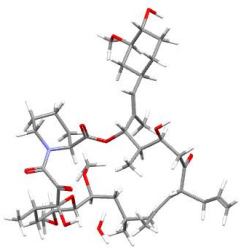
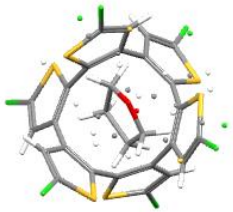
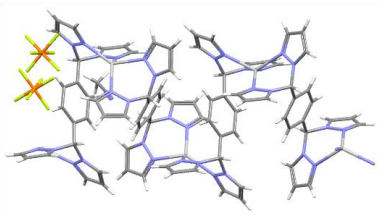
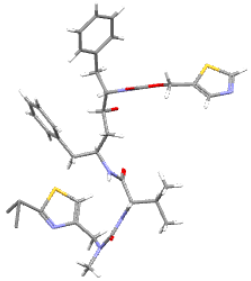
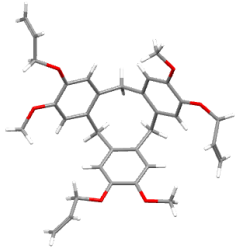
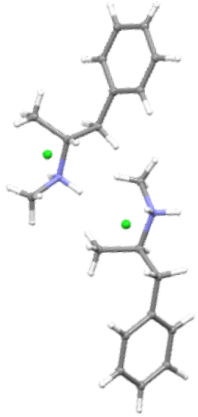
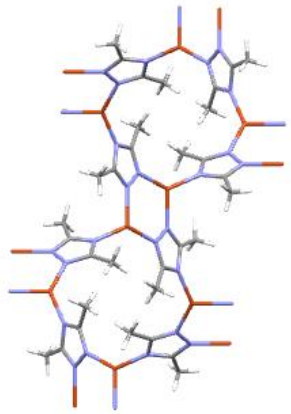
A Journey from Data to Knowledge

Ian Bruno

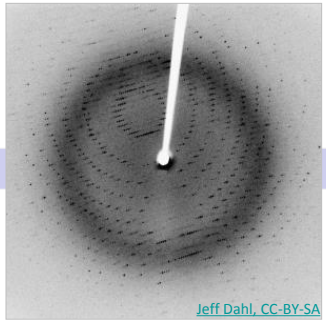
Cambridge Crystallographic Data Centre

@ijbruno

@ccdc_cambridge

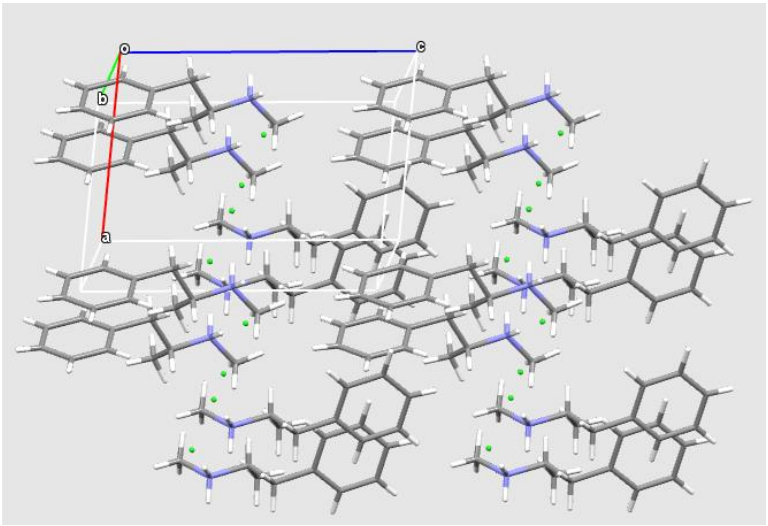
Experimental Data

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



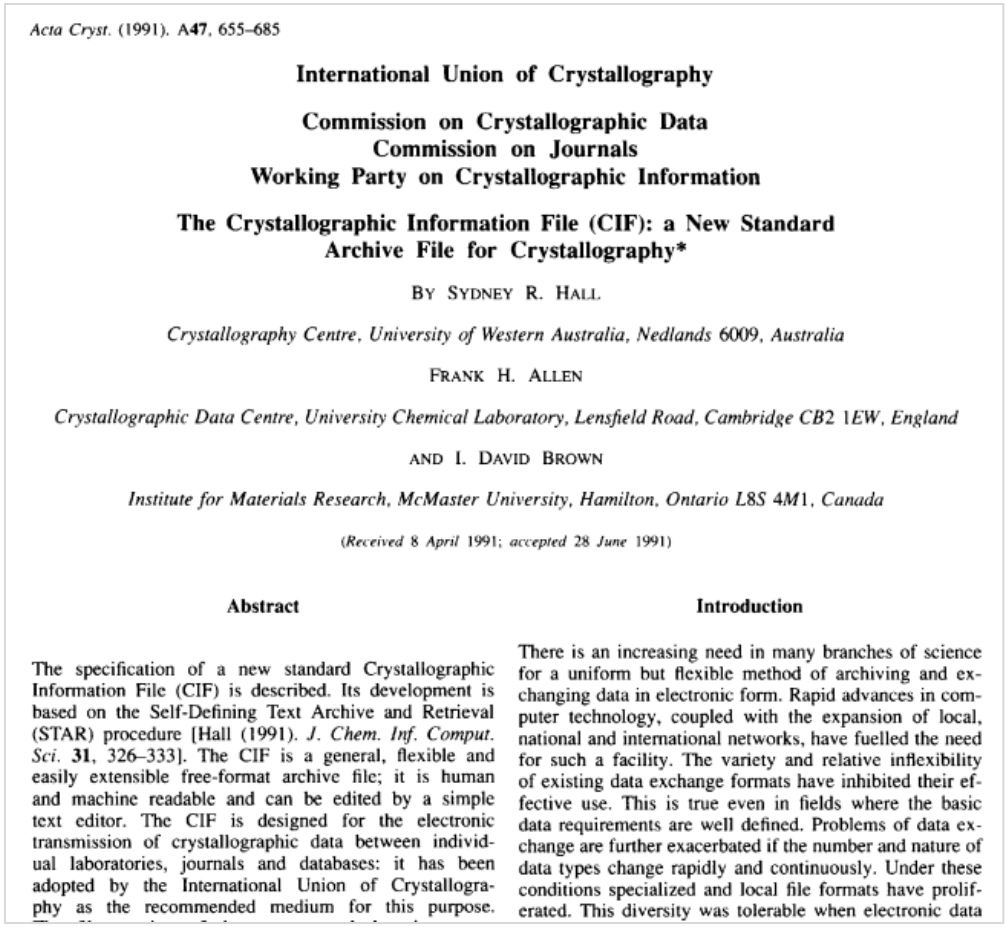

```
_diffrn_ambient_temperature       90(2)
_diffrn_radiation_type            MoK\a
_diffrn_radiation_wavelength      0.71073
_diffrn_radiation_monochromator   graphite
_diffrn_measurement_device_type   'Bruker APEX CCD area-detector'
_diffrn_measurement_method        '\f and \w'
_diffrn_detector_area_resol_mean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

loop_
 _atom_site_type_symbol
 _atom_site_label
 _atom_site_fract_x
  atom_site_fract_y
    om_site_fract_z
   om_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) 0.5363(2) 0.0172(4) Uani d U 1 1 . .
C C1 0.6896(4) 0.7867(6) 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 . .
```



