The Third Dimension

Crystallographic Data and its Application in Scientific Research and Development

Ian Bruno, Cambridge Crystallographic Data Centre, UK

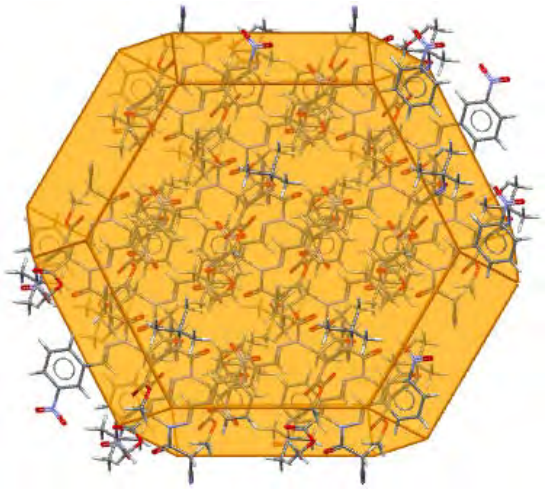
@ccdc_cambridge

New Developments in Chemical Information: 'Best Practice'

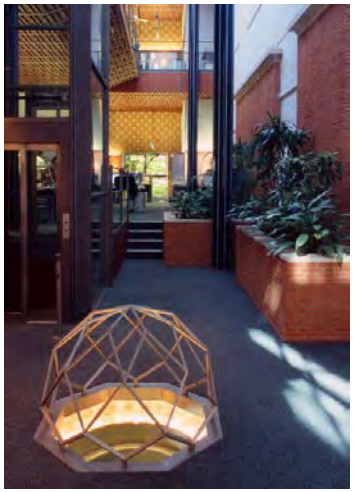
RSC Chemical Information and Computer Applications Group, 3 July 2013

Outline

- Crystal Structure Data
  - why is it important
  - current challenges



- The Cambridge Crystallographic Data Centre
  - current services and applications
  - recent developments



What is the CCDC?

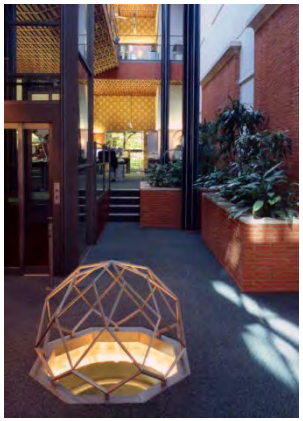
International Data Repository
Archive of crystal structure data
Deposition services

Scientific Software Provider
Search/analysis/visualisation tools
Scientific applications

Collaborative Research Organisation
New methodologies
Fundamental research

Employer of around 45 permanent staff

Scientific editors
Software developers
Applications scientists



An independent not-for-profit organisation established in 1965.

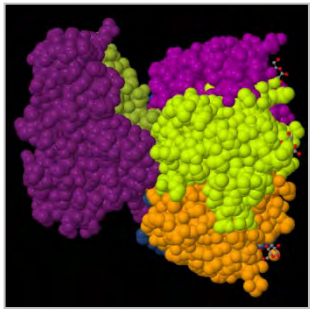
Financially supported by income generated from subscriptions to value-added services.

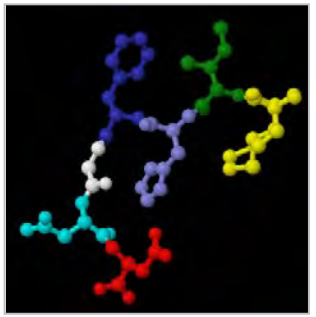
Crystallographic Databases

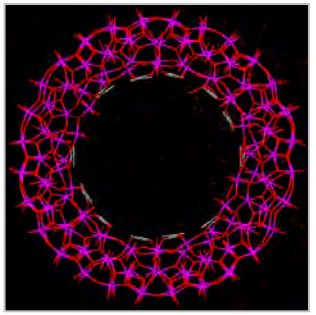
- Biological macromolecules
  - Protein Data Bank (PDB)
  - grant-funded, 16 or so agencies worldwide

- **Organic and metal-organic structures**
  - **Cambridge Structural Database (CSD)**
  - self-supporting, not-for-profit, registered charity

- Inorganic structures
  - ICSD: partnership between FIZ Karlsruhe and NIST
  - CRYSTMET: privately owned (Toth Information Systems)







The Cambridge Structural Database

www.ccdc.cam.ac.uk/services/structure_deposit

CCDC CIF deposition and validation service (v. 1.1)

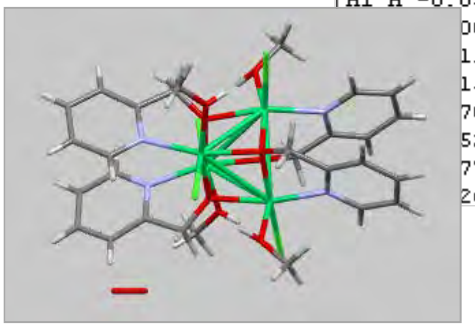
This web service enables submission of CIF files and associated structure factor files for inclusion of structures in the Cambridge Structural Database. Submitting a CIF file via this page will allow the discovery and correction of any problems with the syntax of the file(s).

- Files should be in **CIF, FCF or HKL** format
- At least one CIF file must be included in the submission
- Up to **10 files** may be submitte
- Each file may contain **multiple**
- Where structures relate to a p **publication only**

Data files may also be sent by e-ma

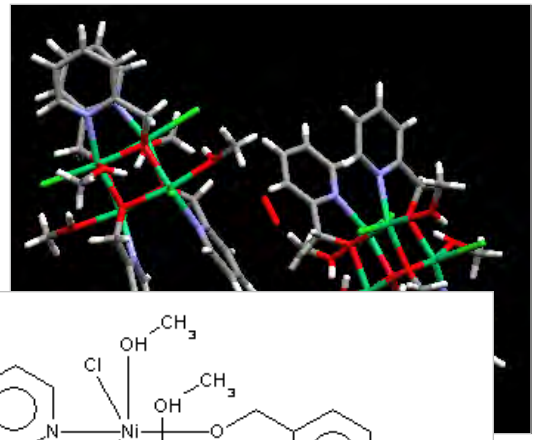
Data are accepted either in conjun **communications** to the Cambridg publication will be held securely in t publishers will be made available to

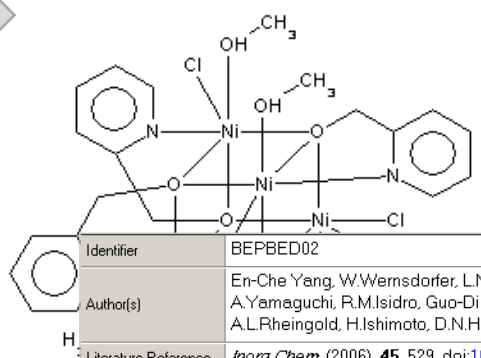
```
loop_
_atom_site_label
_atom_site_type_symbol
_atom_site_fract_x
_atom_site_fract_y
_atom_site_fract_z
C1  C  0.31594(37)  0.75375(20)  0.70189(16)
C2  C  0.19196(41)  0.77066(23)  0.76436(17)
O1  O  0.22259(32)  0.80293(21)  0.83459(13)
C3  C  0.03051(40)  0.74837(26)  0.73926(18)
H1  H -0.05699(40)  0.75910(26)  0.77720(18)
     00500(38)  0.71281(22)  0.66463(17)
     11609(38)  0.69941(22)  0.65036(17)
     1523(28)  0.69499(17)  0.60921(13)
     7057(33)  0.71613(20)  0.62788(16)
     5291(33)  0.70477(20)  0.58724(16)
     7764(37)  0.65275(21)  0.52604(17)
     2644(37)  0.69329(21)  0.48496(17)
```



