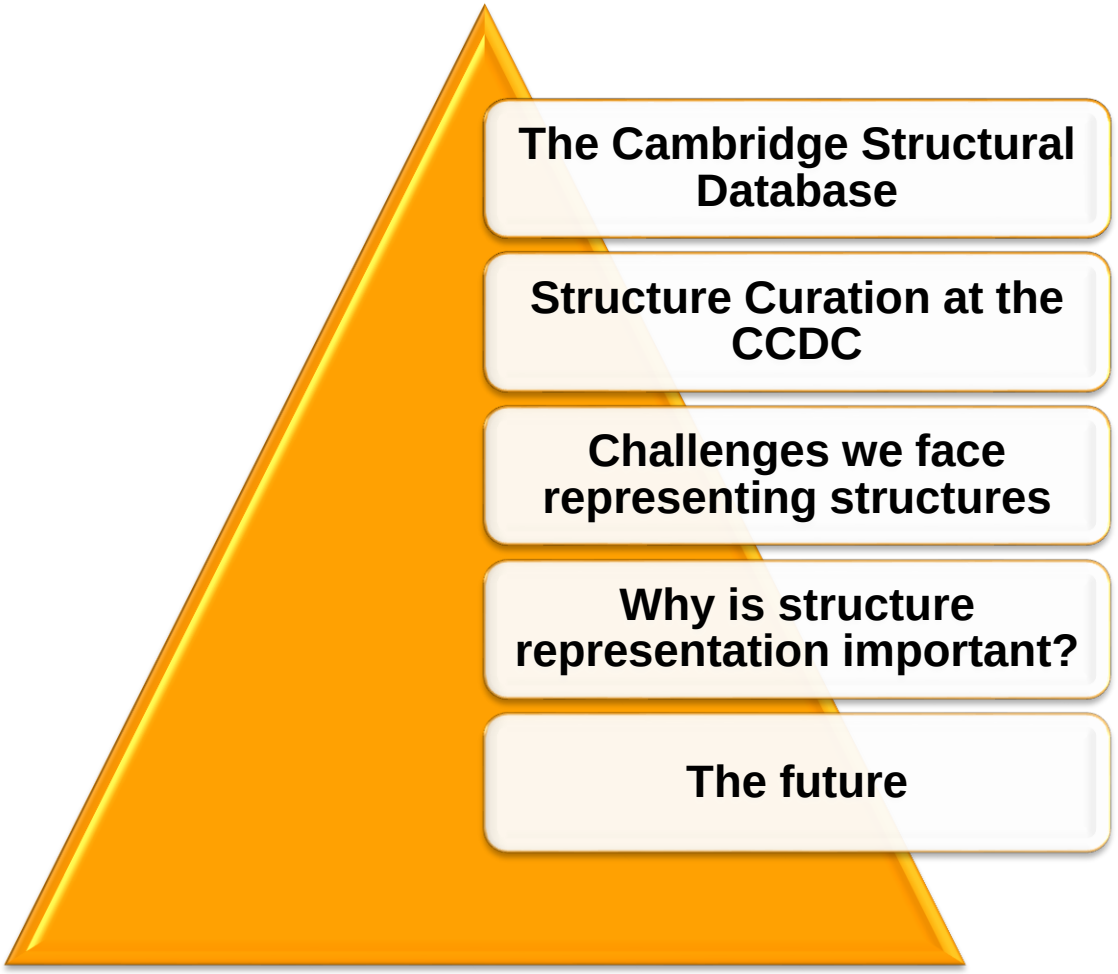
Chemical structure representation challenges encountered when curating the CSD

Matt Lightfoot, Cambridge Crystallographic Data Centre (CCDC)

RSC CICAG MEETING
CHEMICAL STRUCTURE REPRESENTATION: WHAT WOULD DALTON DO NOW?
22 JUNE 2017, UNIVERSITY OF LIVERPOOL

ROYAL SOCIETY OF CHEMISTRY | CHEMICAL INFORMATION AND COMPUTER APPLICATIONS GROUP

Chemical structure representation challenges encountered when curating the CSD



The Cambridge Crystallographic Data Centre

The Cambridge Structural Database (CSD)

A database of organic and metal-organic crystal structures

Established in 1965

~60 staff

Editors and Deposition Coordinators
Software developers
Applications scientists

Cambridge UK • Piscataway NJ

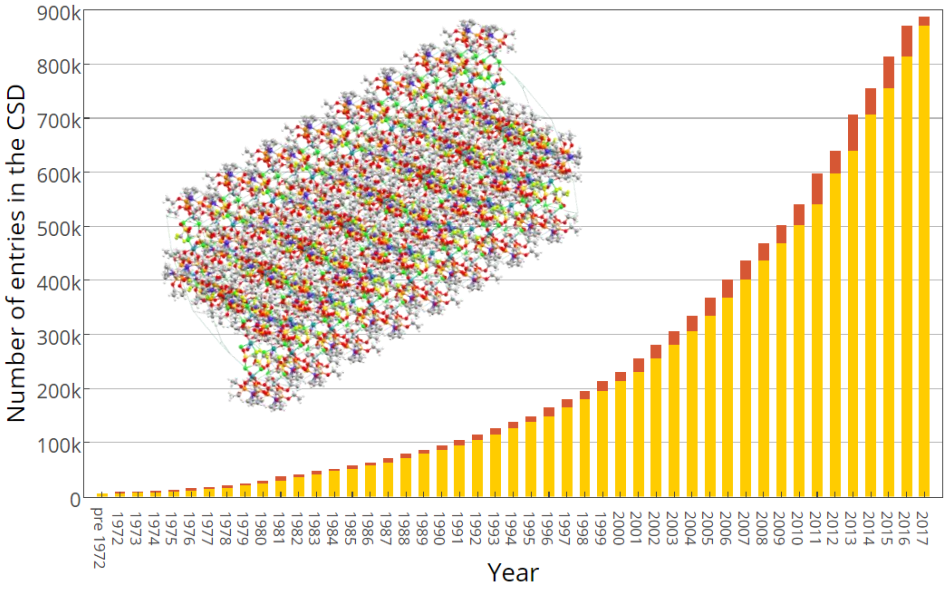
International Data Repository
Archive of crystal structure data
High quality scientific database

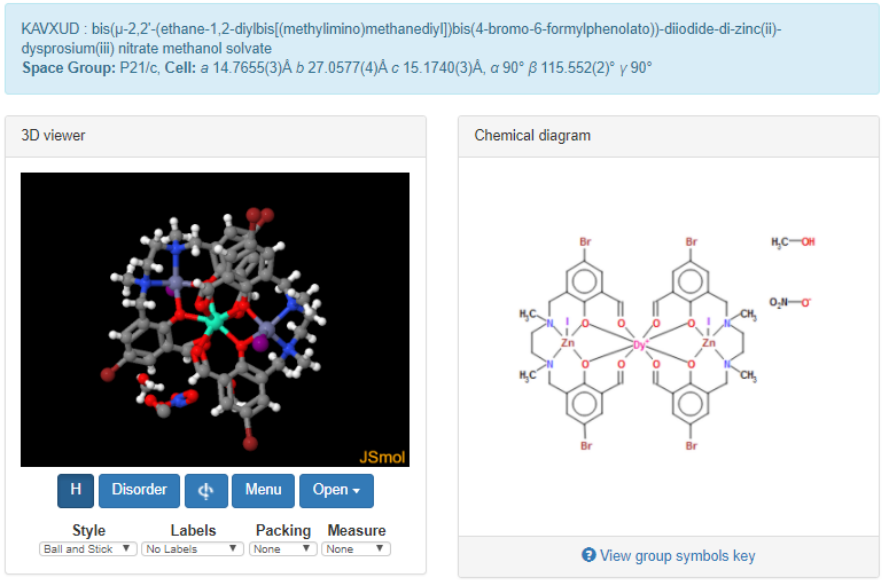
Scientific Software Provider
Search/analysis/visualisation tools
Scientific applications and APIs

Collaborative Research Organisation
New methodologies
Fundamental research

Dedicated to the advancement of chemistry and crystallography for the public benefit through providing high quality information services and software.

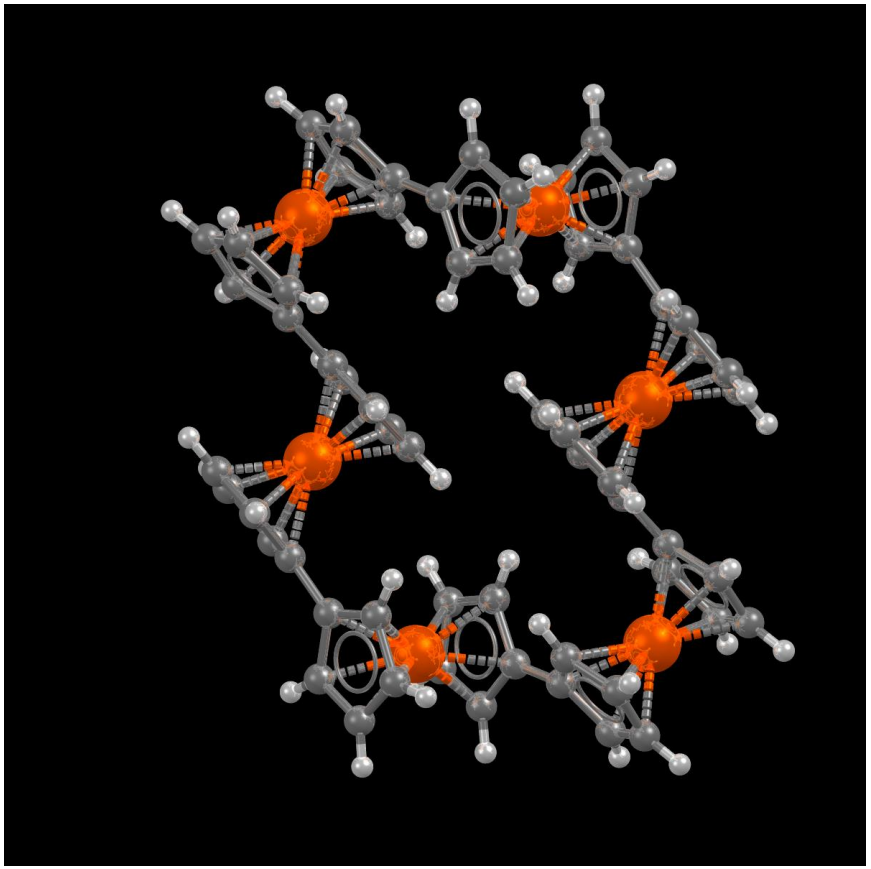
The Cambridge Structural Database





- Over 890,000 small-molecule crystal structures
- Over 60,000 datasets deposited annually
- **Enriched and annotated by experts**
- Structures available for anyone to download
- Links to over 1,000 journals

URAFUQ – A ferrocene Ferris wheel



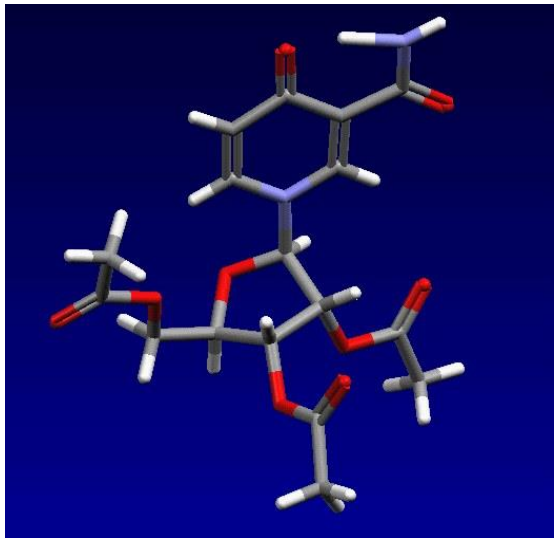
National Onion Rings Day June 22

CCDC 1420914

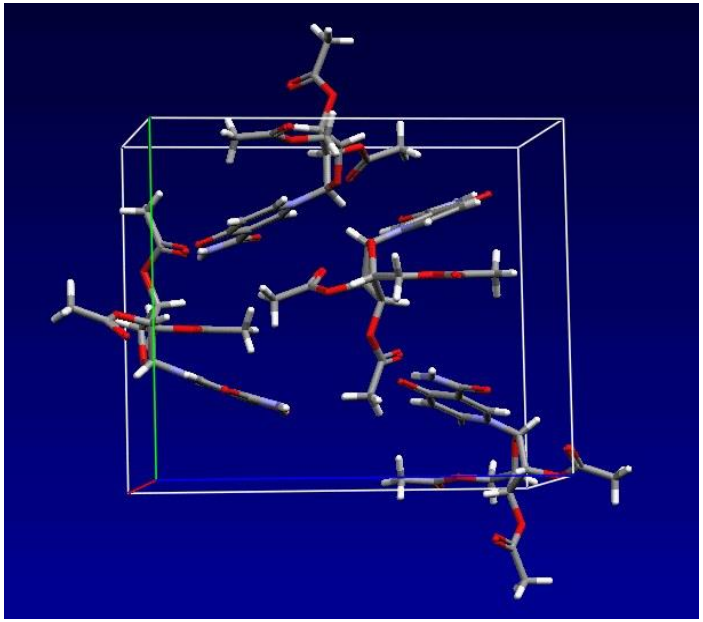
doi:10.5517/ccdc.csd.cc1jpkyq

Michael S. Inkpen, Stefan Scheerer, Michael Linseis, Andrew J. P. White, Rainer F. Winter, Tim Albrecht and Nicholas J. Long
Nature Chemistry, 2016, 8, 825, DOI: 10.1038/nchem.2553