- Experimentally determined 3D coordinates of atoms
- Arrangement of molecules in the crystal
- Captured in a CIF file

Crystallographic Information Framework (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)

Abstract

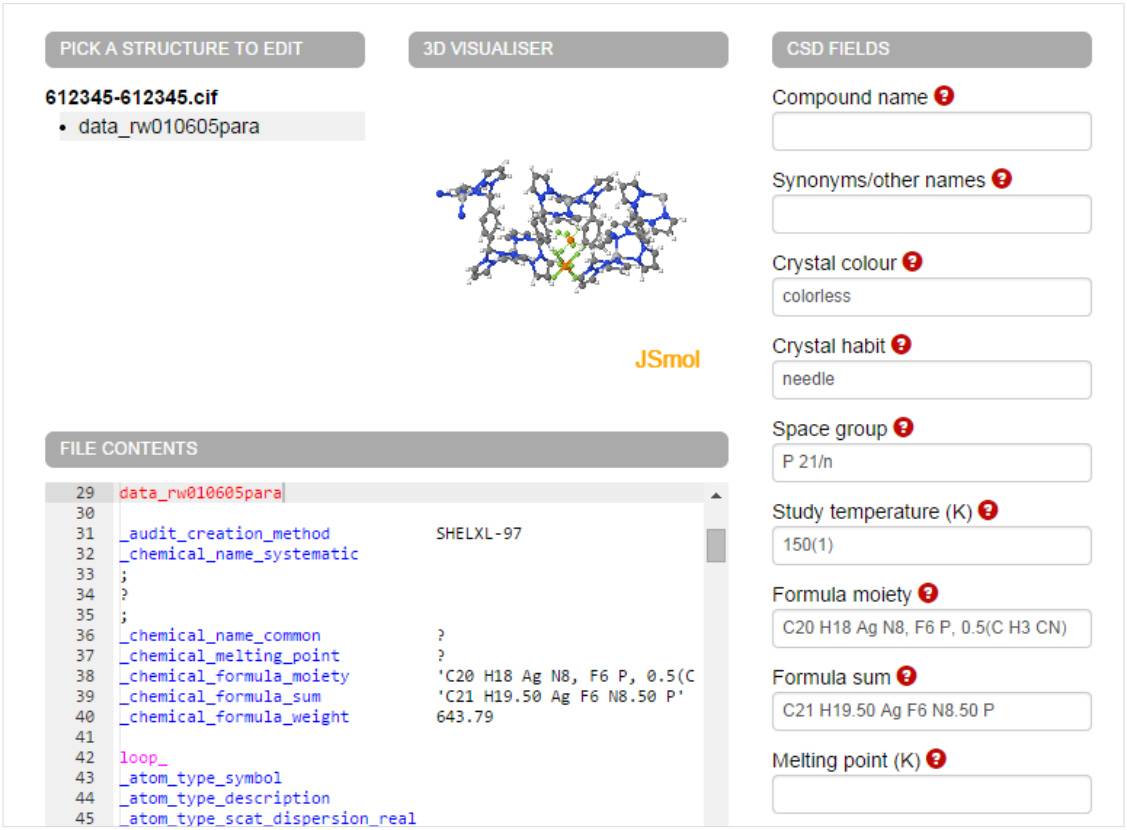
The specification of a new standard Crystallographic Information File (CIF) is described. Its development is based on the Self-Defining Text Archive and Retrieval (STAR) procedure [Hall (1991). *J. Chem. Inf. Comput. Sci.* **31**, 326–333]. The CIF is a general, flexible and easily extensible free-format archive file; it is human and machine readable and can be edited by a simple text editor. The CIF is designed for the electronic transmission of crystallographic data between individual laboratories, journals and databases: it has been adopted by the International Union of Crystallography as the recommended medium for this purpose.

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 data requirements are well defined. Problems of data exchange are further exacerbated if the number and nature of data types change rapidly and continuously. Under these conditions specialized and local file formats have proliferated. This diversity was tolerable when electronic data

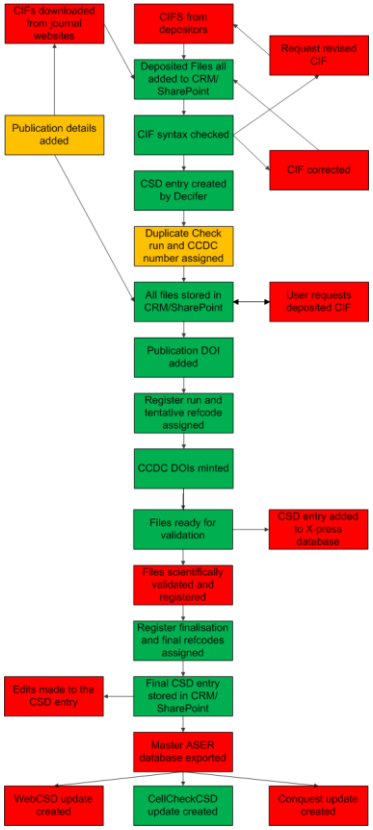
- Standard format for capturing data about structure and experiment
- Maintained by the International Union of Crystallography (IUCr)
- Widely adopted by crystallographic and publishing communities

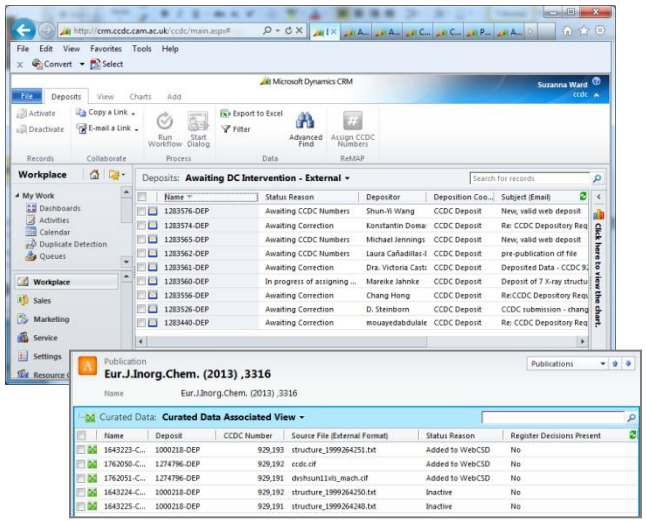
Crystal Structure Deposition



- Researchers deposit CIFs with CCDC
- Deposition services provide basic validation checks
- Depositor receives a unique accession ID
- Pre-publication access for trusted reviewers

