FDA U.S. FOOD & DRUG ADMINISTRATION

Usage of InChI in SPL Substance Indexing Files

Yulia Borodina

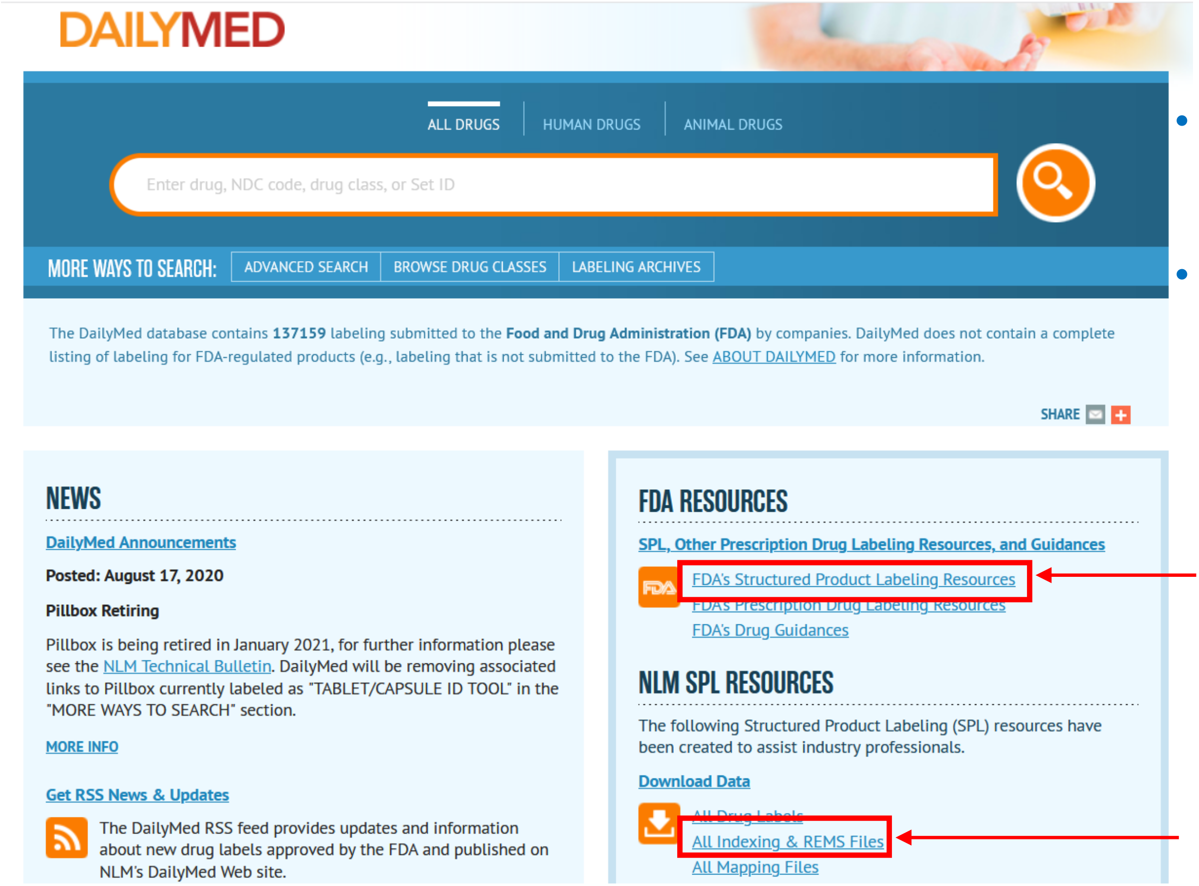
Office of Health Informatics

NIH InChI Workshop, March 22-24, 2021

Disclaimer

The views and opinions presented here represent those of the speaker and should not be considered to represent advice or guidance on behalf of the Food and Drug Administration.

SPL Substance Indexing Files



DAILYMED

ALL DRUGS | HUMAN DRUGS | ANIMAL DRUGS

Enter drug, NDC code, drug class, or Set ID

MORE WAYS TO SEARCH: ADVANCED SEARCH | BROWSE DRUG CLASSES | LABELING ARCHIVES

The DailyMed database contains **137159** labeling submitted to the **Food and Drug Administration (FDA)** by companies. DailyMed does not contain a complete listing of labeling for FDA-regulated products (e.g., labeling that is not submitted to the FDA). See ABOUT DAILYMED for more information.

