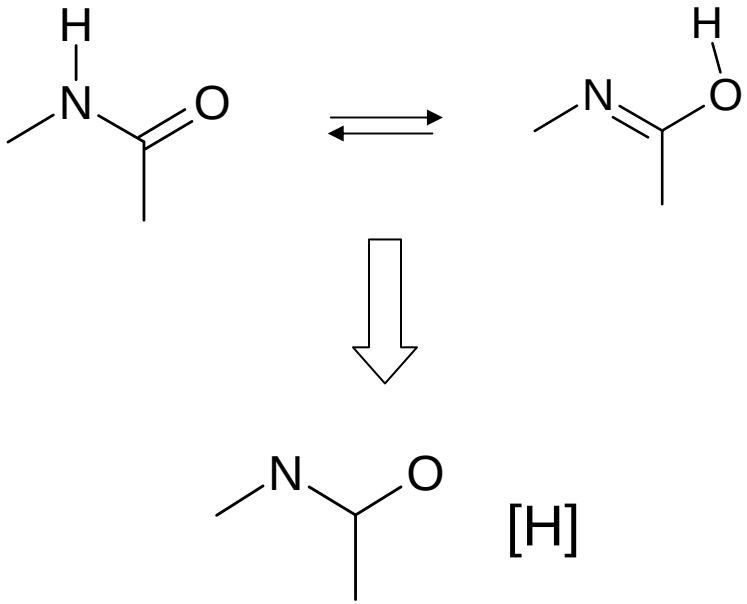


4 Connectivity ‘Sublayers’

- Disconnect metals and H-atoms
  - Skeleton
- Reconnect fixed H-atoms
  - Tautomerism
- Reconnect mobile H-atoms (optional)
  - All connections fixed
- Reconnect metals (optional)
  - Represent bonds to metals

Tautomer Sublayer

H-migration between 1,3-heteroatoms



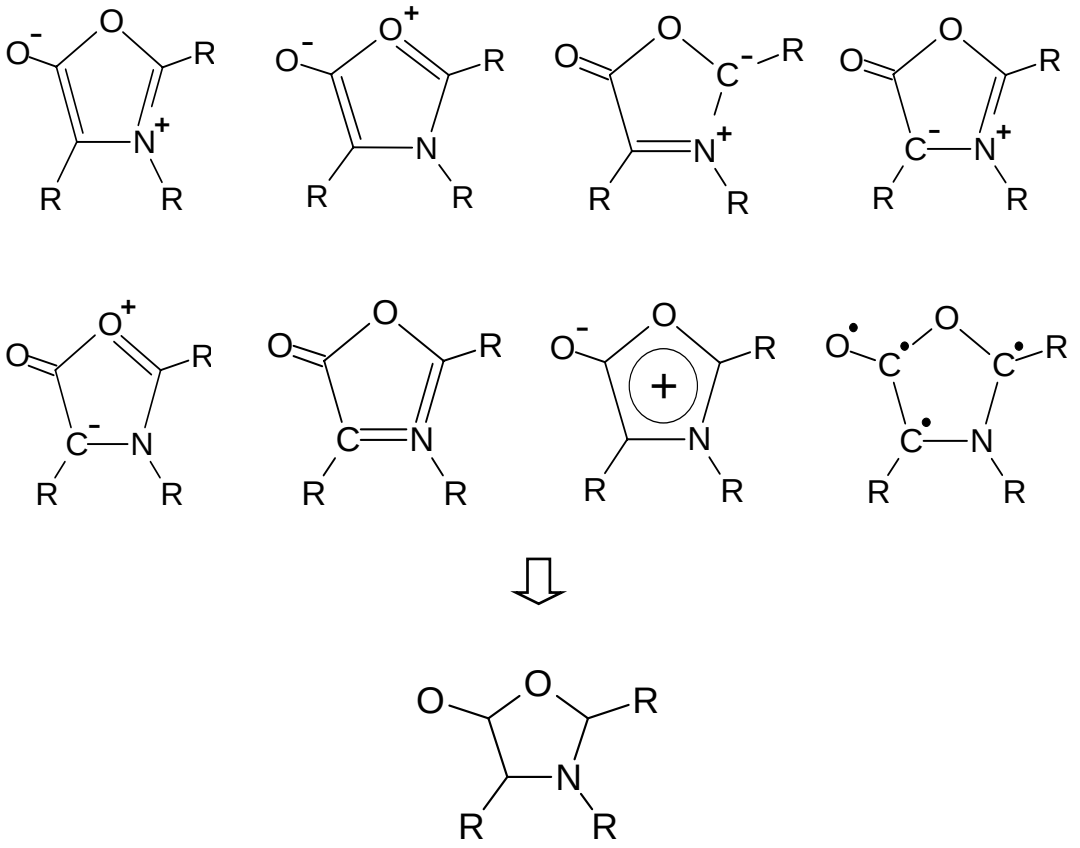
Stereochemical Sublayers

- sp^2 – double bond
- sp^3 – tetrahedral
- {others added later}

- relative, absolute or racemic

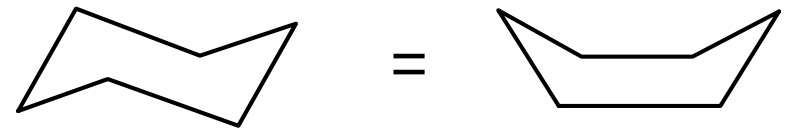
Simplify - Ignore Electrons

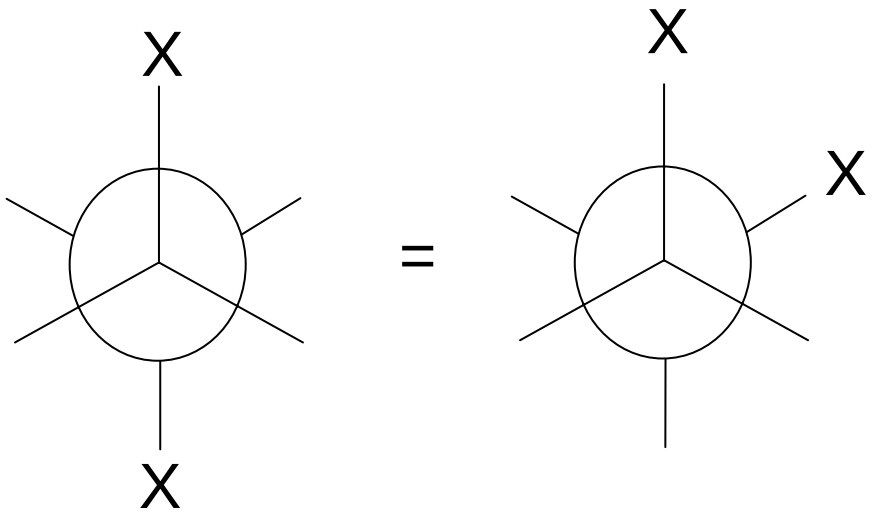
Münchnones



Simplify - Limit Stereo

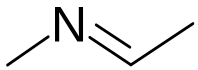
Assume Free Rotation Around Single Bonds

