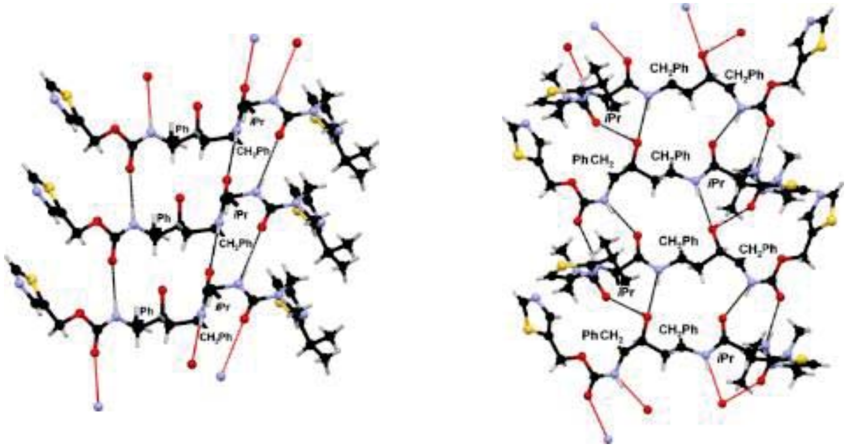
Database-driven discovery

Suzanna Ward and Eric Rogers

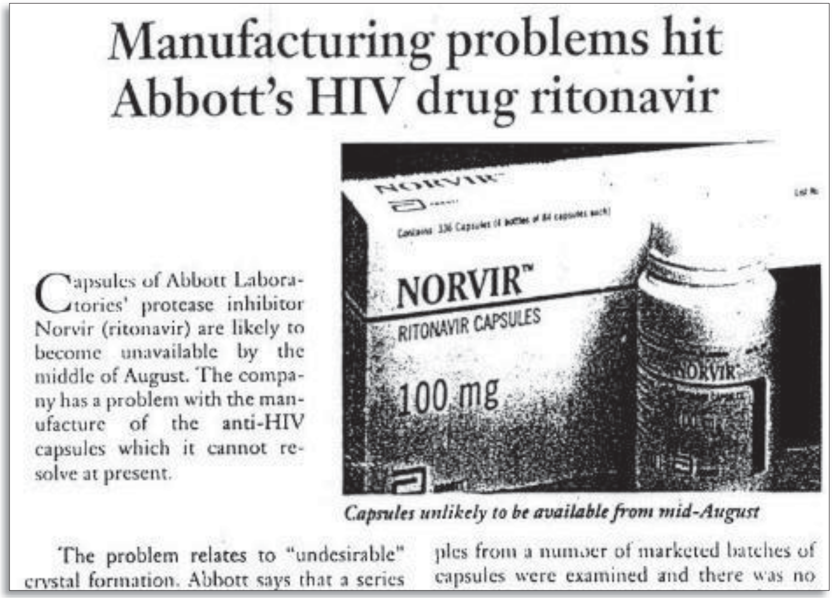
The Cambridge Crystallographic Data Centre

Creation of the CSD

Can Structural Knowledge Mitigate Risk?



Different interactions → Different solubility Different stability



Manufacturing problems hit Abbott's HIV drug ritonavir

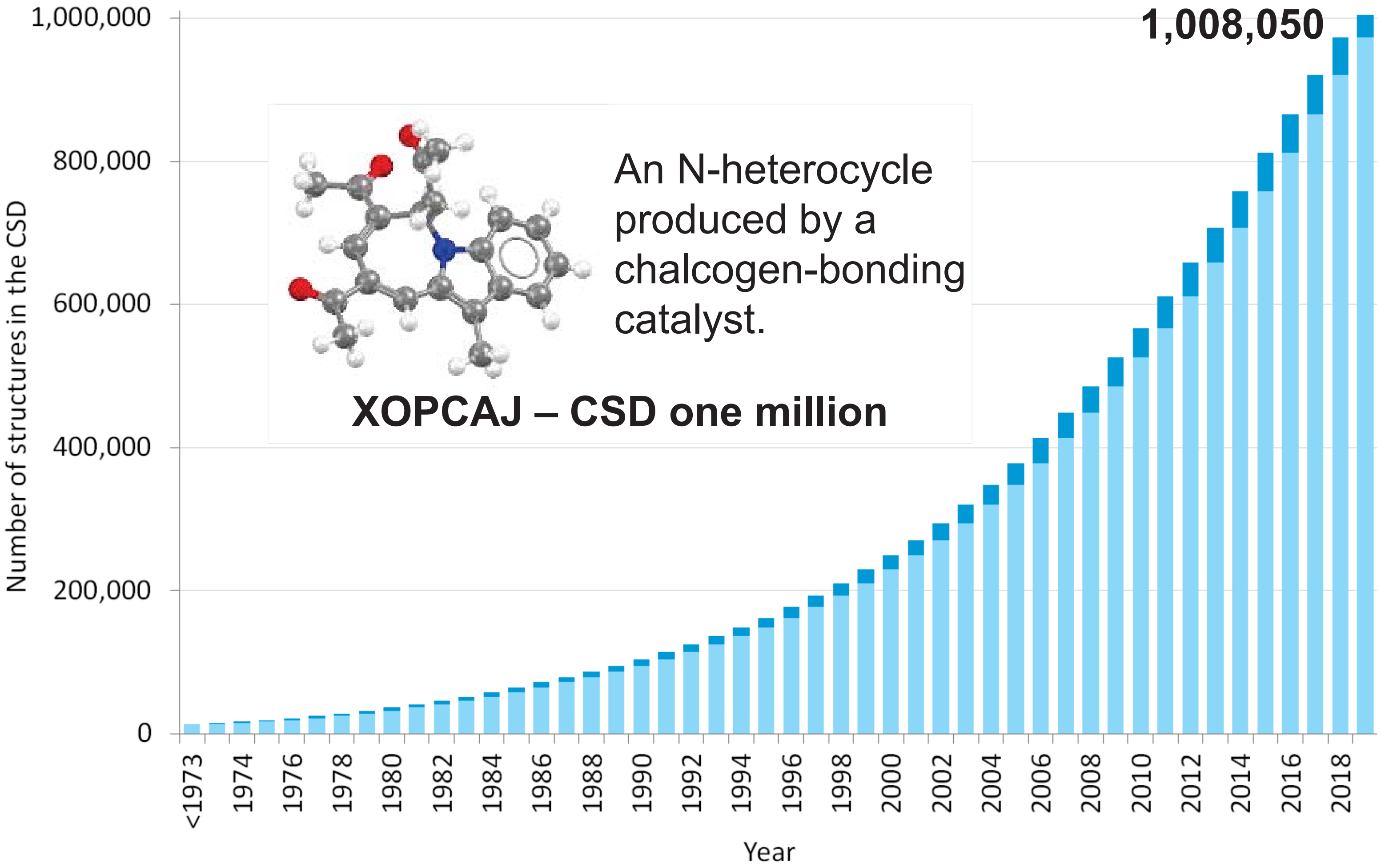
Capsules of Abbott Laboratories' protease inhibitor Norvir (ritonavir) are likely to become unavailable by the middle of August. The company has a problem with the manufacture of the anti-HIV capsules which it cannot resolve at present.

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

Bauer *et al*. *Pharm. Res.,* (2001) 859, DOI:10.1039/B910882C

The Cambridge Structural Database (CSD)

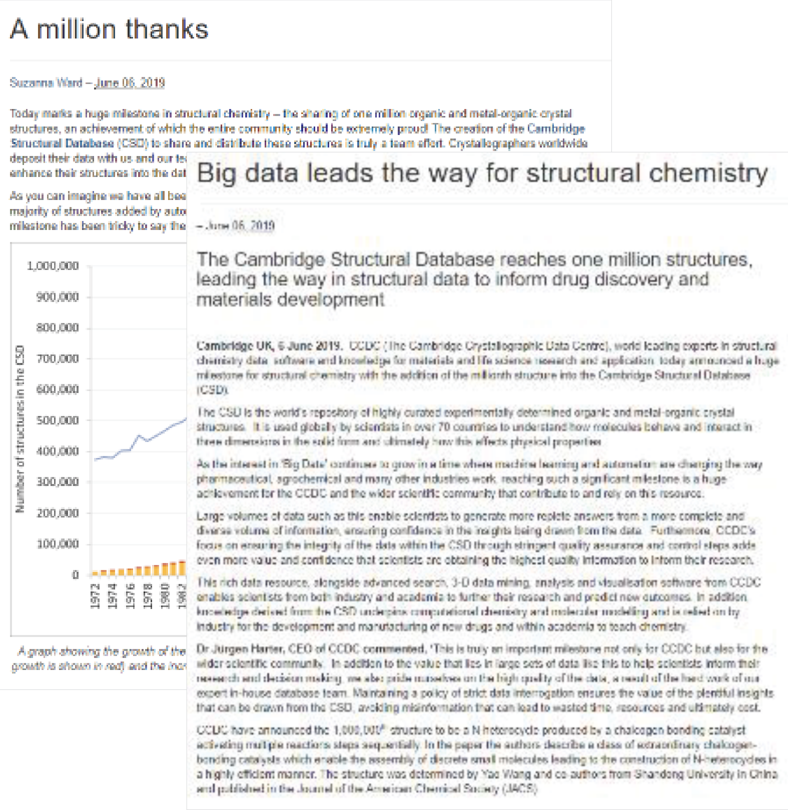


An N-heterocycle produced by a chalcogen-bonding catalyst.

XOPCAJ – CSD one million

- Every published structure
  - Inc. ASAP & early view
  - *CSD Communications*
  - Patents
  - University repositories
- Every entry enriched and annotated by experts
- Discoverability of data and knowledge
- Sustainable for over 54 years

CSD One Million



A million thanks

Suzanna Ward – June 06, 2019

Today marks a huge milestone in structural chemistry – the sharing of one million organic and metal-organic crystal structures, an achievement of which the entire community should be extremely proud! The creation of the Cambridge Structural Database (CSD) to share and distribute these structures is truly a team effort. Crystallographers worldwide deposit their data with us and our te
enhance their structures into the dat

As you can imagine we have all bee
majority of structures added by auto
milestone has been tricky to say the

A graph showing the growth of the
growth is shown in red) and the inc

Big data leads the way for structural chemistry

– June 06, 2019

The Cambridge Structural Database reaches one million structures, leading the way in structural data to inform drug discovery and materials development

Cambridge UK, 6 June 2019. CCDC (The Cambridge Crystallographic Data Centre), world leading experts in structural chemistry data, software and knowledge for materials and life science research and application, today announced a huge milestone for structural chemistry with the addition of the millionth structure into the Cambridge Structural Database (CSD).

The CSD is the world's repository of highly curated experimentally determined organic and metal-organic crystal structures. It is used globally by scientists in over 70 countries to understand how molecules behave and interact in three dimensions in the solid form and ultimately how this effects physical properties.

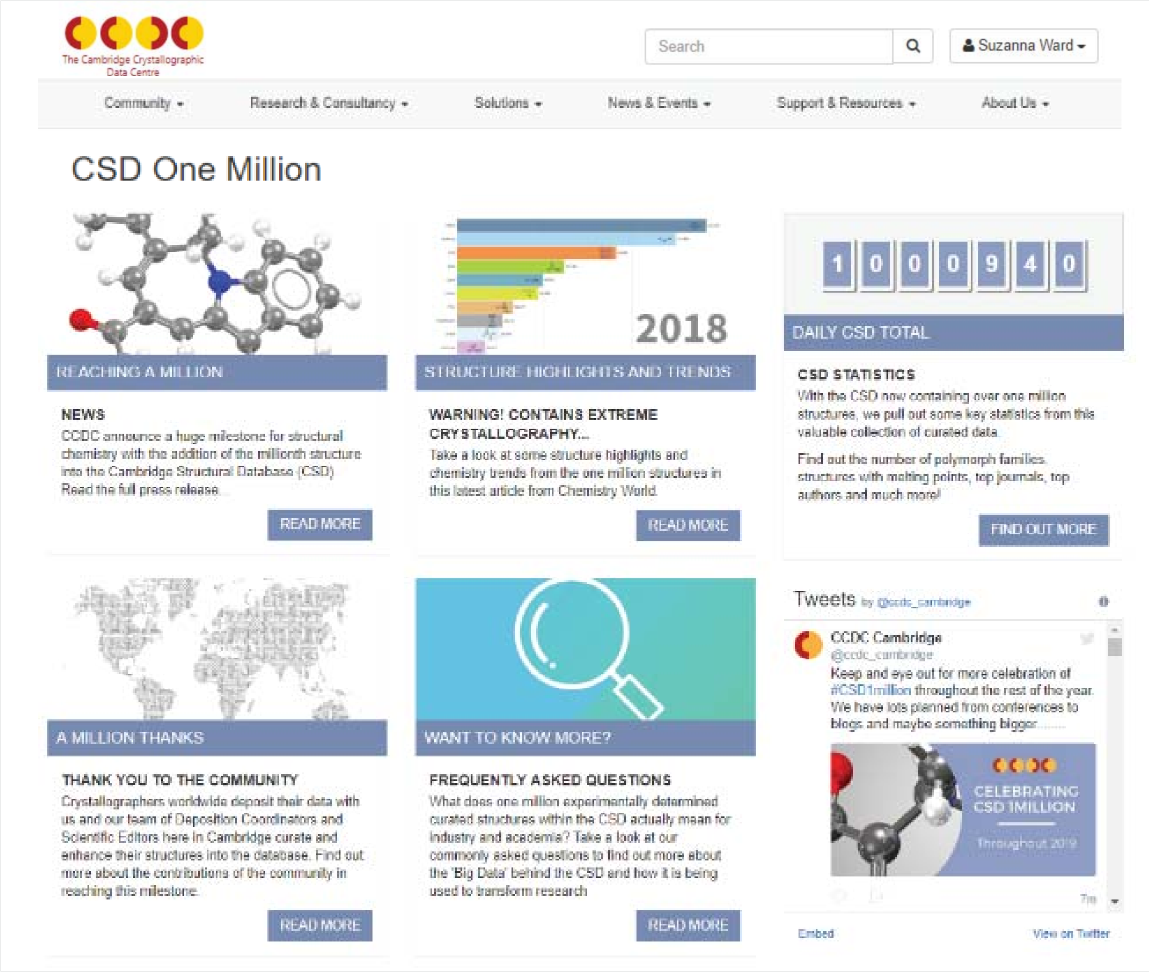
As the interest in 'Big Data' continues to grow in a time where machine learning and automation are changing the way pharmaceutical, agrochemical and many other industries work, reaching such a significant milestone is a huge achievement for the CCDC and the wider scientific community that contribute to and rely on this resource.

Large volumes of data such as this enable scientists to generate more replete answers from a more complete and diverse volume of information, ensuring confidence in the insights being drawn from the data. Furthermore, CCDC's focus on ensuring the integrity of the data within the CSD through stringent quality assurance and control steps adds even more value and confidence that scientists are obtaining the highest quality information to inform their research.

This rich data resource, alongside advanced search, 3-D data mining, analysis and visualisation software from CCDC enables scientists from both industry and academia to further their research and predict new outcomes. In addition, knowledge derived from the CSD underpins computational chemistry and molecular modelling and is relied on by industry for the development and manufacturing of new drugs and within academia to teach chemistry.

Dr Jürgen Harter, CEO of CCDC commented, 'This is truly an important milestone not only for CCDC but also for the wider scientific community. In addition to the value that lies in large sets of data like this to help scientists inform their research and decision making, we also pride ourselves on the high quality of the data, a result of the hard work of our expert in-house database team. Maintaining a policy of strict data interrogation ensures the value of the plentiful insights that can be drawn from the CSD, avoiding misinformation that can lead to wasted time, resources and ultimately cost.

CCDC have announced the 1,000,000th structure to be a N heterocycle produced by a chalcogen bonding catalyst activating multiple reactions steps sequentially. In the paper the authors describe a class of extraordinary chalcogen-bonding catalysts which enable the assembly of discrete small molecules leading to the construction of N-heterocycles in a highly efficient manner. The structure was determined by Yao Wang and co-authors from Shandong University in China and published in the Journal of the American Chemical Society (JACS).



CSD One Million

1 0 0 0 9 4 0

REACHING A MILLION

STRUCTURE HIGHLIGHTS AND TRENDS

DAILY CSD TOTAL

NEWS

CCDC announce a huge milestone for structural chemistry with the addition of the millionth structure into the Cambridge Structural Database (CSD). Read the full press release.

READ MORE

WARNING! CONTAINS EXTREME CRYSTALLOGRAPHY...

Take a look at some structure highlights and chemistry trends from the one million structures in this latest article from Chemistry World.

READ MORE

CSD STATISTICS

With the CSD now containing over one million structures, we pull out some key statistics from this valuable collection of curated data.

Find out the number of polymorph families, structures with melting points, top journals, top authors and much more!

FIND OUT MORE

A MILLION THANKS

WANT TO KNOW MORE?

THANK YOU TO THE COMMUNITY

Crystallographers worldwide deposit their data with us and our team of Deposition Coordinators and Scientific Editors here in Cambridge curate and enhance their structures into the database. Find out more about the contributions of the community in reaching this milestone.

READ MORE

FREQUENTLY ASKED QUESTIONS

What does one million experimentally determined curated structures within the CSD actually mean for industry and academia? Take a look at our commonly asked questions to find out more about the 'Big Data' behind the CSD and how it is being used to transform research.

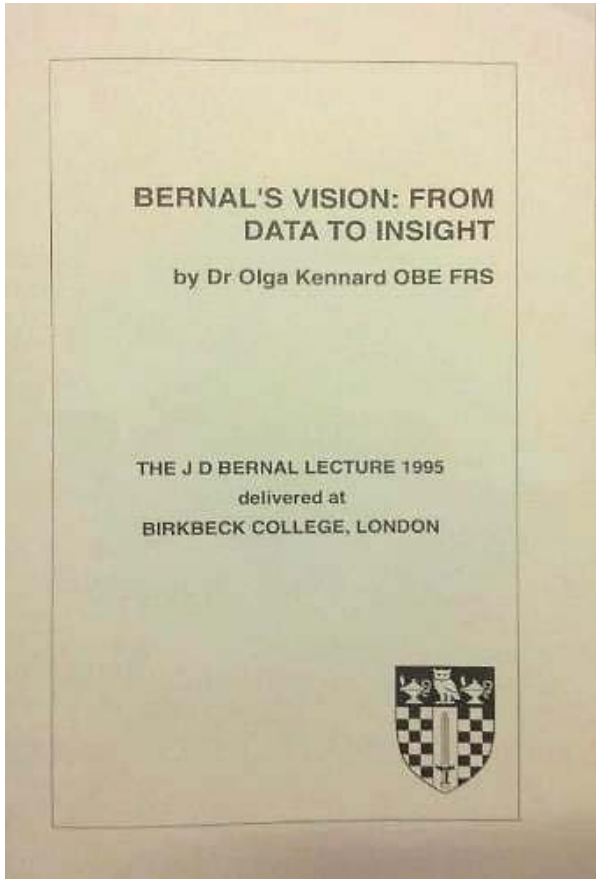
READ MORE

Tweets by @ccdc_cambridge

CCDC Cambridge
@ccdc_cambridge

Keep and eye out for more celebration of #CSD1million throughout the rest of the year. We have lots planned from conferences to blogs and maybe something bigger........

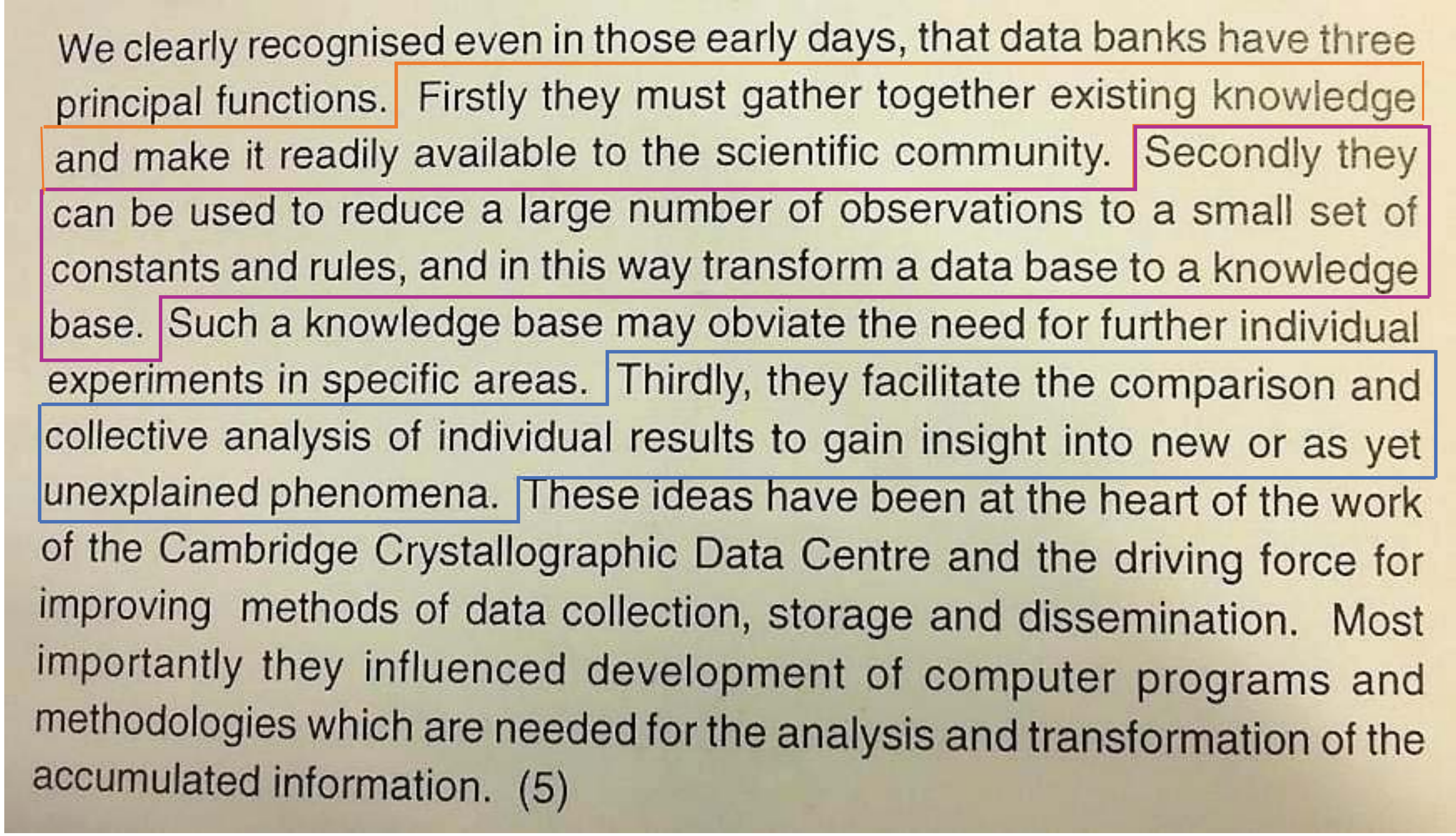
The Vision



BERNAL'S VISION: FROM DATA TO INSIGHT

by Dr Olga Kennard OBE FRS

THE J D BERNAL LECTURE 1995
delivered at
BIRKBECK COLLEGE, LONDON



We clearly recognised even in those early days, that data banks have three principal functions. Firstly they must gather together existing knowledge and make it readily available to the scientific community. Secondly they can be used to reduce a large number of observations to a small set of constants and rules, and in this way transform a data base to a knowledge base. Such a knowledge base may obviate the need for further individual experiments in specific areas. Thirdly, they facilitate the comparison and collective analysis of individual results to gain insight into new or as yet unexplained phenomena. These ideas have been at the heart of the work of the Cambridge Crystallographic Data Centre and the driving force for improving methods of data collection, storage and dissemination. Most importantly they influenced development of computer programs and methodologies which are needed for the analysis and transformation of the accumulated information. (5)

Inside the CSD

Organic
43%

Metal-Organic
57%

At least one transition metal, lanthanide, actinide or any of Al, Ga, In, Tl, Ge, Sn, Pb, Sb, Bi, Po

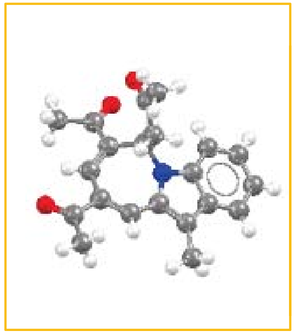
Not Polymeric
89%

Polymeric: 11%

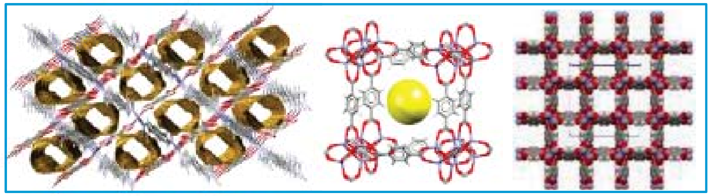
Single Component
56%

Multi Component
44%

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



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

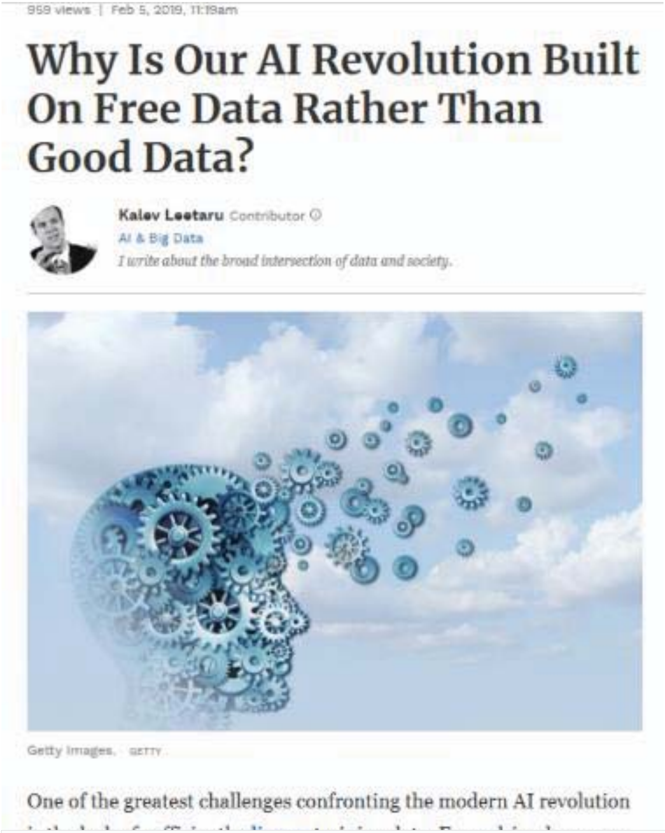


- **Additional data**
- 10,860 polymorph families
- 169,218 melting points
- 840,667 crystal colours
- 700,002 crystal shapes
- 23,622 bioactivity details
- 9,740 natural source data
- > 250,000 oxidation states

- **Links/subsets**
- Drugbank
- Druglike
- MOFs
- PDB ligands
- PubChem
- ChemSpider
- Pesticides

AI and Machine Learning

- AI and machine learning techniques are evolving rapidly
- But the consequences of using poor quality data can be far reaching
  - Incorrect scientific conclusions
  - Wasted investment and effort
  - A loss of trust
  - Ultimately poor business decisions.



959 views | Feb 5, 2019, 11:19am

Why Is Our AI Revolution Built On Free Data Rather Than Good Data?