Data Deposition Challenges

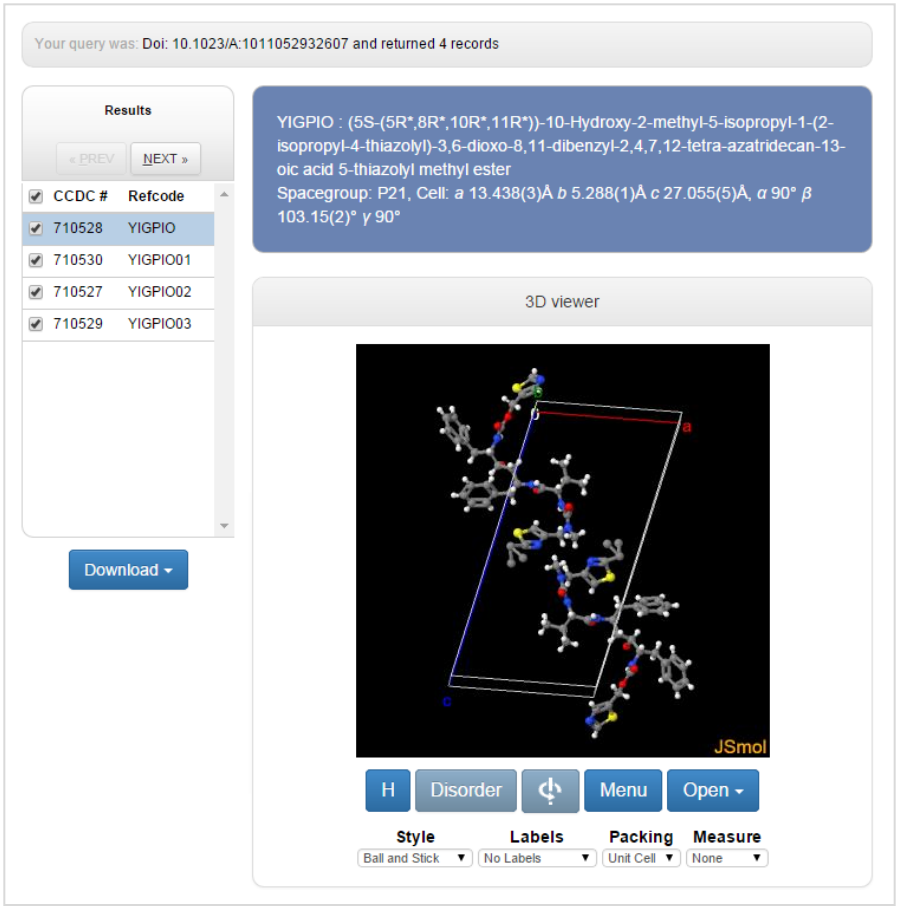




- Syntax errors
- Inaccuracies
- Revisions
- Republications
- Timely release

Over 60,000 structures deposited annually

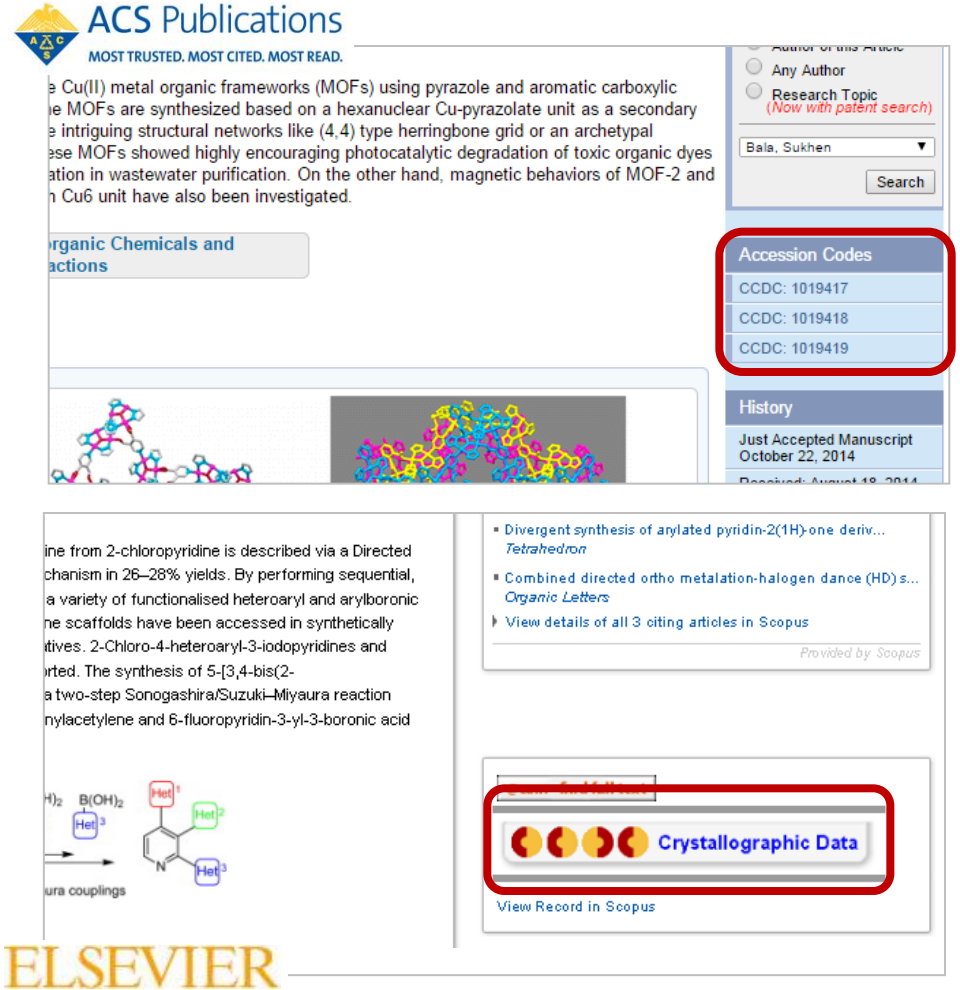
Crystal Structure Dissemination

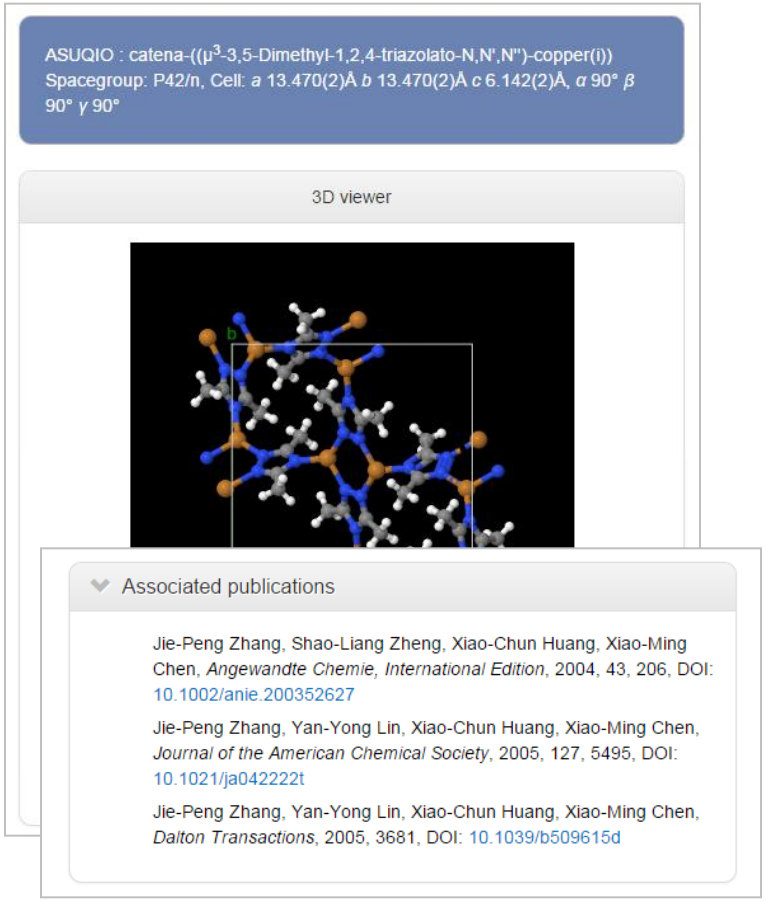


- On publication, deposited data sets freely available for anyone to download
- Individual structures accessible via CCDC Summary Page*
- Web services enable structure lookup and retrieval

* New Summary Page scheduled for launch December 2014

Linking articles and data





Links are in place for ACS, RSC, Elsevier and IUCr journals

DOIs and Data Citation

- Unambiguous and persistent identification of datasets
- Over 500,000 DOIs registered since April 2014
- Foundation for formalising data citation and interoperability

DataCite

10.5517/CCPHZ37

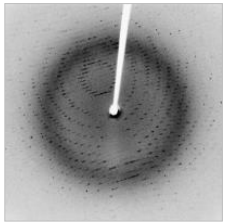
DC1 Data Citation Principles

Data should be considered legitimate, citable products of research…

https://www.force11.org/datacitation

Dataset Publication

CCDC 892348: Experimental Crystal Structure Determination. **A. Crystallographer**, *Cambridge Crystallographic Data Centre* (2013) http://dx.doi.org/10.5517/CCYYKFV



Potential for linking between derived data at CCDC and raw data stored at STFC based on DOIs.

