CCDC

advancing structural science

What's Up

Customer Update Webinar

23rd September 2021

CCDC

advancing structural science

Today's presenters

Abhik Mukhopadhyay

Research and Applications Scientist

Carmen Nitsche

General Manager
CCDC Inc.

Matt Lightfoot

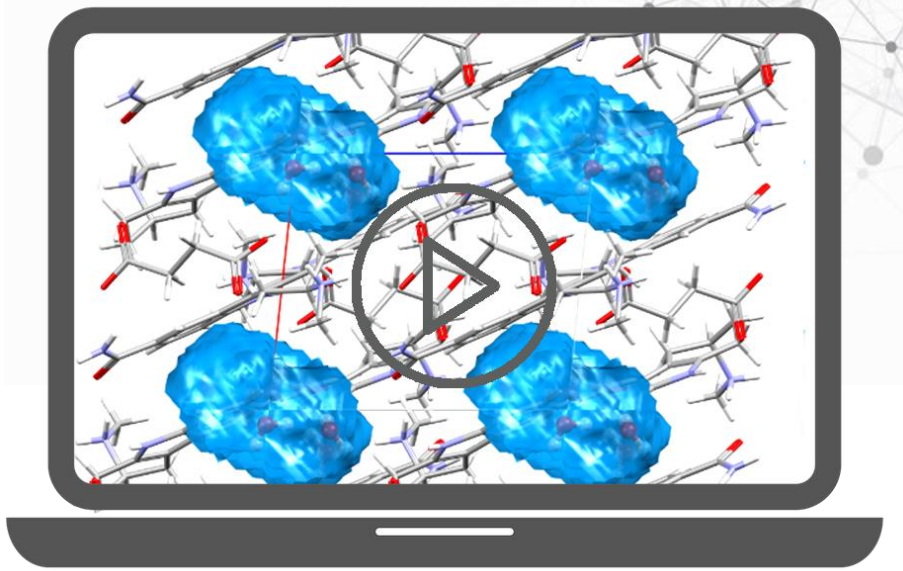
Principal Scientific Editor

CCDC

Overview

In this webinar we will discuss:

- Latest updates and news
- SMILES to 3D structure generation
- How to enhance your CSD depositions
- Q&A: the floor is yours



CCDC

Latest updates

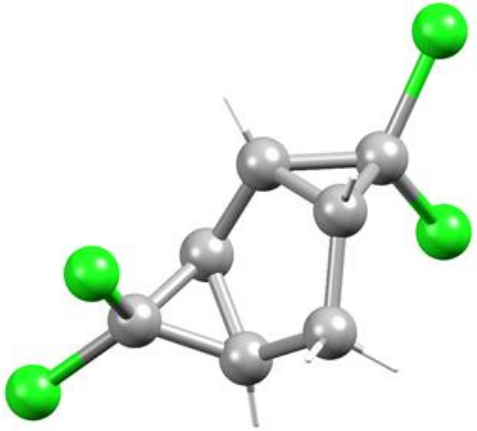
- 2021.2 CSD Release: September 2021
  - CSD-Discovery: the CSD-CrossMiner is now part of the main CSD Portfolio, you can enjoy all our tools within a single installation, which improves usability and search flexibility.
  - CSD-Materials: we have introduced a new component in Mercury, and the CSD Python API, delivering Mogul-like functionality for hydrogen bonds.

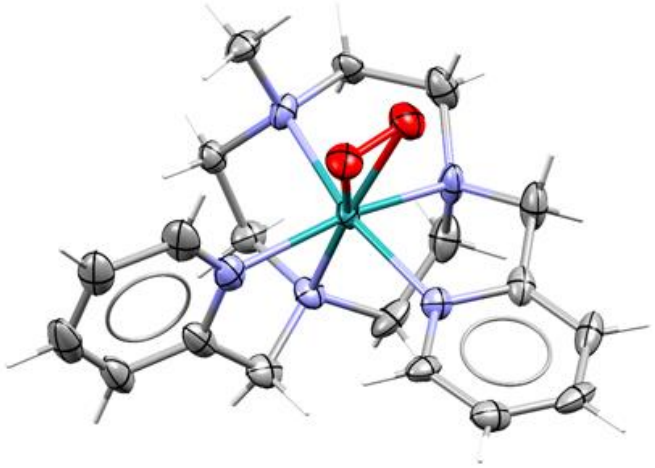
- Read more: https://www.ccdc.cam.ac.uk/solutions/whats-new/

CCDC

Latest updates

- CSD Sketcher v1.0: August 2021 > the first full version of CSD Sketcher, our purpose-built sketcher to enable better database searching in WebCSD.
- CSD Data update: September 2021 > This data update brings you 16,688 new structures (17,283 new entries), including over 300 new entries generated from hardcopy data.





Find out more at www.ccdc.cam.ac.uk/news

CCDC

Upcoming events from CCDC

- CCDC Focus on Metal-Organic Frameworks (MOFs) Virtual Talk and Networking Event – 14th October
- Librarian Teatime: A Focus on CSD Champions – 14th November
  - For Chemistry Librarians in the Americas
- CCDC Virtual Workshops:
  - 2nd November - Deposit your crystallographic data in the CSD
  - 9th November - Learn the basics of Protein Ligand Docking using GOLD
  - 16th November - Intermediate Mogul - Assessing molecular geometries

Register at www.ccdc.cam.ac.uk/News/Events

CCDC

CCDC will be attending

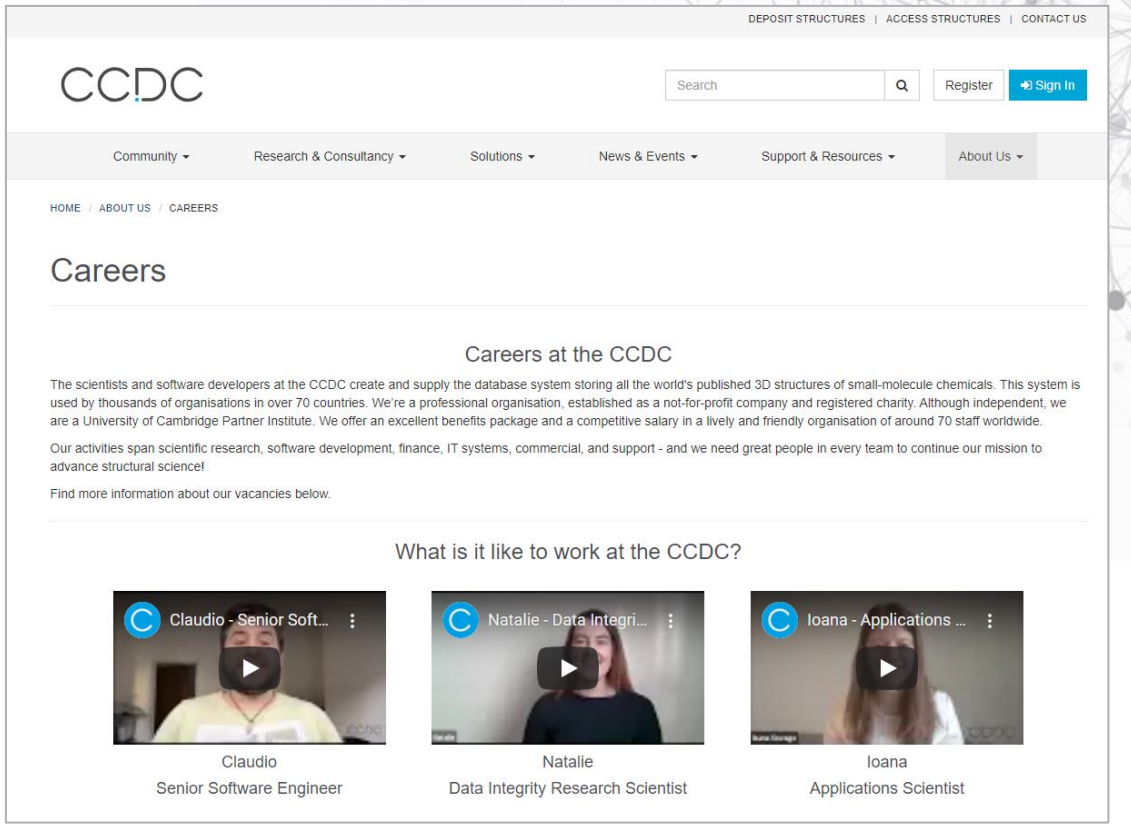
- 4th RSC-BMCS/ RSC-CICAG Artificial Intelligence in Chemistry – **27th to 28th September**
- Chem-Bio Informatics Society Annual Meeting 2021 – **26th October to 28th October**
- Pan-African Crystallographic Conference – **15th November to 19 November**

Register at www.ccdc.cam.ac.uk/News/Events

CCDC

The CCDC team is growing...