The Cambridge Structural Database

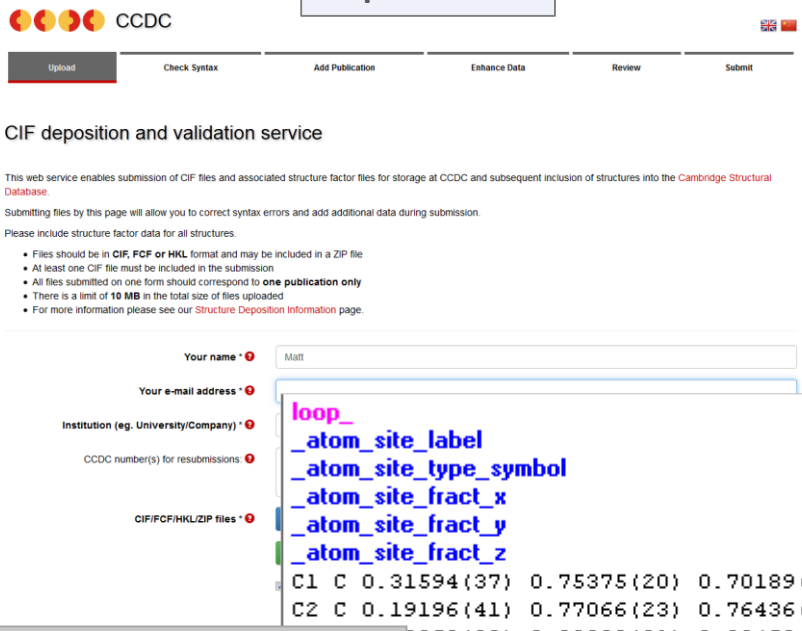


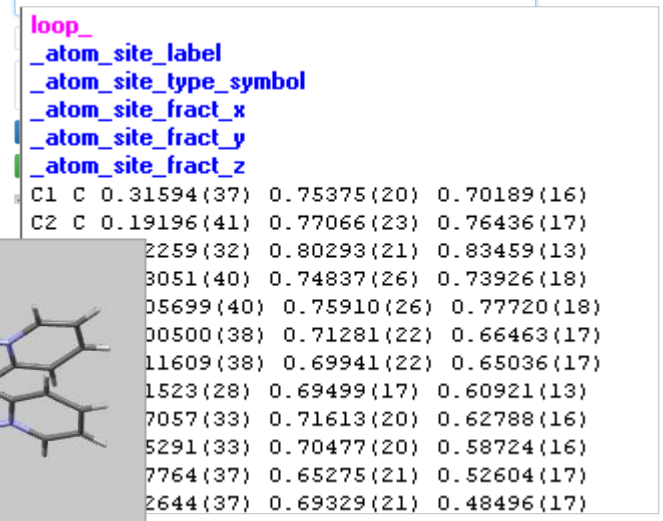
- **CSD provides insights into**
  - molecular dimensions and shape
  - molecular interactions
- **Widely used for**
  - drug design and development
  - design of new materials
  - crystal engineering
  - structure validation

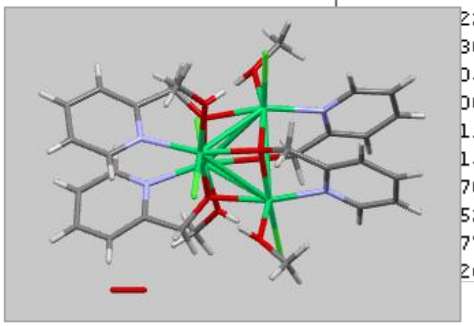


The Cambridge Structural Database

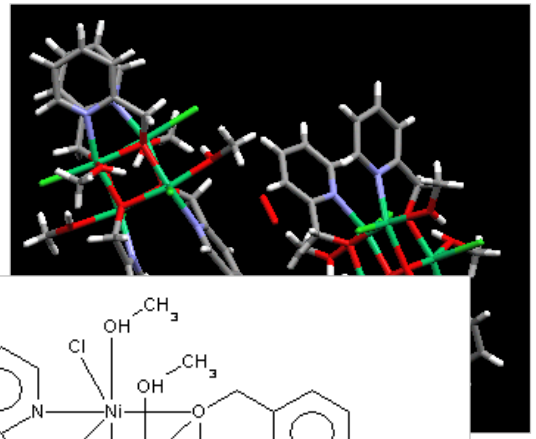
Deposited CIF

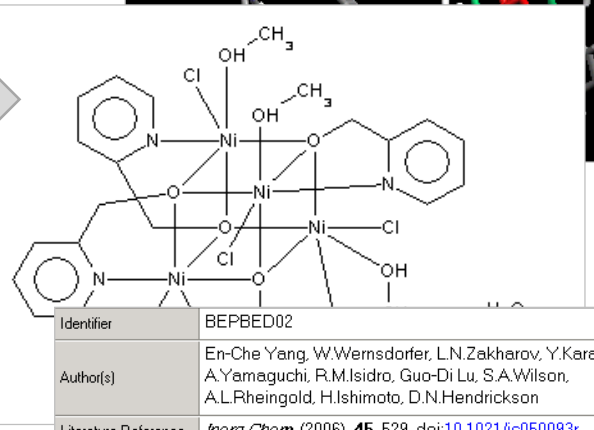






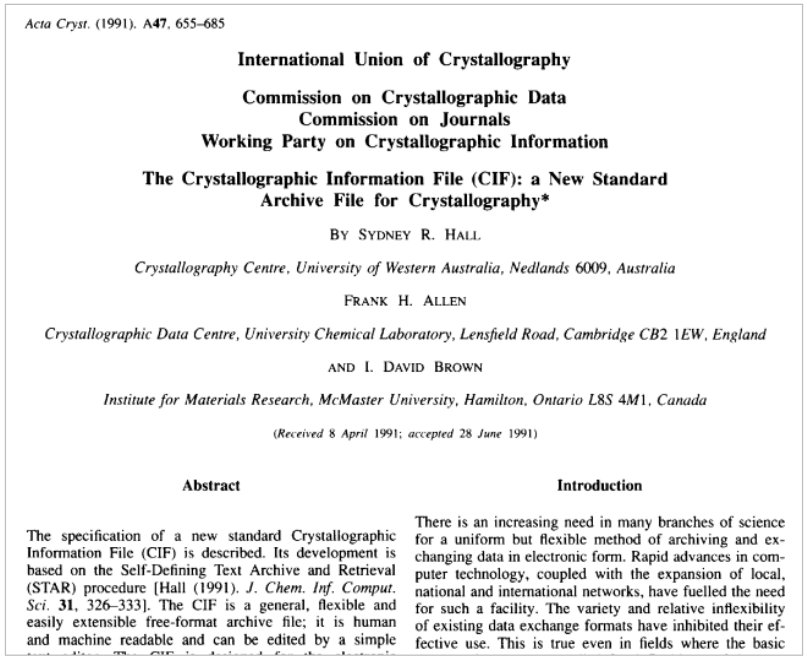
CSD Entry





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.

Community Standards: CIF format



Acta Cryst. (1991). A**47**, 655–685

International Union of Crystallography

Commission on Crystallographic Data
Commission on Journals
Working Party on Crystallographic Information

The Crystallographic Information File (CIF): a New Standard Archive File for Crystallography*

BY SYDNEY R. HALL

Crystallography Centre, University of Western Australia, Nedlands 6009, Australia

FRANK H. ALLEN

Crystallographic Data Centre, University Chemical Laboratory, Lensfield Road, Cambridge CB2 1EW, England

AND I. DAVID BROWN

Institute for Materials Research, McMaster University, Hamilton, Ontario L8S 4M1, Canada

(*Received* 8 *April* 1991; *accepted* 28 *June* 1991)

Abstract

The specification of a new standard Crystallographic Information File (CIF) is described. Its development is based on the Self-Defining Text Archive and Retrieval (STAR) procedure [Hall (1991). *J. Chem. Inf. Comput. Sci.* **31**, 326–333]. The CIF is a general, flexible and easily extensible free-format archive file; it is human and machine readable and can be edited by a simple

Introduction

There is an increasing need in many branches of science for a uniform but flexible method of archiving and exchanging data in electronic form. Rapid advances in computer technology, coupled with the expansion of local, national and international networks, have fuelled the need for such a facility. The variety and relative inflexibility of existing data exchange formats have inhibited their effective use. This is true even in fields where the basic

Syd Hall

Frank Allen

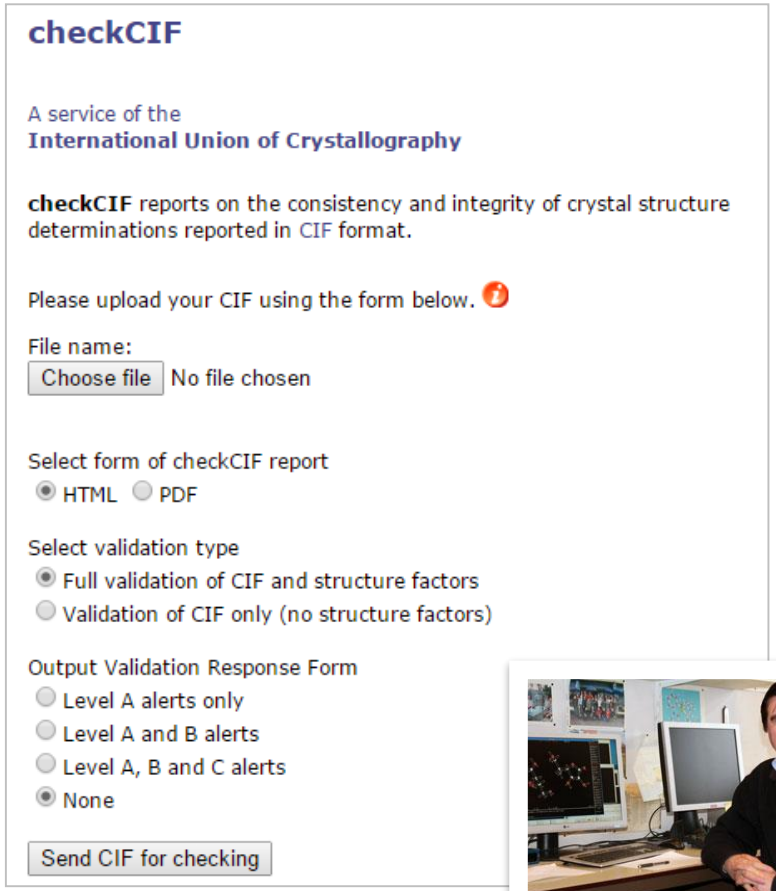
David Brown

- **A standard format for archive and exchange of crystallographic data**
  - derived results
  - raw and processed data
  - experimental conditions

```
loop_
_atom_site_label
_atom_site_type_symbol
_atom_site_fract_x
_atom_site_fract_y
_atom_site_fract_z
_atom_site_U_iso_or_equiv
_atom_site_adp_type
_atom_site_occupancy
_atom_site_symmetry_multiplicity
_atom_site_calc_flag
_atom_site_refinement_flags
_atom_site_disorder_assembly
_atom_site_disorder_group
C11 C1 0.5993(2) 1.0007(7) 0.8131(17) 0.044(3) Uani 0.50 1 d PDU A 1
S1 S 0.5321(3) 0.8260(6) 0.9322(3) 0.0327(11) Uani 0.50 1 d PDU A 1
C2 C 0.5529(4) 0.8802(9) 0.8184(9) 0.029(4) Uani 0.50 1 d PDU A 1
C3 C 0.5286(7) 0.8174(18) 0.7440(7) 0.031(4) Uani 0.50 1 d PDU A 1
H3A H 0.5350 0.8343 0.6771 0.037 Uiso 0.50 1 calc PR A 1
C4 C 0.4918(8) 0.7220(19) 0.7783(8) 0.027(4) Uani 0.50 1 d PDU A 1
C5 C 0.4900(6) 0.7171(14) 0.8779(9) 0.029(4) Uani 0.50 1 d PDU A 1
C12 C1 0.3202(2) 0.4982(6) 1.0830(5) 0.0586(15) Uani 0.50 1 d PDU A 1
S2 S 0.38755(19) 0.6658(5) 0.9578(5) 0.0400(10) Uani 0.50 1 d PDU A 1
```


Acta Cryst., **1991**, A47, 655 doi:10.1107/S010876739101067X

Community Standards: checkCIF



- **Checks consistency and integrity of the data**
- **Generates alerts that should either be corrected or explained**
- **Structure factors enable more rigorous checks**
- **API enables integration with deposition workflows**