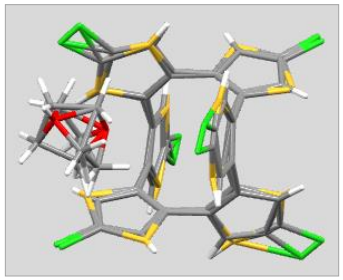
COVERED BY DATA CITATION INDEX THOMSON REUTERS

Coverage of CCDC data by the Thomson Reuters Data Citation Index to be achieved via the DataCite metadata store.

Assigning Chemistry

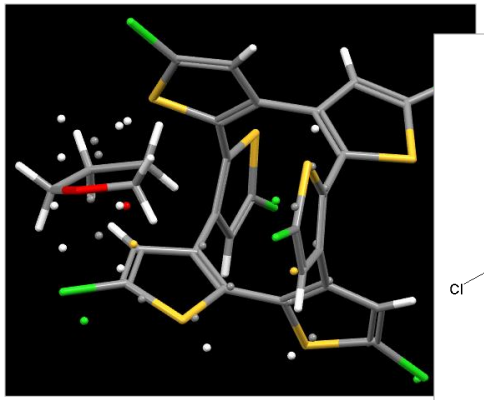
Data deposited with CCDC

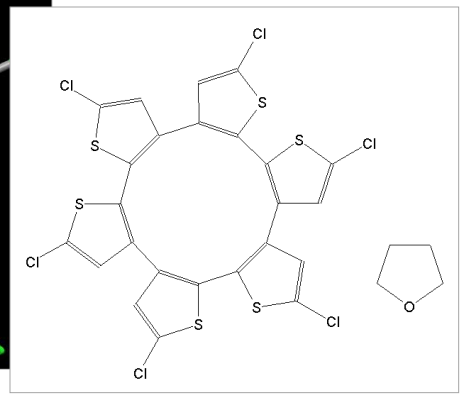
```
loop_
_atom_site_label
_atom_site_type_symbol
_atom_site_fract_x
_atom_site_fract_y
_atom_site_fract_z
_atom_site_U_iso_or_equiv
_atom_site_adp_type
_atom_site_occupancy
_atom_site_symmetry_multiplicity
_atom_site_calc_flag
_atom_site_refinement_flags
_atom_site_disorder_assembly
_atom_site_disorder_group
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
```



- Unambiguous identification of datasets
- Correction of syntax errors
- Provenance and attribution

Entry in Cambridge Structural Database





Identifier	ABOWOD
Literature Reference	M.J.Marsella, Kunsang Yoon, F.S.Tham, *Org.Lett.* (2001), **3**, 2129, doi:10.1021/ol016122t
Formula	$C_{24}H_6Cl_6S_6,2(C_4H_8O)$
Compound Name	trans-1^5,2^5,3^5,4^5,5^5,6^5-Hexachloro-1,2,4,6(3,2),3,5(2,3)-hexathiophenacyclohexaphane tetra hydrofuran solvate
Disorder	The whole cyclophane molecule and the tetrahydrofuran solvent molecule are disordered equally over two sites.