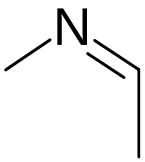


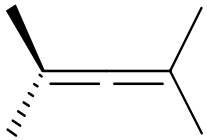


No Conformers

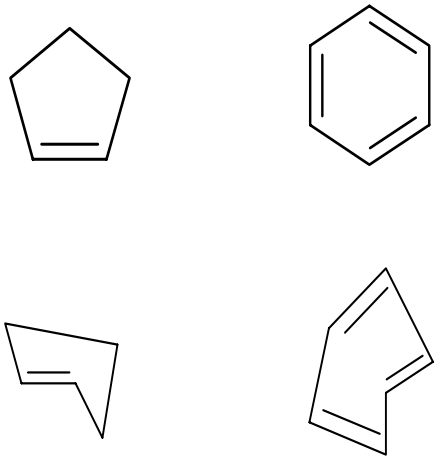
Simplify – Double Bond Stereo





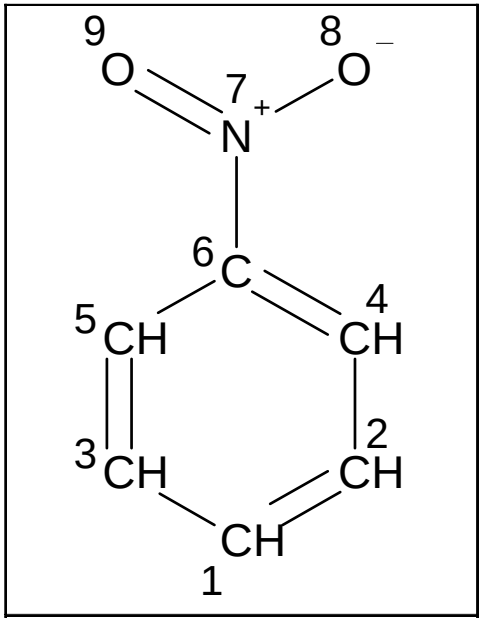


Rules



Ignore stereo for small rings

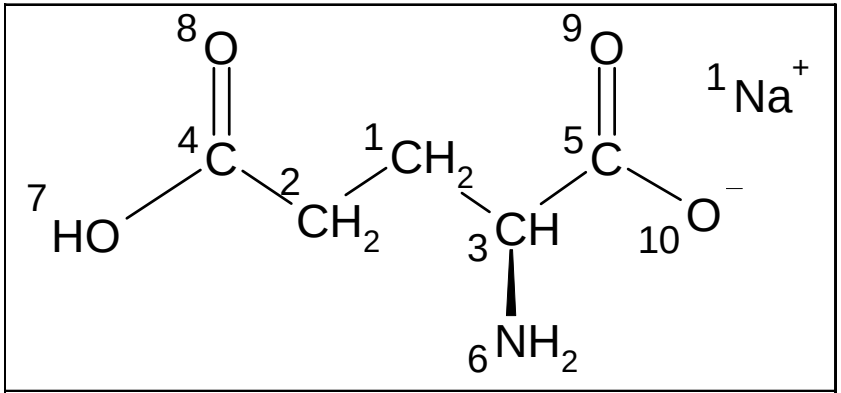
Nitrobenzene



Canonical numbering

Description	Layers
formula	C6H5NO2
connectivity	8-7(9)6-4-2-1-3-5-6
H-atoms	1-5H
charges	

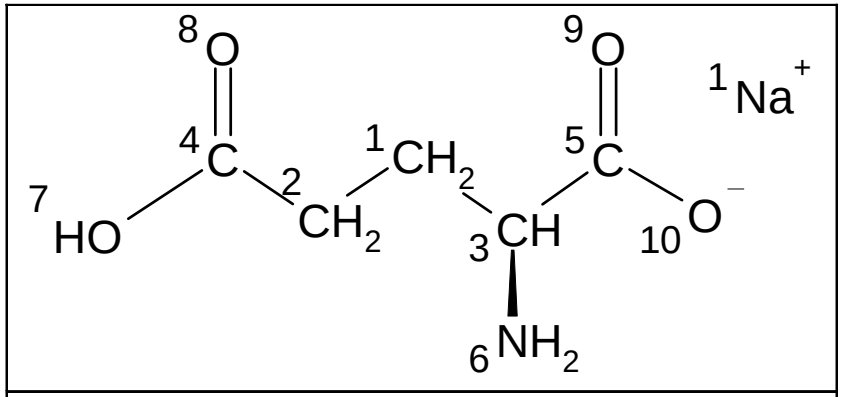
MSG (tautomeric)



Canonical numbering

Description	Layers
formula	**C5H8NO4.Na**
connectivity	**6-3(5(9)10)1-2-4(7)8;**
H-atoms	**1-2H2,3H,6H2(H-,7,8,9,10);**
stereo sp^3	**3-;**
charges	**-1;+1**

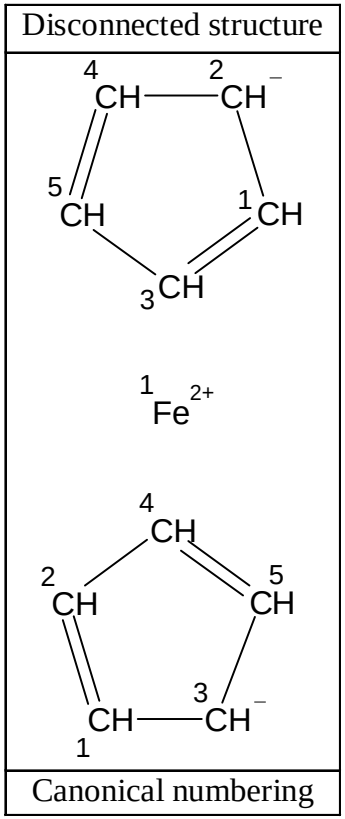
MSG (fixed)

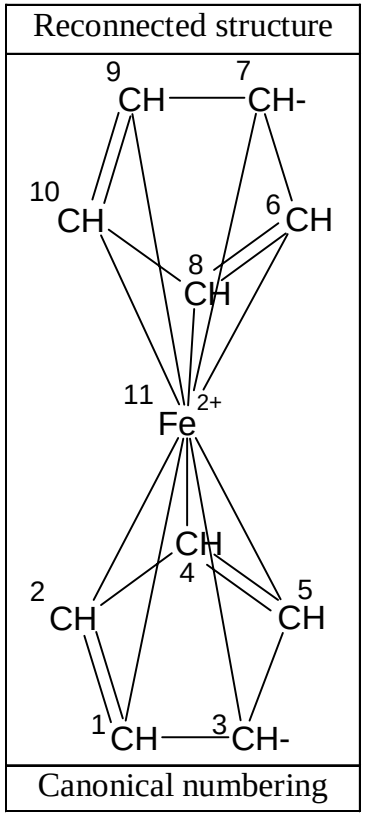


Canonical numbering

Description	Layers
formula	C5H8NO4.Na
connectivity	6-3(5(9)10)1-2-4(7)8;
H-atoms	1-2H2,3H,6H2(H-,7,8,9,10);
stereo sp^3	3-;
H-atoms fixed	7H;
stereo sp^3	3-;
charges	-1;+1