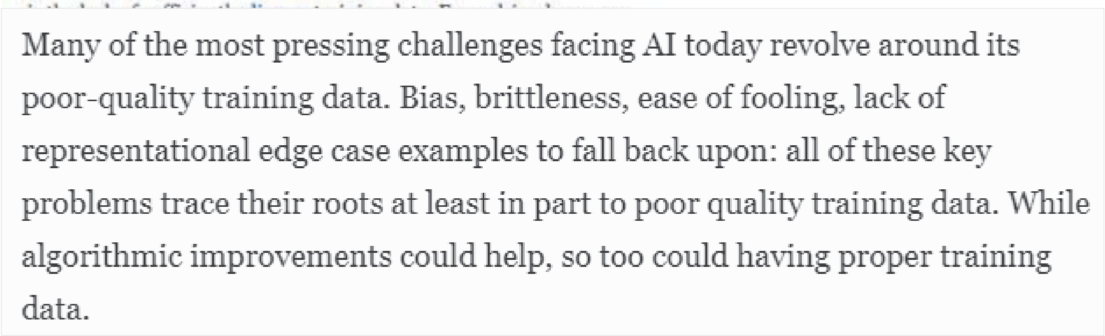
Kalev Leetaru Contributor

AI & Big Data

I write about the broad intersection of data and society.

Getty Images. GETTY

One of the greatest challenges confronting the modern AI revolution

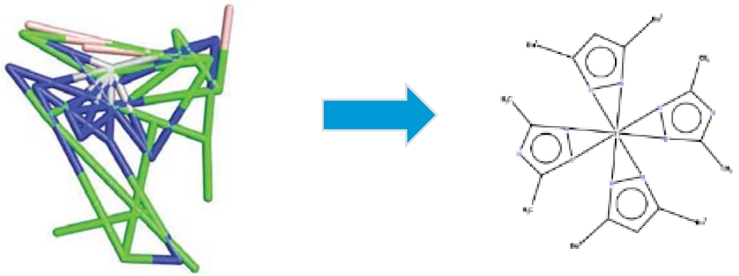


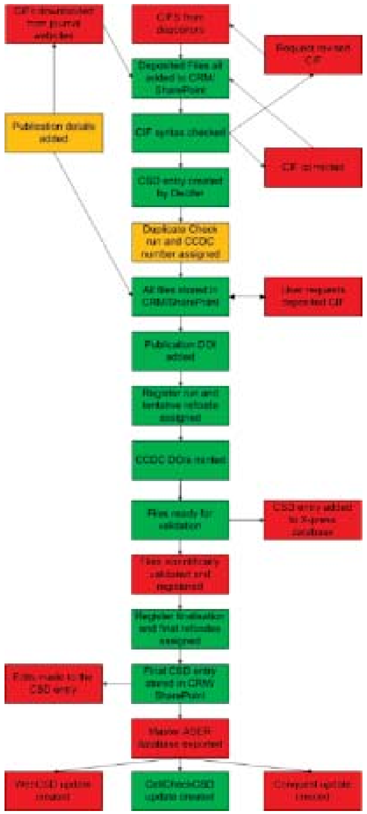
Many of the most pressing challenges facing AI today revolve around its poor-quality training data. Bias, brittleness, ease of fooling, lack of representational edge case examples to fall back upon: all of these key problems trace their roots at least in part to poor quality training data. While algorithmic improvements could help, so too could having proper training data.

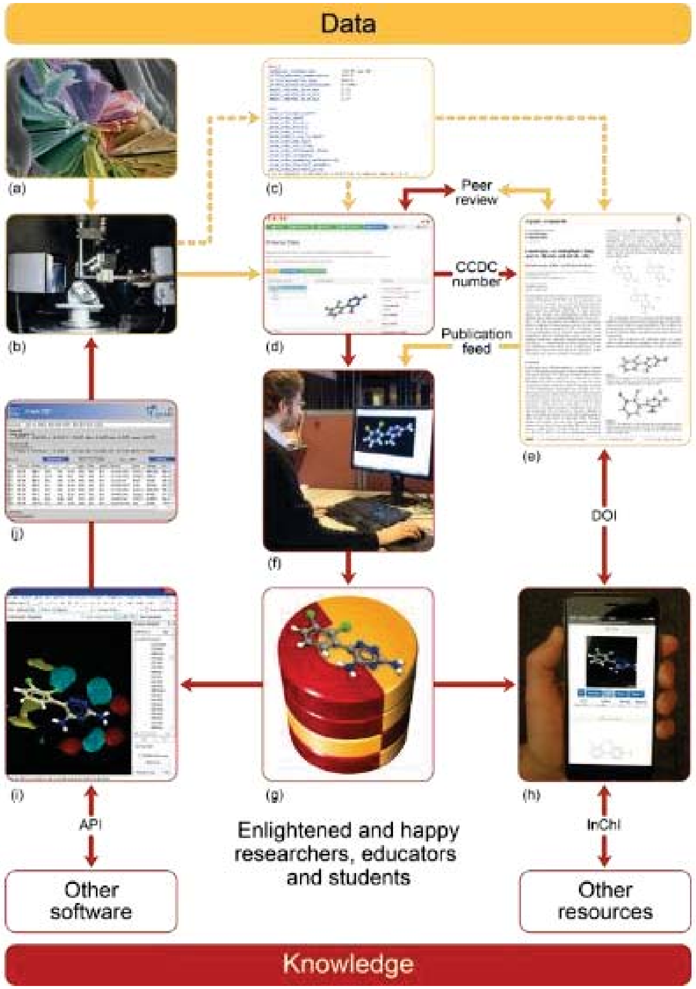
https://www.forbes.com/sites/kalevleetaru/2019/02/05/why-is-our-ai-revolution-built-on-free-data-rather-than-good-data/

Curating the CSD

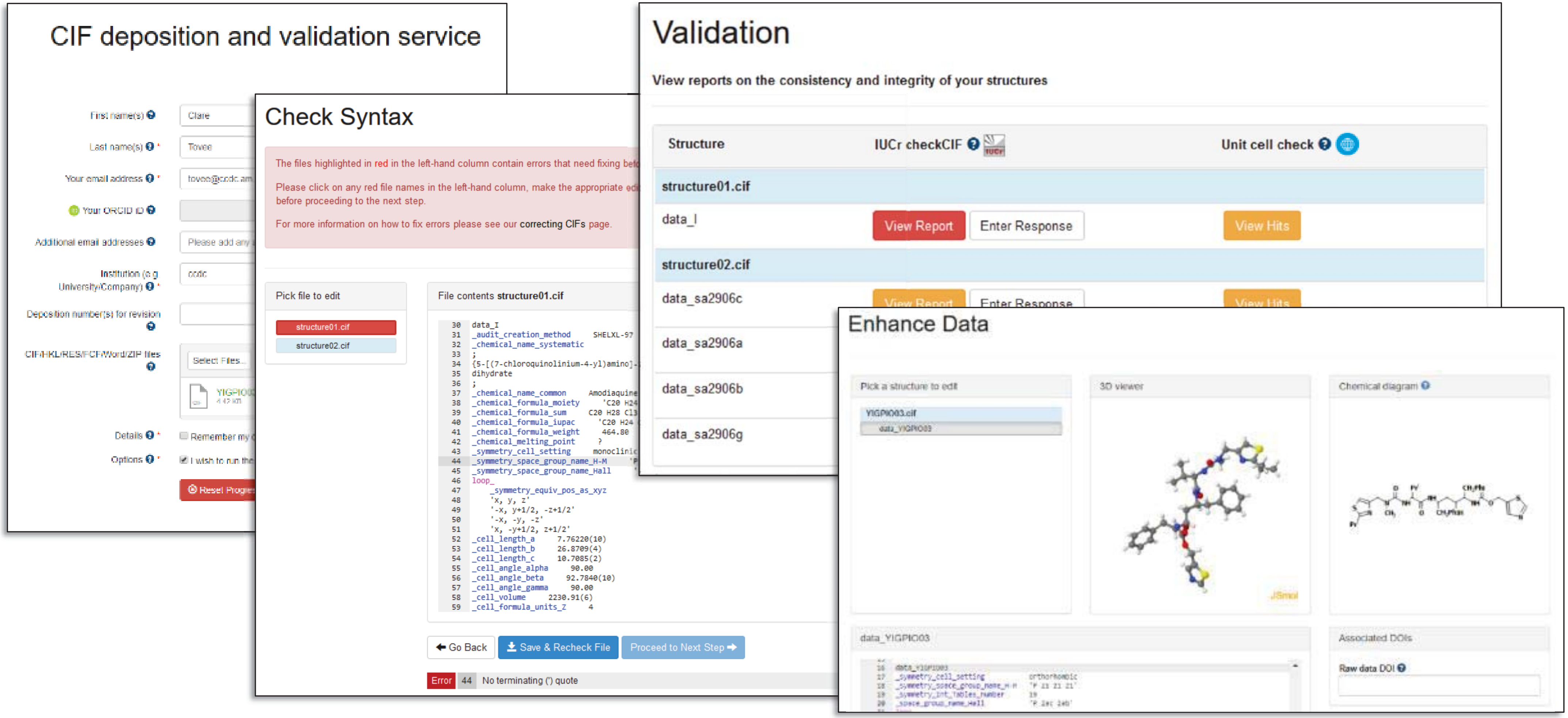
- Each dataset expertly curated
- Datasets enhanced
    - Chemical connectivity
    - Compound names
    - 2D chemical diagrams
    - Additional experimental data
    - Bibliographic information




CIFs downloaded from journal websites
CIFS from depositors
Request revised CIF
Deposited Files all added in CRM/SharePoint
Publication details added
CIF syntax checked
CIF corrected
CSD entry created by DeCifer
Duplicate Check run and CCDC number assigned
All files stored in CRM/SharePoint
User requests deposited CIF
Publication DOI added
Register run and tentative refcode assigned
CCDC DOIs minted
Files ready for validation
CSD entry added to X-press database
Files scientifically validated and registered
Register finalisation and final refcodes assigned
Final CSD entry stored in CRM/SharePoint
Edits made to the CSD entry
Master ASER database exported
WebCSD update created
CellCheckCSD update created
Conquest update created



Depositing the Data



CIF deposition and validation service

First name(s) Clare

Last name(s) * Tovee

Your email address * tovee@ccdc.am

Your ORCID iD

Additional email addresses Please add any

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

Deposition number(s) for revision

CIF/HKL/RES/FCF/Word/ZIP files Select Files... YIGPIO03 4.42 KB

Details * □ Remember my

Options * ☑ I wish to run the

Check Syntax

The files highlighted in red in the left-hand column contain errors that need fixing befo

Please click on any red file names in the left-hand column, make the appropriate edi 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**

```
data_I
_audit_creation_method     SHELXL-97
_chemical_name_systematic
;
{5-[(7-chloroquinolinium-4-yl)amino]-
dihydrate
;
_chemical_name_common     Amodiaquine
_chemical_formula_moiety     'C20 H24
_chemical_formula_sum     C20 H28 Cl3
_chemical_formula_iupac     'C20 H24
_chemical_formula_weight     464.80
_chemical_melting_point     ?
_symmetry_cell_setting     monoclinic
_symmetry_space_group_name_H-M     'P
_symmetry_space_group_name_Hall     '
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     7.76220(10)
_cell_length_b     26.8709(4)
_cell_length_c     10.7085(2)
_cell_angle_alpha     90.00
_cell_angle_beta     92.7840(10)
_cell_angle_gamma     90.00
_cell_volume     2230.91(6)
_cell_formula_units_Z     4
```

Error 44 No terminating (') quote

Validation

View reports on the consistency and integrity of your structures

Structure	IUCr checkCIF		Unit cell check
structure01.cif			
data_I	View Report	Enter Response	View Hits
structure02.cif			
data_sa2906c	View Report	Enter Response	View Hits
data_sa2906a			
data_sa2906b			
data_sa2906g			

Enhance Data

Pick a structure to edit

- YIGPIO03.cif
  - data_YIGPIO03

3D viewer

Chemical diagram

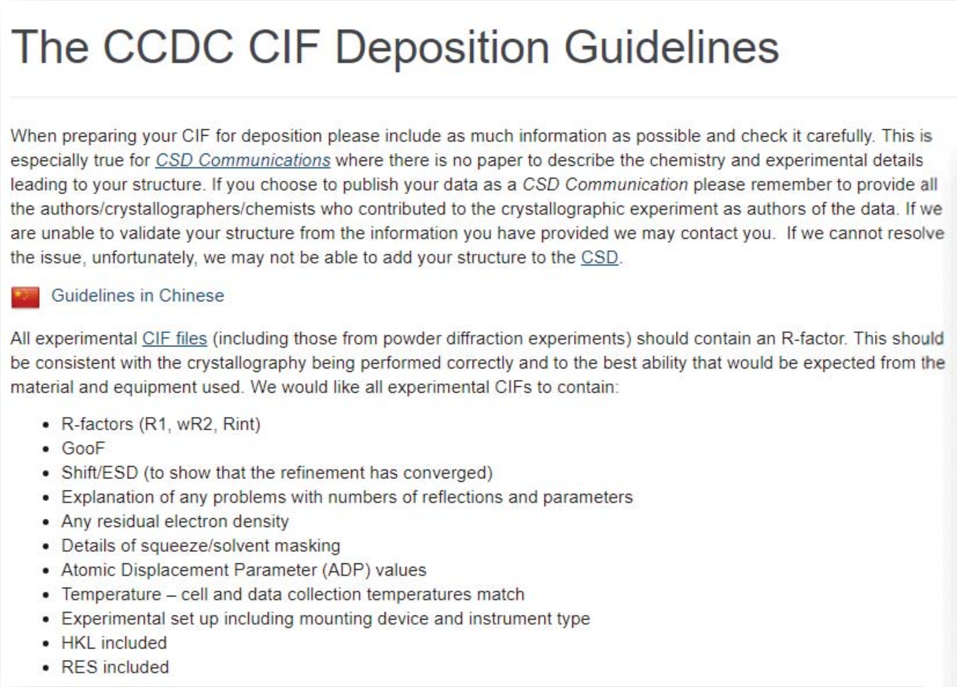
data_YIGPIO03

```
data_YIGPIO03
_symmetry_cell_setting     orthorhombic
_symmetry_space_group_name_H-M     'P 21 21 21'
_symmetry_Int_Tables_number     19
_space_group_name_Hall     'P 2ac 2ab'
```

Associated DOIs

Raw data DOI

Guidelines



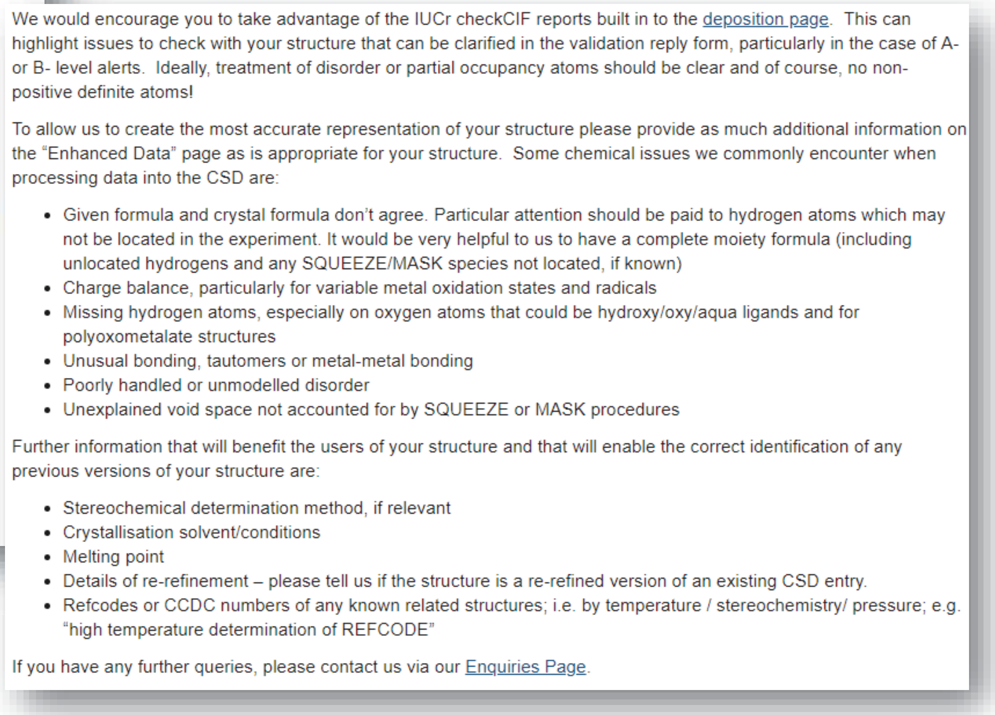
The CCDC CIF Deposition Guidelines

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.

Guidelines in Chinese

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)
- GooF
- Shift/ESD (to show that the refinement has converged)
- Explanation of any problems with numbers of reflections and parameters
- Any residual electron density
- Details of squeeze/solvent masking
- Atomic Displacement Parameter (ADP) values
- Temperature – cell and data collection temperatures match
- Experimental set up including mounting device and instrument type
- HKL included
- RES included



We would encourage you to take advantage of the IUCr checkCIF reports built in to the deposition page. This can highlight issues to check with your structure that can be clarified in the validation reply form, particularly in the case of A- or B- level alerts. Ideally, treatment of disorder or partial occupancy atoms should be clear and of course, no non-positive definite atoms!

To allow us to create the most accurate representation of your structure please provide as much additional information on the "Enhanced Data" page as is appropriate for your structure. Some chemical issues we commonly encounter when processing data into the CSD are:

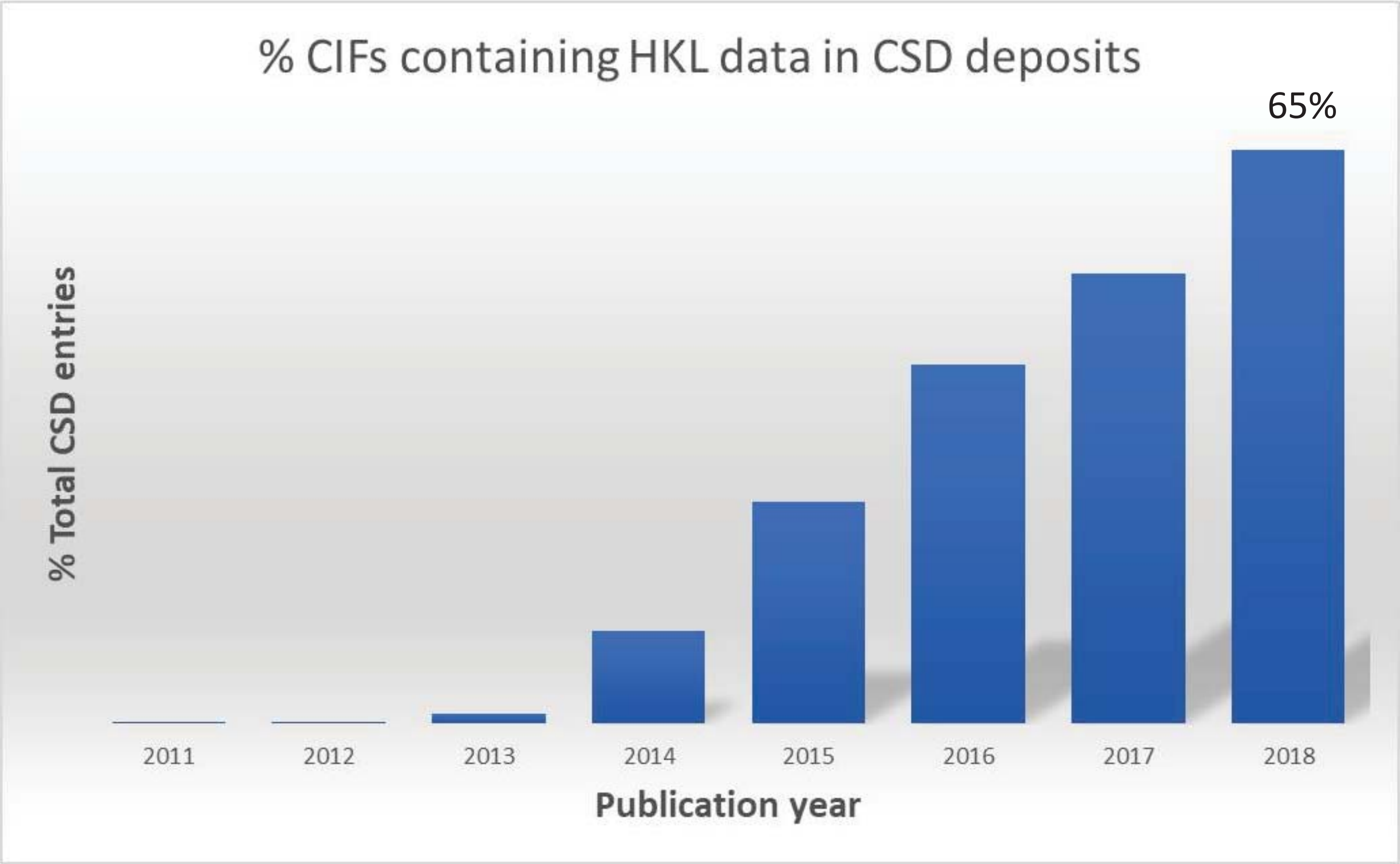
- Given formula and crystal formula don't agree. Particular attention should be paid to hydrogen atoms which may not be located in the experiment. It would be very helpful to us to have a complete moiety formula (including unlocated hydrogens and any SQUEEZE/MASK species not located, if known)
- Charge balance, particularly for variable metal oxidation states and radicals
- Missing hydrogen atoms, especially on oxygen atoms that could be hydroxy/oxy/aqua ligands and for polyoxometalate structures
- Unusual bonding, tautomers or metal-metal bonding
- Poorly handled or unmodelled disorder
- Unexplained void space not accounted for by SQUEEZE or MASK procedures

Further information that will benefit the users of your structure and that will enable the correct identification of any previous versions of your structure are:

- Stereochemical determination method, if relevant
- Crystallisation solvent/conditions
- Melting point
- Details of re-refinement – please tell us if the structure is a re-refined version of an existing CSD entry.
- Refcodes or CCDC numbers of any known related structures; i.e. by temperature / stereochemistry/ pressure; e.g. "high temperature determination of REFCODE"

If you have any further queries, please contact us via our Enquiries Page.

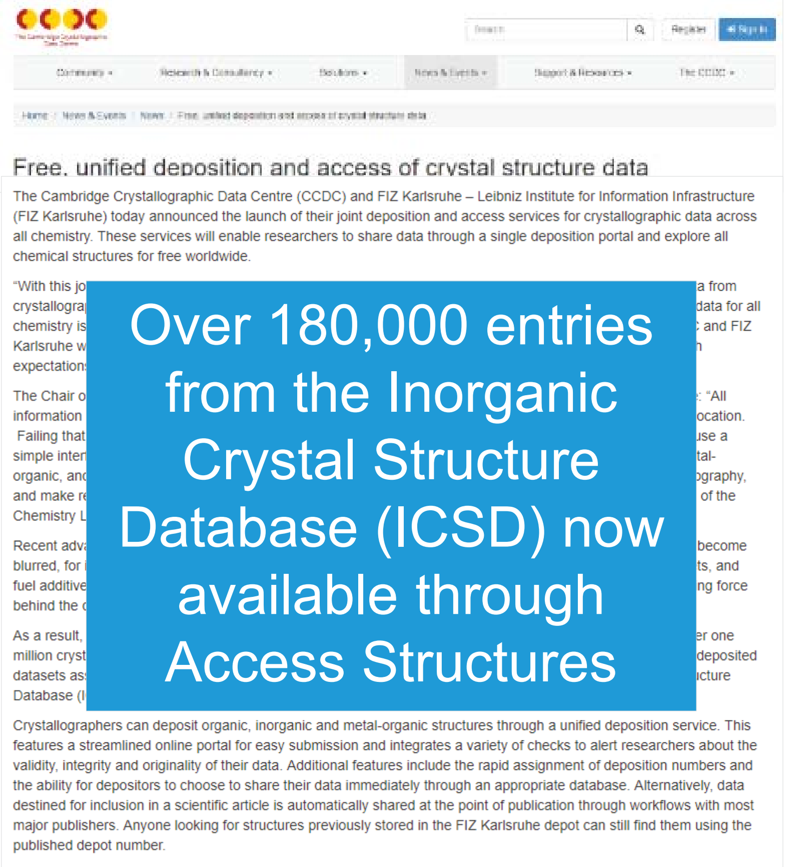
Adoption by the Community



What Else Could We Do?

- **Improved peer review**
  - Mandate crystallographic review of all structure-containing papers
  - Educate reviewers on nature of CheckCIF alerts
- **File requirements**
  - CIF + structure factors
  - Refinement instructions?
  - CheckCIF report?
- **Validation checks**
  - CheckCIF integration
  - Unit cell checks (with HKL checks? Or chemistry check?)
  - Geometry analysis?
- **Additional files available to reviewers?**

Joint CSD and ICSD Services



Free, unified deposition and access of crystal structure data

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.

“With this jo … a from
crystallogra … data for all
chemistry is … and FIZ
Karlsruhe w … h
expectation…

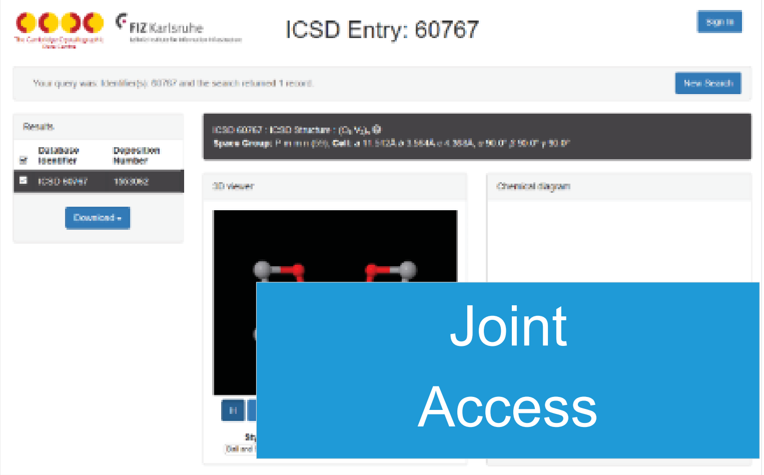
The Chair o … : “All
information … ocation.
Failing that … use a
simple inter … tal-
organic, and … ography,
and make r … of the
Chemistry L…

Recent adv … become
blurred, for i … ts, and
fuel additive … ng force
behind the …

As a result, … er one
million cryst … deposited
datasets as … ucture
Database (I…

Crystallographers can deposit organic, inorganic and metal-organic structures through a unified deposition service. This features a streamlined online portal for easy submission and integrates a variety of checks to alert researchers about the validity, integrity and originality of their data. Additional features include the rapid assignment of deposition numbers and the ability for depositors to choose to share their data immediately through an appropriate database. Alternatively, data destined for inclusion in a scientific article is automatically shared at the point of publication through workflows with most major publishers. Anyone looking for structures previously stored in the FIZ Karlsruhe depot can still find them using the published depot number.

