An Introduction to
The Cambridge Crystallographic Data Centre
The Cambridge Structural Database
Chemistry Data Initiatives

Ian Bruno

Cambridge Crystallographic Data Centre

@ijbruno @ccdc_cambridge

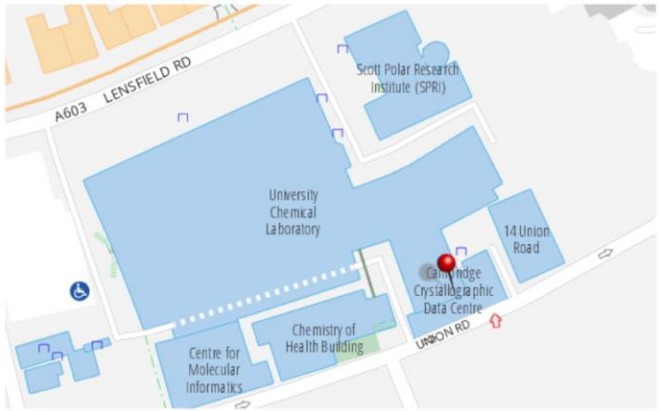
The Cambridge Crystallographic Data Centre

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

Scientific Software Provider
Search/analysis/visualisation tools
Scientific applications

Collaborative Research Organisation
New methodologies
Fundamental research

Education and Outreach
Conferences, Workshops,
Training, Teaching



The Crystallographic Data Centre Cambridge

Olga Kennard, Founding Executive Director

"The database was established in 1965 to fulfil a dream of myself and a great scientist, the polymath J.D. Bernal. We had a passionate belief that the collective use of data would lead to the discovery of new knowledge which transcends the results of individual experiments."

Kennard, O. "From Private Data to Public Knowledge." The Impact of Electronic Publishing on the Academic Community. Ed. I Butterworth. Portland Press Ltd, 1997. 159–166.

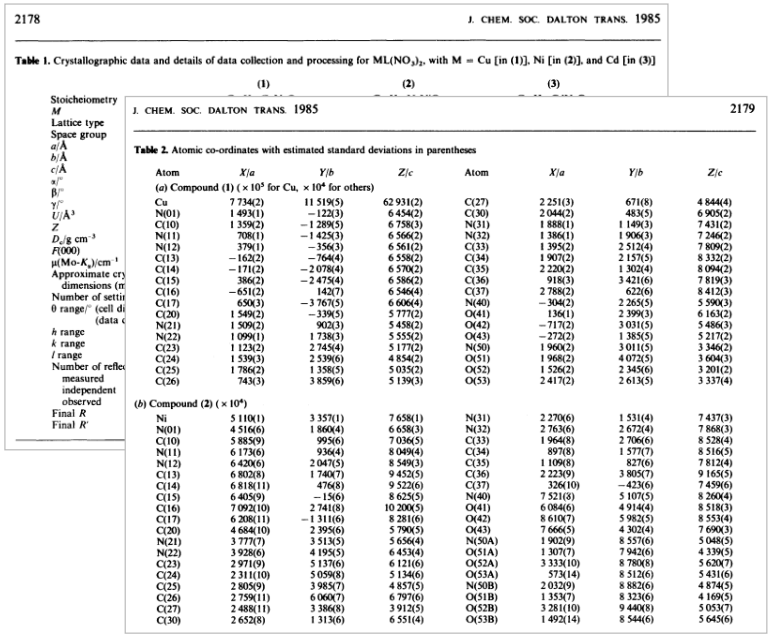
Established 1965 with funding from the Royal Society

Main objective was to assemble a computer-based file of information and data

Data was first published in printed volumes generated from the computer file

Early Publication of Crystal Structures

Hand-typed tables of coordinates in journal articles manually transcribed into database records



2178 J. CHEM. SOC. DALTON TRANS. 1985

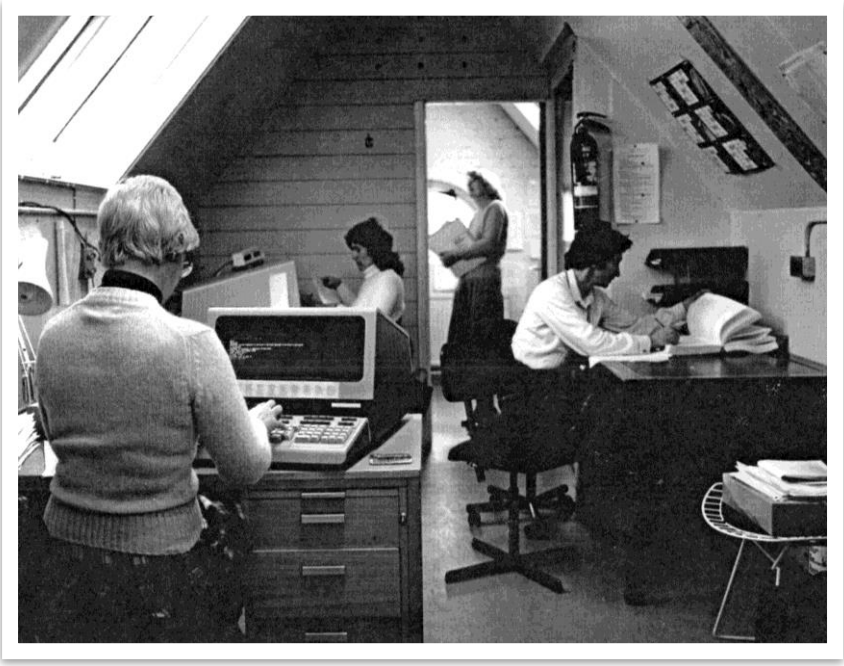
Table 1. Crystallographic data and details of data collection and processing for $ML(NO_3)_2$, with M = Cu [in (1)], Ni [in (2)], and Cd [in (3)]

	(1)	(2)	(3)
Stoicheiometry			
M			
Lattice type			
Space group			
a/Å			
b/Å			
c/Å			
α/°			
β/°			
γ/°			
U/Å^3			
Z			
D_c/g cm^{-3}			
F(000)			
μ(Mo-K_α)/cm^{-1}			
Approximate cry dimensions (m			
Number of settin			
θ range/° (cell di (data c			
h range			
k range			
l range			
Number of reflec measured			
independent			
observed			
Final *R*			
Final *R'*			

J. CHEM. SOC. DALTON TRANS. 1985 2179

Table 2. Atomic co-ordinates with estimated standard deviations in parentheses

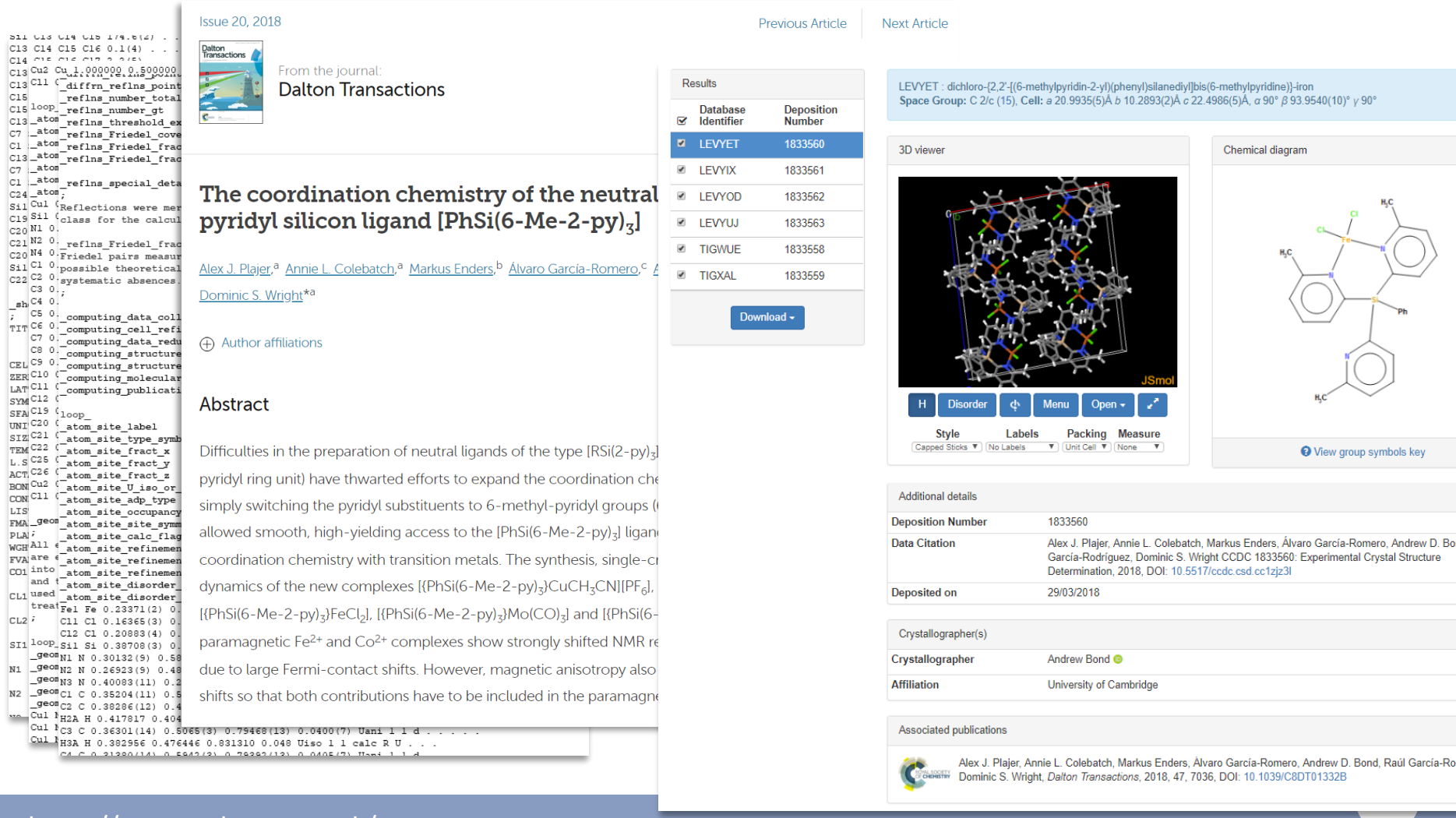
Atom	*X*/*a*	*Y*/*b*	*Z*/*c*	Atom	*X*/*a*	*Y*/*b*	*Z*/*c*
(*a*) Compound (1) (× 10^5 for Cu, × 10^4 for others)							
Cu	7 734(2)	11 519(5)	62 931(2)	C(27)	2 251(3)	671(8)	4 844(4)
N(01)	1 493(1)	−122(3)	6 454(2)	C(30)	2 044(2)	483(5)	6 905(2)
C(10)	1 359(2)	−1 289(5)	6 758(3)	N(31)	1 888(1)	1 149(3)	7 431(2)
N(11)	708(1)	−1 425(3)	6 566(2)	N(32)	1 386(1)	1 906(3)	7 246(2)
N(12)	379(1)	−356(3)	6 561(2)	C(33)	1 395(2)	2 512(4)	7 809(2)
C(13)	−162(2)	−764(4)	6 558(2)	C(34)	1 907(2)	2 157(5)	8 332(2)
C(14)	−171(2)	−2 078(4)	6 570(2)	C(35)	2 220(2)	1 302(4)	8 094(2)
C(15)	386(2)	−2 475(4)	6 586(2)	C(36)	918(3)	3 421(6)	7 819(3)
C(16)	−651(2)	142(7)	6 546(4)	C(37)	2 788(2)	622(6)	8 412(3)
C(17)	650(3)	−3 767(5)	6 606(4)	N(40)	−304(2)	2 265(5)	5 590(3)
C(20)	1 549(2)	−339(5)	5 777(2)	O(41)	136(1)	2 399(3)	6 163(2)
N(21)	1 509(2)	902(3)	5 458(2)	O(42)	−717(2)	3 031(5)	5 486(3)
N(22)	1 099(1)	1 738(3)	5 555(2)	O(43)	−272(2)	1 385(5)	5 217(2)
C(23)	1 123(2)	2 745(4)	5 177(2)	N(50)	1 960(2)	3 011(5)	3 346(2)
C(24)	1 539(3)	2 539(6)	4 854(2)	O(51)	1 968(2)	4 072(5)	3 604(3)
C(25)	1 786(2)	1 358(5)	5 035(2)	O(52)	1 526(2)	2 345(6)	3 201(2)
C(26)	743(3)	3 859(6)	5 139(3)	O(53)	2 417(2)	2 613(5)	3 337(4)
(*b*) Compound (2) (× 10^4)							
Ni	5 110(1)	3 357(1)	7 658(1)	N(31)	2 270(6)	1 531(4)	7 437(3)
N(01)	4 516(6)	1 860(4)	6 658(3)	N(32)	2 763(6)	2 672(4)	7 868(3)
C(10)	5 885(9)	995(6)	7 036(5)	C(33)	1 964(8)	2 706(6)	8 528(4)
N(11)	6 173(6)	936(4)	8 049(4)	C(34)	897(8)	1 577(7)	8 516(5)
N(12)	6 420(6)	2 047(5)	8 549(3)	C(35)	1 109(8)	827(6)	7 812(4)
C(13)	6 802(8)	1 740(7)	9 452(5)	C(36)	2 223(9)	3 805(7)	9 165(5)
C(14)	6 818(11)	476(8)	9 522(6)	C(37)	326(10)	−423(6)	7 459(6)
C(15)	6 405(9)	−15(6)	8 625(5)	N(40)	7 521(8)	5 107(5)	8 260(4)
C(16)	7 092(10)	2 741(8)	10 200(5)	O(41)	6 084(6)	4 914(4)	8 518(3)
C(17)	6 208(11)	−1 311(6)	8 281(6)	O(42)	8 610(7)	5 982(5)	8 553(4)
C(20)	4 684(10)	2 395(6)	5 790(5)	O(43)	7 666(5)	4 302(4)	7 690(3)
N(21)	3 777(7)	3 513(5)	5 656(4)	N(50A)	1 902(9)	8 557(6)	5 048(5)
N(22)	3 928(6)	4 195(5)	6 453(4)	O(51A)	1 307(7)	7 942(6)	4 339(5)
C(23)	2 971(9)	5 137(6)	6 121(6)	O(52A)	3 333(10)	8 780(8)	5 620(7)
C(24)	2 311(10)	5 059(8)	5 134(6)	O(53A)	573(14)	8 512(6)	5 431(6)
C(25)	2 805(9)	3 985(7)	4 857(5)	N(50B)	2 032(9)	8 882(6)	4 874(5)
C(26)	2 759(11)	6 060(7)	6 797(6)	O(51B)	1 353(7)	8 323(6)	4 169(5)
C(27)	2 488(11)	3 386(8)	3 912(5)	O(52B)	3 281(10)	9 440(8)	5 053(7)
C(30)	2 652(8)	1 313(6)	6 551(4)	O(53B)	1 492(14)	8 544(6)	5 645(6)



Kleywegt et al. (1985) *J. Chem. Soc., Dalton Trans*, 2177-2184
doi:10.1039/DT9850002177

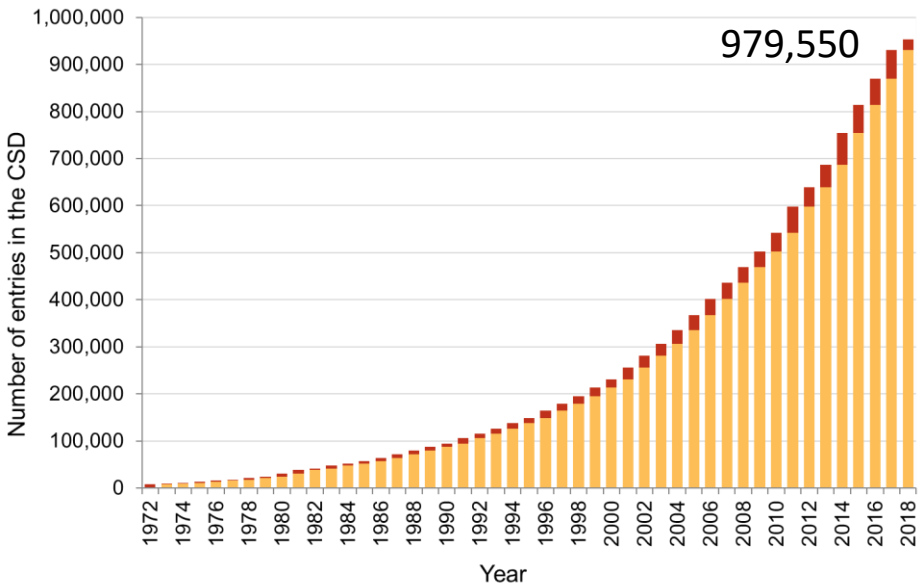
Publication of Crystal Structures Today

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

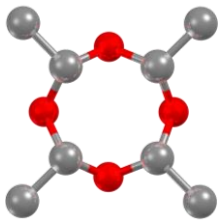


https://www.ccdc.cam.ac.uk/structures

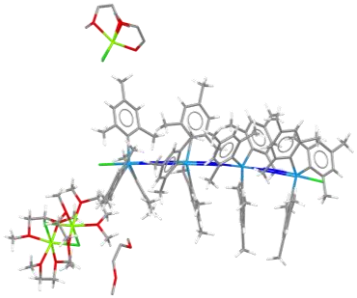
The Cambridge Structural Database



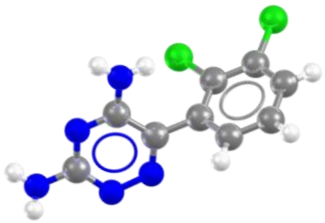
- 970,000+ small-molecule crystal structures
- Over 80,000 datasets deposited annually
- Structures available for anyone to download
- Links to over 1,000 journals
- Enriched and annotated by experts
- Access to data and knowledge