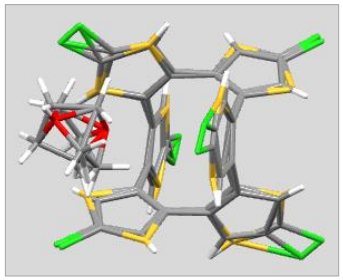
- Assignment of chemistry
- Additional scientific data and metadata
- Review by editorial staff

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

Assigning Chemistry

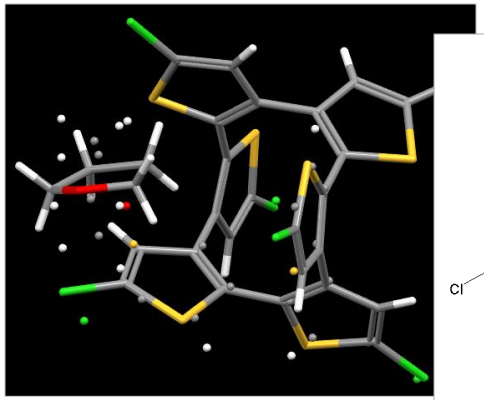
Data deposited with CCDC

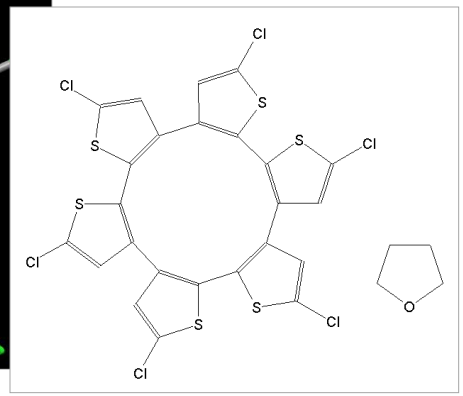
```
loop_
_atom_site_label
_atom_site_type_symbol
_atom_site_fract_x
_atom_site_fract_y
_atom_site_fract_z
_atom_site_U_iso_or_equiv
_atom_site_adp_type
_atom_site_occupancy
_atom_site_symmetry_multiplicity
_atom_site_calc_flag
_atom_site_refinement_flags
_atom_site_disorder_assembly
_atom_site_disorder_group
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
```



- Unambiguous identification of datasets
- Correction of syntax errors
- Provenance and attribution

Entry in Cambridge Structural Database





Identifier	ABOWOD
Literature Reference	M.J.Marsella, Kunsang Yoon, F.S.Tham, *Org.Lett.* (2001), **3**, 2129, doi:10.1021/ol016122t
Formula	$C_{24}H_6Cl_6S_6,2(C_4H_8O)$
Compound Name	trans-1^5,2^5,3^5,4^5,5^5,6^5-Hexachloro-1,2,4,6(3,2),3,5(2,3)-hexathiophenacyclohexaphane tetra hydrofuran solvate
Disorder	The whole cyclophane molecule and the tetrahydrofuran solvent molecule are disordered equally over two sites.

- Assignment of chemistry
- Additional scientific data and metadata
- Review by editorial staff

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

ChemSpider
Search and share chemistry

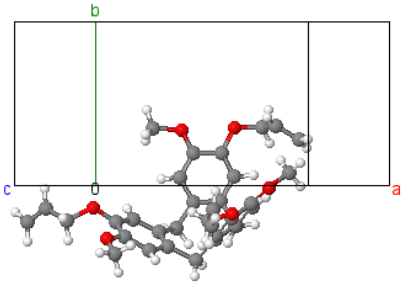
ROYAL SOCIETY OF CHEMISTRY

Crystal CIFs

Add

- Associated Hyperlink: http://www.ccdc.cam.ac.uk/services/structure_request?pid=ccdc:148418&sid=ChemSpider
  Comments: Structure reported in RSC article http://dx.doi.org/10.1039/B000825G and hosted by CCDC with reference 148418
  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
  Approved: No
  Submitted by: Aileen Day

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



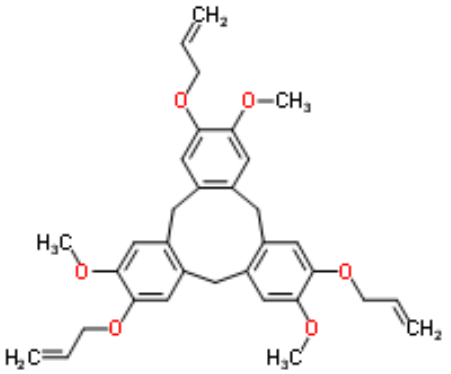
Bulk identification of links between ChemSpider molecules and CCDC Crystal Structures will be facilitated by InChIs.

Other opportunities:

- PubChem
- UniChem (EBI)
- Wikipedia
- Chemical Abstracts

IUPAC

InChI TRUST



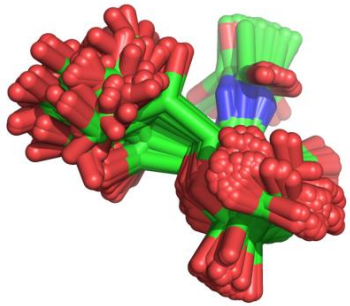
Standard InChI:
`InChI=1S/C33H36O6/c1-7-10-37-31-19-25-13-23-17-29(35-5)33(39-12-9-3)21-27(23)15-24-18-30(36-6)32(38-11-8-2)20-26(24)14-22(25)16-28(31)34-4/h7-9,16-21H,1-3,10-15H2,4-6H3`

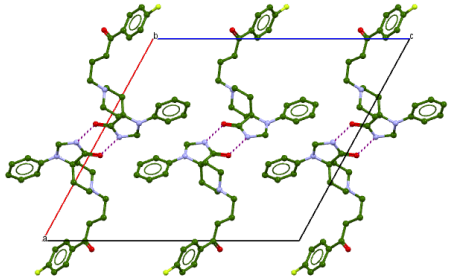
Standard InChIKey:
`IZHKSTHBLQRIOW-UHFFFAOYSA-N`

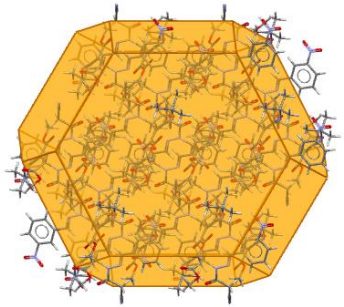
Knowledge-based Solutions

- Provide insights into
  - molecular dimensions and shape
  - molecular interactions

- Applicable to
  - drug design and development
  - design of new materials
  - crystal engineering
  - structure validation

