IUPAC, nomenclature, and chemical representation: From the perspective of a worldwide structural database

Matt Lightfoot, Ian Bruno, Clare Tovee, Suzanna Ward, Seth Wiggin

The Cambridge Crystallographic Data Centre

ACS Fall 2019 Sunday August 25th 2019
Summary

• Introduction to the CCDC and the CSD
• The history of nomenclature in the CSD
• The importance of compound names in the CSD
• Current challenges with nomenclature
• Looking forward

IUPAC’s role in creating the CSD over the past 50 years
# The Cambridge Crystallographic Data Centre

<table>
<thead>
<tr>
<th><strong>International Data Repository</strong></th>
<th><strong>Originated in 1965</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>Archive of crystal structure data</td>
<td>Financially self-supporting</td>
</tr>
<tr>
<td>High quality scientific database</td>
<td>Not-for-profit, UK Registered Charity</td>
</tr>
<tr>
<td><strong>Collaborative Research Organisation</strong></td>
<td><strong>University Partner Institute</strong></td>
</tr>
<tr>
<td>New methodologies</td>
<td></td>
</tr>
<tr>
<td>Fundamental research</td>
<td></td>
</tr>
<tr>
<td><strong>Scientific Software Provider</strong></td>
<td><strong>Dedicated to the</strong></td>
</tr>
<tr>
<td>Search/analysis/visualisation tools</td>
<td><strong>advancement of chemistry</strong></td>
</tr>
<tr>
<td>Scientific applications</td>
<td><strong>and crystallography</strong> for the**</td>
</tr>
<tr>
<td><strong>Education and Outreach</strong></td>
<td><strong>public benefit through high</strong></td>
</tr>
<tr>
<td>Conferences, Workshops, Training, Teaching</td>
<td><strong>quality information services</strong></td>
</tr>
<tr>
<td></td>
<td><strong>and software</strong></td>
</tr>
</tbody>
</table>
The Cambridge Structural Database (CSD)

- Over 1 Million small-molecule crystal structures
- Over 80,000 datasets deposited annually
- Structures available for anyone to download
- Links to over 1,000 journals
- Enriched and annotated by experts
- Access to data and knowledge

XOPCAJ - The millionth CSD structure.
From experiments to knowledge

The aggregation of experimental datasets provides a foundation for resources that enable structural knowledge to be applied to scientific challenges across sectors and domains.

**CSD-System**: Find, analyse and communicate crystal structures

**CSD-Discovery**: Protein and ligand-based design of new drugs

**CSD-Materials**: Behaviour and properties of new materials

Association of chemistry and crystallography is key for enabling discovery of new insights.
Before electronic deposition

Hand-typed tables of coordinates

---

Look up of compound names
Publication of crystal structures today

Electronic data files deposited and disseminated via the Web and linked with journal articles

https://www.ccdc.cam.ac.uk/structures
Searching for structures

- Majority of searches of CSD are substructure searches; however:
  - 16% of all searches in WebCSD are on compound name
Assignment of a chemically meaningful representation is determined using data in the CSD and manual curation.

Sources of names used in the CSD

- CIF or Paper
  - Particularly helpful for capturing stereochemistry and trivial names of drugs and natural products
- Use existing entries in the CSD
- Manually construct the name
- The majority of compounds are automatically named using the naming computer software
Using ACD/Name

• Handles *most* organics well

• Types of difficult cases
  – Symmetry
  – Unusual valences
  – Multicomponent structures
  – Large structures
  – Coordination complexes/ polymers
Adoption of using ACD/Name

- Software speeds up the validation of structures
- CCDC Editors have been using ACD/Name to assist with naming for many years
- An early key issue was how it handled organometallics
  - 62/96 organometallics; 130/156 organics
  - overall success rate of 76%
- CCDC now uses ACD/Name to routinely generate an IUPAC name for most incoming structures
IUPAC and CSD conventions

- Generally use IUPAC name
- For ease of searching we will use semi-systematic names in compound name or synonym field e.g.; Calixarenes, Ferrocene, Cucurbits, Catenanes, Rotaxanes etc.