- Currently hiring for roles including:
  - Software engineers
  - Product Manager
  - Business Transformation Project Manager
  - Sales Executive, US



Find out more and apply at: www.ccdc.cam.ac.uk/theccdcprofile/careers/

CCDC

CSD Software

SMILES to 3D structure generation

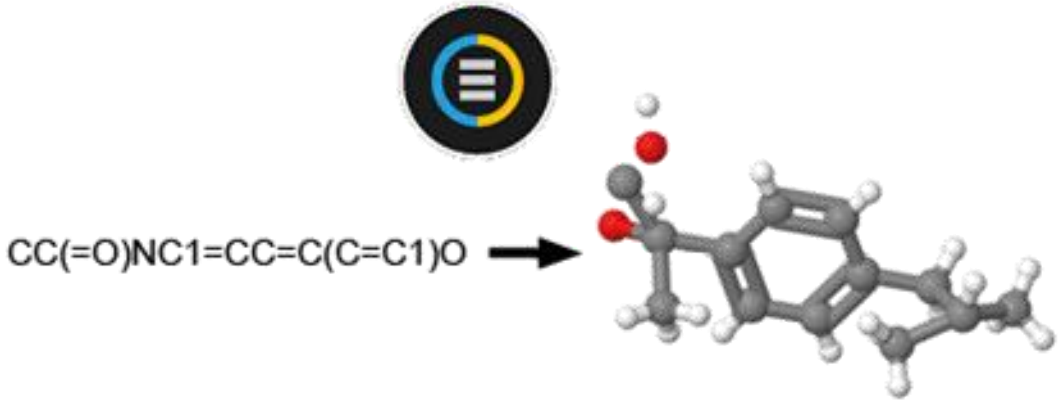
Abhik Mukhopadhyay

Research and Applications Scientist

CCDC

What is it?

- Improvement to the Ligand Preparation workflow in the CSD Python API.
- CCDC Molecule objects can now be created from SMILES strings.
- This improvement will allow to:
  - Use SMILES as a molecule input format to the CSD Python API for the first time.
  - Take a SMILES string and generate a molecule with 3D atom coordinates.
- Generating molecular information from SMILES comes with a CSD-Core license.
- In addition, generating 3D coordinates of the molecule comes with a CSD-Discovery or CSD-Materials licence.



CCDC

Background

- It was not possible before to generate 3D coordinates from SMILES string using CSD Python API.
- Absence of this feature was a bottleneck situation in CADD workflow.
- Addition of this feature will now allow computational chemists to generate 3D coordinates from SMILES string as part of the ligand preparation work-flow of any CADD project, e.g. virtual screening.

CCDC

Why would you use, what are the benefits?

- Users can read SMILES string using CSD Python API and perform other tasks.
- It will be to now easy to incorporate CSD Python API in ligand preparation steps of any CADD workflow.
- Generate 3D coordinates from SMILES using CSD conformer generator that uses CSD knowledgebase.

CCDC

Overview of functionality

- Will allow reading molecule SMILES in CSD Python API.
- Users can generate 3D coordinates from SMILES using CSD conformer generator.
- Conformer Generator functionality will now accept molecules and atoms without coordinates as a starting point for Conformer Generation.
- Key features:
  - Molecules can be read from SMILES strings, with stereochemistry information preserved.
  - 3D conformers can be generated from such molecules and any other molecule without initial 3D coordinates.
  - SMILES strings with stereochemistry information can generated from molecules.

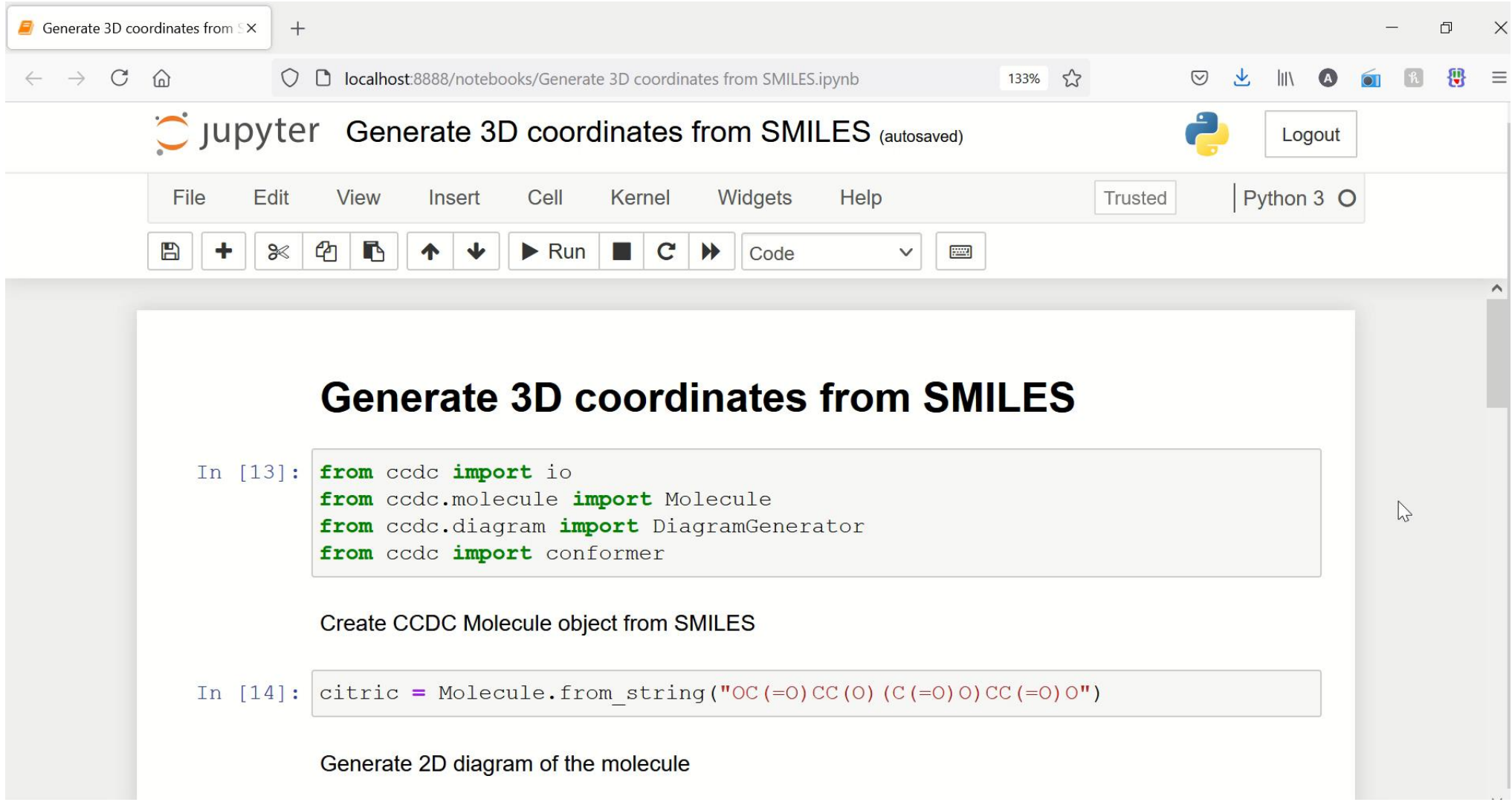
CCDC

How does CSD SMILES to 3D structure work?

- The tool uses the CSD Conformer Generator – uses knowledge from the 1 million+ experimentally derived structures, to predict and generate appropriate conformers - so bond lengths and angles are based on known data.
- The 3D coordinate generator is an iterative, atom-template based process, guided by stereochemistry information and ring positioning heuristics, with continuous optimisation based on CSD geometry distributions.

