Extended Structures, Crystallography and Polymers - Challenges

Clare Tovee, Matthew Lightfoot, Suzanna Ward, Ian Bruno

Cambridge Crystallographic Data Centre

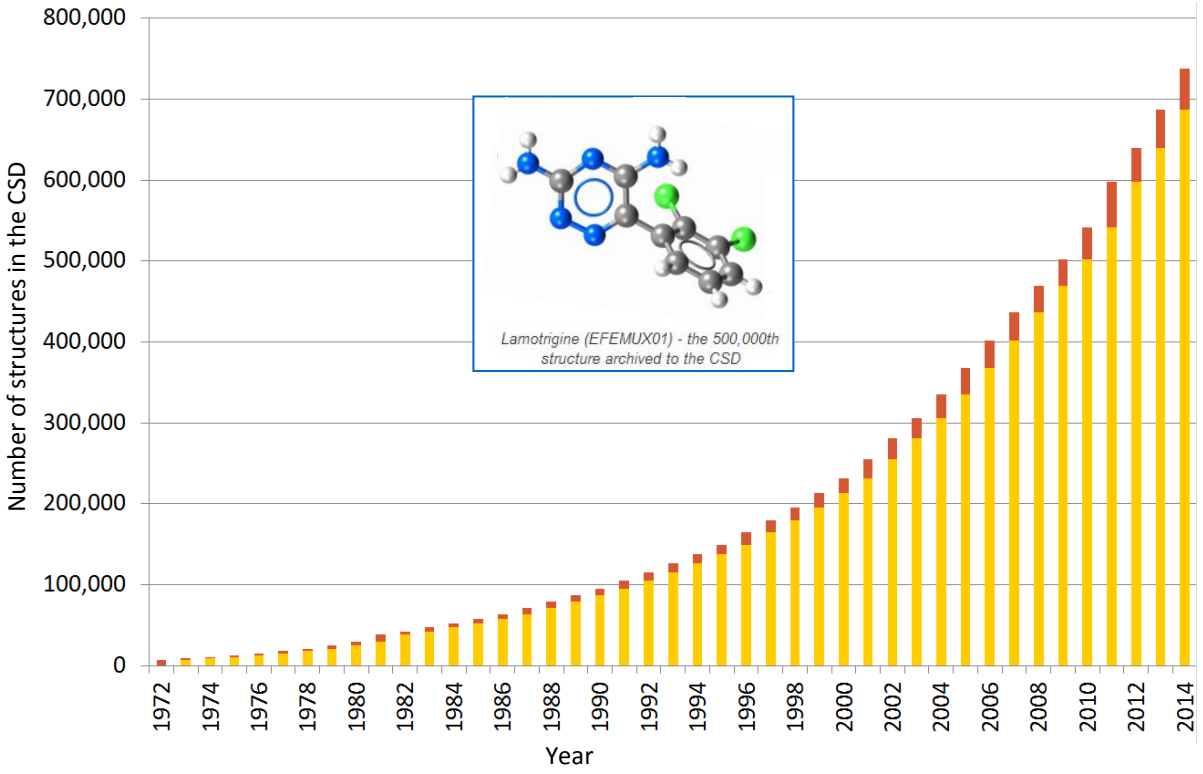
October 2014

The Cambridge Crystallographic Data Centre

- A not-for-profit, charitable institution, established 1965
- Community funded and governed
- A University of Cambridge Partner Institute
- Around 60 members of staff
  - In Cambridge, UK and at Rutgers University, US
- Creates and distribute the Cambridge Structural Database System

Cambridge Structural Database

Worldwide repository of small-molecule organic & metal-organic crystal structures



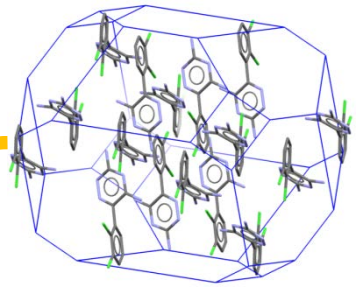
CSD Growth 1970-2014

Now 744,232 entries

Fully comprehensive:

- All published structures
- ASAP & 'Early View'
- Private Communications
- Patents

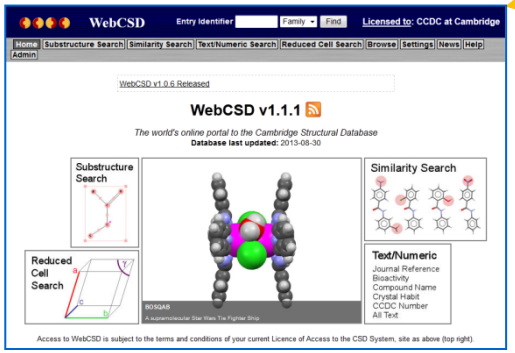
The Cambridge Structural Database System