Deposited CIF

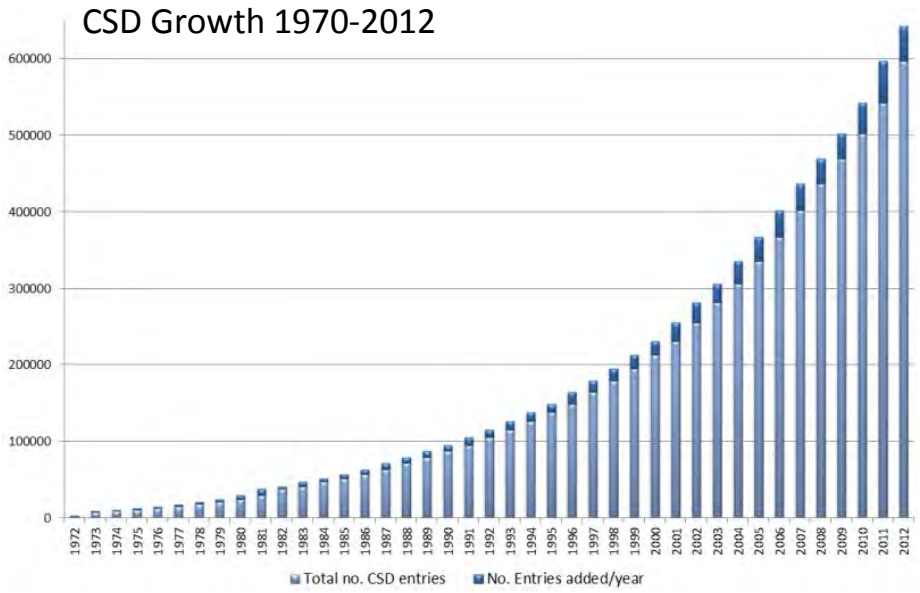
CSD Entry





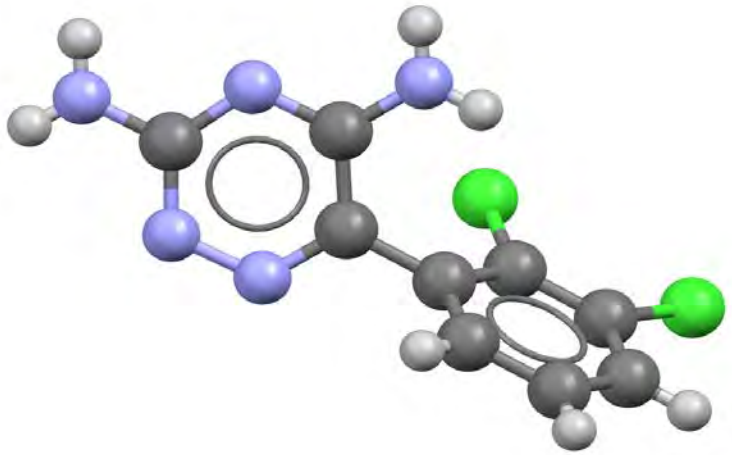
Identifier	BEPBED02
Author(s)	En-Che Yang, W.Wernsdorfer, L.N.Zakharov, Y.Karaki, A.Yamaguchi, R.M.Isidro, Guo-Di Lu, S.A.Wilson, A.L.Rheingold, H.Ishimoto, D.N.Hendrickson
Literature Reference	*Inorg.Chem.* (2006), **45**, 529, doi:10.1021/ic050093r
Formula	$C_{28}H_{40}Cl_4N_4Ni_4O_8H_2O$
Compound Name	tetrakis(μ3-2-(Oxymethyl)pyridine-N,O,O,O)-tetrachloro-tetraki s(methanol)-tetra-nickel(ii) monohydrate
Synonym	
Space Group	I -4 2 d
Cell Lengths	**a** 16.1421(6) **b** 16.1421(6) **c** 29.4689(14)
Cell Angles	**α** 90 **β** 90 **γ** 90
R-Factor (%)	5.21
Disorder	The water molecule is disordered by symmetry.

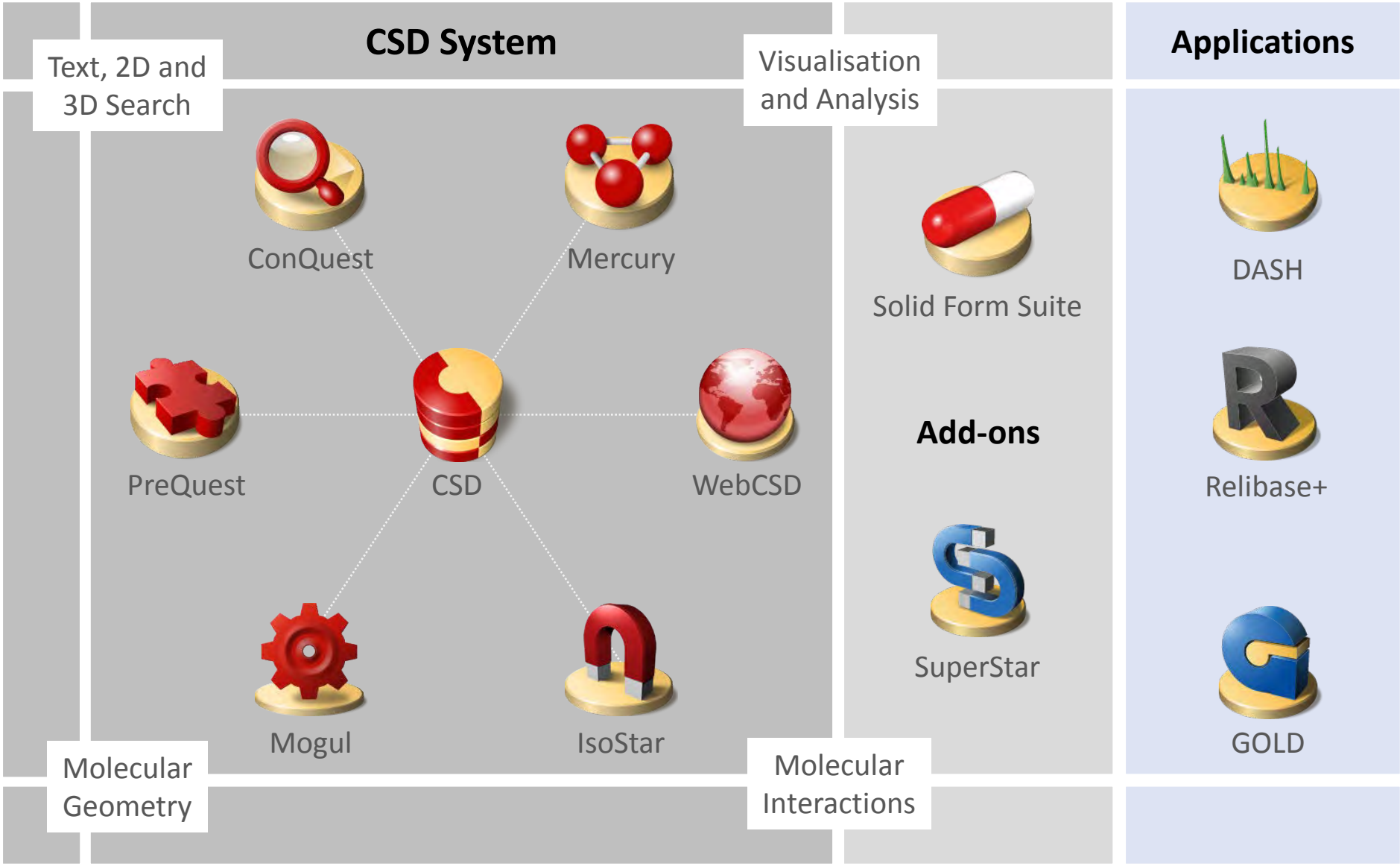
The Cambridge Structural Database



- Dec 2009: 500,000 milestone reached
- June 2013: 673,954 entries

- Lamotrigine
  Acta Cryst., Sect.C:Cryst Struct. Commun. (2009), 65, o460
  Refcode: EFEMUX01

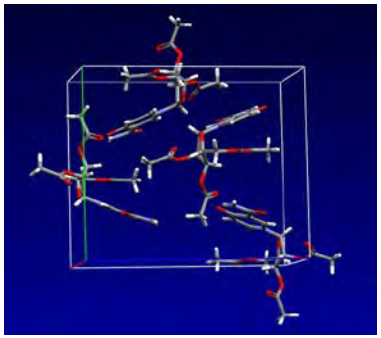



CSD System
Text, 2D and 3D Search
Visualisation and Analysis
Applications
ConQuest
Mercury
Solid Form Suite
DASH
PreQuest
CSD
WebCSD
Add-ons
Relibase+
SuperStar
Mogul
IsoStar
GOLD
Molecular Geometry
Molecular Interactions

Applications of the CSD

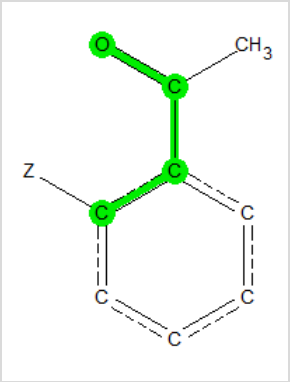
- The CSD provides insights from experimental data
  - molecular dimensions and shape
  - molecular interactions

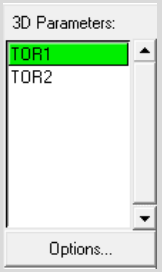
- Widely used for
  - drug design and development
  - design of new materials
  - crystal engineering
  - structure validation



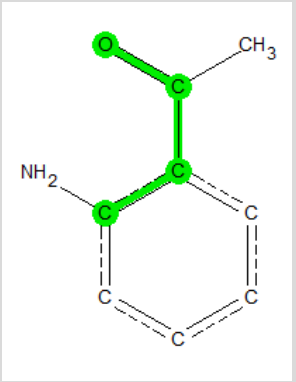
Assessing Molecular Geometry

ConQuest: 3D Substructure Search

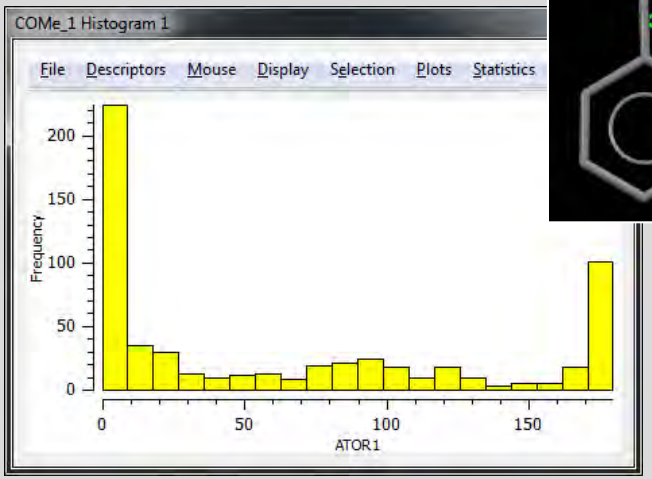




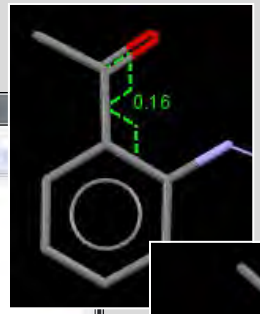
Z = Not H



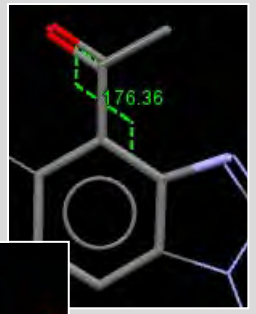
Mercury: Data Analysis and Visualisation