SHARE

NEWS

DailyMed Announcements

Posted: August 17, 2020

Pillbox Retiring

Pillbox is being retired in January 2021, for further information please see the NLM Technical Bulletin. DailyMed will be removing associated links to Pillbox currently labeled as "TABLET/CAPSULE ID TOOL" in the "MORE WAYS TO SEARCH" section.

MORE INFO

Get RSS News & Updates

The DailyMed RSS feed provides updates and information about new drug labels approved by the FDA and published on NLM's DailyMed Web site.

FDA RESOURCES

SPL, Other Prescription Drug Labeling Resources, and Guidances

- FDA's Structured Product Labeling Resources
- FDA's Prescription Drug Labeling Resources
- FDA's Drug Guidances

NLM SPL RESOURCES

The following Structured Product Labeling (SPL) resources have been created to assist industry professionals.

Download Data

- All Drug Labels
- All Indexing & REMS Files
- All Mapping Files

SPL XML schema and implementation guide are here

Substance Indexing Files are here

- Not to be confused with Drug Labels
- Link drug products to ingredient substances
- One Substance Indexing File describes one substance with UNII in ***fully machine-readable format***

About role of InChI

One ring to rule them all, one ring to find them,
data
One ring to bring them all, and in the ~~darkness~~
bind them.

after J.R.R. Tolkien

Historic notes

2011 Task to develop structured *document* format for exchanging information on substances in medicinal products. Prerequisites:

- ISO IDMP 11238 standard for substances
- Data registered in FDA Substance Registration System (SRS, later ***G***SRS)
- Syntactic platform used for Structured Product Labeling (SPL)

2012 **Decision to make InChI the “must be present” characteristic for the exchange format**

2014 Substance Indexing Initiative announced. First small molecules and mixtures in SPL format published on DailyMed

2017 First modified biologics published on DailyMed

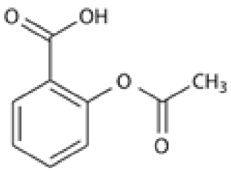
2021 First polymers published on DailyMed

About identification of medicinal substances

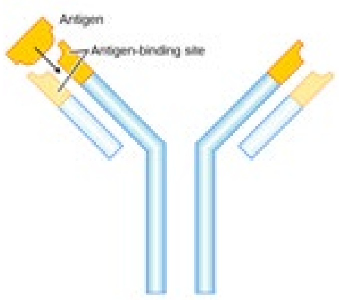
…it is easier for a camel to go through the eye of a needle…

Matthew

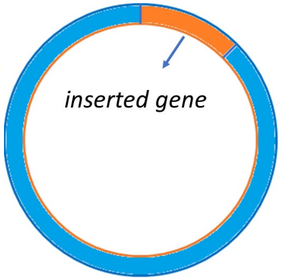
Chemical diversity of medicinal substances



Small molecule



Antibody



DNA vector

Polymer

Botanical extract

Identification of complex (bio)chemical substance using InChI

Comprehensive approach	Complex (Bio)chemical Substance	=	Single InChI + other attributes
Modular approach	Complex (Bio)chemical Substance	=	Complex data model where InChI(s) is/are component(s)

Concerns about *comprehensive* approach

- When complexity of (bio)chemical substance increases it may become problematic to create a unique identifier by exploiting the algorithm designed for much simpler objects

- Implementation of such complex identifiers may influence the original algorithm and make backwards compatibility impossible

About modeling complex data

Everything should be made as simple as possible, but no simpler.

A. Einstein

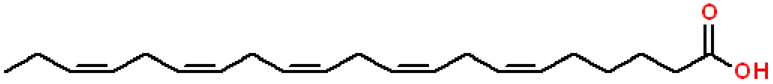
Our data model uses *modular* approach

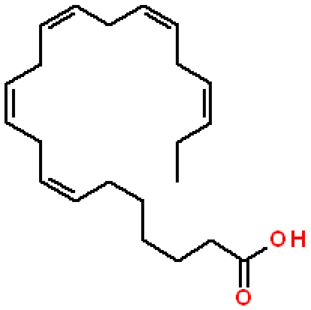
- **Concept of moiety:**
  - ➢ Any part of a substance. Does not have to be a complete functional group. Does not have to be covalently connected to other moieties.
- **Two types of moieties:**
  - Additive moiety
    - ➢ contributes to a whole complex substance
  - Site of interest
    - ➢ delineates features or sites of interests, such as amino acid connection points