Over 180,000 entries from the Inorganic Crystal Structure Database (ICSD) now available through Access Structures



ICSD Entry: 60767

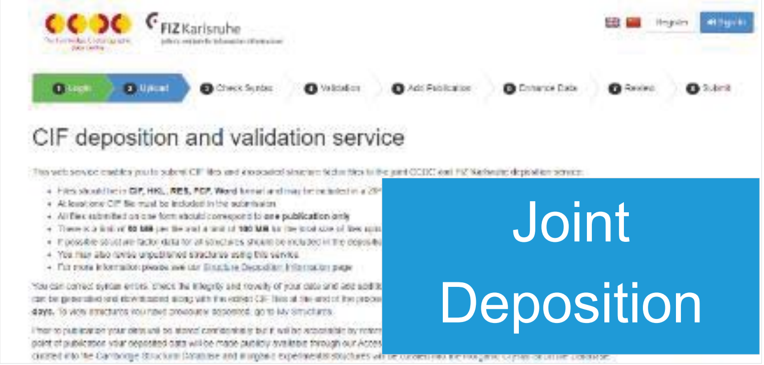
Your query was: Identifier(s): 60767 and the search returned 1 record.

☑	Database Identifier	Deposition Number
☑	ICSD 60767	1503062

ICSD 60767 : ICSD Structure : $(O_5V_2)_n$

Space Group: P m m n (59), **Cell:** a 11.512Å b 3.564Å c 4.368Å, α 90.0° β 90.0° γ 90.0°

Joint Access



CIF deposition and validation service

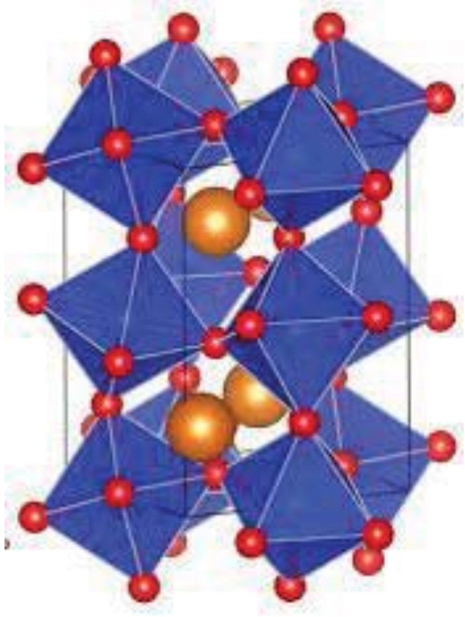
This web service enables you to submit CIF files and associated structure factor files to the joint CCDC and FIZ Karlsruhe deposition service.

- Files should be in **CIF, HKL, RES, FCF, Word** format and may be included in a ZIP
- At least one CIF file must be included in the submission
- All files submitted on one form should correspond to **one publication only**
- There is a limit of **60 MB** per file and a limit of **100 MB** for the total size of files
- If possible structure factor data for all structures should be included in the deposit
- You may also deposit unpublished structures using this service
- For more information please see our Structure Deposition Information page

You can correct syntax errors, check the integrity and novelty of your data and add addi… can be generated and downloaded along with the edited CIF files at the end of the proce… **days.** To view structures you have previously deposited, go to My Structures.

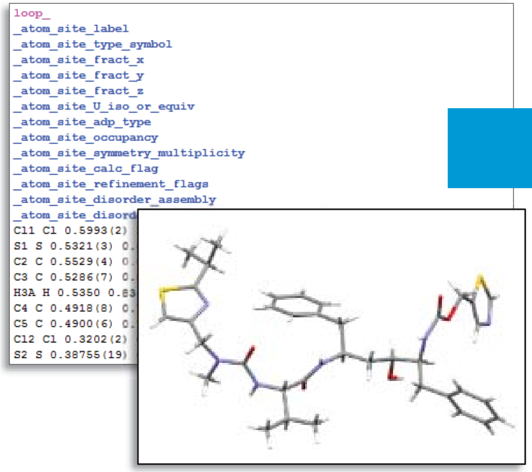
Prior to publication your data will be stored confidentially but it will be accessible by refere… point of publication your deposited data will be made publicly available through our Acces… curated into the Cambridge Structural Database and inorganic experimental structures will be curated into the Inorganic Crystal Structure Database.

Joint Deposition



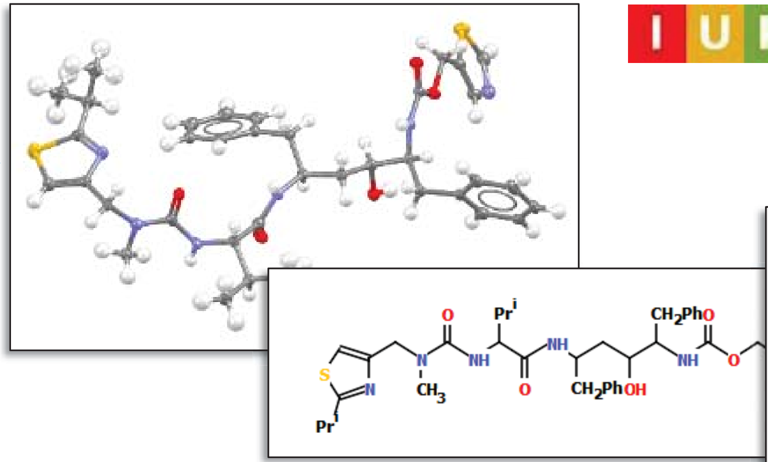
Curation and Chemistry Assignment

Deposited CIF

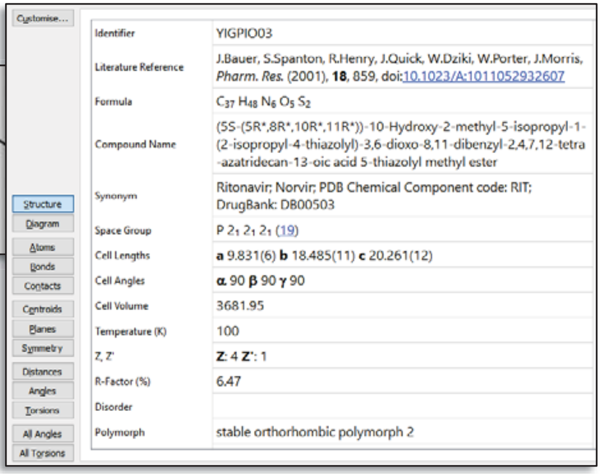


```
loop_
_atom_site_label
_atom_site_type_symbol
_atom_site_fract_x
_atom_site_fract_y
_atom_site_fract_z
_atom_site_U_iso_or_equiv
_atom_site_adp_type
_atom_site_occupancy
_atom_site_symmetry_multiplicity
_atom_site_calc_flag
_atom_site_refinement_flags
_atom_site_disorder_assembly
_atom_site_disord
Cl1 Cl 0.5993(2)
S1 S 0.5321(3) 0.
C2 C 0.5529(4) 0.
C3 C 0.5286(7) 0.
H3A H 0.5350 0.83
C4 C 0.4918(8) 0.
C5 C 0.4900(6) 0.
Cl2 Cl 0.3202(2)
S2 S 0.38755(19)
```

CSD Entry

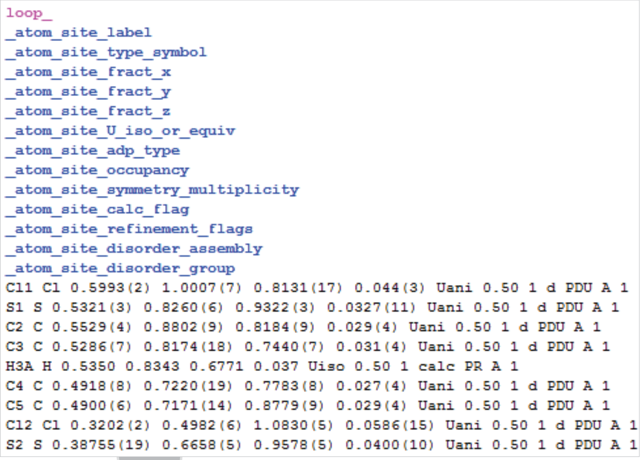


IUPAC



Identifier	YIGPIO03
Literature Reference	J.Bauer, S.Spanton, R.Henry, J.Quick, W.Dziki, W.Porter, J.Morris, *Pharm. Res.* (2001), **18**, 859, doi:10.1023/A:1011052932607
Formula	$C_{37} H_{48} N_6 O_5 S_2$
Compound Name	(5S-(5R*,8R*,10R*,11R*))-10-Hydroxy-2-methyl-5-isopropyl-1-(2-isopropyl-4-thiazolyl)-3,6-dioxo-8,11-dibenzyl-2,4,7,12-tetra-azatridecan-13-oic acid 5-thiazolyl methyl ester
Synonym	Ritonavir; Norvir; PDB Chemical Component code: RIT; DrugBank: DB00503
Space Group	P 2_1 2_1 2_1 (19)
Cell Lengths	**a** 9.831(6) **b** 18.485(11) **c** 20.261(12)
Cell Angles	**α** 90 **β** 90 **γ** 90
Cell Volume	3681.95
Temperature (K)	100
Z, Z'	**Z**: 4 **Z'**: 1
R-Factor (%)	6.47
Disorder	
Polymorph	stable orthorhombic polymorph 2

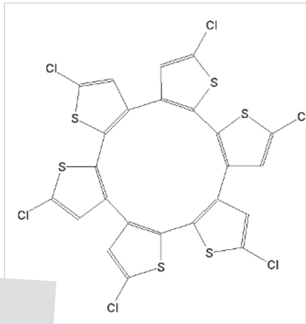
Using the CSD to Help With Curation

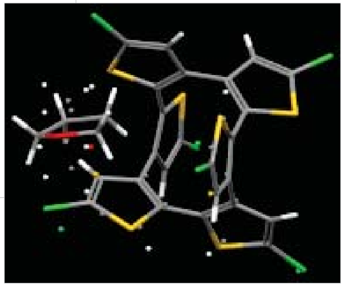


```
loop_
_atom_site_label
_atom_site_type_symbol
_atom_site_fract_x
_atom_site_fract_y
_atom_site_fract_z
_atom_site_U_iso_or_equiv
_atom_site_adp_type
_atom_site_occupancy
_atom_site_symmetry_multiplicity
_atom_site_calc_flag
_atom_site_refinement_flags
_atom_site_disorder_assembly
_atom_site_disorder_group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

An automated probabilistic approach using data in the CSD

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

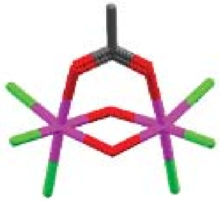




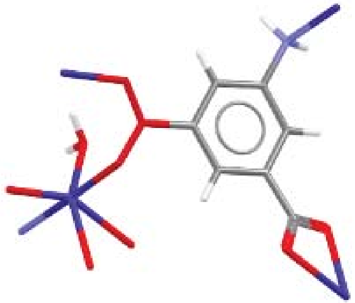
$$P(A|B) = \frac{P(B|A)\,P(A)}{P(B)}.$$

I.J.Bruno, G.P.Shields, R.Taylor, Acta Cryst. (2011). B**67**, 333-349
DOI: 10.1107/S0108768111024608

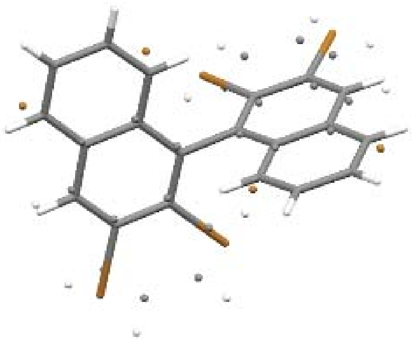
Challenges



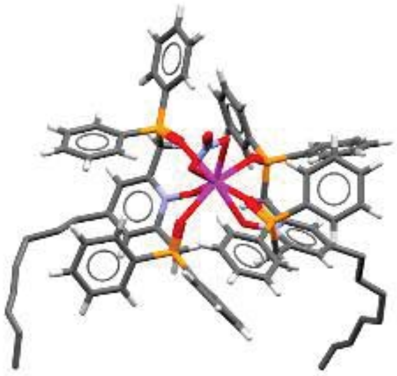
Missing atoms



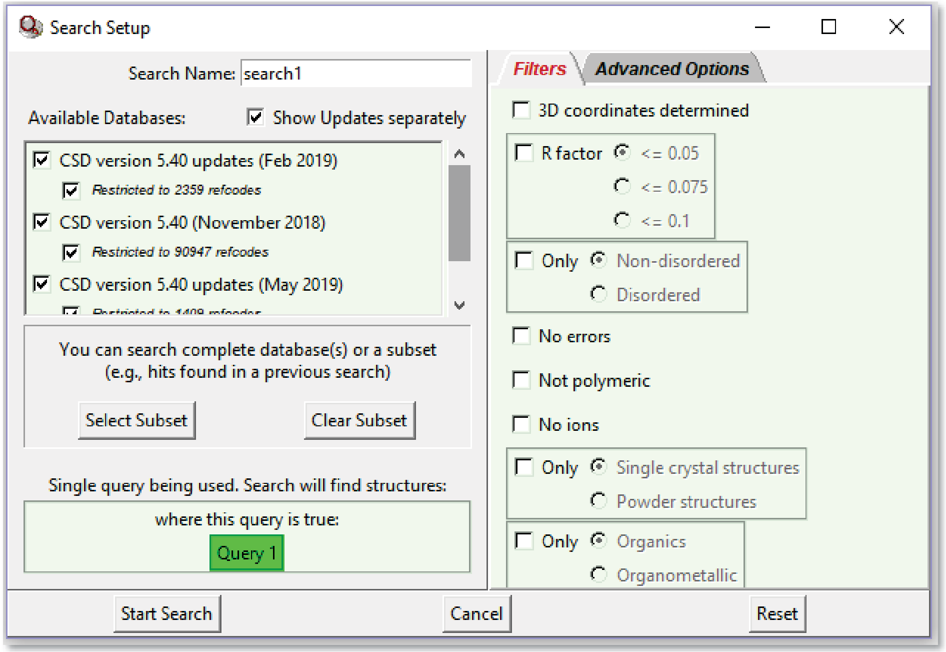
Element assignment



Disorder



Poor geometries



Search Setup

Search Name: search1

Available Databases: ☑ Show Updates separately

☑ CSD version 5.40 updates (Feb 2019)

☑ *Restricted to 2359 refcodes*

☑ CSD version 5.40 (November 2018)

☑ *Restricted to 90947 refcodes*

☑ CSD version 5.40 updates (May 2019)

☑ *Restricted to 1409 refcodes*

You can search complete database(s) or a subset (e.g., hits found in a previous search)

Select Subset | Clear Subset

Single query being used. Search will find structures:

where this query is true:

Query 1

Filters | *Advanced Options*

□ 3D coordinates determined

□ R factor ● <= 0.05 ○ <= 0.075 ○ <= 0.1

□ Only ● Non-disordered ○ Disordered

□ No errors

□ Not polymeric

□ No ions

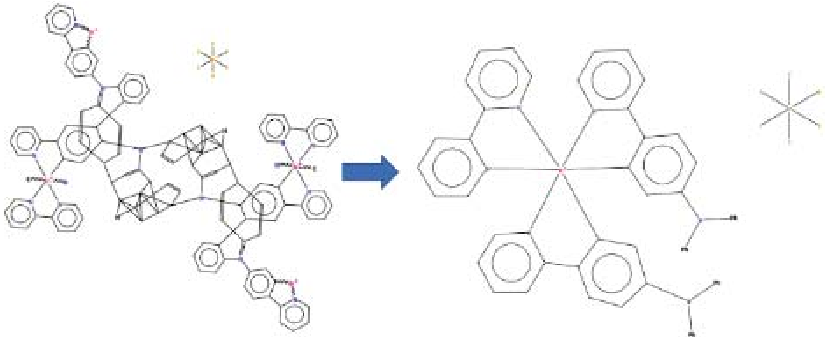
□ Only ● Single crystal structures ○ Powder structures

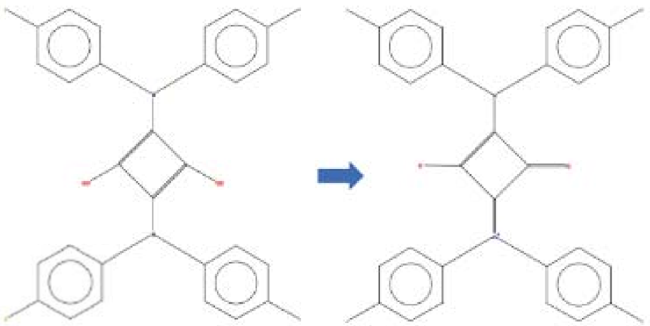
□ Only ● Organics ○ Organometallic

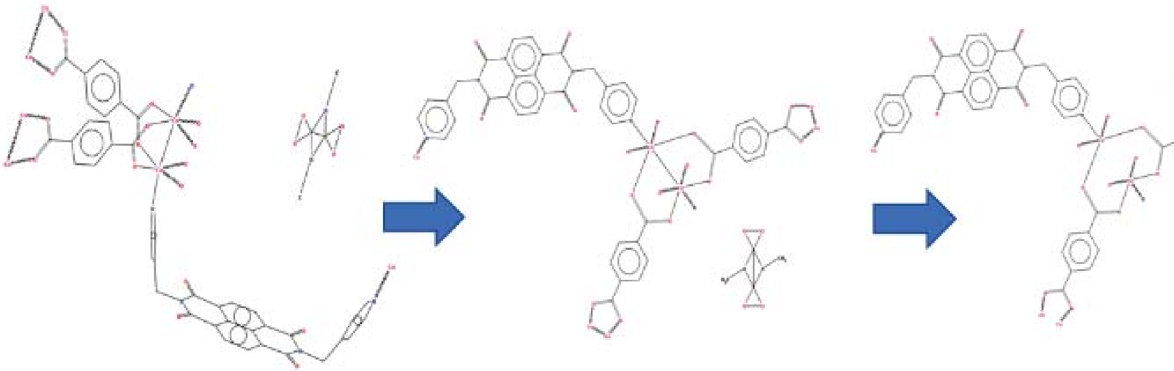
Start Search | Cancel | Reset

The Human Touch

- Each entry looked at by expert Scientific Editors
- Reliability scores focusses editorial efforts
- Manual validation of automated chemical interpretations improves automated methods





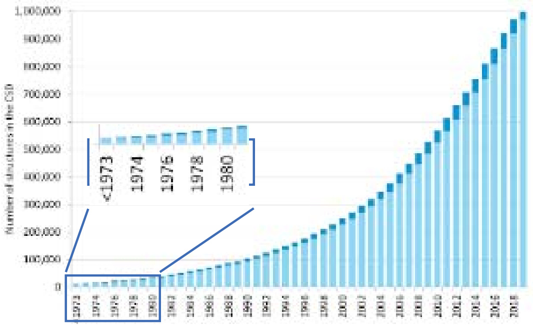


https://www.ccdc.cam.ac.uk/Community/blog/CSD-data-curation-the-human-touch/

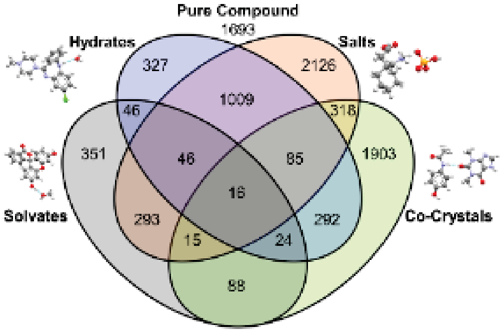
Revisiting Data

Targeted improvements allow improved integrity, consistency, discoverability and value of data

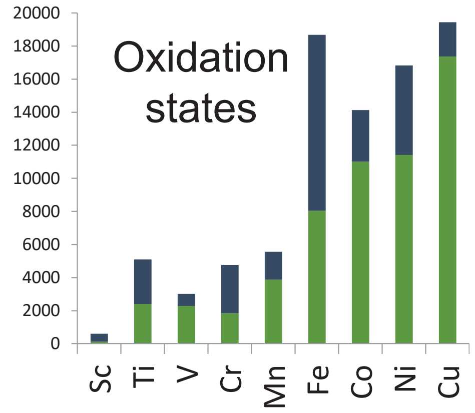
Ensure standardisation of early CSD entries



Creation and maintenance of subsets



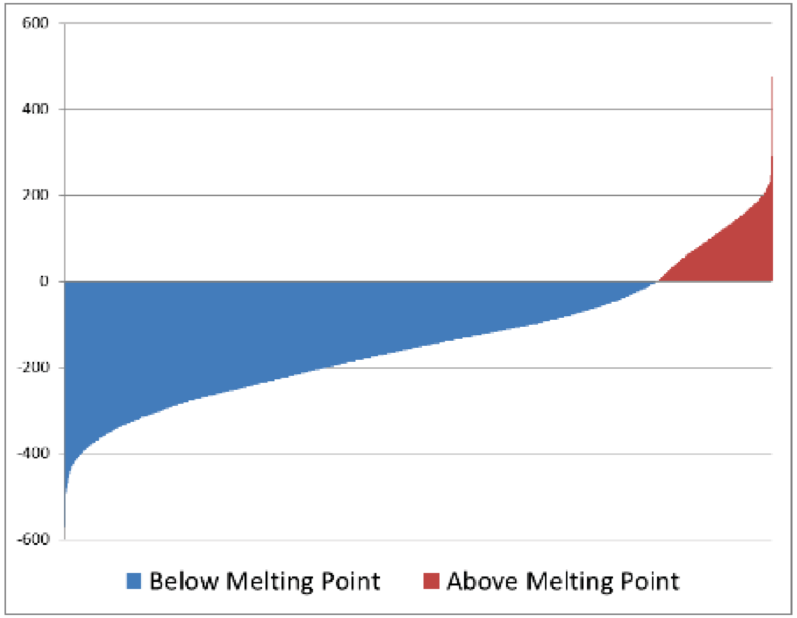
Enrichment of data



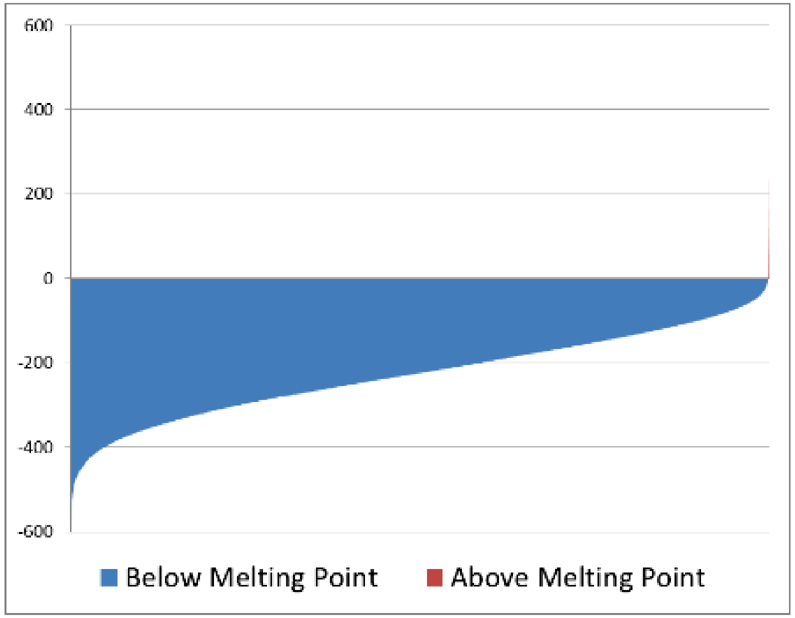
Melting Points in the CSD

>170,000 Melting Points

Study Temperature relative to MP
before additional CCDC validation



Study Temperature relative to MP
after additional CCDC validation



Maintaining Data Integrity

- **Integrity** – Completeness, consistency and trustworthiness

- **Data completeness** – Trends in reporting of metadata
  - Identify CSD Deposit checks and enhancements
  - Identify new filters to allow CSD users to better select fit for purpose data

- **Consistency** – Looking at experimental metadata to identify trends in information supplied

- **Trustworthiness** – Establishing automatic identification of potential cases of misconduct – including fraudulent and plagiarised data

Research integrity is much more than misconduct. *Nature,* 2019, 570, 5-5. DOI:10.1038/d41586-019-01727-0

Following Standard Ethical Practises

- CCDC is now a Member of the **C**ommittee **o**n **P**ublication **E**thics.
- COPE’s objective is "*to educate and advance knowledge in methods of safeguarding the integrity of the scholarly record for the benefit of the public*".
- Membership gives us access to COPE resources and COPE advice – helping us deal with publication ethics and data integrity and issues.

COPE

Member since 2019

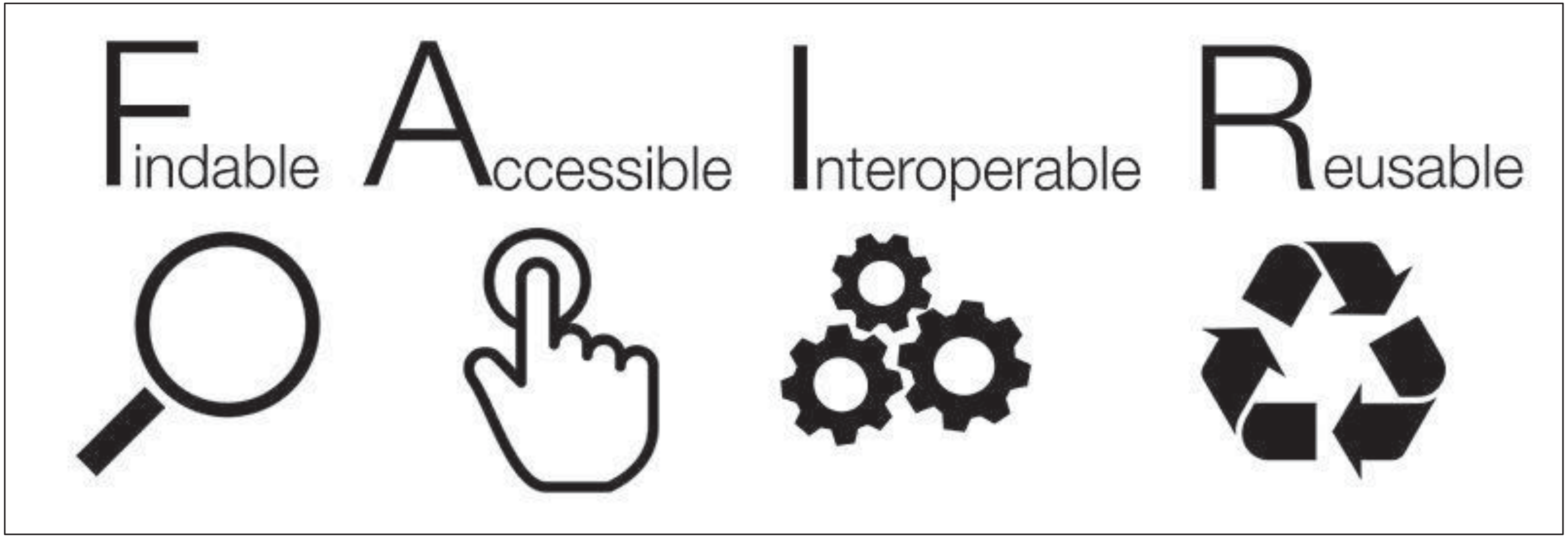
AC00039

https://publicationethics.org/about/governance

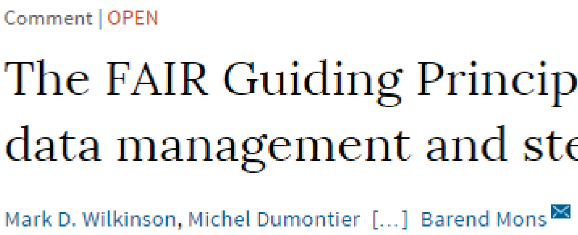
Making Crystallographic Data FAIR

CCDC database workflows

FAIR Data Principles



FAIR Data Principles



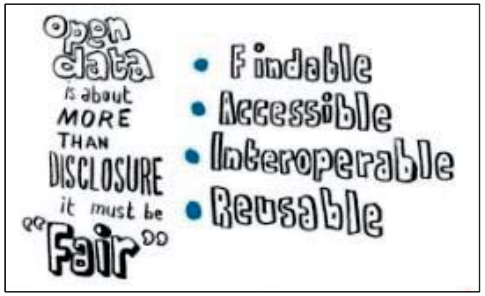
Comment | OPEN

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).

"all research objects should be Findable, Accessible, Interoperable and Reusable (FAIR) both for machines and for people" (Wilkinson, M. D. et al., 2016 : 3)



FORCE11
The Future of Research Communications and e-Scholarship

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

Data FAIRport

http://www.datafairport.org/

FAIR Data Policies and Guidelines

Plan S

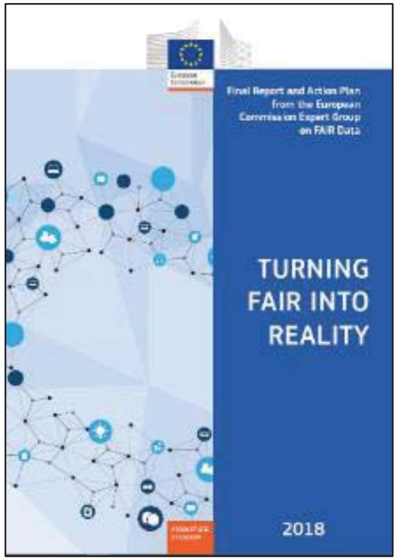
“Although the Plan S principles refer to peer-reviewed scholarly publications, cOAlition S also strongly encourages that research data and other research outputs are made as open as possible and as closed as necessary.”

https://www.coalition-s.org/principles-and-implementation/

European Commission

“..the implementation of FAIR data needs to go hand-in-hand with the principle that data created by publicly-funded research must be as Open as possible and as closed as necessary. The EC and Member States should consider FAIR and Open as complementary concepts and address both in policy. “ *(EU Commission, 2018 : 10)*

EU Commission Expert Group on FAIR Data (2018) Turning FAIR into reality Final Report and Action Plan from the European Commission Expert Group on FAIR Data, EU Commission.
DOI: 10.2777/1524



Final Report and Action Plan from the European Commission Expert Group on FAIR Data

TURNING FAIR INTO REALITY

2018

Funder Research Data Sharing Policies

“Since 2017, all Horizon 2020 projects are part of the Open Research Data Pilot by default. The Principal Investigator must:

- Develop a data management plan in the first 6 months of the project and keep it up-to-date throughout their project;
- Deposit their research data in a suitable research data repository;
- Make sure third parties can freely access, mine, exploit, reproduce and disseminate their data;
- Make clear what tools will be needed to use the raw data to validate research results, or provide the tools themselves.”

erc

European Research Council

Established by the European Commission

“ERC beneficiaries are encouraged to take part in the H2020 Open Research Data Pilot, but this is not compulsory.
Those who take part in the Open Research Data Pilot must adhere to the obligations outlined above.”

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

Publisher research data policies

“The Royal Society of Chemistry believes that, where possible, all data associated with the research in a manuscript should be **freely available** in an **accessible** and **usable** format, enabling other researchers to replicate and build on that research. Therefore, in addition to providing the data required for submission (as detailed above) we encourage authors to deposit as much data as possible that is related to the research in their article. This should be in appropriate and **publicly available repositories**”

https://www.rsc.org/journals-books-databases/journal-authors-reviewers/prepare-your-article/experimental-data/

“It is the practice of IUCr journals to provide **free access** to all supplementary materials and supporting data files deposited with a published article.”

https://journals.iucr.org/services/authorrights.html

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

Data Repositories

Role of Repositories in making data FAIR:

- □ Long-term data preservation
- □ Continued access to data
- □ Assignment of identifiers and DOIs
- □ Added descriptive metadata
- □ Searchable databases
- □ Data curation

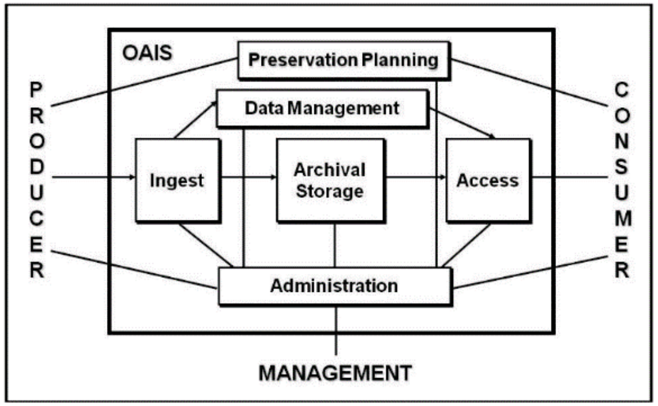
Repository frameworks and systems

DOI assignment

DataCite FIND, ACCESS, AND REUSE DATA

Crossref

Open Archival Information System (OAIS) Reference Model



Global and domain specific membership groups

RDA RESEARCH DATA ALLIANCE

IUPAC

Data preservation policies & plans

Trusted Repository Certification

CORE TRUST SEAL

ICSU WORLD DATA SYSTEM

Digital Preservation Coalition and Brian Lavoie (2014) The Open Archival Information System (OAIS) Reference Model: Introductory Guide (2nd Edition), Digital Preservation Coalition. DOI: 10.7207/twr14-02

Data Repositories

How to find a repository for your research data

□ Search a repository registry

re3data.org
REGISTRY OF RESEARCH DATA REPOSITORIES

https://www.re3data.org/

□ Use a repository recommended by the publisher or funding body

□ Search for repositories among accreditation bodies

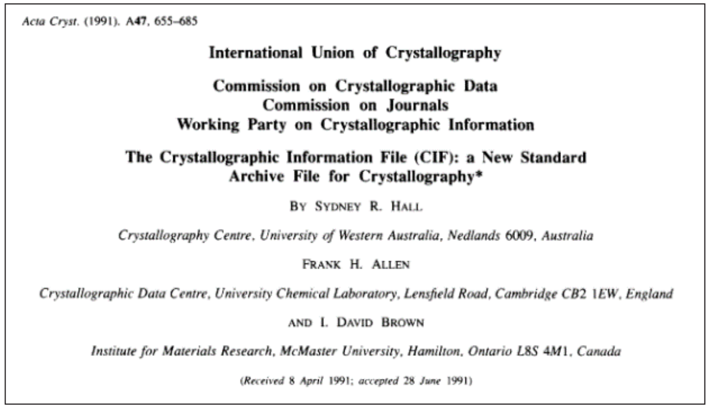
CORE TRUST SEAL

https://www.coretrustseal.org/

ICSU WORLD DATA SYSTEM

https://www.icsu-wds.org/services/certification

Crystallographic Information File: CIF



Acta Cryst. (1991). A47, 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

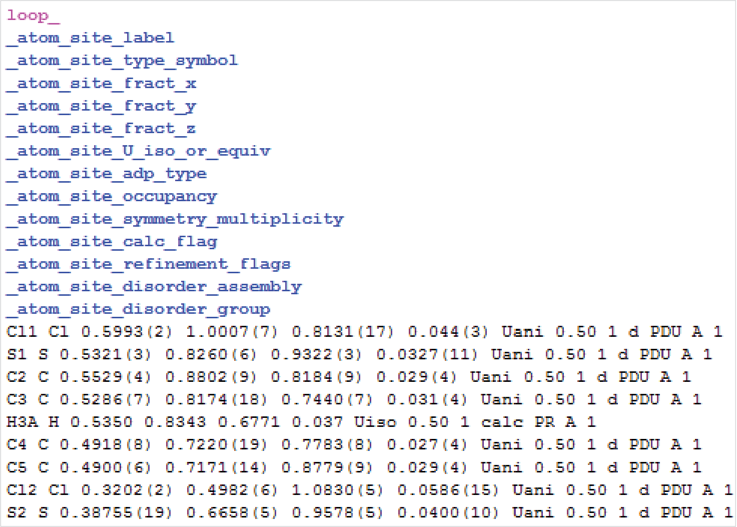
(*Received 8 April 1991; accepted 28 June 1991*)

A standard format for archive and exchange of crystallographic data

- o derived model
- o processed data (structure factors)
- o metadata about raw data (imgCIF)

A standard format for archive and exchange of crystallographic data

- o derived model
- o processed data (structure factors)
- o metadata about raw data (imgCIF)



