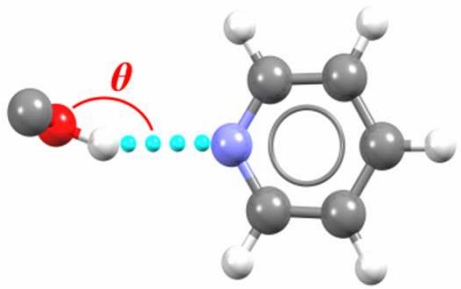
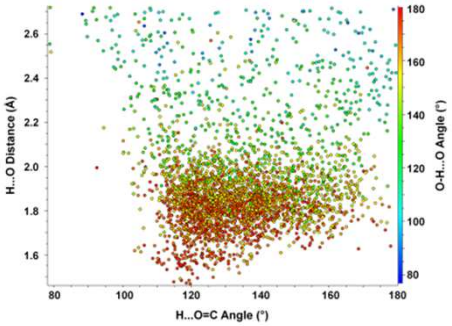




Introduction to the CSD-Enterprise Suite

Cambridge Crystallographic Data Centre (CCDC)

June 5th, 2018

The Cambridge Crystallographic Data Centre

The Cambridge Structural Database (CSD)

A database of organic and metal-organic crystal structures

Established in 1965

~60 staff

Data curators
Software developers
Applications scientists

Cambridge UK • Piscataway NJ

International Data Repository
Archive of crystal structure data
High quality scientific database

Scientific Software Provider
Search/analysis/visualisation tools
Scientific applications

Collaborative Research Organisation
New methodologies
Fundamental research

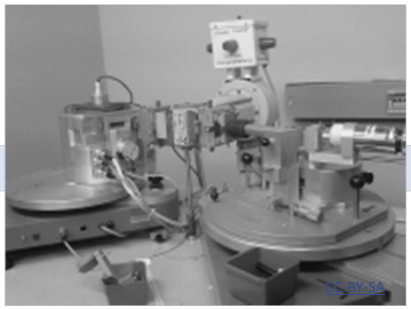
Education and Outreach
Conferences, Workshops,
Bespoke Training, Teaching Materials

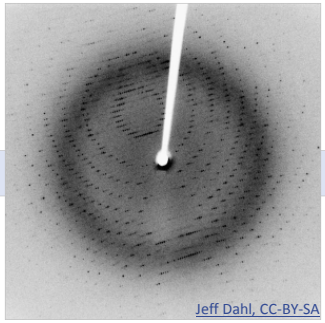
Dedicated to the advancement of chemistry and crystallography for the public benefit through providing high quality information services and software.

From Data to Knowledge

Experimental Data

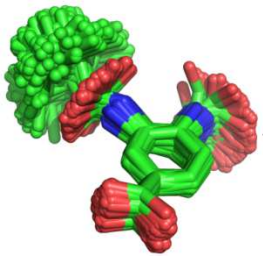
$C_{10}H_{16}N^{+},Cl^{-}$

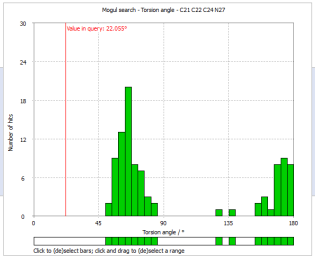


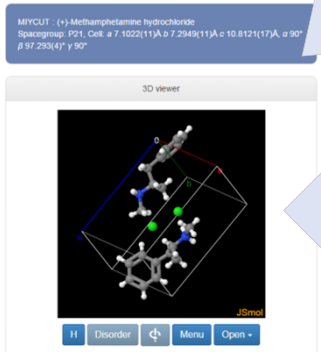


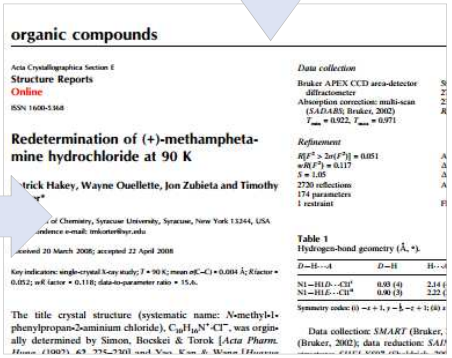
Jeff Dahl, CC-BY-SA

```
_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
 tom_site_fract_z
  m_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 . .
```









organic compounds

Acta Crystallographica Section E
Structure Reports
Online

Redetermination of (+)-methamphetamine hydrochloride at 90 K

Data collection

Refinement

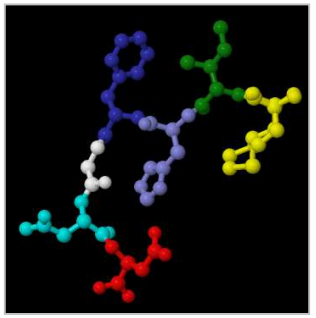
Table 1
Hydrogen-bond geometry (Å, °).

Structural Knowledge

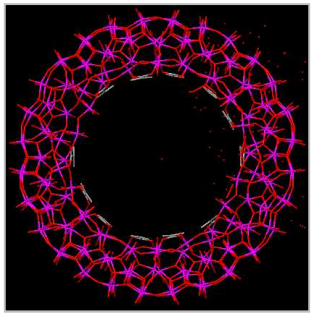
Crystal structure databases

- Cambridge Structural Database
  - organic and metal-organic compounds
  - >900,000 structures
  - Established in 1965
- Inorganic Crystal Structure Database
  - inorganic compounds
  - >185,000 structures
  - Established in 1978
- Protein Data Bank
  - biological macromolecules
  - >120,000 structures
  - Established in 1971