Moiety “simple chemical”

- Is used to define a small molecule
- The structure of this moiety is represented by MOLFILE and/or SMILES
- **InChI is required for unique identification of the structure**
- Small proteins and nucleic acids (up to 999 atoms and 999 bonds) are also represented as simple chemicals

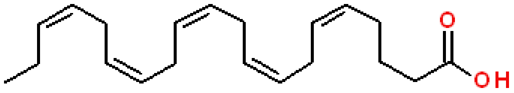
Moiety “mixture component”





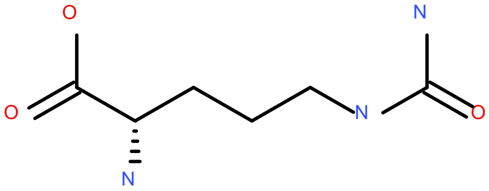
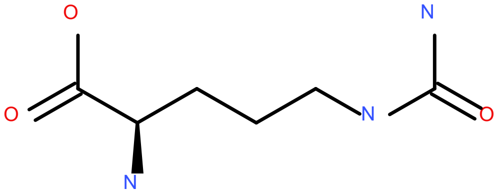
= InChI=1S/C22H34O2/c1-2-3-4-5-6-7-8-9-10-11-12-13-14-15-16-17-18-19-20-21-22(23)24/h3-4,6-7,9-10,12-13,15-16H,2,5,8,11,14,17-21H2,1H3,(H,23,24)/b4-3-,7-6-,10-9-,13-12-,16-15-

+ Amount expressed as mole fraction (f)



Molar mass of mixture: $M = \sum_{i=1}^{N\ of\ components}(f_i M_i)$

Racemate is also a mixture

		InChI		Amount
	=	InChI=1S/C6H13N3O3/c7-4(5(10)11)2-1-3-9-6(8)12/h4H,1-3,7H2,(H,10,11)(H3,8,9,12)/t4-/m0/s1	+	1/2
	=	InChI=1S/C6H13N3O3/c7-4(5(10)11)2-1-3-9-6(8)12/h4H,1-3,7H2,(H,10,11)(H3,8,9,12)/t4-/m1/s1	+	1/2

Molar mass calculation: $M = \frac{1}{2}M_1 + \frac{1}{2}M_2$

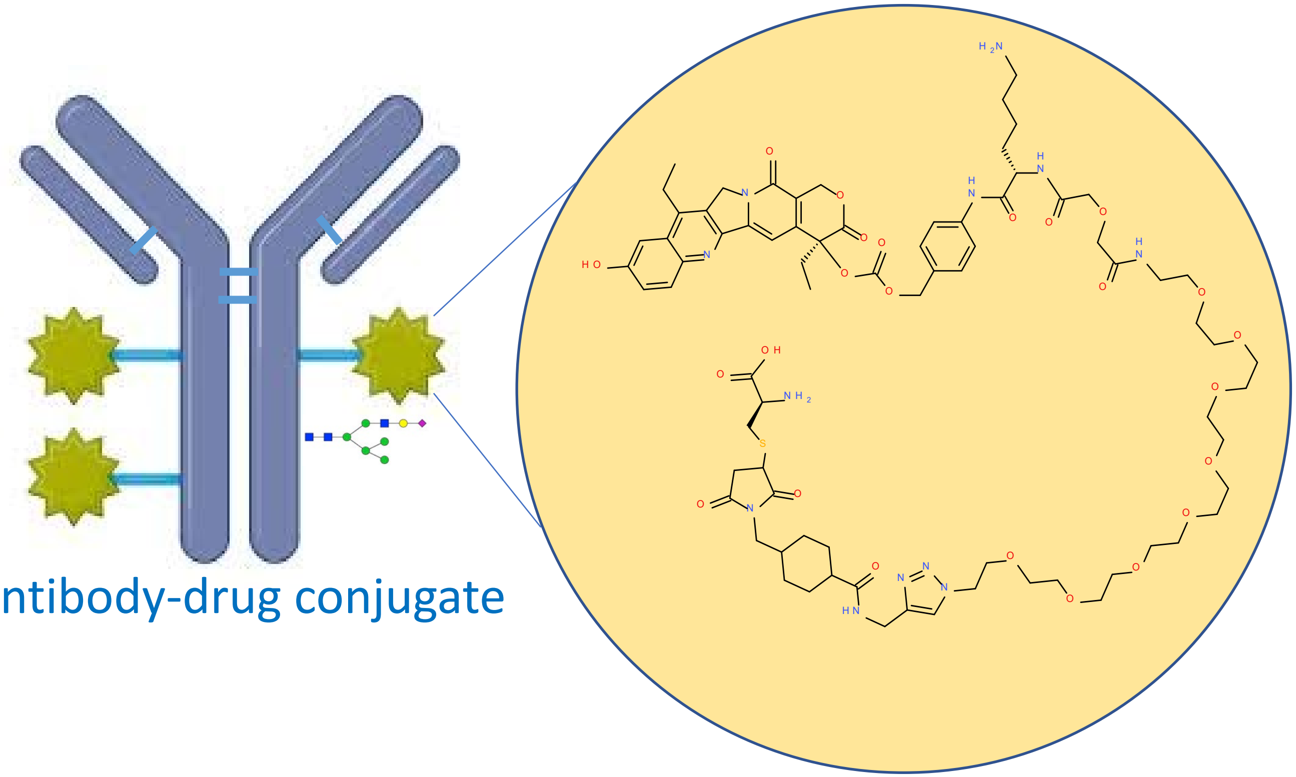
Moiety “protein subunit”