CCDC

Demo


Generate 3D coordinates from
localhost:8888/notebooks/Generate 3D coordinates from SMILES.ipynb
133%
jupyter
Generate 3D coordinates from SMILES (autosaved)
Logout
File Edit View Insert Cell Kernel Widgets Help
Trusted
Python 3
Run
Code
Generate 3D coordinates from SMILES
In [13]:
from ccdc import io
from ccdc.molecule import Molecule
from ccdc.diagram import DiagramGenerator
from ccdc import conformer
Create CCDC Molecule object from SMILES
In [14]:
citric = Molecule.from_string("OC(=O)CC(O)(C(=O)O)CC(=O)O")
Generate 2D diagram of the molecule

Summary

- The process generates a CSD Python API Molecule file from the SMILES string, which could then be saved as a mol2 file.
- This feature is currently available through the CSD Python API.
- Users can control stereochemistry of a molecule during ligand preparation:
  - By using stereochemistry markers on the SMILES input to generate isomeric structures.

CCDC

CSD Community

How to enhance your CSD depositions

Matt Lightfoot

Principal Scientific Editor

CCDC

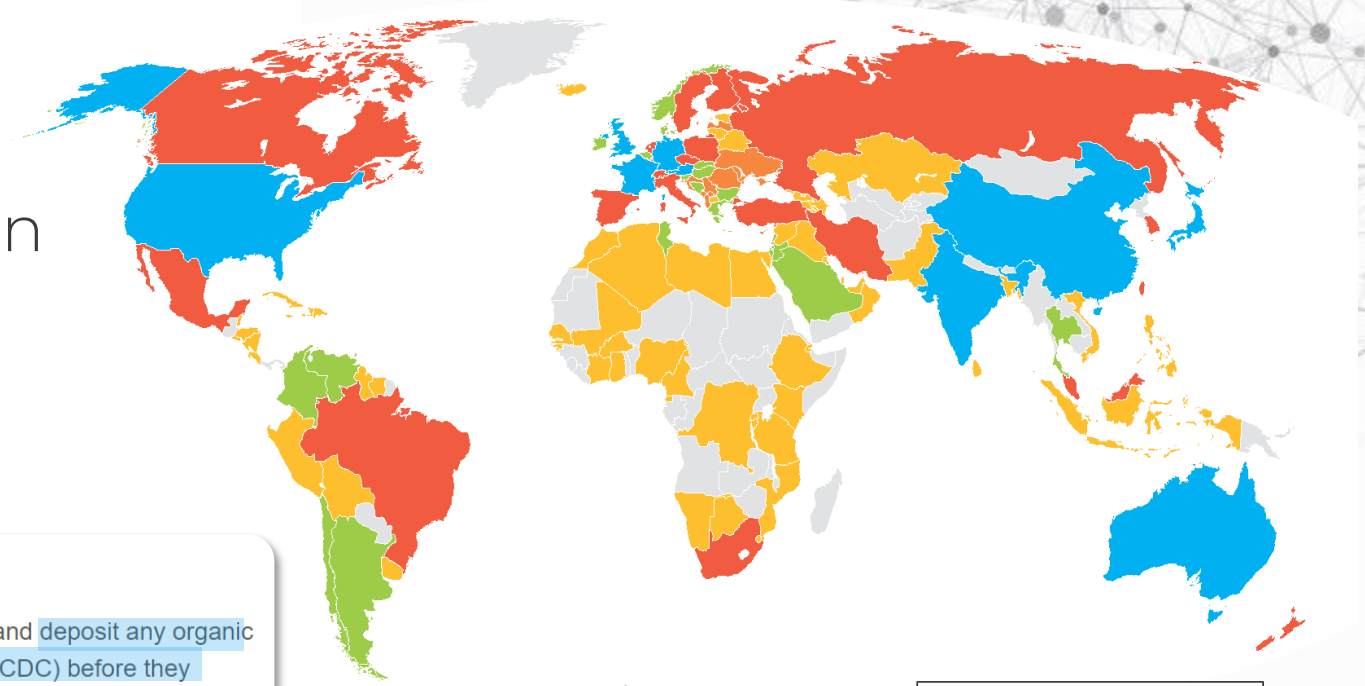
Deposition

- Data deposited pre-publication
  - Enables links to data sets at the point of publication
  - Tailored deposition service

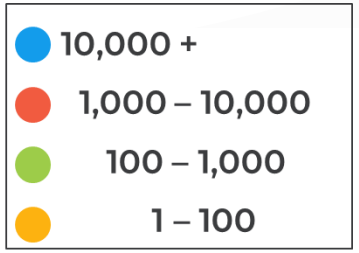
Small molecule single crystal data

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.



Countries contributing data
Analysis based on sampling of email addresses



>30% of depositions from China

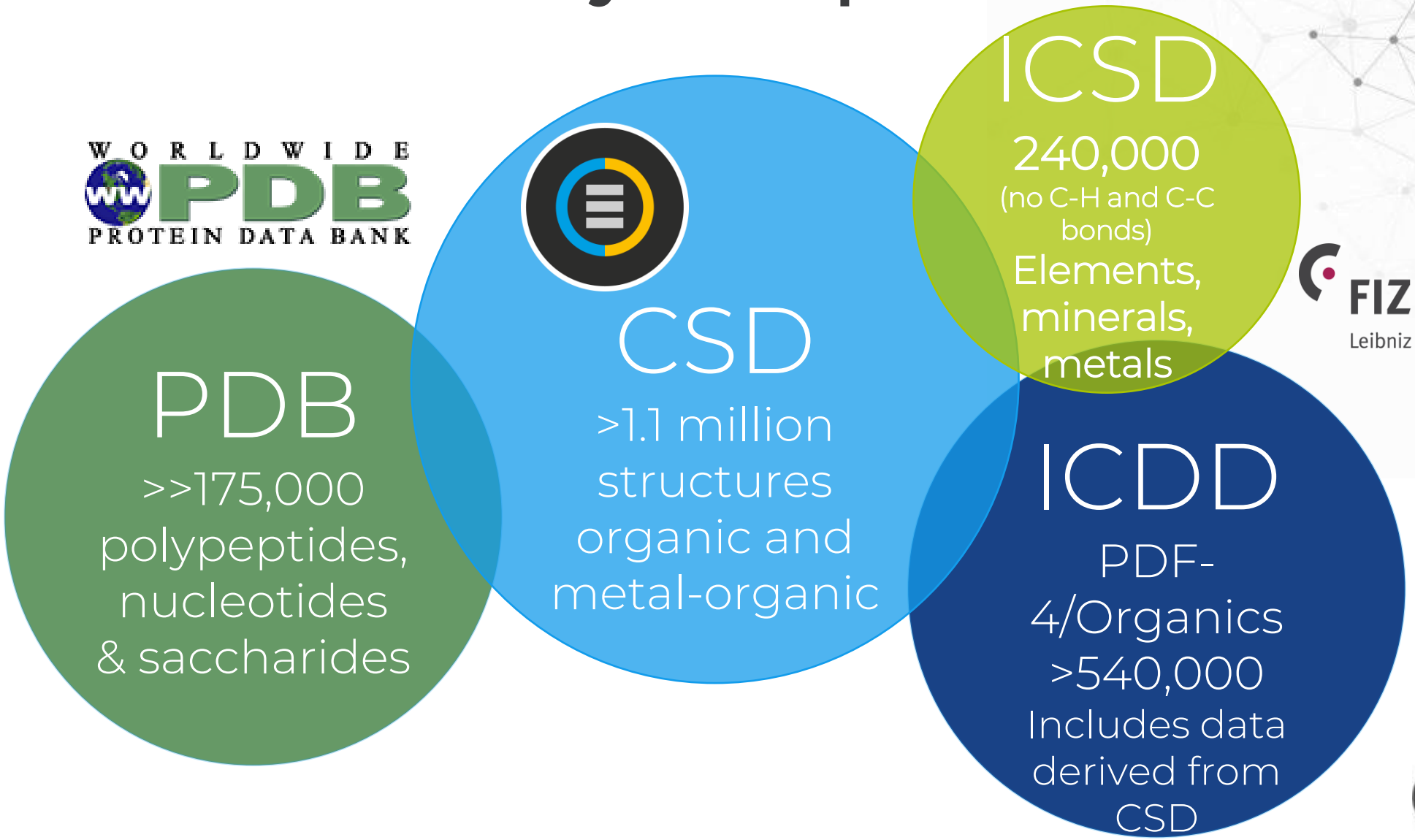
www.ccdc.cam.ac.uk/deposit

When should you deposit data?