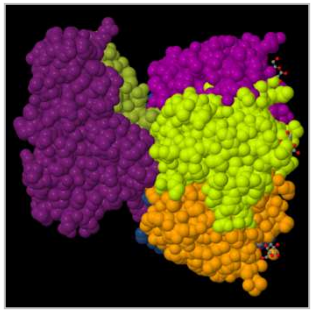
Cambridge Crystallographic Data Centre



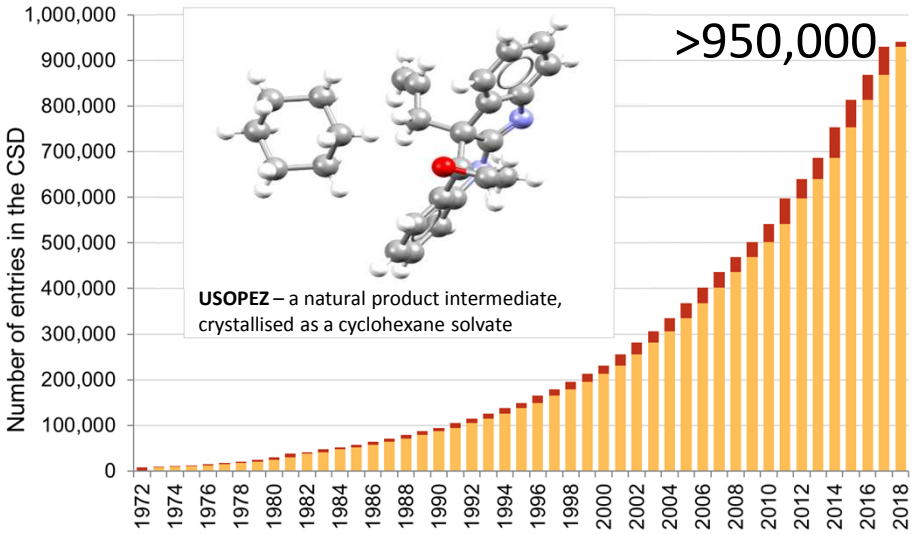
FIZ KARLSRUHE
NIST



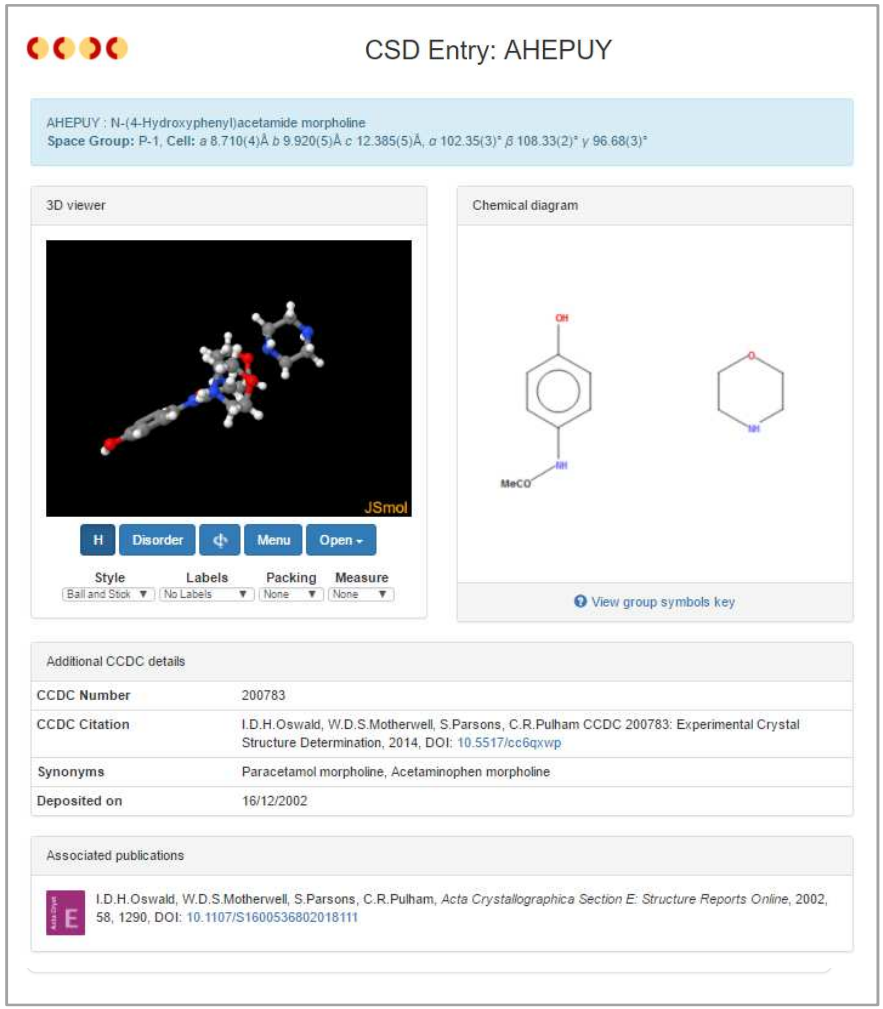
WORLDWIDE PDB
PROTEIN DATA BANK



The Cambridge Structural Database



- 900,000+ small-molecule crystal structures
- Over 80,000 datasets deposited annually
- Structures available for anyone to download
- Links to over 1,000 journals
- Enriched and annotated by experts
- Access to data and knowledge



What's in 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%

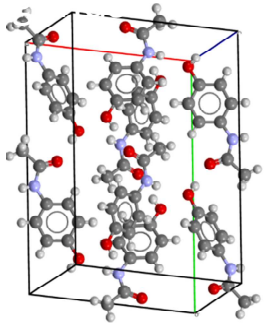
Polymeric: 11%

Single Component
56%

Multi Component
44%

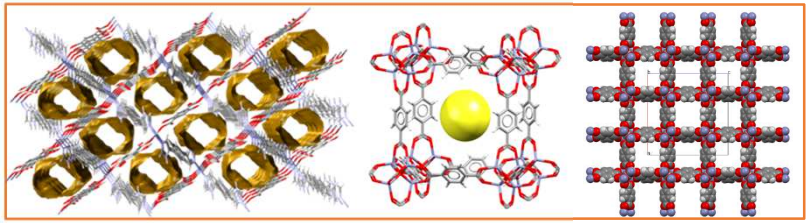
Organic

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



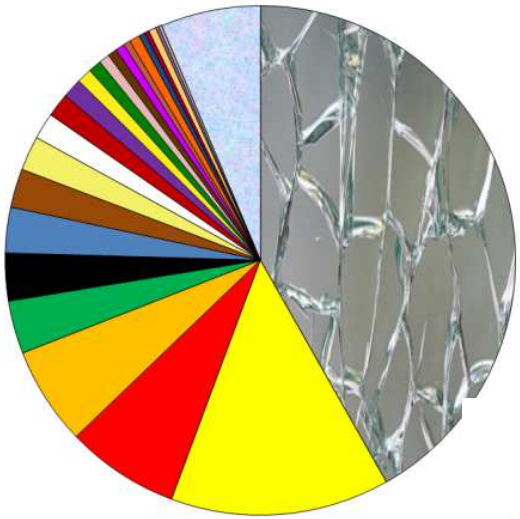
Metal-Organic

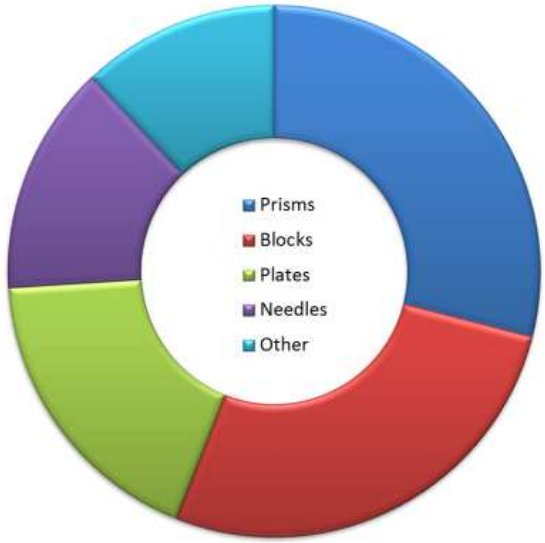
- Metal Organic Frameworks
- Models for new catalysts
- Porous frameworks for gas storage
- Fundamental chemical bonding



What other data is in the CSD?