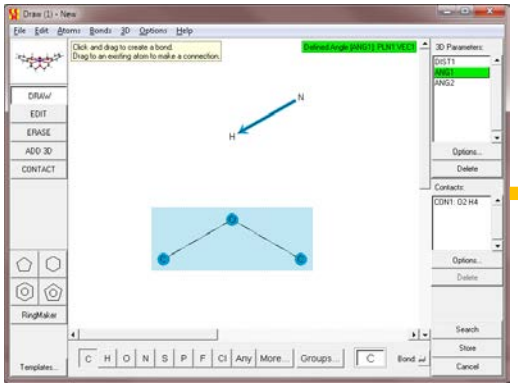
PreQuest: Create (in-house) database

Cambridge Structural Database

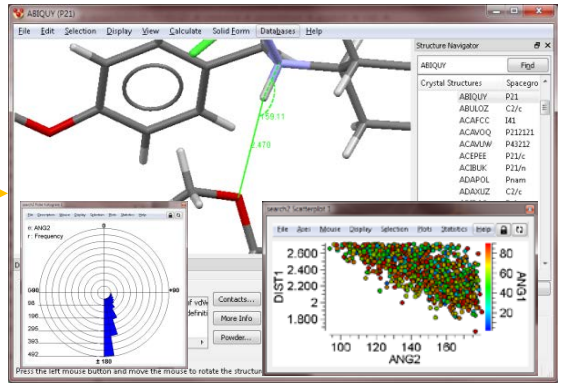
WebCSD: On-line portal to the CSD



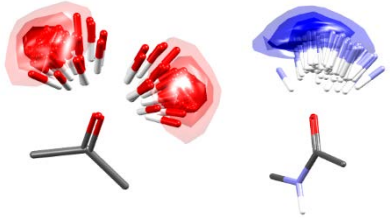
ConQuest: Advanced 3D searching



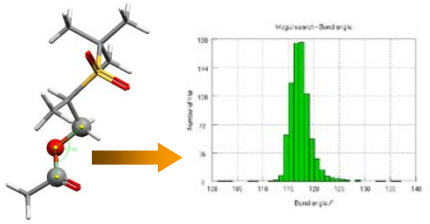
Mercury: Visualisation & data analysis



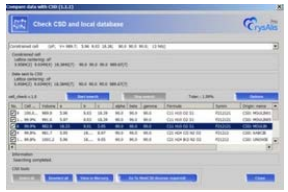
IsoStar: Molecular interaction analysis



Mogul: Molecular geometry analysis

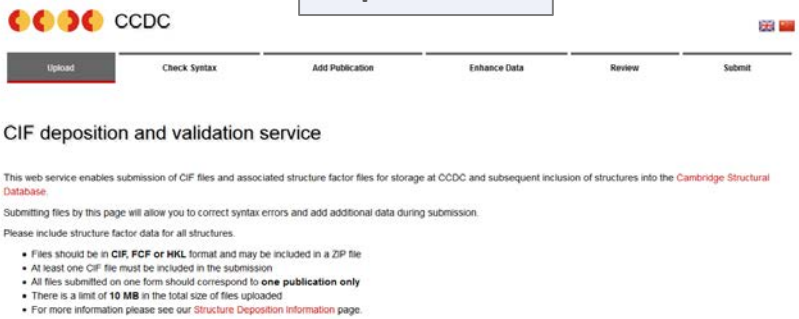


CellCheckCSD

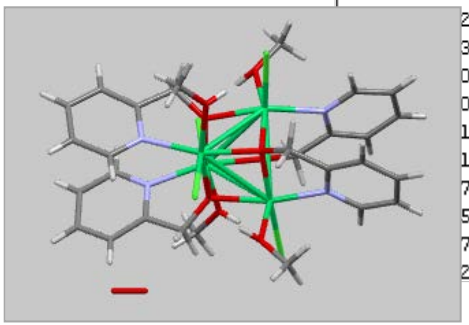


The Cambridge Structural Database

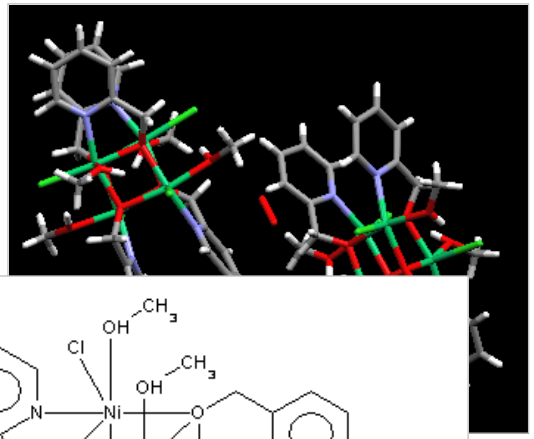
Deposited CIF

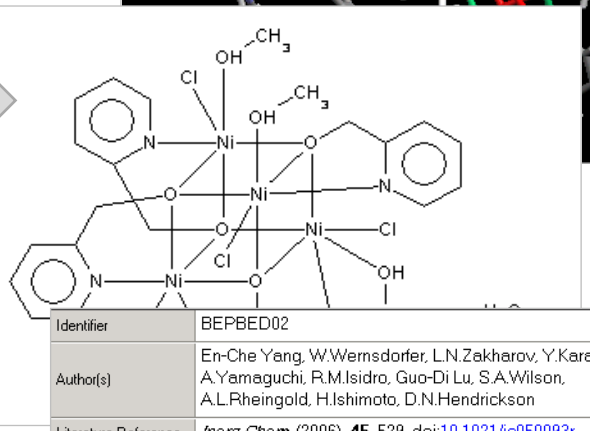



```
loop_
_atom_site_label
_atom_site_type_symbol
_atom_site_fract_x
_atom_site_fract_y
_atom_site_fract_z
C1  C  0.31594(37)  0.75375(20)  0.70189(16)
C2  C  0.19196(41)  0.77066(23)  0.76436(17)
   2259(32)  0.80293(21)  0.83459(13)
   3051(40)  0.74837(26)  0.73926(18)
  05699(40)  0.75910(26)  0.77720(18)
  00500(38)  0.71281(22)  0.66463(17)
  11609(38)  0.69941(22)  0.65036(17)
  1523(28)  0.69499(17)  0.60921(13)
  7057(33)  0.71613(20)  0.62788(16)
  5291(33)  0.70477(20)  0.58724(16)
  7764(37)  0.65275(21)  0.52604(17)
  2644(37)  0.69329(21)  0.48496(17)
```