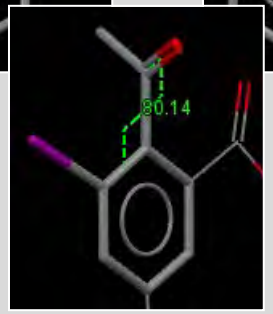


HISGAS01



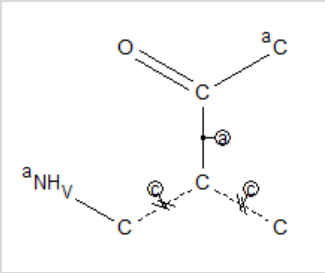
LAHQES





RATFOI

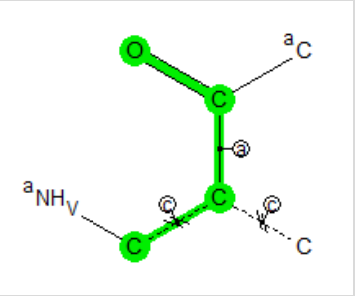
ConQuest: Substructure Query Features



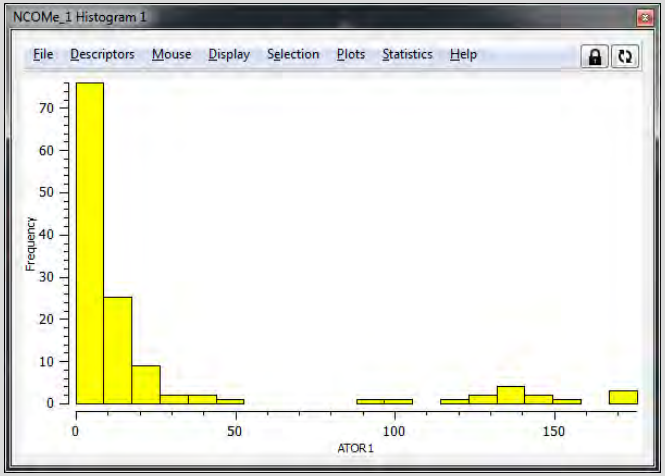
- Variable bond types (e.g. single, double, aromatic)
- Variable number of H (e.g. 1,2)
- Cyclic/Acyclic atoms and bonds

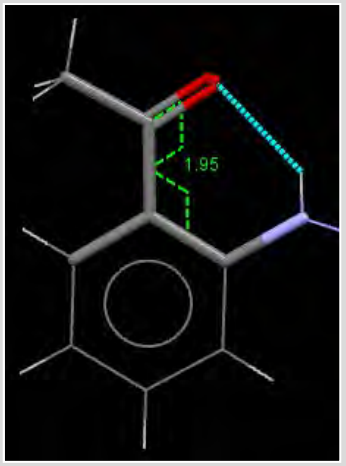
Assessing Molecular Geometry

ConQuest: 3D Substructure Search



Mercury: Data Analysis and Visualisation



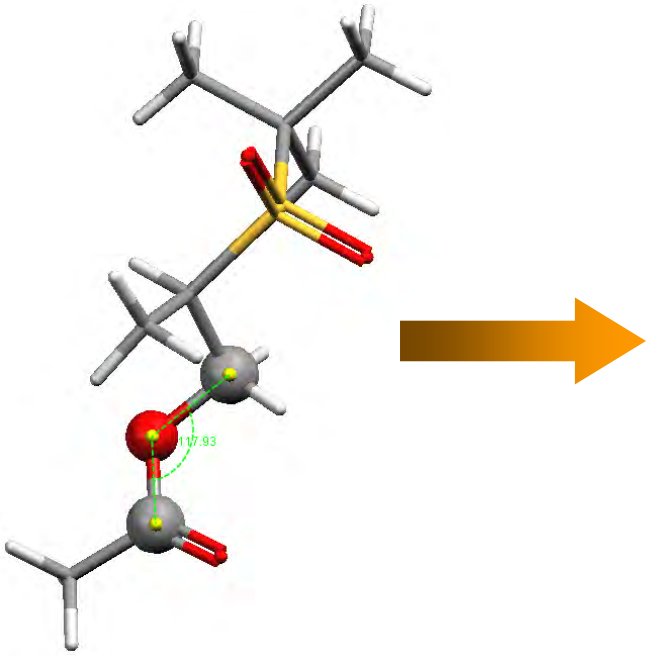


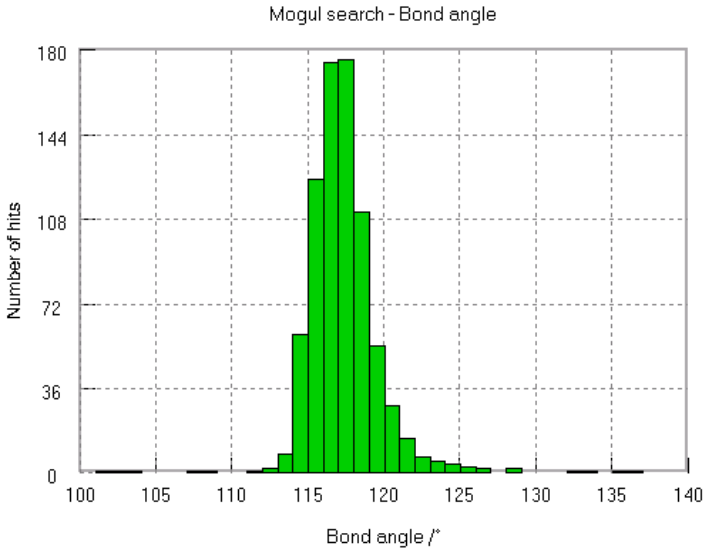
DUDHIU

- Substructure-based analysis offers flexibility and precision
- Can take several iterations to define the right query

Mogul

A Knowledge Base of Molecular Geometries



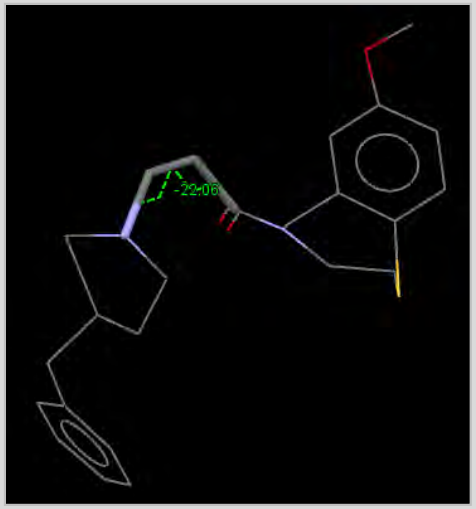


Mogul Overview

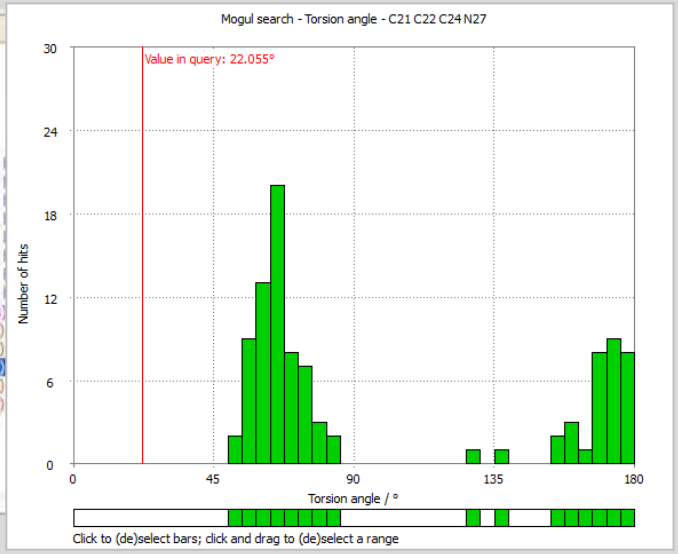
- **Pre-computed libraries:** bond lengths, valence angles, torsion angles and ring conformations derived entirely from the CSD
- **Validates complete geometry:** retrieve distributions, and figures of merit for all fragments in the molecule
- **Fragment Generalisation:** if the fragment specified is rare, Mogul will include CSD results from the most similar fragments
- **Hyperlinking to the CSD:** view CSD entries in specific areas of a distribution
- **Integration with other software:** Instruction file (and now Python API) enables other programs to interact with Mogul

Bruno et al., J. Chem. Inf. Comput. Sci., 44, 2133-2144, 2004

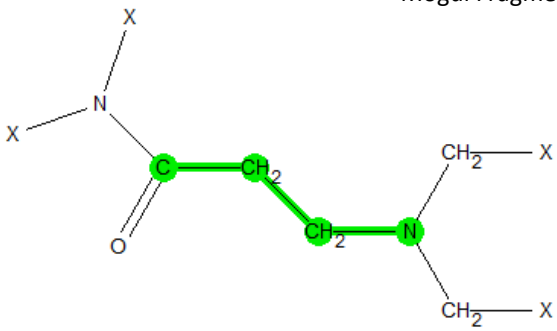
Mogul Torsion Analysis



Type	Molecule	Fragment	Classification
bond			
	LIM__K21_901-B_pdb1hak_1		
angle			
	LIM__K21_901-B_pdb1hak_1		
torsion			
	LIM__K21_901-B_pdb1hak_1		
		C59 O58 C1 C2	Not unusual (enough
		C59 O58 C1 C6	Not unusual (enough
		O23 C21 N11 C12	Not unusual (enough
		C24 C22 C21 N11	Not unusual (enough
		C22 C24 N27 C30	Not unusual (enough
		C22 C24 N27 C34	Not unusual (enough
		C45 C52 C43 C32	Not unusual (enough
		C51 C52 C43 C32	Not unusual (enough
		C22 C21 N11 C12	Not unusual (few hits)
		O23 C21 N11 C10	Unusual (enough hits)
		O23 C21 C22 C24	Unusual (enough hits)
		C21 C22 C24 N27	Unusual (enough hits)
		C31 C32 C43 C52	Unusual (enough hits)
		C33 C32 C43 C52	Unusual (enough hits)
		C22 C21 N11 C10	Unusual (few hits)