```
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
```

CIF as a FAIR data format

Findable

- Searchable fields for identifiers and metadata descriptions

Interoperable

- Standard dictionary and vocabularies
- Standard format for processed and derived data

Accessible

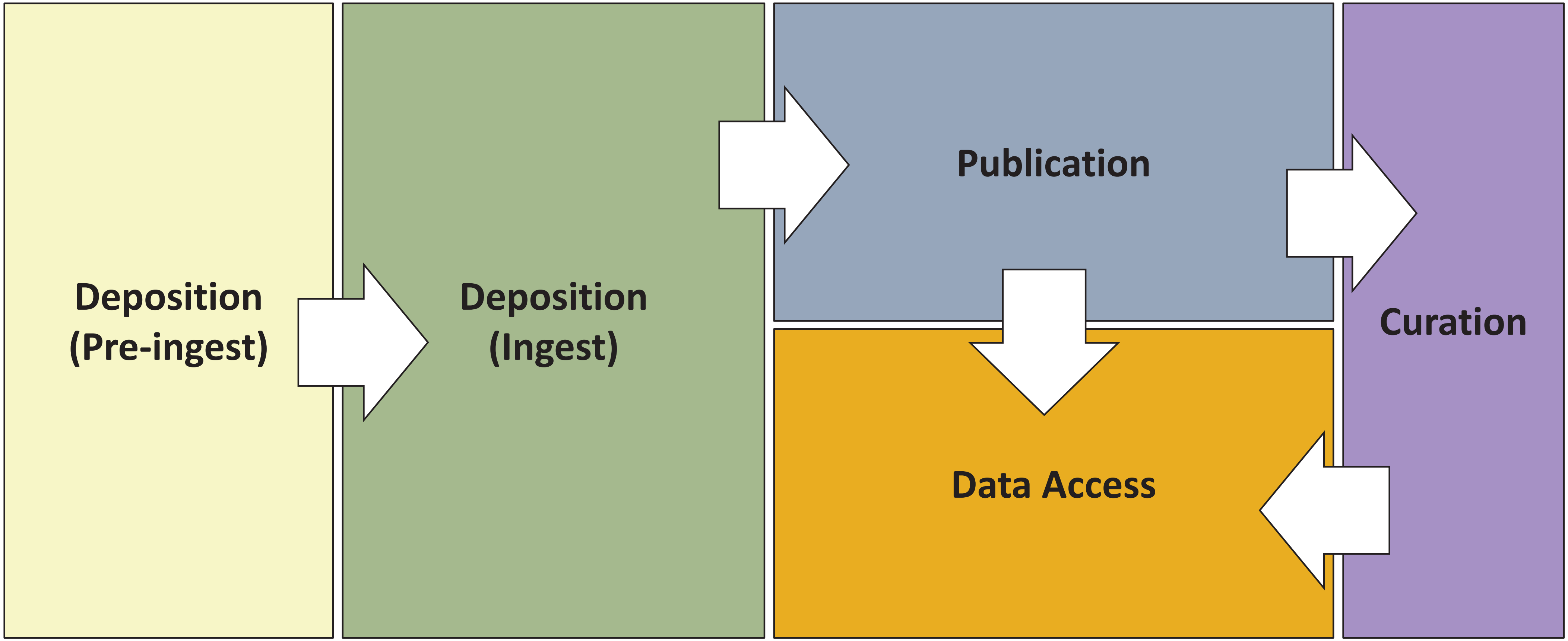
- Trusted searchable data repositories:

- Cambridge Structural Database
- Inorganic Crystal Structure Database
- Protein Data Bank

Reusable

- Data provenance
- Software packages and parameters
- Quality metrics

CCDC Dataset Workflow



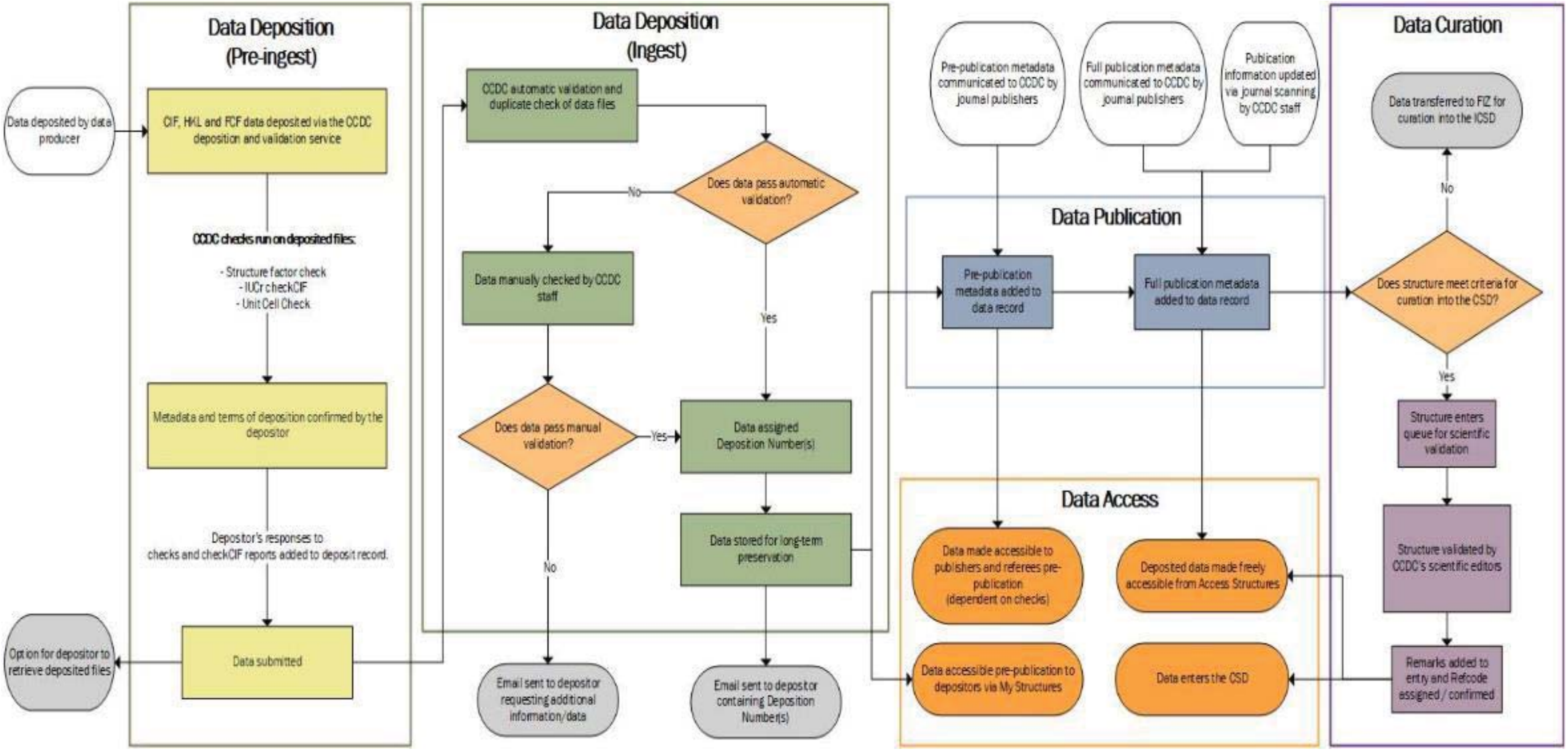
Deposition (Pre-ingest)

Deposition (Ingest)

Publication

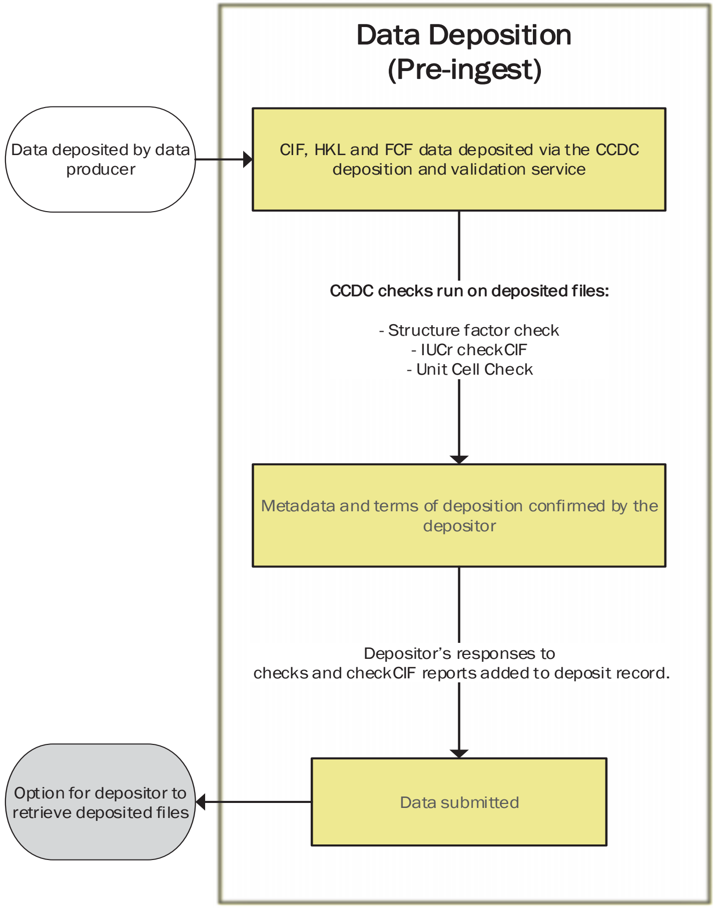
Data Access

Curation


Data Deposition (Pre-ingest)
Data deposited by data producer
CIF, HKL and FCF data deposited via the CCDC deposition and validation service
CCDC checks run on deposited files:
- Structure factor check
- IUCr checkCIF
- Unit Cell Check
Metadata and terms of deposition confirmed by the depositor
Depositor's responses to checks and checkCIF reports added to deposit record.
Data submitted
Option for depositor to retrieve deposited files
Data Deposition (Ingest)
CCDC automatic validation and duplicate check of data files
Does data pass automatic validation?
No
Yes
Data manually checked by CCDC staff
Does data pass manual validation?
Yes
No
Data assigned Deposition Number(s)
Data stored for long-term preservation
Email sent to depositor requesting additional information/data
Email sent to depositor containing Deposition Number(s)
Pre-publication metadata communicated to CCDC by journal publishers
Full publication metadata communicated to CCDC by journal publishers
Publication information updated via journal scanning by CCDC staff
Data Publication
Pre-publication metadata added to data record
Full publication metadata added to data record
Data Access
Data made accessible to publishers and referees pre-publication (dependent on checks)
Deposited data made freely accessible from Access Structures
Data accessible pre-publication to depositors via My Structures
Data enters the CSD
Data Curation
Data transferred to FIZ for curation into the ICSD
No
Does structure meet criteria for curation into the CSD?
Yes
Structure enters queue for scientific validation
Structure validated by CCDC's scientific editors
Remarks added to entry and Refcode assigned / confirmed

Accessible for download at: https://www.ccdc.cam.ac.uk/Community/depositastructure/scientific-data-preservation/

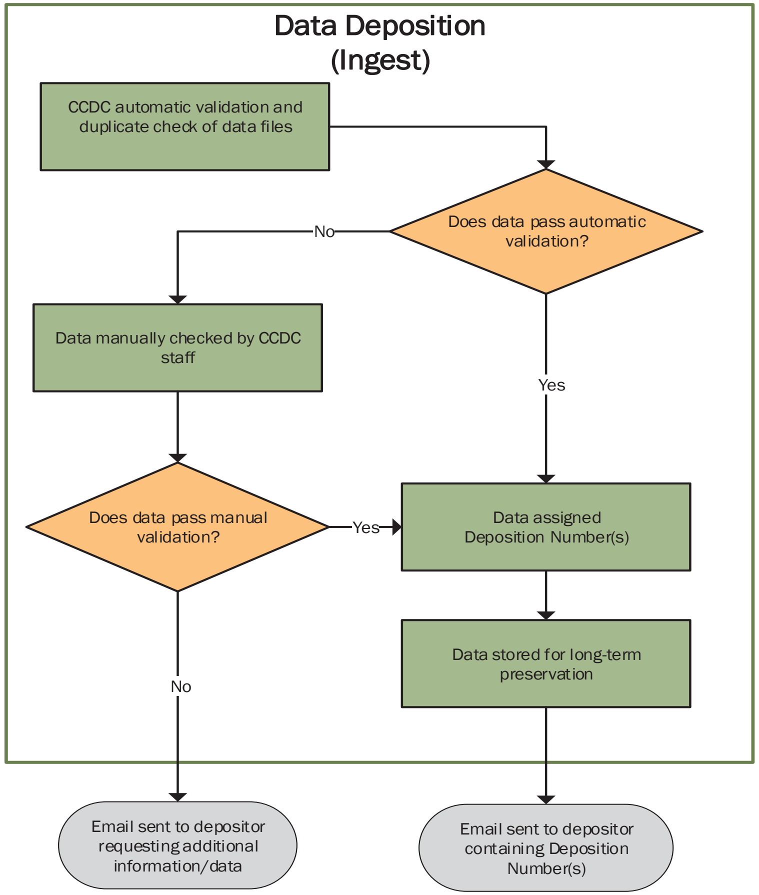
Data Deposition (Pre-ingest)


Data Deposition
(Pre-ingest)
Data deposited by data producer
CIF, HKL and FCF data deposited via the CCDC deposition and validation service
CCDC checks run on deposited files:
- Structure factor check
- IUCr checkCIF
- Unit Cell Check
Metadata and terms of deposition confirmed by the depositor
Depositor's responses to checks and checkCIF reports added to deposit record.
Data submitted
Option for depositor to retrieve deposited files

- Manual User Actions:
  - ❑ Provide personal details and upload files
  - ❑ Fix syntax errors (if any)
  - ❑ Explain why no structure factor data
  - ❑ Add explanations for checkCIF alerts
  - ❑ Add crystallographer details
  - ❑ Provide known publication details
  - ❑ Add additional scientific metadata
  - ❑ Review and confirm

- Automated Actions:
  - ❑ Syntax check
  - ❑ Generation of checkCIF report

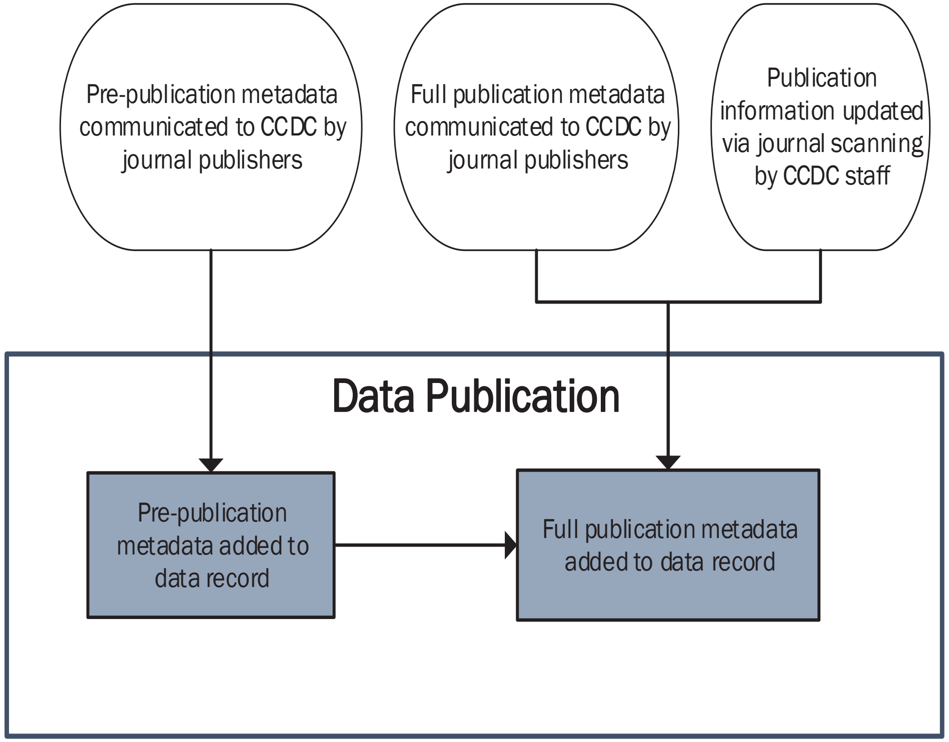
Data Deposition (Ingest)


Data Deposition
(Ingest)
CCDC automatic validation and duplicate check of data files
Does data pass automatic validation?
No
Yes
Data manually checked by CCDC staff
Does data pass manual validation?
Yes
No
Data assigned Deposition Number(s)
Data stored for long-term preservation
Email sent to depositor requesting additional information/data
Email sent to depositor containing Deposition Number(s)

- Automated Actions:
  - ❑ Syntax check
  - ❑ Check for duplicate dataset
  - ❑ Internal record creation
  - ❑ Assigning identifiers

- **Manual CCDC Actions:**
  - ❑ Dealing with non-standard file formats
  - ❑ Investigating duplicate datasets
  - ❑ Updating records with resubmitted data

Data Publication



Pre-publication metadata communicated to CCDC by journal publishers

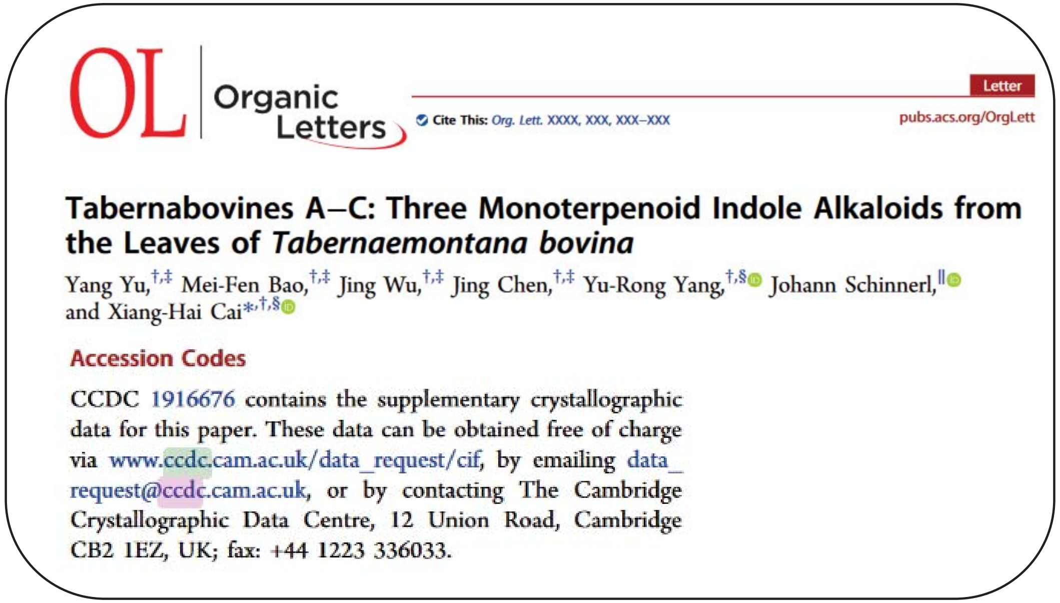
Full publication metadata communicated to CCDC by journal publishers

Publication information updated via journal scanning by CCDC staff

Data Publication

Pre-publication metadata added to data record

Full publication metadata added to data record



OL Organic Letters

Letter

Cite This: *Org. Lett.* XXXX, XXX, XXX−XXX

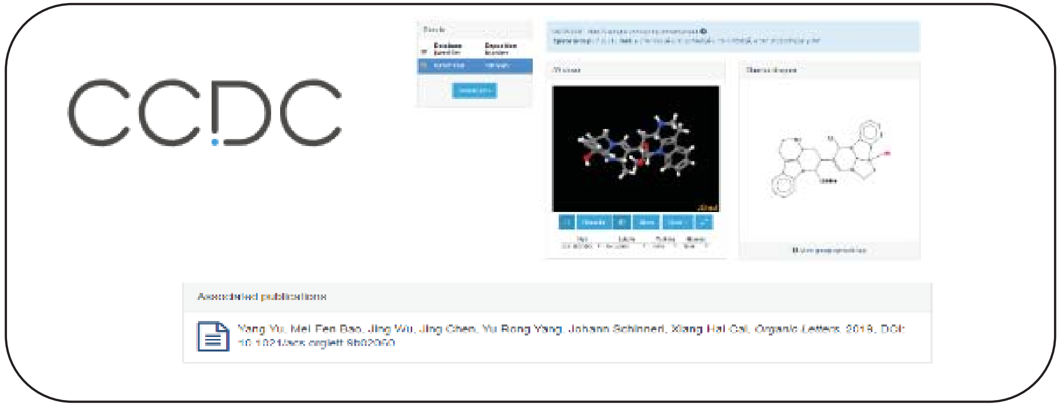
pubs.acs.org/OrgLett

Tabernabovines A−C: Three Monoterpenoid Indole Alkaloids from the Leaves of *Tabernaemontana bovina*

Yang Yu,[†,‡] Mei-Fen Bao,[†,‡] Jing Wu,[†,‡] Jing Chen,[†,‡] Yu-Rong Yang,[†,§] Johann Schinnerl,[∥] and Xiang-Hai Cai*[,†,§]

Accession Codes

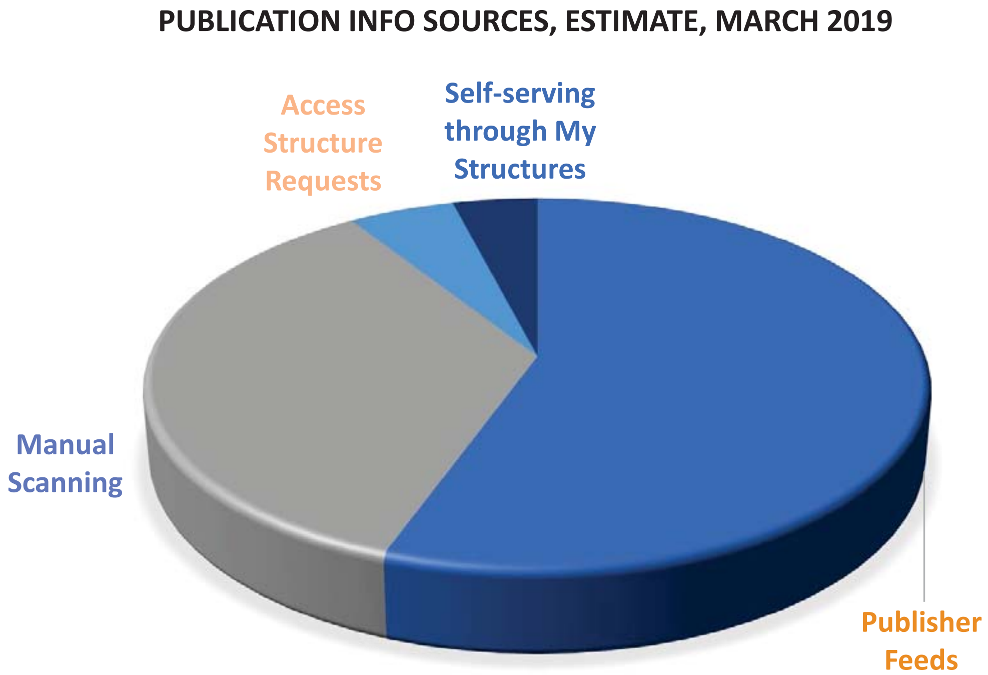
CCDC 1916676 contains the supplementary crystallographic data for this paper. These data can be obtained free of charge via www.ccdc.cam.ac.uk/data_request/cif, by emailing data_request@ccdc.cam.ac.uk, or by contacting The Cambridge Crystallographic Data Centre, 12 Union Road, Cambridge CB2 1EZ, UK; fax: +44 1223 336033.



Associated publications

Yang Yu, Mei Fen Bao, Jing Wu, Jing Chen, Yu Rong Yang, Johann Schinnerl, Xiang Hai Cai, Organic Letters, 2019, DOI: 10.1021/acs.orglett.9b02060

Data Publication



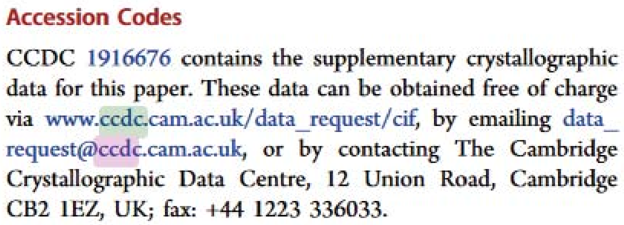
Sources of Publication Information:

- ❑ Pre-publication metadata communicated by journal publisher feeds
- ❑ Full publication metadata communicated and updated by journal publisher feeds

Manual CCDC Actions:

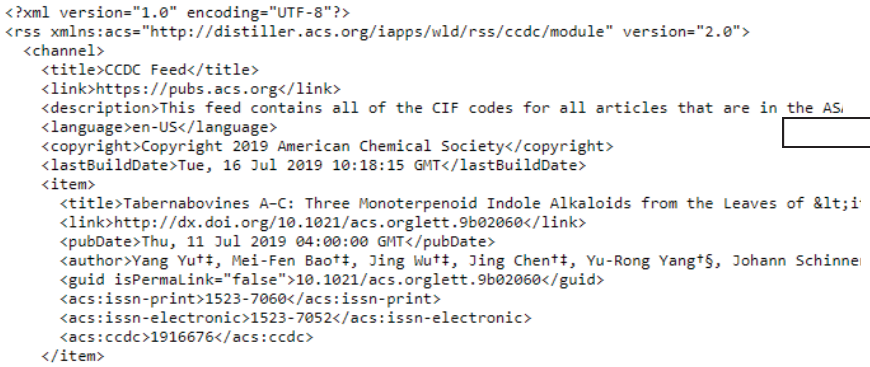
- ❑ Reviewing publication details
- ❑ Publication information updated via journal scanning by CCDC staff
- ❑ Publication information communicated by researchers wanting to access data

Data Publication

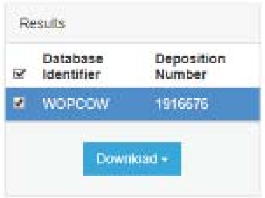


Accession Codes

CCDC 1916676 contains the supplementary crystallographic data for this paper. These data can be obtained free of charge via www.ccdc.cam.ac.uk/data_request/cif, by emailing data_request@ccdc.cam.ac.uk, or by contacting The Cambridge Crystallographic Data Centre, 12 Union Road, Cambridge CB2 1EZ, UK; fax: +44 1223 336033.

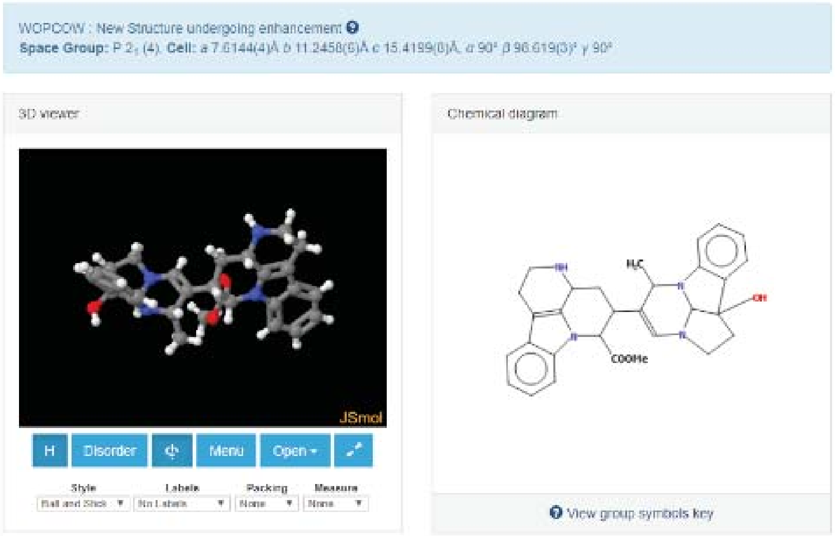


