25–29 August 2013

ECM 28

Data quality and the value of structural databases

Colin Groom

Crystallographic Information and Data Management
A Satellite Symposium to the 28th European Crystallographic Meeting IV.
Towards ever better science

Sunday, August 25, 2013

Contents

- The CCDC
- The CSD
- Creating the CSD
- Curating the CSD
- Current challenges
  - Beyond structural chemists
  - Structure publishing
  - Non-sematic data
- Unanswered questions

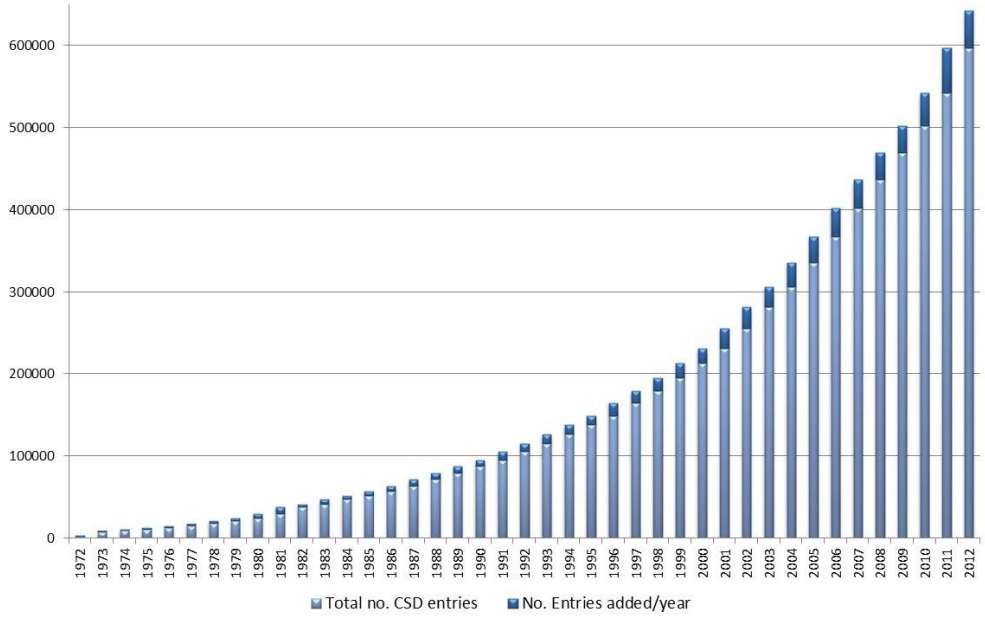
The Cambridge Crystallographic Data Centre

- A not-for-profit, charitable institution, established 1965
- Self-financing and self-administering
  - Funded by contributions from user community
- A University of Cambridge Partner Institute
- A UK Government Independent Research Organisation
- Around 50 members of staff
  - Scientists, software developers, IT, finance and user services

"advancement and promotion of the science of chemistry and crystallography for the public benefit"

The Cambridge Structural Database

- All published small molecule crystal structures
  - Currently 658,047
  - Comprehensive
- All curated
  - Corrected
  - Standardised
  - Enriched
- All searchable
  - ‘Chemistry’ added
  - Sophisticated tools