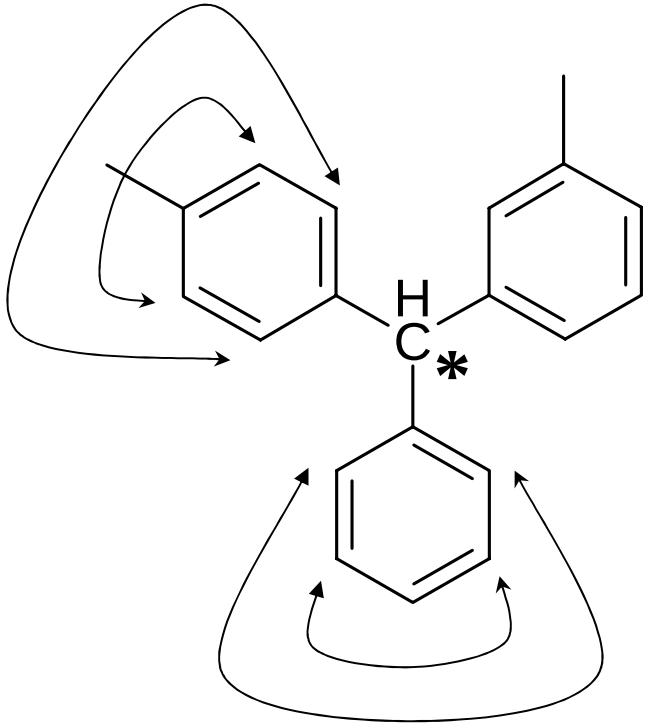
Ferrocene





Description	Layers	Layers
formula	`2C5H5.Fe`	`C10H10Fe`
connectivity	`2*1-2-4-5-3-1;`	`1-2-4-5-3(1)11(1,2,4,5)6-7(11)9(11)10(11)8(6)11`
H-atoms	`2*1-5H;`	`1-10H`
charges	`2*-1;+2`	