DIQMTQSPSSLSASVGDRVTITCRSSQSIVHSVGNTFLEWYQQKPG
KAPKLLIYKVSNRFSGVPSRFSGSGSGTDFTLTISSLQPEDFATYYCFQ
GSQFPYTFGQGTKVEIKRTVAAPSVFIFPPSDEQLKSGTASVVCLLN
NFYPREAKVQWKVDNALQSGNSQESVTEQDSKDSTYSLSSTLTLSK
ADYEKHKVYACEVTHQGLSSPVTKSFNRGEC

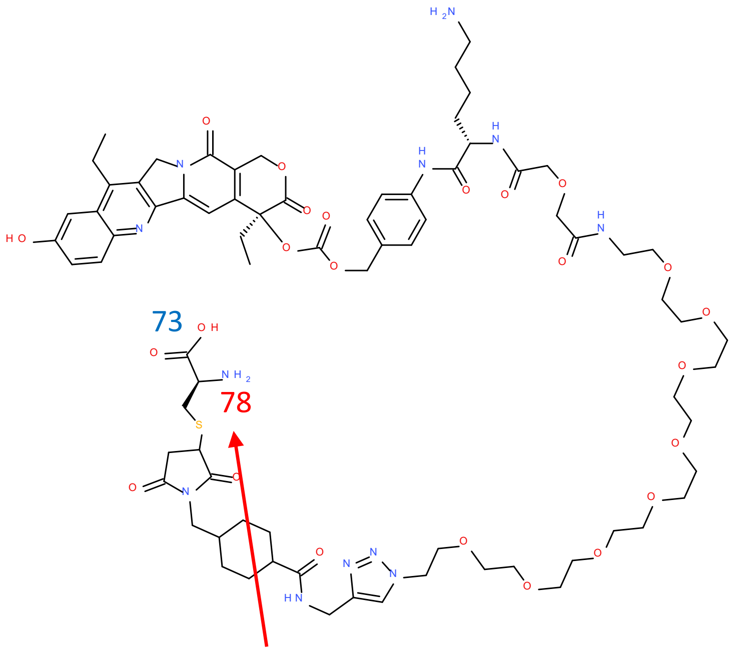
Moiety “polynucleotide”

```
CTGCGCGCTCGCTCGCTCACTGAGGCCGCCCGGGCAAAGCCCGGGCGTCGGGCGACCTTTGGTCGCCCGGCCTCA
GTGAGCGAGCGAGCGCGCAGAGGAGCGCGCAGAGAGGGAGTGGCCAACTCCATCACTAGGGGTTCCTTGTAGTT
AATGATTAACCCGCCATGCTACTTATCTACGTAGCCATGCTCTAGGTACCATTGACGTCAATAATGACGTATGTTCCCAT
AGTAACGCCAATAGGGACTTTCCATTGACGTCAATGGGTGGAGTATTTACGGTAAACTGCCCACTTGGCAGTACATC
AAGTGTATCATATGCCAAGTACGCCCCCTATTGACGTCAATGACGGTAAATGGCCCGCCTGGCATTATGCCCAGTACA
TGACCTTATGGGACTTTCCTACTTGGCAGTACATCTACGTATTAGTCATCGCTATTACCATGGTCGAGGTGAGCCCCAC
GTTCTGCTTCACTCTCCCCATCTCCCCCCCCTCCCCACCCCCAATTTTGTATTTATTTATTTTTTAATTATTTTGTGCAGC
GATGGGGGCGGGGGGGGGGGGGGGCGCGCCAGGCGGGGCGGGGCGGGGCGAGGGGCGGGGCGGGGC
GAGGCGGAGAGGTGCGGCGGCAGCCAATCAGAGCGGCGCGCTCCGAAAGTTTCCTTTTATGGCGAGGCGGCGGC
GGCGGCGGCCCTATAAAAAGCGAAGCGCGCGGCGGGCGGGAGTCGCTGCGCGCTGCCTTCGCCCCGTGCCCCGC
TCCGCCGCCGCCTCGCGCCGCCCGCCCCGGCTCTGACTGACCGCGTTACTCCCACAGGTGAGCGGGCGGGACGGC
CCTTCTCCTCCGGGCTGTAATTAGCGCTTGGTTTAATGACGGCTTGTTTCTTTTCTGTGGCTGCGTGAAAGCCTTGAG
GGGCTCCGGGAGGGCCCTTTGTGCGGGGGGAGCGGCTCGGGGCTGTCCGCGGGGGGACGGCTGCCTTCGGGG
GGGACGGGGCAGGGCGGGGTTCGGCTTCTGGCGTGTGACCGGCGGCTCTAGAGCCTCTGCTAACCATGTTCATGC
CTTCTTCTTTTTCCTACAGCTCCTGGGCAACGTGCTGGTTATTGTGCTGTCTCATCATTTTGGCAAAGAATTGGATCCT
