The Cambridge Structural Database – Developments in deposition and access

Matt Lightfoot, Suzanna Ward
The Cambridge Crystallographic Data Centre (CCDC)
4th December 2018
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,
Bespoke Training, Teaching Materials

**Employer of ~60 staff**
Scientific editors
Software developers
Applications scientists
Cambridge UK
USA

**Originated in 1965**
Financially self-supporting
Not-for-profit, UK Registered Charity
University Partner Institute
www.ccdc.cam.ac.uk

*Dedicated to the advancement of chemistry and crystallography for the public benefit through providing high quality information services and software.*
The Cambridge Structural Database

- One of 1st - METALD
- 250,000th - IBEZUK
- 500,000th - EFEMUX01
- 750,000th - ZOYBIA
- 1 Millionth structure

Number of entries in the CSD

Publication Year

https://www.ccdc.cam.ac.uk/CCDCStats/
### What’s in the CSD?

<table>
<thead>
<tr>
<th>Organic</th>
<th>Metal-Organic</th>
</tr>
</thead>
<tbody>
<tr>
<td>43%</td>
<td>57%</td>
</tr>
<tr>
<td>Not Polymeric</td>
<td>Polymeric: 11%</td>
</tr>
<tr>
<td>89%</td>
<td></td>
</tr>
<tr>
<td>Single Component</td>
<td>Multi Component</td>
</tr>
<tr>
<td>56%</td>
<td>44%</td>
</tr>
</tbody>
</table>

**Organic**
- Drugs
- Agrochemicals
- Pigments
- Explosives
- Protein ligands

**Metal-Organic**
- Metal Organic Frameworks
- Models for new catalysts
- Porous frameworks for gas storage
- Fundamental chemical bonding
New Zealand

New Zealand Manuka Honey (Leptospermum scoparium)

SAVQOX

WAZHAJ
Worldwide data sharing

- 100,000 datasets deposited annually
- >10,000 depositors
- ~1 million datasets
- >1,000 journals
  - >450,000 publications
  - >380,000 authors
- Peer review data access
- Data available to all at the point of publication

Countries contributing data

Analysis based on a sampling of e-mail address domains
Deposition guidelines

- Data deposited pre-publication
  - Deposition when crystallographer most engaged
  - Enables links to data sets at the point of publication
  - Tailored deposition service

Small molecule single crystal data

Authors should present their crystal data in a CIF (Crystallographic Information File) format and deposit any organic or organometallic structural information with the Cambridge Crystallographic Data Centre (CCDC) before they submit their manuscript to us. Data will be held in the CCDC’s confidential archive until publication of the article, when data for organic and organometallic compounds will be entered into the Cambridge Structural Database. Authors are encouraged to deposit inorganic crystal structures with the ICSD, hosted by FIZ Karlsruhe.

During submission of a manuscript to the Royal Society of Chemistry using our online submission system, authors will be asked to provide CCDC reference numbers. CIFs should not be submitted with the manuscript (these should have already been deposited with the CCDC/ICSD, see above). Any revised CIFs obtained subsequently should be deposited directly with the CCDC before the revised manuscript is submitted to us. CCDC or ICSD numbers should be included in the manuscript prior to submission.
New CIF guidelines for deposition

The CCDC CIF Deposition Guidelines

When preparing your CIF for deposition please include as much information as possible and check it carefully. This is especially true for CSD Communications where there is no paper to describe the chemistry and experimental details leading to your structure. If you choose to publish your data as a CSD Communication please remember to provide all the authors/crystallographers/chemists who contributed to the crystallographic experiment as authors of the data. If we are unable to validate your structure from the information you have provided we may contact you. If we cannot resolve the issue, unfortunately, we may not be able to add your structure to the CSD.