```
<?xml version="1.0" encoding="UTF-8"?>
<rss xmlns:acs="http://distiller.acs.org/iapps/wld/rss/ccdc/module" version="2.0">
  <channel>
    <title>CCDC Feed</title>
    <link>https://pubs.acs.org</link>
    <description>This feed contains all of the CIF codes for all articles that are in the AS
    <language>en-US</language>
    <copyright>Copyright 2019 American Chemical Society</copyright>
    <lastBuildDate>Tue, 16 Jul 2019 10:18:15 GMT</lastBuildDate>
    <item>
      <title>Tabernabovines A-C: Three Monoterpenoid Indole Alkaloids from the Leaves of &lt;i
      <link>http://dx.doi.org/10.1021/acs.orglett.9b02060</link>
      <pubDate>Thu, 11 Jul 2019 04:00:00 GMT</pubDate>
      <author>Yang Yu†‡, Mei-Fen Bao†‡, Jing Wu†‡, Jing Chen†‡, Yu-Rong Yang†§, Johann Schinne
      <guid isPermaLink="false">10.1021/acs.orglett.9b02060</guid>
      <acs:issn-print>1523-7060</acs:issn-print>
      <acs:issn-electronic>1523-7052</acs:issn-electronic>
      <acs:ccdc>1916676</acs:ccdc>
    </item>
```



Results

☑	Database Identifier	Deposition Number
☑	WOPCOW	1916676

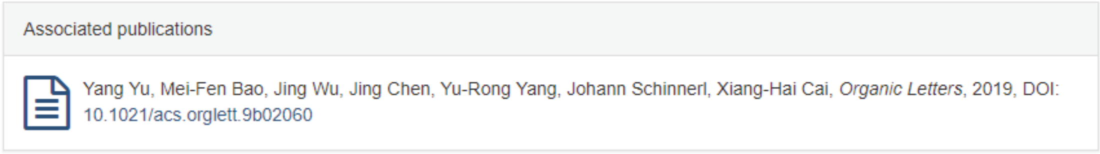


WOPCOW : New Structure undergoing enhancement

Space Group: P 2_1 (4), **Cell:** *a* 7.6144(4)Å *b* 11.2458(6)Å *c* 15.4199(8)Å, *α* 90° *β* 96.619(3)° *γ* 90°

3D viewer

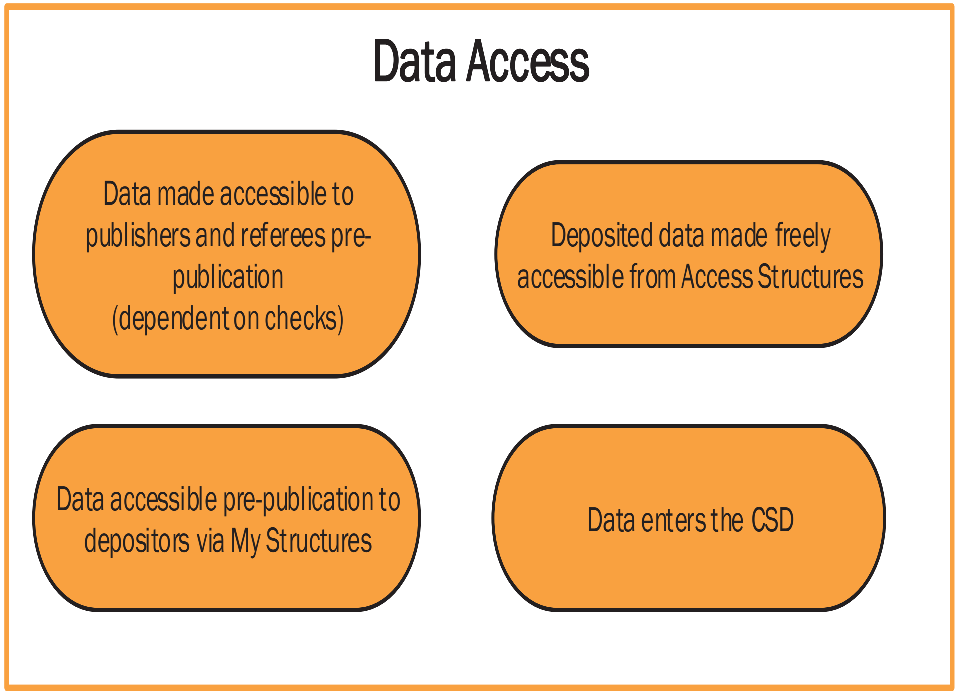
Chemical diagram



Associated publications

Yang Yu, Mei-Fen Bao, Jing Wu, Jing Chen, Yu-Rong Yang, Johann Schinnerl, Xiang-Hai Cai, *Organic Letters*, 2019, DOI: 10.1021/acs.orglett.9b02060

Data Access


Data Access

Data made accessible to publishers and referees pre-publication (dependent on checks)

Deposited data made freely accessible from Access Structures

Data accessible pre-publication to depositors via My Structures

Data enters the CSD

- Data is made accessible:
  - ❑ Pre-publication to reviewers and depositors to facilitate preparation of manuscripts
  - ❑ Immediately through Access Structures once data is published
  - ❑ Through the CSD once curated by CCDC's scientific editors.

- Processes for making data more findable and interoperable:
  - ❑ Identifiers added to data entries
  - ❑ Links from publication articles to data created
  - ❑ Links from CCDC datasets to other databases

Standard Identifiers and Interoperability

DC1 Data Citation Principles

Data should be considered legitimate, citable products of research…

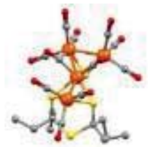
https://www.force11.org/datacitation

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

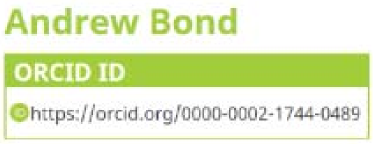
- The CCDC registers DOIs for datasets through DataCite
- Metadata for CCDC datasets is openly accessible via DataCite
- Foundation for interoperability and formalising data citation



10.5517/CCPHZ37

ORCID IDs for Researchers

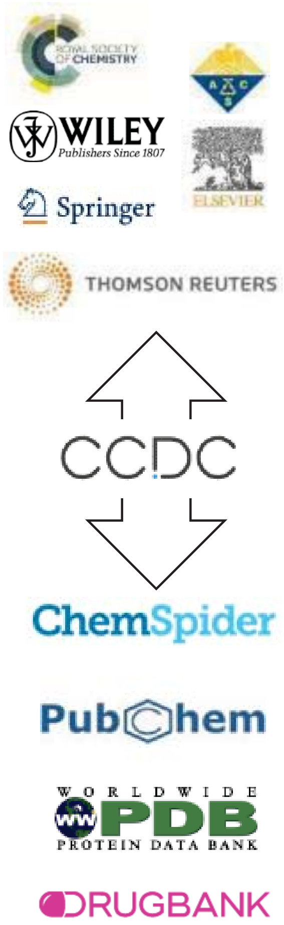
At least 30% of current CSD depositors provide an ORCID ID



Andrew Bond

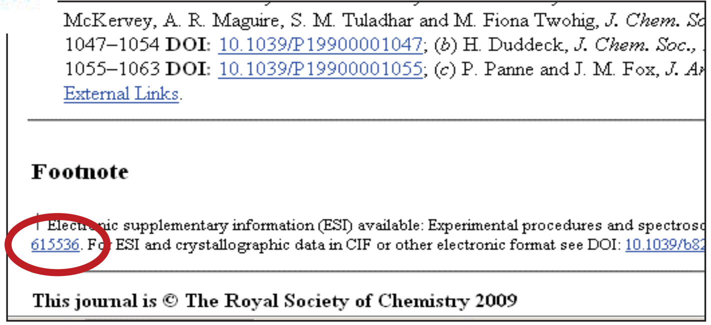
ORCID ID

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


ROYAL SOCIETY OF CHEMISTRY
WILEY Publishers Since 1807
Springer
ELSEVIER
THOMSON REUTERS
CCDC
ChemSpider
PubChem
WORLDWIDE PDB PROTEIN DATA BANK
DRUGBANK

Links from Articles to CCDC Data

ROYAL SOCIETY OF CHEMISTRY



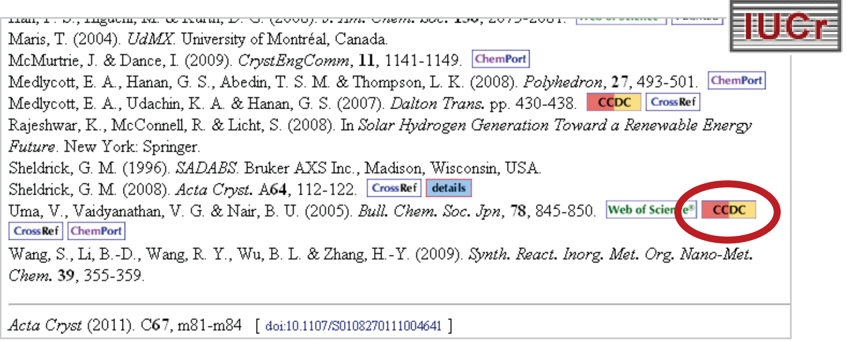
McKervey, A. R. Maguire, S. M. Tuladhar and M. Fiona Twohig, *J. Chem. Sc* 1047–1054 **DOI:** 10.1039/P19900001047; (*b*) H. Duddeck, *J. Chem. Soc.,* 1055–1063 **DOI:** 10.1039/P19900001055; (*c*) P. Panne and J. M. Fox, *J. Ar* External Links.

Footnote

† Electronic supplementary information (ESI) available: Experimental procedures and spectrosc 615536. For ESI and crystallographic data in CIF or other electronic format see DOI: 10.1039/b82

This journal is © The Royal Society of Chemistry 2009

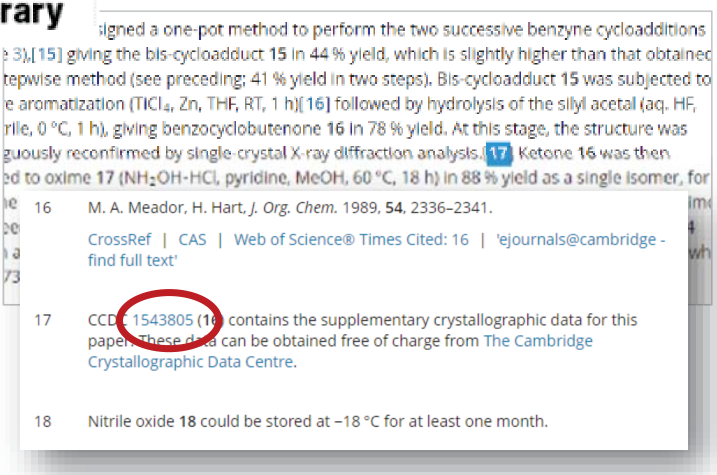
IUCr



Maris, T. (2004). *UdMX.* University of Montréal, Canada.
McMurtrie, J. & Dance, I. (2009). *CrystEngComm*, **11**, 1141-1149. ChemPort
Medlycott, E. A., Hanan, G. S., Abedin, T. S. M. & Thompson, L. K. (2008). *Polyhedron*, **27**, 493-501. ChemPort
Medlycott, E. A., Udachin, K. A. & Hanan, G. S. (2007). *Dalton Trans.* pp. 430-438. CCDC CrossRef
Rajeshwar, K., McConnell, R. & Licht, S. (2008). In *Solar Hydrogen Generation Toward a Renewable Energy Future*. New York: Springer.
Sheldrick, G. M. (1996). *SADABS.* Bruker AXS Inc., Madison, Wisconsin, USA.
Sheldrick, G. M. (2008). *Acta Cryst.* A**64**, 112-122. CrossRef details
Uma, V., Vaidyanathan, V. G. & Nair, B. U. (2005). *Bull. Chem. Soc. Jpn*, **78**, 845-850. Web of Science® CCDC CrossRef ChemPort
Wang, S., Li, B.-D., Wang, R. Y., Wu, B. L. & Zhang, H.-Y. (2009). *Synth. React. Inorg. Met. Org. Nano-Met. Chem.* **39**, 355-359.

Acta Cryst (2011). C**67**, m81-m84 [doi:10.1107/S0108270111004641]

Wiley Online Library



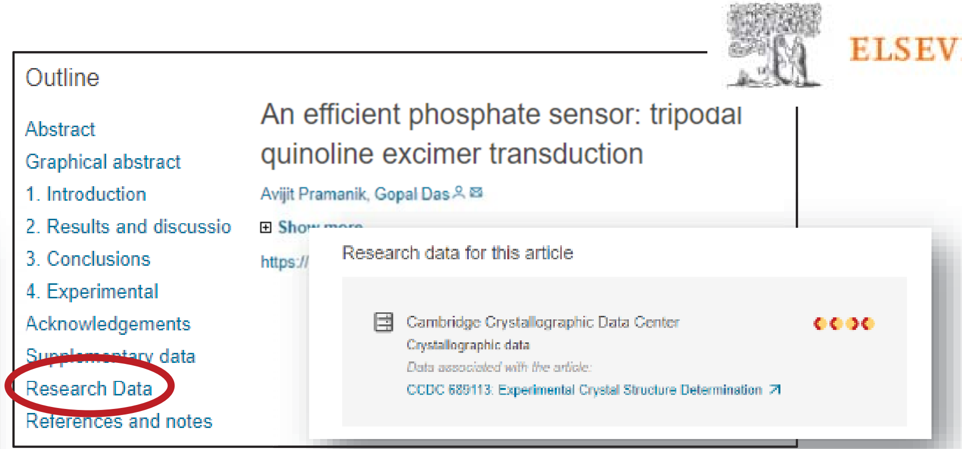
signed a one-pot method to perform the two successive benzyne cycloadditions 3),[15] giving the bis-cycloadduct **15** in 44 % yield, which is slightly higher than that obtained tepwise method (see preceding; 41 % yield in two steps). Bis-cycloadduct **15** was subjected to e aromatization ($TiCl_4$, Zn, THF, RT, 1 h)[16] followed by hydrolysis of the silyl acetal (aq. HF, rile, 0 °C, 1 h), giving benzocyclobutenone **16** in 78 % yield. At this stage, the structure was guously reconfirmed by single-crystal X-ray diffraction analysis.[17] Ketone **16** was then ed to oxime **17** (NH_2OH-HCl, pyridine, MeOH, 60 °C, 18 h) in 88 % yield as a single isomer, for

16 M. A. Meador, H. Hart, *J. Org. Chem.* 1989, **54**, 2336–2341.
CrossRef | CAS | Web of Science® Times Cited: 16 | 'ejournals@cambridge - find full text'

17 CCDC 1543805 (16) contains the supplementary crystallographic data for this paper. These data can be obtained free of charge from The Cambridge Crystallographic Data Centre.

18 Nitrile oxide **18** could be stored at −18 °C for at least one month.

ELSEVIER



Outline

Abstract
Graphical abstract
1. Introduction
2. Results and discussio
3. Conclusions
4. Experimental
Acknowledgements
Supplementary data
Research Data
References and notes

An efficient phosphate sensor: tripodal quinoline excimer transduction

Avijit Pramanik, Gopal Das

⊞ Show more

https://

Research data for this article

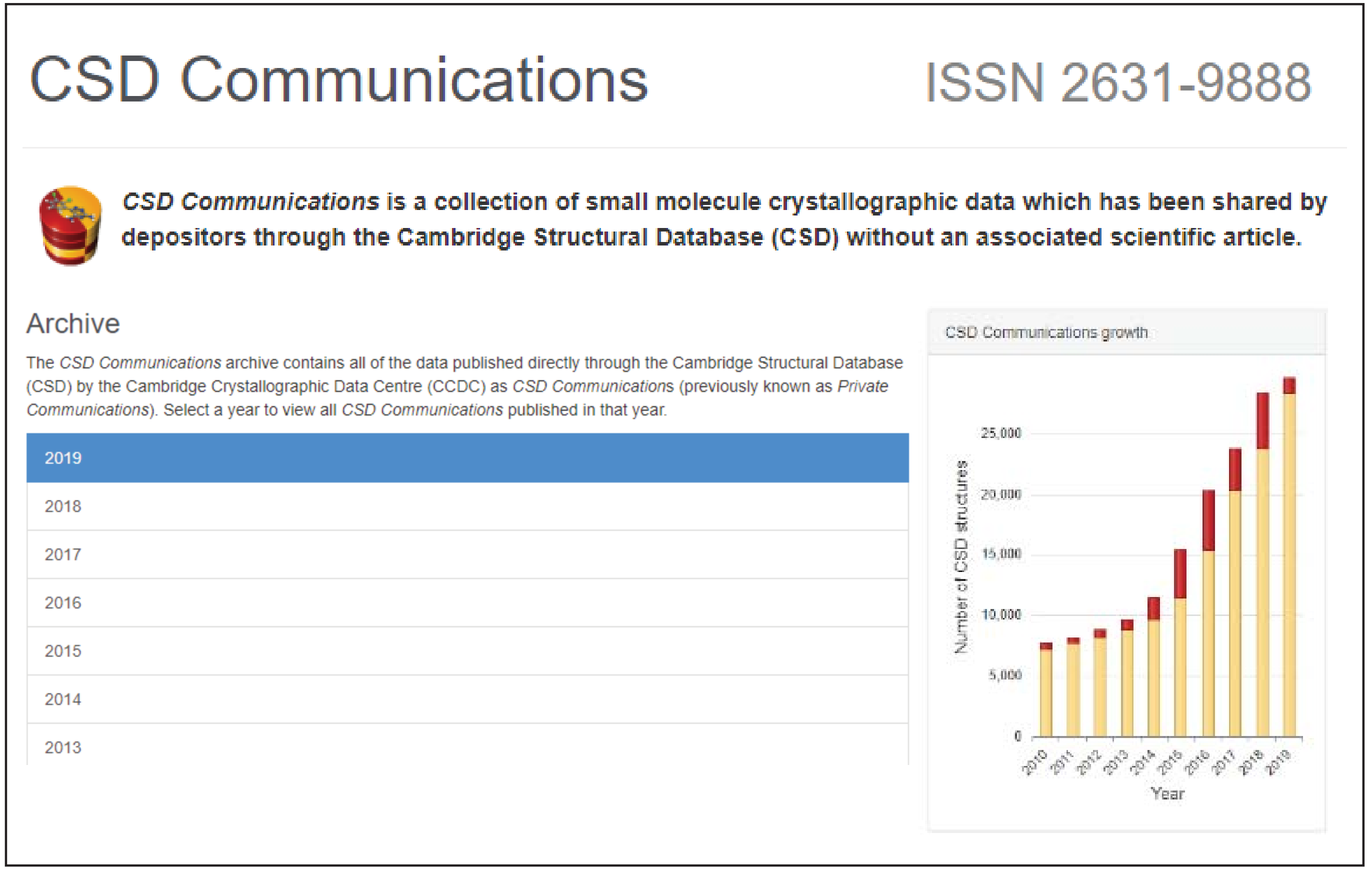
Cambridge Crystallographic Data Center
Crystallographic data
Data associated with the article:
CCDC 689113: Experimental Crystal Structure Determination ↗

SCHOLIX

Making data accessible though the CSD

Data not published in a scientific journal can be curated into the CSD and made available to the community as a **CSD Communication**

Structures from your PhD **thesis** can be made publicly available through the CSD.



CSD Communications

ISSN 2631-9888

CSD Communications **is a collection of small molecule crystallographic data which has been shared by depositors through the Cambridge Structural Database (CSD) without an associated scientific article.**

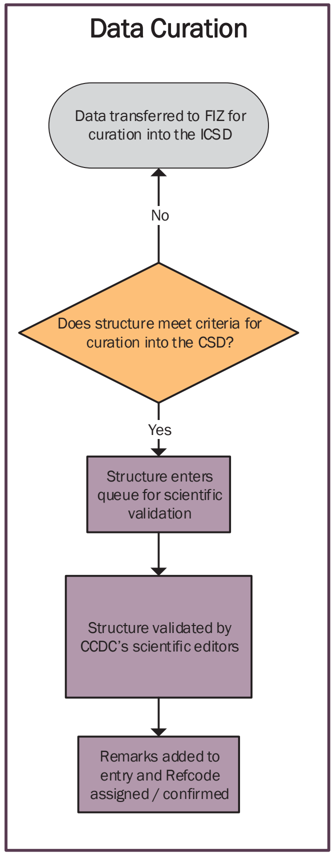
Archive

The *CSD Communications* archive contains all of the data published directly through the Cambridge Structural Database (CSD) by the Cambridge Crystallographic Data Centre (CCDC) as *CSD Communications* (previously known as *Private Communications*). Select a year to view all *CSD Communications* published in that year.

- 2019
- 2018
- 2017
- 2016
- 2015
- 2014
- 2013

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

Data Curation



Data Curation

Data transferred to FIZ for curation into the ICSD

No

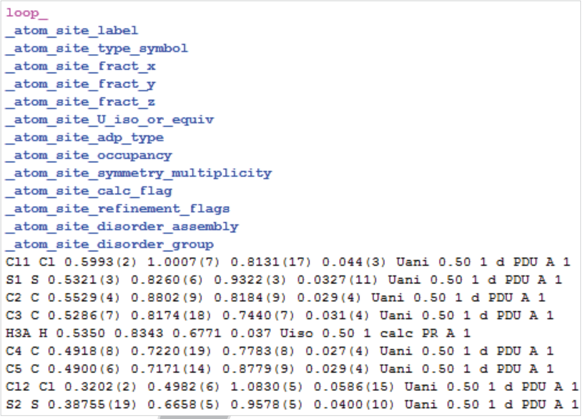
Does structure meet criteria for curation into the CSD?

Yes

Structure enters queue for scientific validation

Structure validated by CCDC's scientific editors

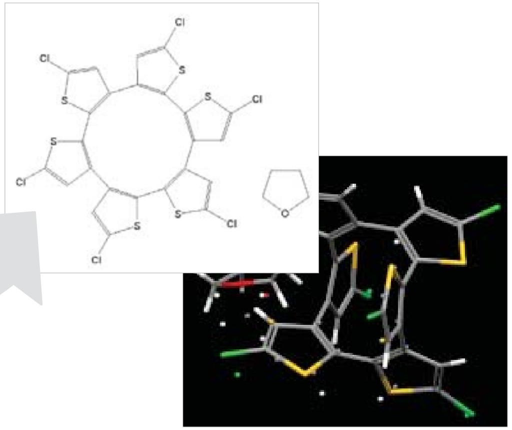
Remarks added to entry and Refcode assigned / confirmed