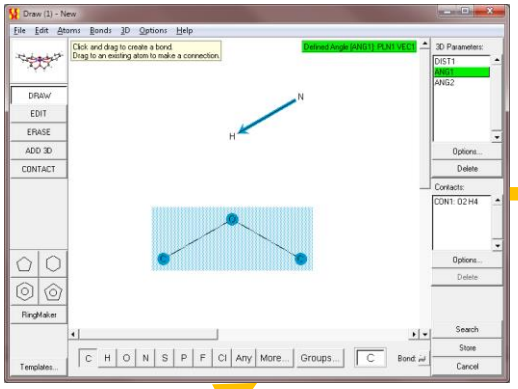


The Cambridge Structural Database System

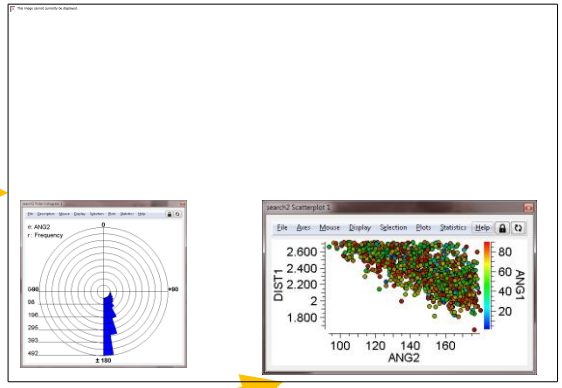
PreQuest: Create (in-house) database

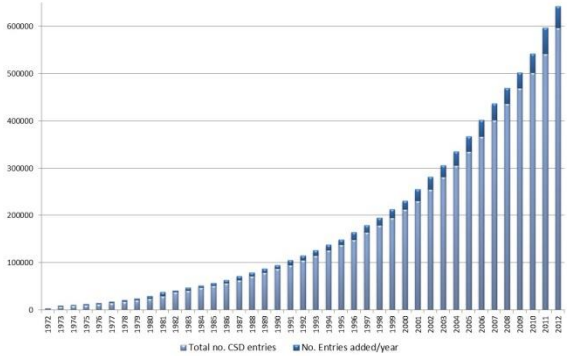
Cambridge Structural Database

ConQuest: Advanced 3D searching

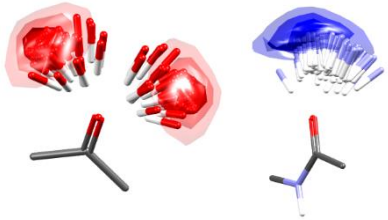


Mercury: Visualisation & data analysis

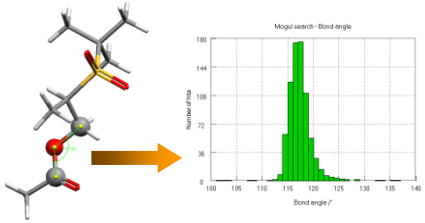




IsoStar: Molecular interaction analysis

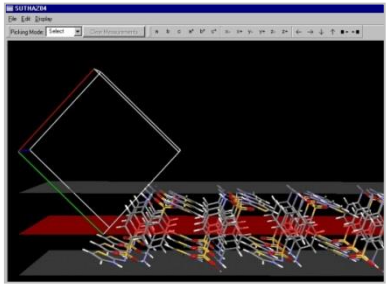


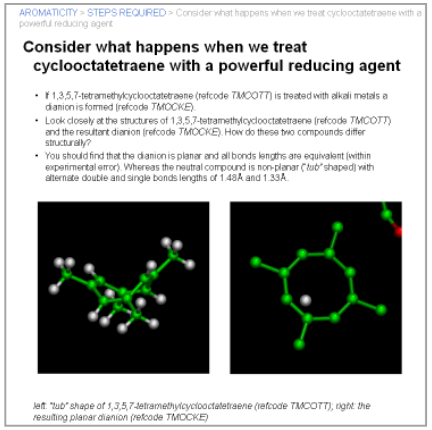
Mogul: Molecular geometry analysis



Services provided

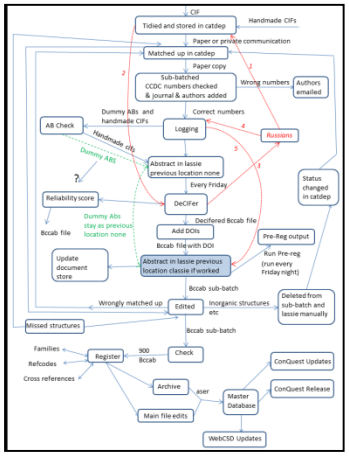
- Access to original deposited data
  - Individual data sets available to anyone
  - No cost archiving and dissemination of crystallographers output
    - c.f. Publication APCs
- Software tools
  - enCIFer (validation of CIFs)
  - Mercury (crystal structure visualisation)
  - Ligand dictionaries
- Targeted subsets of curated data
  - Teaching subsets
  - PDB subset
- CSD System