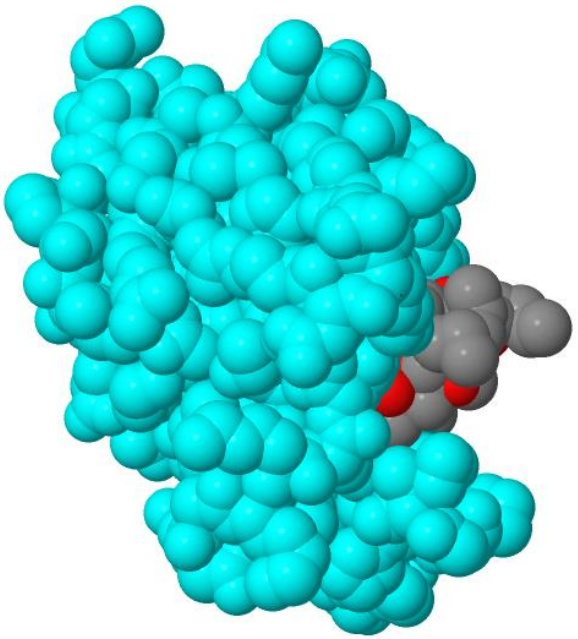


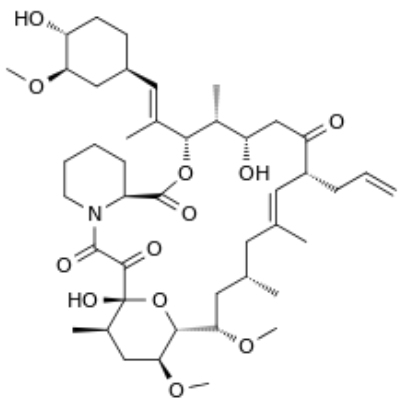




Crystal Structure Knowledge Helps Drug Design

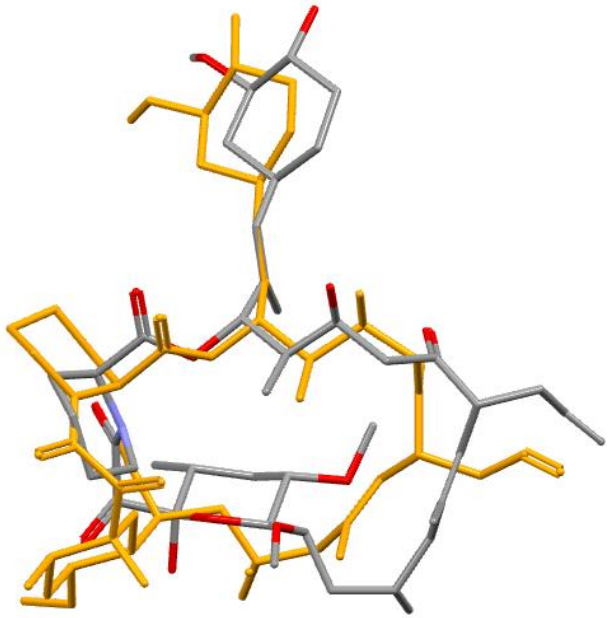
PDB 1BKF complexed with PDB Ligand FK5





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

CSD FINWEE10 overlayed on PDB Ligand FK5



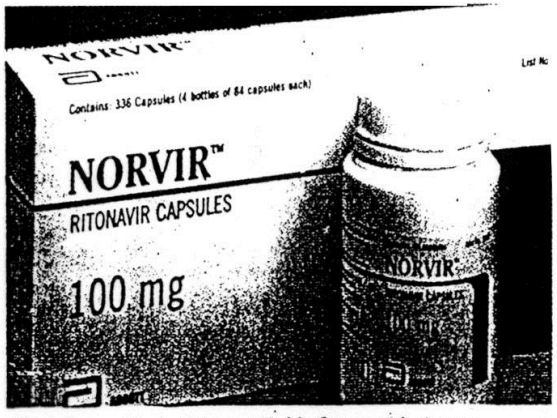
Understanding factors that influence the shape of molecules helps identify better drug candidates

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

Crystal Structure Knowledge Mitigates Risk

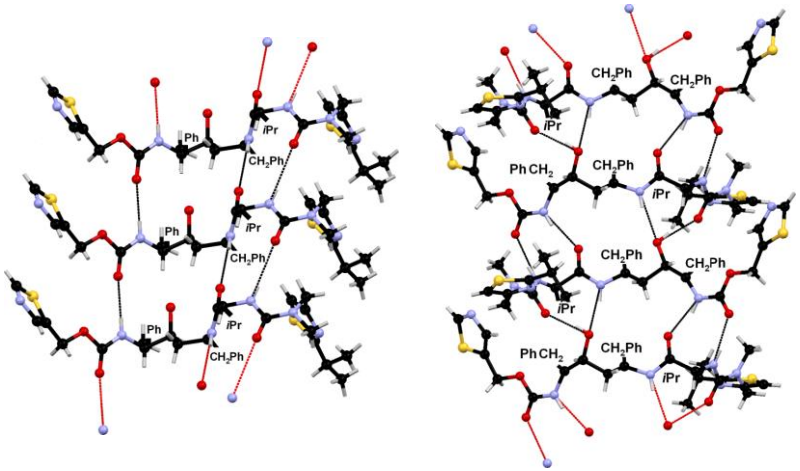
Manufacturing problems hit Abbott's HIV drug ritonavir

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



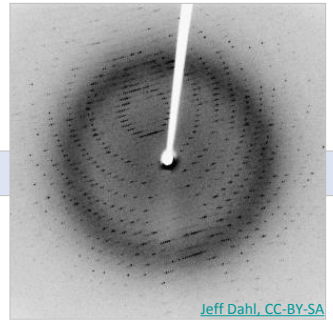
Different crystal forms, different interactions, different solubility, different stability.

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

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

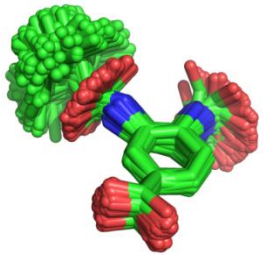
Experimental Data

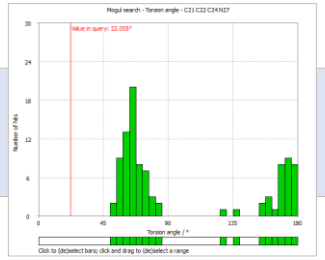
$C_{10}H_{16}N^{+},Cl^{-}$

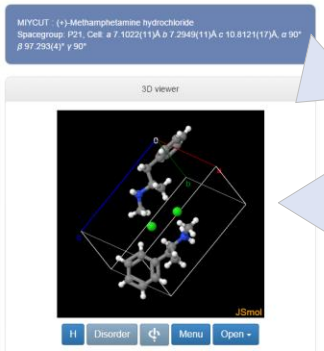


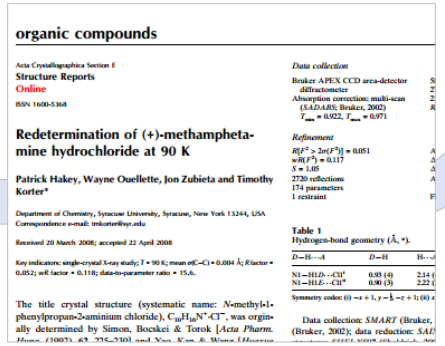

```
_diffrn_ambient_temperature      90(2)
_diffrn_radiation_type           MoK\a
_diffrn_radiation_wavelength     0.71073
_diffrn_radiation_monochromator  graphite
_diffrn_measurement_device_type  'Bruker APEX CCD area-detector'
_diffrn_measurement_method       '\f and \w'
_diffrn_detector_area_resol_mean 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) 0.5363(2) 0.0172(4) Uani d U 1 1 . .
C C1 0.6896(4) 0.7867(6) 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(6) Uani d U 1 1 . .
C C4 0.8700(4) 0.5637(4) 0.7481(3) 0.0236(6) Uani d U 1 1 . .
```









organic compounds

Acta Crystallographica Section E
Structure Reports Online
ISSN 1600-5368

Redetermination of (+)-methamphetamine hydrochloride at 90 K

Patrick Hakey, Wayne Ouellette, Jon Zubieta and Timothy Korter*

Department of Chemistry, Syracuse University, Syracuse, New York 13244, USA
Correspondence e-mail: tmkorter@syr.edu