```
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



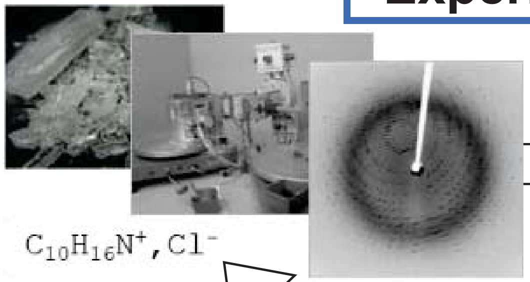
Enablers of FAIR Crystallographic Data

- **Standard formats and identifiers**
  - ❑Crystallographic Information Format and dictionaries (CIF)
  - ❑Standard Identifiers and associated infrastructure (DOIs, ORCID, InChI…)
- **Community stakeholders**
  - ❑Instrument providers and software developers adopting standards
  - ❑Publishers and editors encouraging use of standards for publication
  - ❑Repositories and databases providing access to enriched data
  - ❑International Unions supporting and promoting standards
  - ❑Individual researchers and others championing research data standards
- **Tools and services that make it easy to make data FAIR**

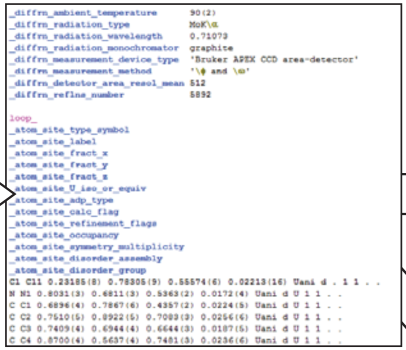
From Data to Knowledge

From Experiment to Knowledge

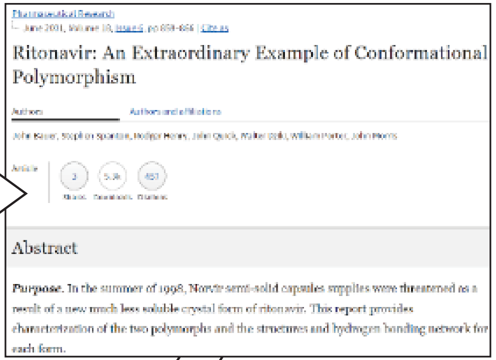
Experiment



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



```
_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(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 . .
```



Pharmaceutical Research
June 2001, Volume 18, Issue 6, pp 859–866 | Cite as

Ritonavir: An Extraordinary Example of Conformational Polymorphism

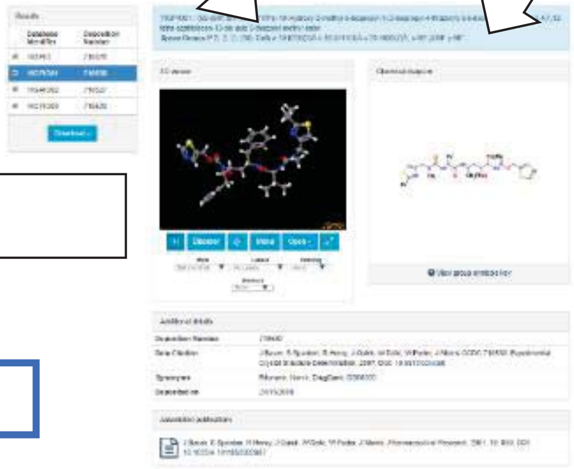
Authors | Authors and affiliations

John Bauer, Stephen Spanton, Rodger Henry, John Quick, Walter Dziki, William Porter, John Morris

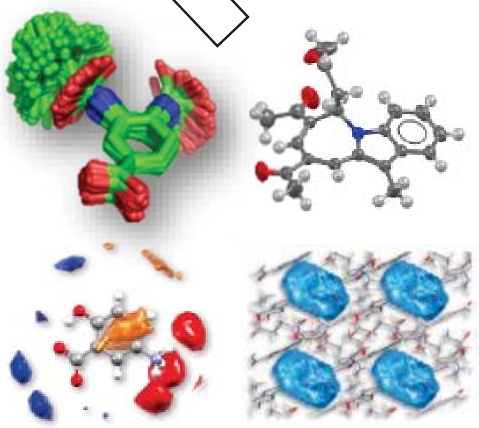
Article | 3 Shares | 5.3k Downloads | 487 Citations

Abstract

Purpose. In the summer of 1998, Norvir semi-solid capsules supplies were threatened as a result of a new much less soluble crystal form of ritonavir. This report provides characterization of the two polymorphs and the structures and hydrogen bonding network for each form.



CCDC



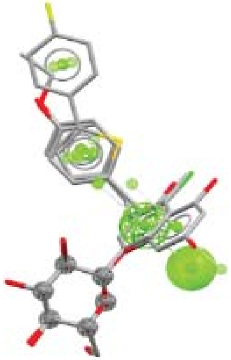
Knowledge

Advanced Services and Software

CSD-Enterprise

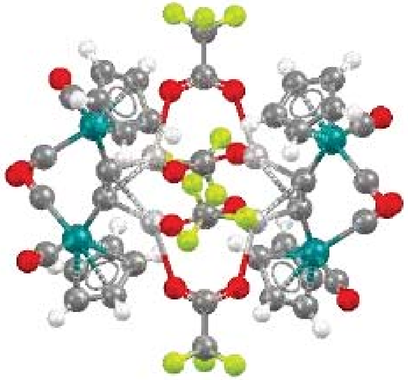
All CCDC applications and software

CSD-Discovery



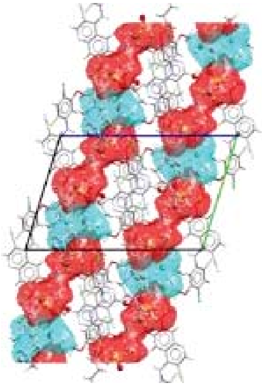
To discover new molecules with pharmaceutical applications

CSD-System



To search, visualise, analyse and communicate structural data

CSD-Materials

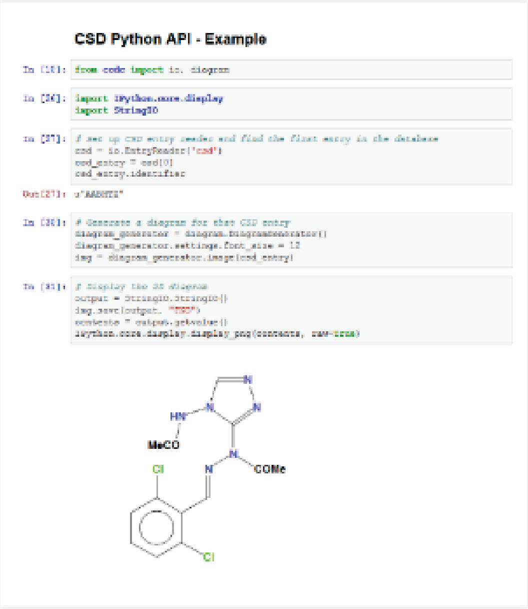


To understand and predict solid form stability and properties

The Cambridge Structural Database

Generating Insights

- The **CSD Python API** enables you to create tailored scripts using full array of CSD functionality
- Answer targeted research questions or integrate access with other software
- Functions include:
    - Full search capabilities
    - Geometry analysis
    - Interaction analysis
    - Descriptor calculation
    - 2D diagram generation

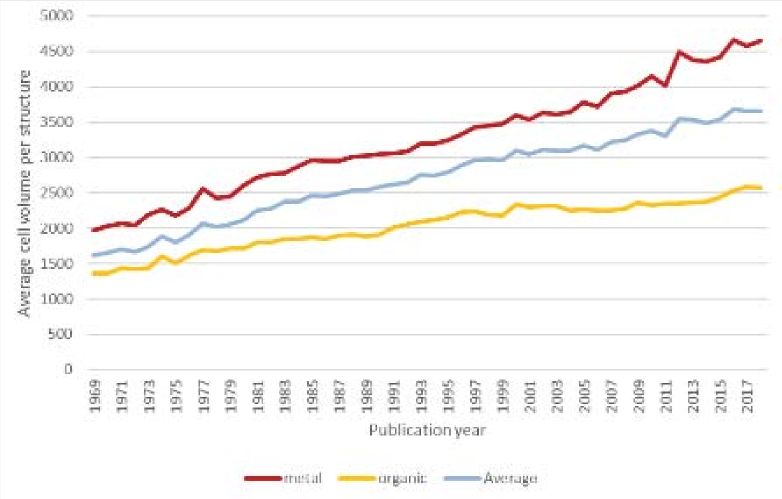


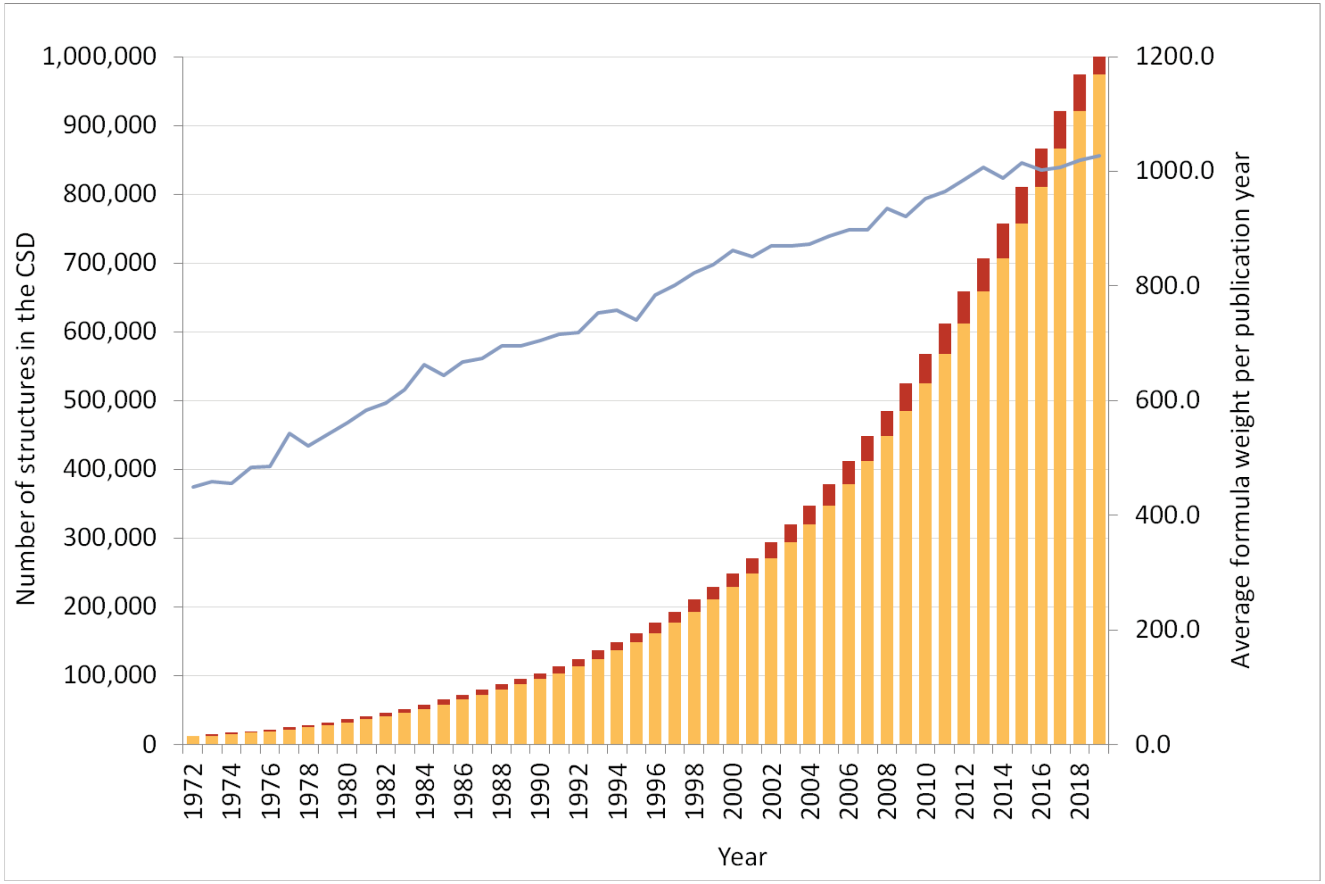
CSD Python API - Example

```
In [18]: from ccdc import io, diagram

In [26]: import IPython.core.display
         import StringIO

In [27]: # Set up CSD entry reader and find the first entry in the database
         csd = io.EntryReader('csd')
         csd_entry = csd[0]
         csd_entry.identifier

Out[27]: u'AABHTZ'

In [30]: # Generate a diagram for that CSD entry
         diagram_generator = diagram.DiagramGenerator()
         diagram_generator.settings.font_size = 12
         img = diagram_generator.image(csd_entry)

In [31]: # Display the 2D diagram
         output = StringIO.StringIO()
         img.save(output, "PNG")
         contents = output.getvalue()
         IPython.core.display.display_png(contents, raw=True)
```