CSD Entry

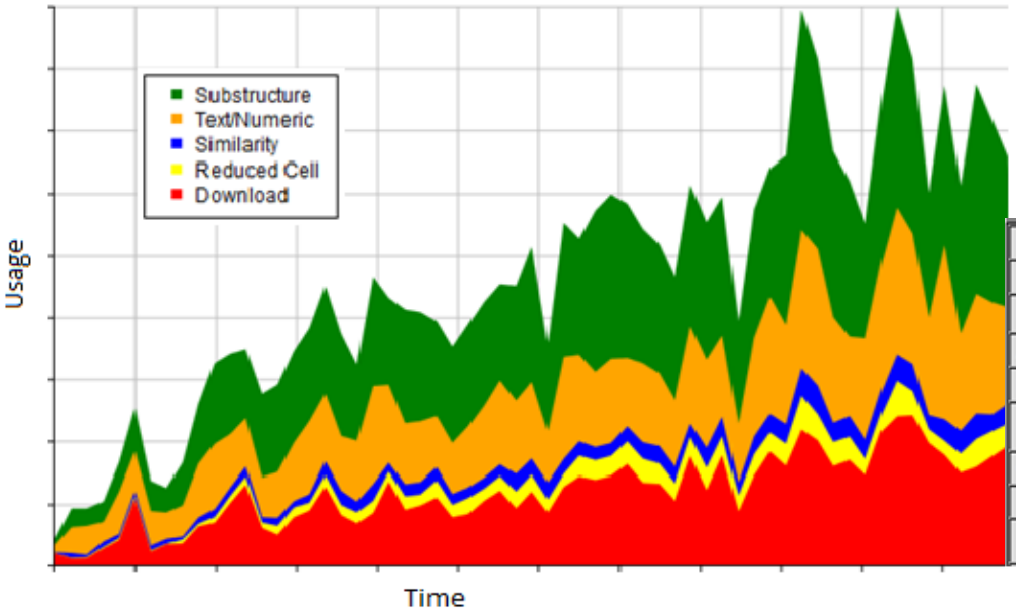


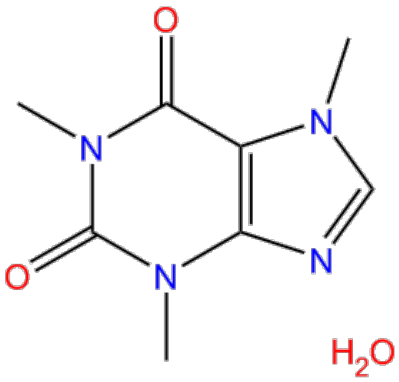


Identifier	BEPBED02
Author(s)	En-Che Yang, W.Wernsdorfer, L.N.Zakharov, Y.Karaki, A.Yamaguchi, R.M.Isidro, Guo-Di Lu, S.A.Wilson, A.L.Rheingold, H.Ishimoto, D.N.Hendrickson
Literature Reference	*Inorg.Chem.* (2006), **45**, 529, doi:10.1021/ic050093r
Formula	$C_{28}H_{40}Cl_4N_4Ni_4O_8.H_2O$
Compound Name	tetrakis(μ3-2-(Oxymethyl)pyridine-N,O,O,O)-tetrachloro-tetrakis(methanol)-tetra-nickel(ii) monohydrate
Synonym	
Space Group	I -4 2 d
Cell Lengths	**a** 16.1421(6) **b** 16.1421(6) **c** 29.4689(14)
Cell Angles	**α** 90 **β** 90 **γ** 90
R-Factor (%)	5.21
Disorder	The water molecule is disordered by symmetry.

Searching for structures

- Majority of searches of CSD are substructure searches; however:
  - 16% of all searches in WebCSD are on compound name
- Searching by compound name can still be useful, especially for complex structures and those with common trivial names





Identifier	CAFINE
Source Database	as531be
Reliability Score	★★★★ Explain score
Author(s)	D.J.Sutor
Reference	*Acta Crystallogr.* (1958), **11**, 453, doi:10.1107/S0365110X58001286
Formula	$C_8 H_{10} N_4 O_2.H_2 O$
Compound	1,3,7-Trimethyl-purine-2,6-dione monohydrate
Synonym	**Caffeine** monohydrate
Space Group	$P\ 2_1/a$

Common name searches in the database

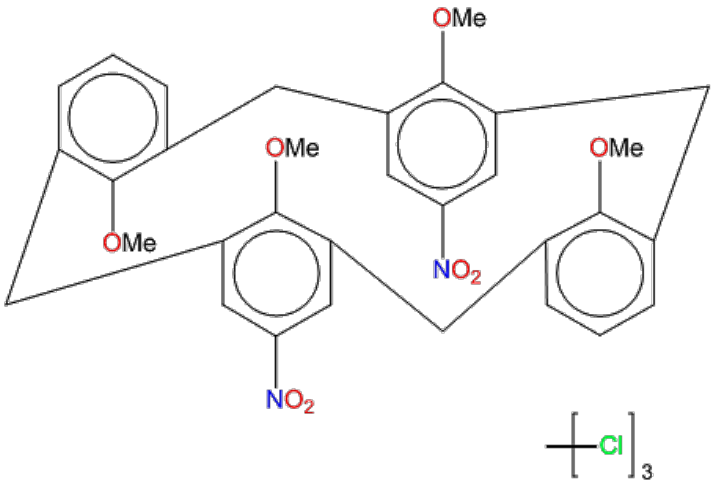
Common name searches in the database

Common name searches in the database

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.; Fullerenes, Calixarenes, Ferrocene, Cucurbits, Catenanes, Rotaxanes etc.

Classifiers
- Carbohydrate
- Nucleoside or Nucleotide
- Amino-acid, peptide and complexes
- Porphyrin, Corrin Complexes
- Steroid
- High Polymer

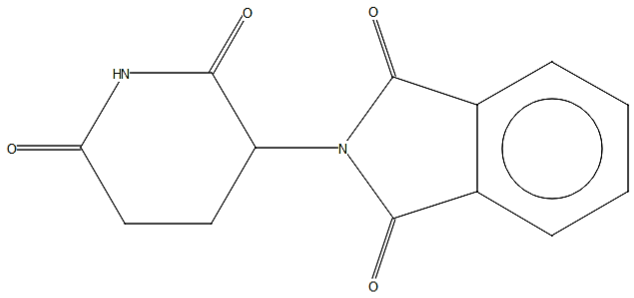
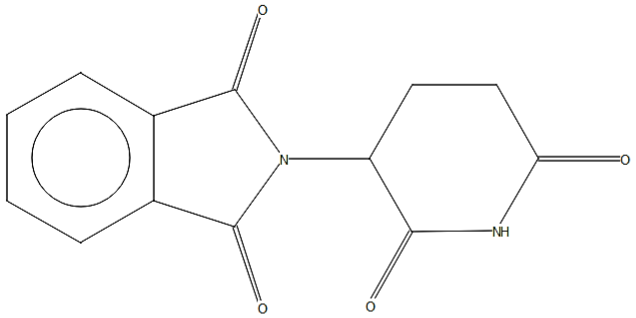