Received 20 March 2008; accepted 22 April 2008

Key indicators: single-crystal X-ray study; T = 90 K; mean σ(C–C) = 0.004 Å; R factor = 0.052; wR factor = 0.118; data-to-parameter ratio = 15.6.

The title crystal structure (systematic name: *N*-methyl-1-phenylpropan-2-aminium chloride), $C_{10}H_{16}N^{+}\cdot Cl^{-}$, was originally determined by Simon, Bockeel & Torok [*Acta Pharm.*

Data collection
Bruker APEX CCD area-detector diffractometer
Absorption correction: multi-scan (*SADABS*; Bruker, 2002)
T_{min} = 0.922, T_{max} = 0.971

Refinement
$R[F^2 > 2\sigma(F^2)]$ = 0.051
$wR(F^2)$ = 0.117
S = 1.05
2720 reflections
174 parameters
1 restraint

Table 1
Hydrogen-bond geometry (Å, °).

D—H···A	D—H	H···A
N1—H1D···Cl1^{i}	0.93 (4)	2.14 (
N1—H1E···Cl1^{ii}	0.90 (3)	2.22 (

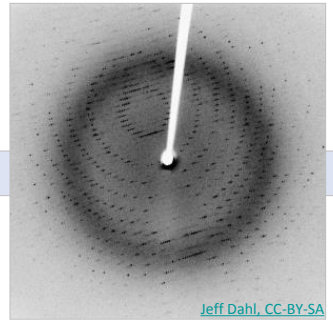
Symmetry codes: (i) −x + 1, y − ½, −z + 1; (ii) x

Data collection: *SMART* (Bruker, (Bruker, 2002); data reduction: *SAI*

Structural Knowledge

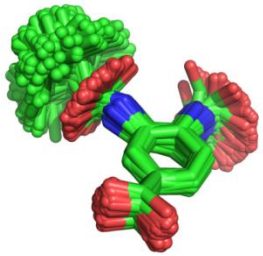
Experimental Data

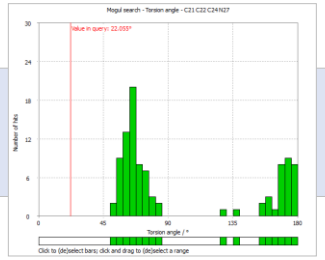
$C_{10}H_{16}N^{+},Cl^{-}$

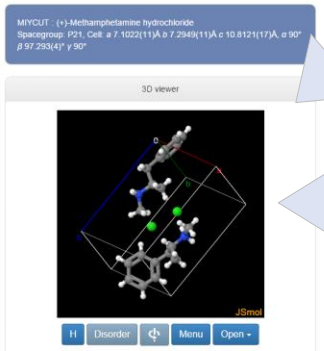


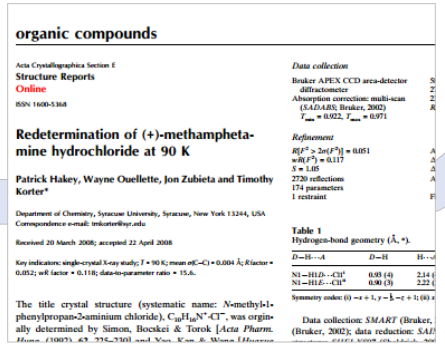

```
_diffrn_ambient_temperature      90(2)
_diffrn_radiation_type           MoK\a
_diffrn_radiation_wavelength     0.71073
_diffrn_radiation_monochromator  graphite
_diffrn_measurement_device_type  'Bruker APEX CCD area-detector'
_diffrn_measurement_method       '\f and \w'
_diffrn_detector_area_resol_mean 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) 0.5363(2) 0.0172(4) Uani d U 1 1 . .
C C1 0.6896(4) 0.7867(6) 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(6) Uani d U 1 1 . .
C C4 0.8700(4) 0.5637(4) 0.7481(3) 0.0236(6) Uani d U 1 1 . .
```









organic compounds

Acta Crystallographica Section E
Structure Reports Online
ISSN 1600-5368

Redetermination of (+)-methamphetamine hydrochloride at 90 K

Patrick Hakey, Wayne Ouellette, Jon Zubieta and Timothy Korter*

Department of Chemistry, Syracuse University, Syracuse, New York 13244, USA
Correspondence e-mail: tmkorter@syr.edu

Received 20 March 2008; accepted 22 April 2008

Key indicators: single-crystal X-ray study; T = 90 K; mean σ(C–C) = 0.004 Å; R factor = 0.052; wR factor = 0.118; data-to-parameter ratio = 15.6.

The title crystal structure (systematic name: *N*-methyl-1-phenylpropan-2-aminium chloride), $C_{10}H_{16}N^{+}\cdot Cl^{-}$, was originally determined by Simon, Bockeel & Torok [*Acta Pharm.*

Data collection
Bruker APEX CCD area-detector diffractometer
Absorption correction: multi-scan (*SADABS*; Bruker, 2002)
T_{min} = 0.922, T_{max} = 0.971

Refinement
$R[F^2 > 2\sigma(F^2)]$ = 0.051
$wR(F^2)$ = 0.117
S = 1.05
2720 reflections
174 parameters
1 restraint

Table 1
Hydrogen-bond geometry (Å, °).