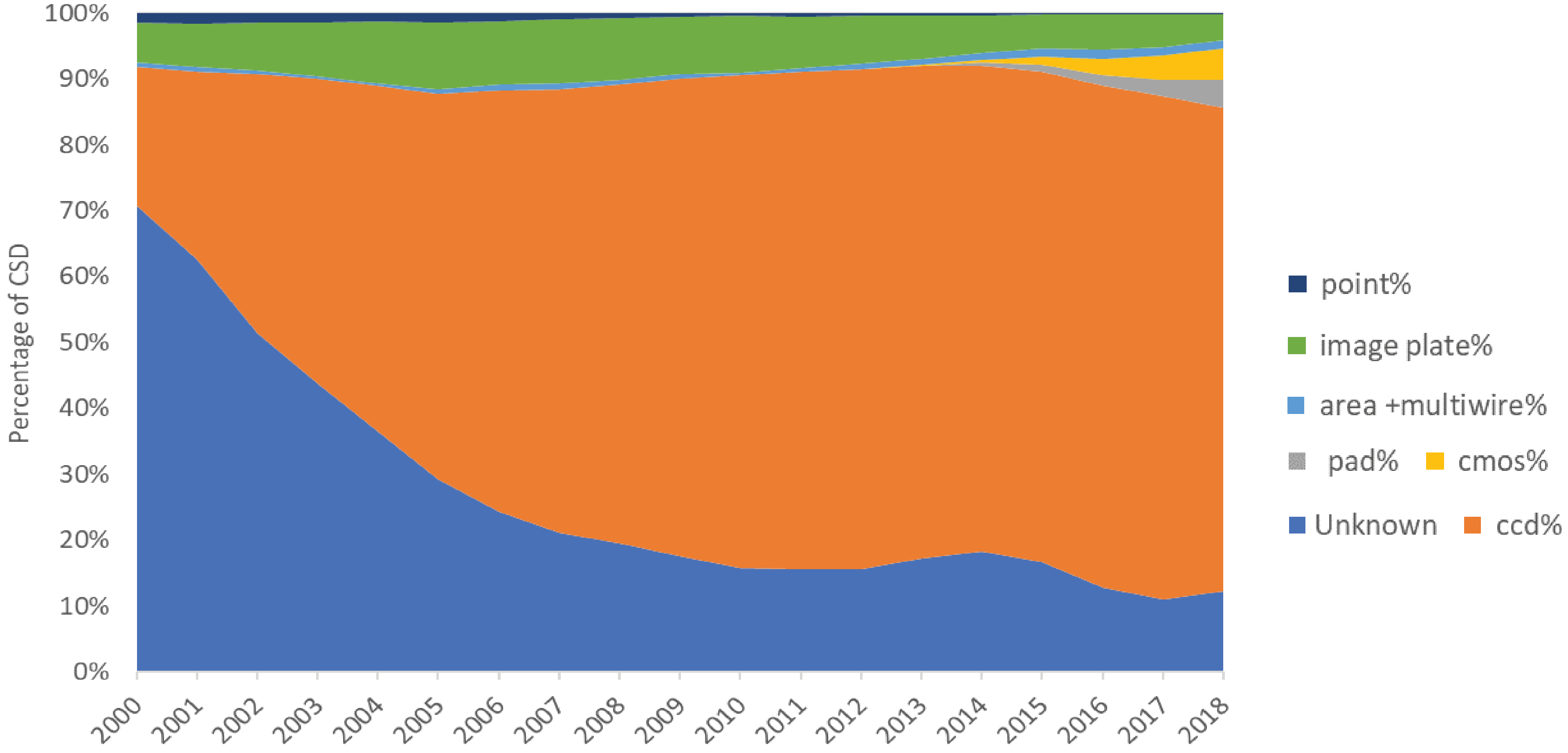
Increasing Complexity

- Increasing:
  - Formula weights
  - Unit Cells
  - Number of elements

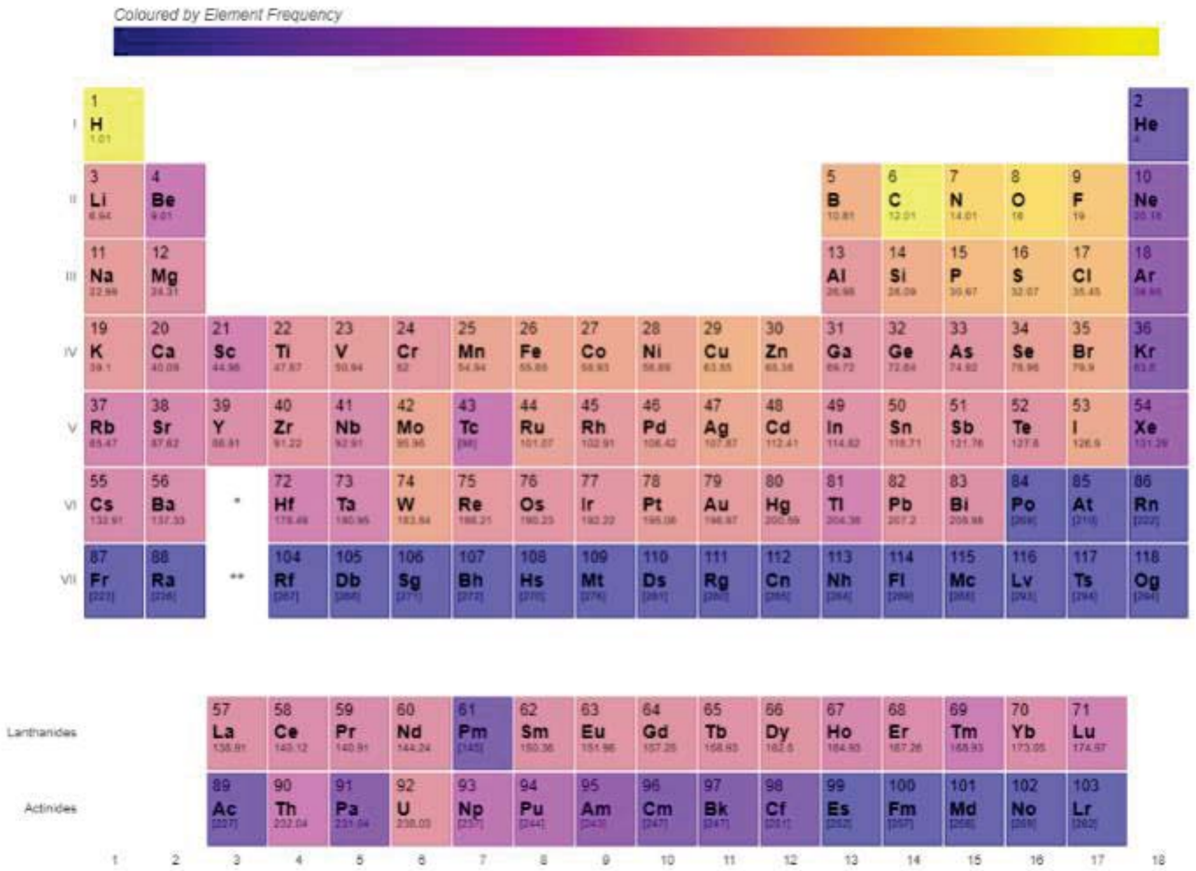


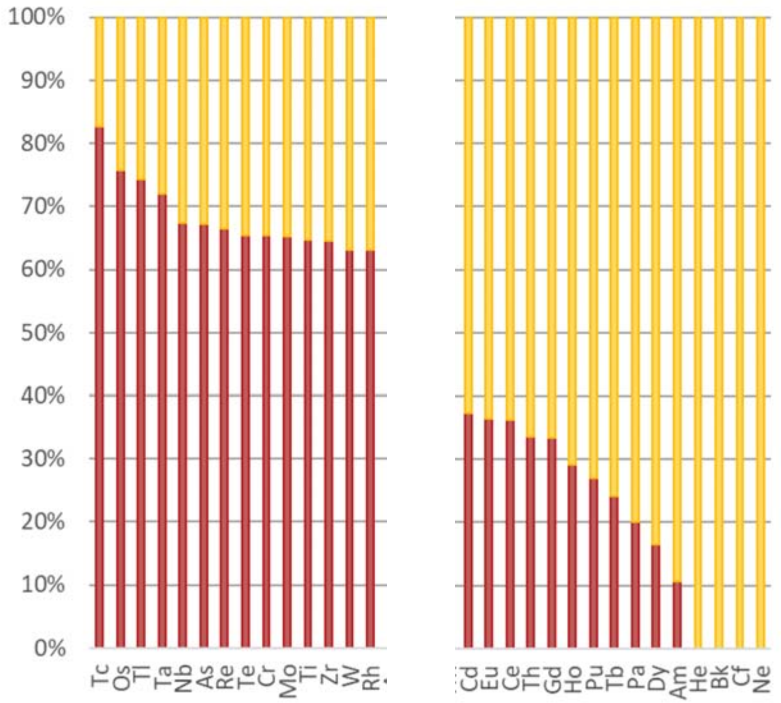


Trends in Experimentation



Elements in the CSD

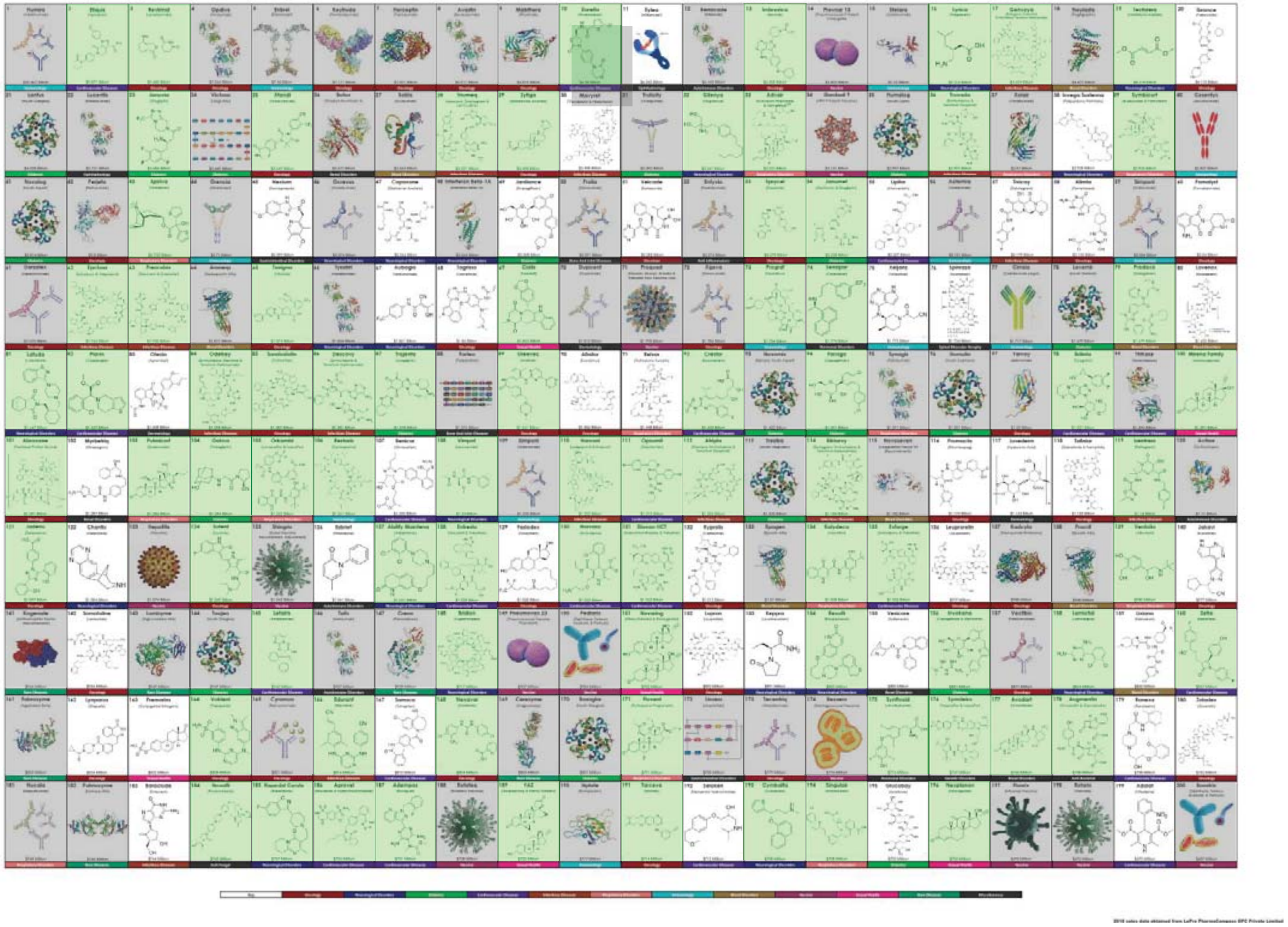




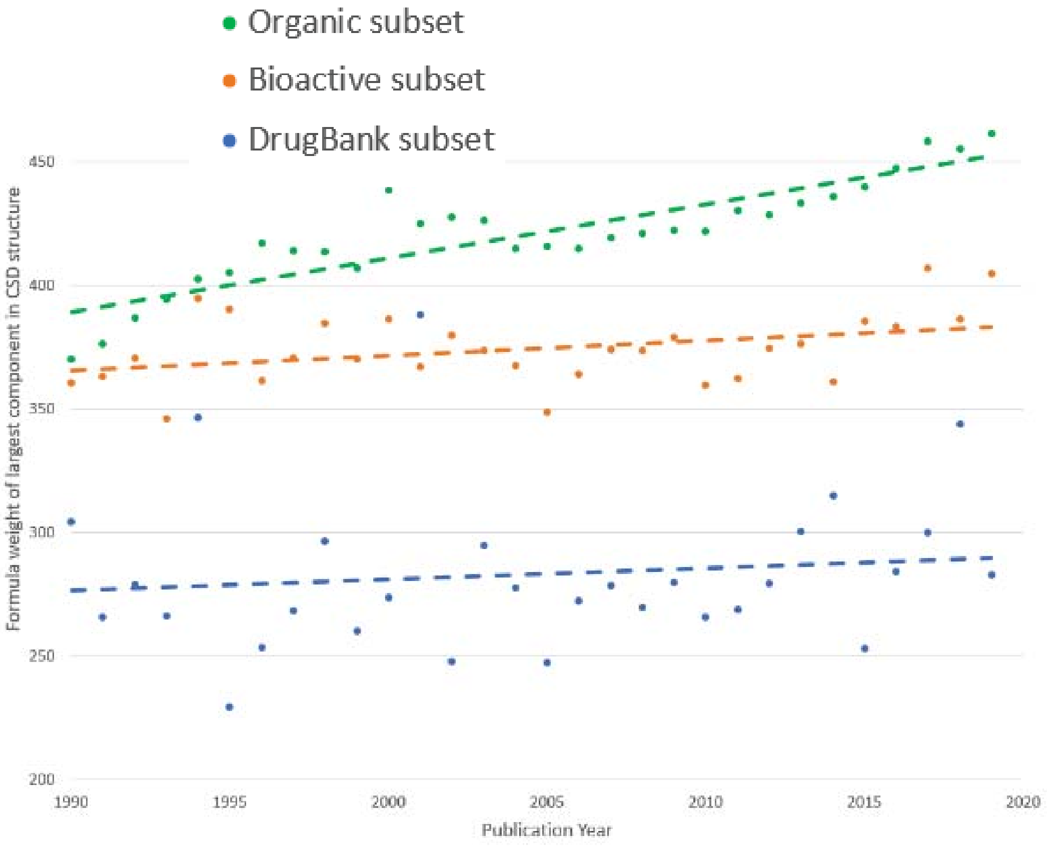
Percentage of structures that contain each element from before (red) and after (yellow) 2009

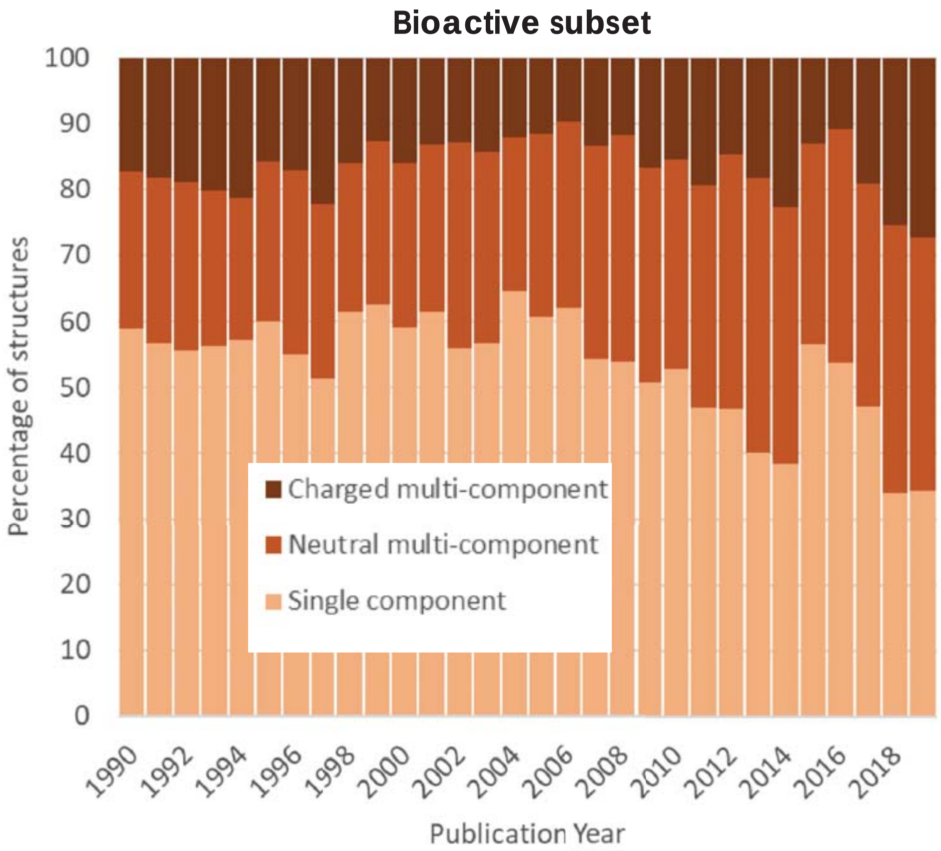
Drugs

- Top 200 Pharmaceutical Products
  - **By retail sales in 2018**
  - Produced by the Njarðarson Group
  - The University of Arizona
  - Drugs already in the CSD coloured green
- *J. Chem. Ed.* **2010**, *87*, 1348



Identifying Trends in Drug Structures



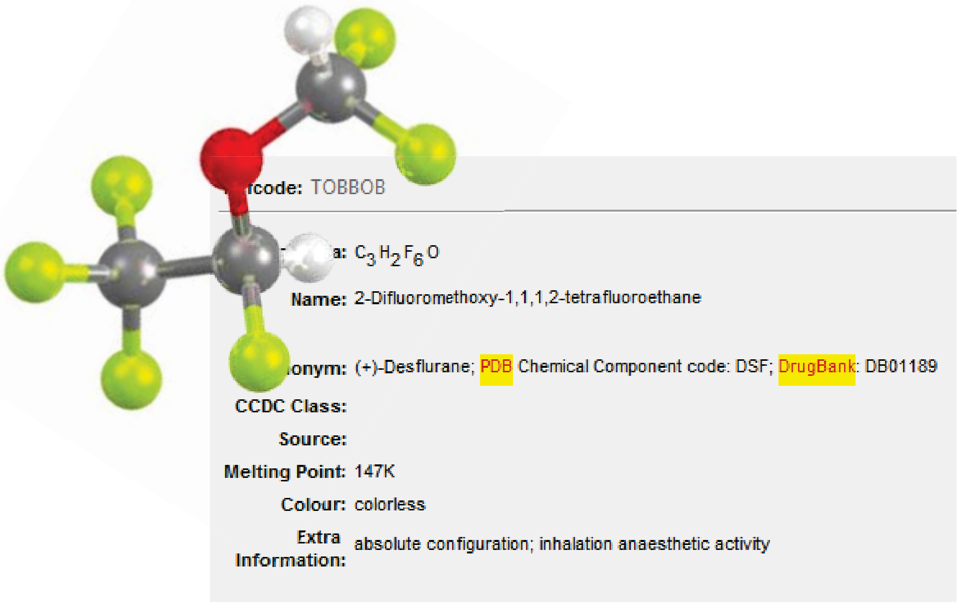


https://www.ccdc.cam.ac.uk/Community/blog/insights-into-drug-like-compounds-from-crystal-data/

The CSD and the PDB

Linking

- Between CSD and PDB ligands



fcode: TOBBOB

a: $C_3H_2F_6O$

Name: 2-Difluoromethoxy-1,1,1,2-tetrafluoroethane

onym: (+)-Desflurane; PDB Chemical Component code: DSF; DrugBank: DB01189

CCDC Class:

Source:

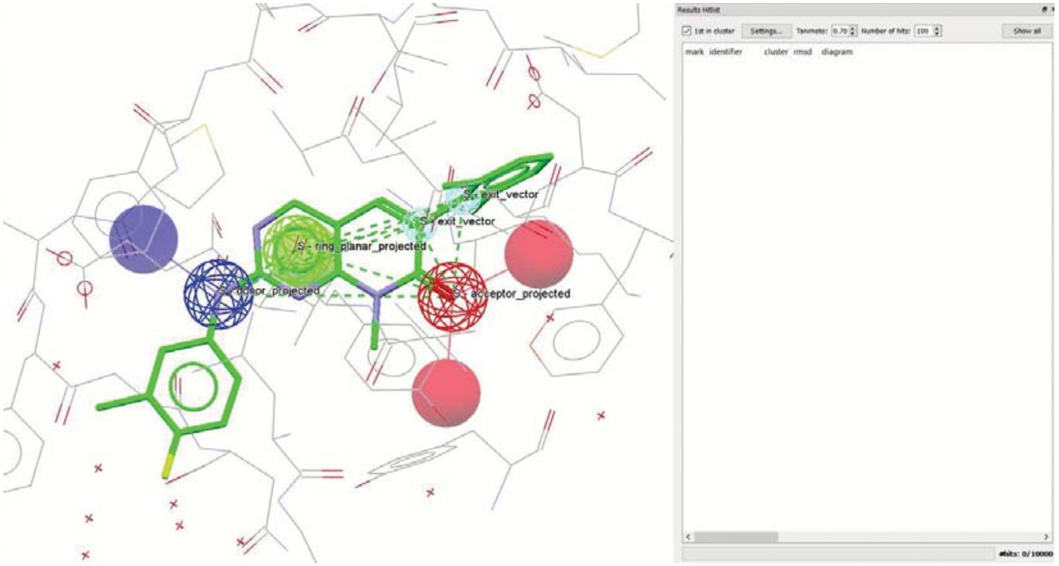
Melting Point: 147K

Colour: colorless

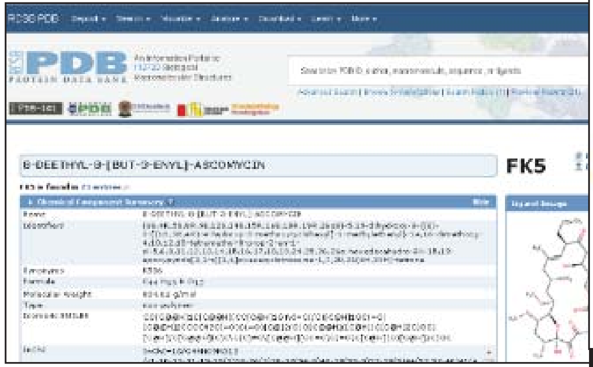
Extra Information: absolute configuration; inhalation anaesthetic activity

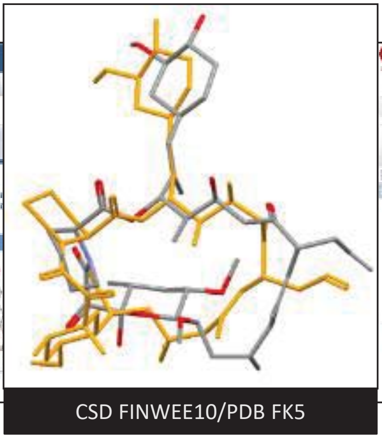
CSD-CrossMiner

- Pharmacophore query tool
- Searches the CSD and PDB

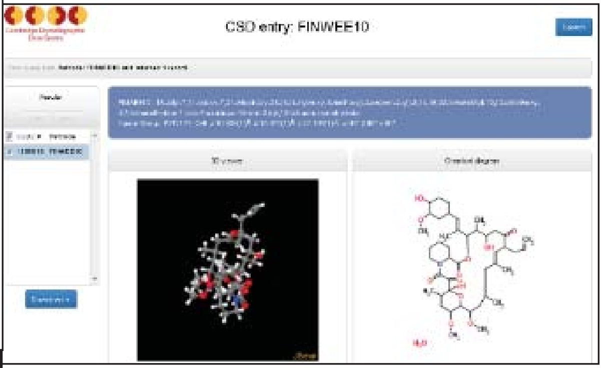


Using the Data

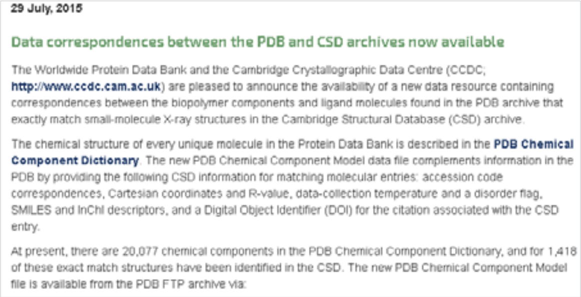




CSD FINWEE10/PDB FK5



Match PDB ligands to best representative CSD molecules



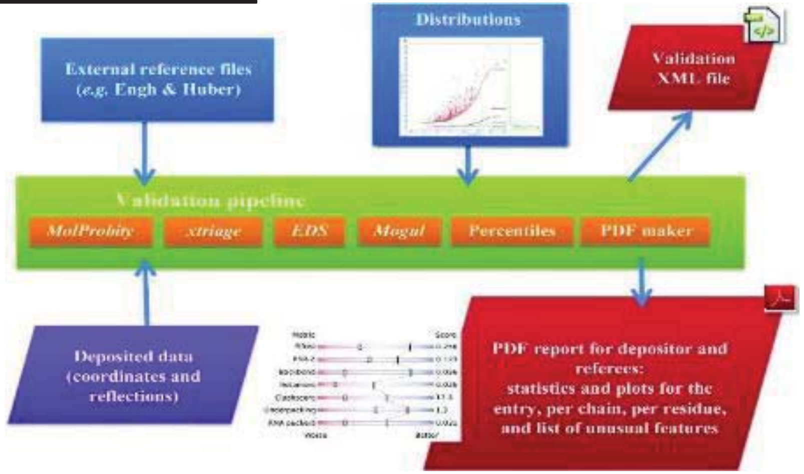
29 July, 2015

Data correspondences between the PDB and CSD archives now available

The Worldwide Protein Data Bank and the Cambridge Crystallographic Data Centre (CCDC; **http://www.ccdc.cam.ac.uk**) are pleased to announce the availability of a new data resource containing correspondences between the biopolymer components and ligand molecules found in the PDB archive that exactly match small-molecule X-ray structures in the Cambridge Structural Database (CSD) archive.

The chemical structure of every unique molecule in the Protein Data Bank is described in the **PDB Chemical Component Dictionary**. The new PDB Chemical Component Model data file complements information in the PDB by providing the following CSD information for matching molecular entries: accession code correspondences, Cartesian coordinates and R-value, data-collection temperature and a disorder flag, SMILES and InChI descriptors, and a Digital Object Identifier (DOI) for the citation associated with the CSD entry.

At present, there are 20,077 chemical components in the PDB Chemical Component Dictionary, and for 1,418 of these exact match structures have been identified in the CSD. The new PDB Chemical Component Model file is available from the PDB FTP archive via:

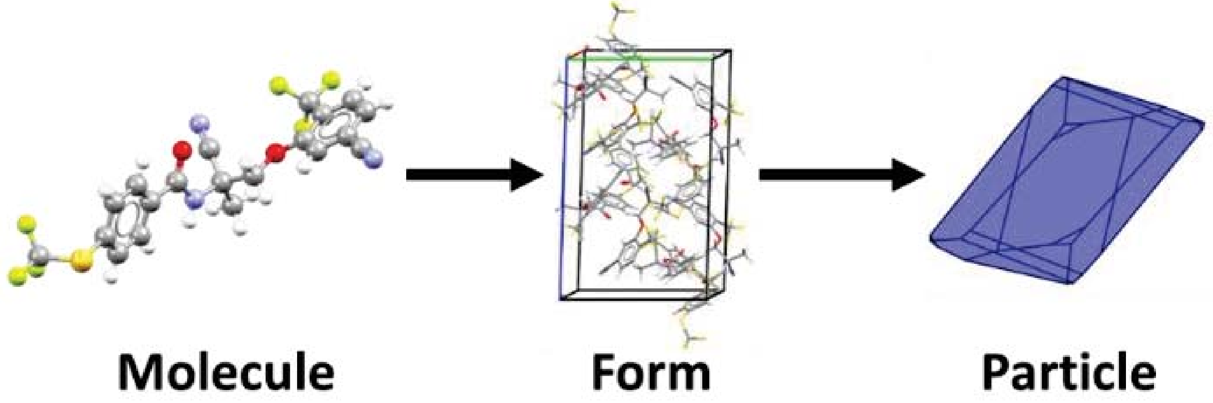


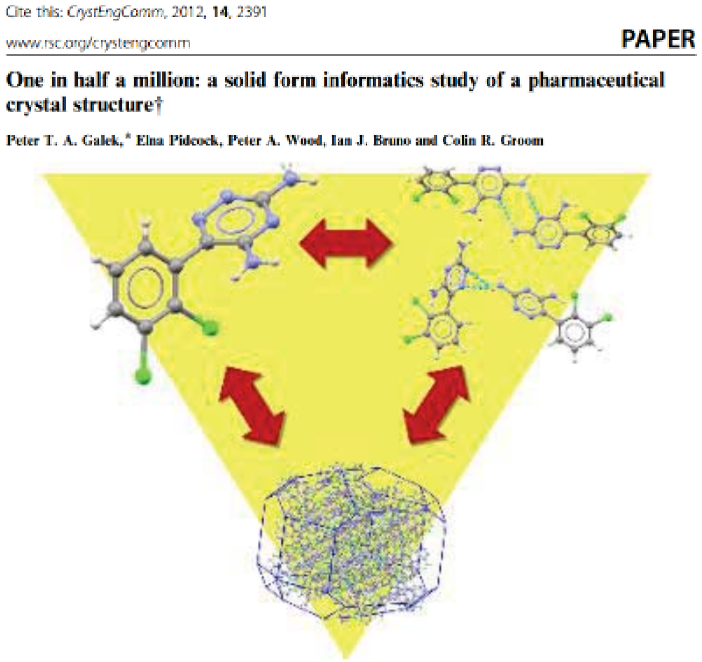
Solid Form Informatics

- The term “**solid form informatics**” first introduced in mid-2000s by Bob Docherty (Pfizer):

Use of structural knowledge to inform key decisions in pharmaceutical development

- Solid form informatics now a key part of the solid form development workflow at most major pharmaceutical companies




Cite this: *CrystEngComm*, 2012, **14**, 2391

www.rsc.org/crystengcomm

PAPER

One in half a million: a solid form informatics study of a pharmaceutical crystal structure†

Peter T. A. Galek,[a] Elna Pidcock, Peter A. Wood, Ian J. Bruno and Colin R. Groom

JPP
Journal of Pharmacy
And Pharmacology

Research Paper

The integration of solid-form informatics into solid-form selection

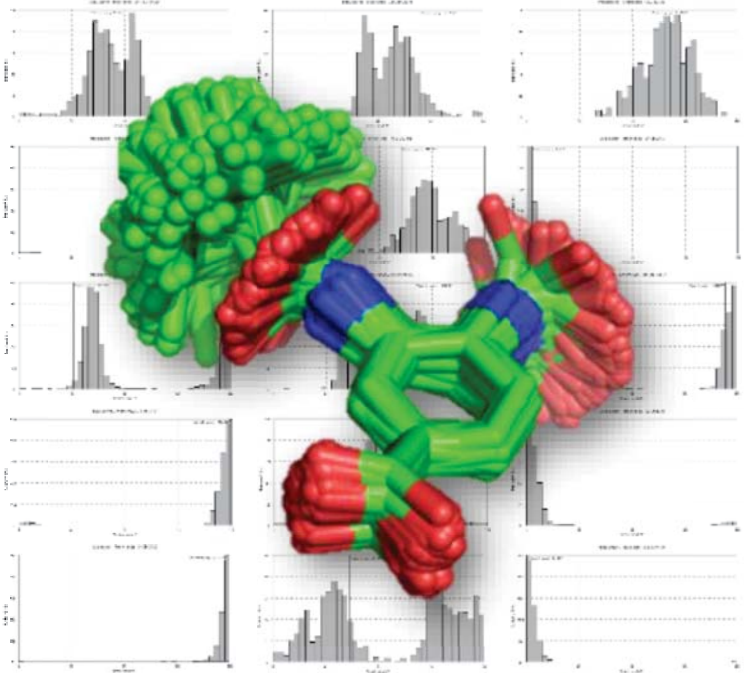
Neil Feeder[a], Elna Pidcock[a], Anthony M. Reilly[a], Ghazala Sadiq[a], Cheryl L. Doherty[b], Kevin R. Back[b], Paul Meenan[c] and Robert Docherty[b]

[a]The Cambridge Crystallographic Data Centre, Cambridge, [b]Pharmaceutical Science, Pfizer Global R&D, Sandwich, UK and [c]Pharmaceutical Science, Pfizer Global R&D, Groton, USA

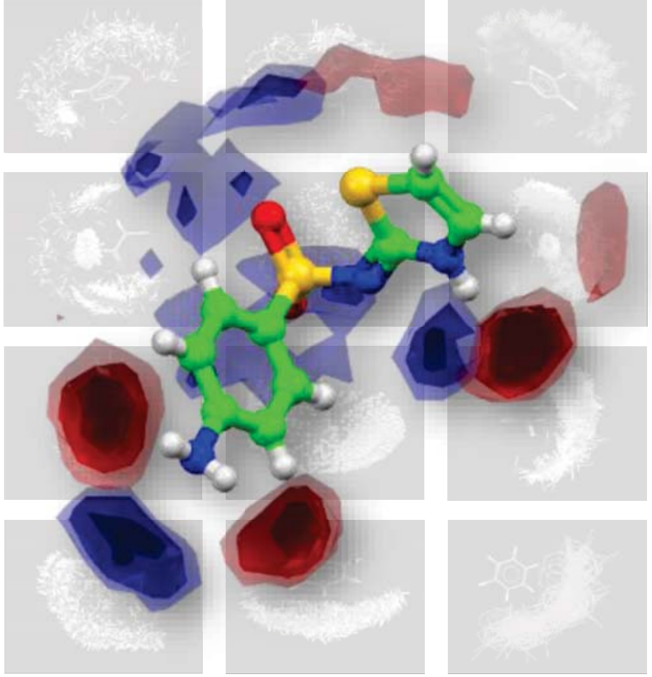
From Data to Knowledge

Individual data points from different datasets combine to provide information that aids in the discovery and optimisation of new chemical entities

Molecular Shape



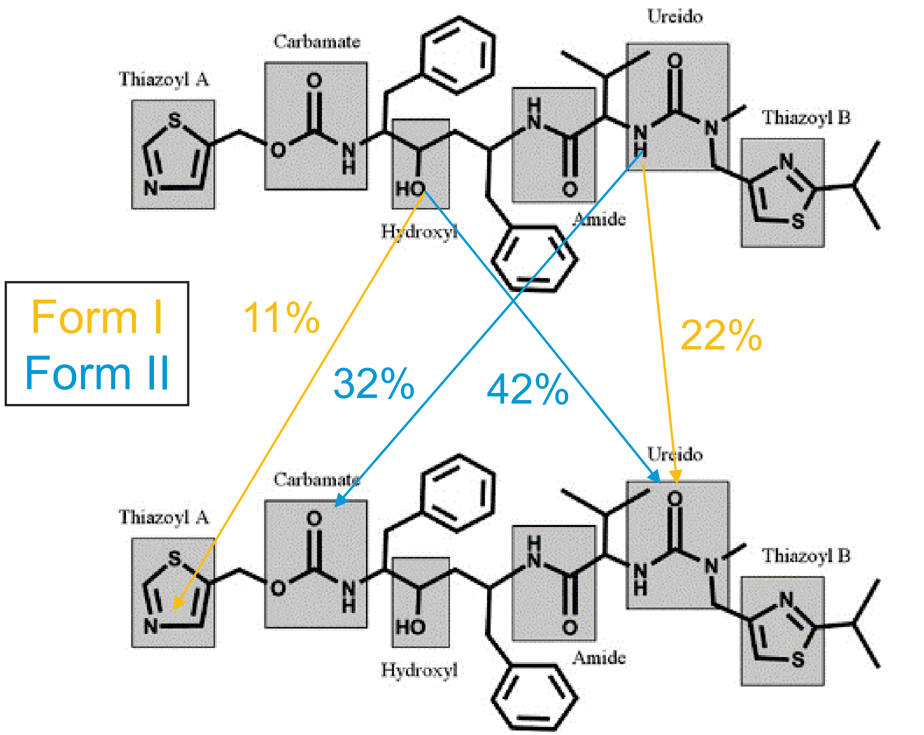
Molecular Interactions



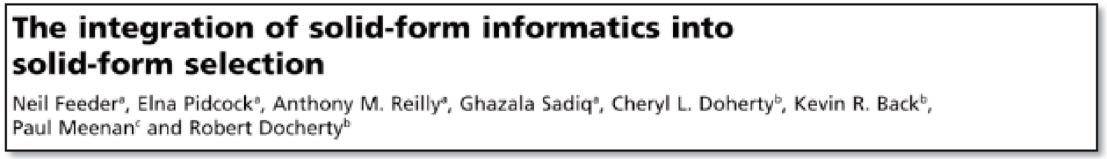
Taylor *et al. J. Chem. Inf. Model.*, (2014) 54 (9), 2500. Wood, P. A. *et al. CrystEngComm* (2013) **15**, 65

Predicting Unlikely Interactions

Predictive analytics is used to identify the likelihood of specific molecular interactions occurring from similar crystal structures