AGCTTGATATCGAATTCCTGCAGCCCGGCGGCACCATGGCGGATACTCTCCCTTCGGAGTTTGATGTGATCGTAATAG
GGACGGGTTTGCCTGAATCCATCATTGCAGCTGCATGTTCAAGAAGTGGCCGGAGAGTTCTGCATGTTGATTCAAG
AAGCTACTATGGAGGAAACTGGGCCAGTTTTAGCTTTTCAGGACTATTGTCCTGGCTAAAGGAATACCAGGAAAAC
AGTGACATTGTAAGTGACAGTCCAGTGTGGCAAGACCAGATCCTTGAAAATGAAGAAGCCATTGCTCTTAGCAGGA
AGGACAAAACTATTCAACATGTGGAAGTATTTTGTTATGCCAGTCAGGATTTGCATGAAGATGTCGAAGAAGCTGGT
GCACTGCAGAAAAATCATGCTCTTGTGACATCTGCAAACTCCACAGAAGCTGCAGATTCTGCCTTCCTGCCTACGGA
GGATGAGTCATTAAGCACTATGAGCTGTGAAATGCTCACAGAACAAACTCCAAGCAGCGATCCAGAGAATGCGCTA
GAAGTAAATGGTGCTGAAGTGACAGGGGAAAAAGAAAACCATTGTGATGATAAAACTTGTGTGCCATCAACTTCAG
CAGAAGACATGAGTGAAAATGTGCCTATAGCAGAAGATACCACAGAGCAACCAAAGAAAAACAGAATTACTTACTC
ACAAATTATTAAAGAAGGCAGGAGATTTAATATTGATTTAGTATCAAAGCTGCTGTATTCTCGAGGATTACTAATTGAT
```

Moiety “structural modification” references a substituent defined in the same document



Definition of substituent relies on InChI *and* InChI canonical atom numbering



=

InChI=1S/C76H104N12O24S/c1-3-55-56-37-54(89)16-17-61(56)83-68-57(55)43-87-63(68)38-59-58(71(87)95)45-110-74(99)76(59,4-2)112-75(100)111-44-50-10-14-52(15-11-50)81-70(94)62(7-5-6-18-77)82-66(91)47-109-46-65(90)79-19-21-101-23-25-103-27-29-105-31-33-107-35-36-108-34-32-106-30-28-104-26-24-102-22-20-86-42-53(84-85-86)40-80-69(93)51-12-8-49(9-13-51)41-88-67(92)39-64(72(88)96)113-48-60(78)73(97)98/h10-11,14-17,37-38,42,49,51,60,62,64,89H,3-9,12-13,18-36,39-41,43-48,77-78H2,1-2H3,(H,79,90)(H,80,93)(H,81,94)(H,82,91)(H,97,98)/t49?,51?,60-,62-,64?,76-/m0/s1