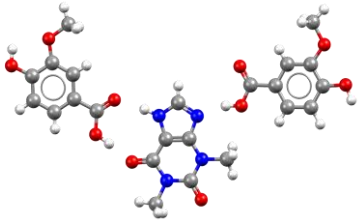
One of 1st – METALD
Metaldehyde – published 1936



250,000th – IBEZUK
Conducting metal-dinitrogen polymer



500,000th – EFEMUX01
Lamotrigine – an anti-convulsant drug



750,000th – ZOYBIA
Co-crystal of vanillic acid and theophylline

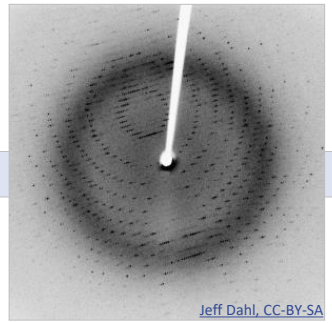
? **1,000,000th structure expected in 2019 #CSD1Million**

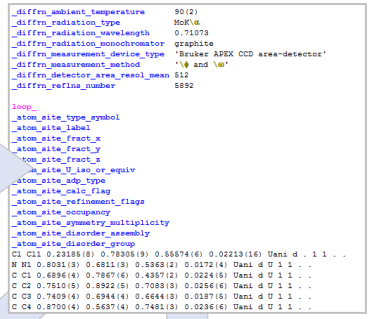
as at 28 January 2019

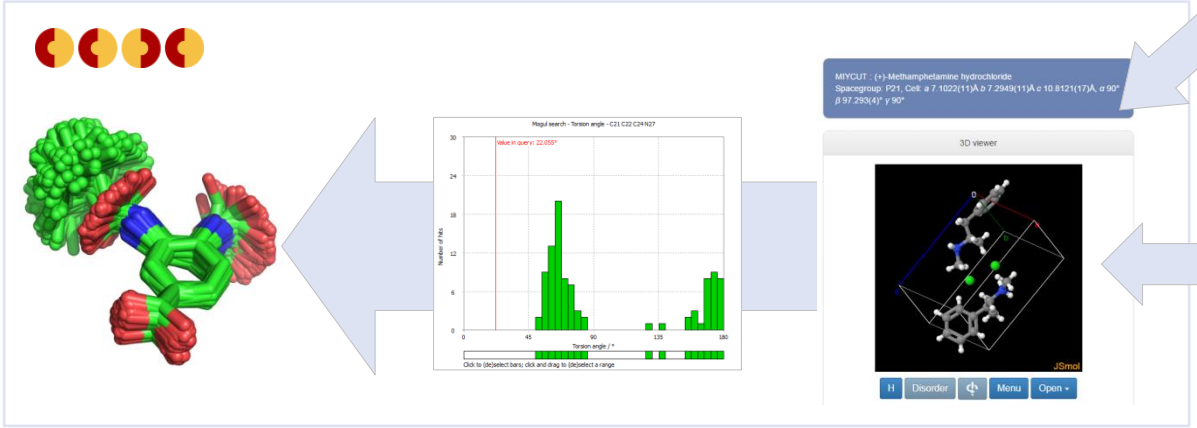
From Data to Knowledge

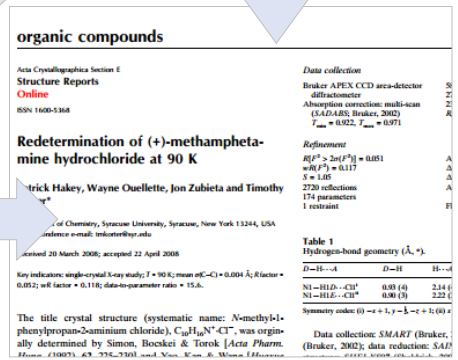
Experimental Data

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





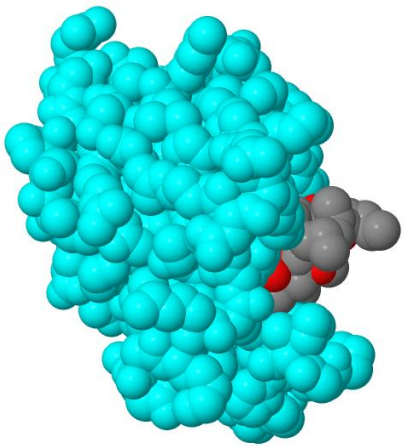


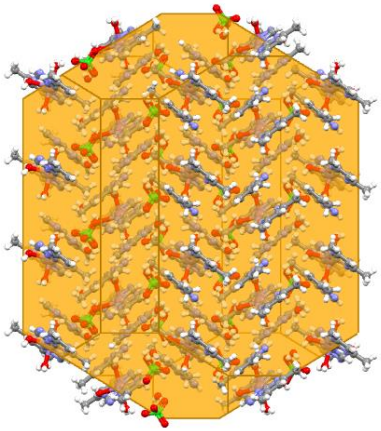


Structural Knowledge

Structural Chemistry Insights

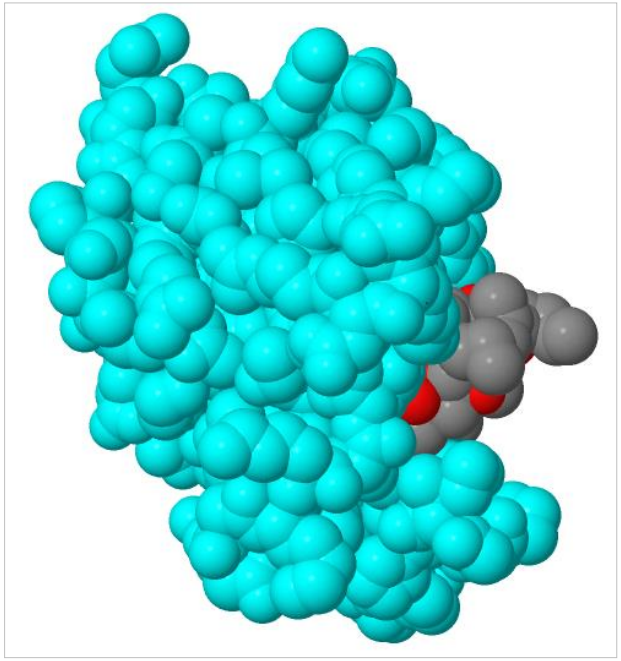
- **Crystal structure data gives insights into:**
  - ❑ Molecular dimensions and shape
  - ❑ Molecular interactions
  - ❑ Solid form properties
- **Applicable to various domains including:**
  - ❑ Drug design and development
  - ❑ Design of new materials
  - ❑ Crystal engineering
  - ❑ Structure validation

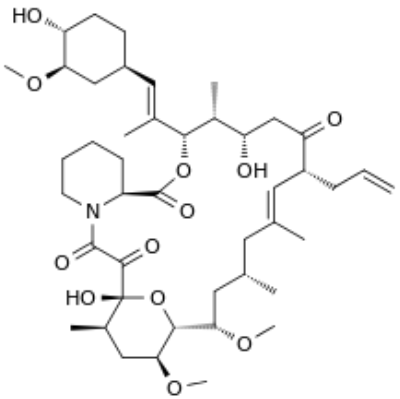




Crystal Structure Knowledge Helps Drug Design

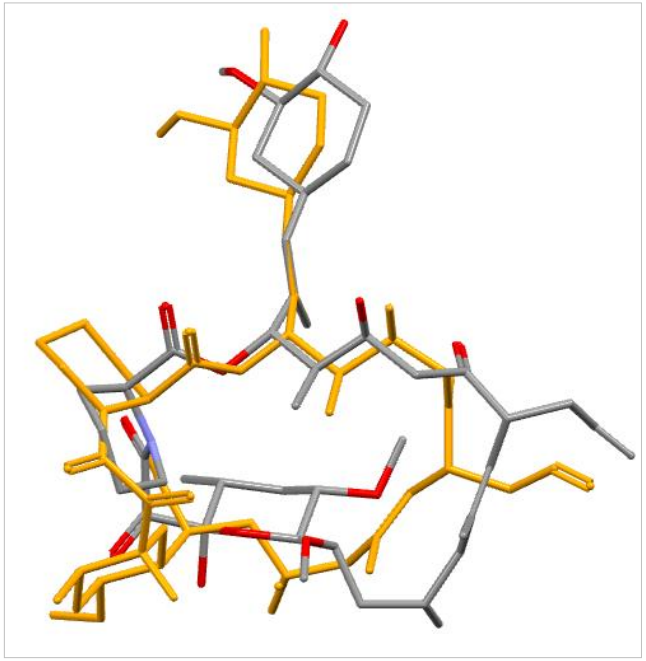
PDB 1BKF complexed with PDB Ligand FK5





Tacrolimus: An immunosuppressive drug. Also used in the treatment of skin conditions.

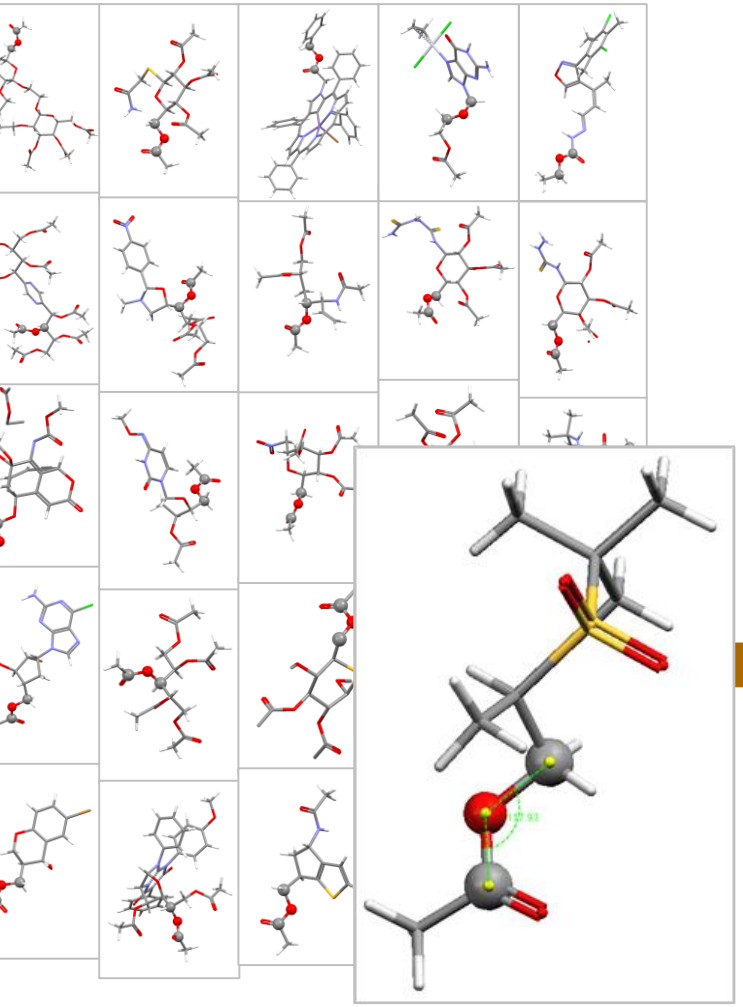
CSD FINWEE10 overlayed on PDB Ligand FK5



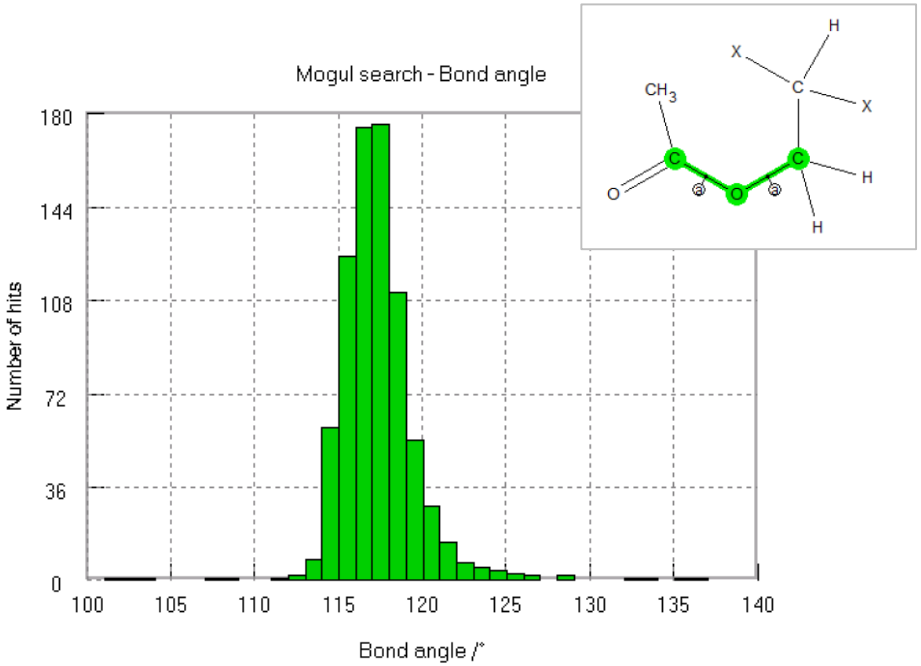
Understanding factors that influence the shape of molecules helps identify better drug candidates

Molecular Recognition of Protein−Ligand Complexes: Applications to Drug Design. R.Babine and S. Bender. Chem. Rev. (1997) doi:10.1021/cr960370z

Mogul A Knowledge Base of Molecular Geometries

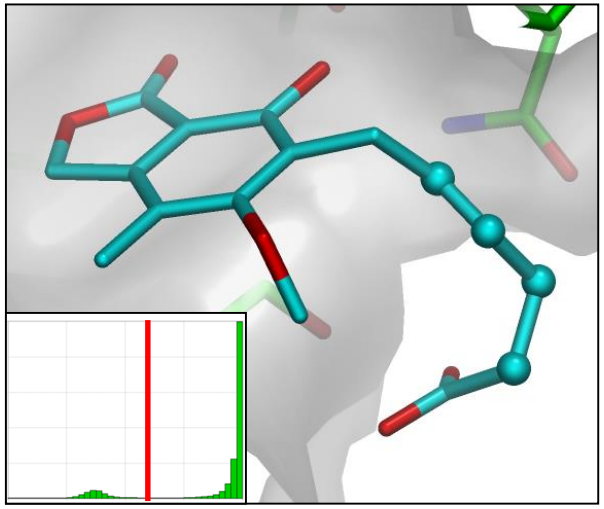


Millions of experimentally observed bond lengths, valence angles and torsions are combined into distributions that indicate preferred geometries of structural features



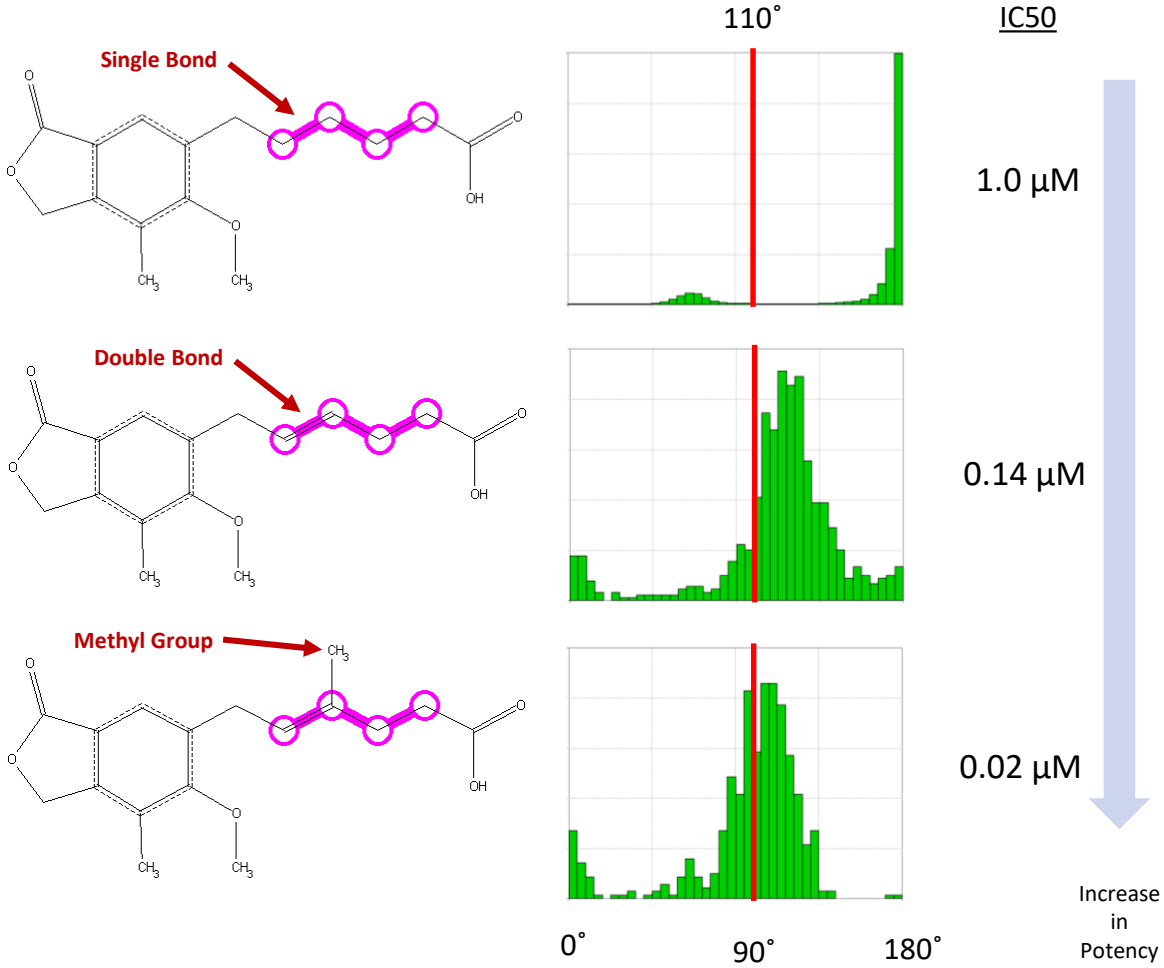
Bruno et al., J. Chem. Inf. Comput. Sci., 44, 2133-2144, 2004