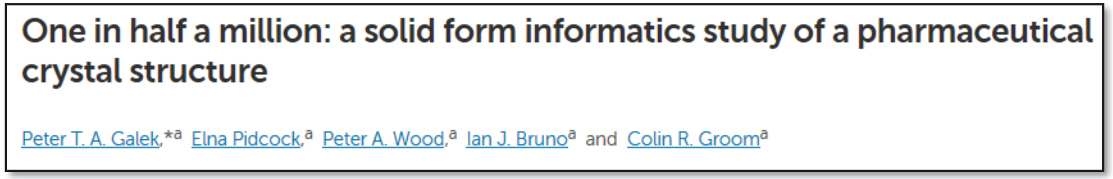


Galek *et al*, *CrystEngComm,* (2009), 11, 2634 - 2639



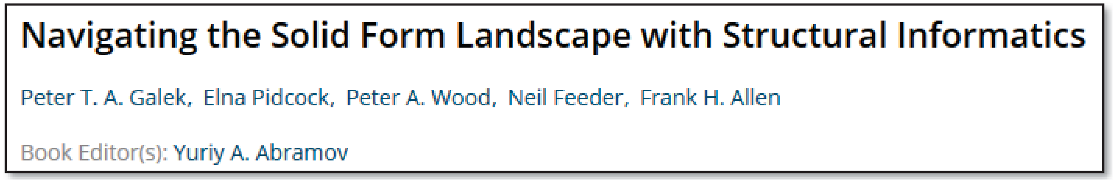
The integration of solid-form informatics into solid-form selection

Neil Feeder[a], Elna Pidcock[a], Anthony M. Reilly[a], Ghazala Sadiq[a], Cheryl L. Doherty[b], Kevin R. Back[b], Paul Meenan[c] and Robert Docherty[b]



One in half a million: a solid form informatics study of a pharmaceutical crystal structure

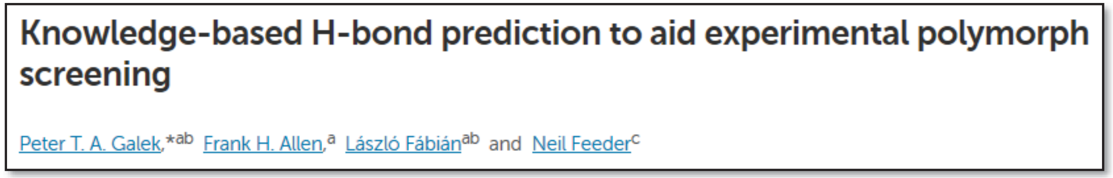
Peter T. A. Galek,[*a] Elna Pidcock,[a] Peter A. Wood,[a] Ian J. Bruno[a] and Colin R. Groom[a]



Navigating the Solid Form Landscape with Structural Informatics

Peter T. A. Galek, Elna Pidcock, Peter A. Wood, Neil Feeder, Frank H. Allen

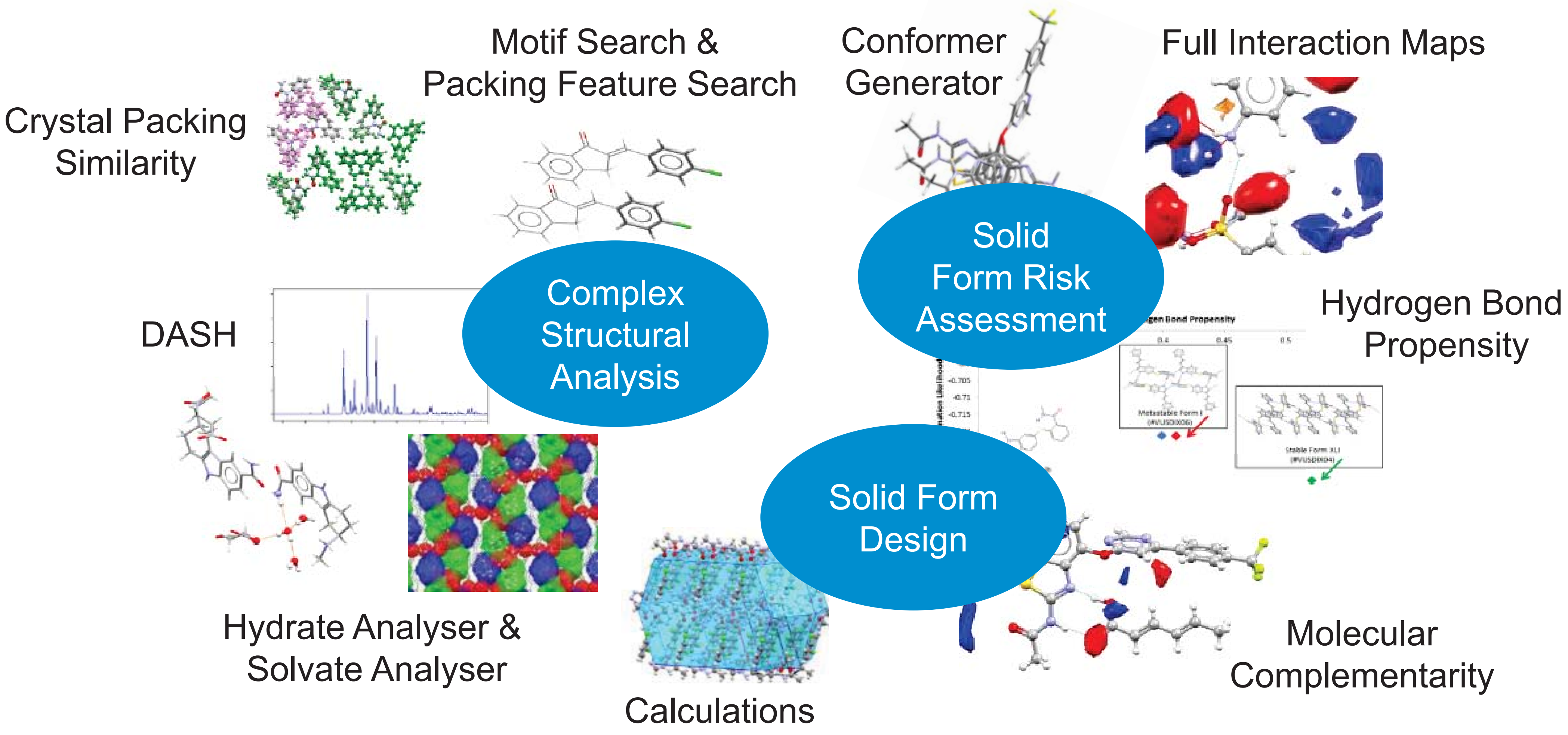
Book Editor(s): Yuriy A. Abramov



Knowledge-based H-bond prediction to aid experimental polymorph screening

Peter T. A. Galek,[*ab] Frank H. Allen,[a] László Fábián[ab] and Neil Feeder[c]

CSD-Materials: Targeted Solutions



Crystal Packing Similarity

Motif Search & Packing Feature Search

Conformer Generator

Full Interaction Maps

DASH

Complex Structural Analysis

Solid Form Risk Assessment

Hydrogen Bond Propensity

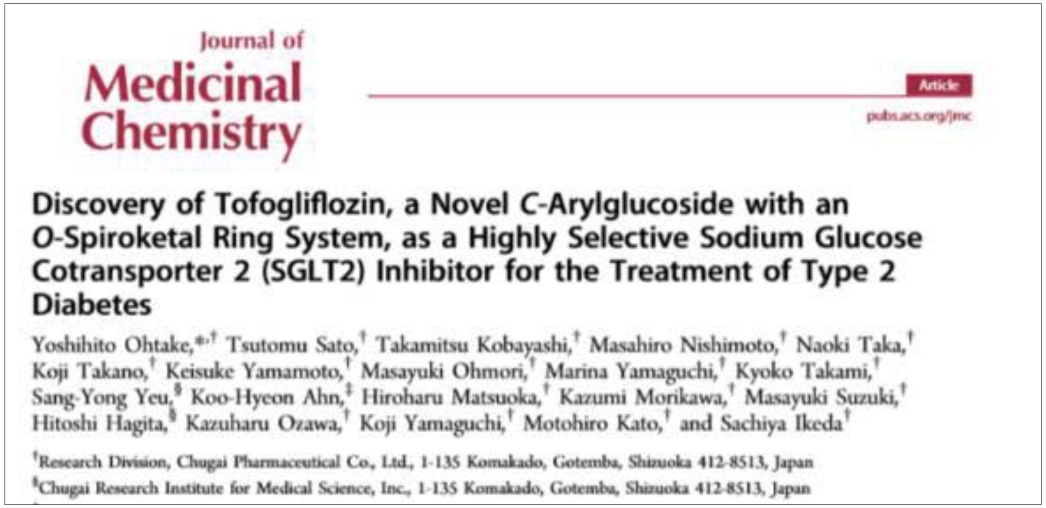
Solid Form Design

Hydrate Analyser & Solvate Analyser

Calculations

Molecular Complementarity

Individual value



Journal of Medicinal Chemistry

Article

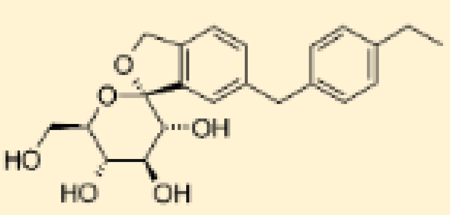
pubs.acs.org/jmc

Discovery of Tofogliflozin, a Novel *C*-Arylglucoside with an *O*-Spiroketal Ring System, as a Highly Selective Sodium Glucose Cotransporter 2 (SGLT2) Inhibitor for the Treatment of Type 2 Diabetes

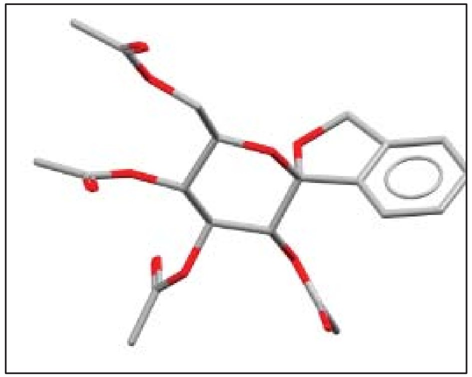
Yoshihito Ohtake,*[†] Tsutomu Sato,[†] Takamitsu Kobayashi,[†] Masahiro Nishimoto,[†] Naoki Taka,[†] Koji Takano,[†] Keisuke Yamamoto,[†] Masayuki Ohmori,[†] Marina Yamaguchi,[†] Kyoko Takami,[†] Sang-Yong Yeu,[§] Koo-Hyeon Ahn,[‡] Hiroharu Matsuoka,[†] Kazumi Morikawa,[†] Masayuki Suzuki,[†] Hitoshi Hagita,[§] Kazuharu Ozawa,[†] Koji Yamaguchi,[†] Motohiro Kato,[†] and Sachiya Ikeda[†]

[†] Research Division, Chugai Pharmaceutical Co., Ltd., 1-135 Komakado, Gotemba, Shizuoka 412-8513, Japan

[§] Chugai Research Institute for Medical Science, Inc., 1-135 Komakado, Gotemba, Shizuoka 412-8513, Japan



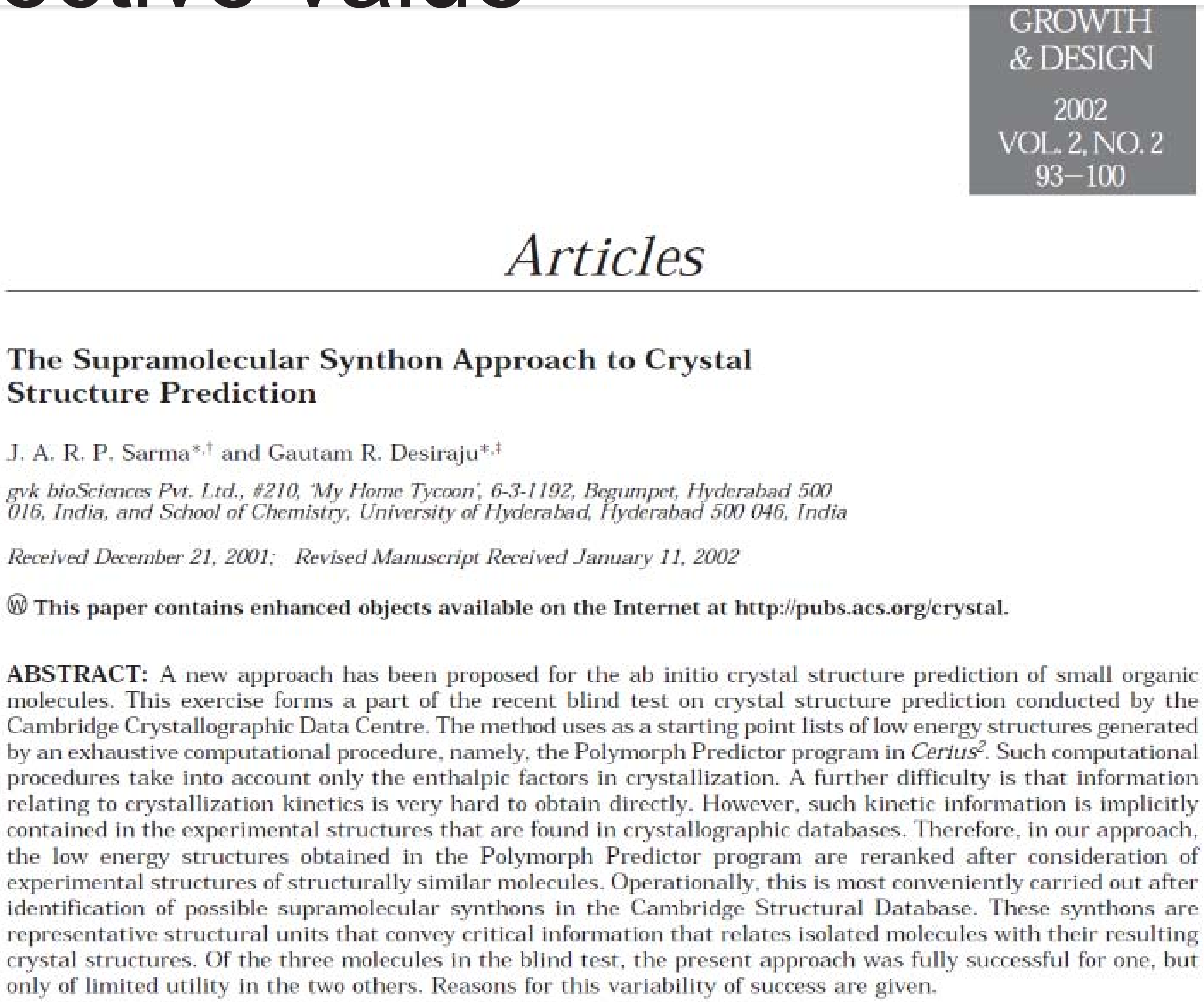
16d



"A **search** of the Cambridge Structural **Database** using a series of **pharmacophore queries** led to the discovery of an O-spiroketal C-arylglucoside scaffold. Subsequent chemical examination combined with computational modelling resulted in the identification of the clinical candidate 16d (CSG452, tofogliflozin), which is currently under phase III clinical trials."

Yoshihito Ohtake et al *Journal of Medicinal Chemistry* 2012 *55* (17), 7828-7840 (Roche, Chugai)

Collective value



GROWTH
& DESIGN
2002
VOL. 2, NO. 2
93–100

Articles

The Supramolecular Synthon Approach to Crystal Structure Prediction

J. A. R. P. Sarma*,† and Gautam R. Desiraju*,‡

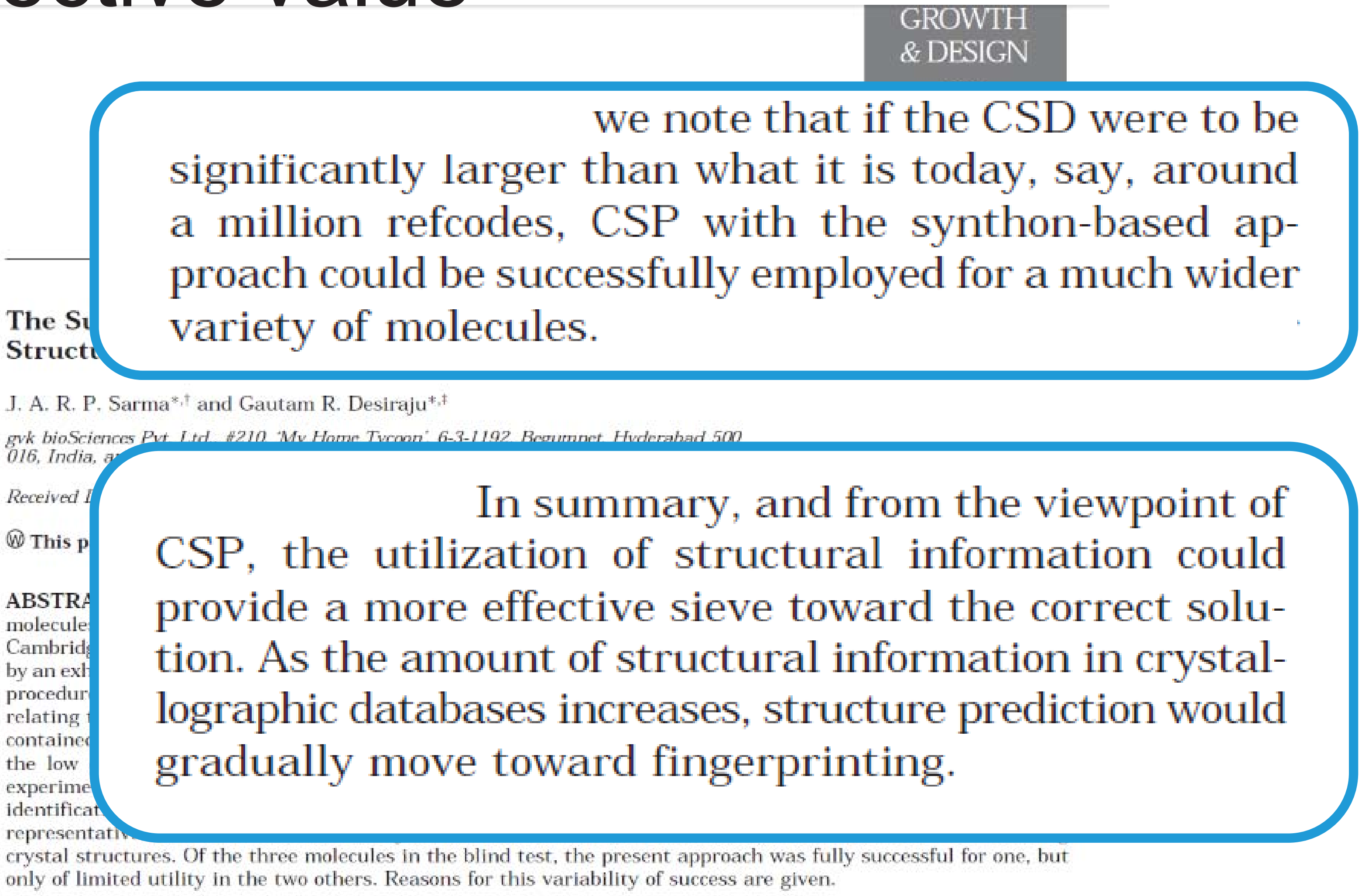
gvk bioSciences Pvt. Ltd., #210, 'My Home Tycoon', 6-3-1192, Begumpet, Hyderabad 500 016, India, and School of Chemistry, University of Hyderabad, Hyderabad 500 046, India

Received December 21, 2001; Revised Manuscript Received January 11, 2002

Ⓦ **This paper contains enhanced objects available on the Internet at http://pubs.acs.org/crystal.**

ABSTRACT: A new approach has been proposed for the ab initio crystal structure prediction of small organic molecules. This exercise forms a part of the recent blind test on crystal structure prediction conducted by the Cambridge Crystallographic Data Centre. The method uses as a starting point lists of low energy structures generated by an exhaustive computational procedure, namely, the Polymorph Predictor program in *Cerius*^{2}. Such computational procedures take into account only the enthalpic factors in crystallization. A further difficulty is that information relating to crystallization kinetics is very hard to obtain directly. However, such kinetic information is implicitly contained in the experimental structures that are found in crystallographic databases. Therefore, in our approach, the low energy structures obtained in the Polymorph Predictor program are reranked after consideration of experimental structures of structurally similar molecules. Operationally, this is most conveniently carried out after identification of possible supramolecular synthons in the Cambridge Structural Database. These synthons are representative structural units that convey critical information that relates isolated molecules with their resulting crystal structures. Of the three molecules in the blind test, the present approach was fully successful for one, but only of limited utility in the two others. Reasons for this variability of success are given.

Collective value



GROWTH
& DESIGN

The Su
Structu

J. A. R. P. Sarma[*,†] and Gautam R. Desiraju[*,‡]

*gvk bioSciences Pvt. Ltd., #210 'My Home Tycoon', 6-3-1192, Begumpet, Hyderabad 500
016, India, a*

Received I

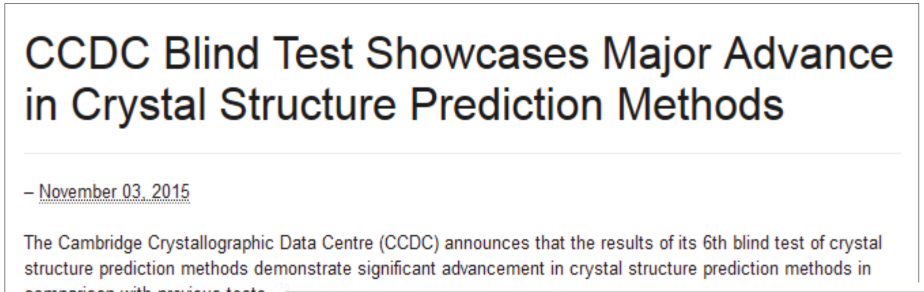
Ⓦ **This p**

ABSTRA
molecule
Cambrid
by an exh
procedur
relating
containe
the low
experime
identificat
representati
crystal structures. Of the three molecules in the blind test, the present approach was fully successful for one, but only of limited utility in the two others. Reasons for this variability of success are given.

we note that if the CSD were to be significantly larger than what it is today, say, around a million refcodes, CSP with the synthon-based approach could be successfully employed for a much wider variety of molecules.

In summary, and from the viewpoint of CSP, the utilization of structural information could provide a more effective sieve toward the correct solution. As the amount of structural information in crystallographic databases increases, structure prediction would gradually move toward fingerprinting.

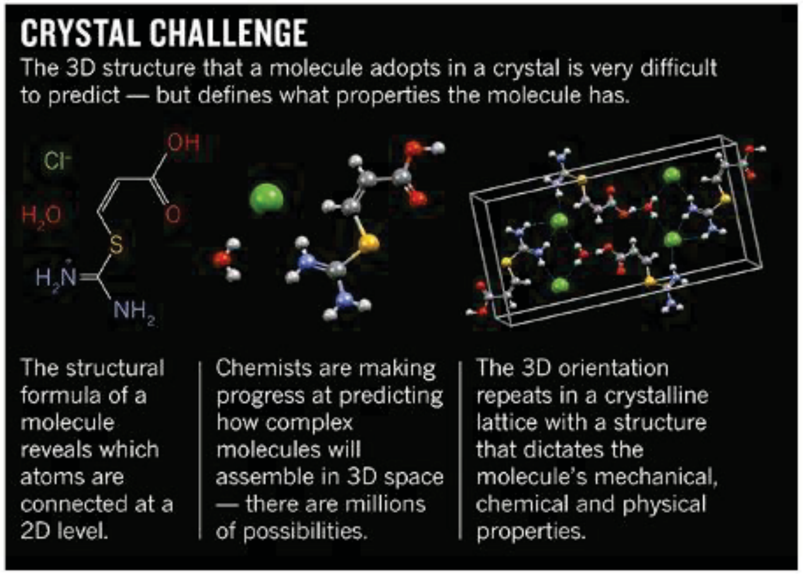
Using the Collection



CCDC Blind Test Showcases Major Advance in Crystal Structure Prediction Methods

– November 03, 2015

The Cambridge Crystallographic Data Centre (CCDC) announces that the results of its 6th blind test of crystal structure prediction methods demonstrate significant advancement in crystal structure prediction methods in comparison with previous tests. of polymorphs, salts and hydrat experimental structures were pr



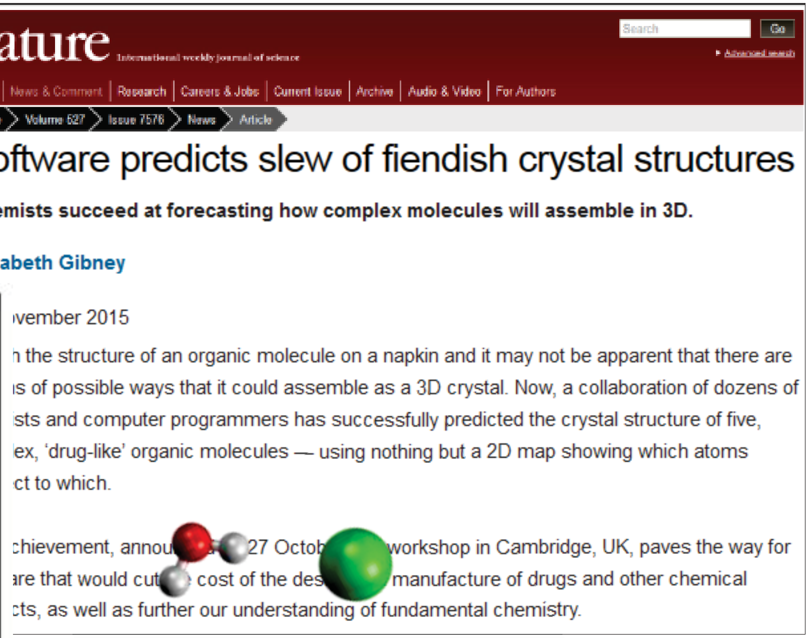
CRYSTAL CHALLENGE

The 3D structure that a molecule adopts in a crystal is very difficult to predict — but defines what properties the molecule has.

The structural formula of a molecule reveals which atoms are connected at a 2D level.

Chemists are making progress at predicting how complex molecules will assemble in 3D space — there are millions of possibilities.

The 3D orientation repeats in a crystalline lattice with a structure that dictates the molecule's mechanical, chemical and physical properties.



nature International weekly journal of science
Search
Go
Advanced search
Home | News & Comment | Research | Careers & Jobs | Current Issue | Archive | Audio & Video | For Authors
Archive > Volume 527 > Issue 7576 > News > Article

Software predicts slew of fiendish crystal structures

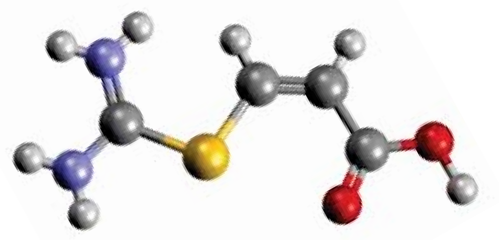
Chemists succeed at forecasting how complex molecules will assemble in 3D.

Elizabeth Gibney

vember 2015

the structure of an organic molecule on a napkin and it may not be apparent that there are is of possible ways that it could assemble as a 3D crystal. Now, a collaboration of dozens of sts and computer programmers has successfully predicted the crystal structure of five, ex, 'drug-like' organic molecules — using nothing but a 2D map showing which atoms ct to which.

chievement, annou 27 Octob workshop in Cambridge, UK, paves the way for are that would cut cost of the des manufacture of drugs and other chemical cts, as well as further our understanding of fundamental chemistry.



Thank you