- 9,959 polymorphic families
- 165,617 melting points
- 747,577 crystal colours
- 614,544 crystal shapes
- 21,939 bioactivity details
- 9,727 natural source data
- 46,238 entries classified
- > 250,000 oxidation states
  - Including 240,666 transition metal oxidation states



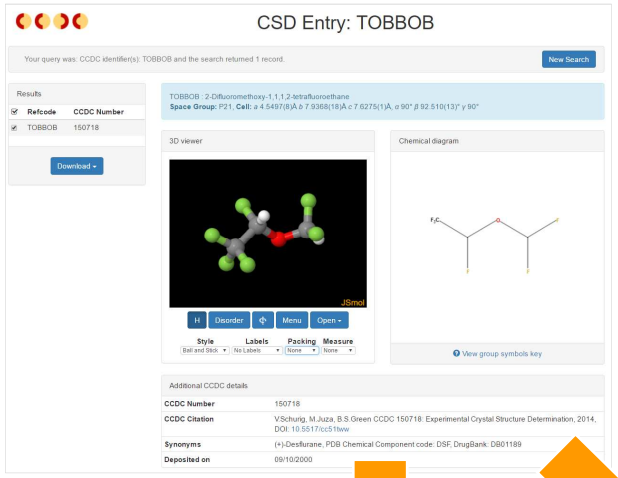


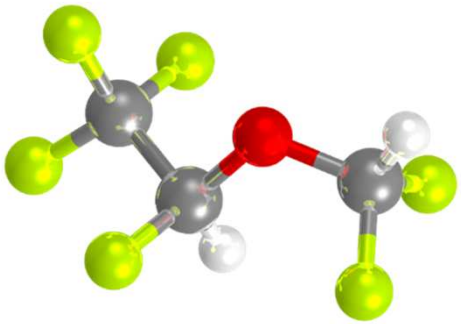
Top 200 small molecule drugs

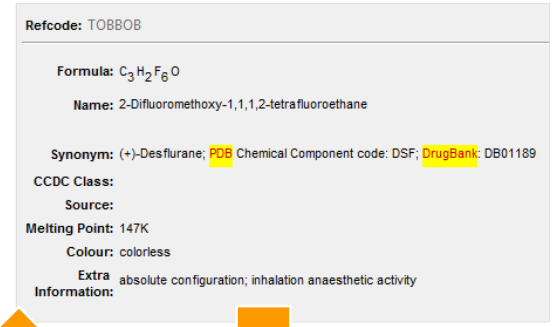
Green = Exact matches in the CSD (118 drugs)

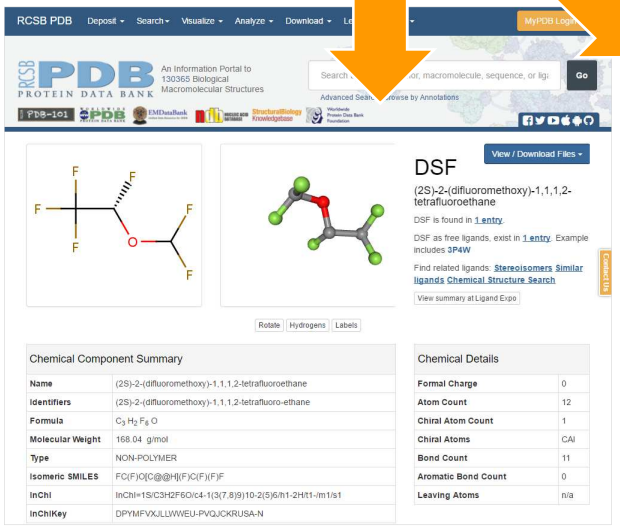
Grey = Exact match not in the CSD but most likely something very similar is