Mogul Fragment



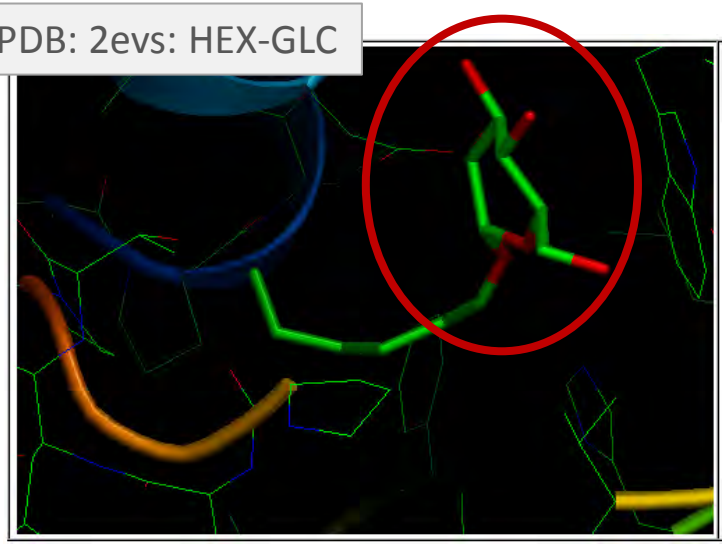
Relevance	Number	Contribution
1.00	4	12.5%
0.91	28	87.5%
	21	65.6%
	4	12.5%
	3	9.4%

If there are insufficient instances of the exact fragment, Mogul will identify related ones and rank them by relevance.

Relevance is a measure of similarity based on chemical features considered to have most impact on geometry.

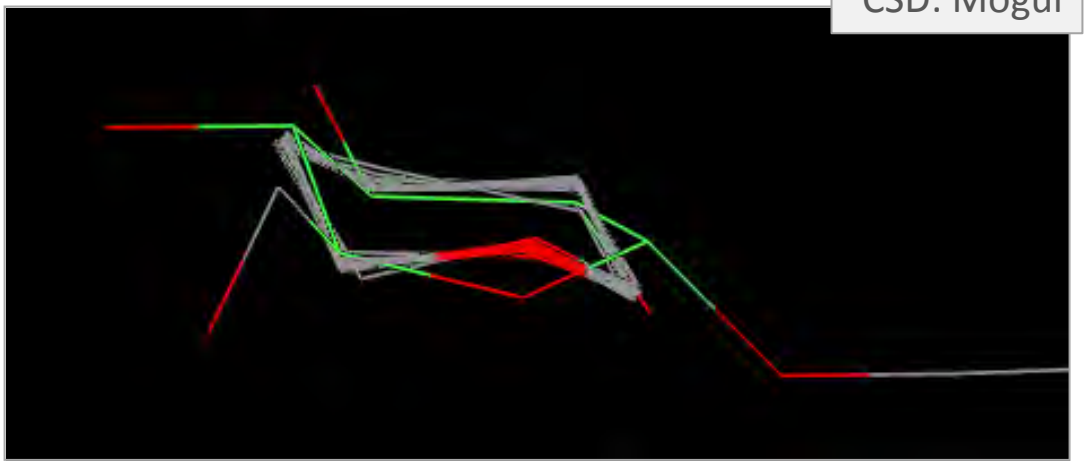
Mogul Ring Analysis

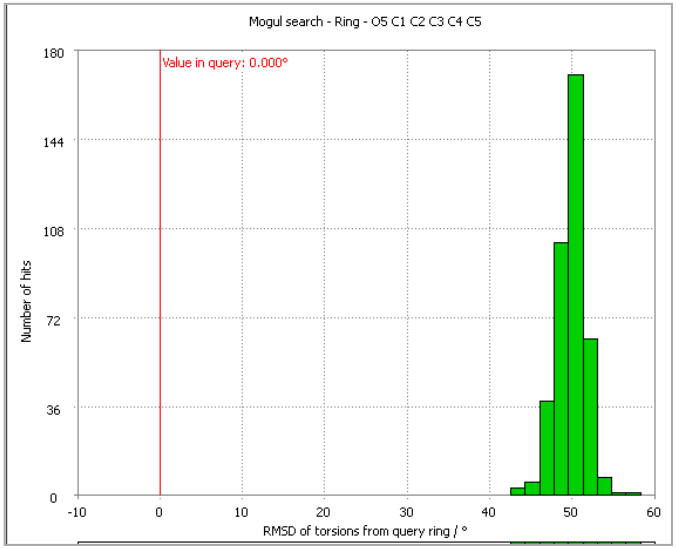
PDB: 2evs: HEX-GLC



- angle
  - LIM__HEX...
    - C4 C3 C2 — Not unusual (enough hits) 926
    - C3 C4 C5 — Not unusual (enough hits) 828
    - O5 C1 C2 — Unusual (enough hits) 3421
    - C1 C2 C3 — Unusual (enough hits) 3296
    - O5 C5 C4 — Unusual (enough hits) 891
    - C1 O5 C5 — Unusual (enough hits) 5171
- torsion
  - LIM__HEX...
- ring
  - LIM__HEX...
    - O5 C1 C2 C3 C4 C5 — Unusual (enough hits) 390

CSD: Mogul





PDB & CCDC Collaboration

September 2011

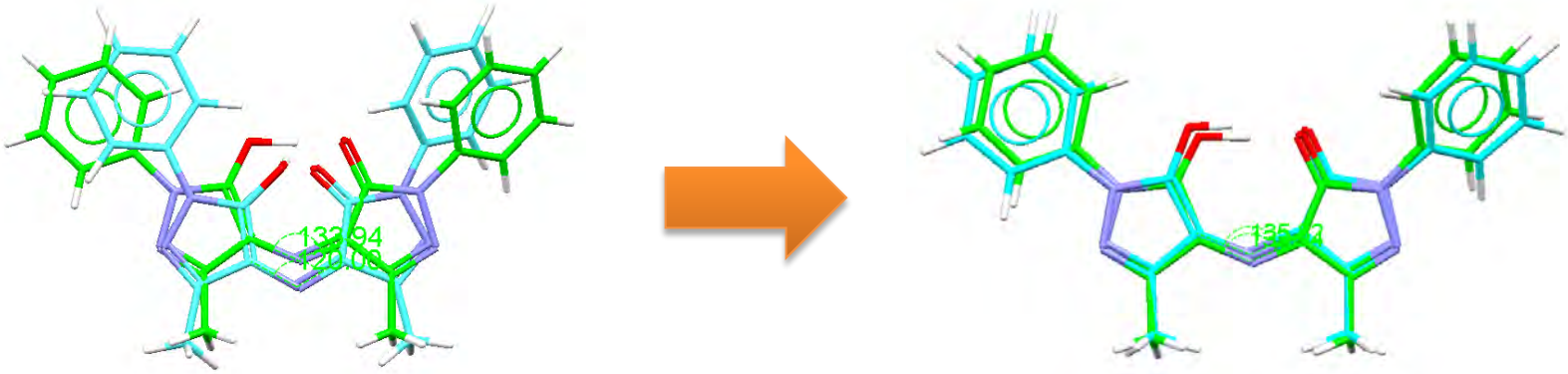
- Knowledge transfer
- Software exchange

  Collaborative project to use CSD data to provide quality measures of ligand geometries deposited to the PDB
- Data exchange

  Make public structures that match CSD ligands in the PDB

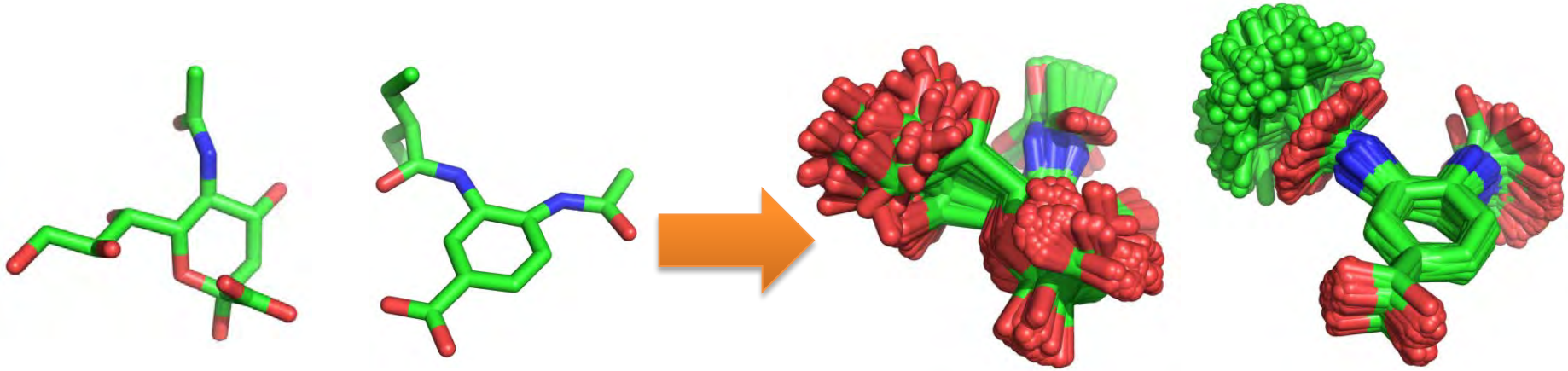
WORLDWIDE PDB PROTEIN DATA BANK

Molecule Minimisation



Current work at CCDC that exploits Mogul geometry distributions.