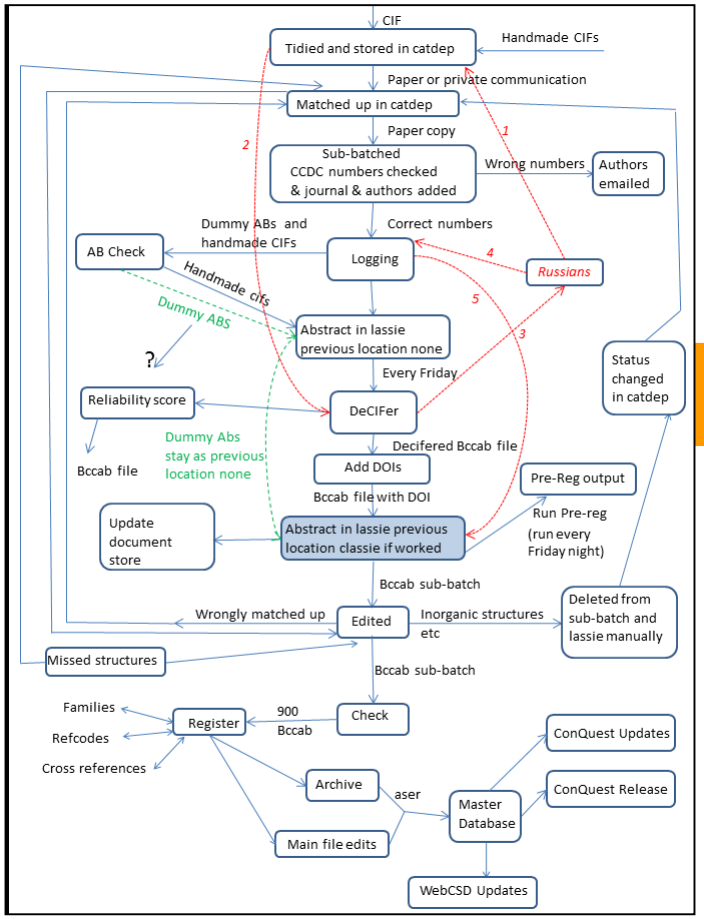


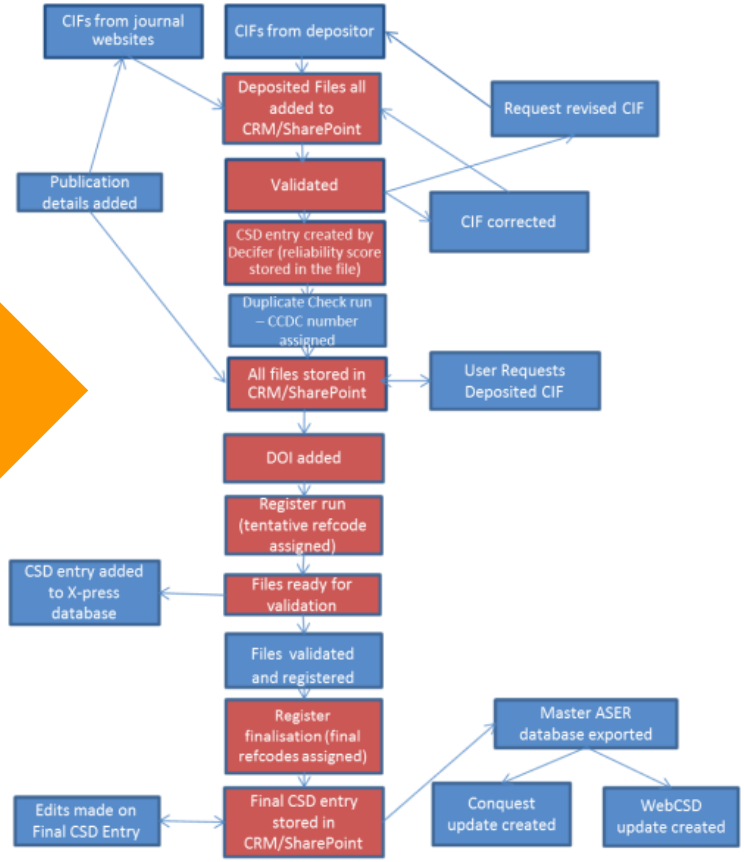
Creating the Cambridge Structural Database

- Structures from depositors and journals
  - Referees, embargoes, revision
- Curation
  - Coordinates to ‘chemistry’
  - Duplication
  - Enrichment
- Distribution and access
  - Searching, analysis, application

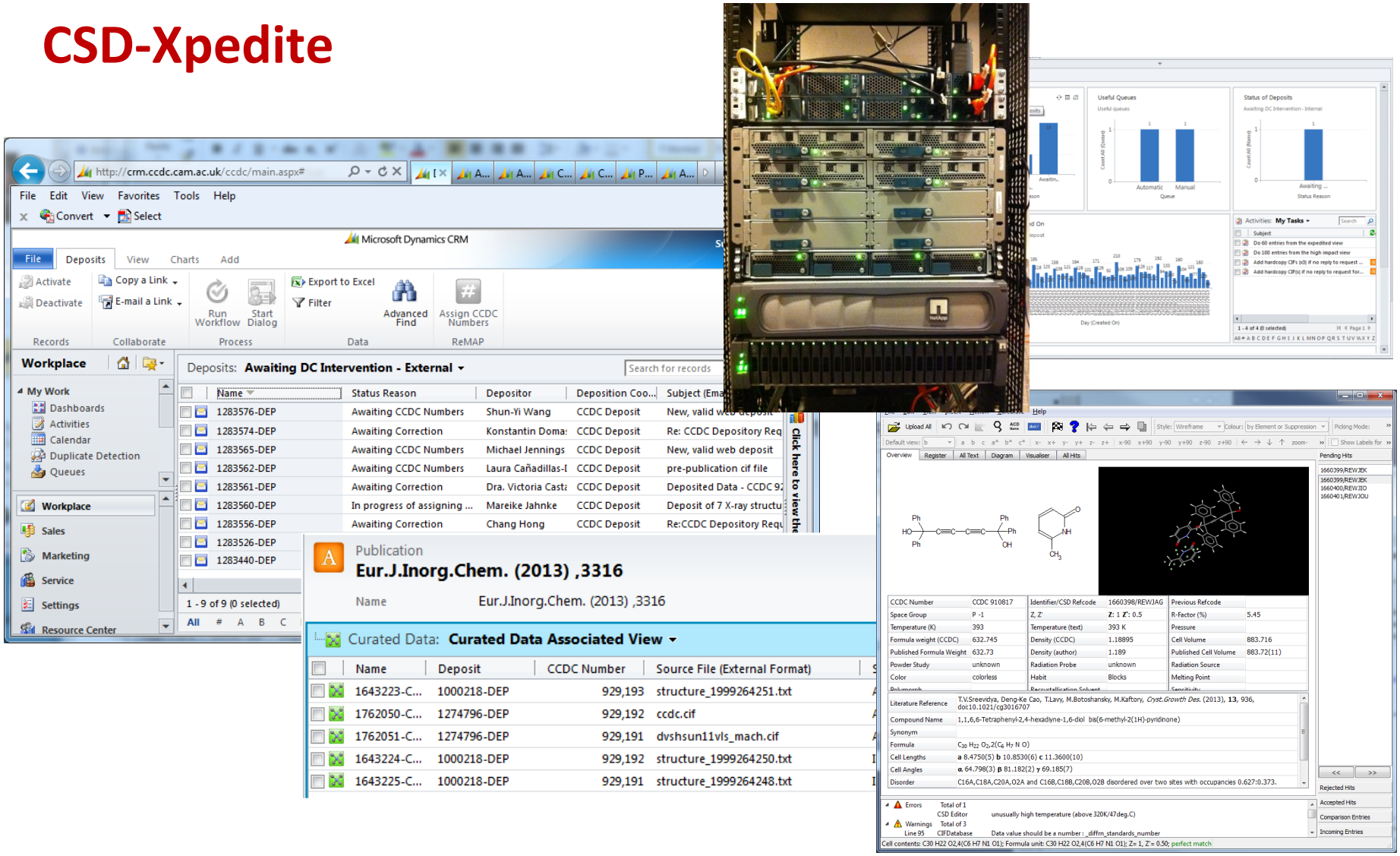


CSD-Xpedite


CIF
Tidied and stored in catdep
Handmade CIFs
Paper or private communication
Matched up in catdep
Paper copy
Sub-batched
CCDC numbers checked
& journal & authors added
Wrong numbers
Authors emailed
Dummy ABs and
handmade CIFs
Correct numbers
AB Check
Logging
Russians
Handmade CIFs
Dummy ABS
Abstract in lassie
previous location none
?
Every Friday
Status
changed
in catdep
Reliability score
DeCIFer
Dummy Abs
stay as previous
location none
Decifered Bccab file
Bccab file
Add DOIs
Pre-Reg output
Bccab file with DOI
Update
document
store
Abstract in lassie previous
location classie if worked
Run Pre-reg
(run every
Friday night)
Bccab sub-batch
Wrongly matched up
Edited
Inorganic structures
etc
Deleted from
sub-batch and
lassie manually
Missed structures
Bccab sub-batch
Families
Refcodes
Cross references
Register
900
Bccab
Check
ConQuest Updates
Archive
aser
Master
Database
ConQuest Release
Main file edits
WebCSD Updates
1
2
3
4
5