+

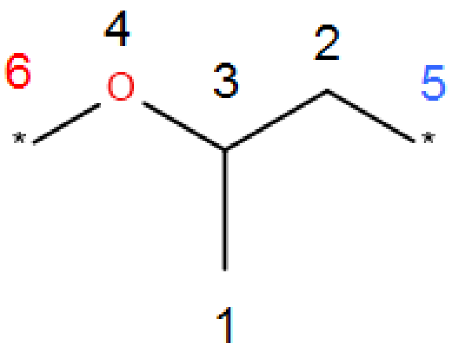
InChI canonical atom numbers → Moiety “Amino Acid Connection Points”: 78, 73

Moiety “polymer”

- Is used to define a stochastic (non-template driven) polymer
- Has one or more sub-moieties **“Structural Repeat Unit” (SRU)**
- Has amount associated with each SRU. Amount can be a range
- Does not include end groups
- End groups are moieties “structural modification”

Definition of SRU relies on InChI=1B/ (v. 1.06) and InChI canonical atom numbering

InChI Canonical SRU



InChI canonical atom numbers

=

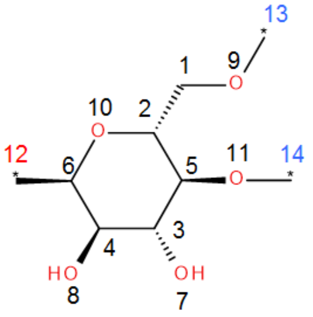
InChI with Experimental Beta Option – pseudo atoms

InChI=1B/C3H6OZz2/c1-3(2-5)4-6/h3H,2H2,1H3

+

Moiety “Linear SRU Connection Points”: 6, 5

SRU can be non-linear

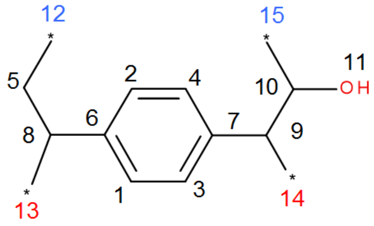


=

InChI=1B/C6H9O5Zz3/c7-3-4(8)6(12)10-2(1-9-13)5(3)11-14/h2-8H,1H2/t2-,3-,4-,5-/m1/s1

+

Moiety “Branched SRU Connection Points”: 12, 13, 14



=

InChI=1B/C10H10OZz4/c11-10(15)9(14)7-3-1-6(2-4-7)8(13)5-12/h1-4,8-11H,5H2

+

Moiety “Cross-linked SRU Connection Points”: 13, 14, 12,15

Hash code

- Concatenation of InChIs and other characteristics of moieties sorted in lexicographical orders feeds into a hashing algorithm and a unique hash is computed
- MD5 hash code is currently used
- Hash is a 32 ASCII HEX character string (displayed as a GUID-like string)
- The hash code is added to each Substance Indexing File as follows

```
<code code="1bc32748-ede8-3a96-fd46-a5497a4683ad“codeSystem="2.16.840.1.113883.3.2705" />
```

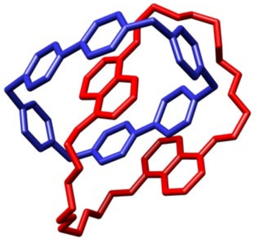
Summary

- SPL Substance Indexing File is a structured document that utilizes a modular approach in which InChI is ***not a single identifier*** of the substance but rather ***a contributor*** to a more complex data model
- All structural moieties represented by atoms and bonds are identified by their InChI
- Other structural moieties, such as protein subunits, are uniquely identified by the letter notation code
- Moieties of type “site of interest” use InChI canonical atom numbers
- All structural moieties and their modifications are uniquely identified within one document
- Linking between moieties is unambiguously defined, so that a complete molecular structure can be recreated
- Since all moieties are uniquely defined, it is possible to build independent canonical identifiers and hash codes including layered hash codes directly from the files