Byproducts:
Stereogenic Centers and Equivalent Atoms

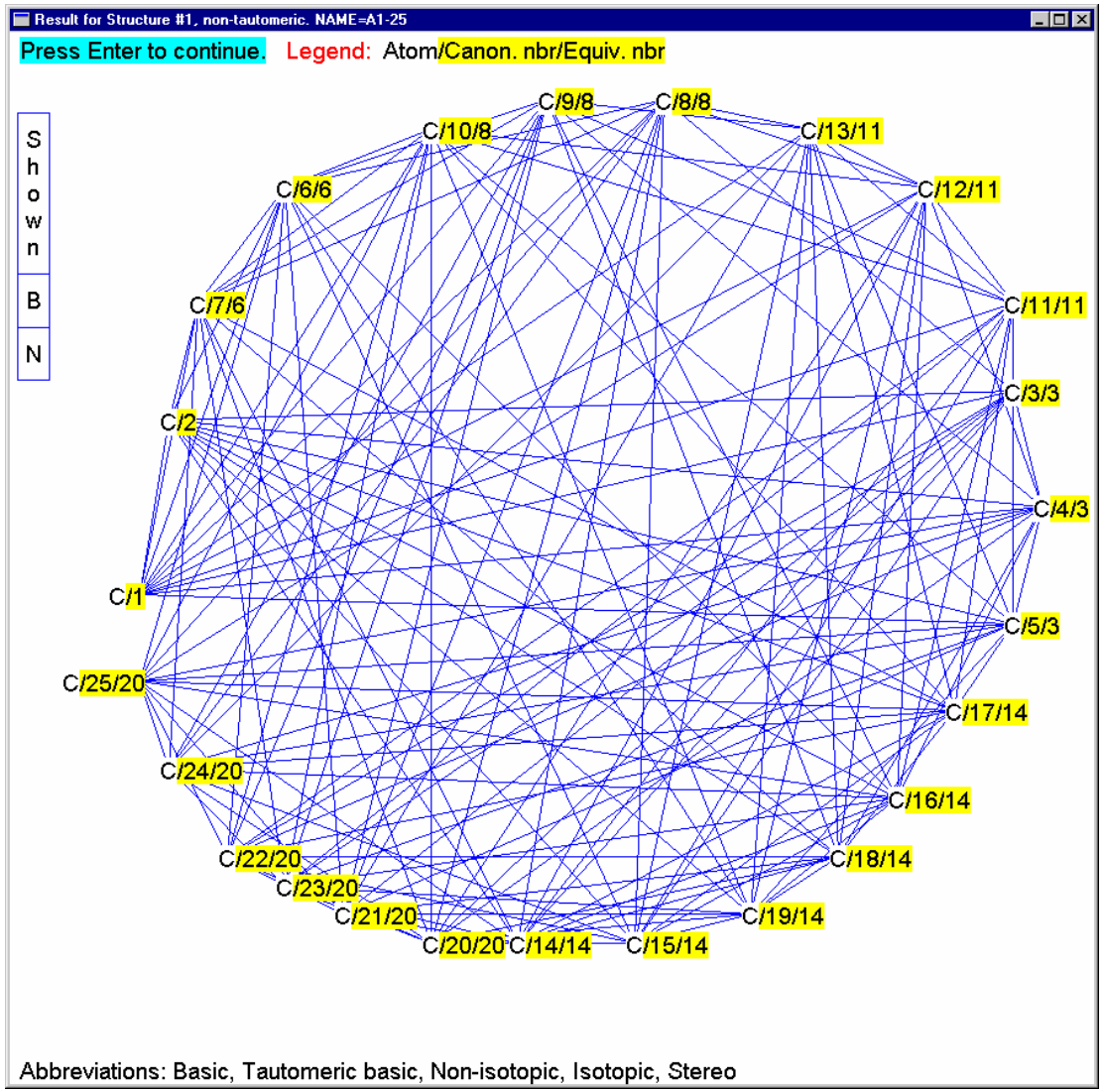


- Aids structure validation

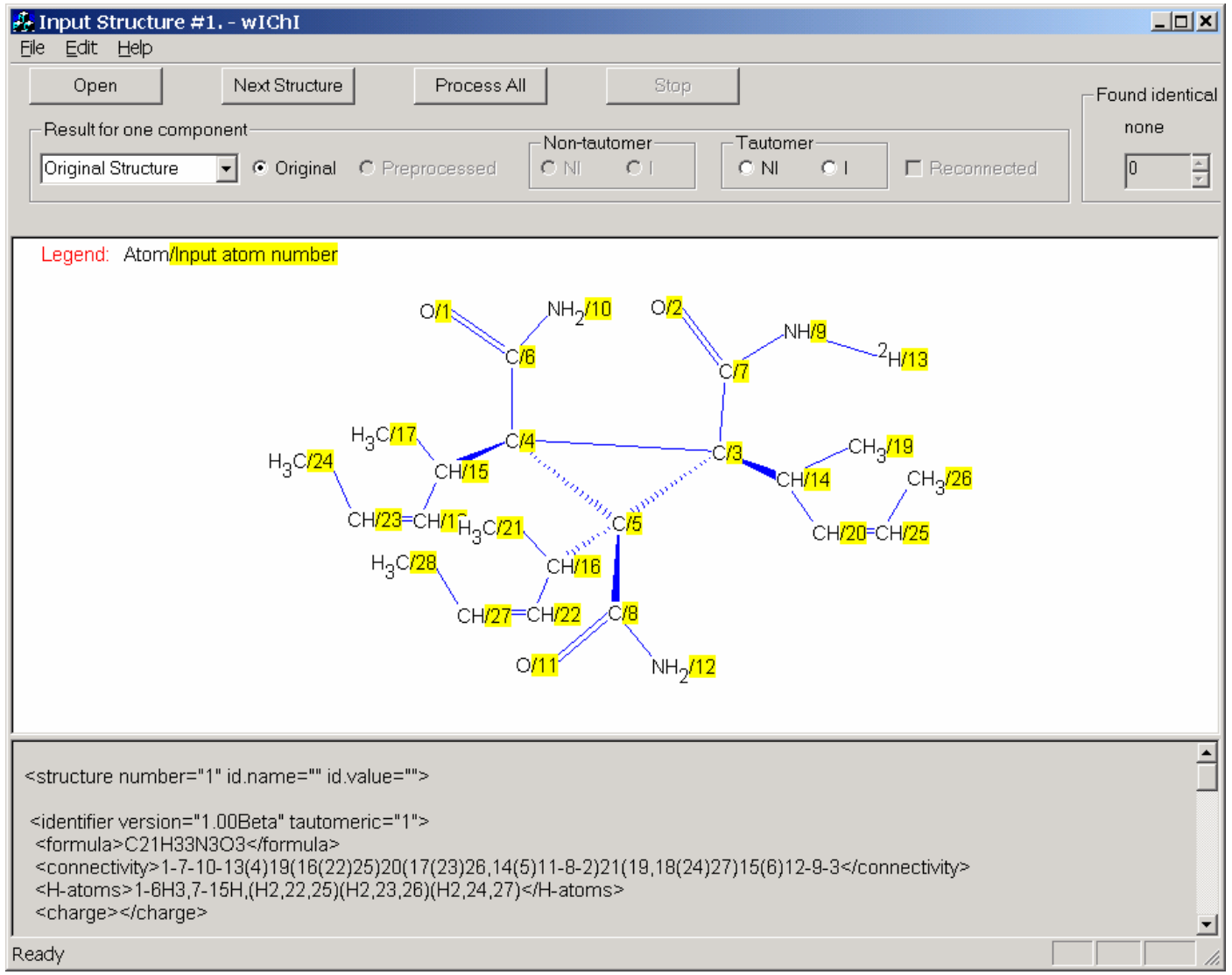
Auxiliary Output

- Warnings/Errors
  - Unusual valences
  - Unrecognized input
- ‘Reversibility’
  - Coordinates
  - Bond/Charge Location

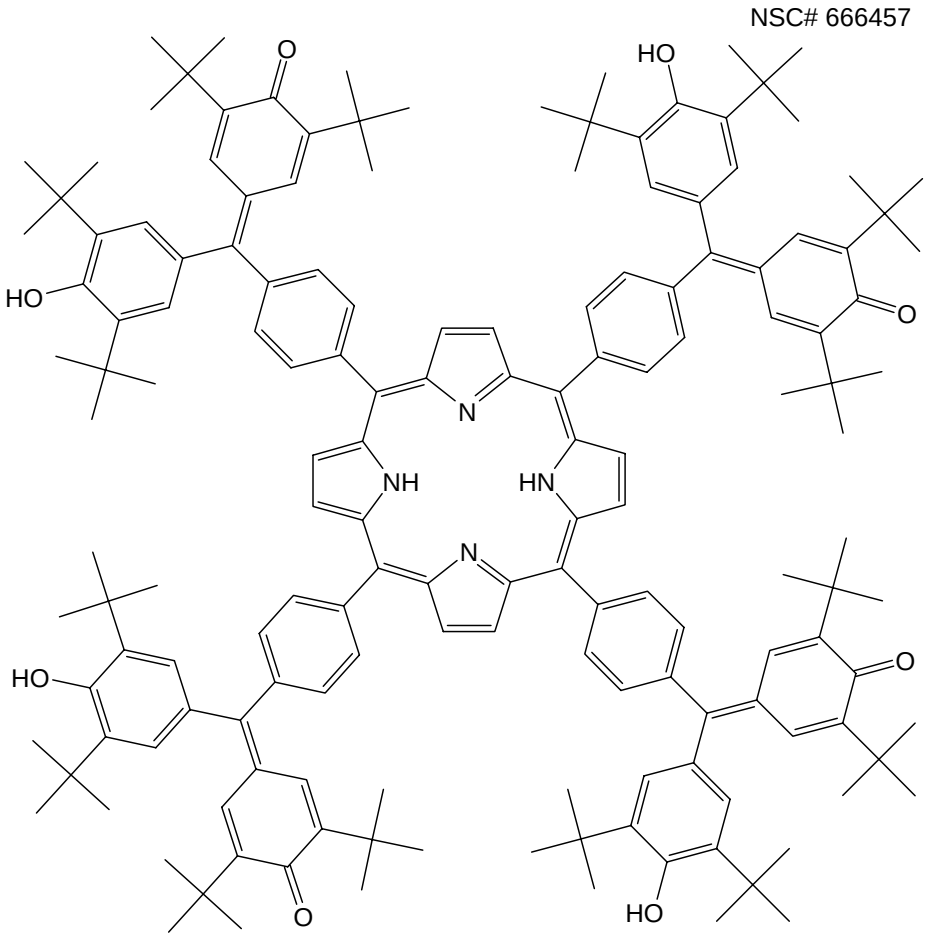
Testing - OK


Result for Structure #1, non-tautomeric. NAME=A1-25
Press Enter to continue.
Legend: Atom/Canon. nbr/Equiv. nbr
Shown
B
N
C/9/8
C/8/8
C/10/8
C/13/11
C/6/6
C/12/11
C/7/6
C/11/11
C/3/3
C/2
C/4/3
C/1
C/5/3
C/25/20
C/17/14
C/24/20
C/16/14
C/22/20
C/18/14
C/23/20
C/21/20
C/19/14
C/20/20
C/14/14
C/15/14
Abbreviations: Basic, Tautomeric basic, Non-isotopic, Isotopic, Stereo

Beta Testing



Performance:
Most Challenging NCI-NIH Structure



50 ms – 2 GHz PC

IChI FAQs

- How can you represent chemistry without electrons?
  - Chemistry is not represented, just identity
  - Whole molecule properties are easily added.
- Do big molecules have big IChIs?
  - Yes, just like systematic names
- How to handle other tautomers, substructures,..?
  - Other software
- Is IChI reversible?
  - Partly - contains only data needed for ‘naming’
  - Auxiliary fields can carry other information
- Is IChI extensible?
  - New layers can add refinement

IChI Capabilities

- Identify compounds at the known level of detail
- Convention-free (mostly)
- Generate quickly from structure
- Contains all essential connectivity information
- Simple ASCII representation

I U P A C

Current Project

Committee on Printed and Electronic Publications

Started

Oct. 2002

Number: 2002-022-1-024

Title: Standard XML data dictionaries for chemistry

Task Group
Chairman: Steve Stein

Members: Kirill Degtyarenko, Jeremy Frey, Francois Gilardoni, Jiri Jirat, Robert Lancashire, Alan McNaught, Peter Murray-Rust, Miloslav Nic, and Henry Rzepa