**Compound:** 25,26,27,28-Tetramethoxy-5,17-dinitropentacyclo[19.3.1.1^3,7.1^9,13.1^15,19]octacosa-1(25),3(28),4,6,9(27),10,12,15(26),16,18,21,23-dodecaene chloroform solvate

**Synonym:** 25,26,27,28-Tetramethoxy-5,17-dinitrocalix(4)arene chloroform solvate
Challenges

• Providing consistent chemical representations
• Large complicated structures
• Polymeric structures
• New types of structures
• Changes to the rules for naming compounds
• Polymorphs and stereoisomers
• New nomenclature definitions for naming groups of structures
Reliable input representations

- How best to reliably represent organometallics?
  - dative vs covalent bonds?
  - explicit hydrogens/valencies?
  - dummy atoms?
  - zero-order bonds?

PubChem SDF

- PUBCHEM_NONSTANDARDBOND

MOL V3000

<table>
<thead>
<tr>
<th>Bond Type</th>
<th>Code</th>
</tr>
</thead>
<tbody>
<tr>
<td>Single</td>
<td>1</td>
</tr>
<tr>
<td>Double</td>
<td>2</td>
</tr>
<tr>
<td>Triple</td>
<td>3</td>
</tr>
<tr>
<td>Quadruple</td>
<td>4</td>
</tr>
<tr>
<td>Dative</td>
<td>5</td>
</tr>
<tr>
<td>Complex</td>
<td>6</td>
</tr>
<tr>
<td>Ionic</td>
<td>7</td>
</tr>
</tbody>
</table>

* excluding query bond types

ACD/Labs MOL V2000 Extensions

- ZZF 3 1 41 2 42 3 43
- ZZH 1 5 2 3 4 5 6
- ZZH 2 5 7 8 9 10 11
- ZZH 3 5 12 14 15 16 17
- ZZE 2 42 18 43 18
Consistent structure representation

CSD: Representations of Imidazolin-2-ylidene metal carbenes in the CSD 5.38 (N = 9468)


Organometallics 2007, 26, 18, 4684-4687, doi:10.1021/om700498w
Challenges – Cyclic and large structures

 Difficulty naming structures with more than 15 rings or 255 atoms

[2]-bis(μ₆-2,6,10,20,24,28,38,42,46(5,2),
3,7,11,21,25,29,39,43,47(2,5)–
octadecapyridina-1,12,19,30,37,48(1,2)–
hexabenzenacyclotetrapentacontaphane-
15,33,51-triene)-hexa-iron-catenane

Star of David catenane

N,N'-bis(6-(6-(6-(6-(6-(4-Isobutoxy-8-nitroquinoline-2-
carboxamido)-4-isobutoxyquinoline-2-
carboxamido)pyridine-2-carbamoyl)pyridine-2-
carboxamido)pyridine-2-yl) 1,8-diaza-4,5-di-
isobutoxy-9-methyl-2,7-anthracenedicarboxamide
Challenges - Clusters

hexakis(tetra-n-butylammonium) tris(μ⁶-oxido)-bis(μ⁴-(tris(oxidomethyl)methyl)imino)-tricontakis(μ²-oxido)-hexadecakis(aqua)-dodeca-molybdenum-hexa-vanadium acetonitrile solvate
Challenges - Polymers

- To name polymers we
  - break the compound down into components of ligands and metals
  - manually construct the name

- ACD occasionally attempts to create a whole name, however it doesn’t always
  - recognise polymer bonds
  - give stoichiometry of ligands or anions
New types of structures

• Graphene-type structures

Nanographene
hexa-n-butyl-C54-nanographene
tetrakis(n-propoxy)-octakis(t-butyl)-nanographene
Stereochemistry

- CSD add stereochemical descriptors if given by the author in the cif or paper
- Group families together of same compound
- Link by cross-referencing diastereomers/enantiomers, but not same group
Polymorphism

Same tautomer – different crystal forms – different interactions – different stability

Ritonavir

ACRDIN
- polymorph II
- polymorph III
- polymorph IV
- polymorph V
- polymorph VI
- polymorph VII

PAMMNC
- α polymorph
- β polymorph
- γ polymorph
- δ polymorph
- ε polymorph
- θ polymorph