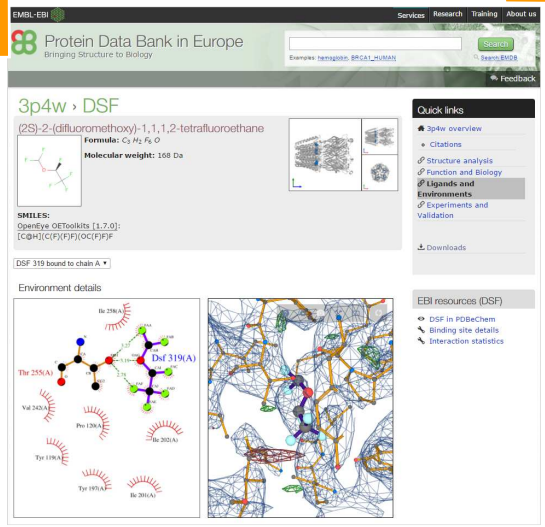
Accessing drugs and ligands in the CSD

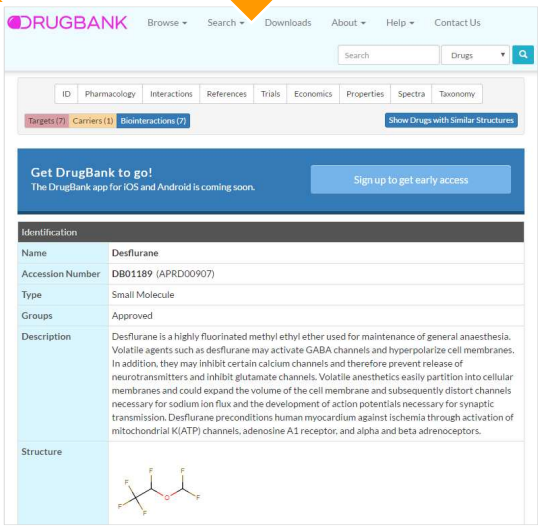






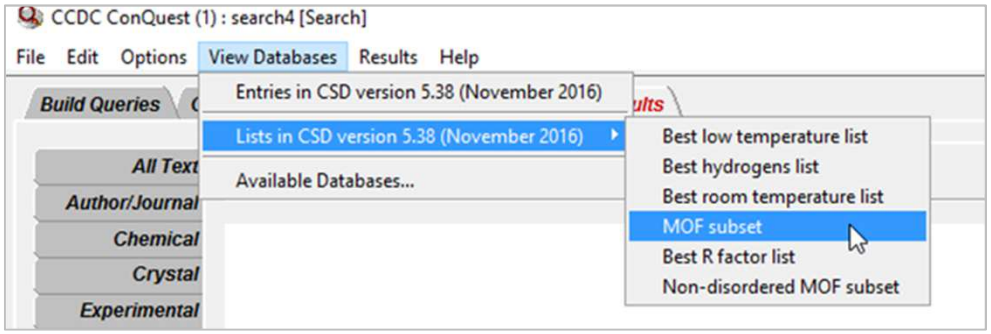


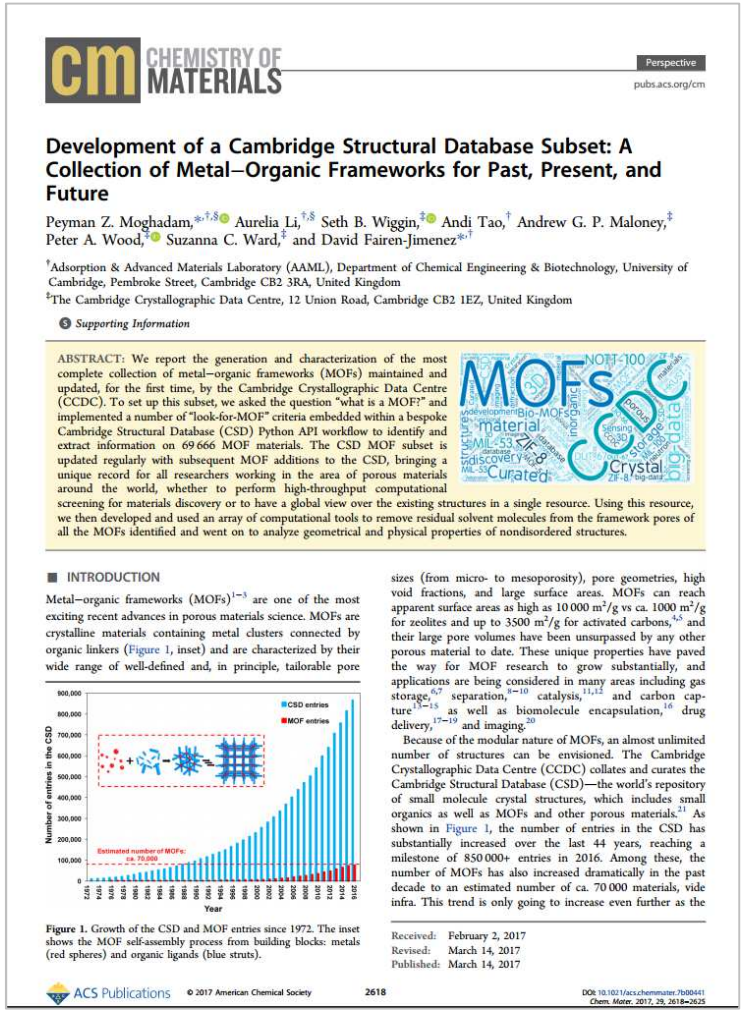




Metal-Organic Frameworks

- Two CSD subsets identifying all metal-organic frameworks within the database
  1. MOF list >**75,000** entries intended to be as widely-applicable as possible
  2. Non-disordered list >**61,000** entries for high-throughput calculations
     - Both subsets updated quarterly
- Python script available to download





CHEMISTRY OF MATERIALS

Perspective

pubs.acs.org/cm

Development of a Cambridge Structural Database Subset: A Collection of Metal–Organic Frameworks for Past, Present, and Future

Peyman Z. Moghadam,*[†,§] Aurelia Li,[†,§] Seth B. Wiggin,[‡] Andi Tao,[†] Andrew G. P. Maloney,[‡] Peter A. Wood,[‡] Suzanna C. Ward,[‡] and David Fairen-Jimenez*[,†]

[†]Adsorption & Advanced Materials Laboratory (AAML), Department of Chemical Engineering & Biotechnology, University of Cambridge, Pembroke Street, Cambridge CB2 3RA, United Kingdom

[‡]The Cambridge Crystallographic Data Centre, 12 Union Road, Cambridge CB2 1EZ, United Kingdom

Supporting Information

ABSTRACT: We report the generation and characterization of the most complete collection of metal–organic frameworks (MOFs) maintained and updated, for the first time, by the Cambridge Crystallographic Data Centre (CCDC). To set up this subset, we asked the question "what is a MOF?" and implemented a number of "look-for-MOF" criteria embedded within a bespoke Cambridge Structural Database (CSD) Python API workflow to identify and extract information on 69 666 MOF materials. The CSD MOF subset is updated regularly with subsequent MOF additions to the CSD, bringing a unique record for all researchers working in the area of porous materials around the world, whether to perform high-throughput computational screening for materials discovery or to have a global view over the existing structures in a single resource. Using this resource, we then developed and used an array of computational tools to remove residual solvent molecules from the framework pores of all the MOFs identified and went on to analyze geometrical and physical properties of nondisordered structures.

INTRODUCTION

Metal–organic frameworks (MOFs)[1−3] are one of the most exciting recent advances in porous materials science. MOFs are crystalline materials containing metal clusters connected by organic linkers (Figure 1, inset) and are characterized by their wide range of well-defined and, in principle, tailorable pore sizes (from micro- to mesoporosity), pore geometries, high void fractions, and large surface areas. MOFs can reach apparent surface areas as high as 10 000 m^2/g vs ca. 1000 m^2/g for zeolites and up to 3500 m^2/g for activated carbons,[4,5] and their large pore volumes have been unsurpassed by any other porous material to date. These unique properties have paved the way for MOF research to grow substantially, and applications are being considered in many areas including gas storage,[6,7] separation,[8−10] catalysis,[11,12] and carbon capture[13−15] as well as biomolecule encapsulation,[16] drug delivery,[17−19] and imaging.[20]

Because of the modular nature of MOFs, an almost unlimited number of structures can be envisioned. The Cambridge Crystallographic Data Centre (CCDC) collates and curates the Cambridge Structural Database (CSD)—the world's repository of small molecule crystal structures, which includes small organics as well as MOFs and other porous materials.[21] As shown in Figure 1, the number of entries in the CSD has substantially increased over the last 44 years, reaching a milestone of 850 000+ entries in 2016. Among these, the number of MOFs has also increased dramatically in the past decade to an estimated number of ca. 70 000 materials, vide infra. This trend is only going to increase even further as the

Figure 1. Growth of the CSD and MOF entries since 1972. The inset shows the MOF self-assembly process from building blocks: metals (red spheres) and organic ligands (blue struts).

Received: February 2, 2017
Revised: March 14, 2017
Published: March 14, 2017

CSD-Community

- CSD-Community provides a selection of products and services free of charge for the benefit of the scientific community
- This includes:

enCIFer
CIF checking, editing and visualisation software from the CCDC

Relibase
Easy Searching for Protein-Ligand Complexes

CSDSymmetry
The most complete collation of observed molecular and crystallographic symmetry properties to date

Deposit Structures
Upload your data to the CCDC for inclusion in the Cambridge Structural Database

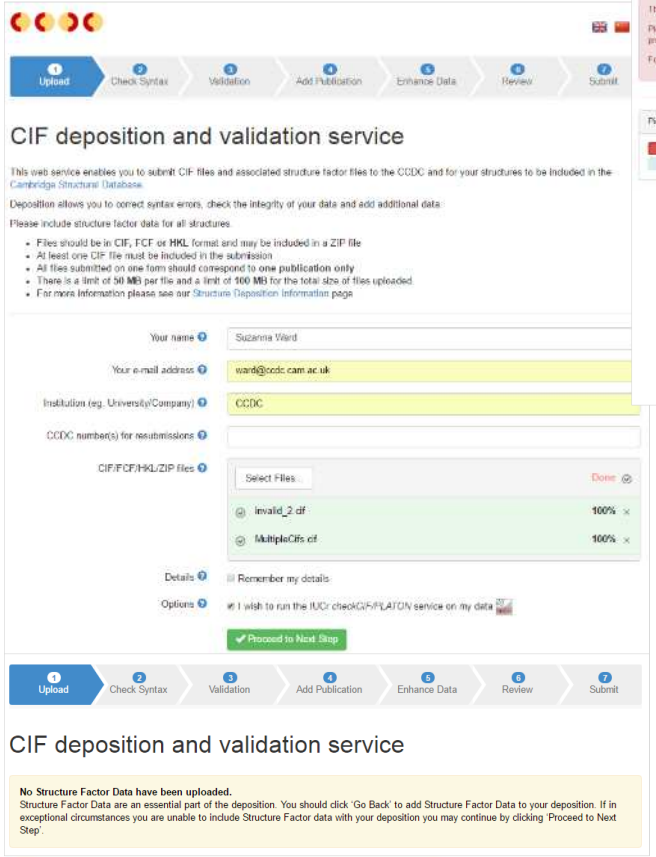
Mercury
Crystal Structure Visualisation and Exploration Made Easy