Much of checkCIF based on components of PLATON developed by Ton Spek, Utrecht University

checkCIF is sponsored by

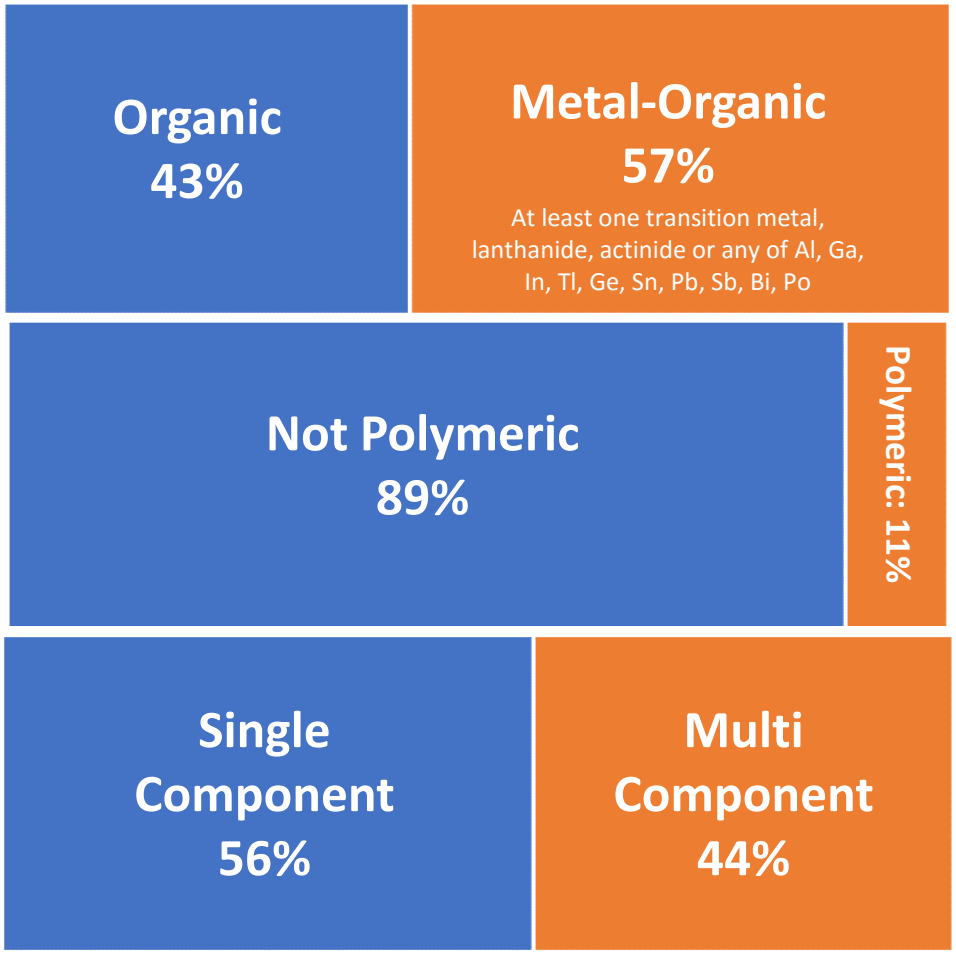
Crystallography Journals Online

ELSEVIER

WILEY

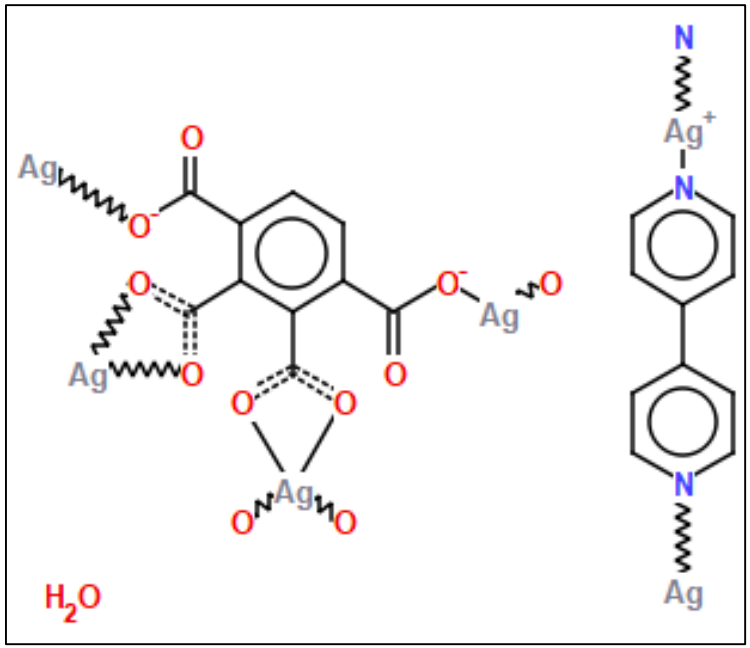
ROYAL SOCIETY OF CHEMISTRY

What's in the CSD?

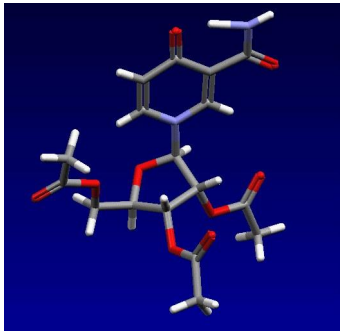


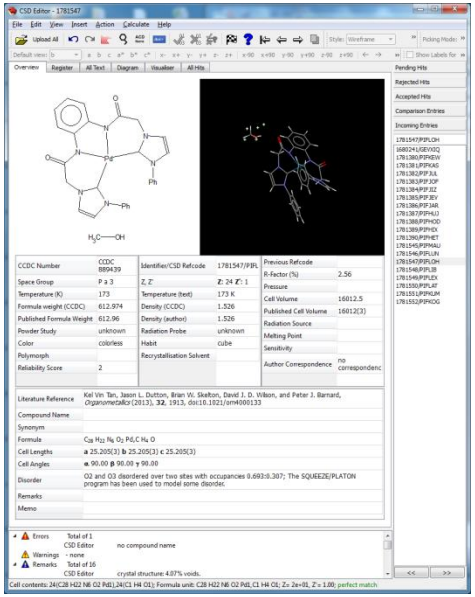
Dalton Transactions, 2011, 40, 10071, DOI: 10.1039/c1dt10830a

DACKID: CCDC 817568
doi:10.5517/ccdc.csd.cc1lwvv5



Structure Curation at CCDC

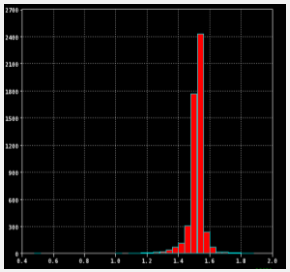




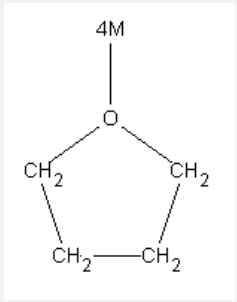
- Automatic syntax checking
- Duplication and revision checks
- **Automated enrichment – Using Decifer**
  - **Resolves disorder**
  - **For polymers, identifies monomeric unit**
  - **Deduces chemical structure from the coordinates**
  - **Assigns a reliability score**
  - **Generates a chemical diagram and chemical name**
- **Expert review and validation by editorial team**

Decifer: Automatic Assignment of Chemistry

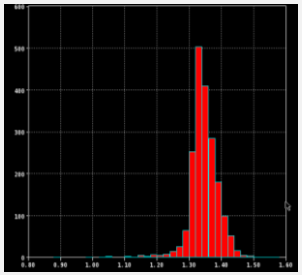
Competing Evidence



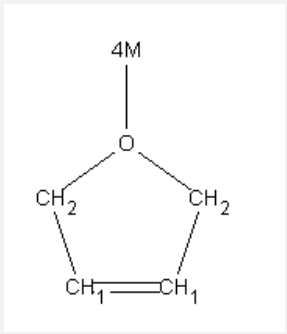
C-C, prob = 0.004



7208 hits



C=C, prob = 0.20



2 hits

Bayes' Theorem

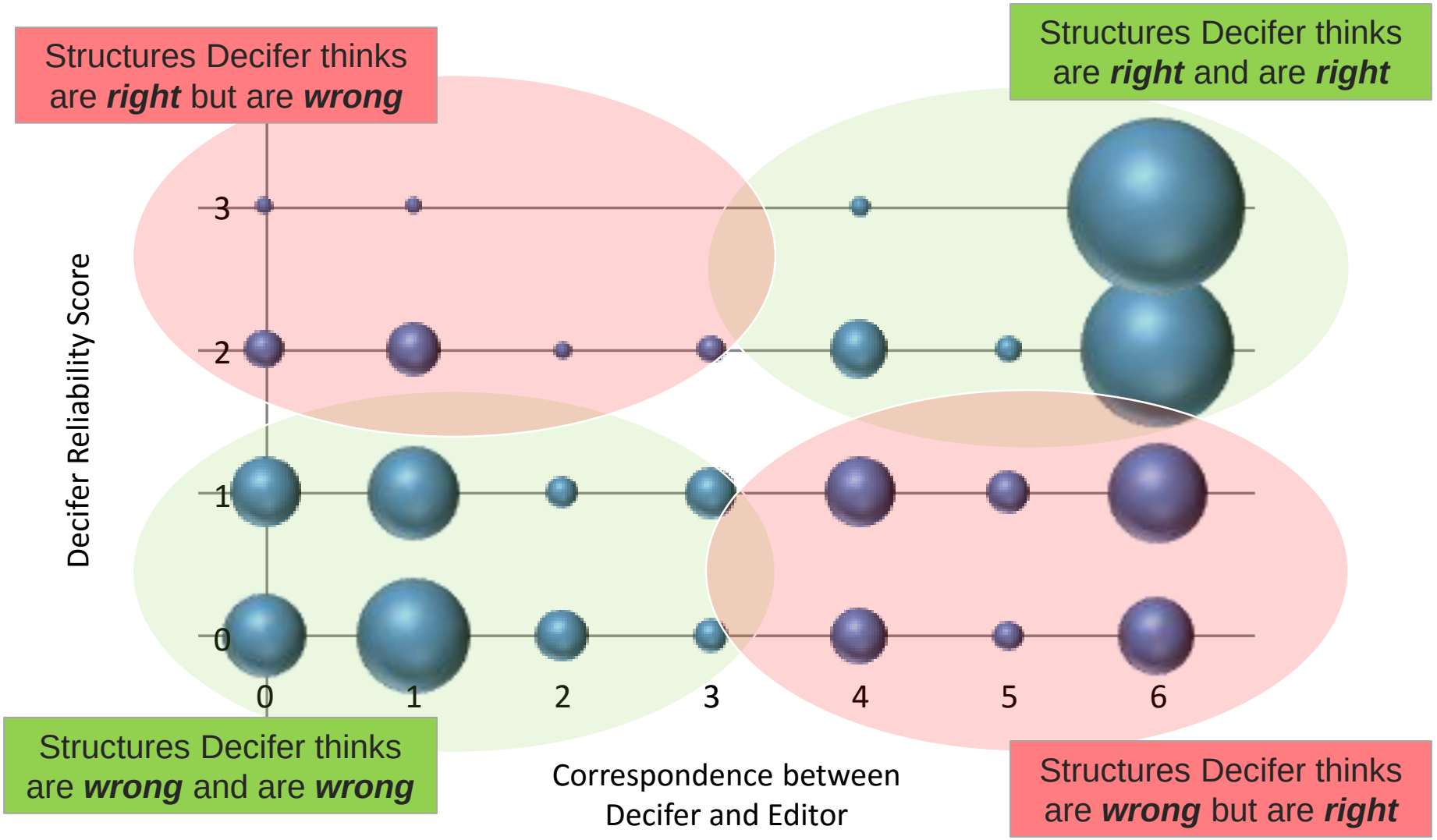
$$P(A|B) = \frac{P(B|A)\,P(A)}{P(B)}.$$

Chemical Assignment + Reliability Report

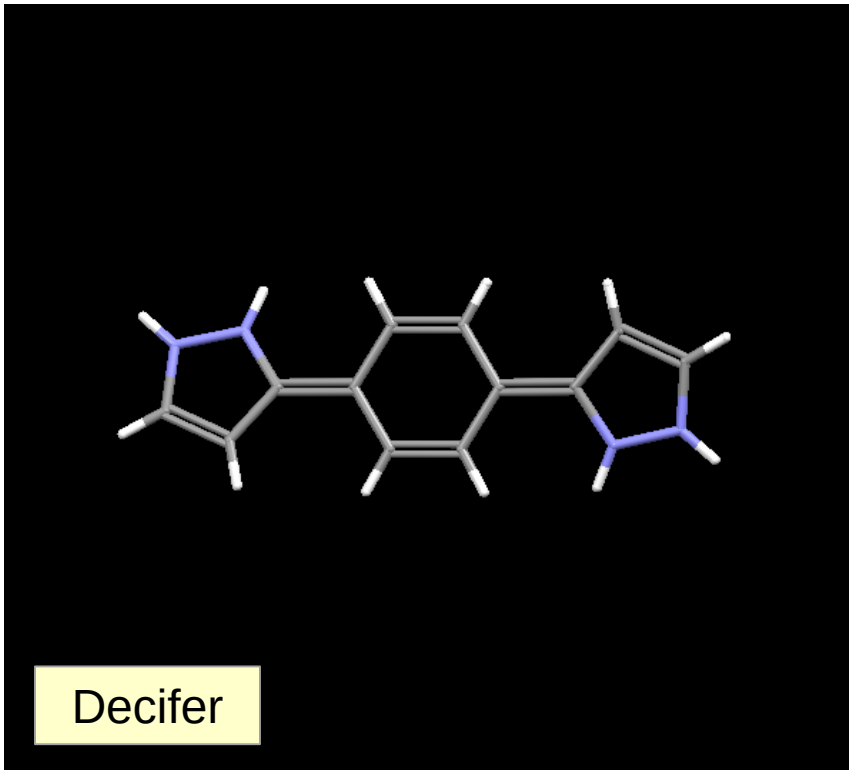
```
Low probability bond lengths:
C5-C6   1.405, av(CSD) = 1.505, prob = 0.001
C2-C3   1.345, av(CSD) = 1.514, prob = 0.001
C3-C4   1.338, av(CSD) = 1.514, prob = 0.001
C3-C6   1.798, av(CSD) = 1.546, prob = 0.001
Reliability level: 2
```

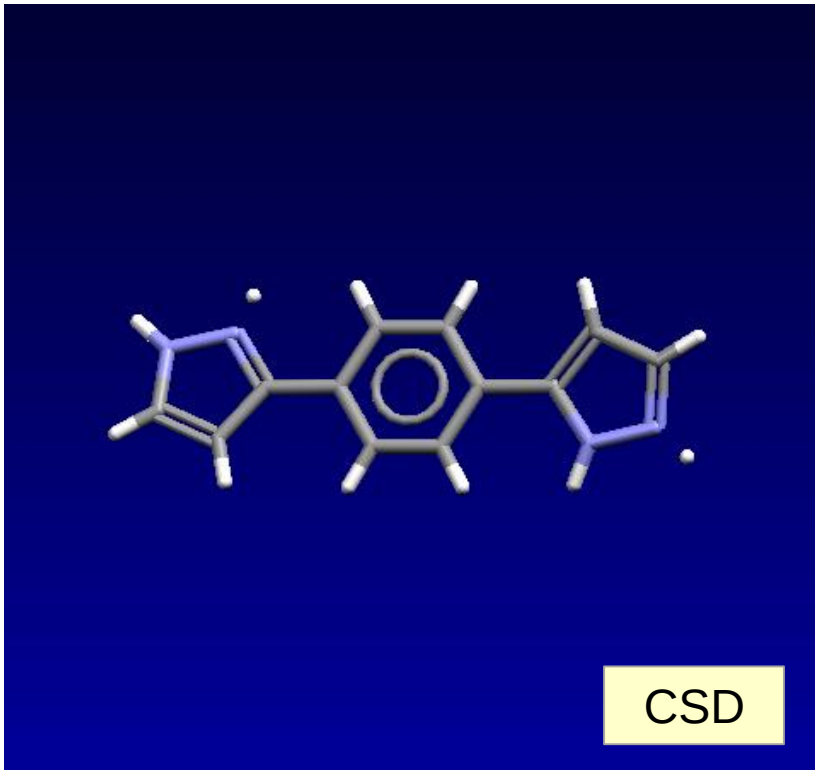

Bruno, Shields and Taylor (2011) *Acta Cryst.*, **B67**, 333-349 10.1107/S0108768111024608