Knowledge-driven Conformer Design



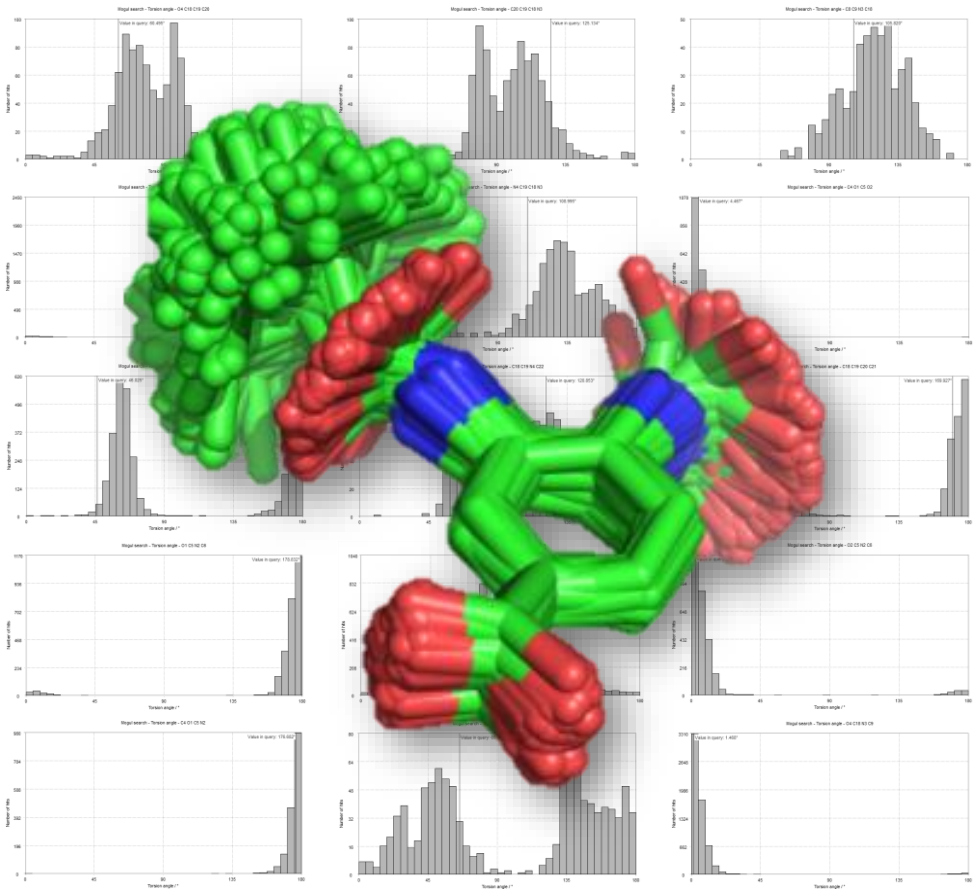
The immmunosuppressant mycophelonic acid binds to inosine monophosphate dehydrogenase with an affinity of 1.0 µM.

The CSD reveals that the highlighted torsion at 110˚ is unfavourable.



Small Molecule Conformational Preferences Derived from Crystal Structure Data. A Medicinal Chemistry Focused Analysis
Ken A. Brameld, Bernd Kuhn, Deborah C. Reuter, and Martin Stahl, *J. Chem. Inf. Model.*, 2008, 48, 1-24

Knowledge-based Conformer Generation



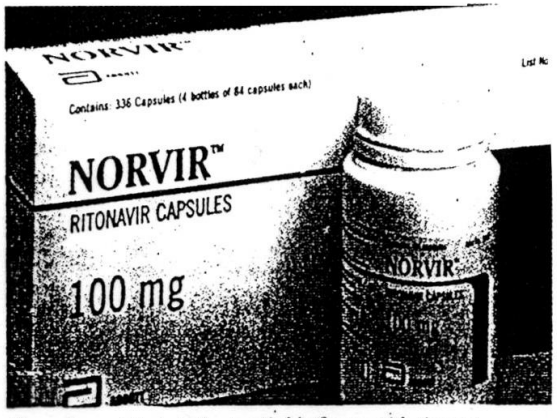
- Bond, angle and torsion distributions derived from experimental structures are used to produce realistic ensembles of low energy structures.
- The CSD Conformer Generator can be used to minimise molecular conformations and generate diverse conformer subsets based on experimental data
- Applicable to drug discovery and drug development

Knowledge-Based Libraries for Predicting the Geometric Preferences of Druglike Molecules
Robin Taylor, Jason Cole, Oliver Korb, and Patrick McCabe. J. Chem. Inf. Model., 2014, 54 (9), pp 2500–2514 DOI: 10.1021/ci500358p

Mitigating Risk in Drug Development

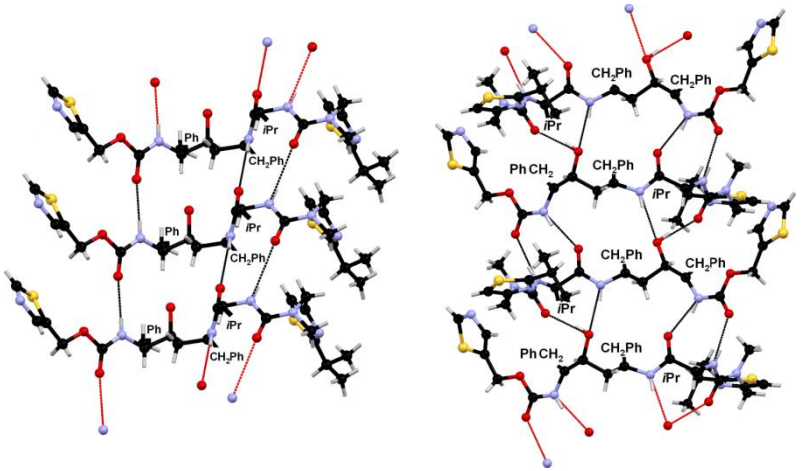
Manufacturing problems hit Abbott's HIV drug ritonavir

Capsules of Abbott Laboratories' protease inhibitor Norvir (ritonavir) are likely to become unavailable by the middle of August. The company has a problem with the manufacture of the anti-HIV capsules which it cannot resolve at present.



Capsules unlikely to be available from mid-August

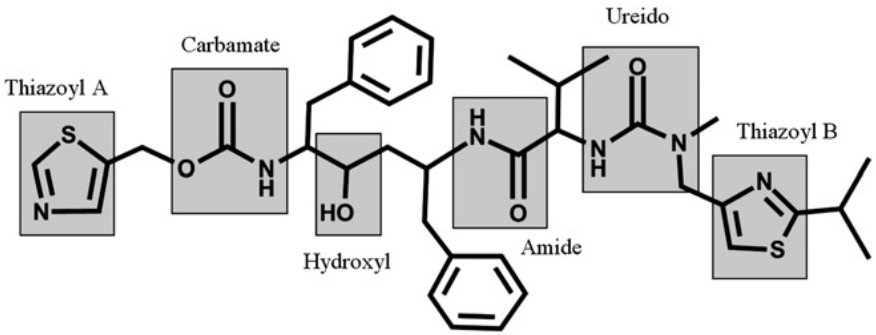
The problem relates to "undesirable" crystal formation. Abbott says that a series ples from a number of marketed batches of capsules were examined and there was no



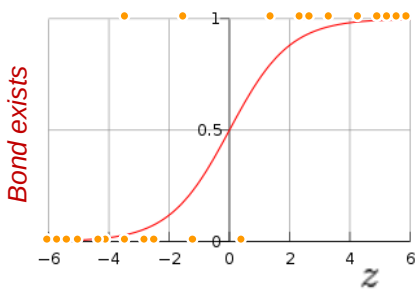
Different crystal forms, different interactions, different solubility, different stability.

Knowing the likelihood of specific molecular interactions occurring helps assess the risk of undesirable crystal formation

Hydrogen Bond Propensity: Ritonavir



Predictive analytics is used to identify feasible and unusual crystal packings based on information from known crystal structures of molecules similar to the target.



Propensity Π (H-bond exists) $= \dfrac{1}{1 + e^{-Z}}$

$Z = \beta_0 + \beta_1.DG1 + \beta_2.AG1 + \beta_3.DSD + \ldots$

DG1, AG1, DSD: Explanatory variables

Logistic Regression

Table 1 Propensity predictions for potential donor–acceptor combinations in ritonavir (as labelled in Fig. 1), and observed hydrogen bonds in either polymorphic form

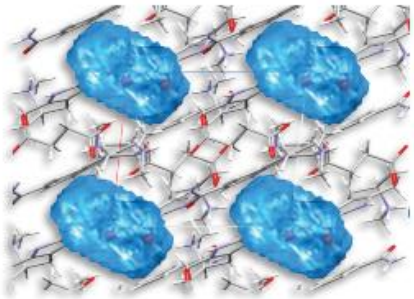
Donor	Acceptor	π	$\pm$ [a]	Form I	Form II
amide	carbamate	0.618	0.094	✗	✗
amide	hydroxyl	0.551	0.052	✗	✓
carbamate	carbamate	0.538	0.090	✓	✗
hydroxyl	carbamate	0.537	0.090	✗	✗
amide	amide	0.501	0.055	✓	✗
amide	ureido	0.499	0.072	✗	✗
carbamate	hydroxyl	0.470	0.078	✗	✗
hydroxyl	hydroxyl	0.469	0.037	✗	✗
carbamate	amide	0.420	0.083	✗	✓
hydroxyl	amide	0.419	0.045	✗	✗
carbamate	ureido	0.418	0.088	✗	✗
hydroxyl	ureido	0.417	0.058	✗	✓
ureido	carbamate	0.319	0.086	✗	✓
ureido	hydroxyl	0.263	0.041	✗	✗
ureido	amide	0.225	0.040	✗	✗
ureido	ureido	0.224	0.044	✓	✗
amide	thiazoyl a	0.152	0.054	✗	✗
amide	thiazoyl b	0.142	0.050	✗	✗
carbamate	thiazoyl a	0.115	0.044	✗	✗
hydroxyl	thiazoyl a	0.114	0.039	✓	✗
carbamate	thiazoyl b	0.107	0.041	✗	✗
hydroxyl	thiazoyl b	0.106	0.036	✗	✗

[a] The error bars of the coefficient value: the value falls within this range at the 95% confidence level, based on a χ^2 distribution.

Knowledge-Based H-Bond Prediction to aid experimental polymorph screening. P.Galek et al. CrystEngComm. (2009) doi:10.1039/B910882C

CSD-Materials

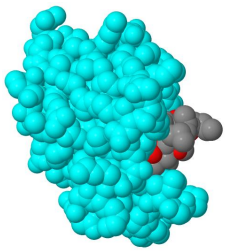
Informatics-based solutions that aid in the understanding and prediction of solid form stability and properties



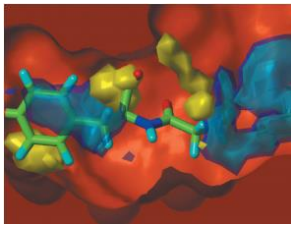
- Discover preferred interactions and engineer changes to satisfy these requirements using *Full Interaction Maps*
- Interpret crystal packing and compare with CSD data using powerful *packing feature*, *similarity* and *motif searches*, and *hydrogen bond propensity analysis*
- Understand the effects of hydration on your lattice with the *hydrate analyser*
- Explore the structures of potential co-crystals using the *molecular complementarity tool*

Developed in partnership with Industry through CCDC's Crystal Form Consortium

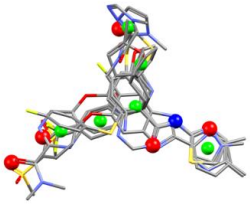
CSD-Discovery



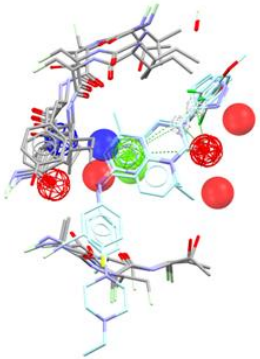
GOLD: Protein-ligand docking - virtual screening, lead optimisation and binding mode prediction



SuperStar: Knowledge-based prediction of intermolecular interactions based on data from small molecule crystal structures



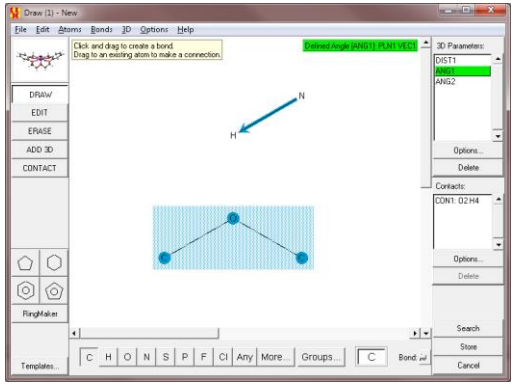
Ligand Overlay: applying structural knowledge to identify common binding modes, interactions and geometries of structurally diverse ligands



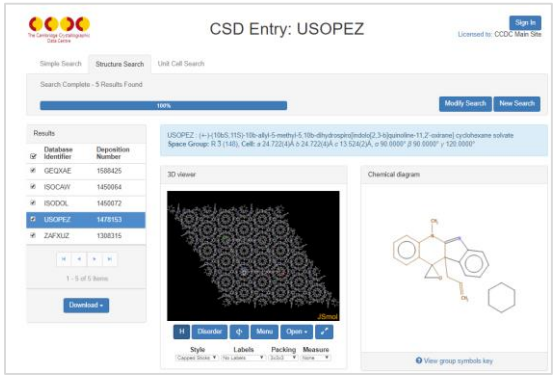
CSD-CrossMiner: Fast and flexible pharmacophore searching across the CSD and the PDB for lead optimisation and scaffold hopping

CSD-System

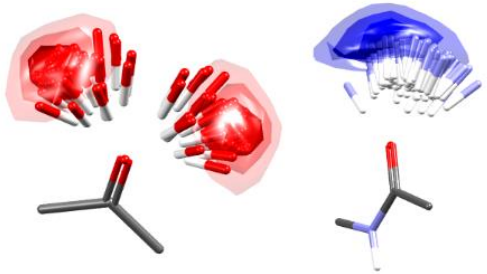
ConQuest: Advanced 3D searching



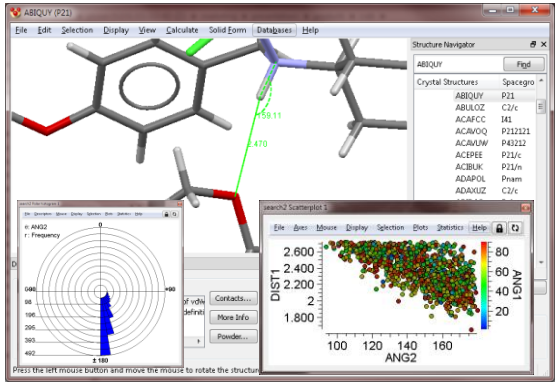
WebCSD: On-line portal to the CSD



IsoStar: Molecular interaction analysis



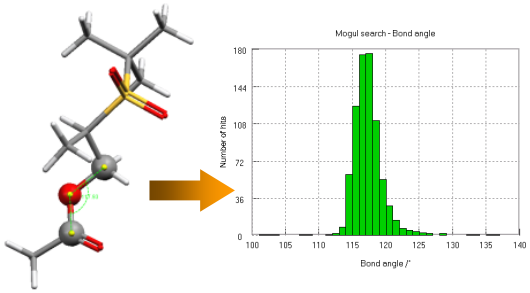
Mercury: Visualisation & data analysis



CSD Python API: Custom search and analysis

```
import sys

from ccdc import conformer
from ccdc import io

args = parser.parse_args()

mol_reader = io.MoleculeReader(args.inmolfn)
engine = conformer.GeometryAnalyser()

molecules = []
min_unusual_torsions = sys.maxint
for (idx, molecule) in enumerate(mol_reader):
    molecule.standardise_aromatic_bonds()
    molecule.standardise_delocalised_bonds()

    # Do the analysis
    geometry_analysed_molecule = engine.analyse_molecule(molecule)

    # Count number of unusual torsions
    molecule.unusual_torsions = [
        t for t in geometry_analysed_molecule.analysed_torsions
            if t.unusual and t.enough_hits]
    num_unusual_torsions = len(molecule.unusual_torsions)
    molecule.num_unusual_torsions = num_unusual_torsions
    molecules.append(molecule)

    if num_unusual_torsions < min_unusual_torsions:
        min_unusual_torsions = num_unusual_torsions
```