Compound: 25,26,27,28-tetramethoxy-5,17-dinitropentacyclo[19.3.1.1^{3,7}.1^{9,13}.1^{15,19}]octacosa-1(25),3(28),4,6,9(27),10,12,15(26),16,18,21,23-dodecaene chloroform solvate

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

Stereochemistry and Polymorphs

Overview | Register | All Text | Diagram | Visualiser | All Hits

Chemical Diagram		
Number Of Coordinates	29	29
Compound Name	(R,S)-2-(2,6-dioxopiperidin-3-yl)-1H-isoindole-1,3(2H)-dione	rac-2-(2,6-Dioxo-3-piperidinyl)-1H-isoindole-1,3(2H)-dione
Synonym	Thalidomide	Thalidomide; N-(2,6-dioxo-3-piperidyl)phalimide
Formula	$C_{13}H_{10}N_2O_4$	$C_{13}H_{10}N_2O_4$
Cell Lengths	a 8.3156(3) b 9.9732(4) c 14.5740(5)	a 20.679(5) b 8.042(2) c 14.162(5)
Cell Angles	α 90.0000 β 102.762(2) γ 90.0000	α 90 β 102.86(3) γ 90
Space Group	P $2_1/n$	C 2/c
Temperature (K)	93.1	Room Temp.(283-303)
Polymorph	alpha polymorph	beta polymorph
Cross References		STEREOMER: for stereoisomer see [1091777-CUR]
CCDC Number	CCDC 1009508	CCDC 628395
Literature Reference	Toshiya Suzuki, Masahito Tanaka, Motoo Shiro, Norio Shibata, Tetsuya Osaka, and Toru Asahi, *Phase Transitions* (2010), **83**, 223, doi:10.1080/01411591003605986	M.R.Caira, S.A.Botha, D.R.Flanagan, *J.Chem.Cryst.* (1994), **24**, 95, doi:10.1007/BF01665353

Search Results

- CCDC Number
- Reduced Cell
- Qualified Formula
- Qualified 2D Connectivity
- Exact 2D Connectivity

Solvents

Matched

Recommendation

Same

Polymorph

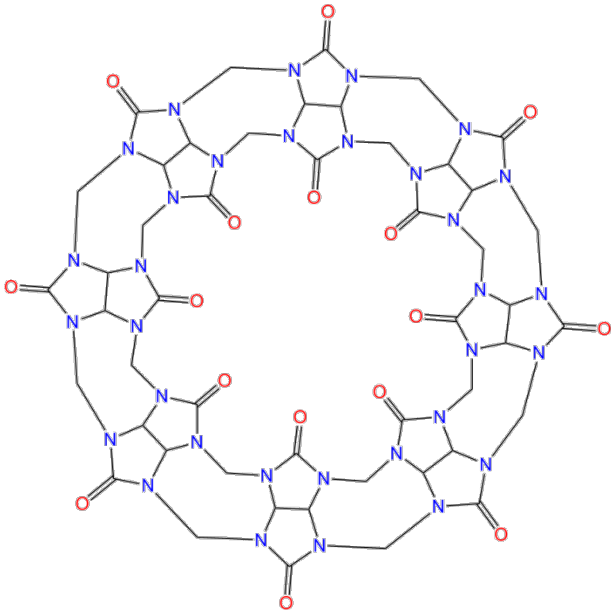
Different

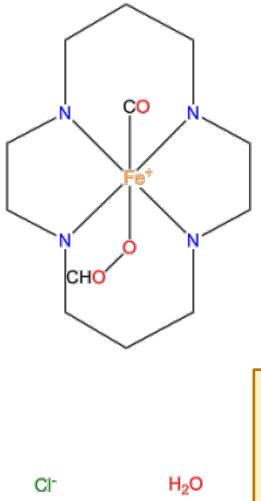
Stereoisomer

Probably Same

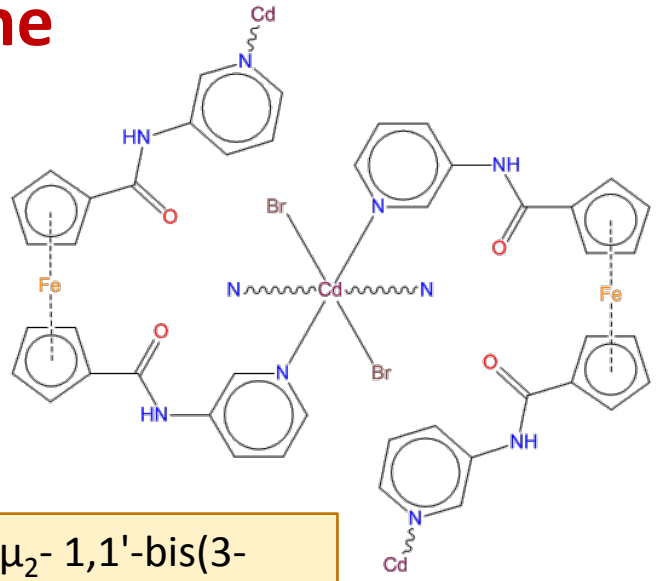
- CSD add stereochemical descriptors if given by the author in the CIF or paper
- Polymorphs – use descriptors and class together
- Group families together of same compound
- Link by cross-referencing diastereomers/enantiomers, but not same group