Decifer Validation



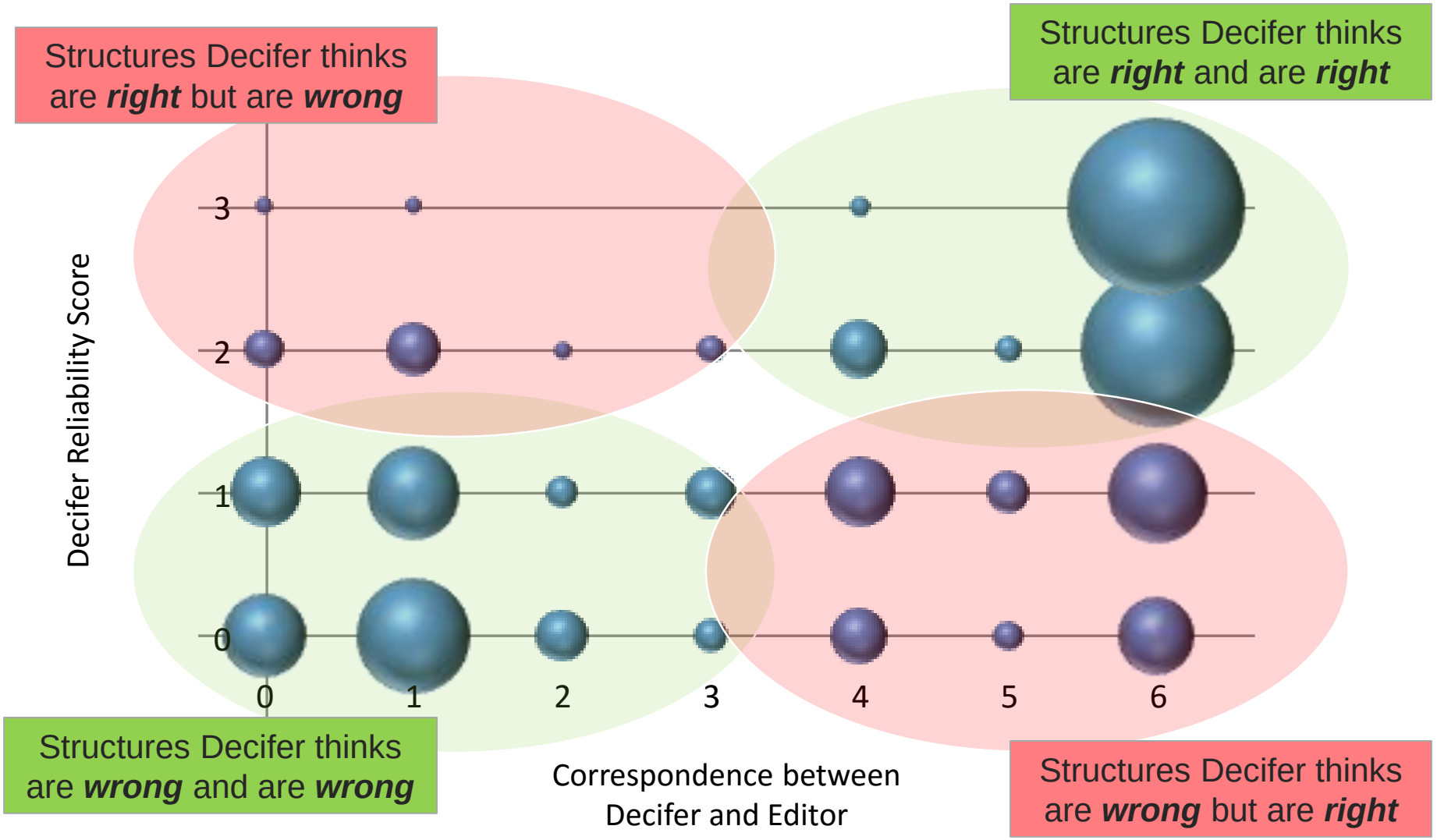
Where Decifer is wrong about being right



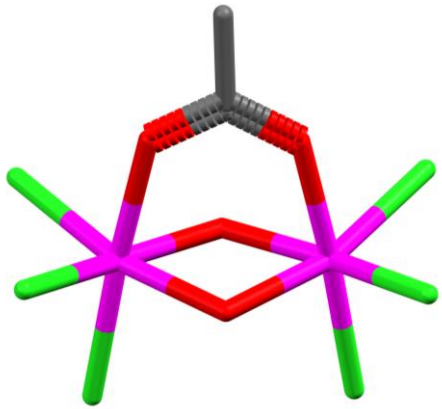


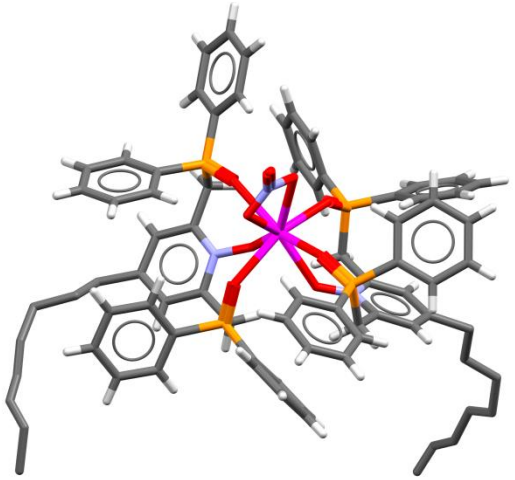
H atom disorder but Decifer does not recognise structure as being disordered

Decifer Validation



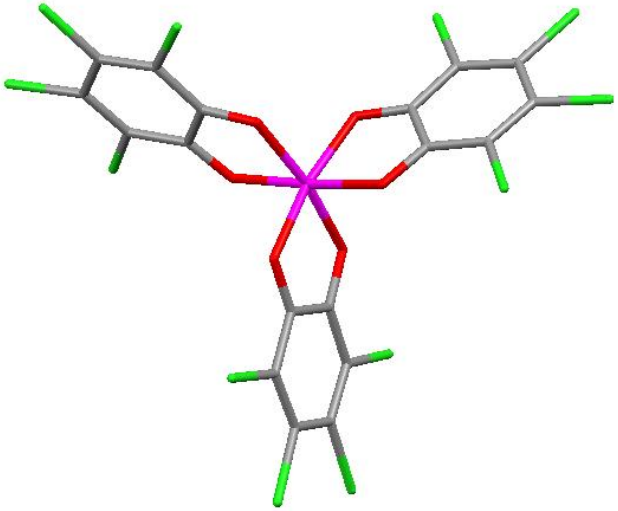
Challenges

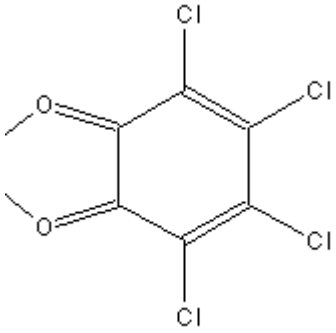


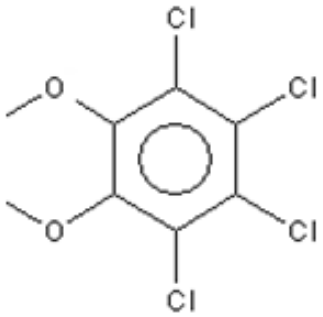


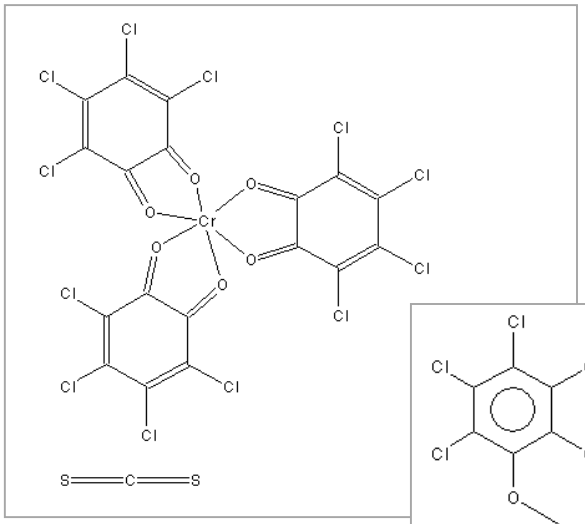
- **Missing H atoms**
  - one bridging oxygen is hydroxy
  - the other has no H atoms
- **Poor geometry**
  - carbon chains include C-C bonds around 1.3 A
  - suggests double but are in fact single

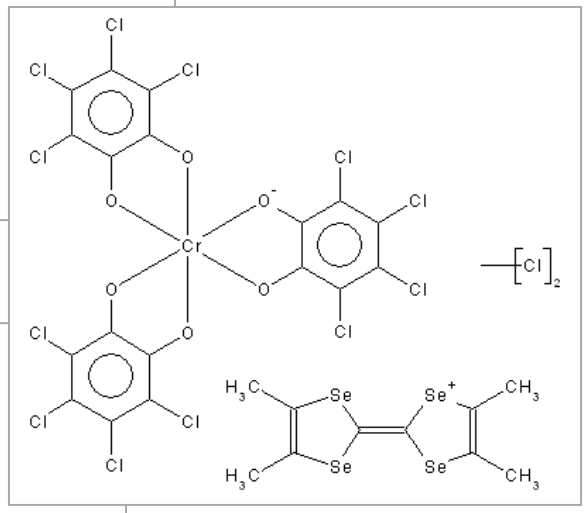
Challenges: Oxidation State

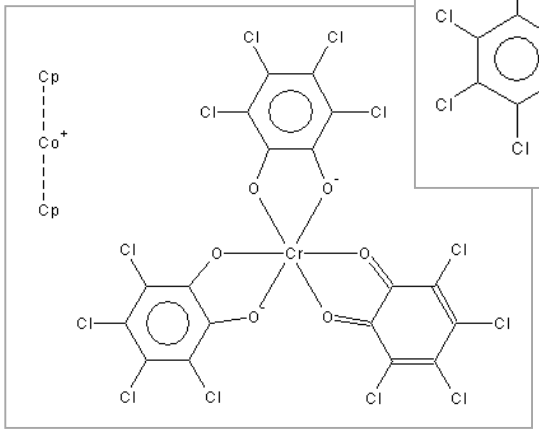










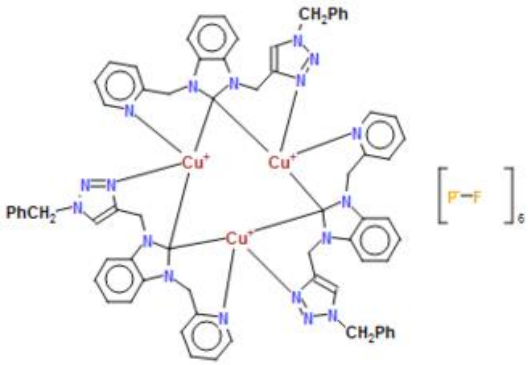


CBZQCR DOXQAI GUNFOK

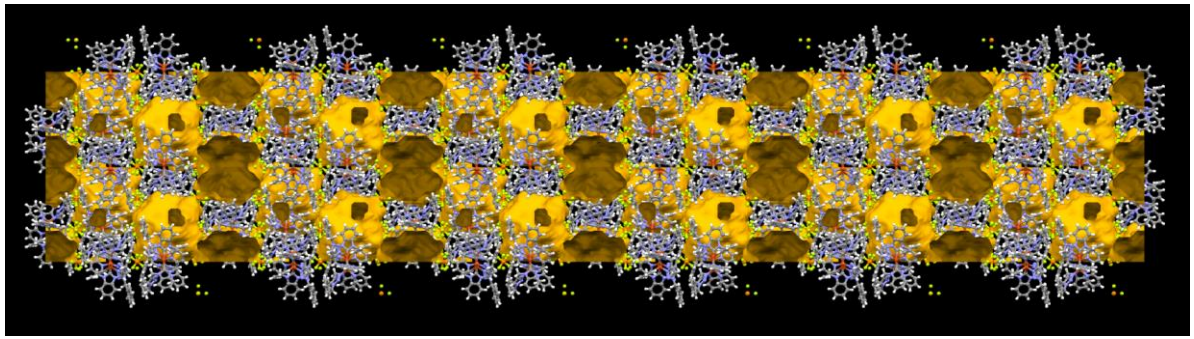
Challenges - Crystallographic Disorder

IBECAW : CCDC 1424017 doi:10.5517/ccdc.csd.cc1jst15

tris(μ-1-((1-benzyl-1H-1,2,3-triazol-4-yl)methyl)-3-((pyridin-2-yl)methyl)-1,3-dihydro-2H-benzimidazol-2-ylidene)-tri-copper **tris(hexafluorophosphate)** **unknown solvate**



The structure is disordered. *The SQUEEZE/PLATON program has been used to model some disorder. The unit cell was found to contain a* ***void of 9116 A^3 holding 3767 electrons***.