Mogul: Molecular geometry analysis



CCDC Knowledge-based Software Solutions

CSD-System: Crystallographers, structural chemists, educators

Find, analyse and communicate crystal structures

CSD-Discovery

Medicinal chemists, computational chemists, structural biologists

Protein and ligand-based design of new molecules

CSD-Materials

Solid form experimentalists, crystallization scientists

Behaviour and properties of new materials

CSD-Enterprise

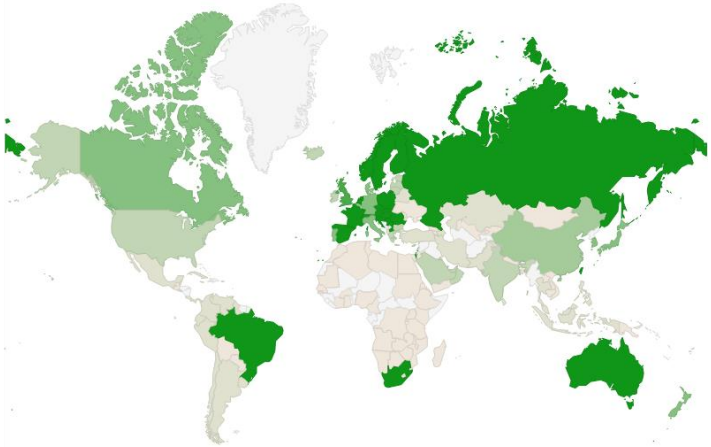
All CCDC application software
(available to all Academics)

Academic Access to Value-added Services

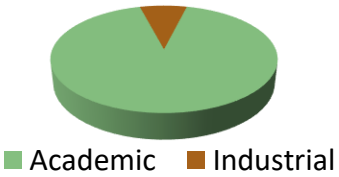
- **Country-wide licences**
  - via National Affiliated Centres
- **Campus-wide licences**
  - often through University Libraries
- **Individual Researchers and Groups**
  - cost can be included in grant applications
- **CCDC Subsidy**
  - Supporting research in developing countries via the FAIRE programme

24 countries with academic country-wide licences

Colour-coding reflects nature of academic arrangements

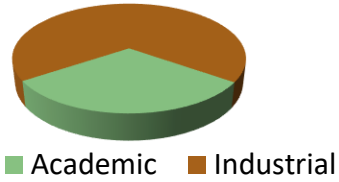


User Community



Number of customers, 2018

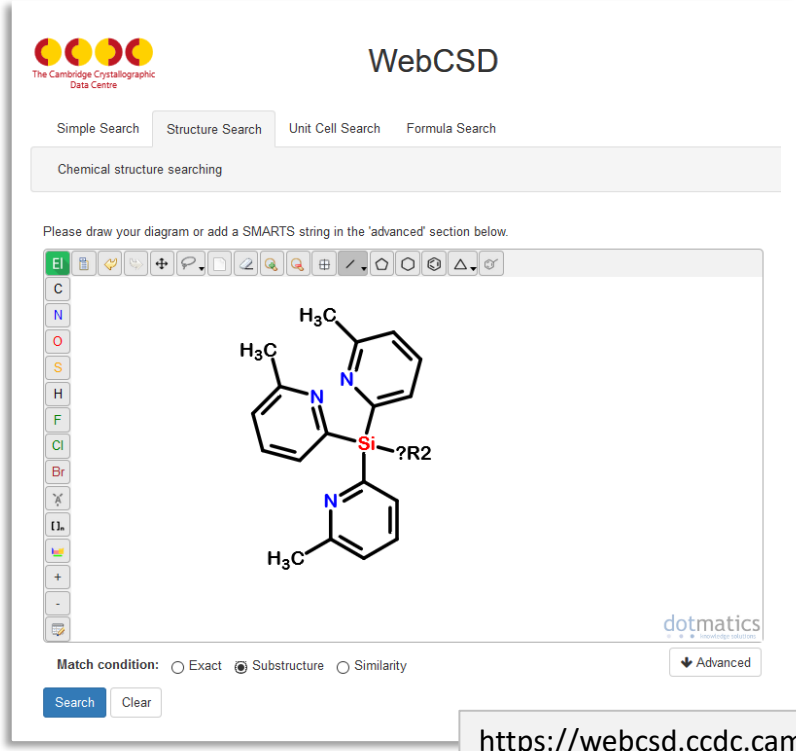
Revenue



CSD and Software, 2018

Chemistry Department Access to CSD-Enterprise

- **Download CSD onto your machine** (Linux, Windows, OS X) following these instructions: https://www.ch.cam.ac.uk/computing/software/cambridge-structural-database-system
- CSD-System Components **installed on MCS machines** in the Chemistry Library and around the University:
  - CSD Conquest
  - CSD DASH
  - CSD 2017 encipher
  - CSD 2017 GOLD
  - CSD 2017 Hermes
  - CSD 2017 IsoStar
  - CSD 2017 IsoStar Server
  - CSD 2017 Mercury
  - CSD 2017 Mogul
  - CSD 2017 PreQuest
  - CSD 2017 Python API



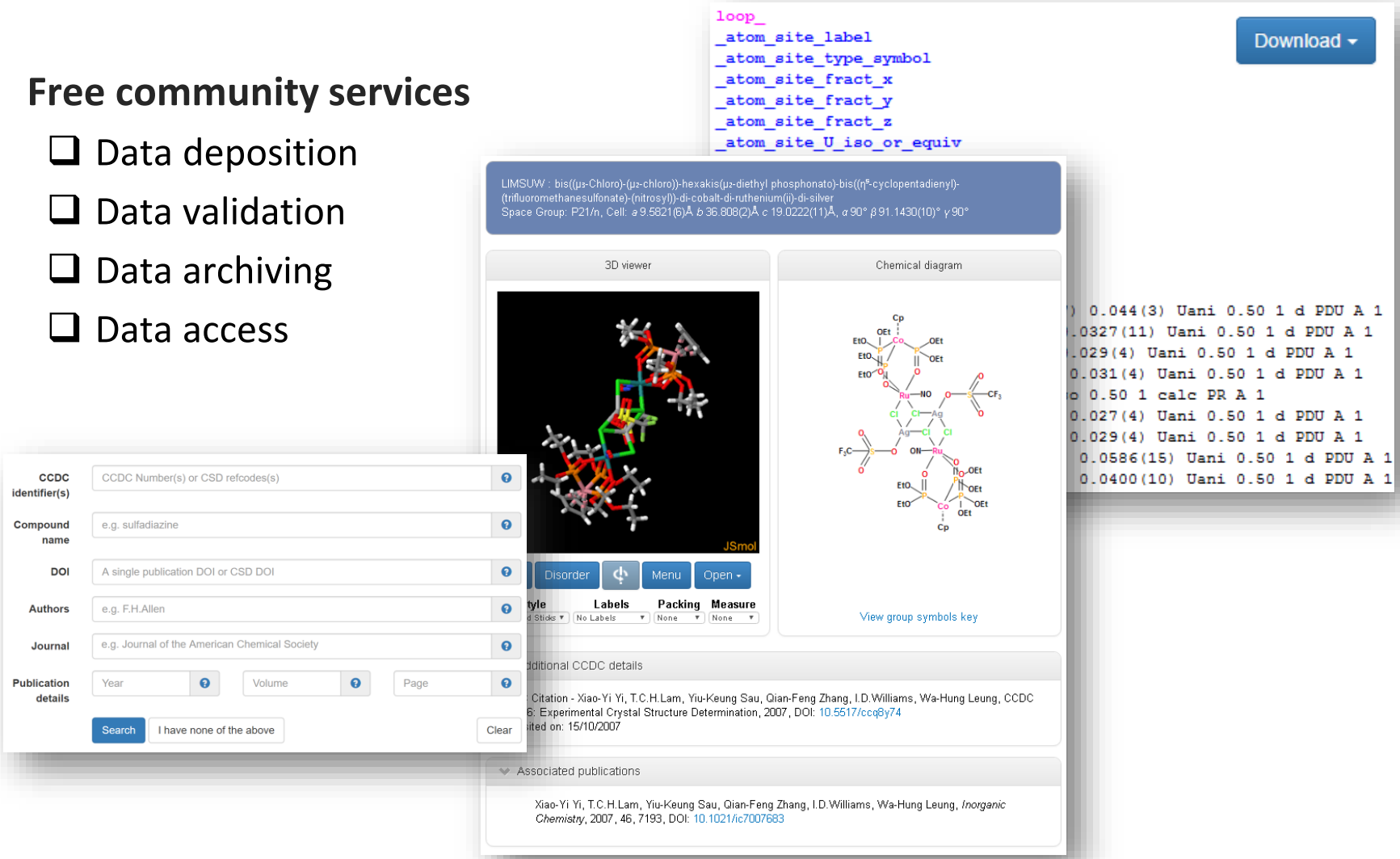
https://webcsd.ccdc.cam.ac.uk/

Sharing Crystallographic Data

Principles and Practices

CSD-Community Services

- **Free community services**
  - ❑ Data deposition
  - ❑ Data validation
  - ❑ Data archiving
  - ❑ Data access



https://www.ccdc.cam.ac.uk/Community/csd-community/

https://www.data.cam.ac.uk/funders

Funder Research Data Sharing Policies

"Publicly funded research data are a public good, (...), which should be made openly available with as few restrictions as possible in a timely and responsible manner (...)."

UK Research and Innovation

"Research data that supports publications must be stored for 10 years."

BBSRC

"EPSRC has the strictest policy on research data sharing…"

EPSRC

- All publications … should have a statement describing how to access underlying data
- Data should be stored for at least 10 years or for 10 years from the last request for access to the data
- All data supporting research publications should be cited with the use of persistent links, for example DOIs (Digital Object Identifiers)…
- Research data should be accompanied by metadata … to allow the discovery of data.

https://www.data.cam.ac.uk/funders/epsrc-funded-researchers

FAIR Data Principles

Comment | OPEN

The FAIR Guiding Principles for scientific data management and stewardship

Mark D. Wilkinson, Michel Dumontier [...] Barend Mons

Wilkinson, M. D. *et al.* The FAIR Guiding Principles for scientific data management and stewardship. *Sci. Data* 3:160018 doi: 10.1038/sdata.2016.18 (2016).

DATA SHOULD BE

Findable | **Interoperable**

Accessible | **Reusable**

BY HUMANS AND MACHINES

Data FAIRport

http://www.datafairport.org/

FORCE11
The Future of Research Communications and e-Scholarship

https://www.force11.org/group/fairgroup/fairprinciples

Aspects of FAIR Data

Findable

- Globally unique and persistent **identifiers**
- Rich **metadata descriptions**
- (Meta)data available in a **searchable resource**

Interoperable

- Standard **formats** for representation
- Use of FAIR **vocabularies**
- **References to other (meta)data**

Accessible

- (Meta)data **retrievable by their identifier**
- Standard, **open communication protocols**
- **Metadata accessible** even when data are not

Reusable

- Described with a plurality of attributes
  - data usage **licenses**
  - detailed **provenance**
  - domain-relevant **community standards**

CCDC Deposition Services

1 Login | 2 Upload | 3 Check Syntax | 4 Validation | 5 Add Publication | 6 Enhance Data | 7 Review | 8 Submit

CIF deposition and validation service

First name(s) *: Ian

Last name(s) *: Bruno

Your email address *: bruno@ccdc.cama.cuk

Your ORCID iD: 0000-0003-4901-9936 | Create or Connect your ORCID iD

Additional email addresses: Please add any additional email addresses

Institution (e.g. University/Company) *:

Deposition number(s) for revision:

CIF/HKL/RES/FCF/Word/ZIP files: Select Files... ✓ Done

- structure01.cif 22.82 KB
- structure02.cif 146.20 KB

Details *: ☐ Remember my details

Options *: ☑ I wish to run the IUCr *checkCIF/PLATON* service on my data

Reset Progress | Proceed to Next Step →

CCDC Deposition Services

1 Login | 2 Upload | 3 Check Syntax | 4 Validation | 5 Add Publication | 6 Enhance Data | 7 Review | 8 Submit

CIF dep

First name(s)

Last name(s)

Your email address

Your ORCID iD

Additional email addresses

Institution (e
University/Company)

Deposition number(s) for revisi

CIF/HKL/RES/FCF/Word/ZIP fi

No Structure Factor data

Structure Factor data are a
you are unable to embed S

If in exceptional circumstan
Next Step', however you mu
embedded into your depos

Reason why your deposition

Check Syntax

The files highlighted in red in the left-hand column contain errors that need fixing before proceeding.

Please click on any red file names in the left-hand column, make the appropriate edits and then click the 'Save & Recheck File' button before proceeding to the next step.

For more information on how to fix errors please see our correcting CIFs page.

Pick file to edit

- structure01.cif
- structure02.cif

File contents **structure01.cif**

```
30  data_I
31  _audit_creation_method     SHELXL-97
32  _chemical_name_systematic
33  ;
34  {5-[(7-chloroquinolinium-4-yl)amino]-2-hydroxybenzyl}diethylammonium dichloride
35  dihydrate
36  ;
37  _chemical_name_common     Amodiaquine dihydrochloride dihydrate'
38  _chemical_formula_moiety     'C20 H24 Cl N3 O 2+, 2(Cl -), 2(H2 O)'
39  _chemical_formula_sum     C20 H28 Cl3 N3 O3'
40  _chemical_formula_iupac     'C20 H24 Cl N3 O 2+, 2Cl -, 2H2 O'
41  _chemical_formula_weight     464.80
42  _chemical_melting_point     ?
43  _symmetry_cell_setting     monoclinic
44  _symmetry_space_group_name_H-M     'P 21/c
45  _symmetry_space_group_name_Hall     '-P 2ybc'
46  loop_
47      _symmetry_equiv_pos_as_xyz
48      'x, y, z'
49      '-x, y+1/2, -z+1/2'
50      '-x, -y, -z'
51      'x, -y+1/2, z+1/2'
52  _cell_length_a     7.76220(10)
53  _cell_length_b     26.8709(4)
54  _cell_length_c     10.7085(2)
55  _cell_angle_alpha     90.00
56  _cell_angle_beta     92.7840(10)
57  _cell_angle_gamma     90.00
58  _cell_volume     2230.91(6)
59  _cell_formula_units_Z     4
```

Go Back | Save & Recheck File | Proceed to Next Step

Error 44 No terminating (') quote

Crystallographic Information File: CIF

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)

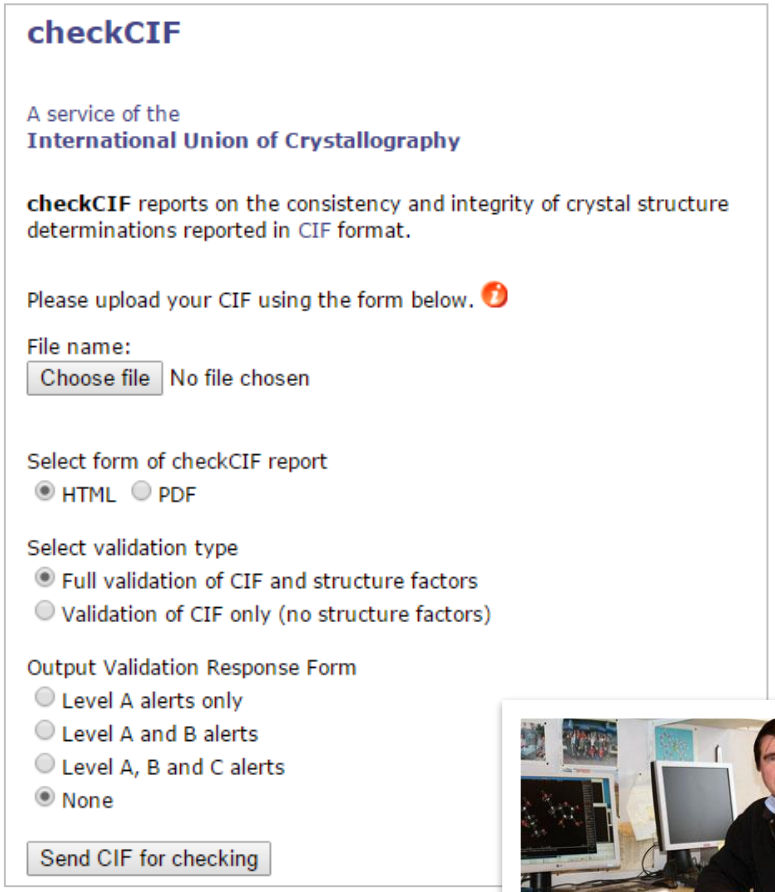
- **A standard format for archive and exchange of crystallographic data**
  - derived model
  - processed data (structure factors)
  - metadata about raw data (imgCIF)