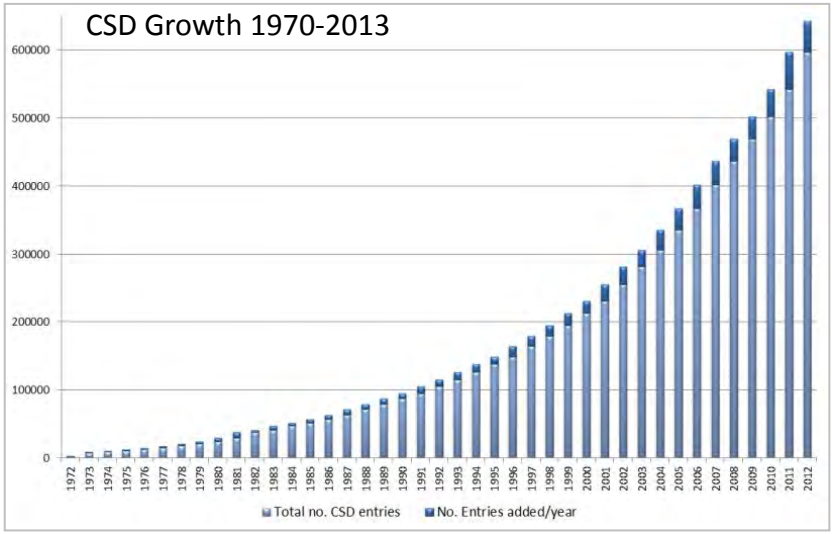
Conformer Generation



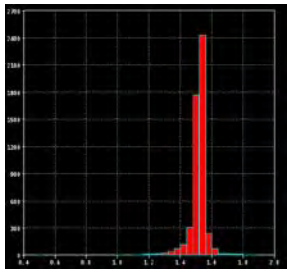
Internal Challenges

CCDC Challenges

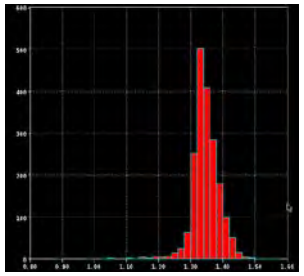
- Throughput is increasing
- Complexity/diversity is increasing
- Issues faced with deposited data
  - disorder
  - poor geometry
  - polymeric structures
  - incomplete chemical representation