CIFs from journal
websites
CIFs from depositor
Deposited Files all
added to
CRM/SharePoint
Request revised CIF
Publication
details added
Validated
CIF corrected
CSD entry created by
Decifer (reliability score
stored in the file)
Duplicate Check run
– CCDC number
assigned
All files stored in
CRM/SharePoint
User Requests
Deposited CIF
DOI added
Register run
(tentative refcode
assigned)
CSD entry added
to X-press
database
Files ready for
validation
Files validated
and registered
Register
finalisation (final
refcodes assigned)
Master ASER
database exported
Edits made on
Final CSD Entry
Final CSD entry
stored in
CRM/SharePoint
Conquest
update created
WebCSD
update created

CSD-Xpedite

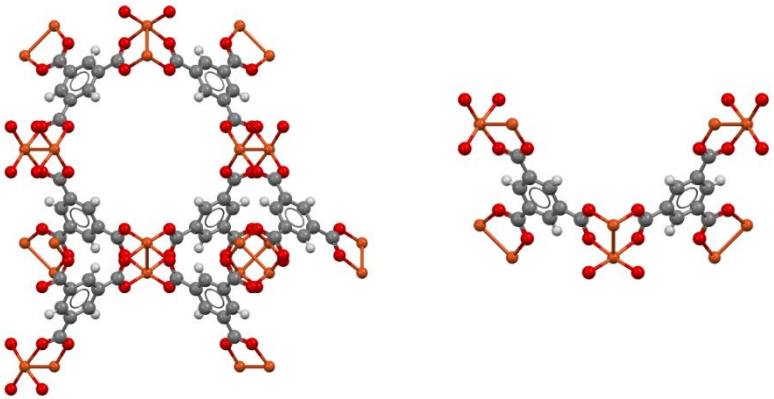

http://crm.ccdc.cam.ac.uk/ccdc/main.aspx#
File Edit View Favorites Tools Help
Convert Select
Microsoft Dynamics CRM
File Deposits View Charts Add
Activate
Deactivate
Copy a Link
E-mail a Link
Run Workflow
Start Dialog
Export to Excel
Filter
Advanced Find
Assign CCDC Numbers
Records
Collaborate
Process
Data
ReMAP
Workplace
My Work
Dashboards
Activities
Calendar
Duplicate Detection
Queues
Workplace
Sales
Marketing
Service
Settings
Resource Center
Deposits: Awaiting DC Intervention - External
Search for records
Name
Status Reason
Depositor
Deposition Coo...
Subject (Ema
1283576-DEP Awaiting CCDC Numbers Shun-Yi Wang CCDC Deposit
1283574-DEP Awaiting Correction Konstantin Doma: CCDC Deposit Re: CCDC Depository Req
1283565-DEP Awaiting CCDC Numbers Michael Jennings CCDC Deposit New, valid web deposit
1283562-DEP Awaiting CCDC Numbers Laura Cañadillas-I CCDC Deposit pre-publication cif file
1283561-DEP Awaiting Correction Dra. Victoria Cast: CCDC Deposit Deposited Data - CCDC 92
1283560-DEP In progress of assigning ... Mareike Jahnke CCDC Deposit Deposit of 7 X-ray structu
1283556-DEP Awaiting Correction Chang Hong CCDC Deposit Re:CCDC Depository Requ
1283526-DEP
1283440-DEP
1 - 9 of 9 (0 selected)
All # A B C
Click here to view the
Publication
Eur.J.Inorg.Chem. (2013) ,3316
Name Eur.J.Inorg.Chem. (2013) ,3316
Curated Data: Curated Data Associated View
Name Deposit CCDC Number Source File (External Format)
1643223-C... 1000218-DEP 929,193 structure_1999264251.txt
1762050-C... 1274796-DEP 929,192 ccdc.cif
1762051-C... 1274796-DEP 929,191 dvshsun11vls_mach.cif
1643224-C... 1000218-DEP 929,192 structure_1999264250.txt
1643225-C... 1000218-DEP 929,191 structure_1999264248.txt
Useful Queues
Automatic
Manual
Queue
Status of Deposits
Awaiting ...
Status Reason
Activities: My Tasks
Do 60 entries from the expedited view
Do 100 entries from the high impact view
Add hardcopy CIFs (x3) if no reply to request ...
Add hardcopy CIF(s) if no reply to request for...
Day (Created On)
Help
Upload All
Overview
Register
All Text
Diagram
Visualiser
All Hits
Pending Hits
1660399/REWJEK
1660399/REWJEK
1660400/REWJIO
1660401/REWJOU
CCDC Number CCDC 910817
Space Group P -1
Temperature (K) 393
Formula weight (CCDC) 632.745
Published Formula Weight 632.73
Powder Study unknown
Color colorless
Identifier/CSD Refcode 1660398/REWJAG
Z, Z' Z: 1 Z': 0.5
Temperature (text) 393 K
Density (CCDC) 1.18895
Density (author) 1.189
Radiation Probe unknown
Habit Blocks
Previous Refcode
R-Factor (%) 5.45
Pressure
Cell Volume 883.716
Published Cell Volume 883.72(11)
Radiation Source
Melting Point
Literature Reference T.V.Sreevidya, Deng-Ke Cao, T.Lavy, M.Botoshansky, M.Kaftory, Cryst.Growth Des. (2013), 13, 936, doi:10.1021/cg3016707
Compound Name 1,1,6,6-Tetraphenyl-2,4-hexadiyne-1,6-diol bis(6-methyl-2(1H)-pyridinone)
Synonym
Formula C30 H22 O2,2(C6 H7 N O)
Cell Lengths a 8.4750(5) b 10.8530(6) c 11.3600(10)
Cell Angles α 64.798(3) β 81.182(2) γ 69.185(7)
Disorder C16A,C18A,C20A,O2A and C16B,C18B,C20B,O2B disordered over two sites with occupancies 0.627:0.373.
Errors Total of 1
CSD Editor unusually high temperature (above 320K/47deg.C)
Warnings Total of 3
Line 95 CIFDatabase Data value should be a number : _diffrn_standards_number
Rejected Hits
Accepted Hits
Comparison Entries
Incoming Entries
Cell contents: C30 H22 O2,4(C6 H7 N1 O1); Formula unit: C30 H22 O2,4(C6 H7 N1 O1); Z= 1, Z'= 0.50; perfect match