- **Data items semantically defined by CIF dictionaries (vocabularies)**
  - crystallisation details
  - instrument details
  - software packages and parameters
  - quality metrics
  - publication details

```
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
Cl1 Cl 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
Cl2 Cl 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

Data Integrity: checkCIF



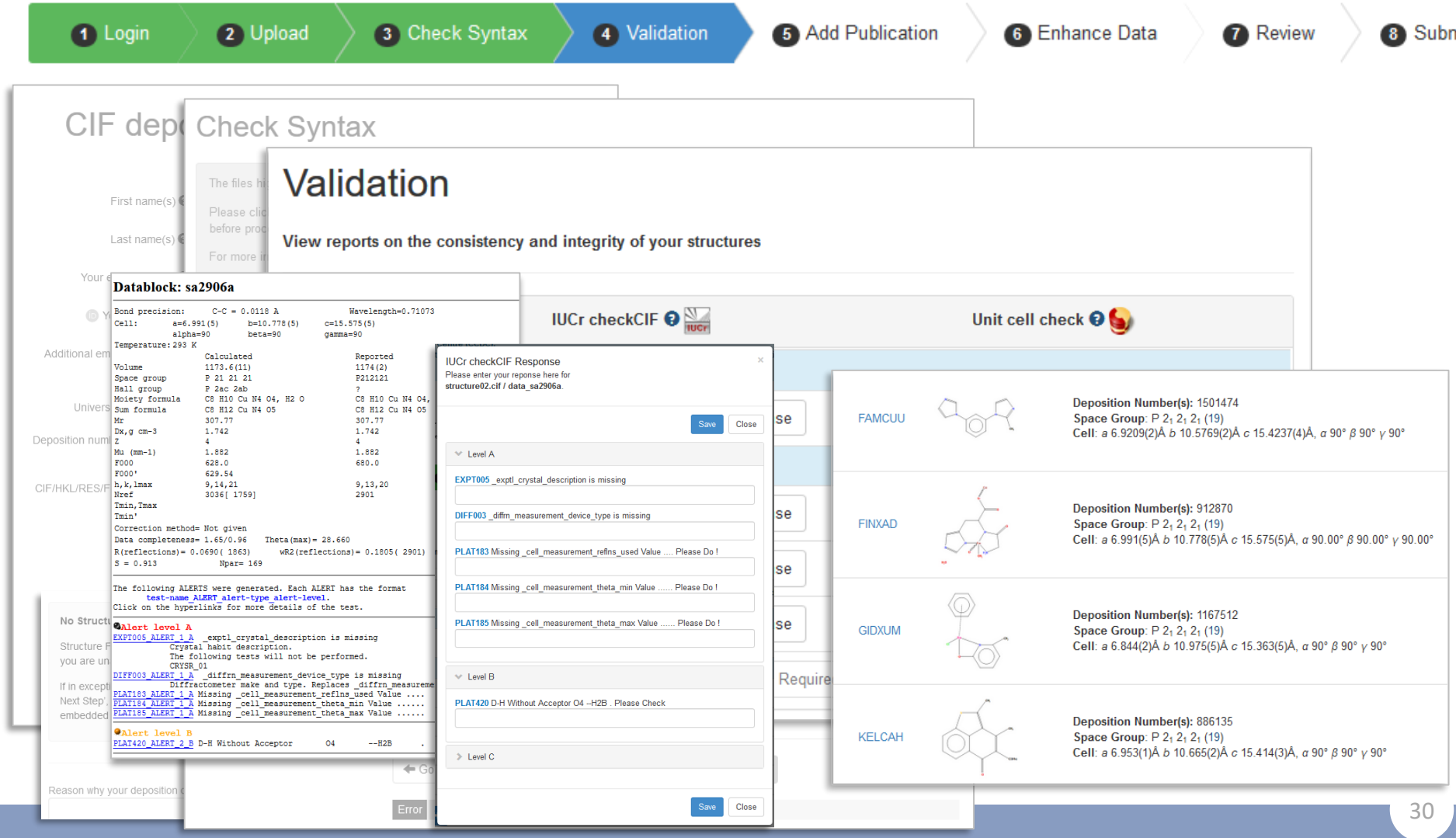
- **Checks consistency and integrity of the data**
- **Generates alerts that should either be corrected or explained**

Level A	Most likely a serious problem - resolve or explain
Level B	A potentially serious problem, consider carefully
Level C	Check. Ensure it is not caused by an omission or oversight
Level G	General information/check it is not something unexpected

checkCIF is sponsored by

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

CCDC Deposition Services



CCDC Deposition Services

1 Login | 2 Upload | 3 Check Syntax | 4 Validation | 5 Add Publication | 6 Enhance Data | 7 Review | 8 Submit

Add Publication

Please check and add/update the publication details shown below.

If you don't know the full publication details then please provide the current list of authors fo

Authors *	Nicola Lavery, Arthur Smith
Journal name	
Volume	Volume
Year	Year
Page	Page
Publication DOI	E.g. 10.14469/hpc/2300
Additional information	

If you do not intend to publish your data in the scientific literatu
immediately through the Cambridge Structural Database (CSD
(ICSD) then please click the 'Publish in a Database' button bel
published in the CSD as a *CSD Communication*. Inorganic dat
Communication.

Publish in a Database

Add Crystallographer Details

Please add the details of the main crystallographer associated with the data below. The email address will be used to notify the crystallographer about this deposition. The name, affiliation, country and if appropriate ORCID iD of the crystallographer may be displayed to users alongside the data.

CSD Communications

The CCDC allows you to publish data directly through the CSD as a *CSD Communication*. Over 5,000 structures were published in this way in 2016 making *CSD Communications* the No. 1 place to publish crystal structures.

Recognition

A citable DOI allows you to receive credit for your structures and add the data to you ORCID record.

If you would like to reference a CSD Communication then we would recommend using the following style of citation

"D.S.Cati, H.Stoeckli-Evans, CCDC 227635: *CSD Communication*, 2004, DOI: 10.5517/cc7mw2s"

Discoverability

Automatic linking via CCDC DOI from third-party repositories, such as DataCite, the Web of Science Data Citation Index and may also be linked from ChemSpider, PubChem and PDB Chemical Component Dictionary

Identifiers and Data Citation

DC1 Data Citation Principles

Data should be considered legitimate, citable products of research…

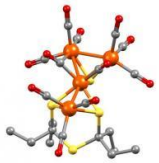
https://www.force11.org/datacitation

Dataset Publication

CCDC 610092: Experimental Crystal Structure Determination. **A. Crystallographer**, *Cambridge Crystallographic Data Centre* (2007)
http://dx.doi.org/10.5517/ccngvdb

DataCite
FIND, ACCESS, AND REUSE DATA

- The CCDC registers DOIs for datasets through DataCite
- Metadata for CCDC datasets is openly accessible via DataCite
- Foundation for interoperability and formalising data citation



10.5517/CCPHZ37

ORCID IDs for Researchers

At least 30% of current CSD depositors provide an ORCID ID

Andrew Bond

ORCID ID

https://orcid.org/0000-0002-1744-0489

Linking Researchers, Datasets and Articles

The Cambridge Crystallographic Data Centre

CSD Entry: GIGMIV

Results

Database Identifier	Deposition Number
GIGMIV	1852044
GIGMOB	1852045

Download

GIGMIV : bis(1,4,7,10-tetraoxacyclododecane)-lithium 3-methyl-1,2-benzodiphosphol-1-ide
Space Group: P $\bar{1}$ (2), **Cell:** *a* 8.4645(4)Å *b* 12.2344(6)Å *c* 14.0658(7)Å, *α* 95.494(3)° *β* 95.567(4)° *γ* 110.021(3)°

Additional details

Deposition Number	1852044
Data Citation	Lily S. H. Dixon, Schirin Hanf, Jessica E. Waters, Andrew D. Bond, Dominic S. Wright CCDC 1852044: Experimental Crystal Structure Determination, 2018, DOI: 10.5517/ccdc.csd.cc2056cs
Deposited on	27/06/2018

Data Citation with Dataset DOI

Crystallographer(s)

Crystallographer	Andrew Bond
Affiliation	University of Cambridge

Link to Crystallographer's ORCID Record

Link to article based on article DOI

Associated publications

Lily S. H. Dixon, Schirin Hanf, Jessica E. Waters, Andrew D. Bond, Dominic S. Wright, *Organometallics*, 2018, 37, 4465, DOI: 10.1021/acs.organomet.8b00480

CCDC Deposition Services

1 Login → 2 Upload → 3 Check Syntax → 4 Validation → 5 Add Publication → 6 Enhance Data → 7 Review → 8 Submit

Add Publication

Please check and add/update the publication details shown below.

If you don't know the full publication details then please provide the c

Authors * Nicola Lavery, Arthur Smith

Journal name

Volume

Year

Page

Publication DOI E.g. 10.14469/hpc/2300

Additional information

If you do not intend to publish y immediately through the Cambr (ICSD) then please click the 'Pu published in the CSD as a CSD Communication.

Publish in a Database

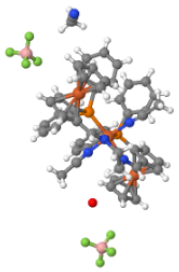
Enhance Data

Pick a structure to edit

structure01.cif
- data_I

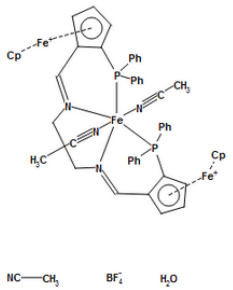
structure02.cif
- data_sa2906c
- data_sa2906a
- data_sa2906b
- data_sa2906g

3D viewer



JSmol

Chemical diagram



data_sa2906c

```
#  For further information on the CCDC, data deposition and
#  data retrieval see:
#                          www.ccdc.cam.ac.uk
#
#  Bona fide researchers may freely download Mercury and enCIFer
#  from this site to visualise CIF-encoded structures and
#  to carry out CIF format checking respectively.
#
data_sa2906c


_audit_creation_method           SHELXL-97
_chemical_name_systematic
;
Bis(acetonitrile-\kN)-{(2Rp,2''Rp)-1,1''-
[1,2-ethanediylbis(nitrilomethylidyne)]bis[2-(diphenylphosphino)-ferrocene]-
\k^4^N,N',P,P'}-iron(ii) tetrafluoroborate acetonitrile water solvate 1/0.5
;
_chemical_name_common            '1704 MB209 at 200K'
_chemical_melting_point          ?
_chemical_formula_moiety
'C52 H48 Fe3 N4 P2, 2(B F4), C2 H3 N, 0.5(H2 O)'
_chemical_formula_sum            'C54 H52 B2 F8 Fe3 N5 O0.50 P2'
_chemical_formula_weight         1182.12

loop_
_atom_type_symbol
```

Associated DOIs

Raw data DOI

Data fields

Compound name: Bis(acetonitrile-\kN)-{(2Rp,2"Rp)-1,1"-[1,2-ethanediylbis(nit

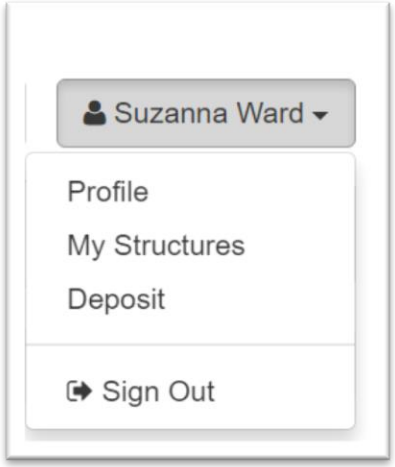
Synonyms/other names

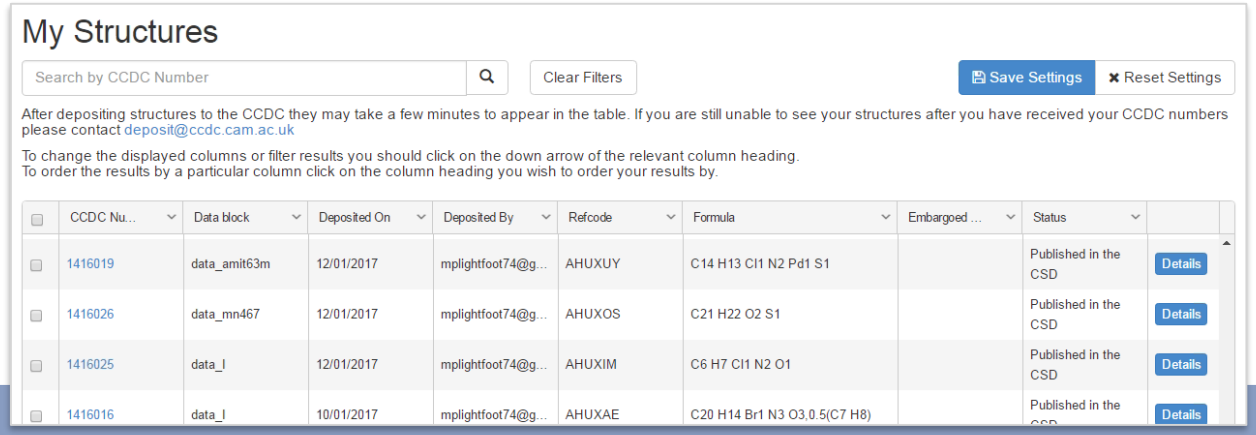
Crystal colour

Crystal habit

My Structures

- **Personalised deposition portal launched December 2016**
  - Ability to log on and view and retrieve depositions
  - Deposit and revise data
  - Edit and update basic information
  - Publish data directly as a *CSD Communication*
  - Share data with co-workers





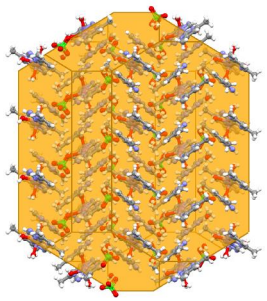
Over 15,000 registered users

Enablers of FAIR Crystallography

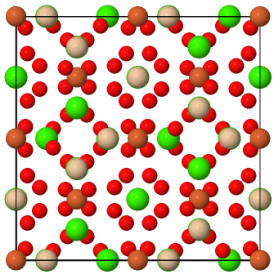
- **Standard File Formats and Vocabularies**
  - ❑ Crystallographic Information File (CIF)
  - ❑ CIF Dictionaries (vocabularies)
  - ❑ checkCIF Data Validation Service
- **Standard Identifiers – disambiguation and interoperability**
  - ❑ Digital Object Identifiers – for articles and data
  - ❑ ORCID iDs – for researchers
- **Trusted searchable data repositories**
  - ❑ Cambridge Structural Database
  - ❑ Inorganic Crystal Structure Database
  - ❑ Protein Data Bank

Crystal Structure Databases

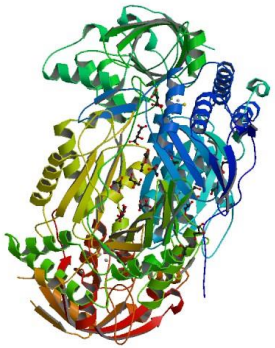
- **Cambridge Structural Database**
  - Organic and Metal-organic compounds
  - >970,000 structures
  - Established in 1965



- **Inorganic Crystal Structure Database**
  - Inorganic compounds
  - >200,000 structures
  - Established in 1978



- **Protein Data Bank**
  - Biological macromolecules
  - >140,000 structures
  - Established in 1971



Free, unified deposition and access of crystal structure data

– July 12, 2018 https://www.ccdc.cam.ac.uk/News/List/2018-07-new-joint-services/

The Cambridge Crystallographic Data Centre (CCDC) and FIZ Karlsruhe – Leibniz Institute for Information Infrastructure (FIZ Karlsruhe) today announced the launch of their joint deposition and access services for crystallographic data across all chemistry. These services will enable researchers to share data through a single deposition portal and explore all chemical structures for free worldwide.

The Cambridge Crystallographic Data Centre

FIZ Karlsruhe
Leibniz Institute for Information Infrastructure

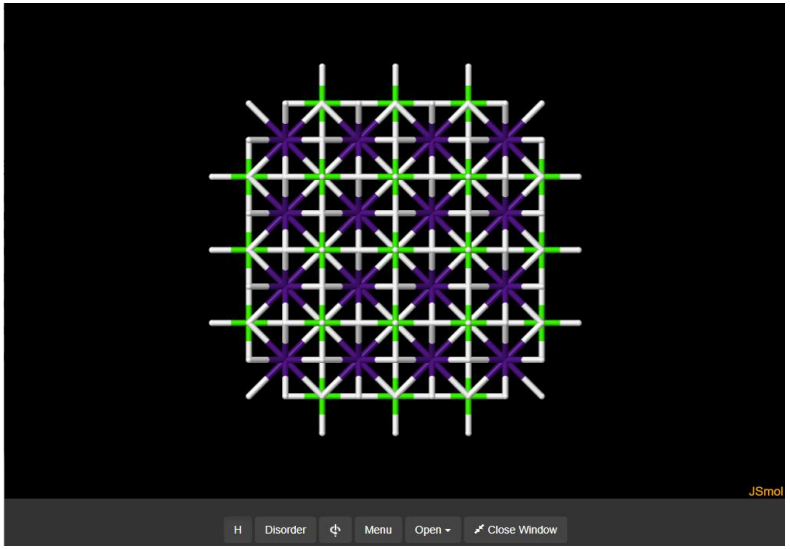
ICSD Entry: 261359

Results

	Database Identifier	Deposition Number
☑	ICSD 261358	1719195
☑	ICSD 261359	1719196
☑	ICSD 261889	1719655
☑	ICSD 261890	1719656
☑	ICSD 422363	1733975
☑	ICSD 422410	1734003
☑	ICSD 422411	1734004

Download

ICSD 261359 : ICSD Structure : H_3 Ca Cs
Space Group: P m $\bar{3}$ m (221), **Cell:** *a* 4.54Å *b* 4.54Å *c* 4.54Å, *α* 90° *β* 90° *γ* 90°



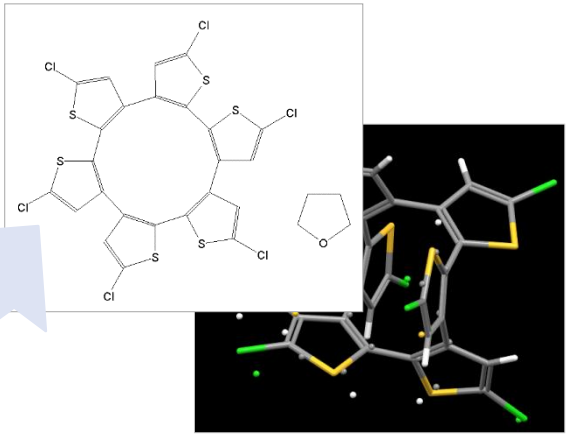
- New Joint Deposition and Access Portal for Organic and Inorganic crystal structures
- CCDC's infrastructure was adapted to accommodate inorganic structures from FIZ
- Implementation greatly aided by use of a common standard format (CIF)

The CSD: Crystallography and Chemistry