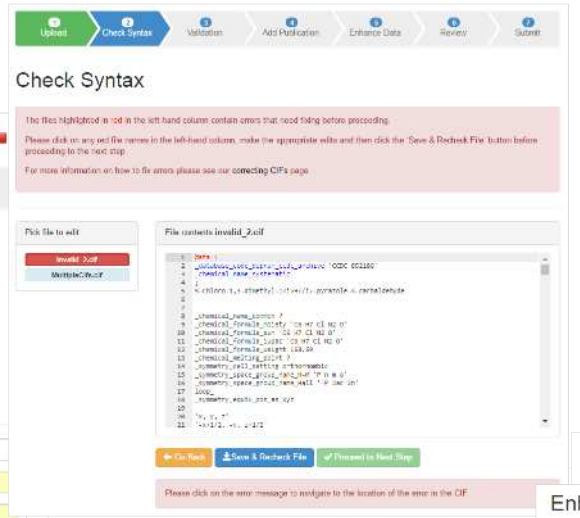
CSD Educational Collection
Free 500 structure subset of the CSD for teaching purposes

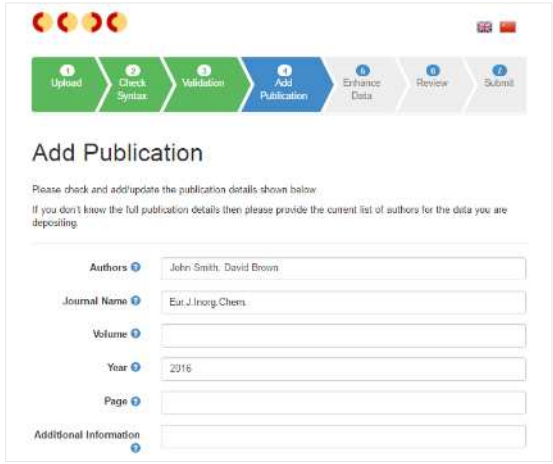
CellCheckCSD
A command-line tool for performing crystal structure reduced cell checks against the CSD

Access Structures
View and retrieve structures in the Cambridge Structural Database

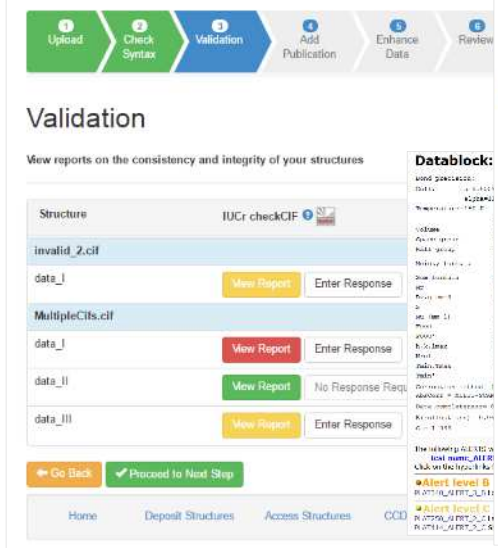
CSD-Deposit

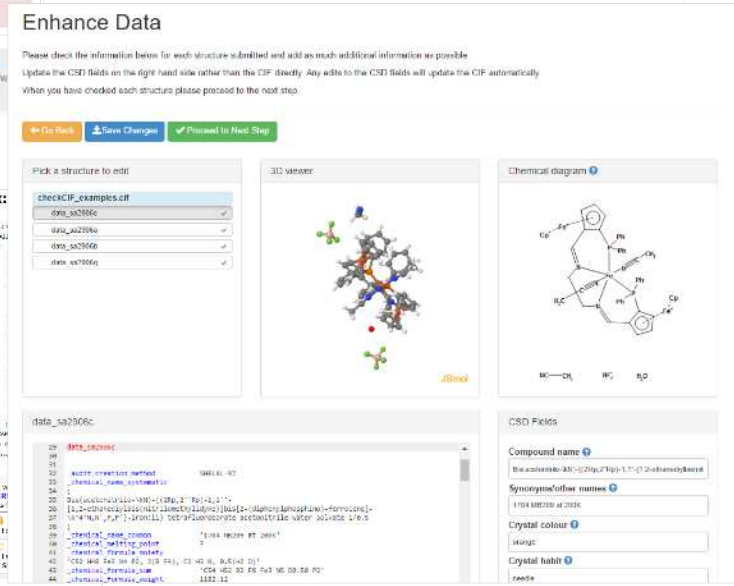






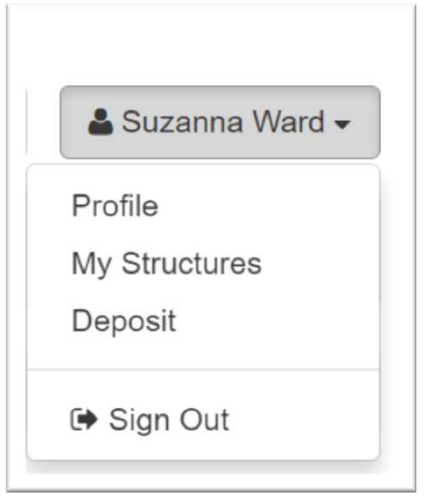
If you do not intend to publish the data in the scientific literature and would like to share the data immediately through the Cambridge Structural Database then please click the 'CSD Communication' button below.

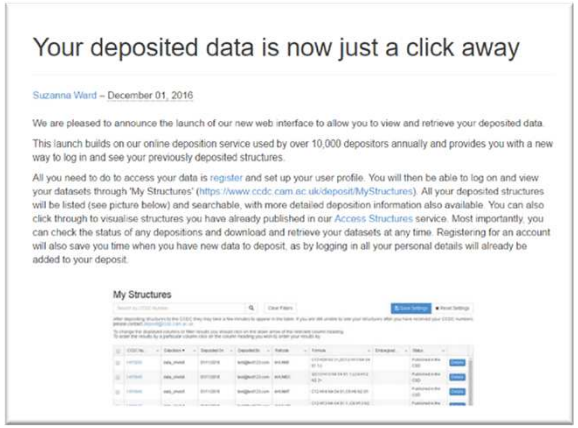


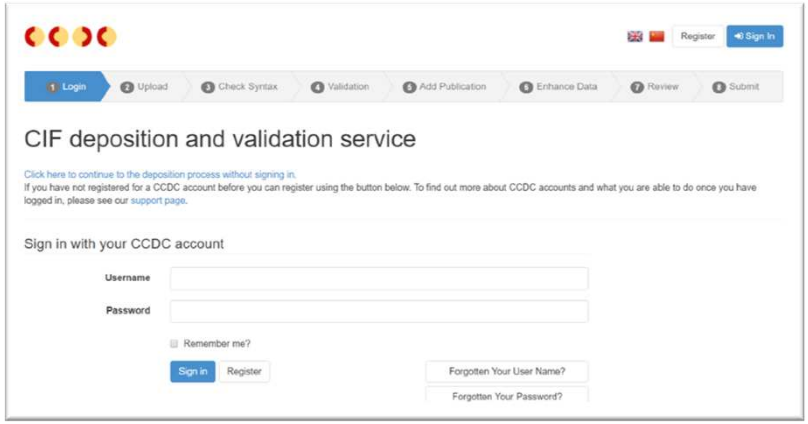


Increasing deposition functionality

- A deposition portal **launched December 1st 2016**
  - Ability to log on and view and retrieve depositions
  - Deposit and revise data
  - Edit and update basic information
  - Publish data directly through a database
  - Share data with co-workers
- Over 15,000 registered users