CSD-Xpedite

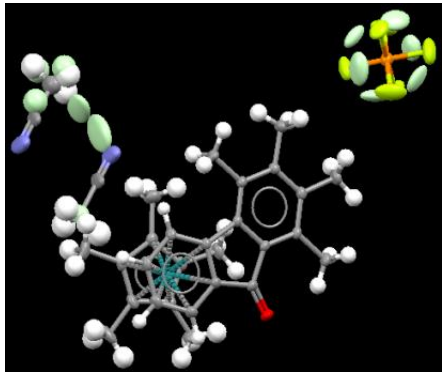
- CSD-Xpedite removes internal processing restrictions
  - Large structures
  - Structure Factors
    - Time to mandate?
- Easier access to deposited data
  - Just CCDC number or publication DOI needed
  - Immediate access
- Better structure comparison and duplicate checking
- Persistent identifiers and links from/to other resources
  - Automatic DOI population
- Faster ‘publication’
  - 10-25 days, down to <10 minutes
- Will allow community curation

Scientific benefits

- Better polymer representation



- ADPs and occupancy factors more visible



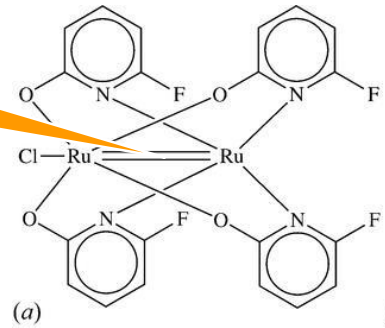
- Consistent interpretations

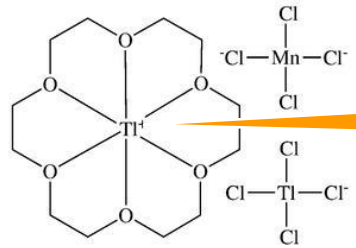
Curating The Cambridge Structural Database

- Curation
  - Coordinates to ‘chemistry’
  - Duplication
  - Families
  - Enrichment
- Distribution and access
  - Searching
  - Analysis
  - Application

Chemical structure from atomic coordinates

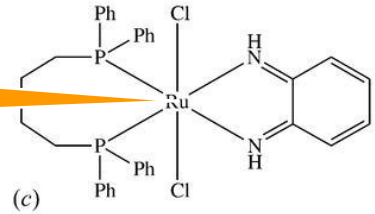
bond order?

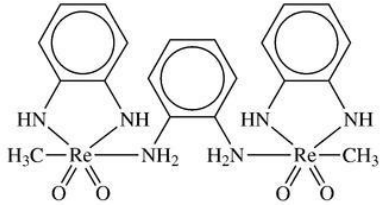




charges?

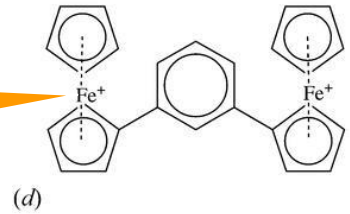
oxidation state?

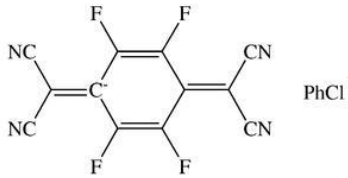




bond type?

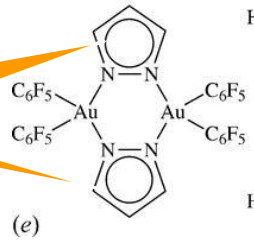
charges?

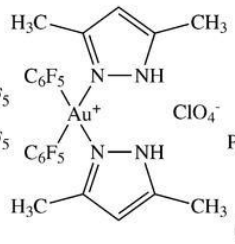




charges?

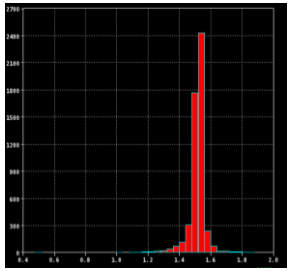
aromatic?