Completion Date: 2005

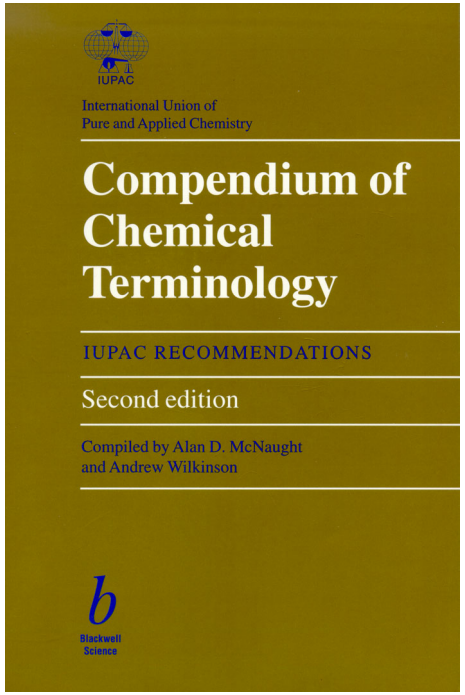

IUPAC
ifcc
International Union of Pure and Applied Chemistry
Clinical Chemistry Division
Commission on Quantities and Units in Clinical Chemistry
and International Federation of Clinical Chemistry
Scientific Division
Committee on Quantities and Units
Compendium of Terminology and Nomenclature of Properties in Clinical Laboratory Sciences
(Recommendations 1995)
J.C. RIGG, S.
R.DYBKÆR A
b
Blackwell Science
International Union of Pure and Applied Chemistry
COMPENDIUM OF ANALYTICAL NOMENCLATURE
DEFINITIVE RULES 1997
Prepared for publication by
János Inczédy, Tamás Lengyel, Allan M. Ure
IUPAC
International Union of
Pure and Applied Chemistry
Compendium of Chemical Terminology
IUPAC RECOMMENDATIONS
Second edition
Compiled by Alan D. McNaught
and Andrew Wilkinson
b
Blackwell Science
INTERNATIONAL UNION OF PURE AND APPLIED CHEMISTRY
PHYSICAL CHEMISTRY DIVISION
Quantities, Units and Symbols in Physical Chemistry
Prepared for publication by
IAN MILLS TOMISLAV CVITAŠ
KLAUS HOMANN NIKOLA KALLAY
KOZO KUCHITSU
SECOND EDITION
BLACKWELL SCIENCE

Utility of Digital ‘Dictionary’

- Traceability
  - Clarity (especially for computers)
- Indexing
  - Effective ‘keywording’
- Accuracy
  - Error checking
- Automated Processing

Goal Color Books as Source of Basic Chemical Terms in XML

- Why IUPAC?
  - International Acceptance
  - Comprehensive
  - Open Process
  - Long-standing
    - Part of its mission



nuclear fusion reaction
A reaction between two light nuclei resulting in the production of a nuclear species heavier than either initial nucleus.
1982, *54*, 1543

nuclear graphite
A *polygranular graphite* material for use in nuclear reactor cores consisting of *graphitic carbon* of very high chemical purity. High purity is needed to avoid absorption of low-energy neutrons and the production of undesirable radioactive species.
Notes:
Apart from the absence of neutron-absorbing impurities, modern reactor graphites are also characterized by a high degree of *graphitization* and no preferred bulk orientation. Such properties increase the dimensional stability of the nuclear graphite at high temperatures and in a high flux of neutrons. The term nuclear graphite is often, but incorrectly, used for any *graphite material* in a nuclear reactor, even if it serves only for structural purposes.
1995, *67*, 498

nuclear isomers
Nuclides having the same *mass number* and *atomic number,* but occupying different nuclear energy states.
1982, *54*, 1545

nuclear level
One of the energy values at which a *nucleus* can exist for an appreciable time ($> 10^{-22}$ s).
1982, *54*, 1547

nuclear magneton
Electromagnetic fundamental physical constant $\mu_N = (m_e/m_p)\mu_B = 5.050\ 7866\ (17) \times 10^{-27}$ J T^{-1}, where m_e is the electron rest mass, m_p the proton rest mass and μ_B the Bohr magneton.
CODATA Bull., 1986, *63*, 1

nuclear particle

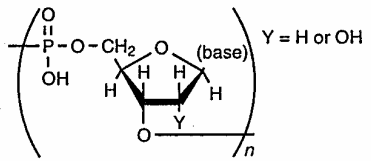
nuclear transition
For a *nucleus* a change from one quantized energy state into another or a *nuclear transformation.*
1982, *54*, 1553

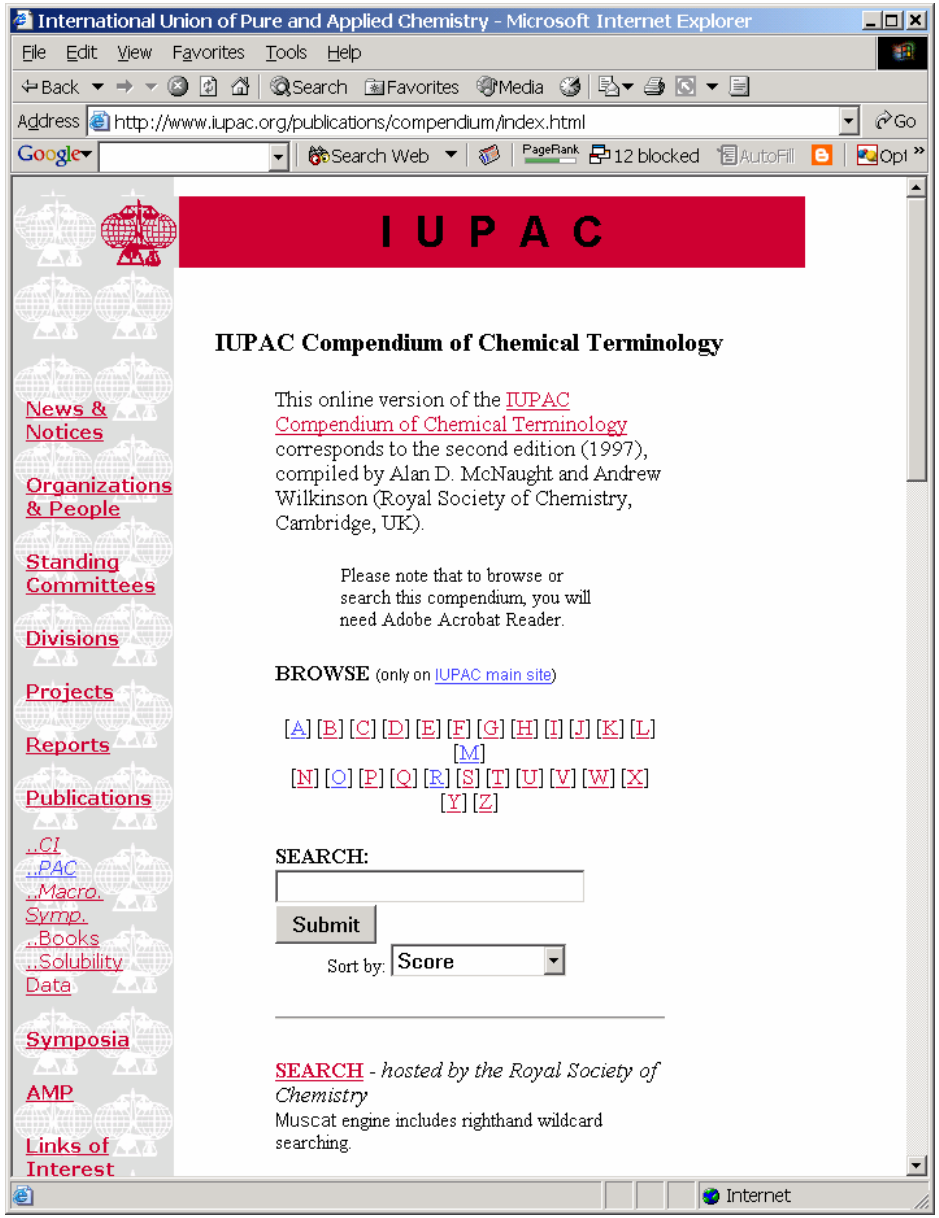
nucleating agent
A material either added to or present in a system, which induces either homogeneous or heterogeneous *nucleation.*
1972, *31*, 608