```
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
Cl1 Cl 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
Cl2 Cl 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
```

Assignment of chemistry is required to make data findable, interoperable and reusable

- A reliable chemical representation is essential for enabling reuse and application of crystallographic data
- Representation is generated at CCDC using a combination of automated processes and manual validation



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

A representation of chemistry can be included in a CIF but in deposited files this is rarely found

Chemistry Data Initiatives

Enablers of FAIR Chemistry

- **Technical Enablers**
  - ❑ Standard Identifiers (InChI)
  - ❑ Open File Formats (Structures)
  - ❑ Standard File Formats (Spectra)
  - ❑ Terminologies/Vocabularies

- **Social Enablers**
  - ❑ Domain Data Activities
  - ❑ General Data Initiatives

Linking Crystal Structures to PubChem

PubChem: A database of chemical molecules and their activities against biological assays

InChI Key: WHGYBXFWUBPSRW-FOUAGVGXSA-N

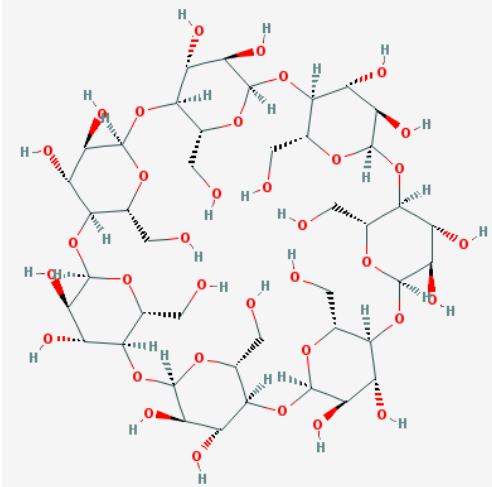
PubChem OPEN CHEMISTRY DATABASE

Compound Summary for CID 444041

PUBCHEM > COMPOUND > BETA-CYCLODEXTRIN

beta-CYCLODEXTRIN

Vendors | Pharmacology | Literature | Patents | Bioactivities



4.3 Crystal Structures

Crystal Structures: 1 of 1	
CCDC Number	762697
Crystal Structure Data	DOI:10.5517/cctln45
Associated Article	DOI:10.1039/C3CE26414A

from The Cambridge Structural Database

Linking Crystal Structures to PubChem

PubChem: A database of chemical molecules and their activities against biological assays

PubChem | OPEN CHEMISTRY DATABASE

Compound Summary for CID 444041

PUBCHEM > COMPOUND > BETA-CYCLODEXTRIN

beta-CYCLODEXTRIN

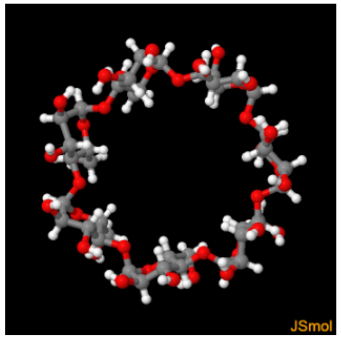
Vendors | Pharmacology | Literature | Patents | Bioactivities

4.3 Crystal Structures

Crystal Structures: 1 of 1	
CCDC Number	762697
Crystal Structure Data	DOI:10.5517/cctln45
Associated Article	DOI:10.1039/C3CE26414A

WEWTOJ : 5,10,15,20,25,30,35-heptakis(hydroxymethyl)-2,4,7,9,12,14,17,19,22,24,27,29,32,34-tetradecaoxaoctacyclo[31.2.2.2^{3,6}.2^{8,11}.2^{13,16}.2^{18,21}.2^{23,26}.2^{28,31}]nonatetracontane-36,37,38,39,40,41,42,43,44,45,46,47,48,49-tetradecol
Space Group: C2, Cell: *a* 19.056(5)Å *b* 24.415(6)Å *c* 15.698(4)Å, *α* 90.00° *β* 109.463(13)° *γ* 90.00°

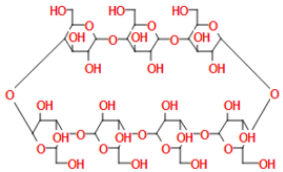
3D viewer



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 Citation - A.I.Ramos,T.M.Braga,P.Silva,J.A.Fernandes,P.Ribeiro-Claro,M.de F.S.Lopes,F.A.A.Paz,S.S.Braga, CCDC 762697: Experimental Crystal Structure Determination, 2013, DOI: 10.5517/cctln45
Deposited on: 20/1/2010

Associated publications

A.I.Ramos,T.M.Braga,P.Silva,J.A.Fernandes,P.Ribeiro-Claro,M.de F.S.Lopes,F.A.A.Paz,S.S.Braga, *CrystEngComm*, 2013, 15, 2822, DOI: 10.1039/C3CE26414A

from The Cambridge Structural Database

Linking enabled by InChI

PubChem OPEN CHEMISTRY DATABASE

Bianthrone

Vendors | Literature | Patents | Bioactivities

4.2 Crystal Structures

Crystal Structures: 1 of 16	
CCDC Number	247865
Crystal Structure Data	DOI:10.5517/cc89xn3

from The Cambridge Structural Database

Crystal Structures: 2 of 16	
CCDC Number	299197
Crystal Structure Data	DOI:10.5517/ccb1bj6
Associated Article	DOI:10.1021/jp061205i

from The Cambridge Structural Database

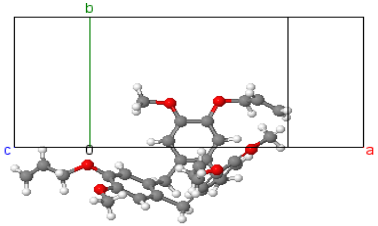
ROYAL SOCIETY OF CHEMISTRY

ChemSpider
Search and share chemistry

Names and identifiers | Properties | Searches | Spectra | Crystal CIFs | Pharma Links | More

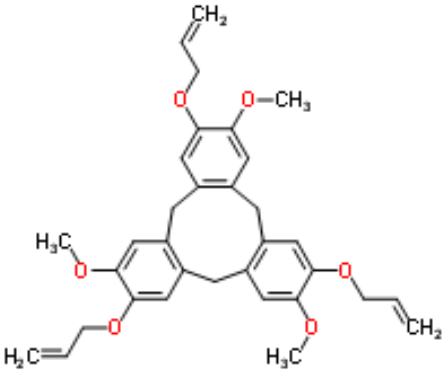
Associated Hyperlink: http://dx.doi.org/10.5517/cc4zfp6
Comments: Structure CCDC 148418 from the Cambridge Structural Database reported in RSC article http://dx.doi.org/10.1039/b000825g
Unit cell: a=14.4429(5)Å, b=8.0609(3)Å, c=24.3908(7)Å, alpha=90.00°, beta=99.510(2)°, gamma=90.00°, T=123(2)K, space group P21/n, Z=4
Submitted by: antony.williams

HM:P21/n
a=14.443Å
b=8.061Å
c=24.391Å
α=90.000°
β=99.510°
γ=90.000°



IUPAC

InChI TRUST



Standard InChI:
InChI=1S/C33H36O6/c1-7-10-37-31-19-25-13-23-17-29(35-5)33(39-12-9-3)21-27(23)15-24-18-30(36-6)32(38-11-8-2)20-26(24)14-22(25)16-28(31)34-4/h7-9,16-21H,1-3,10-15H2,4-6H3

Standard InChIKey:
IZHKSTHBLQRIOW-UHFFFAOYSA-N

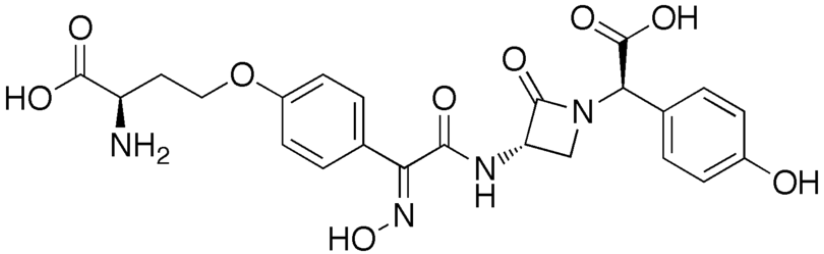
The IUPAC International Chemical Identifier

The IUPAC International Chemical Identifier (InChI™) is a non-proprietary identifier for chemical substances that can be used in printed and electronic data sources thus enabling easier linking of diverse data compilations.

- **Initially developed through an IUPAC project from 2000-2004**
- **Development now overseen by the InChI Trust**
- **IUPAC still involved in the scientific direction**
- **Development driven through IUPAC and other Task Groups**

https://www.inchi-trust.org/

Anatomy of an InChI



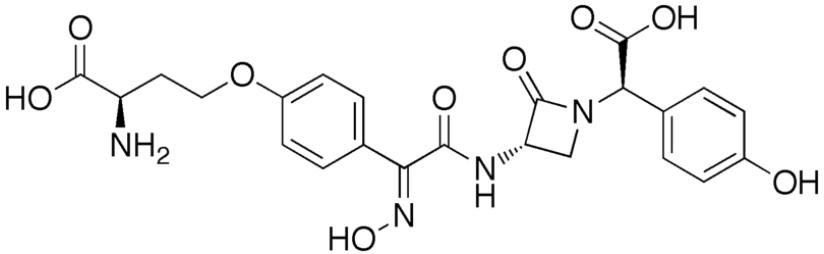
Nocardicin A

From Wikipedia, the free encyclopedia


```
InChI=1S/C23H24N4O9/c24-16(22(31)32)9-10-36-15-7-3-12(4-8-15)18(26-35)20(29)25-17-11-27(21(17)30)19(23(33)34)13-1-5-14(28)6-2-13/h1-8,16-17,19,28,35H,9-11,24H2,(H,25,29)(H,31,32)(H,33,34)/b26-18+/t16-,17+,19-/m1/s1
```


- **InChI layers**
  - Main layer
    - formula
    - connections
    - H atoms
  - Atom and Bond Stereochemistry
  - Isotopic and Fixed Hydrogens (tautomerism)
  - Charge

InChIKeys



Nocardicin A
From Wikipedia, the free encyclopedia

InChIKey=CTNZOGJNVIFEBA-TWTPMLPMSA-N

a hashed version of the full standard InChI

- **InChI layers**
  - Main
    - formula
    - connections
    - H atoms
  - Deprotonation indicator (related to charge)
  - Stereo
  - Isotopic and Fixed hydrogens
  - Standard/non-standard, version number

InChIs can be converted back to structures. InChIKeys cannot.

Millions of InChI and InChIKeys have been generated for structures in e.g. PubChem, ChEMBL, ChemSpider

What's in the CSD?

Organic
43%

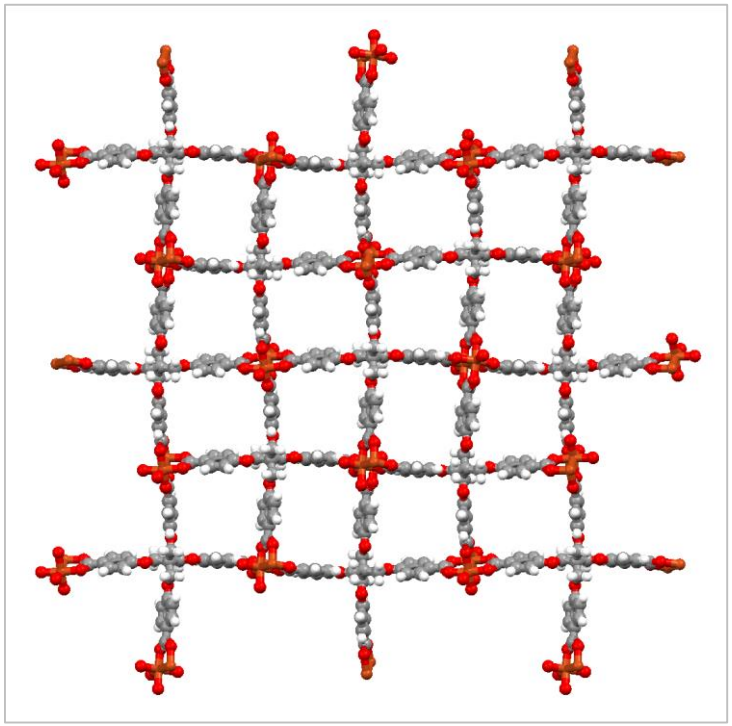
Not Organic
57%
At least one transition metal, lanthanide, actinide or any of Al, Ga, In, Tl, Ge, Sn, Pb, Sb, Bi, Po

Not Polymeric
89%

Polymeric: 11%

Single Component
56%

Multi Component
44%

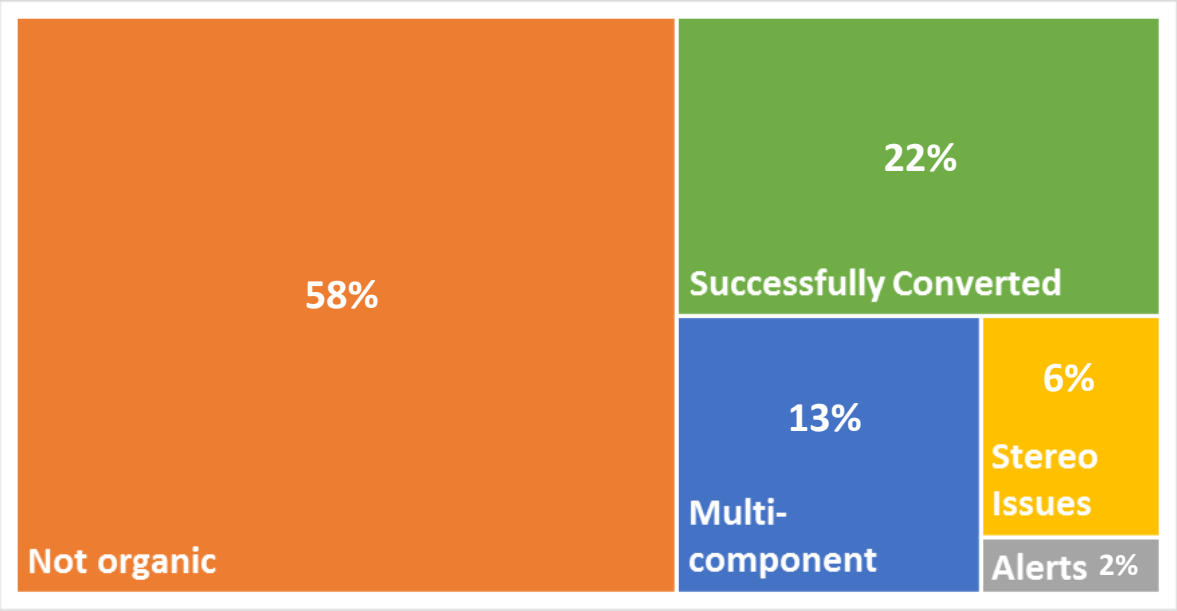


EPOTAF: CCDC 779539 doi:10.5517/ccv55fl

$(C_{33}H_{28}Cu_2O_{14})_n$, $7n(C_3H_7NO)$, $3n(C_4H_8O_2)$, $n(CH_4O)$

InChI Challenges

Generation of reliable InChIs for chemical substances in the CSD



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

Order of filtering entries out:

- Not organic
- Multi-component
- InChI alerts
- Stereochemistry Issues

- Can confidently generate InChIs for ~22% of CSD entries
- If multi-component entries included then ~35% assuming no other issues
- If based on just organic compounds then 52% (up to 82% including multi-component)

InChI Workshop, Cambridge UK, 4-5 February 2019

Organometallics	Variability/uncertainty, e.g.
Mixtures	- Tautomers
Reactions	- Stereochemistry
QR Codes	- Positional Isomers
Educational Resources	- etc.

InChI Chemical Data Standard: Identifiers and Extensions

Professor Jonathan Goodman: InChI Champion

Wed 13 February 2019, 14:30 - 15:30, Room U203 at the Department of Chemistry. Book at https://www.training.cam.ac.uk/chem/event/2855805 or email cmc32@cam.ac.uk.

InChI, InChIKeys, Reactions, Mixtures, QR Codes and more…

Chemistry Structure File Formats

- InChIs are generated using a standard InChI Generator supplied by the InChI Trust
- To reliably generate InChIs a **reliable digital representation of a chemical structure** must be supplied as input

- There are **many different file formats** that aim to electronically represent a chemical structure
- Some are ubiquitously used and can be considered ***de facto* standards** – e.g. MOL/SDF, SMILES
- There are ways of **inter-converting** between different file formats

Open Babel: a chemical toolbox designed to speak the many languages of chemical data

- Common cheminformatics formats
  - Canonical SMILES format (can)
  - Chemical Markup Language (cml, mrv)
  - InChI format (inchi)
  - MDL MOL format (mol, mdl, sdf, sd)
  - Protein Data Bank format (pdb, ent)
  - SMILES format (smi, smiles)
  - Sybyl Mol2 format (ml2, sy2, mol2)
- Other cheminformatics formats
  - Accelrys/MSI Biosym/Insight II CAR format (arc, car)
  - Accelrys/MSI Cerius II MSI format (msi)
  - Accelrys/MSI Quanta CSR format (csr)
  - MCDL format (mcdl)
  - MSI BGF format (bgf)
  - PubChem format (pc)
- Computational chemistry formats
  - ADF cartesian input format (adf)
  - ADF output format (adfout)
  - CAChe MolStruct format (cache, cac)
  - CASTEP format (castep)
  - Cacao Cartesian format (caccrt)
  - Cacao Int
  - DMol3 co

Reads, writes and converts over 110 chemical file formats

Reliable Input Representations

- **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
```