Types of structures we need to name



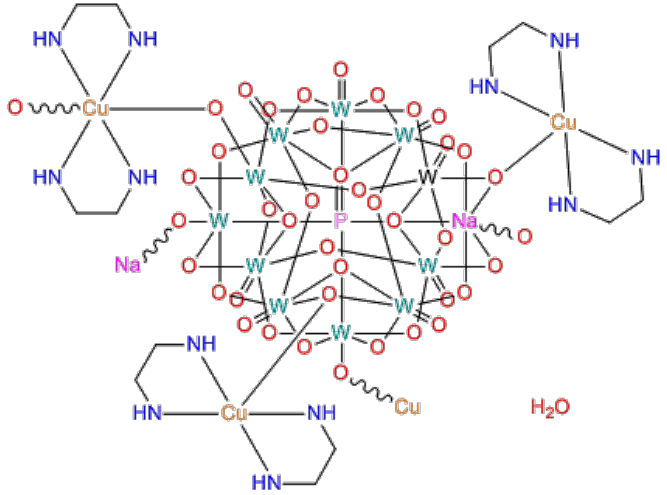


trans-(1,4,8,11-tetraazacyclotetradecane)-carbonyl-formato-iron(ii) cucurbit(8)uril clathrate chloride pentadecahydrate



catena-[bis(μ_2- 1,1'-bis(3-Pyridylcarbamoyl)ferrocene-N,N')-dibromo-cadmium(ii)]

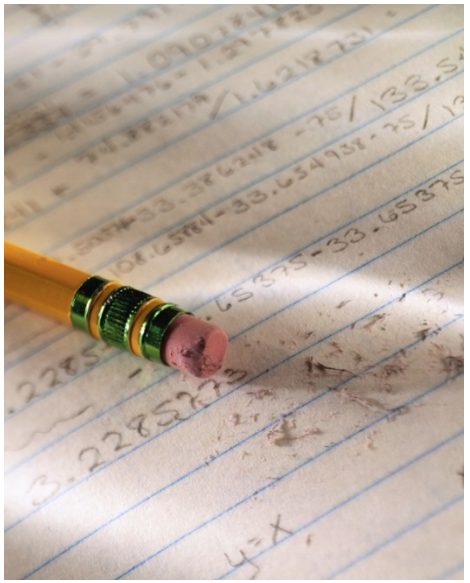
catena-[(μ_{12}-phosphato)-bis(μ_3-oxo)-pentacosakis (μ_2-oxo)-hexakis(1,2-diaminoethane-N,N')-octa-oxo-tri-copper(ii)-sodium-tungsten(vi) dihydrate]



Sources of names used in the CSD

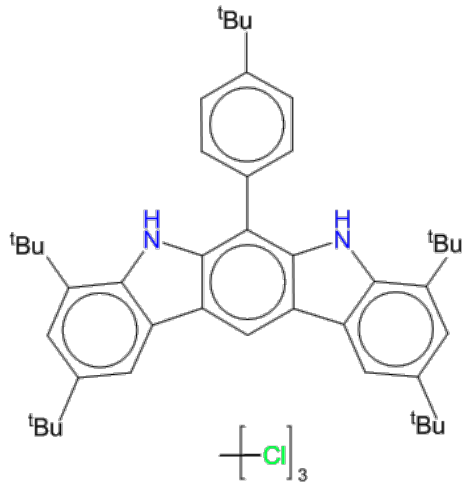
- When an entry is generated any useful information is extracted from the CIF
- The majority of compounds are automatically named using the naming computer software

- 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



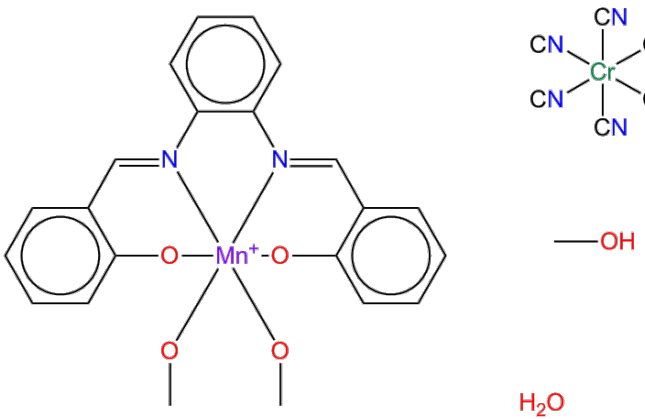
Why not just take the author's compound name?

- Full systematic names for complex compounds are often not given in the CIF or paper
- Examples of names from CIFs and CSD name
  - Paula4
  - Mn complex
  - LNO2
  - University of Somewhere



Compound 8

2,4,8,10-tetra-t-butyl-6-(4-t-butylphenyl)-5,7-dihydroindolo[2,3-b]carbazole chloroform solvate



non-bridge cyano

(2,2'-(1,2-phenylenebis(nitrilomethylidene))diphenolato)-bis(methanol)-manganese methanol solvate hydrate

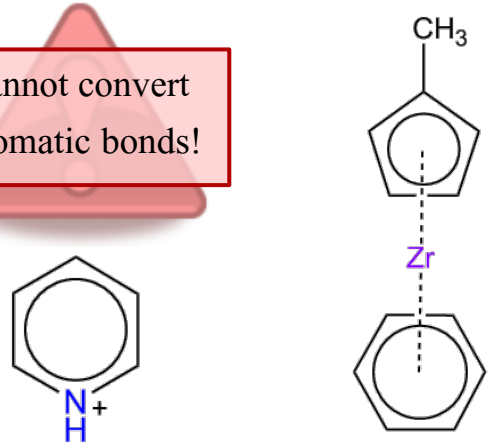
Adoption of using ACD/Name

- Software speeds up the validation of structures
- CCDC Editors have been using ACD/Name to assist with naming for several years
- During 2006, a more systematic evaluation of ACD/Name was carried out
- A 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

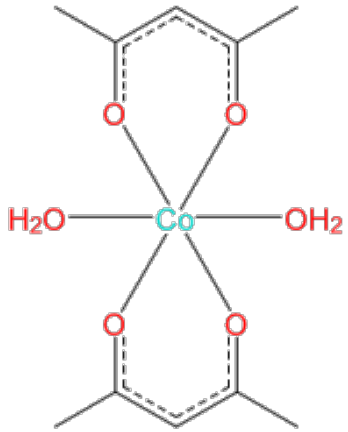
Challenges – Bond types

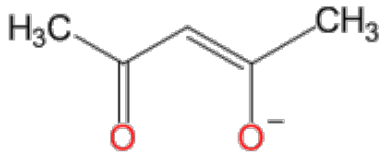
- The CSD has defined aromatic and delocalised bond types, however third-party software uses alternate single/double bonds for aromatic rings and delocalised bonds
- This also occurs for delocalised systems

Cannot convert aromatic bonds!