All experimental CIF files (including those from powder diffraction experiments) should contain an R-factor. This should be consistent with the crystallography being performed correctly and to the best ability that would be expected from the material and equipment used. We would like all experimental CIFs to contain:

- R-factors (R1, wR2, Rint)
Data Deposition
Ability to identify the data producer

- Details will also be embedded into CIFs downloaded from CCDC
- Linked ORCID iD
Integrity and validation checks

Validation

Example 1:
- Structure: example_1.cif
- CheckCIF: View Report, Enter Response
- Unit cell check: View Hits
- 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

Example 2:
- Structure: data_5_93GPa
- CheckCIF: View Report, Enter Response
- Unit cell check: View Hits

CheckCIF Responses:
- Level A:
  - PLATON: check_range: Ranges for angles [1.0, 1.0] Width of range [0.0, 1.0] Width of range [0.0, 1.0]
  - PLATON: check_restraints: Restraints check: Range [0.0, 1.0] Width of range [0.0, 1.0]
- Level B:
  - PLATON: check_inter_D1: H1.1.1, H2.2.2, 2.00
- Level C:
  - PLATON: check_extra_D1: H1.1.1, H2.2.2, 2.00

Chemical diagram:
- View group: View group/chemical_data
Data enrichment: By the depositor

Chemical interpretation of 3D data generated by CCDC software. Downloadable after deposition

Opportunity to link to raw data

Opportunity for crystallographer to provide additional domain-specific data items
Adding data to your Researcher IDs

DataCite Search

suzanna Ward

52 Works

CCDC 653598: Experimental Crystal Structure Determination
Terry L. Threlfall, Simon J. Coles, Suzanna C. Ward & Michael B. Hursthouse
Work published 2014 via Cambridge Crystallographic Data Centre

https://doi.org/10.5517/CCPY3TJ

CCDC 653599: Experimental Crystal Structure Determination
Terry L. Threlfall, Simon J. Coles, Suzanna C. Ward & Michael B. Hursthouse
Work published 2014 via Cambridge Crystallographic Data Centre

ORCID

Suzanna C. Ward

ORCID ID

orcid.org/0000-0002-6062-7472

Also known as
Suzanna Ward

Keywords
Crystal

Citation

CCDC 1008189: Experimental Crystal Structure Determination
2014
DOI: 10.5517/CC12V38

Created
2014-06-16

Source: DataCite

ORCID Connecting Research and Researchers

3,045,119 ORCID IDs and counting. See more...
Deposition portal

- **Launched December 1\textsuperscript{st} 2016**
  - create a CCDC profile, log on and view and retrieve your depositions
  - deposit new data using your account
- **April 2017**
  - Edit and update basic information
  - Assign data as *CSD Communications*
  - Extend embargo periods
- **July 2017**
  - Share data with co-workers
- **October 2017**
  - Revise their structures
  - Add links to raw data files
- **May 2018**
  - Improve the deposition of large files
- **July 2018**
  - Launch of the joint deposition process for CSD and ICSD structures
- **September 2018**
  - Ability of add additional files after deposition
- **Currently 29,000 users**
### My Structures

Search by CCDC Number

After depositing structures to the CCDC they may take a few minutes to appear in the table. If you have any questions please contact deposit@ccdc.cam.ac.uk

To change the displayed columns or filter results you should click on the down arrow of the relevant column. To order the results by a particular column click on the heading you wish to order your results by.