Decifer: Automatic Assignment of Chemistry



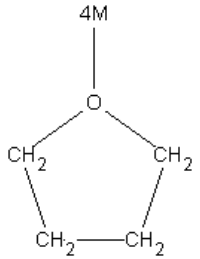
C-C, prob = 0.004



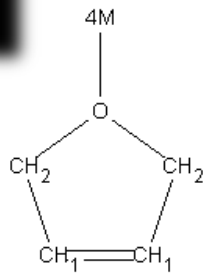
C=C, prob = 0.20

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

Bayes' Theorem



7208 hits



2 hits

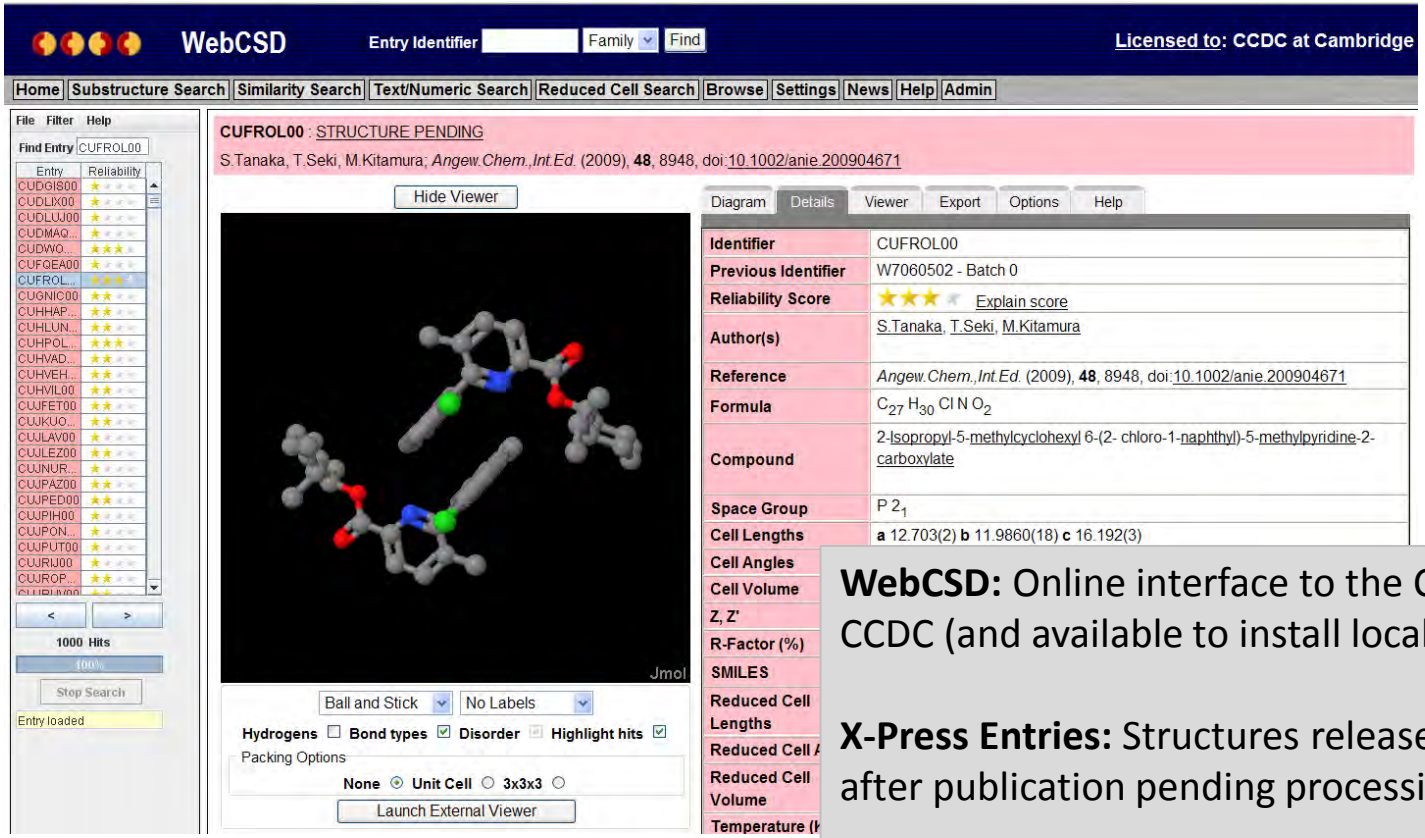
Conflicting Evidence

```
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

CSD X-Press



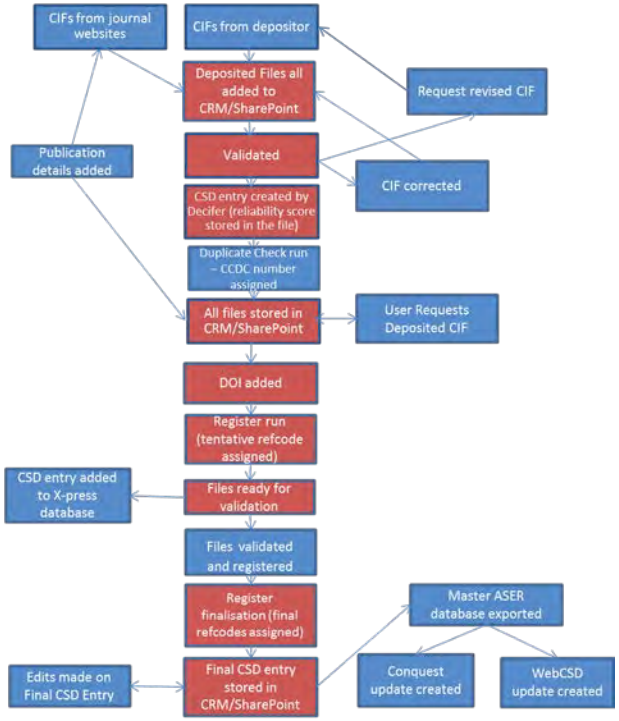
WebCSD: Online interface to the CSD hosted at CCDC (and available to install locally)

X-Press Entries: Structures released immediately after publication pending processing by experts

Reliability Score: Reflects the number of problems identified by automatic processing

CSD X-Pedite

- New internal system for processing entries deployed April 2013
- Internal benefits
  - replaces legacy file formats and software
  - improved workflows and scientific tools
- Future benefits for the scientific community
  - improved deposition processes
  - faster release of entries
  - better representation of the underlying data
  - improved integration with third-party resources



Links from Third Party Resources

RSC | Advancing the Chemical Sciences

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

Footnote

† Electronic supplementary information (ESI) available: Experimental procedures and spe
615536. For ESI and crystallographic data in CIF or other electronic format see DOI: 10.10

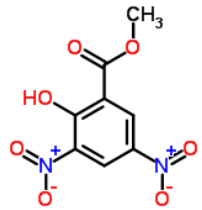
This journal is © The Royal Society of Chemistry 2009

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]

ChemSpider
The free chemical database



Cell 2D 3D Save Zoom

methyl 2-hydroxy-3,5-dinitrobenzoate

ChemSpider ID: **81036**
Molecular Formula: $C_8H_6N_2O_7$
Average mass: 242.142395 Da
Monoisotopic mass: 242.017502 Da
Systematic name
Methyl 2-hydroxy-3,5-dinitrobenzoate
SMILES and InChIs
Cite this record

Data Sources

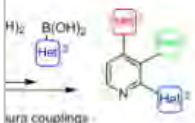
All Data Sources | Biological Data | Publishers | Phys. Properties | Theor. Data | Tox/Envir. Data | Per
Spectral Data | Data Aggregators

Data Source	External ID(s)
Cambridge Crystallographic Data Centre	ABENAX

ELSEVIER

vember 2009. Available online 17 November 2009.

line from 2-chloropyridine is described via a Directed
chanism in 26–28% yields. By performing sequential,
a variety of functionalised heteroaryl and arylboronic
ne scaffolds have been accessed in synthetically
atives. 2-Chloro-4-heteroaryl-3-iodopyridines and
orted. The synthesis of 5-[3,4-bis(2-
a two-step Sonogashira/Suzuki–Miyaura reaction
nylacetylene and 6-fluoropyridin-3-yl-3-boronic acid



ura couplings

Synthesis
- Divergent synthesis of arylated pyridin-2(1H)-one deriv... *Tetrahedron*
- Combined directed ortho metalation-halogen dance (HD) s... *Organic Letters*

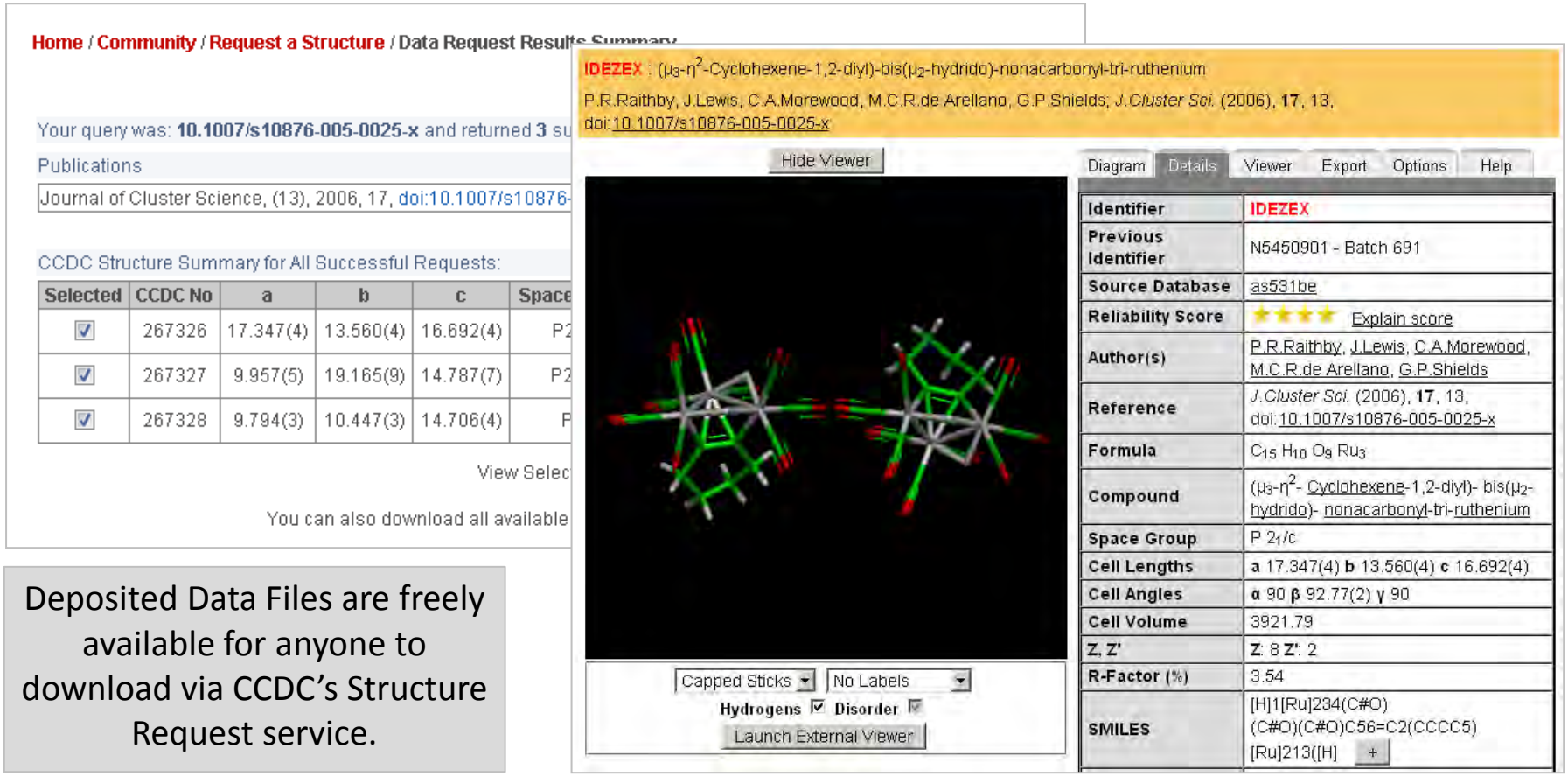
View details of all 3 citing articles in Scopus

Crystallographic Data

View Record in Scopus

Deposited Data

www.ccdc.cam.ac.uk/services/structure_request



Deposited Data Files are freely available for anyone to download via CCDC's Structure Request service.

CSD System subscribers can link through to WebCSD entries

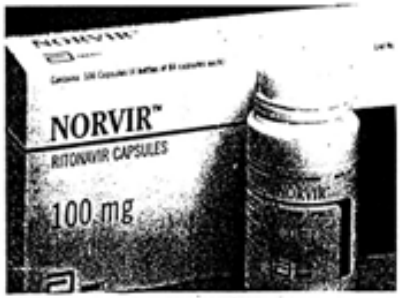
External Challenges

Ritonavir

THE PHARMACEUTICAL JOURNAL (VOL. 261) August 1, 1998

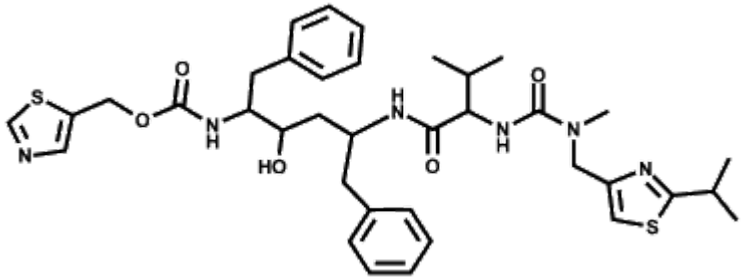
Manufacturing problems hit Abbott's HIV drug ritonavir

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 of recent production batches of Norvir capsules failed the approved test for dissolution, and were not released for marketing. Investigation of the reason for the failure showed the presence of a new crystalline form of ritonavir which affects the way it dissolves, and possibly its absorption. Retained sam-



Steve Lichter… from Abbott, discussed the efforts by Abbott to fix the situation. …Lichter said Abbott invested several million dollars to build new manufacturing facilities but failed. Apparently, they decided the emergence of crystals was not preventable and it would be impossible to manufacture ritonavir in the old form.

National Aids Treatment and Advocacy Project. Extract from report written by Jules Levin, NATAP http://www.natap.org/1998/norvirupdate.html

Rotigotine

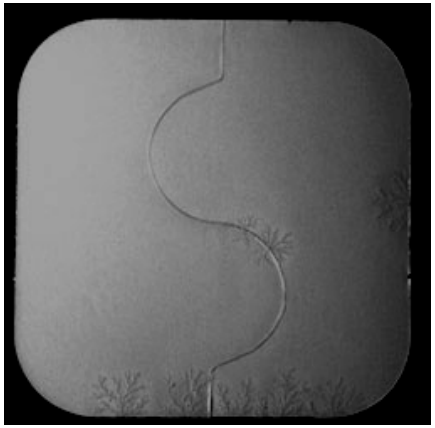
http://ebdgroup.com/partneringnews/2008/04/ucb-shares-drop-dramatically-after-announce-on-neupro-recall/

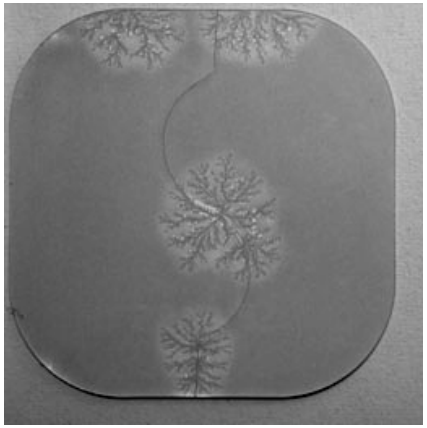
News Older Newer

UCB shares drop dramatically after announce on Neupro recall

April 8th, 2008

The Belgian UCB was recalling its Parkinson's drug Neupro (rotigotine transdermal system) in the United States and some batches from Europe, after uncovering a "deviation from the approved product specification". This decision has set a quick review of its 2008 sales forecast. UCB shares dropped as much as 18.4 percent to EUR 21.60 on 21 March, their lowest level since August 2003. According to UCB, the full effect on UCB's business is not yet known.





