Use and Utility of InChI in PubChem

Evan E. Bolton, Ph.D.
NCBI/NLM/NIH
ACS National Meeting, SLC 2009
bolton@ncbi.nlm.nih.gov
What is PubChem?

• An archive
• Links chemistry and biology
• Search, retrieval, and data analysis tools

**PubChem Compound:** Search unique chemical structures using names, synonyms or keywords. Links to available biological property information are provided for each compound.

**PubChem Substance:** Search deposited chemical substance records using names, synonyms or keywords. Links to biological property information and depositor web sites are provided.

**PubChem BioAssay:** Search bioassay records using terms from the bioassay description, for example "cancer cell line". Links to active compounds and bioassay results are provided.

**Structure Search:** Search PubChem's Compound database using a chemical structure as the query. Structures may be sketched or specified by SMILES, InChI, MOL files, or other formats.

Primary Goal?

To be an on-line resource providing comprehensive information on the biological activities of small molecules.
Key Goals?

Optimize research utility by including other public sources of chemical structure and bioactivity information
Key Goals?

Integrate with other NIH biomedical information resources whenever possible
Where does PubChem get its data?
Our depositors!
Growth in PubChem Depositors

![Graph showing growth in PubChem Depositors over time, with two lines representing Substance and BioAssay categories.](image-url)
Growth in PubChem Substances
Growth in PubChem BioAssays

![Graph showing growth in PubChem BioAssays from June 2005 to March 2009. The graph compares the total count and MLP count over time. The total count shows a steady increase, while the MLP count shows a more erratic but also increasing trend.](image-url)
Growth in Substance Outcomes

![Graph showing growth in substance outcomes over time. The graph compares 'Total' and 'MLP' with deposit dates from June 2005 to March 2009. The y-axis represents Total Count (in millions), while the x-axis shows Deposit Date. The 'Total' line is generally lower than the 'MLP' line, indicating a disparity in growth. The 'MLP' line shows a more significant increase in count towards the end of the graph.]
Growth in PubChem Unique Users
How does PubChem use InChI?
Allowed Deposition Structure Format

**PubChem Substance Deposition using SD File Format**

v1.4.1  

**PubChem Deposition Allowed SD Fields (cont.)**

<table>
<thead>
<tr>
<th>PUBCHEM_EXT_DATASOURCE_INCHI</th>
</tr>
</thead>
<tbody>
<tr>
<td>InChI string to be used to represent the chemical structure for the Substance being deposited. This is an alternate mechanism for providing a structural description of a Substance and will be ignored if a chemical structure with atoms is also provided in the SD file format CTAB section in the same SDF record deposited.</td>
</tr>
</tbody>
</table>

Only a single InChI string is allowed for a given Substance. The expected format is a single line of text containing a valid InChI string.