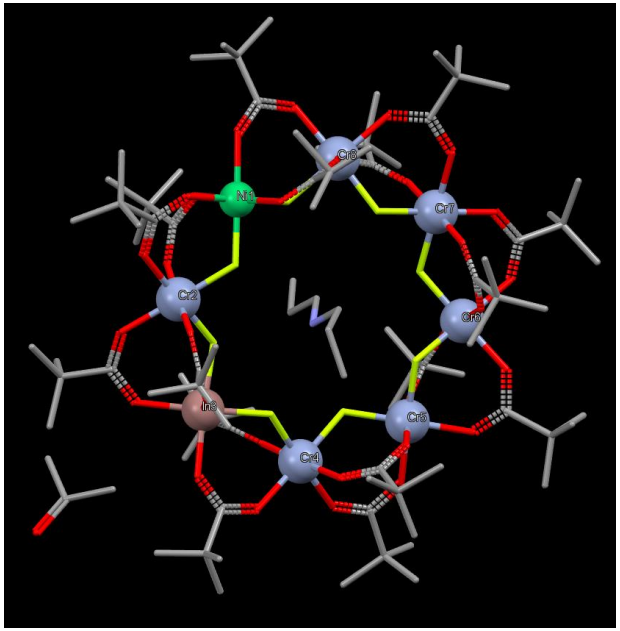


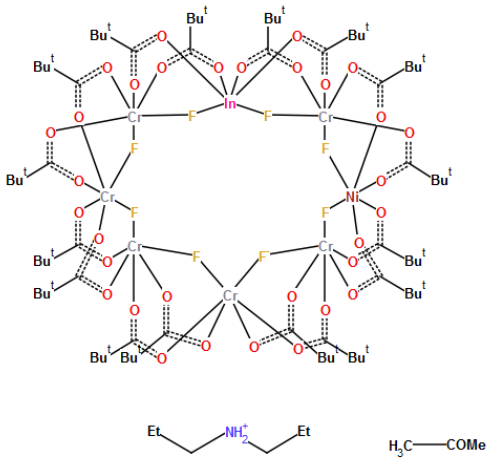
Challenges - Mixed Metal Disorder

GUJYIU: CCDC 705752 doi:10.5517/ccrpd60

Dipropylammonium octakis(m2-fluoro)-hexadecakis(m2-pivalato)-hexa-chromium(iii)-indium(iii)-nickel(ii) acetone solvate

$C_{80}H_{144}Cr_6F_8InNiO_{32}^-$, $C_6H_{16}N^+$, $2(C_3H_6O)$





The metal sites have mixed occupancy Cr:Ni:In.

Each site has been refined with 0.75 Cr, 0.13 Ni,0.13 In

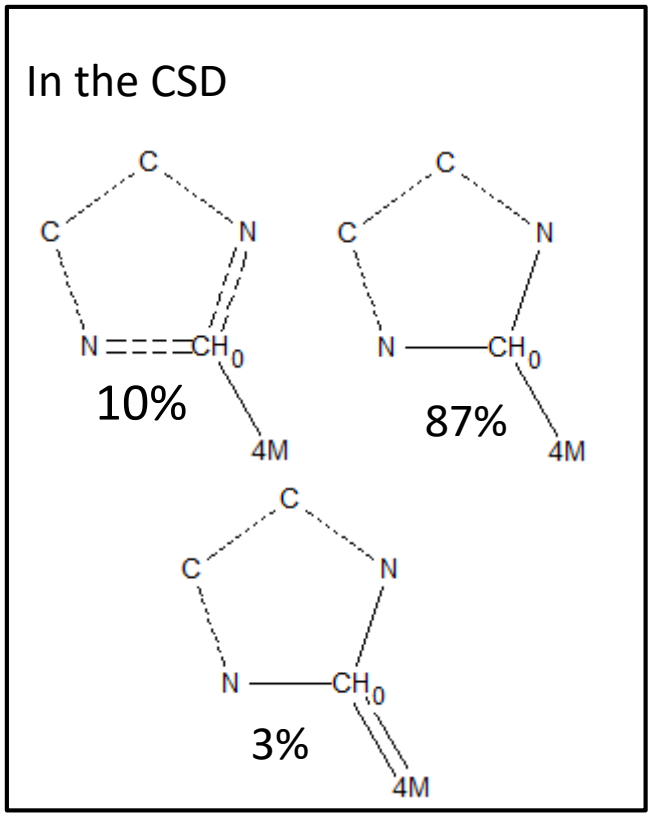
For reasons of charge balance the CSD entry has one of the mixed metal sites assigned as In and one as Ni

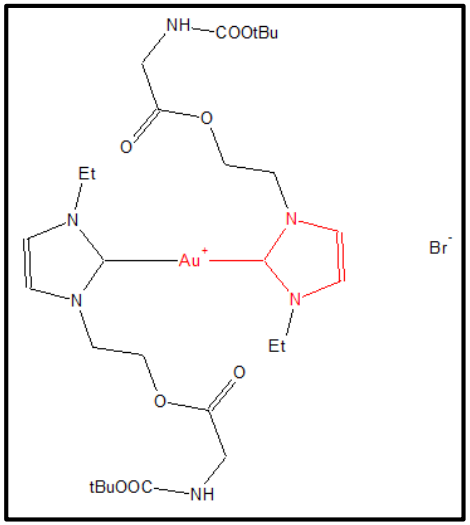
Chemical representation differences

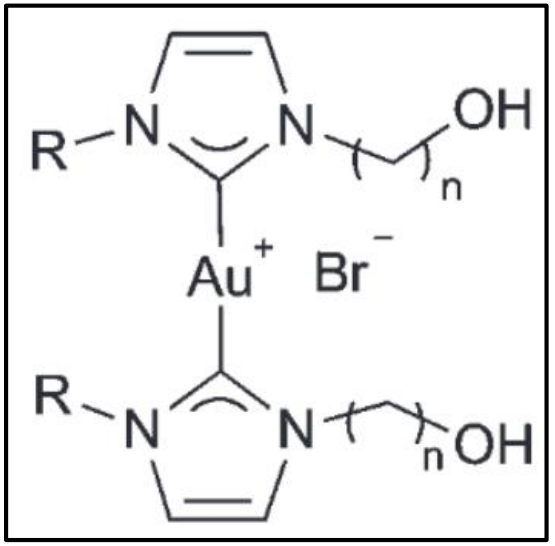
ZAPTOC: CCDC 1520090 doi:10.5517/ccdc.csd.cc1n0s5m

bis(1-(2-((((t-butoxycarbonyl)amino)acetyl)oxy)ethyl)-3-ethylimidazol-2-ylidene)-gold bromide

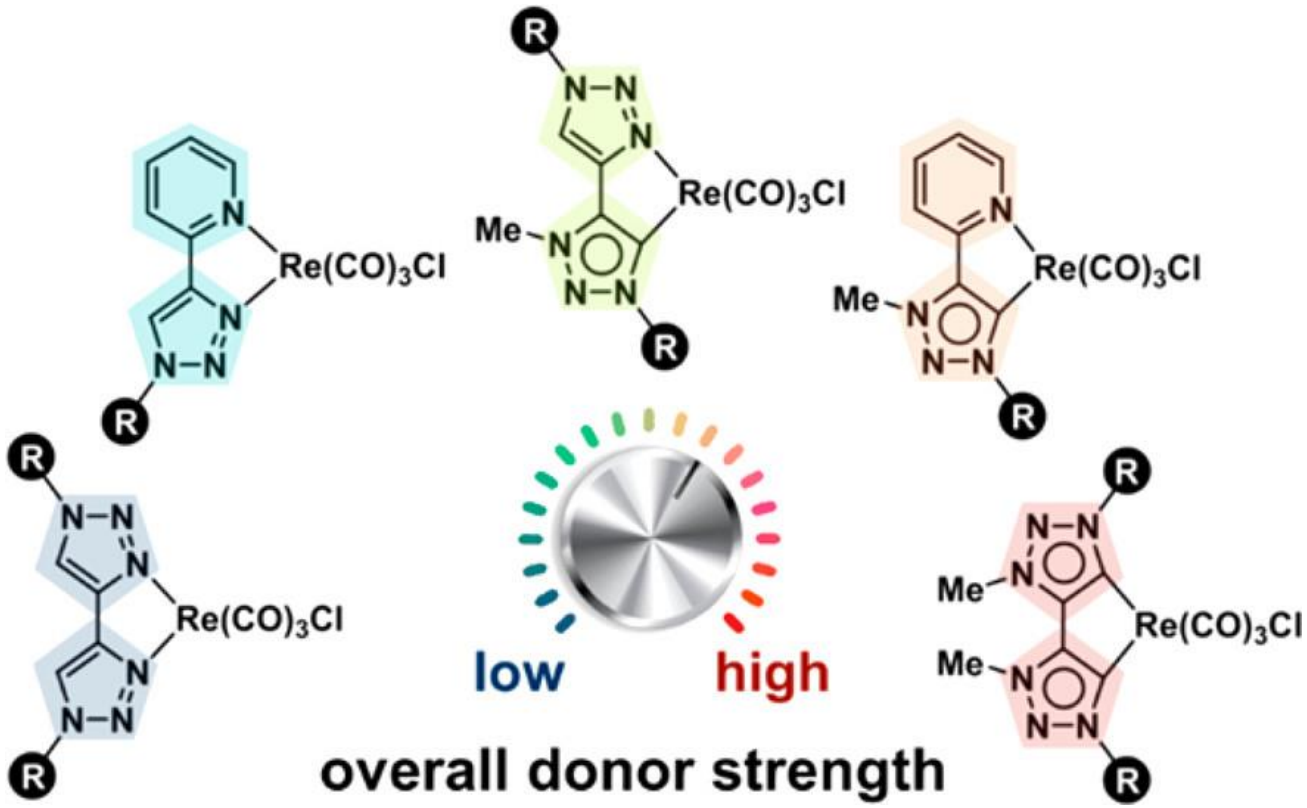
Dalton Transactions, 2017, 46, 2988, DOI: 10.1039/C6DT04834J







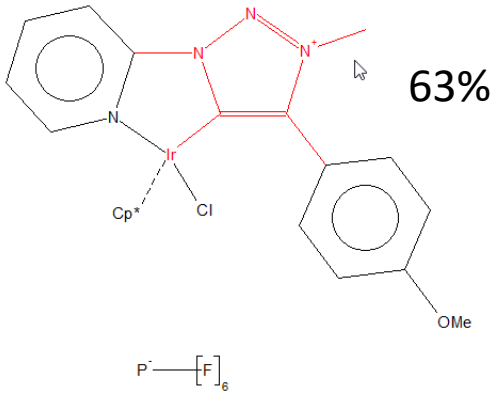
Chemical representation differences



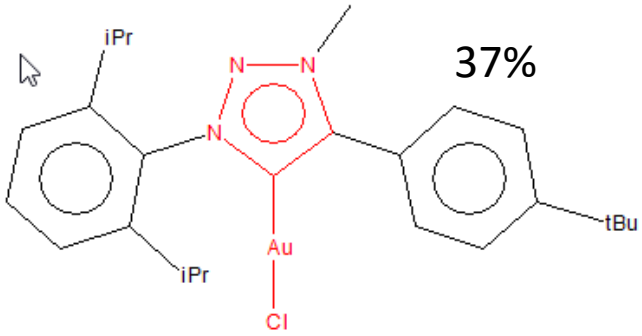
Published in: Lisa Suntrup; Sinja Klenk; Johannes Klein; Sebastian Sobottka; Biprajit Sarkar; *Inorg. Chem.* **2017,** 56, 5771-5783.
DOI: 10.1021/acs.inorgchem.7b00393

Chemical representation differences

In the CSD



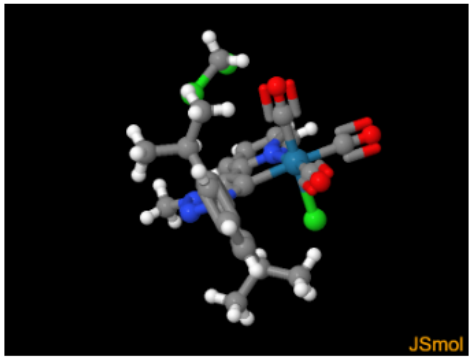
63%



37%

CATLAN : tricarbonyl-chloro-{1-[2,6-di-isopropylphenyl]-3-methyl-4-(pyridin-2-yl)-1H-1,2,3-triazol-3-ium-5-ylidene}-rhenium dichloromethane solvate
Space Group: P-1, **Cell:** *a* 9.342(5)Å *b* 10.520(5)Å *c* 15.983(5)Å, *α* 90.251(5)° *β* 105.580(5)° *γ* 115.435(5)°