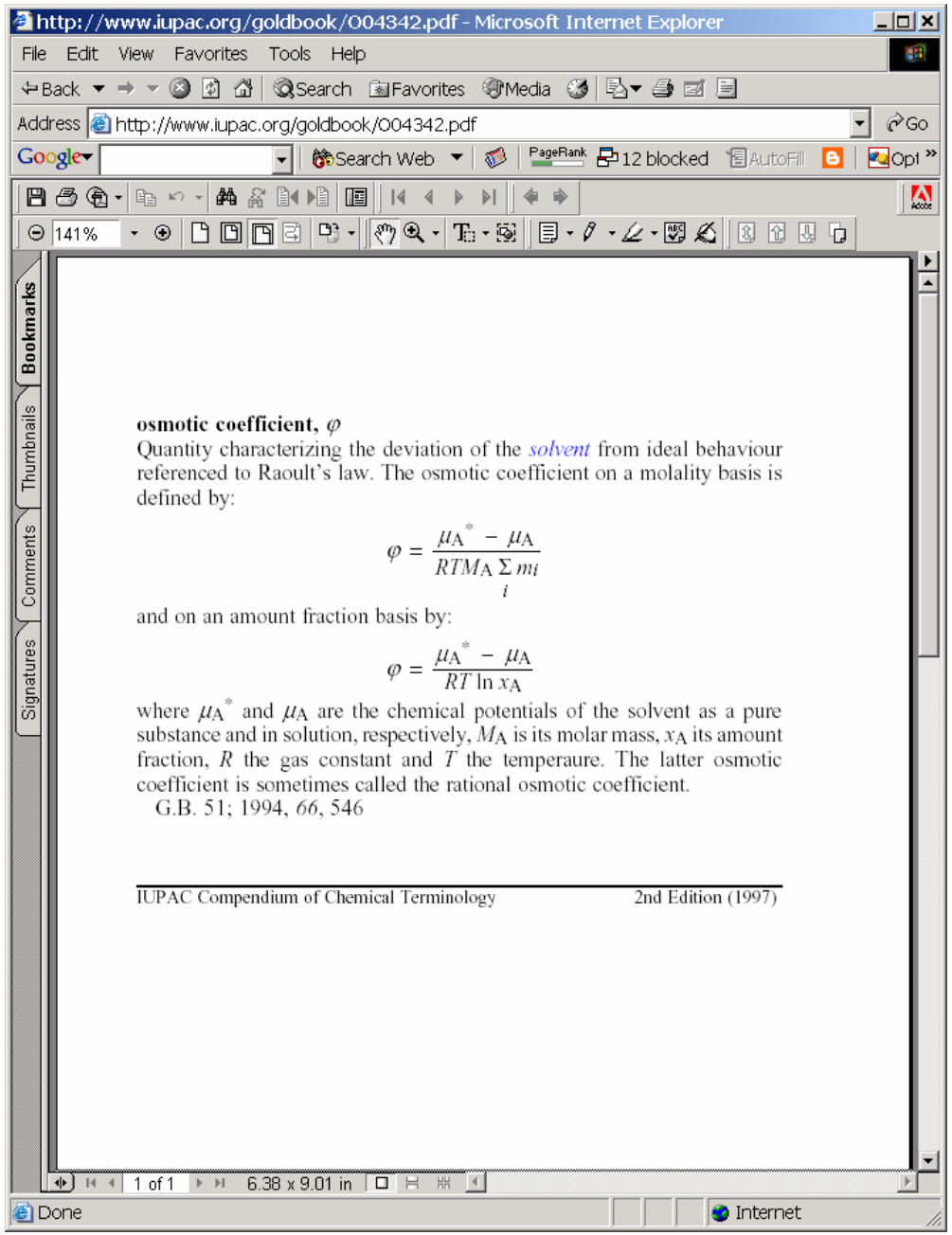
nucleation (in colloid chemistry)
The process by which nuclei are formed in solution. The condensation of a single chemical compound is called homogeneous nucleation. The simultaneous condensation of more than one compound is called simultaneous nucleation. The condensation of a compound on a foreign substance is called heterogeneous nucleation.
O.B. 84; see also 1972, *31*, 608

nucleation and growth
A process in a *phase transition* in which nuclei of a new phase are first formed, followed by the propagation of the new phase at a faster rate.
See *continuous precipitation, discontinuous precipitation.*
1994, *66*, 587

nucleic acids
Macromolecules, the major organic matter of the nuclei of biological cells, made up of *nucleotide* units, and hydrolysable into certain *pyrimidine* or *purine bases* (usually adenine, cytosine, guanine, thymine, uracil), D-ribose or 2-deoxy-D-ribose and phosphoric acid.




International Union of Pure and Applied Chemistry - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites Media
Address http://www.iupac.org/publications/compendium/index.html
Go
Google Search Web 12 blocked AutoFill Opt
I U P A C
News & Notices
Organizations & People
Standing Committees
Divisions
Projects
Reports
Publications
..CI
..PAC
..Macro. Symp.
..Books
..Solubility Data
Symposia
AMP
Links of Interest
IUPAC Compendium of Chemical Terminology
This online version of the IUPAC Compendium of Chemical Terminology corresponds to the second edition (1997), compiled by Alan D. McNaught and Andrew Wilkinson (Royal Society of Chemistry, Cambridge, UK).
Please note that to browse or search this compendium, you will need Adobe Acrobat Reader.
BROWSE (only on IUPAC main site)
[A] [B] [C] [D] [E] [F] [G] [H] [I] [J] [K] [L] [M]
[N] [O] [P] [Q] [R] [S] [T] [U] [V] [W] [X] [Y] [Z]
SEARCH:
Submit
Sort by: Score
SEARCH - hosted by the Royal Society of Chemistry
Muscat engine includes righthand wildcard searching.
Internet



osmotic coefficient, φ

Quantity characterizing the deviation of the *solvent* from ideal behaviour referenced to Raoult's law. The osmotic coefficient on a molality basis is defined by:

$$\varphi = \frac{\mu_A^* - \mu_A}{RTM_A \sum_i m_i}$$

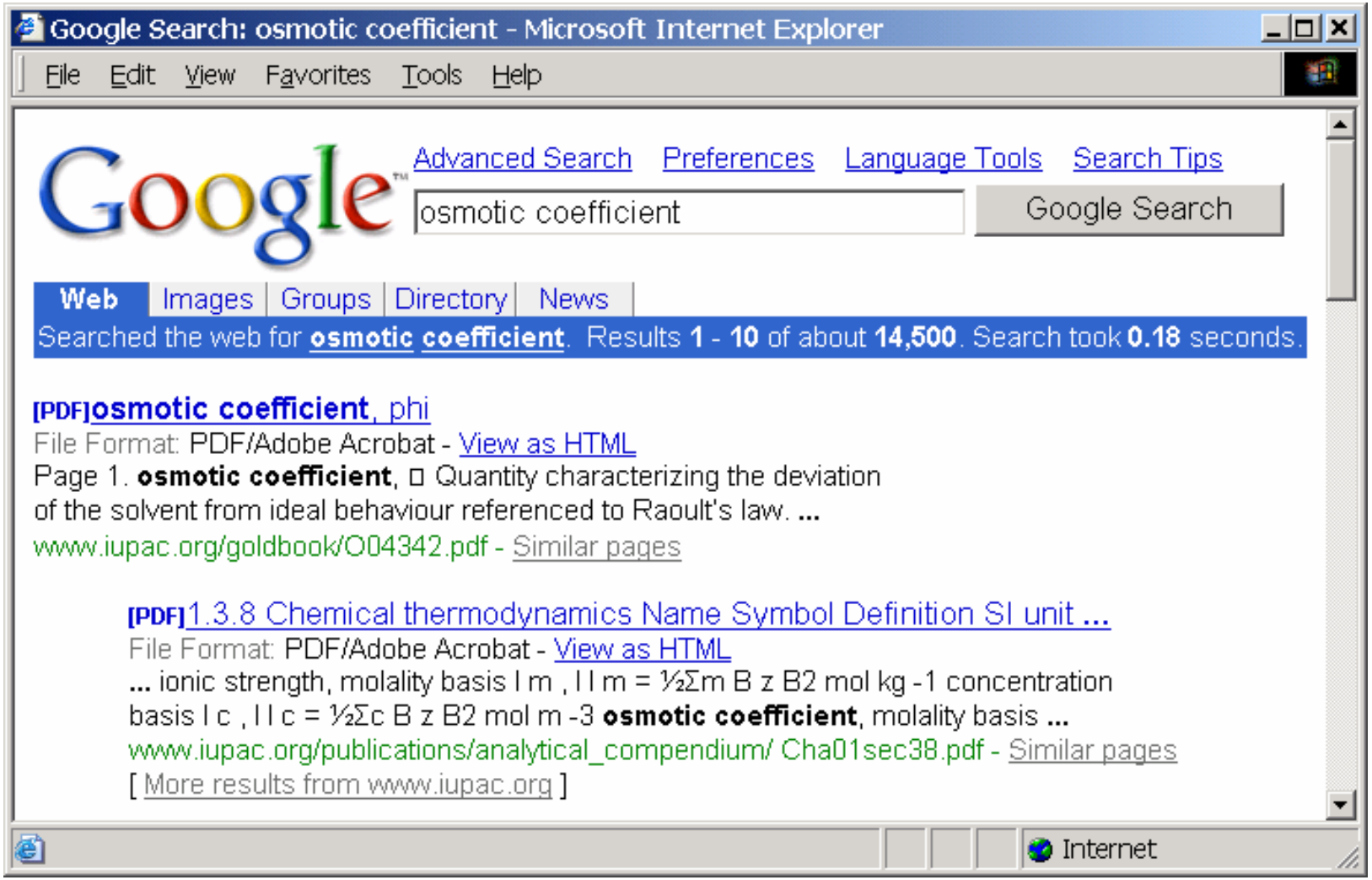
and on an amount fraction basis by:

$$\varphi = \frac{\mu_A^* - \mu_A}{RT \ln x_A}$$

where μ_A^* and μ_A are the chemical potentials of the solvent as a pure substance and in solution, respectively, M_A is its molar mass, x_A its amount fraction, R the gas constant and T the temperaure. The latter osmotic coefficient is sometimes called the rational osmotic coefficient.

G.B. 51; 1994, *66*, 546

The Gold Book is ‘Indexed’ on the Web



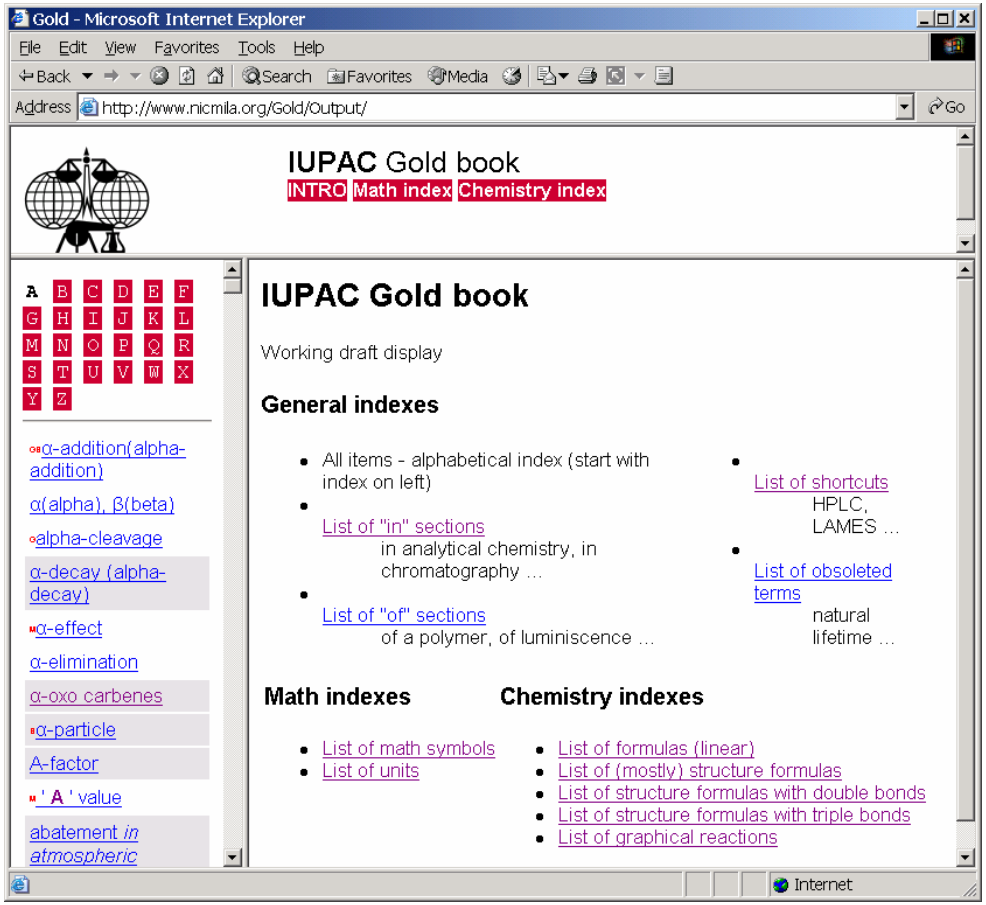
Gold Book in XML

- Provide uniform chemical terminology for XML documents
- Root for digital ‘tags’ in chemistry
- Model for future IUPAC recommendations