- Pre-publication – when you have finished a structure is when you know most about the dataset
- Enables you to check and correct data prior to inclusion in a manuscript
- Deposited data is stored privately at the CCDC prior to publication
- Most publishers require data to be deposited before manuscript submission
  - Enables publishers and referees secure access to the data during the peer review process
  - Enables links to the data to be added at the point of publication

CCDC

What data can you deposit with us?



FIZ Karlsruhe
Leibniz Institute for Information Infrastructure

ICDD

CCDC

Ways to enhance your data during deposition

- The importance of including structures factors
- Data validation and integrity checks and adding responses
- How you can enhance the metadata in your depositions

CCDC

Structure factors

Structure factors: what

The CCDC and FIZ Karlsruhe strongly encourage the incl
Structure Database (ICSD), in line with recommendations

What are structure factors?

Structure factors are created from experimental crystallog
in the structure. For detailed information, see the IUCr's c

Currently, there are two types of information that CCDC a
intensities (.hkl). They can also be appended to the CIF fi

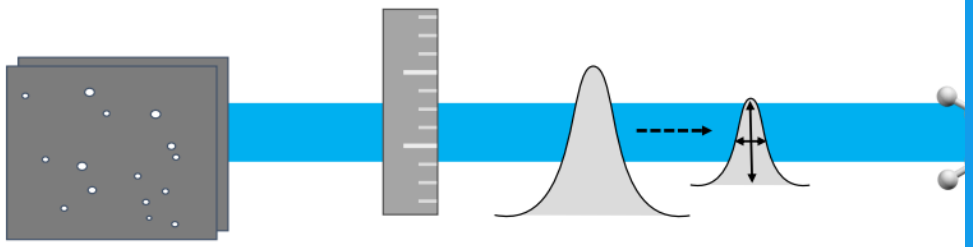
- **The .hkl file** contains the intensity and standard ur
values: h, k and l - to identify the reflection. This inf
- **The .fcf file** contains the structure factors, which a
be recreated using the .hkl and .res file. This last o
constraints that have been applied). The .fcf file is

These files provide more information about the crystal str
crystallographic data, as part of the publications standard

Why is it important to share struct

- It's good data practice - by sharing the data used t
later, then it is already stored, especially if you mov
- CheckCIF (the IUCr's CIF checking service) can d
.res in the CIF allows such checks to be performed
- Several journals require the deposition of structure

A number of different files are created at each stage
crystallographic experiment – here we have listed so
the main ones. Files may have different names/some s
may be combined in different crystallographic softw



Data Collection

Calculating reflection intensities and applying corrections (Integration & Scaling)

Struct
mo

- Diffraction frames (.sfrm, .cbf, .img, etc.)

- .hkl

If the diffraction frames have been deposited in another repository, the DOI can be associated with your dataset during deposition under "Raw Data DOI"

Files for CCDC deposition:

- Mandatory
- Recommended

If no structure factor data has been uploaded, you should:

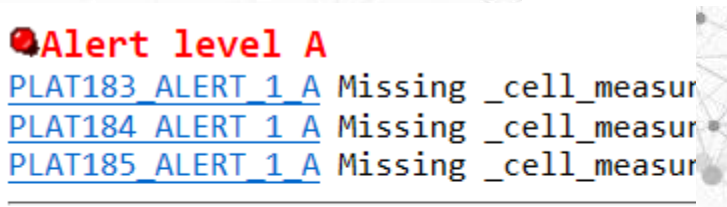
- Go back and upload a new dataset which includes structure factors
- Most refinement software now embeds this into a CIF
- If it is historic data and no structure factor data is available, you need to add a comment before proceeding.

www.ccdc.cam.ac.uk/Community/depositastructure/cif-deposition-guidelines/structure-factor/

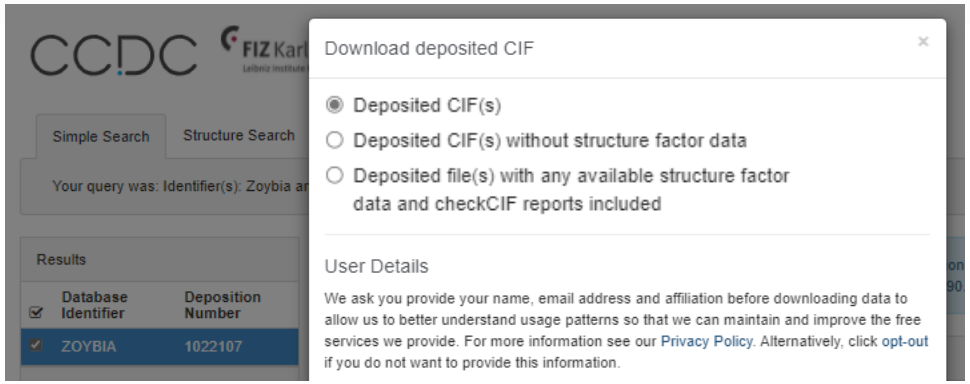
CCDC

Data validation and integrity checks

- IUCr's checkCIF service is embedded into deposition
- Alerts provide potential errors, unusual findings and suggestions for improvement
  - All alerts should be checked
  - Depositors decide to improve their model/dataset or add a comment to provide information about the reason for the alert
  - Comments are embedded into the CIF and available during peer-review and to CSD users after publication
- checkCIF ultimately helps to improve the quality of data published and provides users of the data more information about potential issues in the datasets



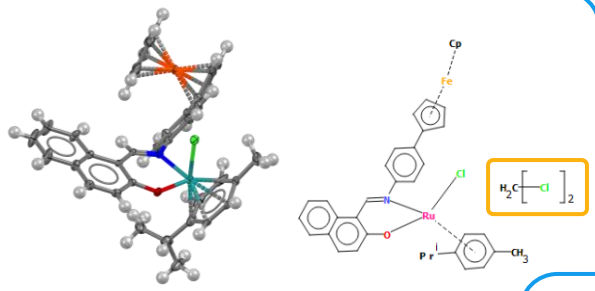
Alerts are hyperlinked to more information about the alert and what action should be taken



https://journals.iucr.org/services/cif/checking/checkfaq.html

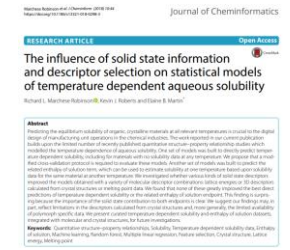
Enhancing key metadata

- Depositors are asked to check and enhance key meta-data
- Helps to improve the quality of key information
- Helps to ensure data is correctly curated and represented in the CSD
- Helps to improve the availability of data
- Ultimately improves the discoverability and re-use of datasets and enables scientists worldwide to learn more from the data

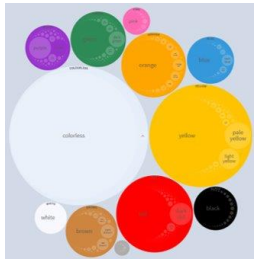


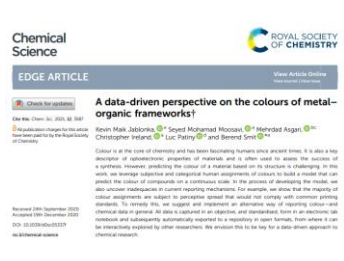
CSD BAGNEG Example of SQUEEZED data being represented correctly