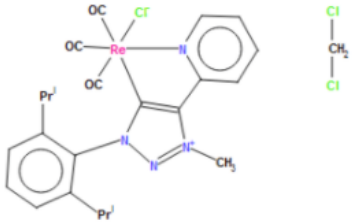
3D viewer



H | Disorder | Menu | Open

Style: Ball and Stick | Labels: No Labels | Packing: None | Measure: None

Chemical diagram



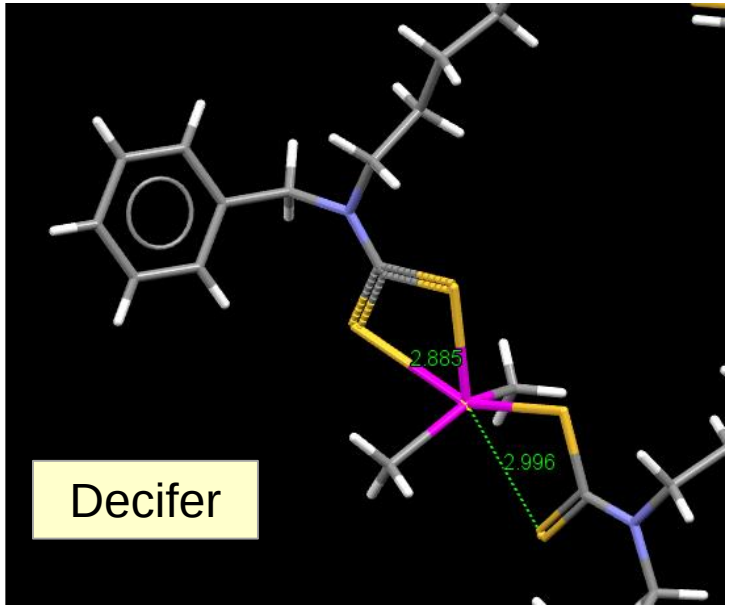
View group symbols key

Additional CCDC details

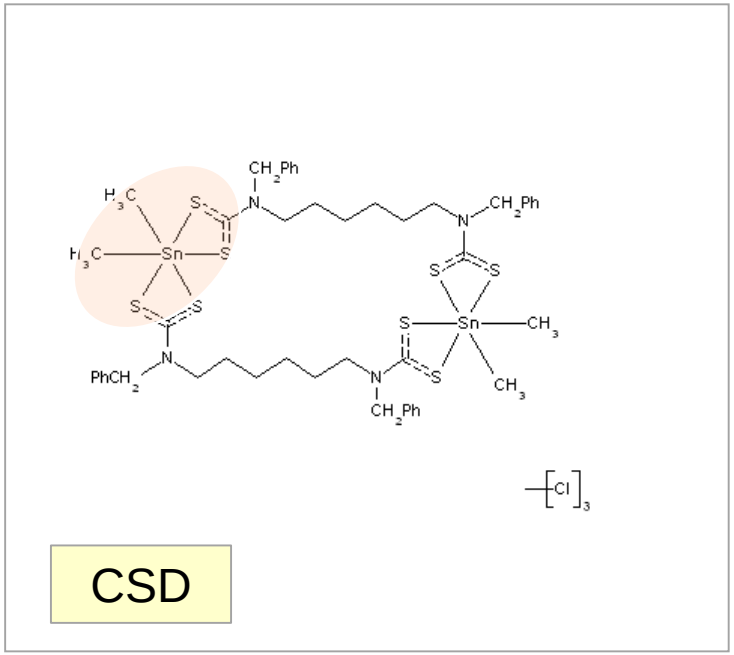
CCDC Number	1442848
CCDC Citation	Lisa Suntrup, Sinja Klenk, Johannes Klein, Sebastian Sobottka, Biprajit Sarkar CCDC 1442848: Experimental Crystal Structure Determination, 2017, DOI: 10.5517/ccdc.csd.cc1kfdhw
Deposited on	17/03/2017

Associated publications

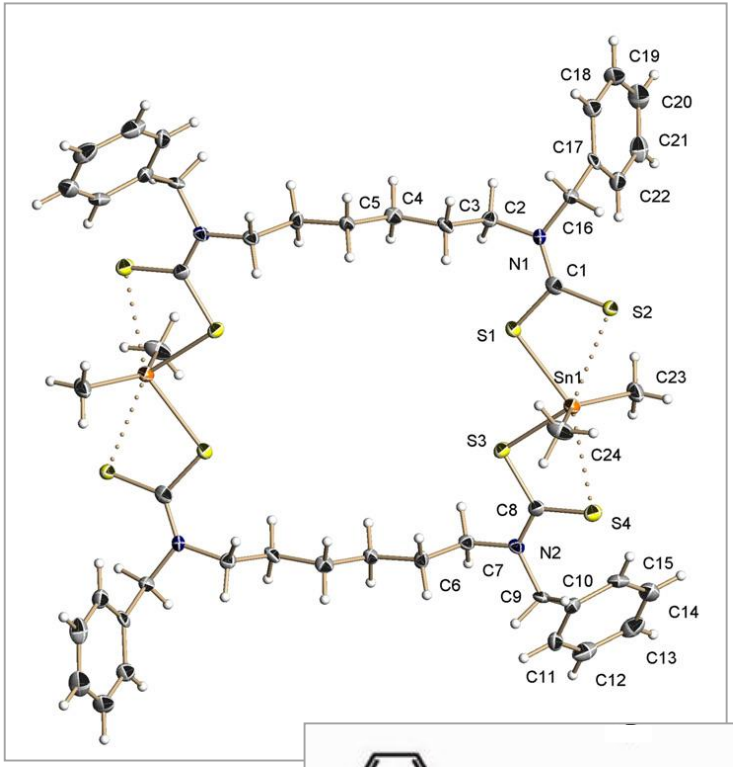
Lisa Suntrup, Sinja Klenk, Johannes Klein, Sebastian Sobottka, Biprajit Sarkar, *Inorganic Chemistry*, 2017, 56, 5771, DOI: 10.1021/acs.inorgchem.7b00393



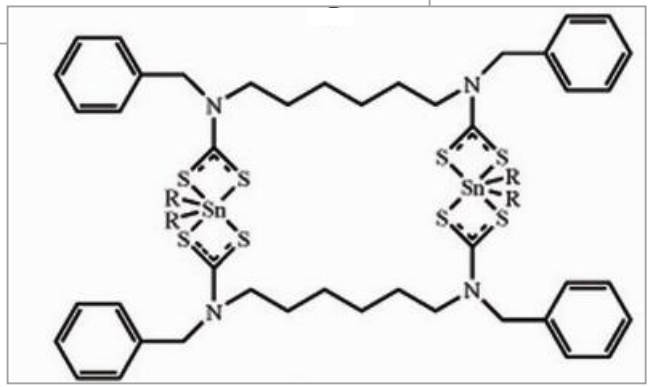
Decifer



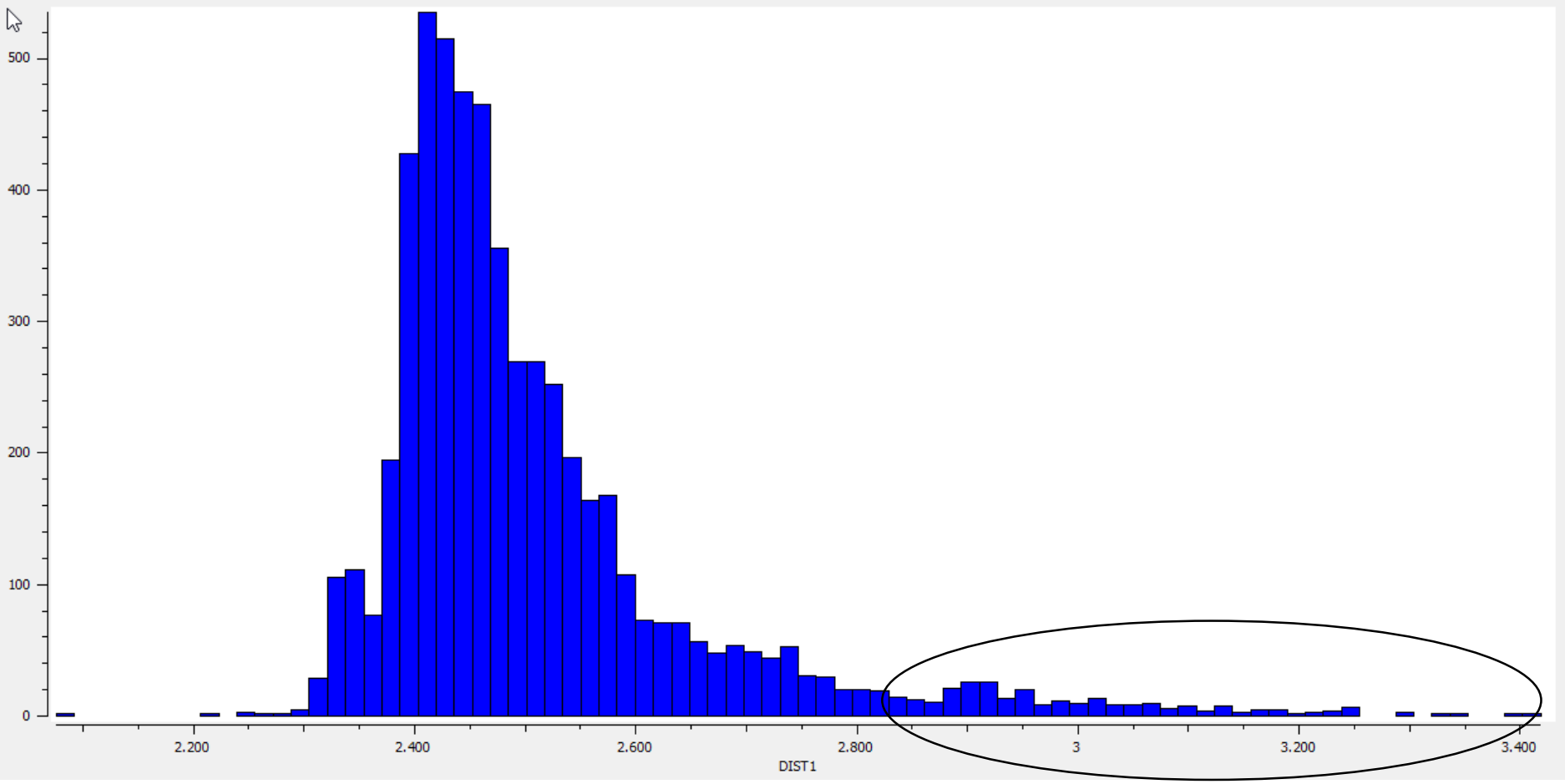
CSD



Paper



Sn-S bond lengths in the CSD

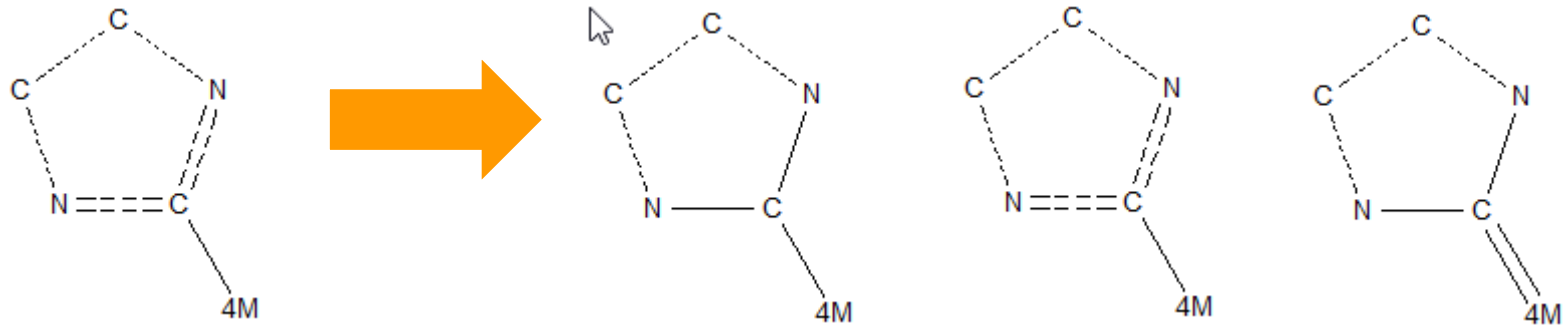


Chemical representation differences

- CSD contains structures **added over 50 years**
- The **representation** of certain types of structures in the literature and the CSD has **changed over this period**
- **Chemistry given by authors is not always consistent** within publications and between publications
- We now have a **more automated approach to assigning chemistry**
- Some structures are **represented in a standard way** in the CSD
- But if structures are represented in various ways in the literature then they are probably **represented with the same differences in the CSD**
- Generally CSD users know of these inconsistencies and **allow for these in the way that they search**
- But this can cause **problems for new and inexperienced users**

Chemical representation differences

- **What could we do to help this?**
- Provide an additional standard chemical representation for all structures?
  - Not a simple process
  - Not always clear what the standard should be
- Improve our searching algorithms to find all representations?