My Structure Details

Revise Structure | Back to My Structures

Suzanna Ward

Datablock: tBu10kbar
Space Group: I2/a, **Cell:** *a* 14.811Å *b* 6.456Å *c* 19.759Å, *α* 90° *β* 94.060(8)° *γ* 90°
Formula: C11 H15 N1 O2, Temperature: 298 K

Additional CCDC Details

Deposition Number	1584792
Refcode	
Compound Name	2-t-butyl-6-[(hydroxyimino)methyl]phenol
Deposited On	09/11/2017
Deposited By	
Additional Depositors	
Status	Unpublished
Embargoed Date	05/03/2019

Publish as a CSD C

Associated Publications

Publication	Created on	Modified on
Suzanna Ward	09/11/2017	09/11/2017

Add Publication

Structure Shared With

Name	Edit Rights	Sharing Rights

Share Structure

Chemical diagram

No Diagram Available

Save Settings | Reset Settings

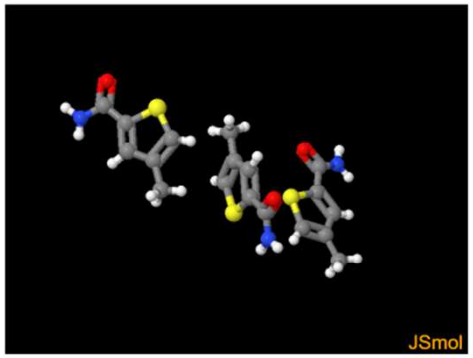
	825800	data_2007may0...	06/06/2011
Details	825789		

1 2 20 items per page

Download Selected 0 | View 0 in Access Structures | Export | Clear selectio

EPAGAE : 4-Methylthiophene-2-carboxamide
Space Group: P $\bar{1}$ (2), **Cell:** *a* 10.0846(3)Å *b* 10.4718(3)Å *c* 10.9149(3)Å, *α* 106.7180(10)° *β* 111.473(2)° *γ* 95.603(2)°

3D viewer



H | Disorder | Menu | Open

Style	Labels	Packing	Measure
Ball and Stick	No Labels	None	None

Chemical diagram

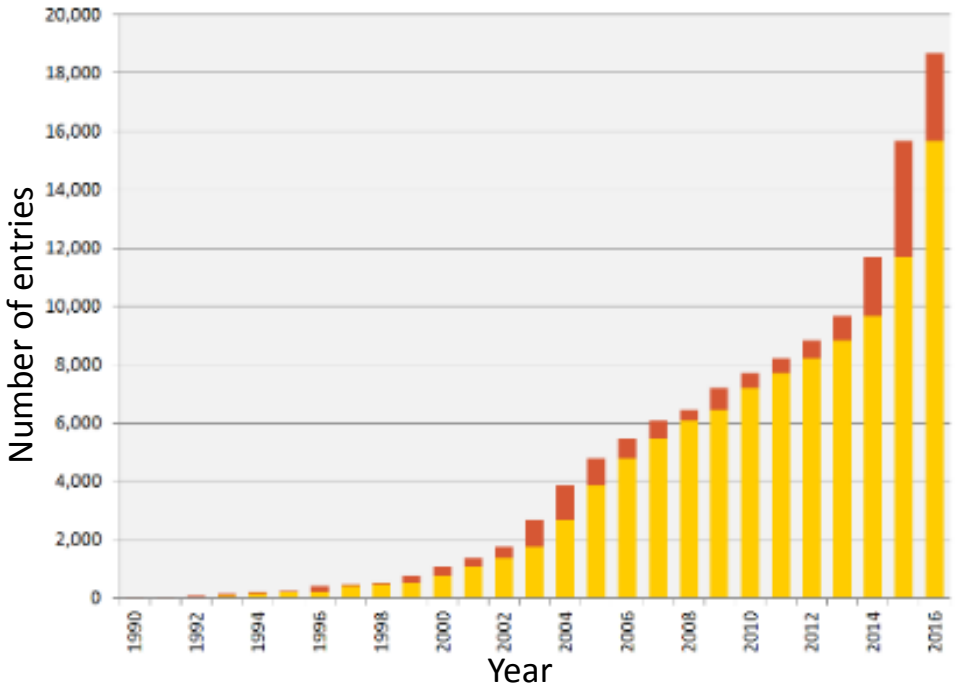
View group symbols key

Additional details

Deposition Number	825792
Data Citation	S.L.Huth, M.B.Hursthouse CCDC 825792: Experimental Crystal Structure Determination, 2014, DOI: 10.5517/ccwq9gb
Deposited on	13/05/2011

CSD Communications

- Allow data to be published directly through the CSD
  - During deposition
  - After Deposition via My Structures
- > 23,000 *CSD Communications*