(η^5 -methylcyclopentadienyl)-(η^6-benzene)-zirconium





Structure should not contain bonds with partial order!

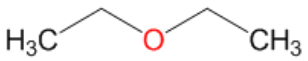
diaqua-bis(acetylacetonato)-cobalt

- CSD convention to use η^5, η^2 etc. to denote pi-bonding in organometallics. The software can name pi systems (but numbers them)

Using ACD/Name

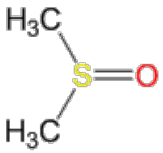
- Handles *most* organics well
- Types of difficult cases
  - Symmetry
  - Unusual valences
  - Multicomponent structures
  - Large structures
  - Coordination complexes/ polymers

Molecular Symmetry



ethoxyethane

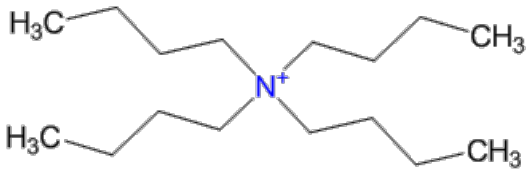
diethyl ether



(methylsulfinyl)methane

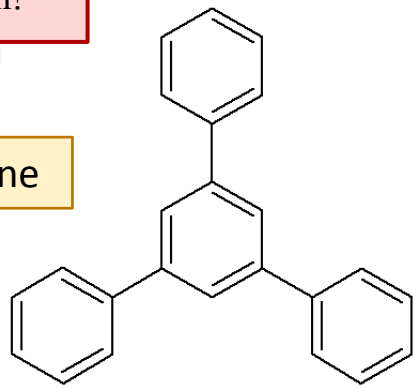
dimethyl sulfoxide

Structure contains assemblies of cyclic systems which are not supported in current version!



N,N,N-tributylbutan-1-aminium

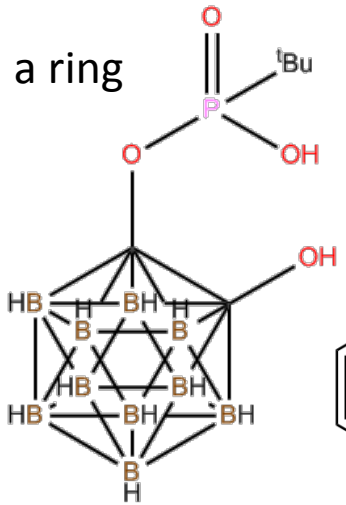
tetra-n-butylammonium



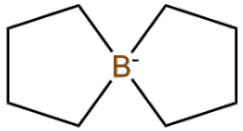
1,3,5-triphenylbenzene

Challenges

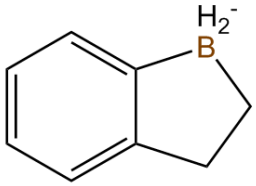
- Unusual valencies, especially for tetra-coordinated boron
  - Particularly when boron is in a ring
  - Or hydrogens are involved
  - Boron cages



1-[(1,1-dimethylethyl)hydroxyphosphinyl]oxy-2-hydroxy-1,2-dicarba-closo-dodecaborane(10)

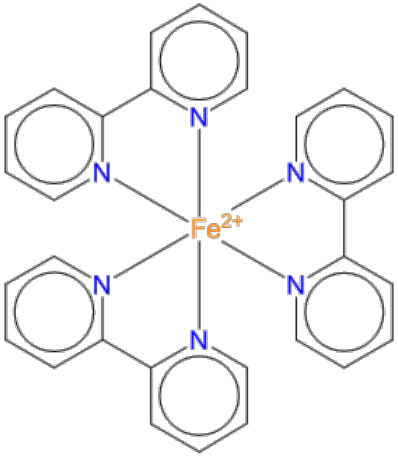


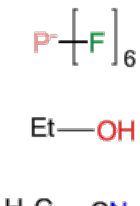
Structure contains atom(s) in disallowed valence state!



Structure contains a fragment which is not allowed in current version!

- ACD struggles to generate a full name for structures with more than three components (excluding water)



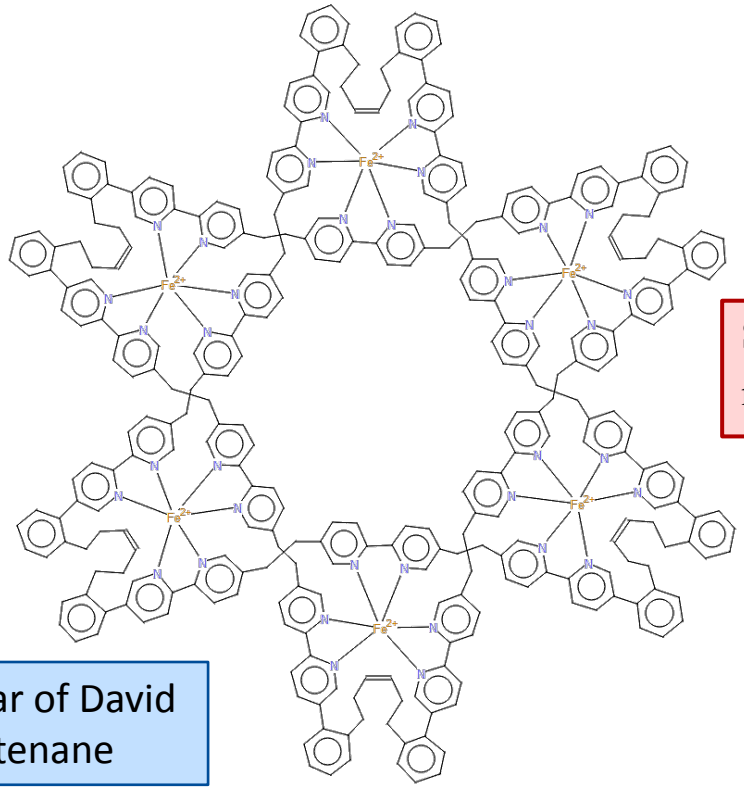


This type of multicomponent structures is not supported in current version!