CUIMDZ
- blue J I2/a polymorph
- blue B trigonal polymorph
- green orthorhombic polymorph
- olive C2/c polymorph

ADEDIX
- monoclinic disappearing polymorph
- tetragonal polymorph

YIGPIO
Metastable monoclinic polymorph 1

YIGPIO01
Stable orthorhombic polymorph 2
Tautomerismic polymorphism

Different tautomers – different crystal forms – different interactions – different stability

Less stable tautomers can sometimes form more stable hydrogen bonded networks

A J Cruz-Cabeza, C Groom, CrystEngComm, 2011, 13, 93
Changes in naming recommendations

• Naming coordination polymers

https://iupac.org/projects/project-details/?project_nr=2011-035-1-800
Changes in naming recommendations

• Challenges with naming coordination polymers
  • Consistency between naming polymers and coordination complexes
  • How to deal with changes to the recommendations that affect existing entries
  • How to deal with indicating deprotonation of ligands in multiple positions
  • Organic polymers that are cross-linked with coordination

https://iupac.org/projects/project-details/?project_nr=2011-035-1-800
Changes in nomenclature rules

https://doi.org/10.1039/9781849733069
Changes in nomenclature rules

(f) The prefixes ‘ureido’ and ‘ureylene’ are not used in preferred IUPAC names. The prefixes ‘carbamoylamino’ and ‘carbonylbis(azanediyl)’, respectively, are recommended for preferred IUPAC names and in general nomenclature.

(g) The prefix ‘guanidino’ is no longer acceptable in preferred IUPAC names; the preferred prefix is ‘carbonimidamido’.

(h) The prefix ‘amidino’ is no longer acceptable in preferred IUPAC names; the preferred prefix is ‘carbamimidoyl’.

(i) The prefix ‘aci-nitro’ for HO-N(O)= is no longer acceptable for preferred IUPAC names. The preferred prefix is ‘hydroxy(oxo)-\(\lambda^5\)-azanylidene’; the prefix recommended in the 1993 Guide (ref. 2) ‘hydroxynitroaryl’ is not acceptable in the context of these recommendations where two free valences must be expressed by the correct ‘ylidene’ or ‘diyl’ type.

(j) Preselected prefixes derived from the preselected parent hydride hydrazine are now formed systematically from hydrazine: ‘hydrazinyl’ for \(\text{H}_2\text{N-NH}^-\); ‘hydrazinylidene’ for \(\text{H}_2\text{N-N}=-\); ‘hydrazinediylidene’ for \(-\text{N-N}=\); and hydrazine-1,2-diyl for \(-\text{NH-NH}^-\). The prefixes ‘hydrazino’, ‘hydrazono’, ‘azino’ and ‘hydrazo’, respectively, are no longer acceptable, even for general nomenclature.

https://doi.org/10.1039/9781849733069
Changes in nomenclature rules

- CSD 5.40 + 2 updates
- Ureido – 449 hits (Carbamoylamino – 54 hits)
- Ureylene – 33 hits (Carbonylbis(azanediyl) – 5 hits)
- Guanidino – 189 hits (Carbonimidamido – 0 hits)
- Amidino – 167 hits (Carbamimidoyl – 315 hits)

https://doi.org/10.1039/9781849733069
Naming of topologies for Metal Organic Frameworks

- IUPAC task force
  - Challenge is to come up with a reliable and consistent topology
  - Different topologies for the same structure can be assigned based on the method used
  - Task force looking at trying to gain a community consensus on the best approach
Summary – the future

• Naming of compounds will continue to be really important
• One of the main challenges is being consistent
  • Consistent between old and new entries
  • Consistent between groups of entries
• IUPAC continues to be very important in tackling these challenges
Thank You

- CINF: One Million Crystal Structures: A Wealth of Structural Chemistry Knowledge
- Tuesday Aug 27 1.30 PM – 5.20 PM
- Wednesday Aug 28 9.00 AM – 12.15 PM
- Wednesday Aug 28 1.30 PM – 4.40 PM
- Grand Ballroom D, Omni San Diego Hotel

http://www.ccdc.cam.ac.uk/
@ccdccambridge
ccdc.cambridge