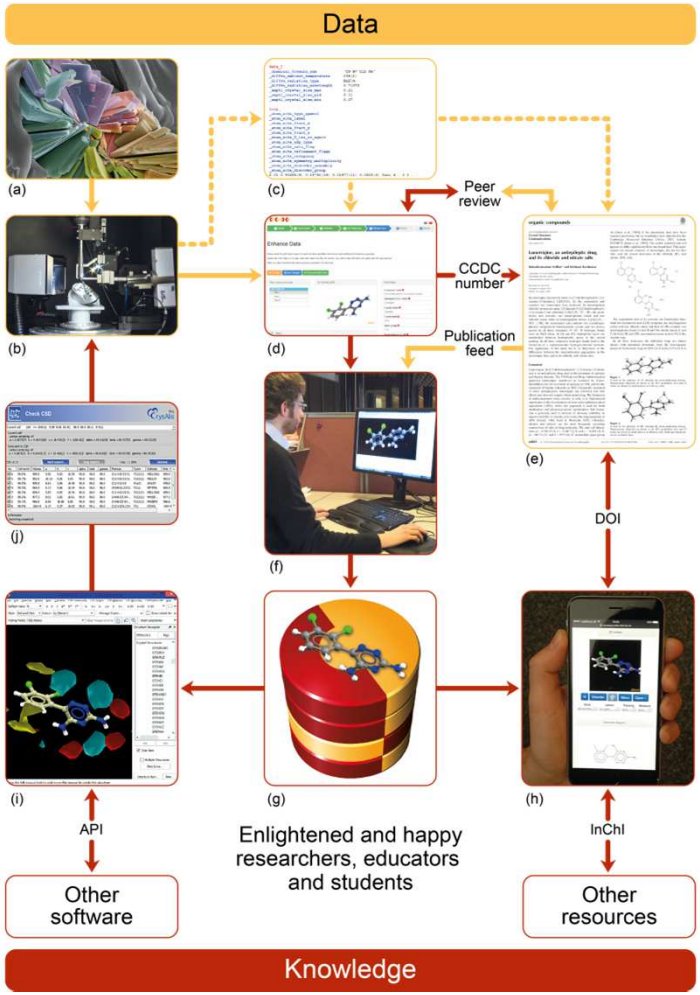
Deposition guidelines

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

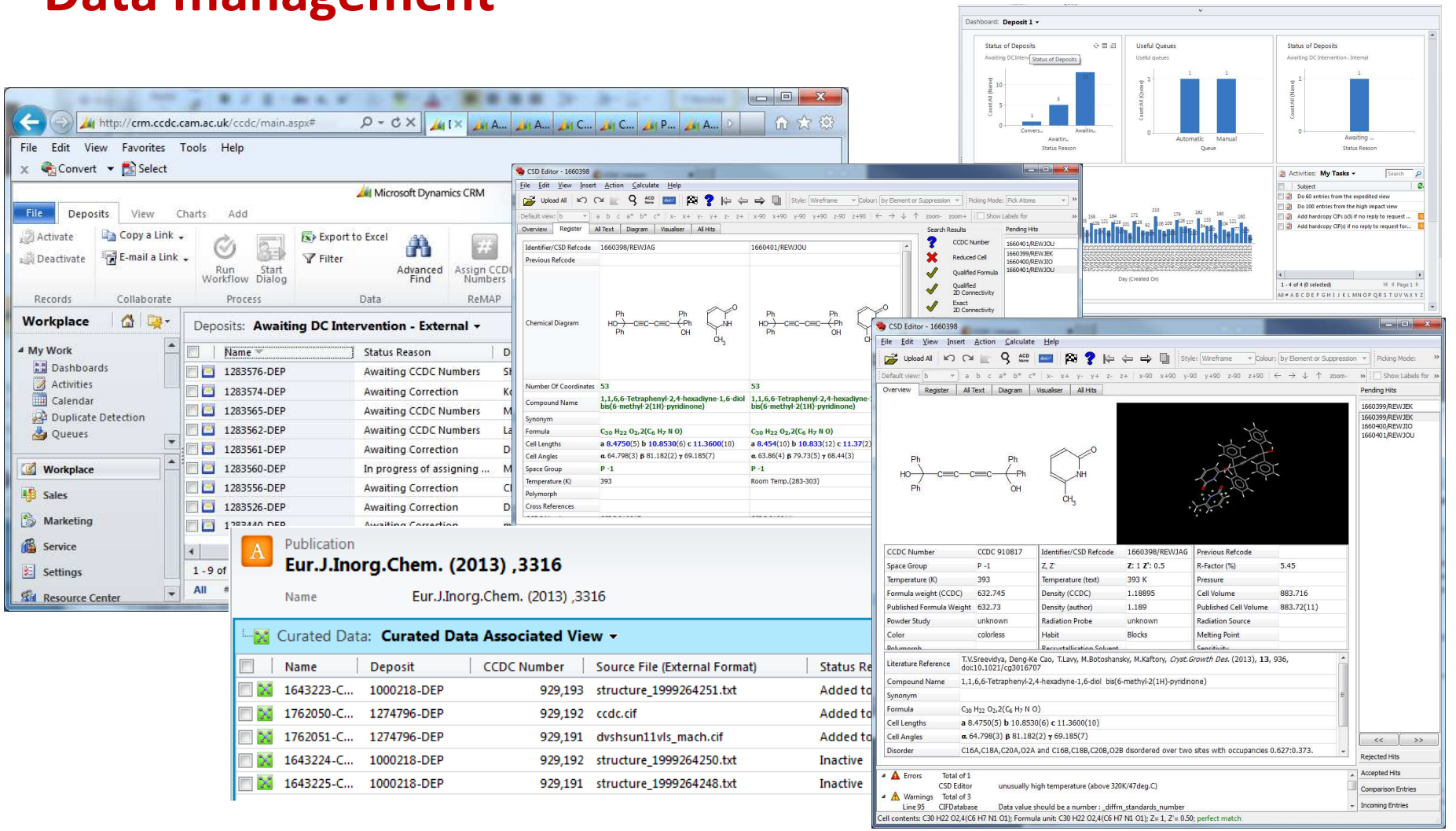
Small molecule single crystal data

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

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



Data management

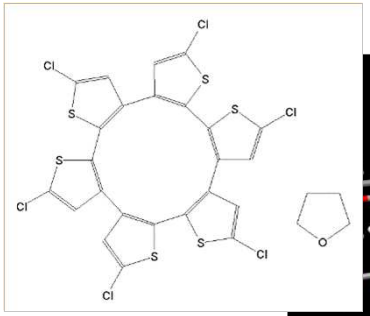


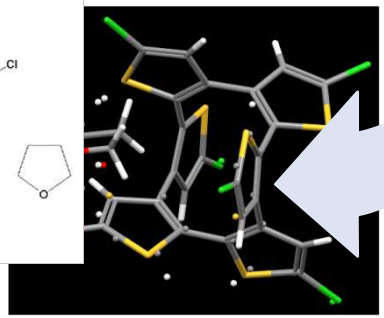
Discovery and re-use - Assignment of chemistry

Assignment of chemistry is important for discovery, data re-use, mining , analysis and interoperability

```
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
```

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





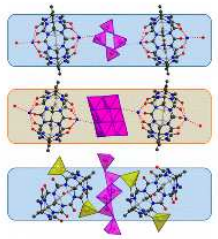
Sandwich-type Inorganic–Organic Hybrid Solids of Iso-polyvanadate Clusters and Decamethylcucurbit[5]uril

Li-Wei Han, Jing-Xiang Lin, Qi Yin, Bahar Karadeniz, Hong-Fang Li, Jian Lü, and Rong Cao

Publication Date (Web): January 15, 2016 (Article)

DOI: 10.1021/acs.cgd.5b01176

CRYSTAL GROWTH & DESIGN



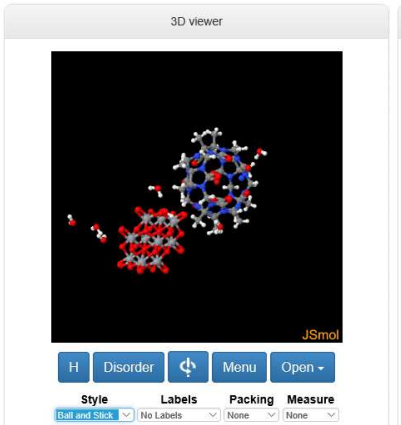
Accession Codes

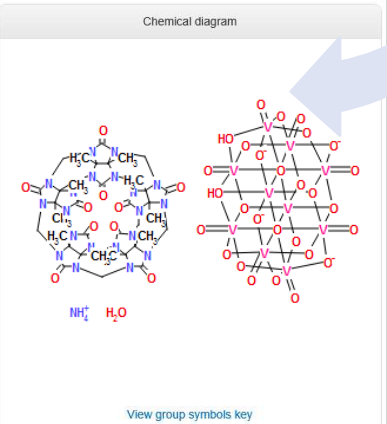
CCDC: 1032491

CCDC: 1032492

CCDC: 1032493

NAJYIJ : tetrakis(ammonium) bis(decamethylcucurbit[5]uril) bis(μ-hydroxy)-octa(oxo)-octadeca(μ-oxido)-deca-vanadium octadecahydrate
Space Group: P-1, Cell: *a* 14.9979(6)Å *b* 16.5740(11)Å *c* 16.6113(8)Å, *α* 63.456(6)° *β* 71.772(4)° *γ* 66.779(5)°