tris(2,2'-bipyridine)-iron(ii) bis(hexafluorophosphate) acetonitrile ethanol solvate

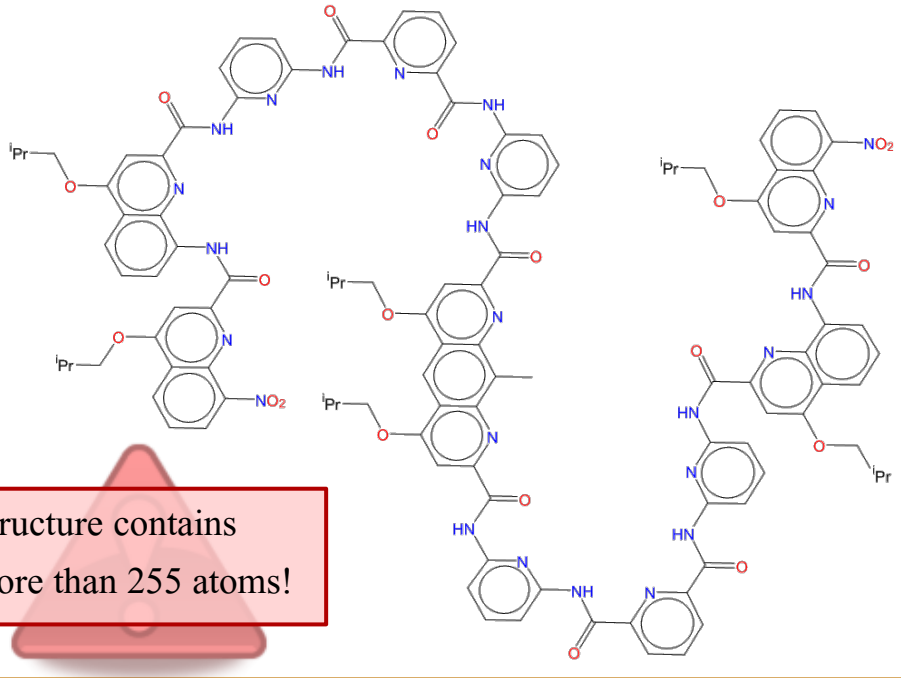
Challenges – Cyclic and Large structures

[2]-bis(μ^6-2,6,10,20,24,28,38,42,46(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

ACD/Name limits: structures with more than 15 rings or 255 atoms

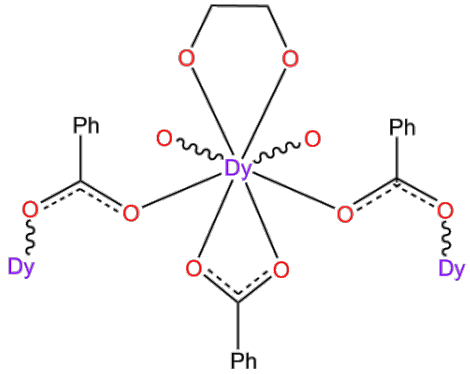


Structure contains more than 255 atoms!

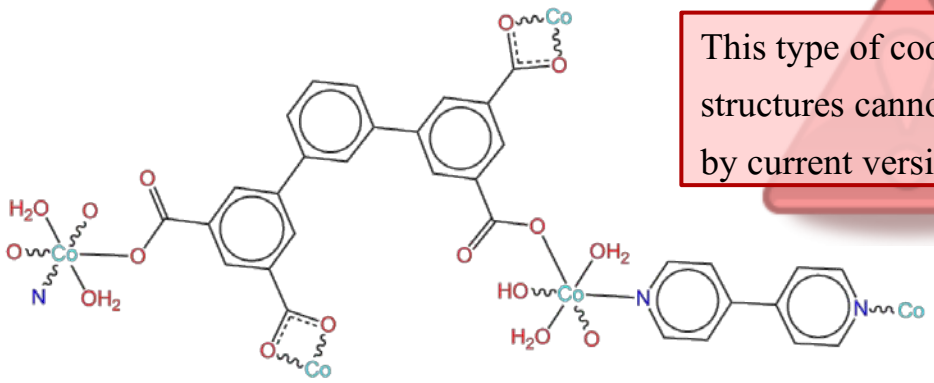
N,N'-bis(6-(6-(6-(8-(4-Isobutoxy-8-nitroquinoline-2-carboxamido)-4-isobutoxyquinoline-2-carboxamido)pyridine-2-carbamoyl)pyridine-2-carboxamido)pyridine-2-yl) 1,8-diaza-4,5-di-isobutoxy-9-methyl-2,7-anthracenedicarboxamide

Challenges - Polymers

- ‘Catena-[‘ denotes polymer containing metal, ‘Poly-[‘ denotes organic polymer – can be useful terms to search
- To name polymers we
  - break the compound down into components of ligands and metals
  - manually construct the name



catena-[bis(μ_2-benzoato)-(benzoato)-(ethane-1,2-diol)-dysprosium]



This type of coordination structures cannot be named by current version!