>170K melting points in the CSD



Additional metadata being used to help predict solubilities



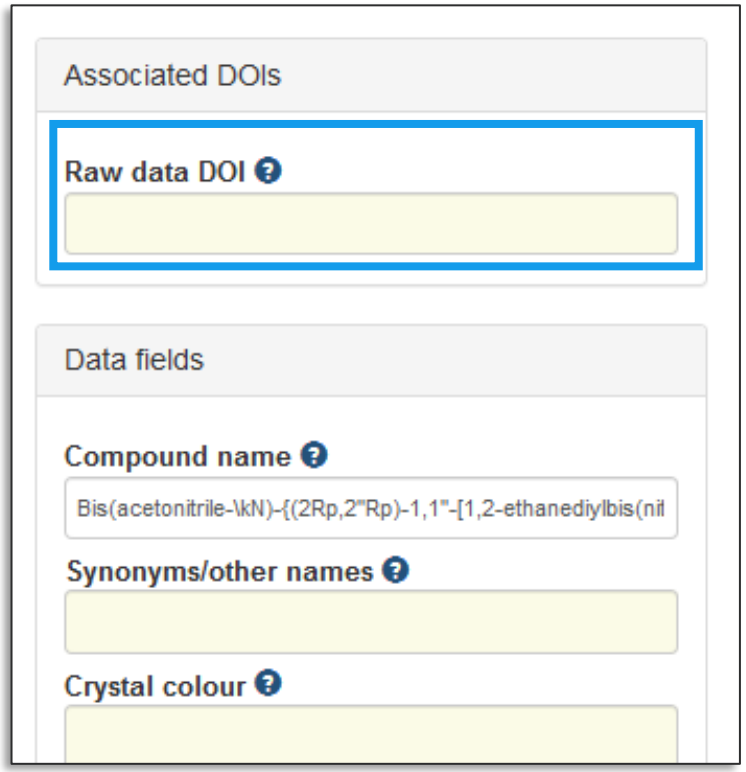


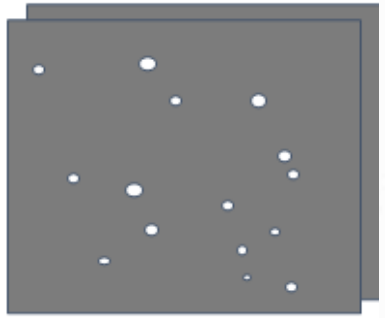
Example of using CSD data to assess the colours of Metal-Organic Frameworks

Chem. Sci., 2021,12, 3587-3598, https://doi.org/10.1039/D0SC05337F,
J Cheminform 10, 44 (2018). https://doi.org/10.1186/s13321-018-0298-3

CCDC

Raw data DOIs





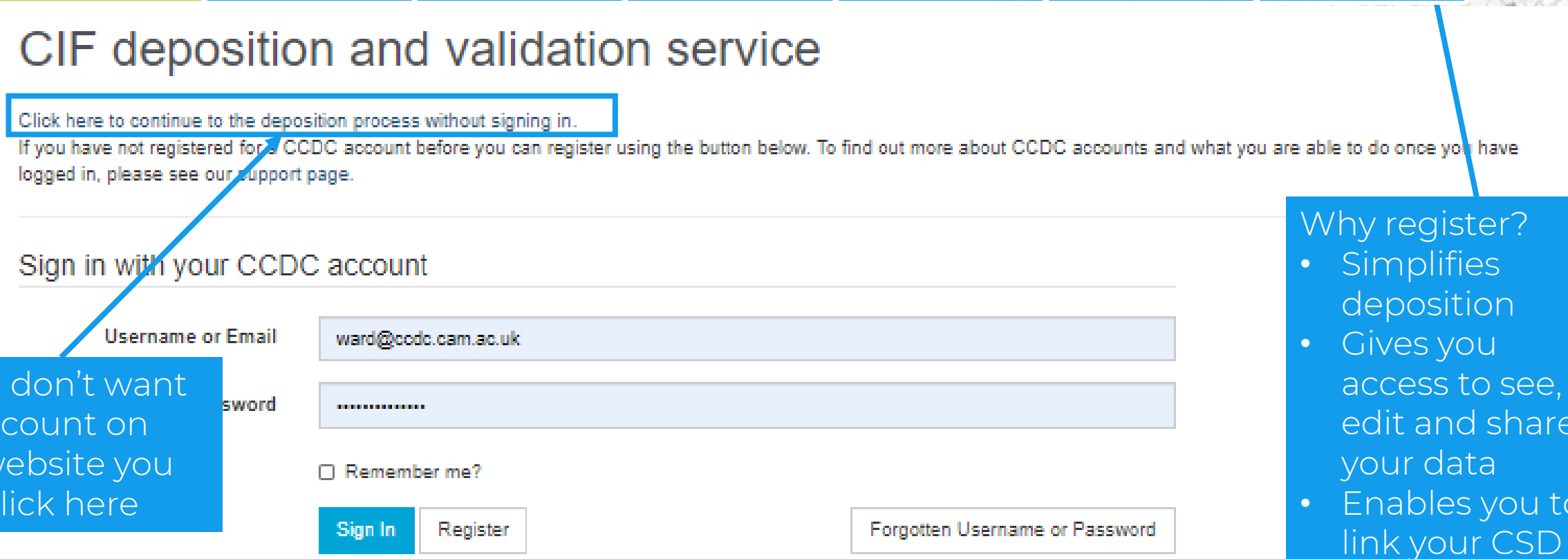
Raw data DOIs are for data created during data collection

- Diffraction frames e.g. .sfrm, .cbf, .img, etc

- We do not accept diffraction frame data during deposition
- But you can choose to upload your diffraction frames in another repository or institution repository
  - If you do publish your diffraction frames you should add the DOI to “Raw data DOI”
- Raw data DOI should not include the DOI of associated articles

CCDC

Deposition



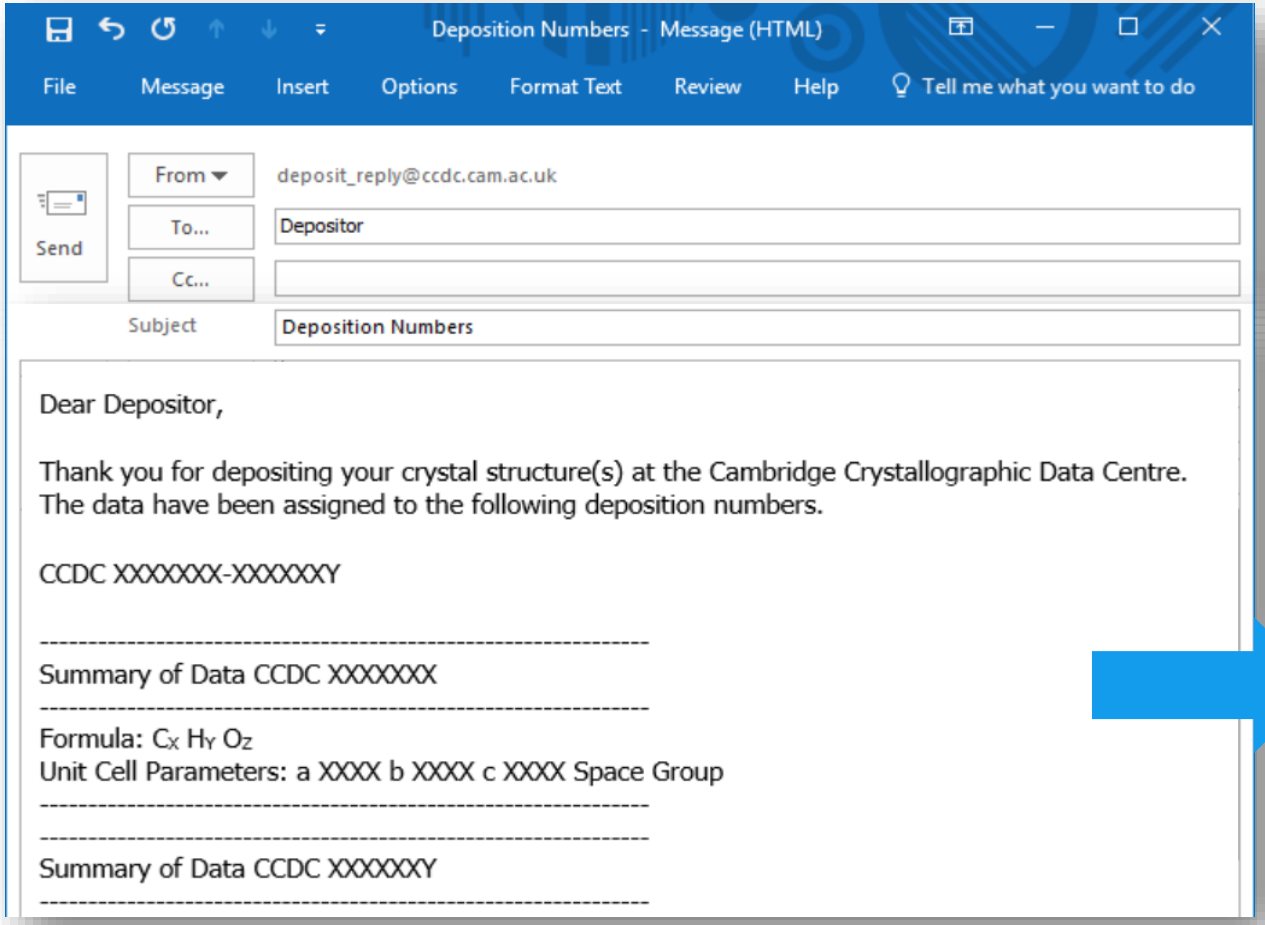
If you don't want an account on our website you can click here