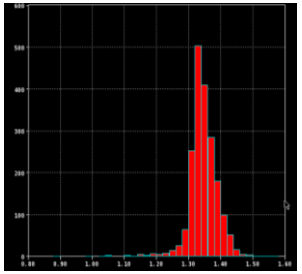


carbene coordinated?

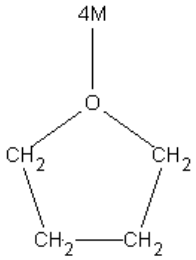
DeCIFer: Automatic Assignment of Chemistry



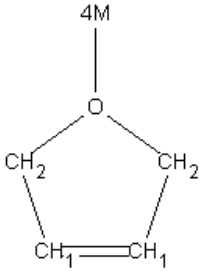
C-C, prob = 0.004



C=C, prob = 0.20



7208 hits



2 hits

Conflicting Evidence

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

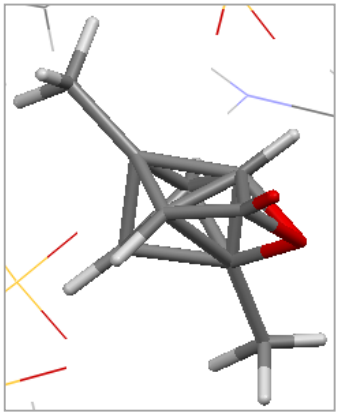
Bayes' Theorem

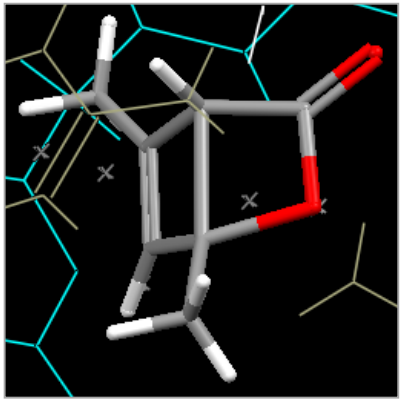
```
Low probability bond lengths:
C5-C6  1.405, av(CSD) = 1.505, prob = 0.001
C2-C3  1.345, av(CSD) = 1.514, prob = 0.001
C3-C4  1.338, av(CSD) = 1.514, prob = 0.001
C3-C6  1.798, av(CSD) = 1.546, prob = 0.001
Reliability level: 2
```

Chemical Assignment + Reliability Report

DeCIFer also attempts to automate resolving of disorder and generating diagrams and names

Avoiding duplication of effort





Under UV radiation the clathrated pyrone molecule converts to a disordered mixture of square-planar 1, 3-dimethylcyclobutadiene and rectangular-bent 1, 3-dimethylcyclobutadiene in van der Waals contact with a carbon dioxide molecule. The ratio of the square-planar to rectangular-bent 1, 3-dimethylcyclobutadiene clathrate is modelled with occupancies 0.6292:0.3708.

Unresolved disorder

Resolved disorder with editorial comment

Current challenges

- Increase ‘discoverability’ of crystal structure data
  - More links – databases, aggregators and publications
  - Improved access to deposited data
- More depositions
- Enriched depositions
  - Diagrams, structure factors
- Community curation

Improve access to deposited data sets

- The CSD and CSD System are ubiquitous in structural chemistry
- Anyone can access individual deposited data sets
- Access to these data sets must be embedded in other resources
  - Priority to remove technical barriers
  - Establish conditions of use consistent with modern age and desires of rights holders
    - By building on existing services that make data freely available to the community

Adoption outside small molecule communities

CANCER

Drug for an 'undruggable' protein

Scientists have long aimed to develop drugs against the cancer-associated protein KRAS, but without success. An approach that targets the oncoprotein's cellular localization reignites lost enthusiasm. SEE LETTER P.638

NICOLE M. BAKER & CHANNING J. DER

Human *RAS* genes have two claims to notoriety. First, they make up the most frequently mutated oncogene family in human cancer, having a prevalence of one in every three cases[1]. Second, despite more than three decades of intensive effort, no effective pharmacological inhibitor of the RAS oncoprotein has reached the clinic. So it is exciting that, on page 638 of this issue, Zimmermann *et al.*[2] report* the identification and characterization of a small-molecule inhibitor that interferes with the localization of KRAS — the RAS isoform most commonly mutated in human cancers — to the plasma membrane surrounding cells[3].

Following their synthesis in the cytoplasm, RAS proteins are initially inactive[4]. They then undergo a series of rapid post-translational modifications that ensure their association with the inner leaflet of the plasma membrane, where these proteins exert their normal, as well as their cancer-associated, signalling activity. Therefore, most efforts aimed at anti-RAS drug discovery have involved indirect approaches to block the activities of proteins that either promote plasma-membrane association of RAS or are components of its downstream signalling pathway.

The key post-translational modification of RAS involves the addition of a 15-carbon farnesyl lipid tail in a reaction catalysed by the farnesyltransferase enzyme. This modification facilitates RAS association with membranes and is essential for proper RAS localization and activity, having prompted intensive efforts in the 1990s to develop farnesyltransferase inhibitors (FTIs).