<table>
<thead>
<tr>
<th>CCDC No.</th>
<th>Data Block</th>
<th>Deposited On</th>
<th>Deposited By</th>
<th>Refcode</th>
</tr>
</thead>
<tbody>
<tr>
<td>1416019</td>
<td>data_amrh53m</td>
<td>12/01/2017</td>
<td><a href="mailto:mplightfoot74@gmail.com">mplightfoot74@gmail.com</a></td>
<td>AHUXUY</td>
</tr>
<tr>
<td>1416026</td>
<td>data_mm457</td>
<td>12/01/2017</td>
<td><a href="mailto:mplightfoot74@gmail.com">mplightfoot74@gmail.com</a></td>
<td>AHUXOS</td>
</tr>
<tr>
<td>1416025</td>
<td>data_j</td>
<td>12/01/2017</td>
<td><a href="mailto:mplightfoot74@gmail.com">mplightfoot74@gmail.com</a></td>
<td>AHUXIM</td>
</tr>
<tr>
<td>1416016</td>
<td>data_j</td>
<td>10/01/2017</td>
<td><a href="mailto:mplightfoot74@gmail.com">mplightfoot74@gmail.com</a></td>
<td>AHUXAE</td>
</tr>
<tr>
<td>926930</td>
<td>data_aau_dotam_</td>
<td>10/01/2017</td>
<td><a href="mailto:mplightfoot74@gmail.com">mplightfoot74@gmail.com</a></td>
<td>AHUXAE</td>
</tr>
<tr>
<td>1416015</td>
<td>data_aau_dotam_</td>
<td>10/01/2017</td>
<td><a href="mailto:mplightfoot74@gmail.com">mplightfoot74@gmail.com</a></td>
<td>AHUXAE</td>
</tr>
<tr>
<td>1416010</td>
<td>data_1n_DMF_1...</td>
<td>21/12/2016</td>
<td><a href="mailto:mplightfoot74@gmail.com">mplightfoot74@gmail.com</a></td>
<td>AHUWIL</td>
</tr>
<tr>
<td>1416009</td>
<td>data_12DMF_90K</td>
<td>21/12/2016</td>
<td><a href="mailto:mplightfoot74@gmail.com">mplightfoot74@gmail.com</a></td>
<td>AHUWEH</td>
</tr>
<tr>
<td>1416008</td>
<td>data_12DMF_180K</td>
<td>21/12/2016</td>
<td><a href="mailto:mplightfoot74@gmail.com">mplightfoot74@gmail.com</a></td>
<td>AHUWAD</td>
</tr>
</tbody>
</table>

CSD Entry: AHUXIM

Your query was: CCDC identifier(s): 1416025. Authors: John Smith and the search returned 1 record.

Results: AHUXIM - New structure undergoing enhancement

**Space Group**: Pmma, Cell: a 13.16(2) A, b 4.05(2) A, c 19.00(2) A, α 90.00°, β 90.00°, γ 90.00°

**Additional CCDC Details**

- **CCDC Number**: 1416025
- **CCDC Citation**: John Smith. CCDC 1416025: Experimental Crystal Structure Determination, 2017. DOI: 10.1107/cdcl2.153231

**Deposited on**: 12/01/2017

**Associated Publications**

- John Smith, Inorganic Chemistry, 2016, 5, 5234
Coming soon in Deposition/My Structures

- Ability to create a subset of your structures
- Ability to share a subset of your structures with your colleagues
- Ability to create a CSD format database of your in-house structures to use in CSDS software.
CSD Communications allow you to publish data directly through the CSD

- Currently 25,766 CSD Communications

From deposition to access through CSD Communications

For more information see https://www.ccdc.cam.ac.uk/Community/depositastrcture/CSDCommunications/

www.ccdc.cam.ac.uk  UK: +44 1223 336408  US: +1 (648) 445 4893
Impact on quality of the CSD?

- *CSD Communications* vs. peer-reviewed structures
- Does the quality of the data depend on the source?
- Common question/challenges we hear

How can I rely on data if it hasn’t been peer-reviewed?

It isn’t the compound I wanted.

It’s only data that isn’t good enough for publication?

Only duplicates are shared through a database?
Selection of data sources

- From Thomson Reuters Web of Knowledge
- High impact factor journals in ISI categories
  - All Journals Sci1, Sci2
  - Chemistry - Multidisciplinary Chem1, Chem2, Chem3
  - Crystallography Cryst1, Cryst2
  - Data based journal Data1
  - Data published through CSD communications CSDComms
  - Average all data CSD
R-factor

Standard reliability metric – measure of agreement between model and experimental data
CheckCIF results

<table>
<thead>
<tr>
<th>Level</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>Most likely a serious problem, resolve or explain</td>
</tr>
<tr>
<td>B</td>
<td>A potentially serious problem, consider carefully</td>
</tr>
<tr>
<td>C</td>
<td>Check. Ensure it is not caused by an omission or oversight</td>
</tr>
<tr>
<td>G</td>
<td>General information/check it is not something unexpected</td>
</tr>
</tbody>
</table>