CIF, MIF and Chemistry Standards

- **CIF has data items for a 2D chemical representation**
  - these are rarely used in deposited data files
  - identified 3 uses in ~480,000 deposited files (2015)
- **MIF: proposed standard similar to CIF for chemical data**

```
_chemical_conn_atom_[]
    _chemical_conn_atom_charge
    _chemical_conn_atom_display_x
    _chemical_conn_atom_display_y
    _chemical_conn_atom_NCA
    _chemical_conn_atom_NH
    _chemical_conn_atom_number
    _chemical_conn_atom_type_symbol
_chemical_conn_bond_[]
    _chemical_conn_bond_atom_1
    _chemical_conn_bond_atom_2
    _chemical_conn_bond_type
```

The Molecular Information File (MIF): Core Specifications of a New Standard Format for Chemical Data. Frank H. Allen, John M. Barnard, Anthony P. F. Cook, Sydney R. Hall (1995) *J. Chem. Inf. Comput. Sci.,* 35 (3), 412–427. 10.1021/ci00025a009

- **Would this help?**

De facto Chemistry Standards

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

MOL V3000

```
1 = single
2 = double
3 = triple
9 = coordination
10 = hydrogen
```

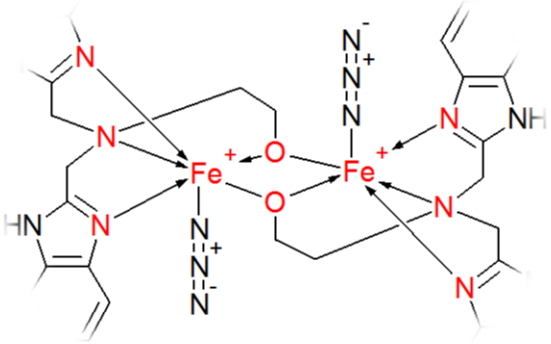
* excluding query bond types

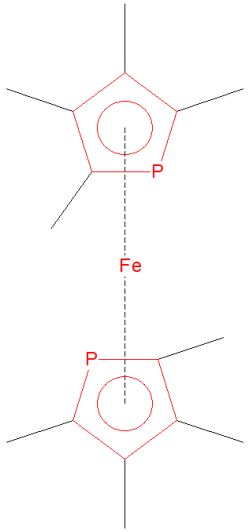
PubChem SDF
PUBCHEM_NONSTANDARDBOND

```
1 Single Bond
2 Double Bond
3 Triple Bond
4 Quadruple Bond
5 Dative Bond
6 Complex Bond
7 Ionic Bond
```

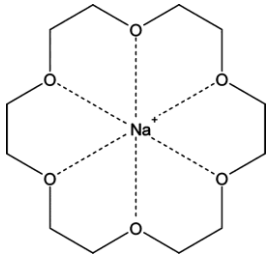
ACD/Labs MOL V2000 Extensions

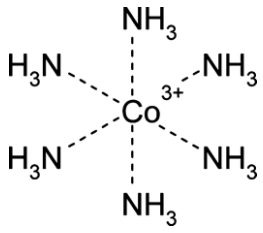
```
M  ZZF  3   1  41   2  42   3  43
M  ZZH   1  5   2   3   4   5   6
M  ZZH   2  5   7   8   9  10  11
M  ZZH   3  5  12  14  15  16  17
M  ZZE  2  42  18  43  18
```

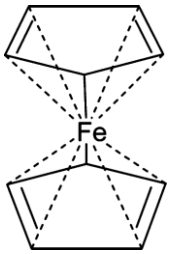




Accurate Specification of Molecular Structures: The Case for Zero-Order Bonds and Explicit Hydrogen Counting. **Alex M. Clark.**
J. Chem. Inf. Model., 2011, 51 (12), 3149. doi:10.1021/ci200488k

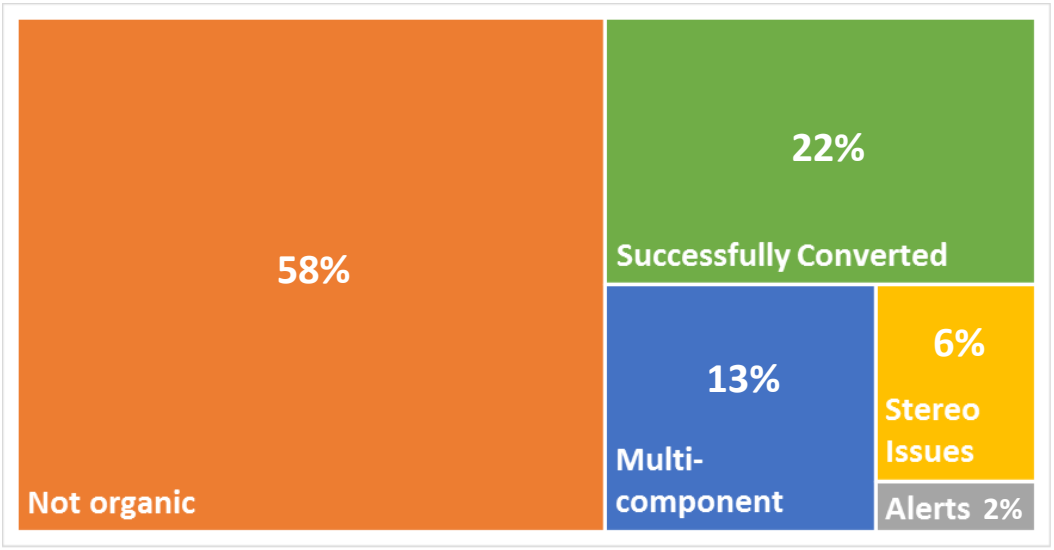






Impact of representation challenges

- **GOAL:** Reliable standard InChI representations to enable intersection of the CSD with e.g. ChemSpider and PubChem



Order of filtering:
- Not organic
- Multi-component
- Inchi alerts
- Stereochemistry Issues

Based on a subset of 495,751 entries from CSD V5.36

InChIs successfully generated for 108,570 entries

- Can confidently generate InChIs for ~22% of CSD entries
- If multi-component entries included then ~35% assuming no other issues

Shared Challenges

- **Current InChI Working Group Discussions**
  - Organometallics
  - Mixtures
  - Polymers
  - Stereochemistry
  - Tautomers

DATE AND TIME

Wed, Aug 16, 2017, 8:30 AM – Fri, Aug 18, 2017, 12:30 PM EDT

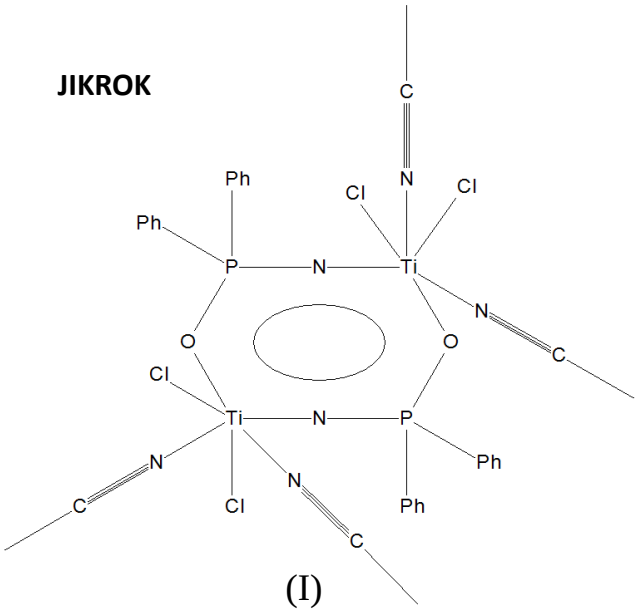
Add to Calendar

LOCATION

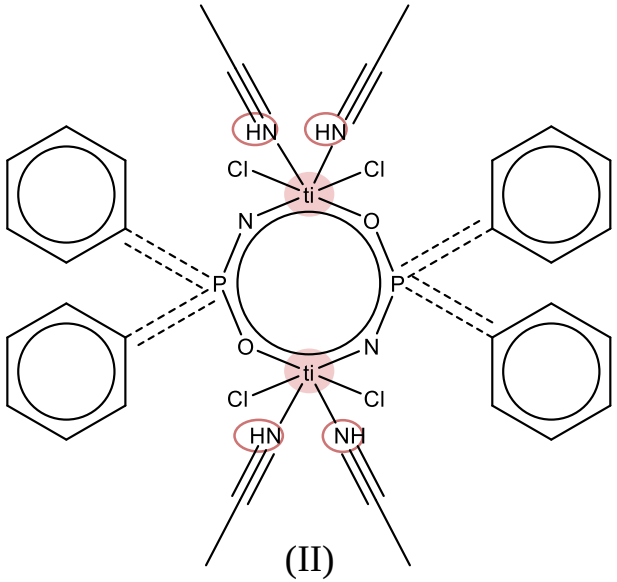
National Institutes of Health
9000 Rockville Pike
Bethesda, MD 20892
United States

Back and forth with SMILES – aromatic metal atoms

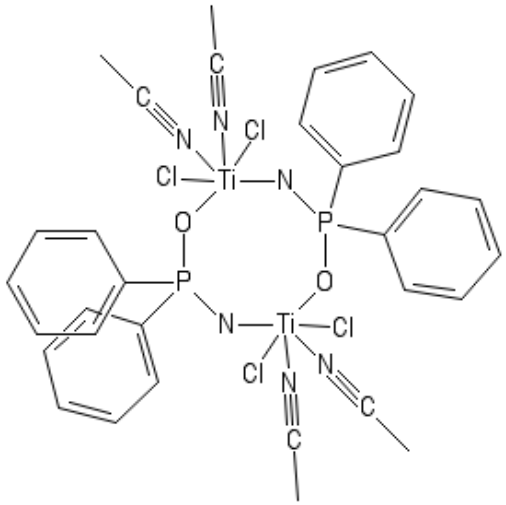
JIKROK



(I)



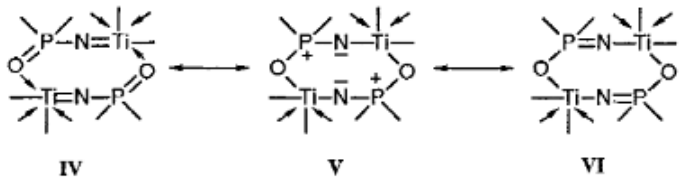
(II)



(III)

SMILES: Cl[ti]1(Cl)(np(o[ti](Cl)(Cl)(np(o1)(c1ccccc1)c1ccccc1)(N#CC)N#CC)(c1ccccc1)c1ccccc1)(N#CC)N#CC

I. CCDC Representation
II. SMILES interpreted by drawing package
III. SMILES interpreted by online sketcher
IV. Representation of chemistry in article



IV V VI

Scheme 2 Resonance forms of compound **7a**, outer ring atoms being omitted for clarity

10.1039/DT9910000663

Why is structure representation important?

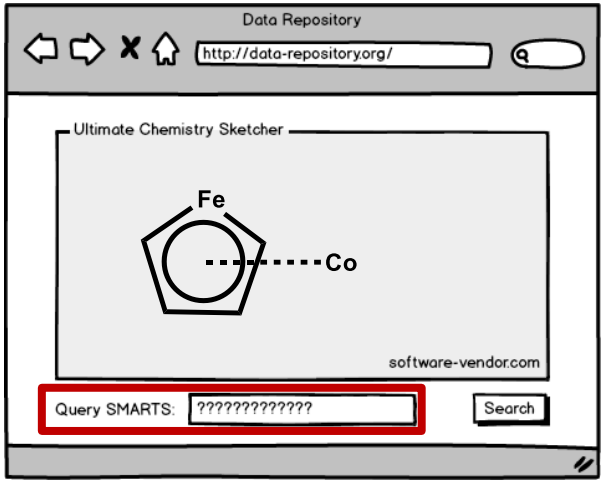
Findable

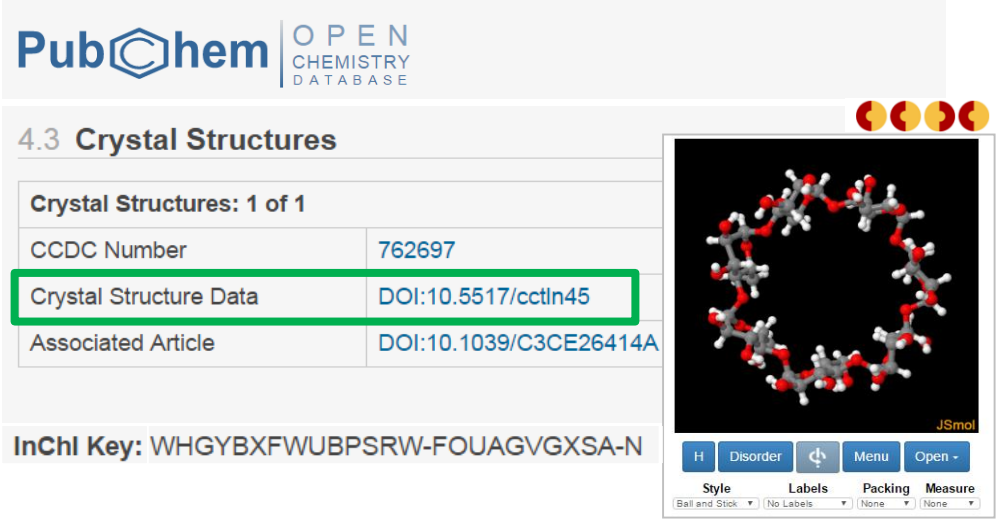
Interoperable

Reusable

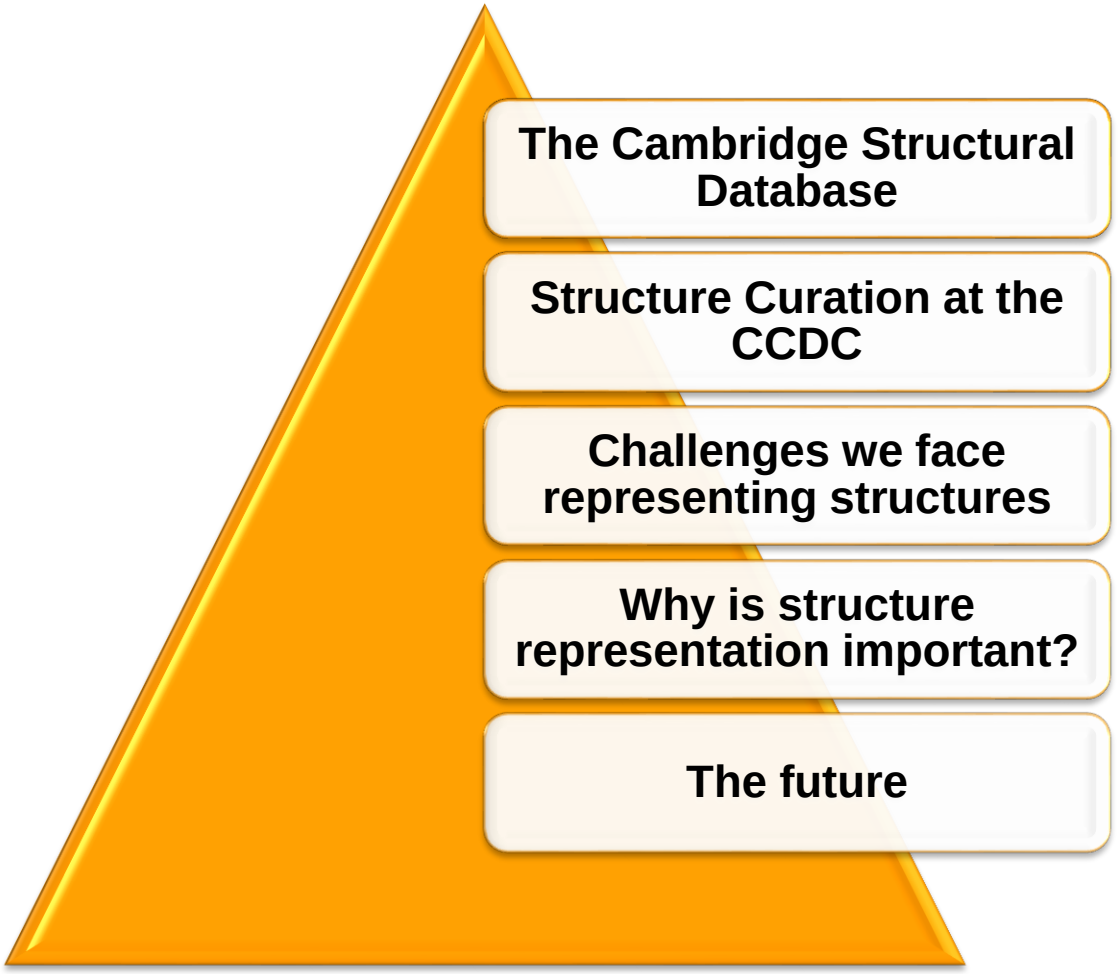
Why is structure representation important?