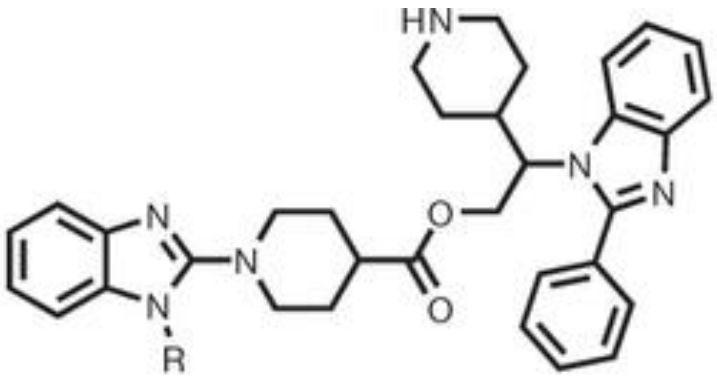
Despite promising results in preclinical studies, however, the results of clinical trials with FTIs were disappointing. The inhibitors blocked membrane association of the HRAS isoform, but lacked antitumour activity in cancers involving mutated KRAS (and NRAS). KRAS could still associate with the plasma membrane through an unexpected compensatory activity of the farnesyltransferase-related enzyme geranylgeranyltransferase-I, which modifies RAS with a geranylgeranyl, rather than a farnesyl, group. This discouraging outcome greatly dampened interest in targeting RAS — and, in particular, its membrane association — for cancer treatment. Instead, ongoing efforts have mainly focused on inhibitors of the RAF–MEK–ERK and the PI3K-AKT signalling cascades downstream of RAS.

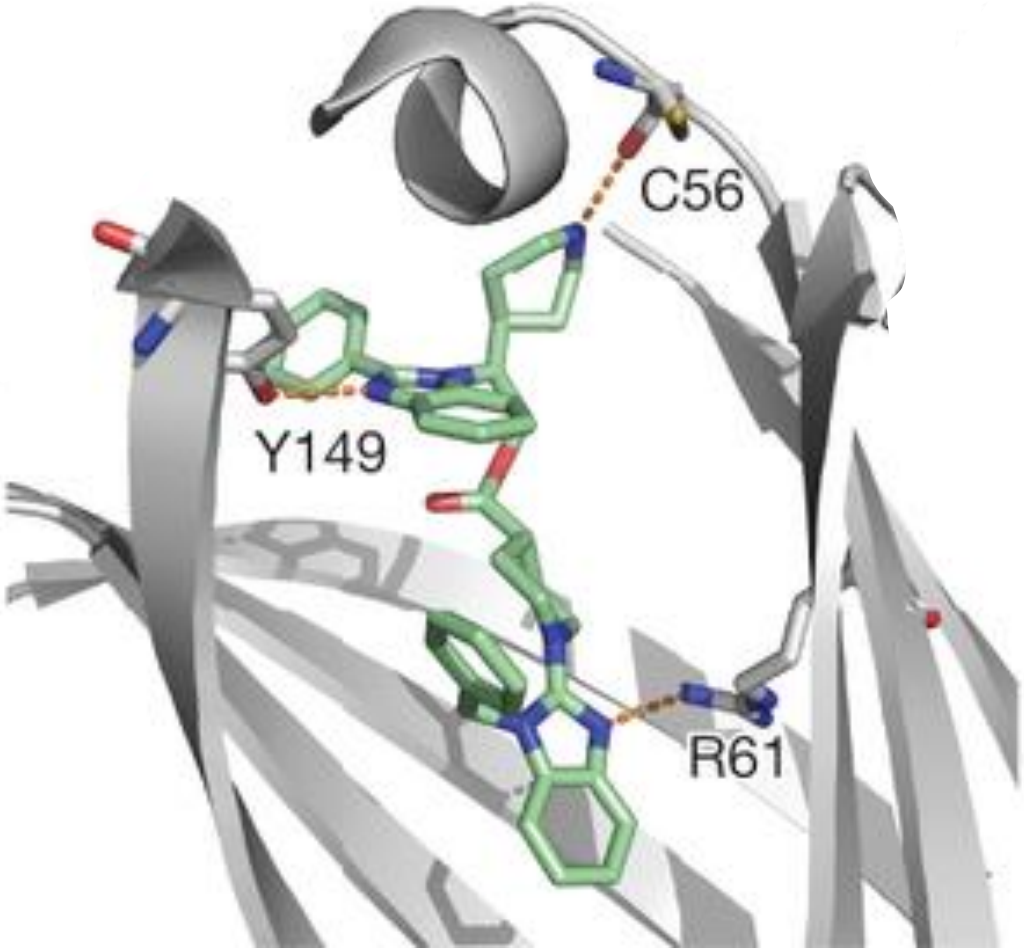
Zimmermann *et al.* describe an approach aimed at disrupting KRAS membrane association that warrants reassessment of the current strategies. The authors identify and characterize a small-molecule inhibitor of PDEδ, a protein that can bind to and regulate the trafficking of RAS and RAS-related proteins to membrane compartments[5–8] (Box 1). Specifically, PDEδ contains a deep, hydrophobic pocket capable of binding the lipid moiety of farnesylated proteins, in particular RAS.

An earlier study[8] found that suppression of PDEδ levels disrupts RAS association with the plasma membrane and impairs the growth of RAS-mutant cancer cells. This finding

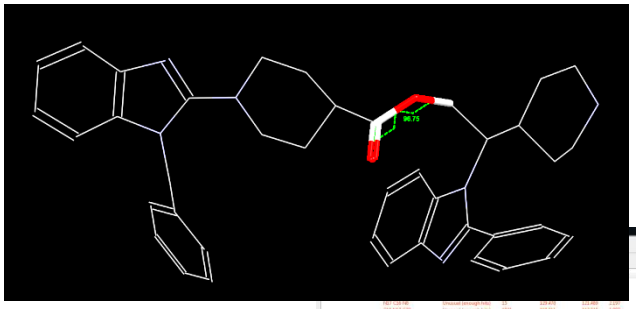
*This article and the paper under discussion[2] were published online on 22 May 2013.