---

**Minimum SDF Formatted Deposition using InChI as Chemical Structure**

```
-OEChem-11050415022D
0 0 0 0 0 0 0 0 0 0 0 9999 V2000
M END
> <PUBCHEM_EXT_DATASOURCE_REGID>
YourUniqueSubstanceID
```

```
> <PUBCHEM_EXT_DATASOURCE_INCHI>
InChI=1S/C9H8O4/c1-6(10)13-8-5-3-2-4-7(8)9(11)12/h2-5H,1H3,(H,11,12)
```

```
Available for Chemical Structures
Provided in Exported Data

> <PUBCHEM_IUPAC_TRADITIONAL_NAME>
2-acetoxybenzoic acid

> <PUBCHEM_NIST_INCHI>
InChI=1S/C9H8O4/c1-6(10)13-8-5-3-2-4-7(8)9(11)12/h2-5H,1H3,(H,11,12)

> <PUBCHEM_NIST_INCHIKEY>
BSYNRYMUTBXSQ-UHFFFAOYSA-N

> <PUBCHEM_XLOGP3>
1.2

> <PUBCHEM_EXACT_MASS>
180.042259

> <PUBCHEM_MOLECULAR_FORMULA>
C9H8O4

> <PUBCHEM_MOLECULAR_WEIGHT>
180.15742

> <PUBCHEM_OPENEYE_CAN_SMILES>
CC(=O)OC1=CC=CC=C1C(=O)O

> <PUBCHEM_OPENEYE_ISO_SMILES>
CC(=O)OC1=CC=CC=C1C(=O)O

> <PUBCHEM_CACTVS_TPSA>
Structure Download Format

PubChem Download Service

Download the records in the format selected below:

Choose a format

Choose a compression type

Retrieve 3D records/images?

Save this job in XML format (e.g. for PUG)
Structure Sketcher Input Format

InChI=1/C9H8O4/c1-6(10)13-8-5-3-2-4-7(8)9(11)12/fH2-5H.1H3,(H,11,12)/f/h11H
Standardization Service Input/Output

PubChem Standardization Service

Submit this job to PubChem's standardization service

**Input**

- Start with SMILES
- Start with InChI
- Start with an SDF file

**Output**

- Get back SMILES
- Get back InChI
- Get back SDF
InChI Compound-based Lookup

CID: 2244

aspirin; Acetylpyrin; Colfarit...
IUPAC: 2-acetyloxybenzoic acid
MW: 180.157420 g/mol | MF: C₉H₈O₄
Tested in BioAssays: All: 242, Active: 2; BioActivity Analysis
Cyclooxygenase Inhibitors... more
InChI Substance-based Lookup

ACETYSALICYLIC ACID; MLS001336046; SMR000059138
Compound ID: 2244
Source: MLSMR (MLS001336046)
IUPAC: 2-acetyloxybenzoic acid
MW: 180.157420 g/mol | MF: C₉H₈O₄
Cyclooxygenase Inhibitors... more

ACETYSALICYLIC ACID; MLS001336045; SMR000059138
Compound ID: 2244
Source: MLSMR (MLS001336045)
InChIKey Compound-based Lookup

CID: 2244

aspirin; Acylopyrin; Cofarit ...
IUPAC: 2-acetyloxybenzoic acid
MW: 180.157420 g/mol | MF: C₉H₈O₄
Tested in BioAssays: All: 242, Active: 2; BioActivity Analysis
Cyclooxygenase Inhibitors... more
### InChIKey Substance-based Lookup

#### ACETYSALICYLIC ACID; MLS001336046; SMR000059138
**Compound ID:** 2244  
**Source:** MLSMR (MLS001336046)  
**IUPAC:** 2-acetyloxybenzoic acid  
**MW:** 180.157420 g/mol | **MF:** C₉H₈O₄  
**Cyclooxygenase Inhibitors**... [more]

#### ACETYSALICYLIC ACID; MLS001336045; SMR000059138
**Compound ID:** 2244  
**Source:** MLSMR (MLS001336045)  
**IUPAC:** 2-acetyloxybenzoic acid  
**MW:** 180.157420 g/mol | **MF:** C₉H₈O₄
InChIKey... quotes not necessary
InChI/InChIKey Indexes

PubChem Compound Preview/Index - Internet Explorer provided by Dell

For: "InChI=1S/C9H8O4/c1-6(10)13-8-5-3-2-4-7(8)9(11)12/h2-5H,1H3,(H,11,12)"

Most Recent Queries

<table>
<thead>
<tr>
<th>InChIKey</th>
<th>Time</th>
<th>Result</th>
</tr>
</thead>
<tbody>
<tr>
<td>C9H8O4/c1-6(10)13-8-5-3-2-4-7(8)9(11)12/h2-5H,1H3,(H,11,12)&quot;</td>
<td>16:28:14</td>
<td>1</td>
</tr>
</tbody>
</table>

Use the pull-down menu to specify a search field.

Click AND OR NOT to add a term to the query box.
Search InChI/InChIKey Index Only

ACETYSALICYLIC ACID; MLS001336046;
SMR000059138
Compound ID: 2244
Source: MLMSR (MLS001336046)
IUPAC: 2-acetyloxybenzoic acid
MW: 180.157420 g/mol | MF: C9H7O4
Cyclooxygenase Inhibitors... more

ACETYSALICYLIC ACID; MLS001336045;
SMR000059138
Compound ID: 2244
How does PubChem use InChI?

• Deposition structure format
• Computed property of compounds
• Text-based structure lookup
• Structure export format
• Structure search input
• Standardization service input/output
InChI/InChIKey Use and Utility

- **InChI**
  - Enabler of data exchange
  - Provides structure canonicalization

- **InChIKey**
  - Compact form for structure lookup
  - Allows “secret” chemical information exchange
InChI Caveats
May not use same InChI/InChIKey

• PubChem InChI provides stereo parity without coordinates (like SMILES)
• PubChem SDF record with 2D coordinates may produce different InChI

InChI=1S/C28H30N2O3.ClH/
c1-6-29-23-15-25-21(13-17(23)4)27(19-11-9-10-12-20(19)28(31)32-8-3)22-14-18(5)24(30-7-2)16-26(22)33-25;/h9-16,29H,6-8H2,1-5H3;1H/b30-24+;

InChI=1S/C28H30N2O3.ClH/
c1-6-29-23-15-25-21(13-17(23)4)27(19-11-9-10-12-20(19)28(31)32-8-3)22-14-18(5)24(30-7-2)16-26(22)33-25;/h9-16,29H,6-8H2,1-5H3;1H
May get more than one record

"InChI=1S/C28H30N2O3.ClH/
c1-6-29-23-15-25-21(13-17(23)4)27(19-11-9-10-12-20(19)28(31)32-8-3)22-14-18(5)
24(30-7-2)16-26(22)33-25;/h9-16,29H,6-8H2,1-5H3;1H"[InChI]
State of InChI

• Standard InChI vs. Non-Standard InChI
• 0D/2D Perception differences
• InChI to Structure to InChI
• Structure to InChI to Structure
• How to solve?
InChI Future

• Can it be everything to everybody?
• What are the tradeoffs?

• What more can it be?

• External organization participation
PubChem Crew …

Steve Bryant
Yulia Borodina
Jie Chen
Svetlana Dracheva
Lewis Geer
Lianyi Han
Jane He
Siqian He
Karen Karapetian
Sunghwan Kim
Wenyao Shi
Ben Shoemaker

Vahan Simonyan
Tugba Suzek
Paul Thiessen
Valery Tkachenko
Jiyao Wang
Yanli Wang
Jewen Xiao
Jian Zhang
Acknowledgments

• Steve Heller
• InChI Developers
• NIST
• IUPAC

• You!