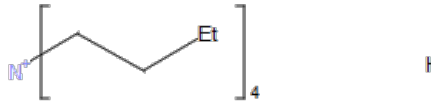
Partial name matches <<<<< ... 1,1':3',1''-terphenyl-3,3'',5,5''-tetracarboxylic acid
Partial name matches <<<<< ... 4,4'-bipyridine
Partial name matches <<<<< ... water
Partial name matches <<<<< ... ammonia

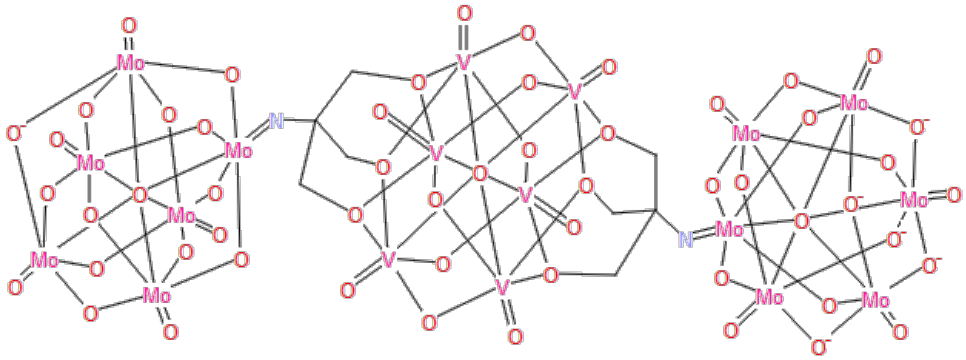
catena-[(μ_4-1,1':3',1''-terphenyl-3,3'',5,5''-tetracarboxylato)-(μ_2-4,4'-bipyridine)-tetra-aqua-di-zinc]

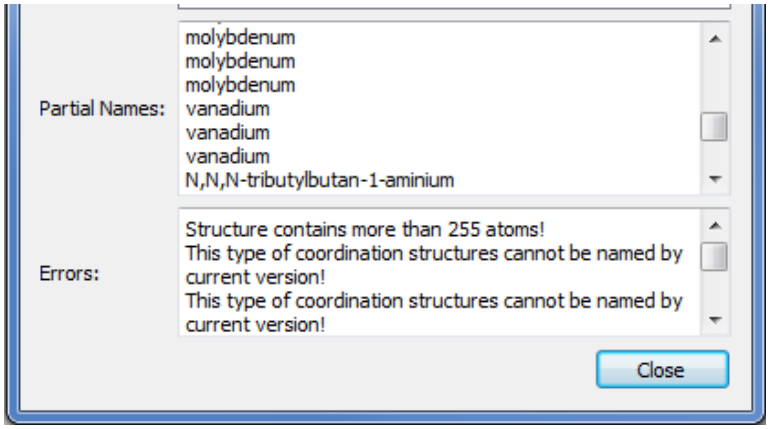
- ACD occasionally attempts to create a whole name, however it doesn’t always
  - recognise polymer bonds
  - give stoichiometry of ligands or anions

Challenges - Clusters

- For complex structures we use a residue formula, which shows the molecular formula of each ligand but not the connections.







$Mo_{12}V_6[mu_6\text{-}O]_3[mu_4\text{-}C_4H_6NO_3\text{-}kappa^1N,kappa^6O,O',O'']_2[mu_2\text{-}O]_{30}[O]_{16}^{-6}$

This type of coordination structures cannot be named by current version!

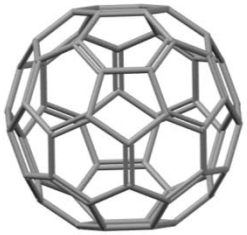
hexakis(tetra-n-butylammonium) tris(μ^6-oxido)-bis(μ^4-(tris(oxidomethyl)methyl)imino)-tricontakis(μ^2oxido)-hexadecakis(aqua)-dodeca-molybdenum-hexa-vanadium acetonitrile solvate

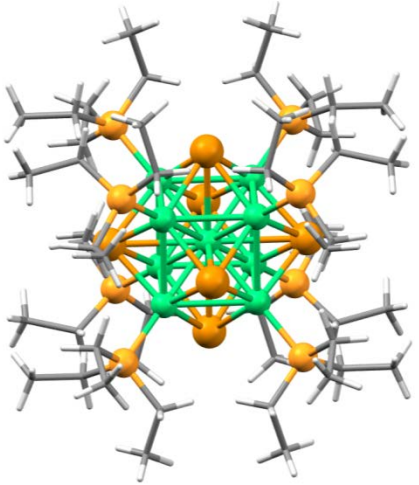
Conclusions and future improvements

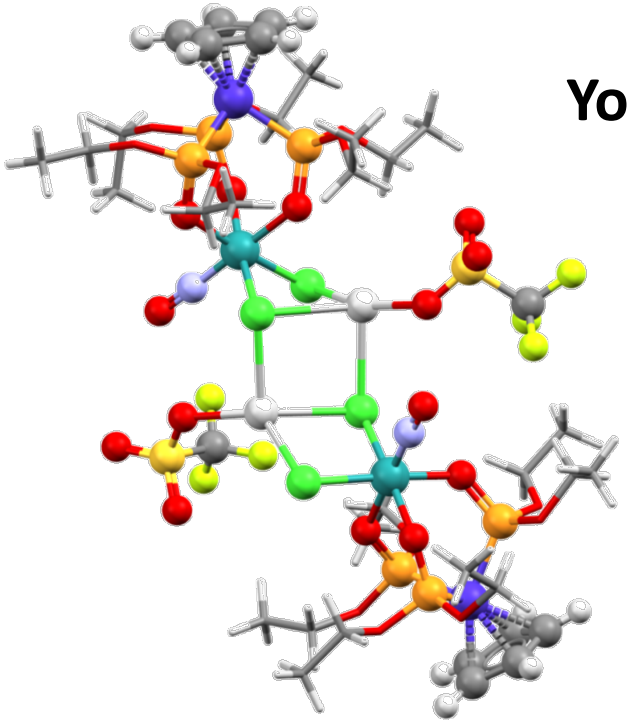
- Organics done well by current software
- Improvements in naming software could help in
  - converting between CSD and ACD bond types
  - recognising symmetry in the structure
  - atoms in unusual valencies
  - extending the size limits
  - putting together a whole name for coordination complexes
  - recognising polymer bonds and constructing a whole name with stoichiometry
- Human knowledge still needed for complex structures
  - Can we improve how we capture trivial names useful for searching large structures?
- As software improves so researchers come up with new unusual structures!

Thank you!

- CCDC staff, especially Matt, Suz, and Ian
- The 289,801 authors of crystal structures
- CICAG RSC organisers







You for listening!