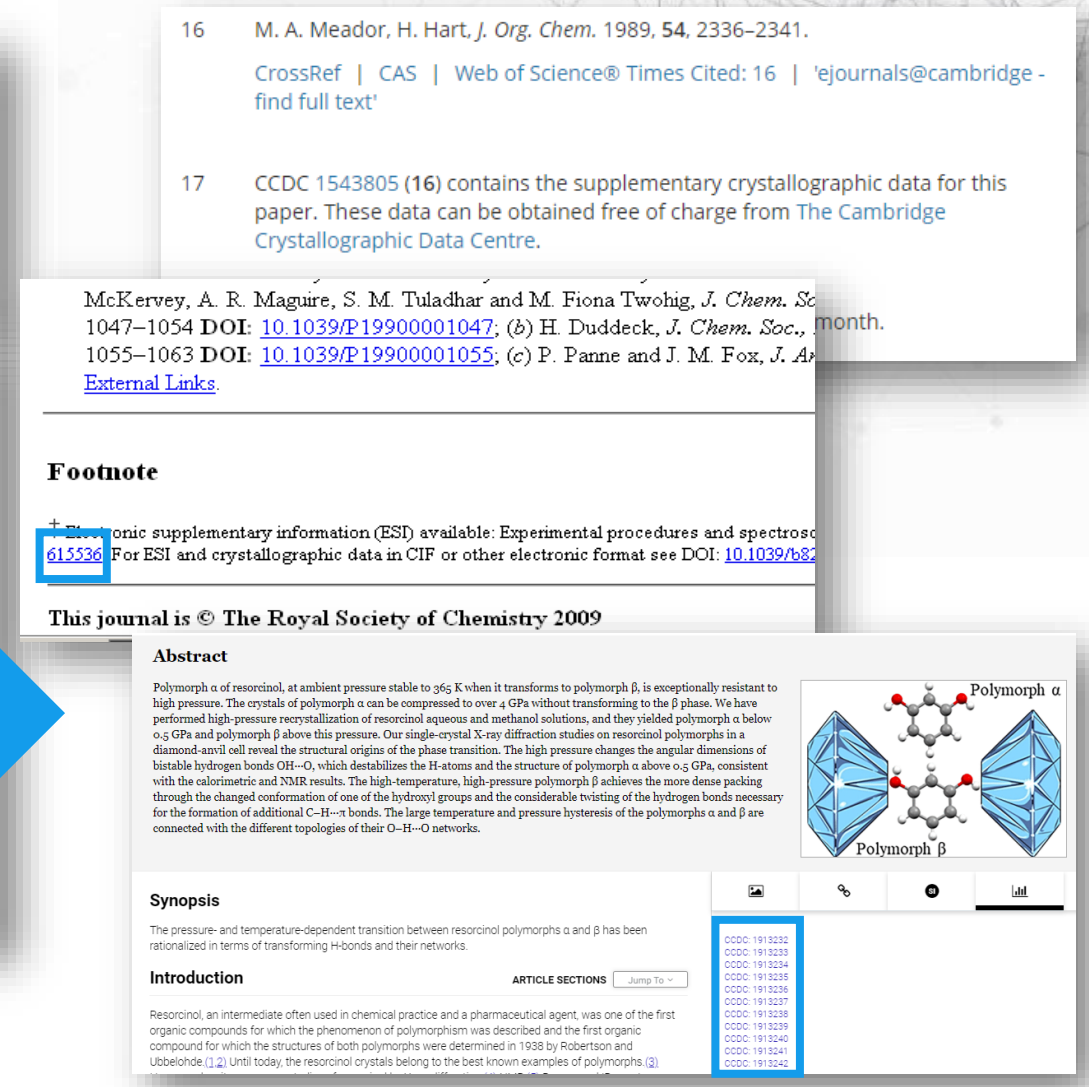
Why register?

- Simplifies deposition
- Gives you access to see, edit and share your data
- Enables you to link your CSD licence to use WebCSD

Dataset identifiers

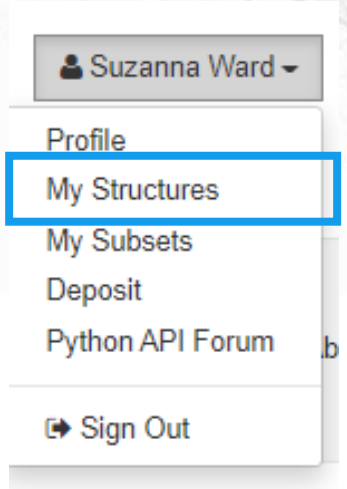
Deposition Numbers should be included in manuscripts





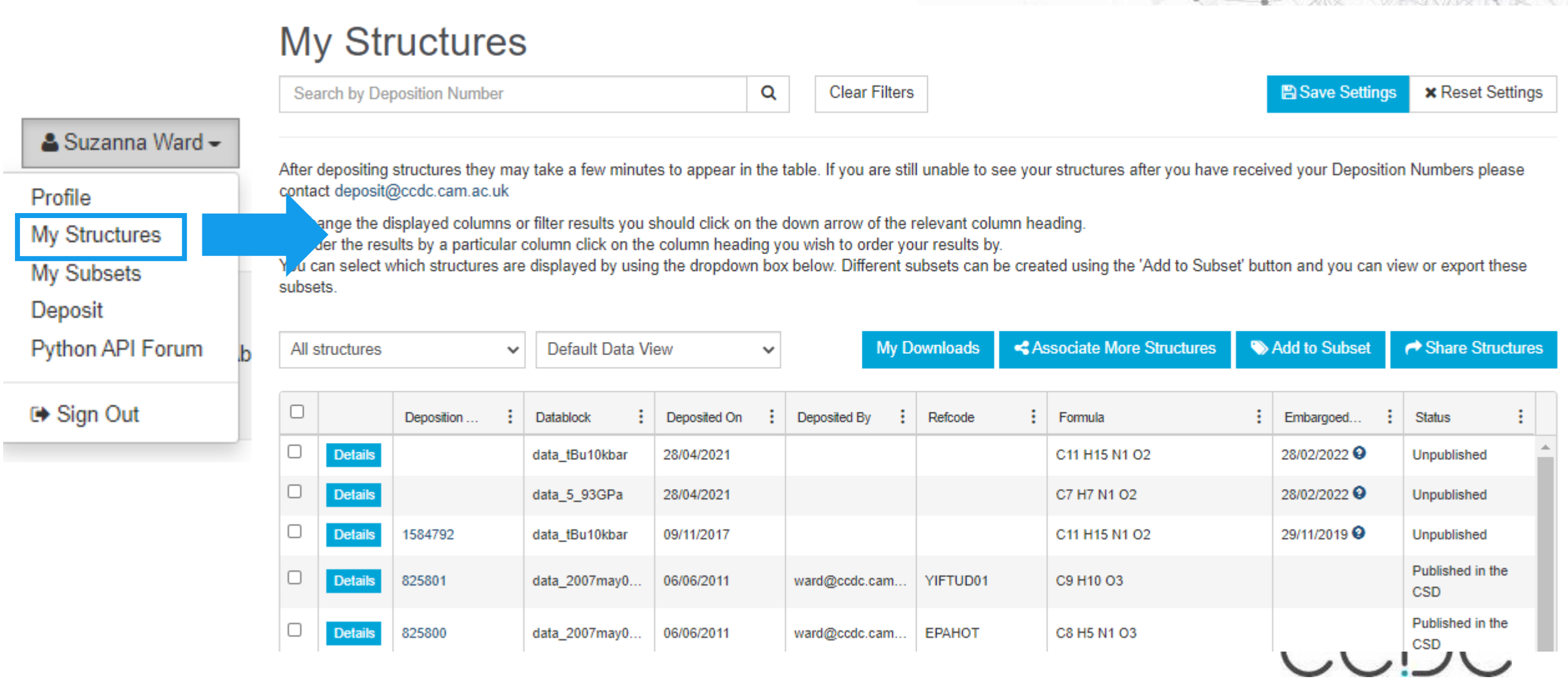
What happens to your data after deposition?

