

Robert E. Belford^{1*}, Ehren C. Bucholtz², Andrew P. Cornell¹, Jordi Cuadros³, Vincent Scalfani⁴ & Martin A. Walker⁵

¹University of Arkansas at Little Rock, ²St.Louis College of Pharmacy, ³IQS Univ. Ramon Llull, ⁴University of Alabama, ⁵SUNY-Potsdam **rebelford@ualr.edu (author for correspondence)

Chemical Identifiers and 21st Century Nomenclature

- Chemical structures can be described as machine readable alphanumeric strings
- Labels provide convenient means of comparing and distinguishing chemicals
- Enables communication across databases, software agents and human beings
- Molecules are assigned a unique text label identifier:
 - Must be unambiguous and always refer to same substance
 - Registry-lookup Identifiers (e.g. CASRN and PubChem CID)
 - Act as pointers for databases which contain the structural information
- Structure Based Identifiers (e.g. nomenclature, SMILES and InChI)
 - Software algorithms can convert the identifier to structure

Structure Based Identifiers

- Structure identification is based on graph theory, where atoms are nodes, and bonds are edges in connection tables

Chemistry and Graph Theory

Henry Heppel (1818-1891)



The Connection Table

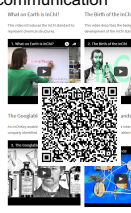
(Adjacency Matrix)



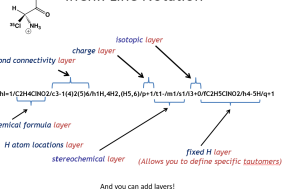
- Software converts connection table file to line notations
 - Closed Standard Identifiers (e.g. SMILES)
 - Use proprietary algorithms to create a line notation
 - May not be canonical and creates problems in communication between databases
 - Open Standard Identifiers (e.g. InChI)
 - Non-proprietary

The InChI Standard

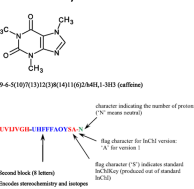
- Structure based approach- anyone can produce
- InChI from structural formula
- Strict uniqueness and canonical
- Non-proprietary, open source, and free access
- Open access to source code
- Hierarchical approach with levels of granularity



InChI: Line Notation



InChI Key



InChI OER

<https://www.inchi-trust.org/oer/>



HOME ABOUT THE INCHI TRUST ABOUT THE INCHI STANDARD DOWNLOADS NEWS RESOURCES CONTACT

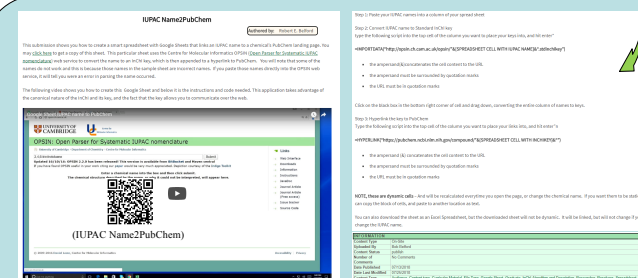
- An **Open Education Resource** devoted to the use of InChI in the chemical sciences
- Repository of two types of content
 - On-site open access materials to download and reuse as you see fit
 - Off-site links to publications related to InChI (may or may not be open access)

| POST DATE | TITLE/LINK | CONTENT TYPE |
|------------|---|--------------|
| 07/26/2018 | InChI - the worldwide chemical structure identifier standard | Off-Site |
| 07/26/2018 | What on the Earth is InChI? - IUPAC 200 Status | Off-Site |
| 07/26/2018 | The Status of the IUPAC Chemical Structure Standard - Today and the Future | On-Site |
| 07/26/2018 | The Status of the InChI Project - 8/25/2011 | On-Site |
| 07/26/2018 | The Status of the International Chemical Identifier standard-InChI 6/3/2014 | On-Site |
| 07/25/2018 | Matlab InChIkey Scripts | Off-Site |
| 07/25/2018 | Identifier conversion on an Excel spreadsheet | On-Site |
| 07/23/2018 | Many InChI and quite some feat | Off-Site |
| 07/16/2018 | Brevi introducción a la digitalización de la información química | Off-Site |
| 07/13/2018 | IUPAC Name2PubChem | On-Site |

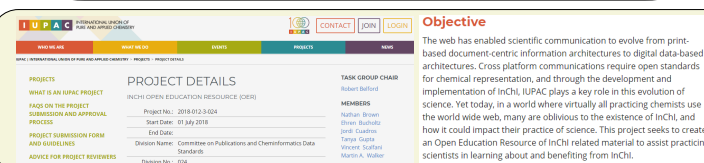
| POST DATE | TITLE/LINK | CONTENT TYPE |
|------------|------------------------------------|--------------|
| 07/26/2018 | IUPAC Name2PubChem | On-Site |

Filters by multiple tags, which reduce display to only content with chosen tags, and the other tags of that content

- Off Site Content provide link with brief overview of the content
- On Site Content Provides ability to download file and gain instructional information
- All content has information box
 - Author/copyright/DOI and other information

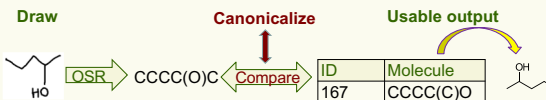


InChI OER Material for IUPAC Name-to-OPSIN-to-PubChem

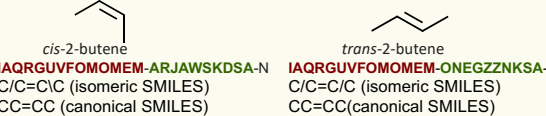


InChIKey & OSR

- Optical Structure Recognition (OSR) converts graphical representations of molecules to line notation
- Goal: Use OSR to grade student drawn structures on quizzes



- Using SMILES for development work as it is human readable
- SMILES presents database issues for alkenes
- Switch to InChIKey, a 27 character hashed version of full InChI
 - Allows for grading on both connectivity and stereochemistry
 - InChIKey is canonical by design with very low probability of two molecules having same InChIKey

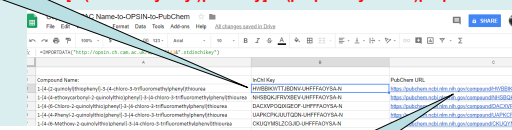


| Functional Group (line notation) | Validation set size | Percent correct computer generated | Percent correct expert drawn |
|----------------------------------|---------------------|------------------------------------|------------------------------|
| alkenes (SMILES) | 80 | 80 | 65 |
| alkenes (InChIKey) | 80 | 97 | 80 |

IUPAC Name-to-OPSIN-to-PubChem

What is:

1-[4-(2-methoxyethyl)phenoxy]-3-(propan-2-ylamino)propan-2-ol



Step 2: Spreadsheet concatenates InChIKey to PubChem URL

The compound is Metoprolol:
PubChem has a wealth of information on it

Please use QR Codes to link to web pages and watch YouTube videos on your cell phone

This work has been supported by IUPAC (project 2018-012-3-024) and an InChI Trust Fellowship

"... we cannot improve the language of any science without at the same time improving the science itself; neither can we, on the other hand, improve a science, without improving the language or nomenclature which belongs to it"

Lavoisier's Preface to *Traité Élémentaire de Chimie*
translation by Robert Kerr (Edinburgh, 1790)