Improvements in InChI would be appreciated

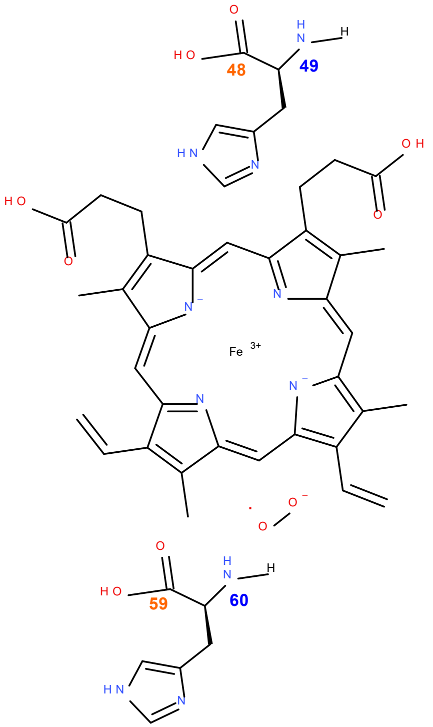
- E/Z stereochemistry of sulfoxides
- Tautomerism
- Organometallic and inorganic compounds
- Special stereo: hindered rotation, square planar, octahedral, etc.

- Topoisomerism ?



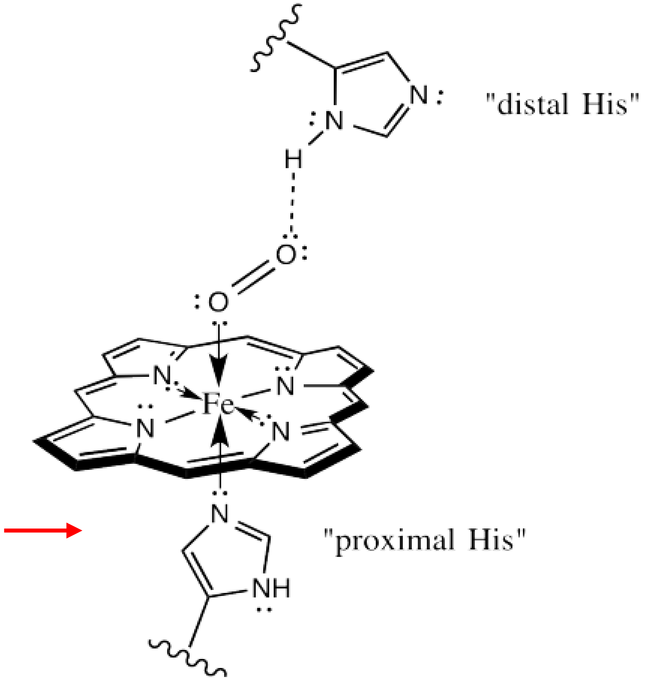
https://en.wikipedia.org/wiki/File:Catenane_Crystal_Structure_ChemComm_page634_1991_commons.png

Example organometallic compound: histidine-heme-histidine complex in Hemoglobin



Currently represented by disconnected fragments

Improvements could include coordination bond handling



Acknowledgments

Substance Indexing

Lavanya Balabhadra
Igor Filippov
Prasad Pallinti
Yuri Pevzner
George WashburnIV

SPL

Yisong Liu
Vedashree Puntambekar
Gunther Schadow
Lonnie Smith
Eva Tu

GSRS

Dammika Amugoda
Larry Callahan
Tyler Peryea
Frank Switzer

InChI 1.06

Steve Heller
Igor Pletnev

InChI Working Groups

Tautomers

Marc Nicklaus
Gerd Blanke
Evan Bolton
Alex M. Clark
Bret Daniel
Devendra Dhaked
Laura Guasch
Wolf-Dietrich Ihlenfeldt
Gregory Landrum
John W. Mayfield
Hitesh Patel
Igor Pletnev
Roger Sayle
Dmitrii Tchekhovskoi

Organometallics

Colin Batchelor
Gerd Blanke
Evan Bolton
Ian Bruno
Andrei Erin
Jane Frommer
Jonathan Goodman
Richard Hartshorn
Hinnerk Rey
Clare Tovee

Stereochemistry

Gerd Blanke
Andrey Erin
Jane Frommer
Burt Leland
Juergen Kammerer
Igor Pletnev
Clare Tovee

Questions?

Yulia.Borodina@fda.hhs.gov