Gold Book – PDF to XML
(implicit to explicit)

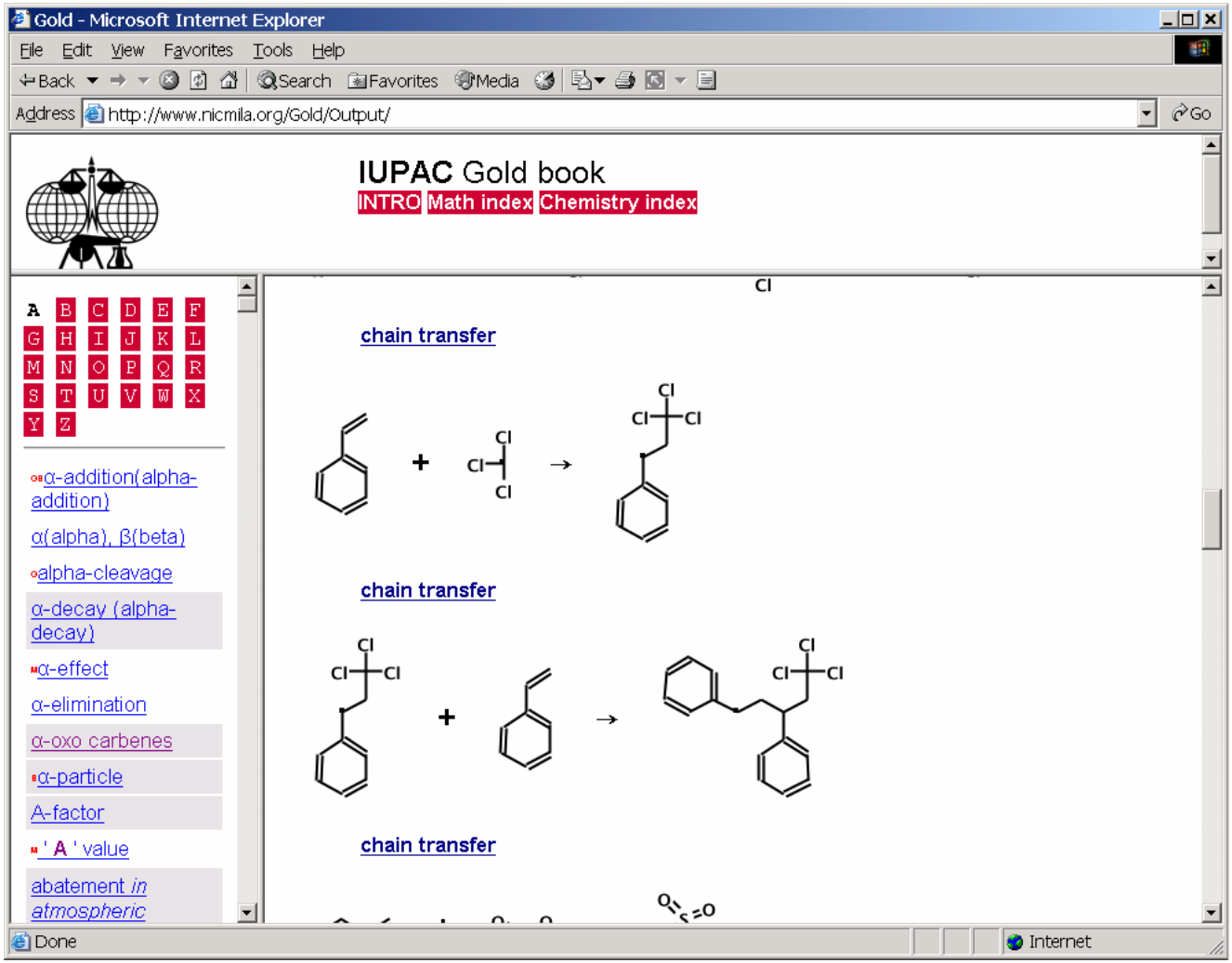
- Text
  - ‘Tag’ data and relations
- Chemical Structures
  - To connection tables/CML/SVG
- Equations
  - To MathML
- Figures & Complex Schemes
  - Redraw in SVG

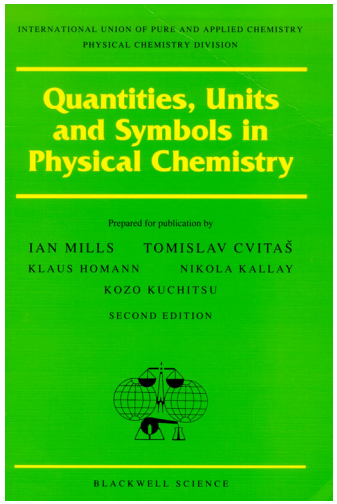
http://www.nicmila.org/Gold/Output/



Miloslav Nic, Jiri Jirat, Czech Republic

Some structures were convertible





Green Book - Promise

Template’ for numeric property validation

- Ensure proper units and representation
- Traceable to IUPAC definition
- Basic Tags for Common Properties
  - Covers 15 ‘fields’ of chemistry

Next

- Nov 12-14 Meeting at NIST
- IChI
  - ‘Final’ Beta Nov. 2003
  - Dissemination
    - Databases, Software
  - Version 2
- XML Data Dictionary
  - Gold Book Conversion
  - Maintenance Method
  - Green Book

Naming follows Recognition

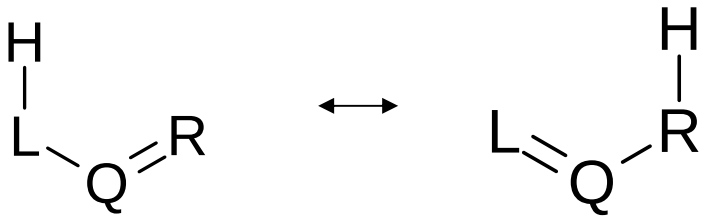
“He called the light Day, and the darkness He called night” (Genesis 1.5)

Adam McLean's Gallery of alchemical images
www.levity.com/alchemy

Green Book - Promise

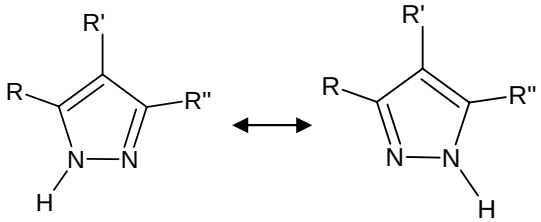
- Periodic Table and Relative Molar Masses
  - Originating digital source
  - Integrate with relevant IUPAC recommendations
- Provide root of chemical information ‘tree’
  - Spectroscopy, electrochemistry, thermochemistry, catalysis, …

Tautomer Rules



L,R = N, O, S, Se, Te

Q = C, N, S, P, …



$-OH$ $-O^-$ Salts