"A previously undiscovered polymorph of rotigotine with an altered three dimensional conformation and subsequently a differed packing arrangement of rotigotine molecules"

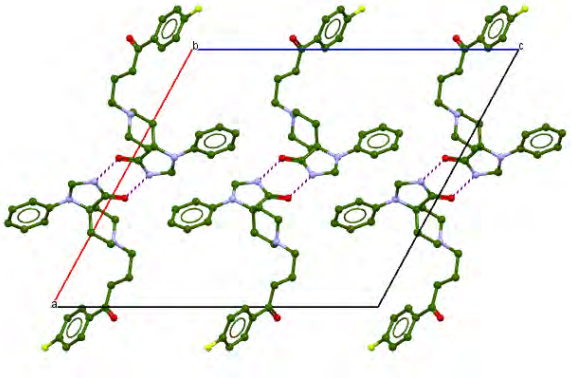
"The polymorph forms over time at room temperature, but does not develop significantly at colder temperatures"

Rotigotine transdermal system: An update JA Zackman, L Hakes, C Arth, L Bauer, L Dewulf
PO 1.145 61st Annual Meeting of the American Academy of Neurology, Seattle, 25 April-2 May 2009, *Neurology,* (2009) 72 (suppl 3)

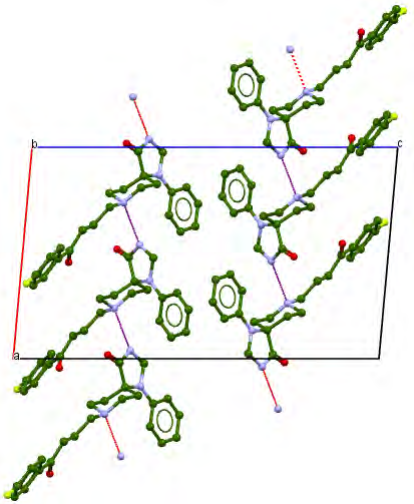
Polymorphism

- Polymorphism is the ability of a solid material to exist in more than one form or crystal structure
  - some polymorphic forms are more stable than others
  - formation depends on various conditions in the crystallisation process

- More than 80% of marketed drugs are polymorphic[1]
  - latent polymorphism has proved to be very costly
  - an important consideration in patent protection



FBPAZD: Spiperone, Form ii



FBPAZD01: Spiperone, Form i

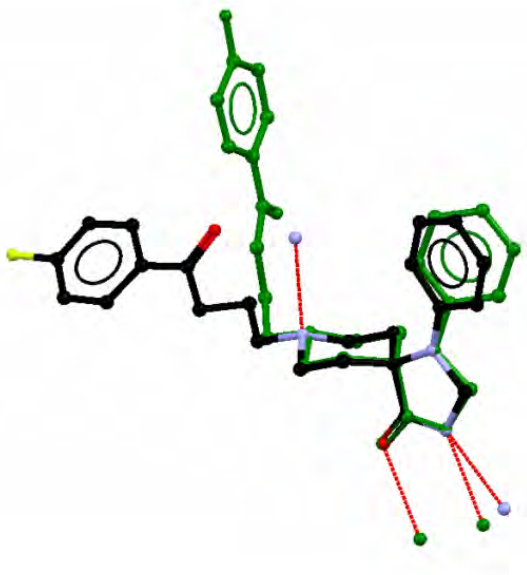
[1] Lohani, S.; Grant, D. J. W. Thermodynamics of Polymorphs. In *Polymorphism, Hilfiker, R., Ed.; Wiley,* 2006

Mitigating Risk: Solid Form Informatics

- The Cambridge Structural Database - an alternative definition:

  "An ensemble of free energy minima offering collective knowledge of intra and intermolecular properties from millions of discrete observations"

- The prevalent conserved features are *characteristic* of stable structures
  - geometric trends, conformation
  - atom/group interactions
  - hydrogen bonding
  - crystal packing

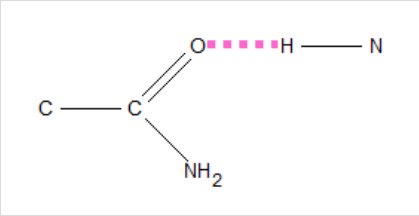


Spiperone

FBPAZD : FBPAZD01

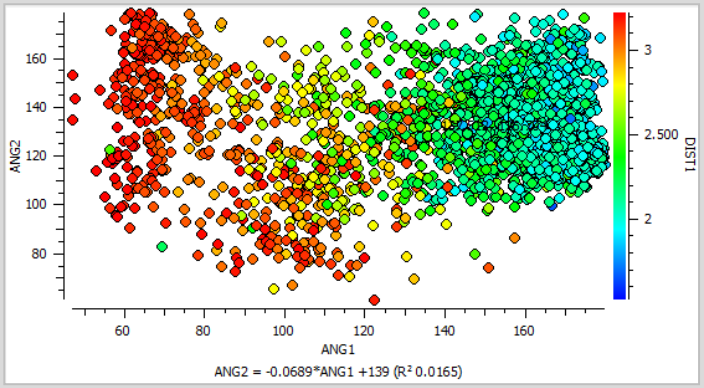
Interaction Analysis

ConQuest

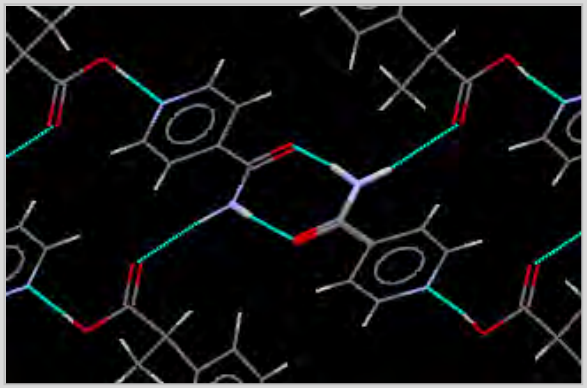


Find non-bonded contacts within sum of vdW radii +/- tolerance

Mercury



RONDAA

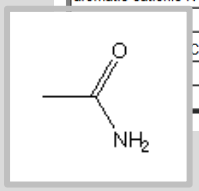


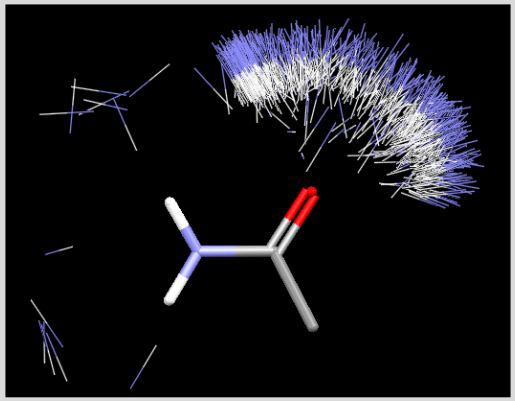
IsoStar: A Knowledge Base of Inter-molecular Interactions

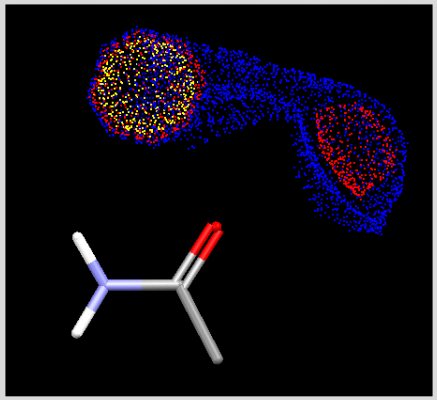
General| C,H only| N-H| O-H| Other N or O| Sulfur| Halo/halide| Amino acid

N-H				
Links to statistical data	Links to theoretical energy data			
Contact Group	*CSD*	*PDB*	*Stats*	*Theory*
any NH	4922	1561		
any uncharged NH	4924	*		
any cationic NH	331	101		
aromatic cationic N-H	44	*		
	198	759		
C(ar)-NH2	2498	*		
	99	101		
	5	*		

* This search has not been done.

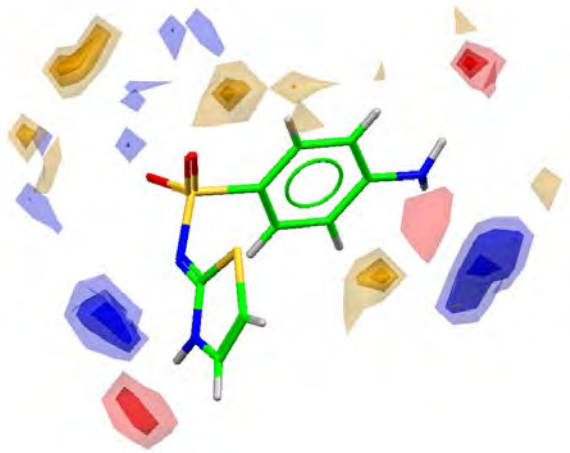






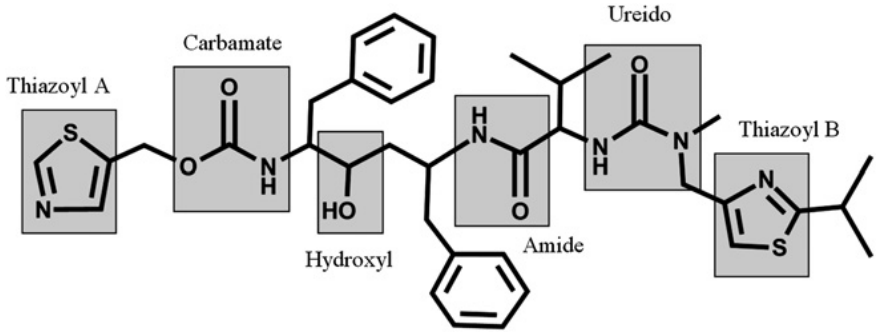
CCDC Solid Form Module

- An add-on to Mercury that aids with the assessment and discovery of new crystal forms:
  - hydrogen-bond motif analysis
  - crystal packing analysis
  - hydrogen bond propensity prediction
  - full interaction maps
  - coformer selection

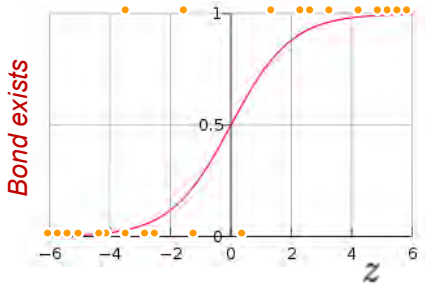


- Development guided by a consortium of industrial scientists
  - assess risks associated with an active ingredient
  - inform decisions about further experimental work

Hydrogen Bond Propensity



Predictive analytics is used to identify feasible and unusual crystal packings based on information from the known crystal structures of molecules similar to the target.



Propensity Π (bond exists) $= \dfrac{1}{1 + e^{-z}}$

$Z = \beta_0 + \beta_1.DG1 + \beta_2.AG1 + \beta_3.DSD + \ldots$

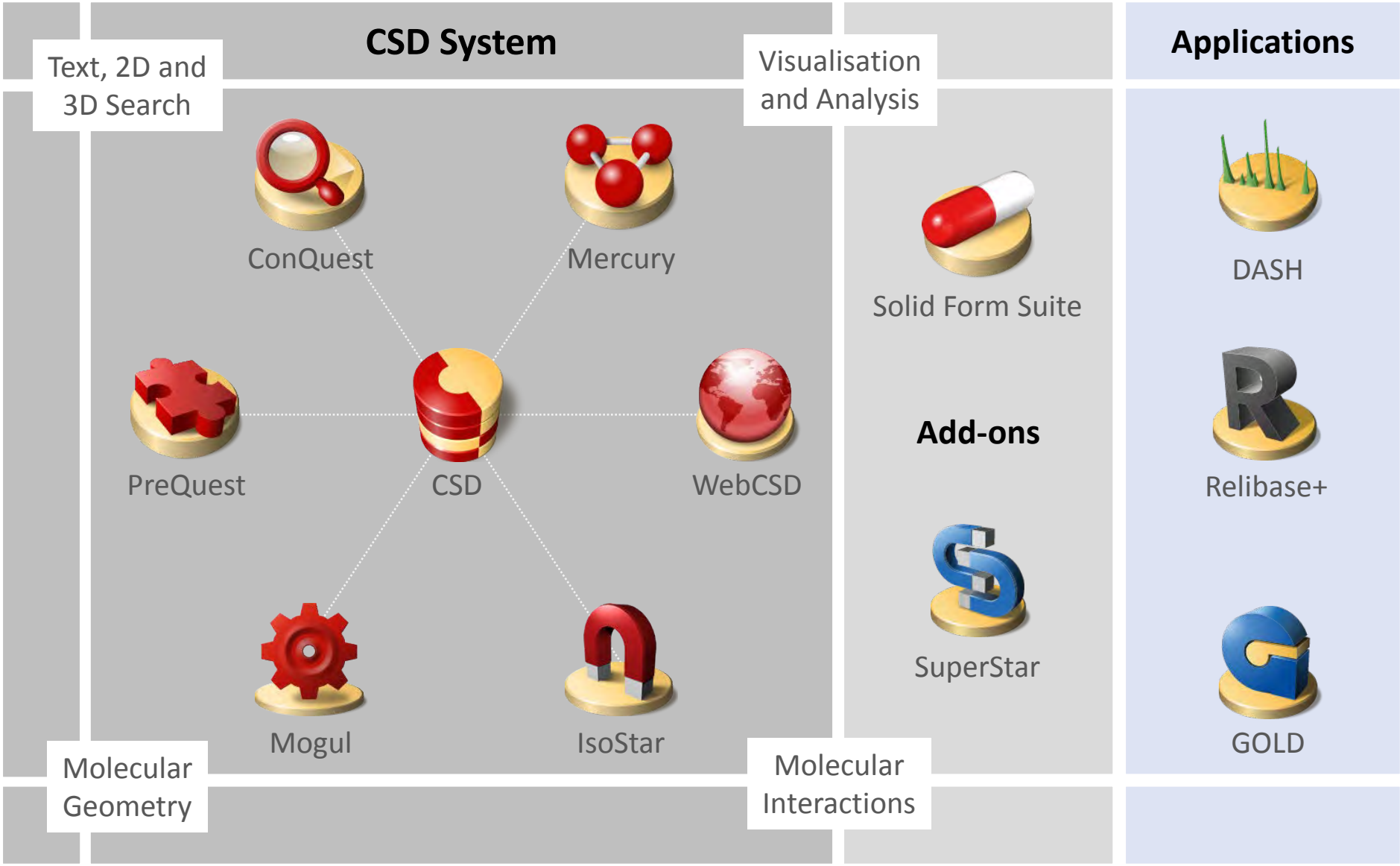
DG1, AG1, DSD: Explanatory variables

Table 1 Propensity predictions for potential donor–acceptor combinations in ritonavir (as labelled in Fig. 1), and observed hydrogen bonds in either polymorphic form

Donor	Acceptor	π	$\pm$ [a]	Form I	Form II
amide	carbamate	0.618	0.094	✗	✗
amide	hydroxyl	0.551	0.052	✗	✓
carbamate	carbamate	0.538	0.090	✓	✗
hydroxyl	carbamate	0.537	0.090	✗	✗
amide	amide	0.501	0.055	✓	✗
amide	ureido	0.499	0.072	✗	✗
carbamate	hydroxyl	0.470	0.078	✗	✗
hydroxyl	hydroxyl	0.469	0.037	✗	✗
carbamate	amide	0.420	0.083	✗	✓
hydroxyl	amide	0.419	0.045	✗	✗
carbamate	ureido	0.418	0.088	✗	✗
hydroxyl	ureido	0.417	0.058	✗	✓
ureido	carbamate	0.319	0.086	✗	✓
ureido	hydroxyl	0.263	0.041	✗	✗
ureido	amide	0.225	0.040	✗	✗
ureido	ureido	0.224	0.044	✓	✗
amide	thiazoyl a	0.152	0.054	✗	✗
amide	thiazoyl b	0.142	0.050	✗	✗
carbamate	thiazoyl a	0.115	0.044	✗	✗
hydroxyl	thiazoyl a	0.114	0.039	✓	✗
carbamate	thiazoyl b	0.107	0.041	✗	✗
hydroxyl	thiazoyl b	0.106	0.036	✗	✗

[a] The error bars of the coefficient value: the value falls within this range at the 95% confidence level, based on a χ^2 distribution.

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


CSD System
Text, 2D and 3D Search
Visualisation and Analysis
Applications
ConQuest
Mercury
Solid Form Suite
DASH
PreQuest
CSD
WebCSD
Add-ons
Relibase+
SuperStar
Mogul
IsoStar
GOLD
Molecular Geometry
Molecular Interactions

The CCDC Python API

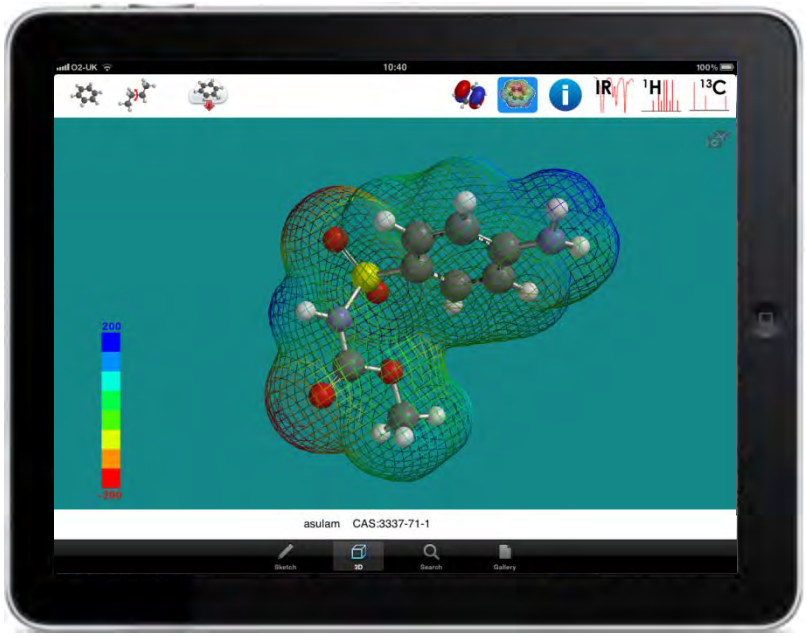
- Programmatic access to CCDC data and functionality
- Intended to facilitate:
  - integration with third party applications and internal workflows
  - development of Web Services and pipeline components
  - analyses not currently possible through existing interfaces
- *Currently in development*