Ave. number of alerts
Molecular complexity

![Molecular complexity chart]

- Sci1
- Sci2
- Chem1
- Chem2
- Chem3
- Cryst1
- Cryst2
- Data
- CSDComms

**Ave. % unusual bonds**

**Ave. number of bonds per structure**
Data integrity and fraud checks

• 1 year CCDC funded post-doc has started

• Outputs/aims include:
  – Identification of potentially fraudulent CIF data at deposition and in archive
  – Investigation into data integrity in the CSD
    • Additional checks, flags and options added to Deposition, Referee Service and available through CSD-System
Should we have more checks on all deposited data?

**Structure**

<table>
<thead>
<tr>
<th>CCDC 1234567</th>
<th>View Report</th>
<th>View Report</th>
<th>0 duplicates</th>
<th>View Report</th>
<th>View Report</th>
</tr>
</thead>
</table>

**Summary**

- 0 syntax issues
- 1 crystallographic issue
- 2 chemical issues
- 0 space group issues

**CheckCIF**

- 0 duplicates

**Geometry check**

**Interaction check**
Accessing data

Welcome to Access Structures, the CCDC’s advanced search engine. When searching for details in more than one field, the search will be performed more efficiently.

More advanced search functionality and additional information can be found through the CSD-System and ICSD, respectively.
WebCSD – more advanced searching
WebCSD - More information visible to users
Joint deposition and access for inorganics too

Over 180,000 entries from the Inorganic Crystal Structure Database (ICSD)
Annual targeted database improvements

- Greater efficiency allows targeted improvements to further increase integrity, consistency, discoverability and usage of data
  - 50,000 existing entries improved in 2015
  - 74,000 existing entries improved in 2016
  - 62,000 existing entries improved in 2017
  - 80,000 existing entries improved in 2018

Study Temperature relative to Melting Point

Predominately due to MP reported in °C not K.

Targeted improvements to metal-organics

Add Oxidation state values to CSD entries

Ensure consistency of early CSD entries
Metal-Organic Frameworks

- Two CSD subsets identifying all metal-organic frameworks within the database
  1. MOF list >80,000 entries intended to be as widely-applicable as possible
  2. Non-disordered list >65,000 entries for high-throughput calculations
    - Both subsets updated quarterly
- Python script available to download


Peyman Z. Moghadam,1,* Aurelia Li,1,2 Seth B. Wiggin,1,2 Ando Tao,1 Andrew G. P. Maloney,1 Peter A. Wood,3 Suzanna C. Ward,1 and David Fairen-Jimenez1
1Airborne & Advanced Materials Laboratory (AAML), Department of Chemical Engineering & Biochemistry, University of Cambridge, Pembroke Street, Cambridge CB2 1EW, United Kingdom
2The Cambridge Crystallographic Data Centre, Union Road, Cambridge CB2 1EZ, United Kingdom

Abstract: We report the generation and characterisation of the most complete collection of metal–organic frameworks (MOFs) studied and updated, for the first time, by the Cambridge Crystallographic Data Centre (CCDC). To set up this subset, we asked the question “what is a MOF?” and implemented a number of “back-of-the-envelope” criteria embedded within a bespoke Cambridge Structural Database (CSD) Python API workflow to identify and extract information on 49,468 MOF materials. The CSD MOF subset is updated regularly with subsequent MOF additions to the CSD, bringing a unique record for all researchers working in the area of porous materials around the world, whether they perform high-throughput computational screening for materials discovery or to have a global view over the existing structures in a single resource. Using this resource, we then developed and used a suite of computational tools to remove residual solvent molecules from the framework pores of all the MOFs identified and went on to analyze geometrical and physical properties of non-disordered structures.
Accessing data from general resources

Worldwide online access
Summary

• The CSD is a worldwide community resource
• It is important that we develop services that help you
• What are your priorities for deposition and access?

• Please let us know
  • lightfoot@ccdc.cam.ac.uk
  • support@ccdc.cam.ac.uk

• Thank you for your attention