- It is stored privately until published in an associated scientific article or you decide to publish directly through the CSD or ICSD
- If you have registered on our website you can access your data at any time using our My Structures service
  - Enables you to view and retrieve datasets
  - Enables you to edit and update datasets
  - Enables you to share your datasets with co-workers
  - Enables you to create your own institutional database
  - Enables you to publish your data directly through the CSD/ICSD
- If you have not published your data after one year, we will email you to ask if you would like to publish it



CCDC

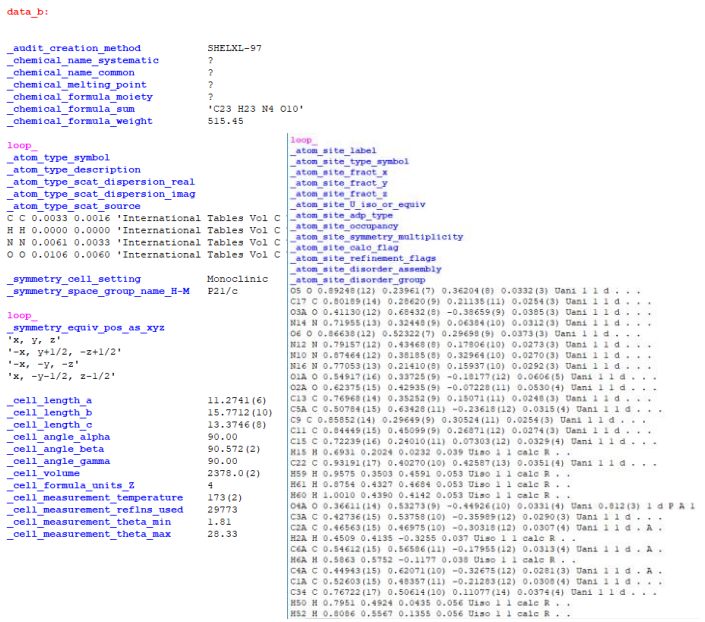
My Structures – view, edit, retrieve, share



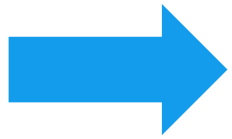
Curating the data into the CSD and ICSD

After publication your data is enhanced and curated

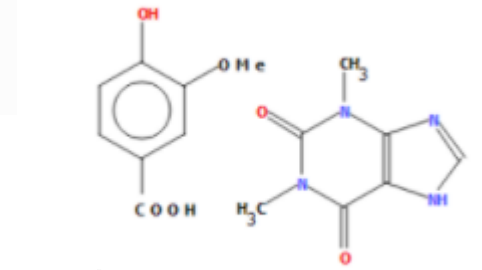
- Every entry undergoes both automated and manual processes
- Enhanced and curated by experts at CCDC
- Ensuring accessibility, discoverability and re-usability of the data

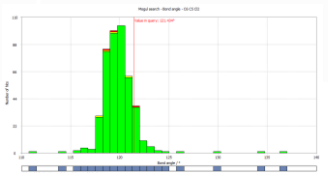


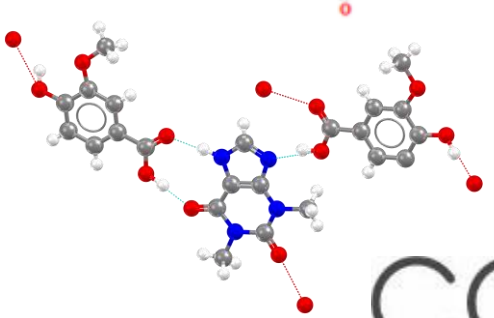
```
data_b:

_audit_creation_method            SHELXL-97
_chemical_name_systematic         ?
_chemical_name_common             ?
_chemical_melting_point           ?
_chemical_formula_moiety          ?
_chemical_formula_sum             'C23 H23 N4 O10'
_chemical_formula_weight          515.45

loop_
_atom_type_symbol
_atom_type_description
_atom_type_scat_dispersion_real
_atom_type_scat_dispersion_imag
_atom_type_scat_source
C C 0.0033 0.0016 'International Tables Vol C
H H 0.0000 0.0000 'International Tables Vol C
N N 0.0061 0.0033 'International Tables Vol C
O O 0.0106 0.0060 'International Tables Vol C

_symmetry_cell_setting            Monoclinic
_symmetry_space_group_name_H-M    P21/c

loop_
_symmetry_equiv_pos_as_xyz
'x, y, z'
'-x, y+1/2, -z+1/2'
'-x, -y, -z'
'x, -y-1/2, z-1/2'

_cell_length_a                    11.2741(6)
_cell_length_b                    15.7712(10)
_cell_length_c                    13.3746(8)
_cell_angle_alpha                 90.00
_cell_angle_beta                  90.572(2)
_cell_angle_gamma                 90.00
_cell_volume                      2378.0(2)
_cell_formula_units_Z             4
_cell_measurement_temperature     173(2)
_cell_measurement_reflns_used     29773
_cell_measurement_theta_min       1.81
_cell_measurement_theta_max       28.33

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
O5 O 0.89248(12) 0.23961(7) 0.36204(8) 0.0332(3) Uani 1 1 d . . .
C17 C 0.80189(14) 0.28620(9) 0.21135(11) 0.0254(3) Uani 1 1 d . . .
O3A O 0.41130(12) 0.69432(8) -0.38659(9) 0.0385(3) Uani 1 1 d . . .
N14 N 0.71955(13) 0.32448(9) 0.06304(10) 0.0312(3) Uani 1 1 d . . .
O6 O 0.86638(12) 0.52322(7) 0.29698(9) 0.0373(3) Uani 1 1 d . . .
N12 N 0.79157(12) 0.43468(8) 0.17806(10) 0.0273(3) Uani 1 1 d . . .
N10 N 0.87464(12) 0.38185(8) 0.32964(10) 0.0270(3) Uani 1 1 d . . .
N16 N 0.77053(13) 0.21410(8) 0.15937(10) 0.0292(3) Uani 1 1 d . . .
O1A O 0.54917(16) 0.33725(9) -0.18177(12) 0.0606(5) Uani 1 1 d . . .
O2A O 0.62375(15) 0.42935(9) -0.07228(11) 0.0530(4) Uani 1 1 d . . .
C13 C 0.76968(14) 0.35252(9) 0.15071(11) 0.0248(3) Uani 1 1 d . . .
C5A C 0.50784(15) 0.63428(11) -0.23618(12) 0.0315(4) Uani 1 1 d . . .
C9 C 0.85852(14) 0.29649(9) 0.30524(11) 0.0254(3) Uani 1 1 d . . .
C11 C 0.84449(15) 0.45099(9) 0.26871(12) 0.0274(3) Uani 1 1 d . . .
C15 C 0.72239(16) 0.24010(11) 0.07303(12) 0.0329(4) Uani 1 1 d . . .
H15 H 0.6931 0.2024 0.0232 0.039 Uiso 1 1 calc R . .
C22 C 0.93191(17) 0.40270(10) 0.42587(13) 0.0351(4) Uani 1 1 d . . .
H59 H 0.9575 0.3503 0.4591 0.053 Uiso 1 1 calc R . .
H61 H 0.8754 0.4327 0.4684 0.053 Uiso 1 1 calc R . .
H60 H 1.0010 0.4390 0.4142 0.053 Uiso 1 1 calc R . .
O4A O 0.36411(14) 0.53273(9) -0.44926(10) 0.0331(4) Uani 0.812(3) 1 d P A 1
C3A C 0.42736(15) 0.53758(10) -0.35989(12) 0.0290(3) Uani 1 1 d . . .
C2A C 0.46563(15) 0.46975(10) -0.30318(12) 0.0307(4) Uani 1 1 d . A .
H2A H 0.4509 0.4135 -0.3255 0.037 Uiso 1 1 calc R . .
C4A C 0.54612(15) 0.56586(11) -0.17955(12) 0.0313(4) Uani 1 1 d . A .
H4A H 0.5863 0.5752 -0.1177 0.038 Uiso 1 1 calc R . .
C4A C 0.44943(15) 0.62071(10) -0.32675(12) 0.0281(3) Uani 1 1 d . A .
C1A C 0.52603(15) 0.48357(11) -0.21283(12) 0.0308(4) Uani 1 1 d . . .
C34 C 0.76722(17) 0.50614(10) 0.11077(14) 0.0374(4) Uani 1 1 d . . .
H50 H 0.7951 0.4924 0.0435 0.056 Uiso 1 1 calc R . .
H52 H 0.8086 0.5567 0.1355 0.056 Uiso 1 1 calc R . .
```