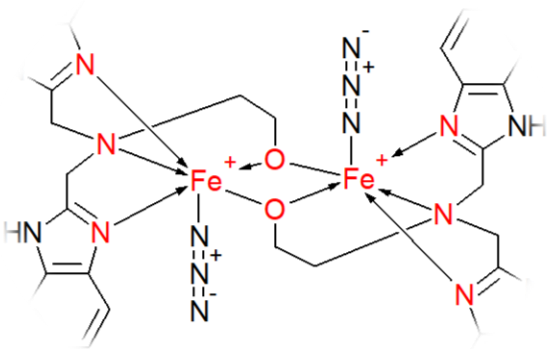
* excluding query bond types

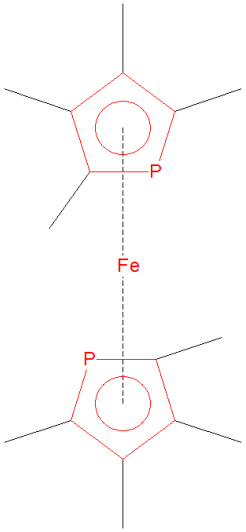
PubChem SDF
PUBCHEM_NONSTANDARDBOND

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

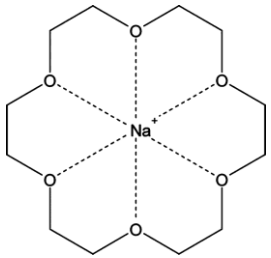
ACD/Labs MOL V2000 Extensions

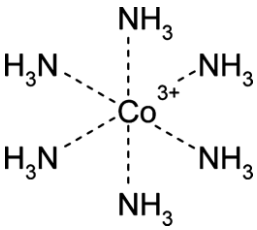
```
M  ZZF  3   1  41   2  42   3  43
M  ZZH   1  5   2   3   4   5   6
M  ZZH   2  5   7   8   9  10  11
M  ZZH   3  5  12  14  15  16  17
M  ZZE  2  42  18  43  18
```

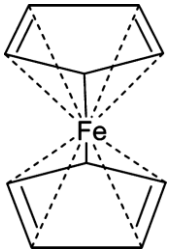




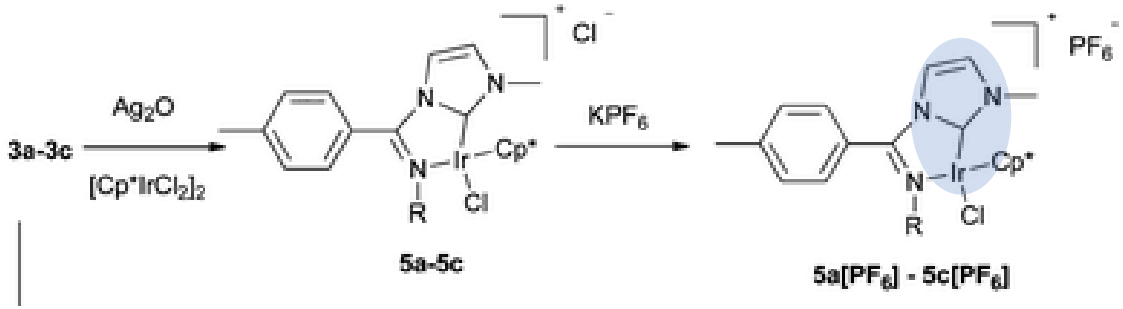
Accurate Specification of Molecular Structures: The Case for Zero-Order Bonds and Explicit Hydrogen Counting. **Alex M. Clark.**
J. Chem. Inf. Model., 2011, 51 (12), 3149. doi:10.1021/ci200488k



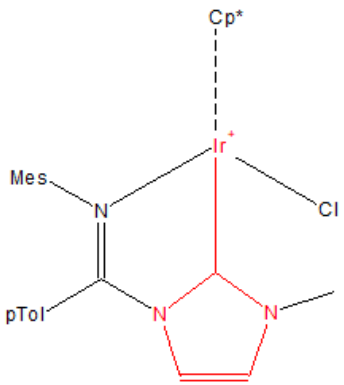




Consistent Structure Representation



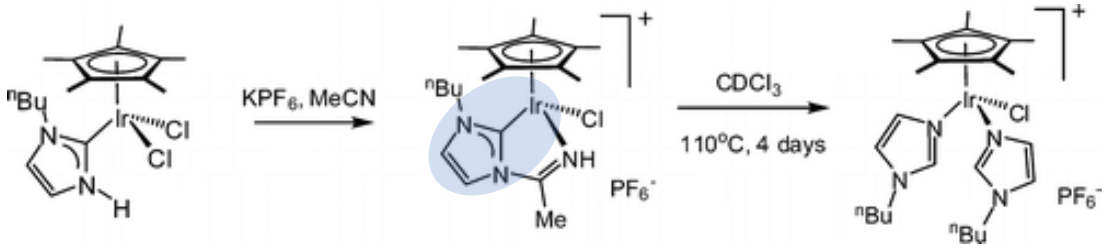
Dalton Trans., 2012,41, 14557-14567, doi:10.1039/C2DT31989F



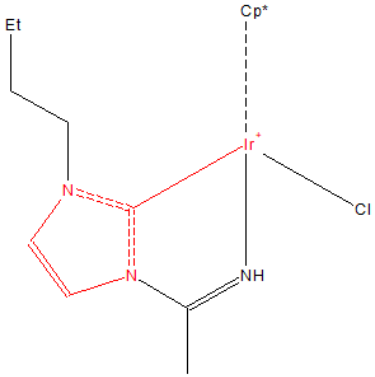
CSD

85%

ECIWUK CCDC:872879 10.5517/ccy99dx



Organometallics 2007, 26, 18, 4684-4687, doi:10.1021/om700498w



12%

LIMXAH CCDC:664254 10.5517/ccq96kr

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

Publishing Chemical Structures

Fragment-Based Approach to Targeting Inosine-5′-monophosphate Dehydrogenase (IMPDH) from *Mycobacterium tuberculosis*

Article Options

Supporting Information

Fragment-Based Approach to Targeting Inosine-5′-monophosphate Dehydrogenase

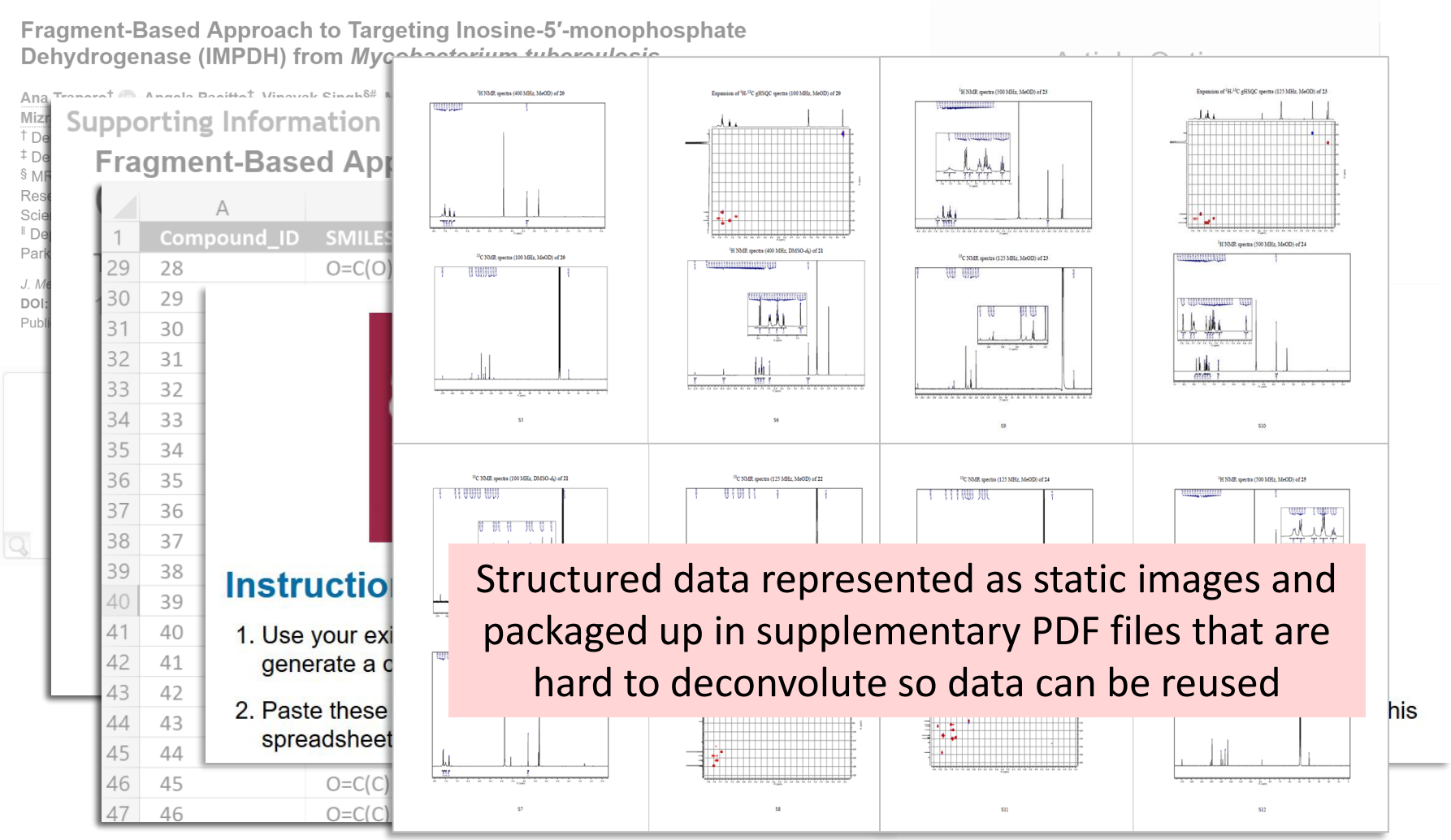
	A	B	C
1	Compound_ID	SMILES	Mth IMPDH IC50 (uM)
29	28	O=C(O)CSC1=NC(C2=CC=CC=C2)=CN1	–
30	29		
31	30		
32	31		
33	32		
34	33		
35	34		
36	35		
37	36		
38	37		
39	38		
40	39		
41	40		
42	41		
43	42		
44	43		
45	44		
46	45	O=C(C)NC1=NC(C2=CC=C(Br)C=C2)=CN1	–
47	46	O=C(C)NC1=NC(C2=CC=CC=C2)=CN1	–

https://pubs.acs.org/page/jmcmar/submission/jmcmar_mfstrings.html

Instructions for Authors

1. Use your existing chemical drawing programs (e.g., ChemDraw, ACD ChemSketch, Marvin Sketch) to generate a computer-readable SMILES formula for each compound presented in your article.
2. Paste these formulas into the spreadsheet template, along with basic information about each compound. This spreadsheet will provide a machine-readable version of the key data presented in the article's tables.

Publishing Chemical Spectra

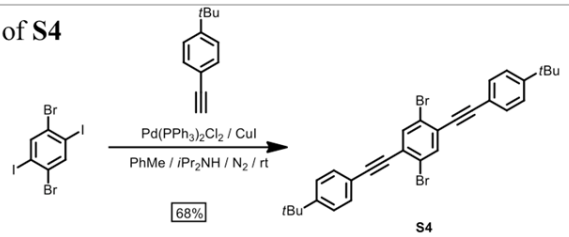


Structured data represented as static images and packaged up in supplementary PDF files that are hard to deconvolute so data can be reused

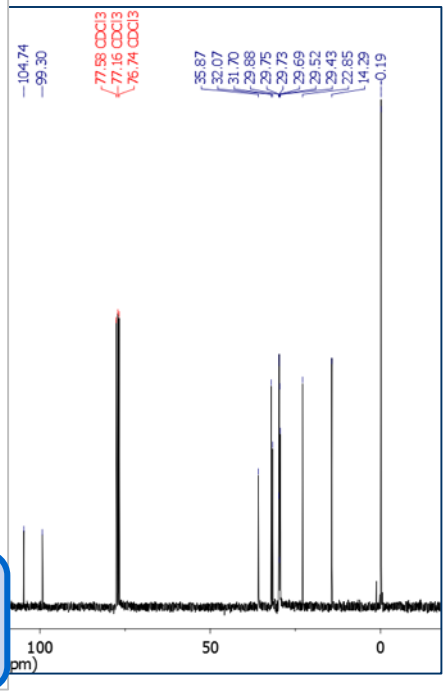
Chemistry Data Publication Workflow (simplified)

Chemical Sample → Spectrometer → Raw Data → Analysis Software → Processed Data → Supplemental Information → Article → Human Readers

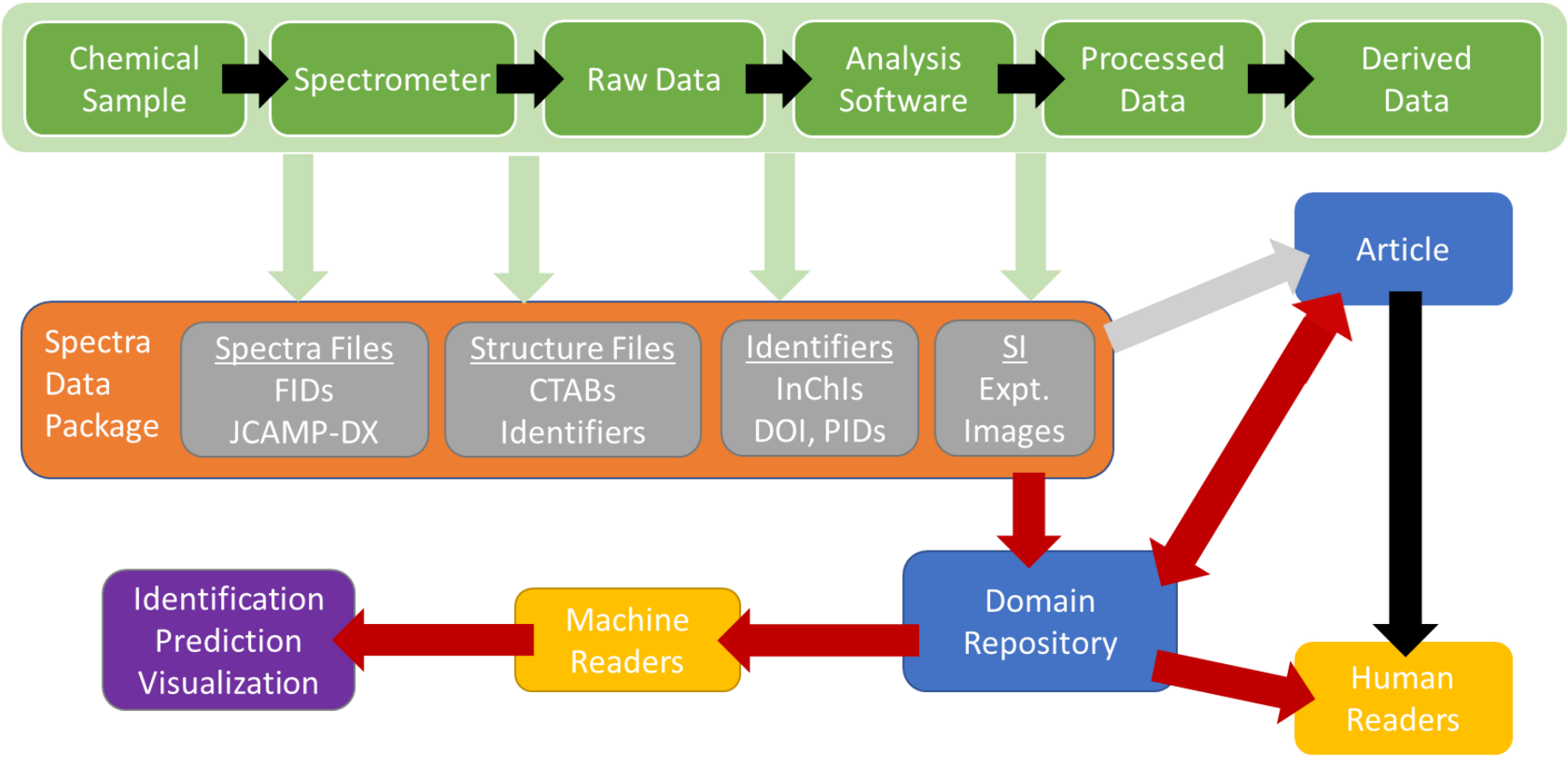
Scheme S6. Synthesis of **S4**



Synthesis of S4: Anhydrous PhMe (31 mL) and freshly distilled *i*Pr_2NH (10 mL) were added to a 100 mL flask and sparged with N_2 for 20 min. 1,4-dibromo-2,5-diiodobenzene (3.000 g, 6.15 mmol), 1-(*tert*-butyl)-4-ethynylbenzene (2.433 g, 15.38 mmol), $Pd(PPh_3)_2Cl_2$ (216 mg, 0.308 mmol) and CuI (117 mg, 0.615 mmol) were added to the solution, in sequence. The mixture was stirred at rt for 14 h. The crude reaction mixture was filtered through a pad of SiO_2 gel and washed with additional CH_2Cl_2 (200 mL). Evaporation of the solvent provided the crude product as a solid, which was purified by chromatography (SiO_2, hexanes) to provide **S4** (2.310 g, 68% yield) as a white powder. ^1H and ^{13}C NMR spectroscopy were consistent with the report of Hseuh *et al.* **S4:** ^1H NMR (300 MHz, $CDCl_3$) δ 7.77 (s, 2H), 7.58 – 7.47 (m, 4H), 7.46 – 7.35 (m, 4H), 1.34 (s, 18H). ^{13}C NMR (75 MHz, $CDCl_3$) δ 152.68, 136.05, 131.70, 126.58, 125.63, 123.81, 119.46, 97.03, 86.49, 77.58, 77.16, 76.74, 35.06, 31.30.