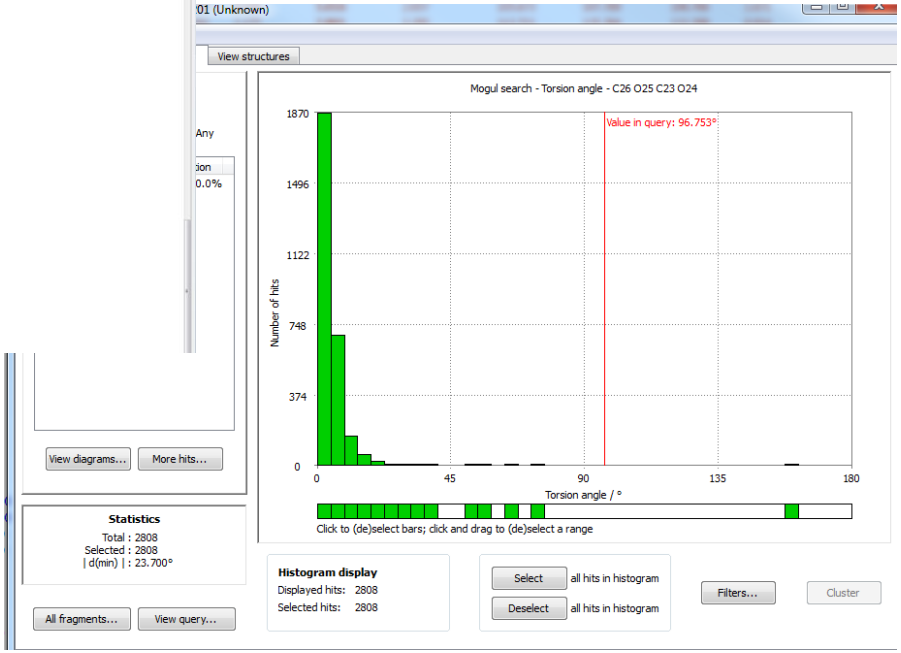
Adoption outside small molecule communities





Adoption outside small molecule communities





Adoption outside small molecule communities

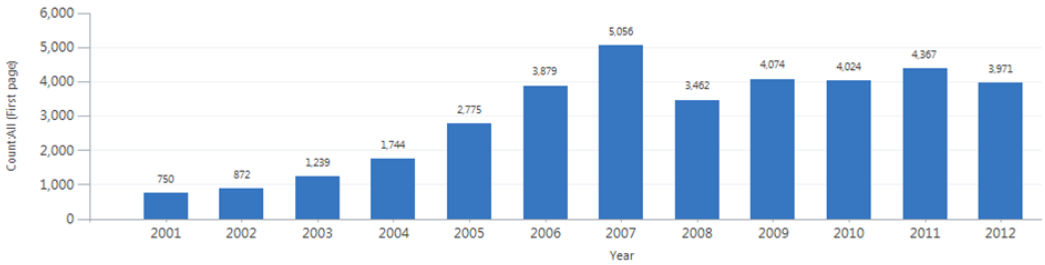
- Structural biology community still struggling with chemistry
- A failure by the small molecule community, including the CCDC to communicate value
  - Some success through collaboration
    - PDB
    - Global Phasing
    - COOT
    - CCP4
- Do current business models contribute to this ‘bad science’?
  - The need for academic contributions restricts sharing
  - Lack of clarity of ‘permissions’

Current challenges

- Increase ‘discoverability’ of crystal structure data
  - More links – databases, aggregators and publications
  - Improved access to deposited data
- More depositions
- Enriched depositions
  - Diagrams, structure factors
- Community curation

Structure publishing

- Pre 2000 CCDC and IUCr recognised many valuable structures were never published
  - IUCr launched *Acta Crystallographica Section* E Structure Reports Online
  - 350 (extra?) structures per month



- Recent reduction in structures from *Acta E*
  - No sign yet that they are going elsewhere
- CCDC will assign DOIs to entries
- CSD to be listed in Thomson Reuters Data Citation Index
- What else should we do to encourage ‘publication’

Current challenges

- Increase ‘discoverability’ of crystal structure data
  - More links – databases, aggregators and publications
  - Improved access to deposited data
- More depositions
- Enriched depositions
  - Diagrams, structure factors
- Community curation

Difficulties with non-semantic data

- CSD Editors still extract data from PDFs
  - Some automatic, some manual
- Utterly stupid way of working
  - Semantic data -> non-semantic, obfuscated data -> semantic data
- Heroic efforts to extract data from PDFs
  - Is this the right place to expend effort?
  - All data could be semantic, especially as supplementary

Current challenges

- Increase ‘discoverability’ of crystal structure data
  - More links – databases, aggregators and publications
  - Improved access to deposited data
- More depositions
- Enriched depositions
  - Diagrams, structure factors
- Community curation

Unanswered Questions

- Time to mandate processed data deposition?
  - Who decides?
- Is free access to individual CIFs enough?
- Should the community continue to fund the CSD at point of use
  - Consumer pays
  - This requires access and sharing restrictions
- Should research councils ‘pay’
  - Currently the model for many countries
    - Eases move to Open Data, raises questions of sustainability
  - Desperately confused situation in the UK
- Could the entire crystallography community adopt a single *modus operandi*
  - Only Open Access publishing
  - All (processed) data available (semantic form)
  - APCs for papers
  - APCs for data

Acknowledgements

- Ian Bruno – Strategic Relationships Manager
- Suzanna Ward - CSD Group Manager
- Matthew Lightfoot – Editor in Chief of the CSD
- CCDC staff
- The 253,000 authors of structure containing publications