- **Consistent and reliable representation of chemical structures is important for**
  - interoperability between tools and data resources
  - data discovery and reuse
  - effective communication of science





Chemical structure representation challenges encountered when curating the CSD



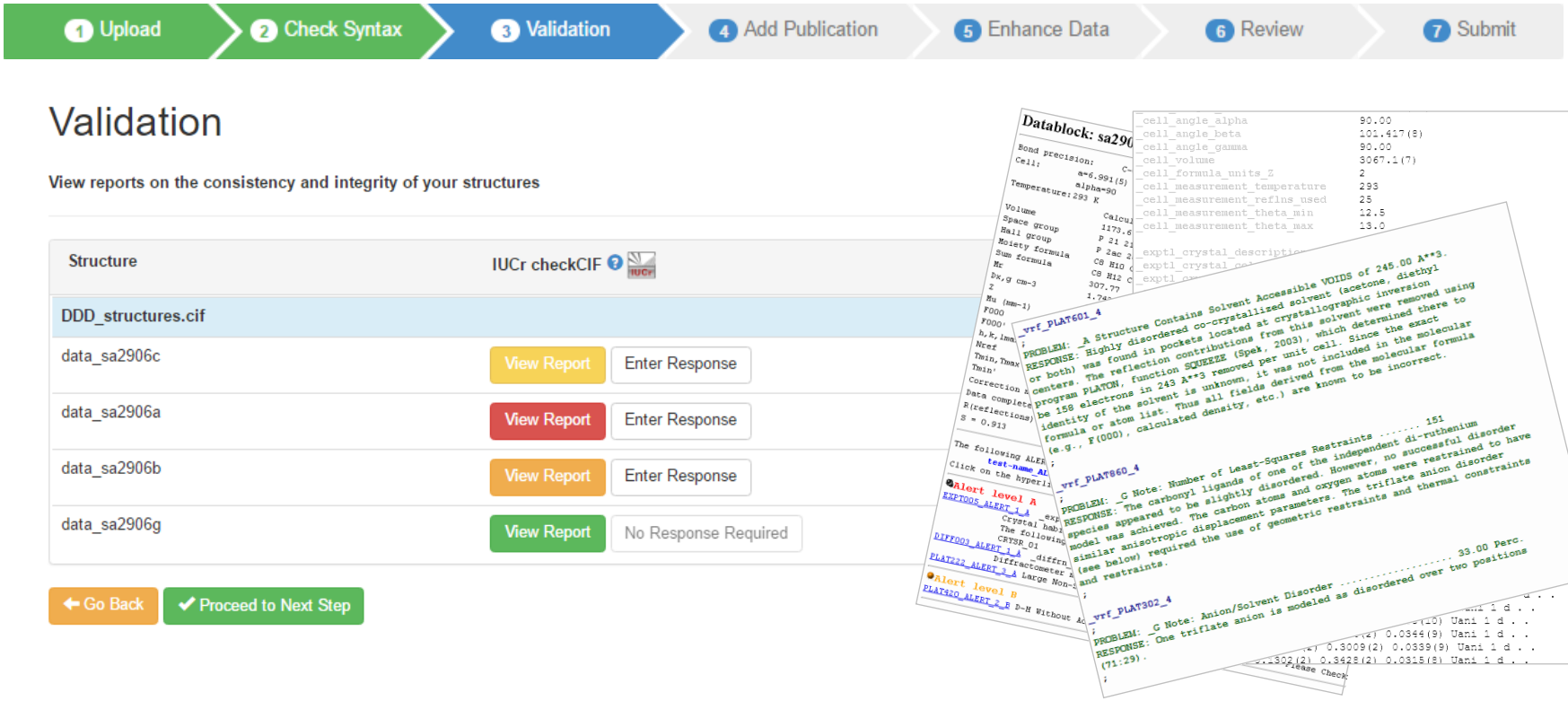
Who's responsible for getting it right?

Those responsible for providing infrastructure:

- ❑ Support the development of standards based on researcher needs
  - Funders
  - Professional bodies
  - Institutions
- ❑ Adopt standards and implement them in systems and workflows
  - Tool providers
  - Repositories
  - Publishers

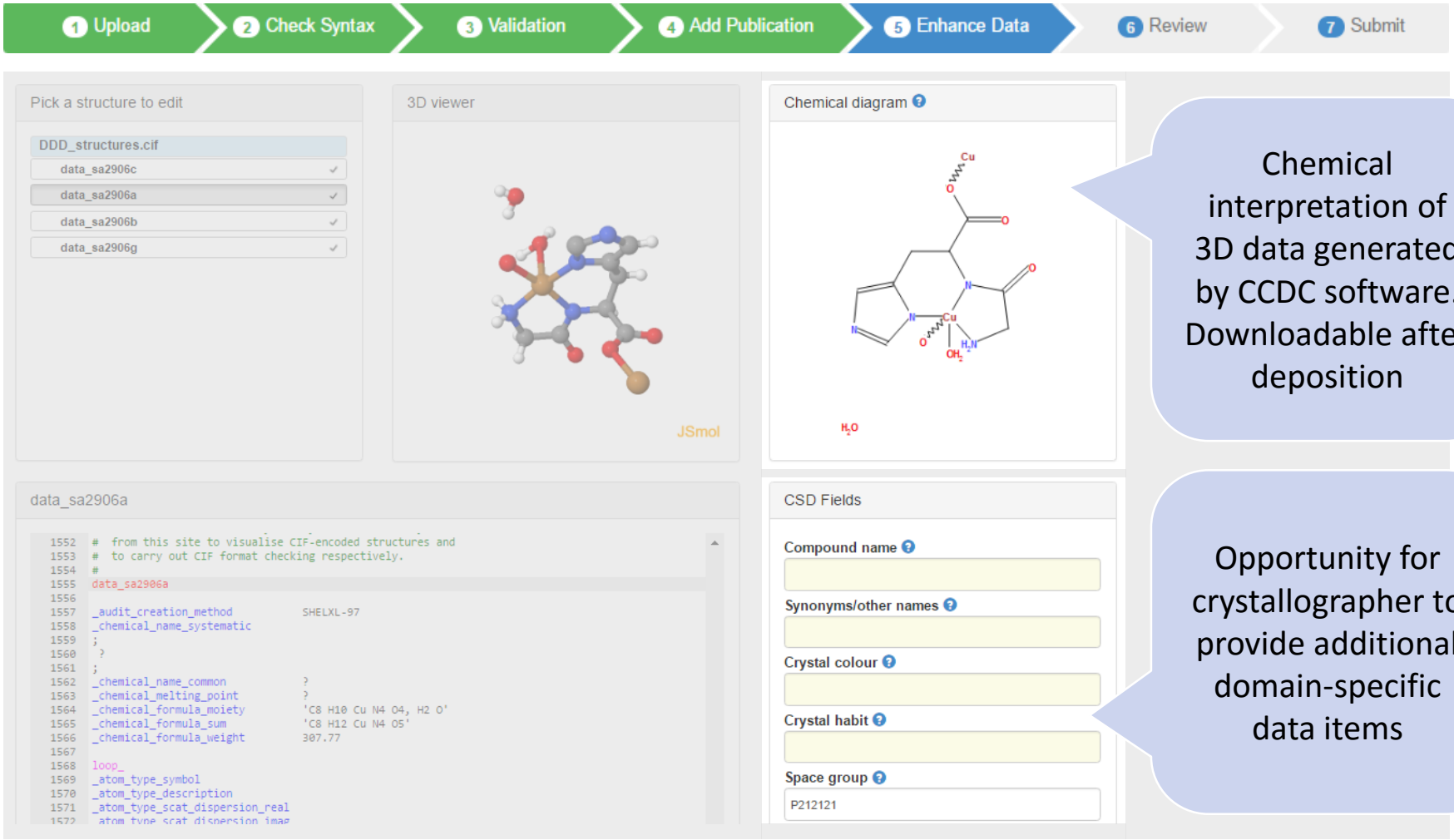
Aim - Enable researchers to conform to standards without really trying!

Integration of integrity checks



Validation reports now available to depositors and reviewers pre-publication and researchers post-publication alongside the data

Data enrichment: By the depositor



Connecting the experimental and publishing workflows

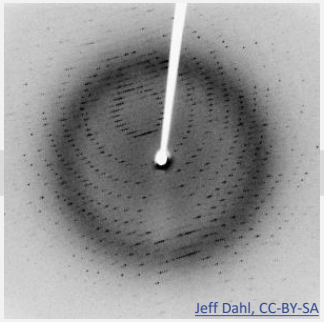
- We all want to make the data more reusable and a more consistent chemical representation helps with this.

- All parts of the experimental and publishing workflows should be as interconnected as possible to help this.

- Everyone involved in the process potentially has a role to play in this.

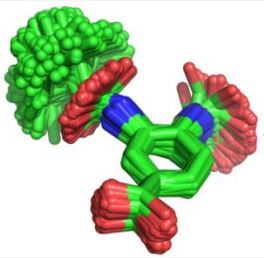
Experimental Data

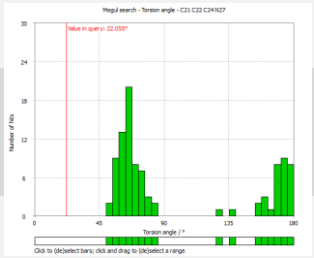
$C_{10}H_{16}N^{+},Cl^{-}$

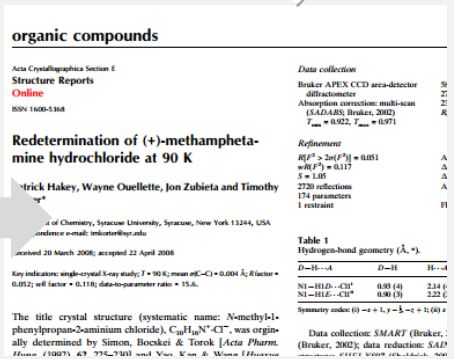



```
_diffrn_ambient_temperature        90(2)
_diffrn_radiation_type             MoK\a
_diffrn_radiation_wavelength       0.71073
_diffrn_radiation_monochromator    graphite
_diffrn_measurement_device_type    'Bruker APEX CCD area-detector'
_diffrn_measurement_method         '\f and \w'
_diffrn_detector_area_resol_mean   512
_diffrn_reflns_number              5892

loop_
_atom_site_type_symbol
_atom_site_label
_atom_site_fract_x
_atom_site_fract_y
_atom_site_fract_z
_atom_site_U_iso_or_equiv
_atom_site_adp_type
_atom_site_calc_flag
_atom_site_refinement_flags
_atom_site_occupancy
_atom_site_symmetry_multiplicity
_atom_site_disorder_assembly
_atom_site_disorder_group
Cl Cl1 0.23185(8) 0.78305(9) 0.55574(6) 0.02213(16) Uani d . 1 1 . .
N N1 0.8031(3) 0.6811(3) 0.5368(2) 0.0172(4) Uani d U 1 1 . .
C C1 0.6896(4) 0.7867(4) 0.4357(2) 0.0224(5) Uani d U 1 1 . .
C C2 0.7510(5) 0.8922(5) 0.7083(3) 0.0256(6) Uani d U 1 1 . .
C C3 0.7409(4) 0.6944(4) 0.6644(3) 0.0187(5) Uani d U 1 1 . .
C C4 0.8700(4) 0.5637(4) 0.7481(3) 0.0236(6) Uani d U 1 1 . .
```



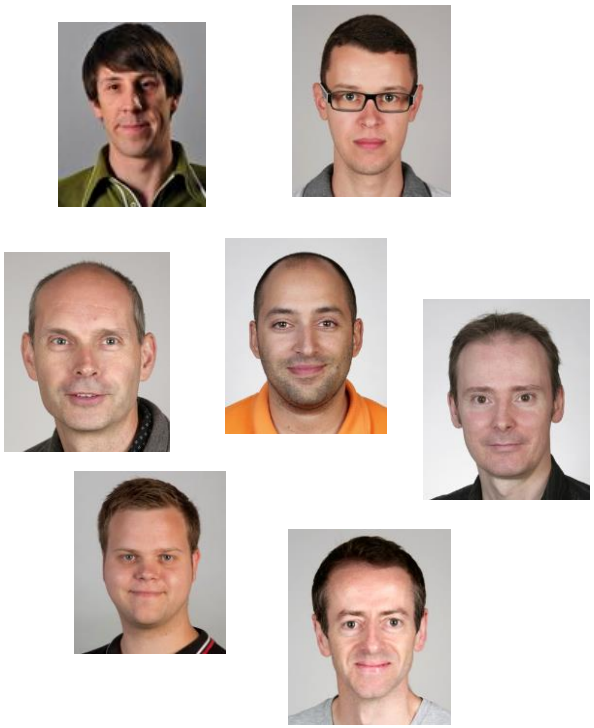




Scientific Knowledge

Acknowledgements

- Colleagues at CCDC



- Thank you for your attention