```
def main(infile, outfile, local_density_cutoff, min_obs):
    """The main program for filtering a set of conformers."""
    mogul_engine = create_mogul_engine()

    # Set the current best (minimum number of unusual torsions) to a large
    # number.
    best_so_far = sys.maxint

    # Open a file to write molecules to.
    out = ccdc.io.MoleculeWriter(outfile)
    with ccdc.io.MoleculeReader(infile) as mol_reader:
        for mol in mol_reader:

            # Do the Mogul analysis.
            mogul_checked_mol = mogul_engine.analyse_molecule(mol)

            # Find the number of unusual torions according to the input
            # criteria.
            unusual_torsions = [
                t for t in mogul_checked_mol.mogul_torsions
                  if (t.local_density < local_density_cutoff
                      and t.nhits > min_obs)
            ]
            num_unusual_torsions = len(unusual_torsions)

            # Should the conformer be kept or discarded.
            if num_unusual_torsions > best_so_far:
                continue
            elif num_unusual_torsions == best_so_far:
                out.write(mogul_checked_mol)
            elif num_unusual_torsions < best_so_far:
                # We have found something better, which means we have to
                # discard all the conformers written out already. We therefore
                # close the output writer and open it again.
                out.close()
                out = ccdc.io.MoleculeWriter(outfile)
                out.write(mogul_checked_mol)

                # Make sure that we update the variable that we are using for
                # the comparision.
                best_so_far = num_unusual_torsions
```

Mobile Devices

- Partnership with Wavefunction to provide access to CSD through iSpartan

- iSpartan offers
  - conformational analysis
  - calculated NMR and IR spectra
  - display of molecular orbitals
  - electrostatic potential maps
  - *coming soon: CSD search*



The Cambridge Crystallographic Data Centre

International Data Repository
Archive of crystal structure data
Deposition services

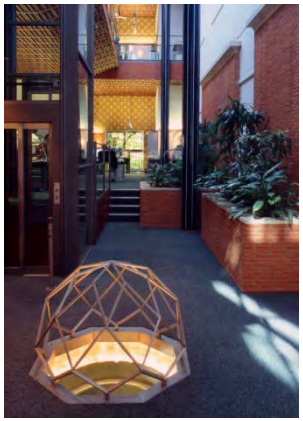
Scientific Software Provider
Search/analysis/visualisation tools
Scientific applications

Collaborative Research Organisation
New methodologies
Fundamental research

@ccdc_cambridge

ccdc.cambridge

LinkedIn



Ian Bruno

bruno@ccdc.cam.ac.uk