D—H···A	D—H	H···A
N1—H1D···Cl1^{i}	0.93 (4)	2.14 (
N1—H1E···Cl1^{ii}	0.90 (3)	2.22 (

Symmetry codes: (i) −x + 1, y − ½, −z + 1; (ii) x

Data collection: *SMART* (Bruker, (Bruker, 2002); data reduction: *SAI*

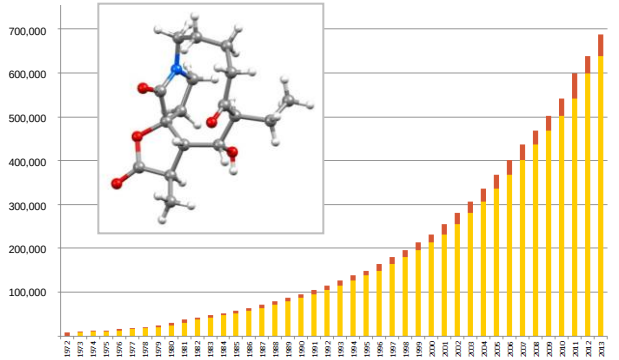
Structural Knowledge

Crystal Structure Growth

- Cambridge Structural Database
  - organic and metal-organic compounds
  - **750,422 structures**

ESTABLISHED 1965

Growth of the CSD



- Inorganic Crystal Structure Database
  - inorganic compounds
  - **173,473 structures**

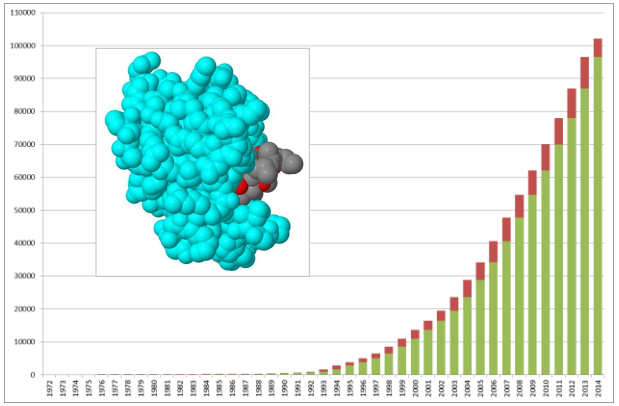
NIST

FIZ KARLSRUHE

- Protein Data Bank
  - biological macromolecules
  - **105,025 structures**

WORLDWIDE PDB
PROTEIN DATA BANK

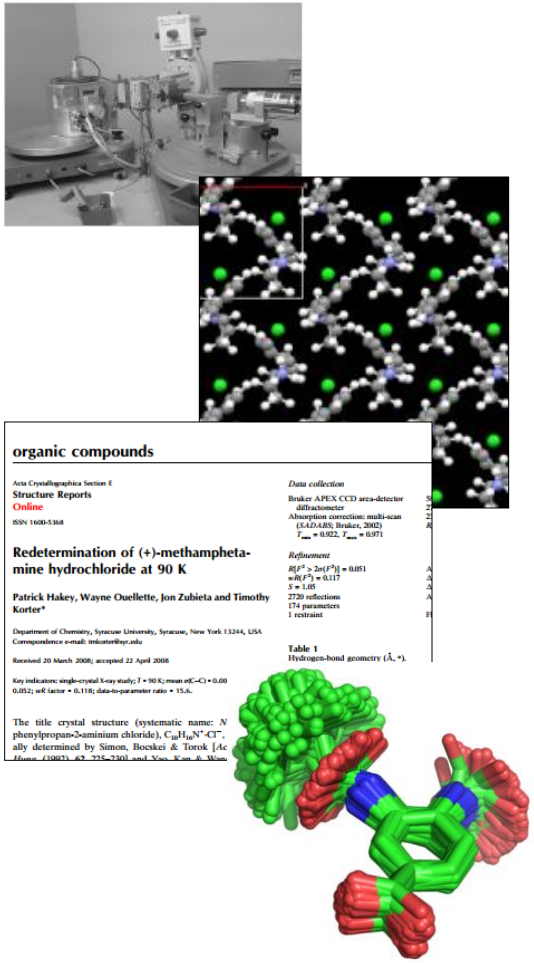
Growth of the PDB



CSD: 18 Nov 2014; ICSD: 27 Oct 2014; PDB: 25 Nov 2014: Overall: 1,028,920

A Collaborative Journey

- Academic research community
- Instrument manufacturers
- Publishers
- Data repositories
- Industrial partners
- Scientific software vendors



We've come a long way…

- Essential foundations
  - standard file formats
  - rich metadata
  - standard identifiers
  - interoperable services
  - sustainability

RDA-WDS Data Publishing Activities

- Workflows
- Publishing Data Services
- Data Bibliometrics
- Cost Recovery

… there are new roads ahead

- Essential foundations
  - standard file formats
  - rich metadata
  - standard identifiers
  - interoperable services
  - sustainability

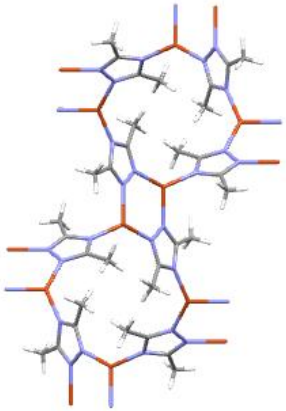
- Future opportunities
  - unpublished data
  - enhanced deposition services
  - new publication pathways
  - data metrics and impact

RDA-WDS Data Publishing Activities

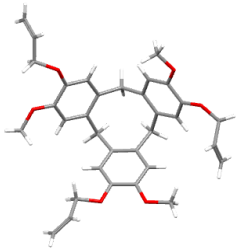
- Workflows
- Publishing Data Services
- Data Bibliometrics
- Cost Recovery

Time's Up!

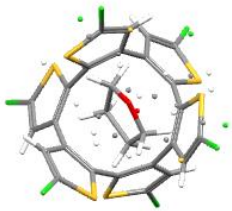
- **About your speaker:**
  - **Name:** Ian Bruno
  - **Company:** Cambridge Crystallographic Data Centre
  - **Email:** bruno@ccdc.cam.ac.uk
  - **Social Media:** @ijbruno @ccdc_cambridge



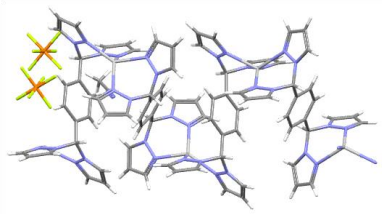
10.5517/cc790j1



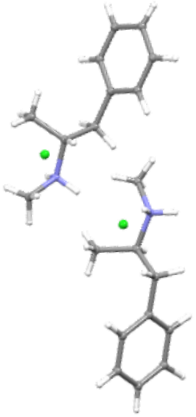
10.5517/cc4zfp6



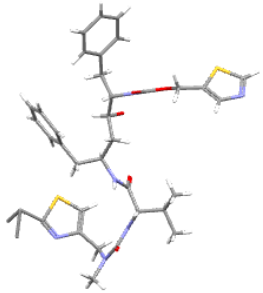
10.5517/cc5jk3b



10.5517/ccnk62g



10.5517/ccr3zh9



10.5517/ccrvc86