Chemistry Data Publication Workflow (an alternative)



Standards for Spectra: JCAMP-DX

http://www.jcamp-dx.org/

- JCAMP-DX is a standard for spectroscopic data developed in the 1980s

> **JCAMP-DX is a standard file form for exchange of infrared spectra and related chemical and physical information between spectrometer data systems of different manufacture, main-frame time-sharing systems, general purpose lab computers, and personal computers. It is compatible with all media: telephone, magnetic and optical disk, magnetic tape, and even the printed page (via optical reader).**

JCAMP-DX: A Standard Form for Exchange of Infrared Spectra in Computer Readable Form
Applied Spectroscopy (1988)

- It has undergone a number of enhancements to address new instrumentation needs, as well as extensions to new spectroscopic methods.
- Vendors have added custom extensions for their own instruments; new metadata standards, such as ORCIDs and InChIs have been developed.
- Working group of IUPAC Subcomittee on Cheminformatics Data Standards looking at updating the standard

Open Data FAQs for chemists

https://www-library.ch.cam.ac.uk/open-data-faqs-chemists

The following FAQs have been asked by members of the Department of Chemistry and answered by members of the Open Data team at the University.

If you have any amendments or further questions you would like to ask please contact the Librarian at the Department of Chemistry, Clair Castle, at library@ch.cam.ac.uk, in the first instance.

The Open Data team can also be contacted at info@data.cam.ac.uk, http://www.data.cam.ac.uk/.

FAQs

What would open data for a typical synthetic organic chemistry paper look like?

For a synthetic paper you might include the output files from NMR, UV/Vis, and IR measurements (for example). These should be in a format that others can use, so the data should
experiments, wouldn't meet this criteria). So f
example) so that future users can replot the

Lab books form an important record of the ex
least the detailed methodology for the experi

Output files for NMR, UV/Vis, IR etc. should be in formats that others can use – images of graphs, especially of NMR experiments, don't meet this criteria

However, if it would be too time consuming and costly to digitise the lab books then you can simply create a meta-data record on the repository so that future users can contact you to physically access your lab books.

What would open data for a typical molecular dynamics based paper look like?

For a computational paper you might include the input and output files from the calculations. Whether you need to include binary output files (which are often produced but hardly ever analysed) is left at your discretion, but if you feel that these files are necessary for the interpretation of the results then they should also be included.

If you have performed a whole suite of experiments, all of which are similar, then it might only be necessary to provide the input files and a couple of example output files. Future researchers can then scrutinize a sample of your output and then re-run all your input files if they wish to do so.

Convert your files into an open data format

NMR spectroscopy data from TopSpin

These instructions are for converting NMR spectroscopy data from TopSpin to a text file in the internationally accepted open data format JCAMP-DX (http://www.jcamp-dx.org).

In TopSpin
- File
- Save
- Save data set in a JCAMP-DX file
- OK
- Optional: Change name and directory
- Leave "Type of archive file = JCAMP DIFF/DUP"
- Change "Include these data types =" to "FID+All_PROCNOS"
- Leave "JCAMP version = 6.0"
- OK

Save as JCAMP-DX – with FIDs

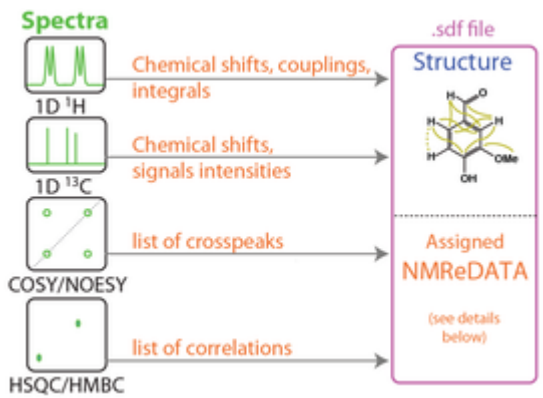
https://www-library.ch.cam.ac.uk/convert-your-files-open-data-format

http://nmredata.org/

NMReDATA initiative

GENERATE, STORE AND SHARE THE DATA EXTRACTED FROM SET OF NMR SPECTRA ASSOCIATED TO A COMPOUND

The goal of the NMReDATA initiative is to improve the FAIRness and quality of the NMR data available to the community.



Important benefits of the new format

- Improved quality of the NMR data
- Easier inclusion of NMR data in reports and articles
- Simplified referee work
- Compatibility with electronic storage in database
- Easier comparison of dataset
- Improved searchability of NMR data

Mnova Documentation Toolkit – Mpublish

- Researchers with access to Mnova can zip up publication quality images with raw data
- Publisher can digitally sign the zip file so data can be viewed by reviewers and readers without the need for a Mnova licence

Mestrelab Research

http://resources.mestrelab.com/documentation-toolkit-project-mpublish/

FAIR Publishing Guidelines for Spectral Data and Chemical Structures in Support of Chemistry and Related Disciplinary Communities

funded by NSF, Orlando FL, March 2019

WORKSHOP GOALS:

1. **Workflow**: develop digital data publishing model across stakeholders
2. **Guidelines**: formulate consistent guidelines for publishing FAIR chemical data for common data types
3. **Value Proposition**: review re-use cases for chemical characterization data
4. **Coalition**: initiate process for ongoing coordination and stakeholder engagement

Publishers • Databases • Repositories • Software Developers • Researchers • Librarians
Standards Organisations • Data Initiatives

Chemistry Publication Guidelines

What data should be published? How should it be validated? Where should it be stored?

- Idea to survey journal requirements for chemistry data initiated at an **IUPAC/RDA workshop**
- Sampling of requirements undertaken by **Vin Scalfani**, University of Alabama Libraries

- Discussion of survey at an **session of the ACS Division of Chemical Information**
- Prompted an **IUPAC taskforce** looking at requirements for a publication of spectra
- Issues relating to file formats discussed at **IUPAC/CODATA workshop** on publishing FAIR data
- Workflow discussions to be advanced at and **NSF Workshop** on Publishing Guidelines for Spectral Data and Chemical Structures

IUPAC Chemical Terminology

- **Blue Book**
  - Nomenclature of Organic Chemistry
- **Red Book**
  - Nomenclature of Inorganic Chemistry
- **White Book**
  - Biochemical Nomenclature
- **Orange Book**
  - Analytical Terminology
- **Purple Book**
  - Compendium of Polymer Terminology and Nomenclature
- **Silver Book**
  - Compendium of Terminology and Nomenclature of Properties Clinical Laboratory Sciences
- **Green Book**
  - Quantities, Units and Symbols in Physical Chemistry

“Digital” Chemical Terminology

https://goldbook.iupac.org

- > 7000 terms with authoritative definitions, spanning the whole range of chemistry – with DOIs
- Source documents include *IUPAC Color Books* and recommendations published in *Pure and Applied Chemistry*
- Currently undergoing stabilization and development to provide a foundation for future application

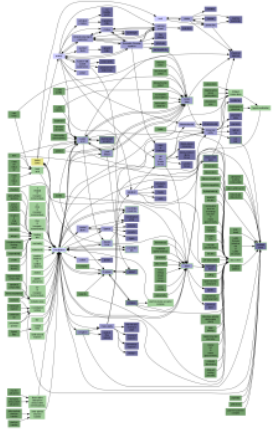
IUPAC GOLD BOOK

dative bond

The coordination bond formed upon interaction between molecular species, one of which serves as a donor and the other as an acceptor of the electron pair to be shared in the complex formed, *e.g*, the N→B bond in $H_3N{\rightarrow}BH_3$. In spite of the analogy of dative bonds with covalent bonds, in that both types imply sharing a common electron pair between two vicinal atoms, the former are distinguished by their significant polarity, lesser strength, and greater length. The distinctive feature of dative bonds is that their minimum-energy rupture in the gas phase or in inert solvent follows the heterolytic bond cleavage path.

Source:
PAC, 1999, *71*, 1919 *(Glossary of terms used in theoretical organic chemistry)* on page 1933

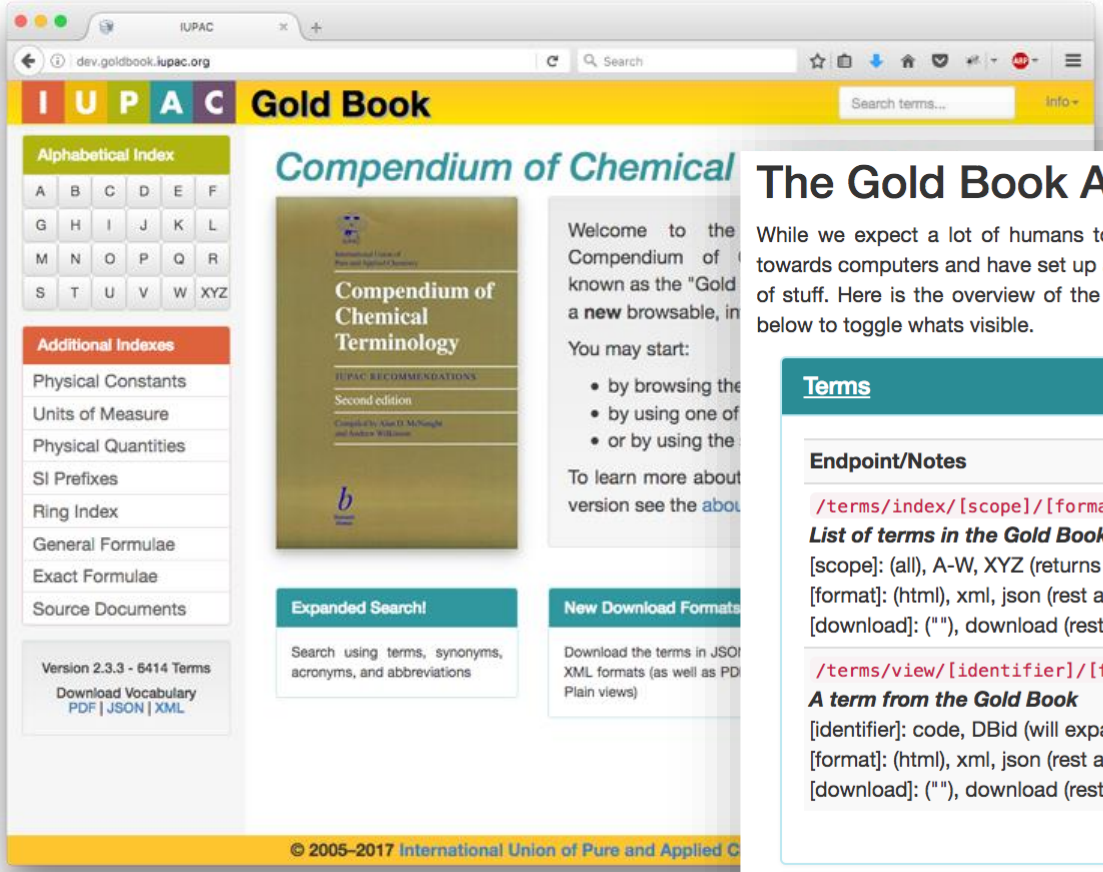
Interactive Link Maps

First Level	Second Level	Third Level
		

Cite as:
IUPAC. Compendium of Chemical Terminology, 2nd ed. (the "Gold Book"). Compiled by A. D. McNaught and A. Wilkinson. Blackwell Scientific Publications, Oxford (1997). XML on-line corrected version: http://goldbook.iupac.org (2006-) created by M. Nic, J. Jirat, B. Kosata; updates compiled by A. Jenkins. ISBN 0-9678550-9-8. https://doi.org/10.1351/goldbook.

Last update: 2014-02-24; version: 2.3.3.

DOI of this term: https://doi.org/10.1351/goldbook.D01523.



The Gold Book API

Alpha API v0.1 (6/30/17)

While we expect a lot of humans to stop by the Gold Book, its about time that the vocabulary be friendly towards computers and have set up an application programming interface (API) so they may download a bunch of stuff. Here is the overview of the API and we are working on additional documentation. (click the headers below to toggle whats visible.

Terms

Endpoint/Notes	Example(s)
`/terms/index/[scope]/[format]/[download]` ***List of terms in the Gold Book*** [scope]: (all), A-W, XYZ (returns to referring page if no data) [format]: (html), xml, json (rest are ignored) [download]: (""), download (rest are ignored)	/terms/index/all (just "terms" works too) /terms/index/C/xml /terms/index/XYZ/json/download
`/terms/view/[identifier]/[format]/[download]` ***A term from the Gold Book*** [identifier]: code, DBid (will expand this) [format]: (html), xml, json (rest are ignored) [download]: (""), download (rest are ignored)	/terms/view/A00001 /terms/view/P04409/json /terms/view/ZT07132/xml/download

Sources (Click to show)

Ack: S. Chalk

Chemistry Data Initiatives

- **Enablers for FAIR Chemistry data**
  - ❑ Standard Identifiers (InChI)
  - ❑ Open File Formats (Structures)
  - ❑ Standard File Formats (Spectra)
  - ❑ Terminologies/Vocabularies
- **Active Communities**
  - ❑ InChI Trust
  - ❑ IUPAC Committees and Subcommittees
  - ❑ ACS Division of Chemical Information (CINF)
  - ❑ RDA Chemistry Research Data Interest Group (CRDIG)
  - ❑ GO-FAIR Chemistry Implementation Network (ChIN)

Chemistry Data Interest Group: DIGChem

https://bit.ly/digchem-activity

DIGChem

- Home
- DIGChem Events
- Publishing Guidelines
- Survey
- WorkflowTools
- JCAMP-DX
- NMR/Spectra Repositories
- OpenStructures
- DataCite Recommendations
- Education
- Professional Training
- Cheminformatics Color Book
- Gold Book Website

The Data Interest Group/Chemistry: DIGChem

Coming in 2018:

- Supporting FAIR Exchange of Chemical Data Through Standards Development, Amsterdam, July 16-17, 2018.
- Activities at ACS/Boston, August 19-23, 2018
- International Data Week, Gaborone, Botswana, November 5-8, 2018

More (Future/Past)

The Data Interest Group/Chemistry is an effort to f
order to accomplish this vision, CRDIG is analyzing
data repositories, evaluating and updating existing
advocating for and educating researchers, librarian
Anyone in the broader chemistry community who is
conjunction with the International Union of Pure ar
Data Standards (SCDS) and the Research Data Allia
Discussions of the interest group have been held at
Information (CINF), at the RSC Chemical Informatic
General Assembly, at RDA Plenaries, and at the Bei

Chemistry Research Data IG

RDA RESEARCH DATA ALLIANCE

IG

Group details

Status: Recognised & Endorsed
Chair (s): Leah McEwen, Stuart Chalk, Ian Bruno, David Martinsen, Richard Kidd

Building the social and technical bridges to enable open data sharing

https://bit.ly/digchem

FAIR Data Initiatives

A GO FAIR Chemistry Implementation Network (ChIN) has been recently endorsed

GO FAIR

Implementation Networks

- The **GO FAIR** movement aims to implement the Internet of FAIR Data and Services
- Related to the European Open Science Cloud (EOSC), involving partners outside of the EU

Pistoia Alliance

Lowering Barriers to Innovation in Life Sciences R&D

Implementation of FAIR Data Principles for Pharma and Life Sciences

Kees van Bochove, The Hyve | 07/06/2018 - 05:54

Initiating | Tags: Standards, Knowledge Management

Unmet Needs:
The advent of ML / AI for pharma is very promising, however without a basic amount of metadata and smart annotation of existing data assets, the algorithms cannot make much headway. Several years of IMI knowledge management projects, experiments with data warehouses, data lakes etc. have made it clear that proper semantic annotation of data assets is a hard and resource intensive but very important hurdle to overcome.

Forthcoming Events

- **InChI Worshop: InChI Opportunities**
  Department of Chemistry, Cambridge, February 4-5, 2019
- **InChI Chemical Data Standard: Identifiers and Extensions**
  Department of Chemistry, Cambridge, February 13, 2019, 14:30 U203
- **FAIR Publishing Guidelines for Spectral Data and Chemical Structures**
  Orlando, March 28-29, 2019
- **IUPAC General Assembly and World Congress**
  Celebrating 100 Years of IUPAC, Paris, July 5-12, 2019
  Special Symposium: Digital Chemistry and the Lab of the Future
- **InChI Symposium**
  San Diego, August 23-24, 2019 (Followed by ACS National Meeting)
  https://www.eventbrite.com/e/inchi-symposium-tickets-52810788490

One Million Crystal Structures: A Wealth of Structural Chemistry Knowledge
Symposium being planned for Fall 2019 ACS Meeting
https://callforpapers.acs.org/sandiego2019/CINF

Summary

- **Chemistry Data Initiatives**
  - ❑ Initiatives aimed at supporting FAIR publication of chemistry data
  - ❑ Motivated by guiding principles arising from global data initiatives
  - ❑ Challenges being actively addressed by a range of community groups
- **Cambridge Crystallographic Data Centre**
  - ❑ Sharing crystallographic data and knowledge since 1965
  - ❑ Supporting FAIR publication and access to crystal structures
  - ❑ Adopting and supporting community data standards
- **Cambridge Structural Database**
  - ❑ The world's repository of small molecule crystallographic data
  - ❑ Providing knowledge and insights applicable across chemistry
  - ❑ Available to you at the Department of Chemistry

The Cambridge Crystallographic Data Centre

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

Scientific Software Provider
Search/analysis/visualisation tools
Scientific applications

Collaborative Research Organisation
New methodologies
Fundamental research

Education and Outreach
Conferences, Workshops,
Training, Teaching

@ccdc_cambridge

ccdc.cambridge

http://www.ccdc.cam.ac.uk/

Enriching Chemistry with Crystallographic Data and Knowledge

THANK YOU!