Author(s)	A.Jacobs, F.M.A.Noa
Reference	CrystEngComm (2015), 17, 98
Publication DOI	10.1039/C4CE01795A
Deposition	CCDC 1022107
Formula	$2(C_8H_8O_4)$, $C_7H_8N_4O_2$
Compound	bis(4-Hydroxy-3-methoxybenzoic acid) 1,3-dimethyl-3,7-dihydro-1H-purine-2,6-dione
Synonym	bis(Vanillic acid) theophylline
Spacegroup	**Name:** P21/c **Number:** 14
Cell	**a:** 11.274(<1) **b:** 15.771(1) **c:** 13.375(<1) **alpha:** 90.00 **beta:** 90.57(<1) **gamma:** 90.00 **Volume:** 2377.967
Reduced Cell	**a:** 11.274 **b:** 13.375 **c:** 15.771 **alpha:** 90.00 **beta:** 90.00 **gamma:** 90.57 **Volume:** 2377.967
Molecular Volume	594.492
Chemical Units	2
Z, Z'	**Z:** 4.0 **Z':** 1.0
R-Factor (%)	4.7
Disorder	C7A,O4A and C9A,O5A disordered over two sites with occupancies 0.812:0.188.
Temperature (K)	173
Density	**CCDC:** 1.443 **Author:** 1.44
Intensity Meas	diffractometer
Average Sigma (C-C)	0.001-0.005Å
Colour	colorless
Habit	rectangle
Recryst. Solvent	water







CCDC

Where to find more guidance

The CCDC CIF Deposition Guidelines

This page is designed to help you deposit your data, check your CIFs and provide you more information as you go through our deposition process. Click on the boxes below to navigate to the section of interest.

Basic guidelines	Required CIF fields	Authorship guidelines	Checking CIFs
Correcting CIFs	Useful resources	Structure Factors	Translated guidelines

Basic Guidelines

When preparing your CIF (What is a 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 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.

In this section

- The CCDC CIF Deposition Guidelines.pdf
- Structure factors: what are they and why share them?
- The Benefits of Data Sharing

Related Topics

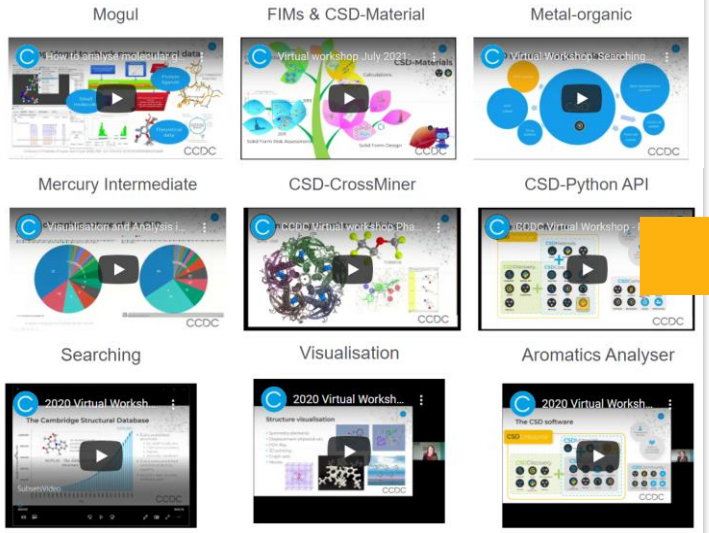
- Deposit a Structure
- Structure Deposition Information
- Correcting CIFs
- CSD Communications

https://www.ccdc.cam.ac.uk/Community/depositastructure/cif-deposition-guidelines/

CCDC

Keen to learn more about CSD deposition?

CCDC Virtual Workshops



Next series Nov 2021

November 2021 series

Session 1 – Depositing data into the CSD

- 2nd November
- 90 min interactive session
- Expert tutors on hand to help
- Spaces limited so register now!

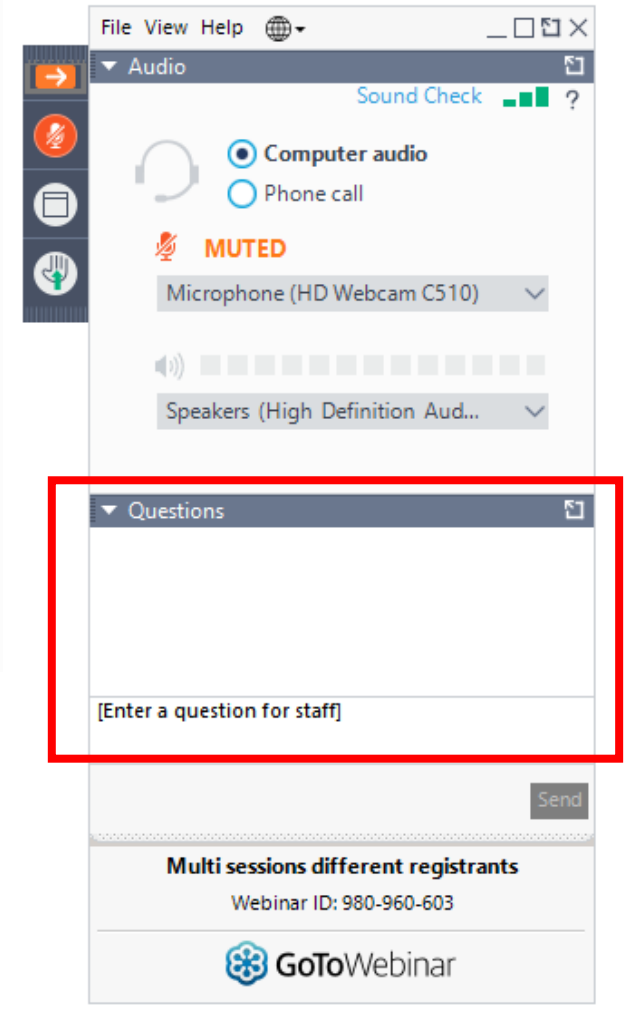
Show one — Try one — Explore more

https://www.ccdc.cam.ac.uk/News/Events/

CCDC

Q&A

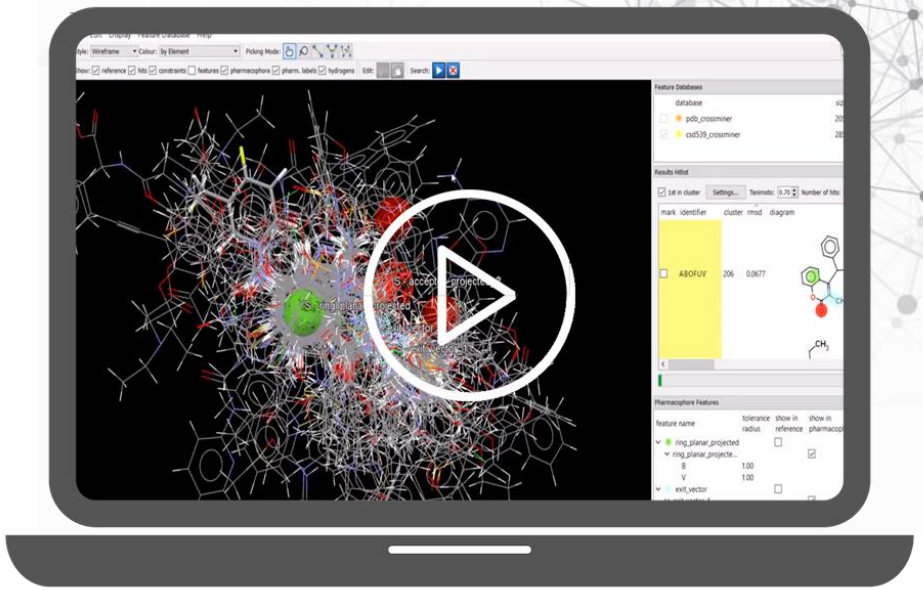
- Type your questions in the box as shown



CCDC

Next What's Up Webinar

- Next webinar: **November 18th**
  - CSD-CrossMiner: new features from 2021.2 release
  - DASH: what has changed?



- Follow us on social media
- Send us your ideas and news

  hello@ccdc.cam.ac.uk

Thank you

hello@ccdc.cam.ac.uk

The Cambridge Crystallographic Data Centre
12 Union Road, Cambridge CB2 1EZ, United Kingdom
Registered Charity No. 800579

CCDC