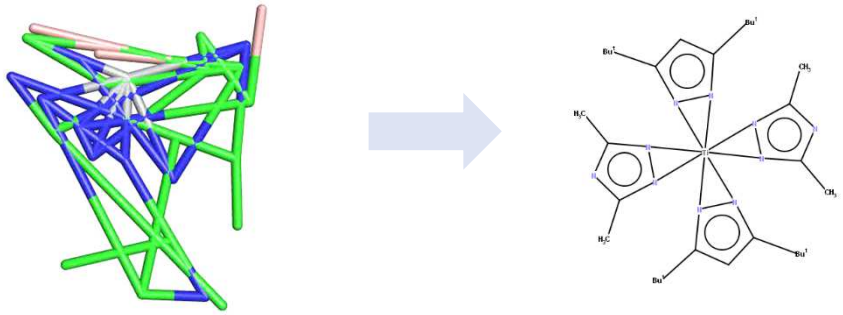




Enrichment and curation

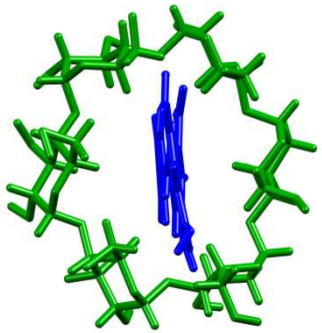
2D chemical diagrams:

Generated from existing curated CSD entries, CCDC diagram generation software, integrated third party software, manual curation



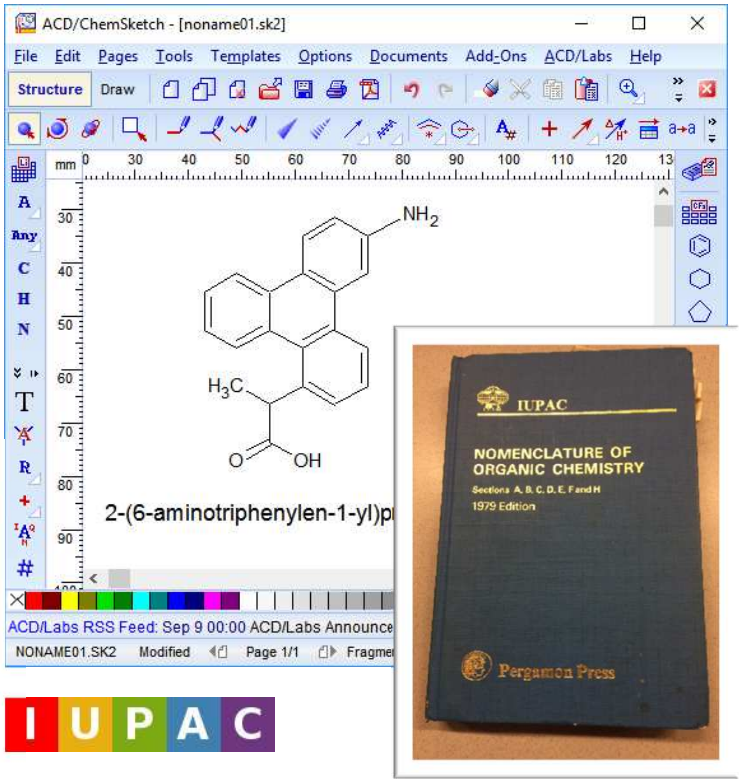
Classes:

ARACEB : (2)-(1,4-bis(2-(3,5-Dicarboxyphenyl)vinyl)benzene)-(alpha-cyclodextrin)-**rotaxane** dodecahydrate



Compound names:

Based on IUPAC conventions using ACDname and manual curation

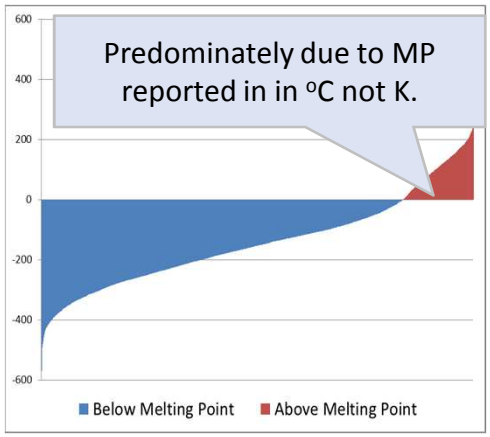


Annual targeted database improvements

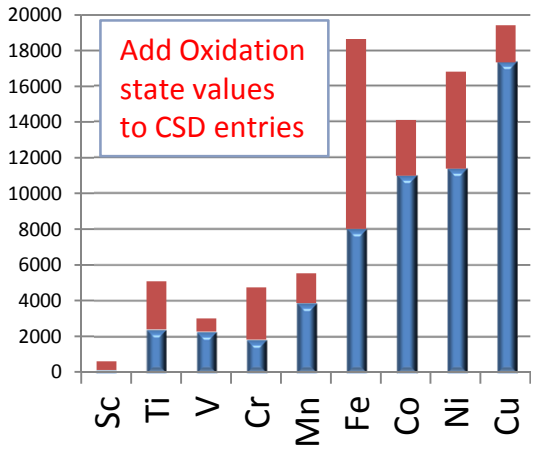
- Greater efficiency allows targeted improvements to further increase integrity, consistency, discoverability and usage of data
  - 50,000 existing entries improved in 2015
  - 74,000 existing entries improved in 2016
  - 62,00 existing entries improved in 2017

Targeted improvements to drug subset

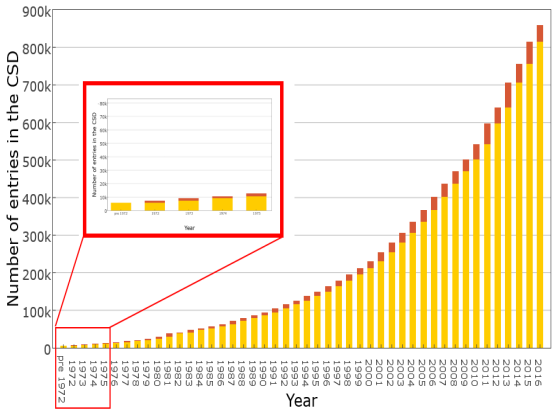
Study Temperature relative to Melting Point



Targeted improvements to metal-organics

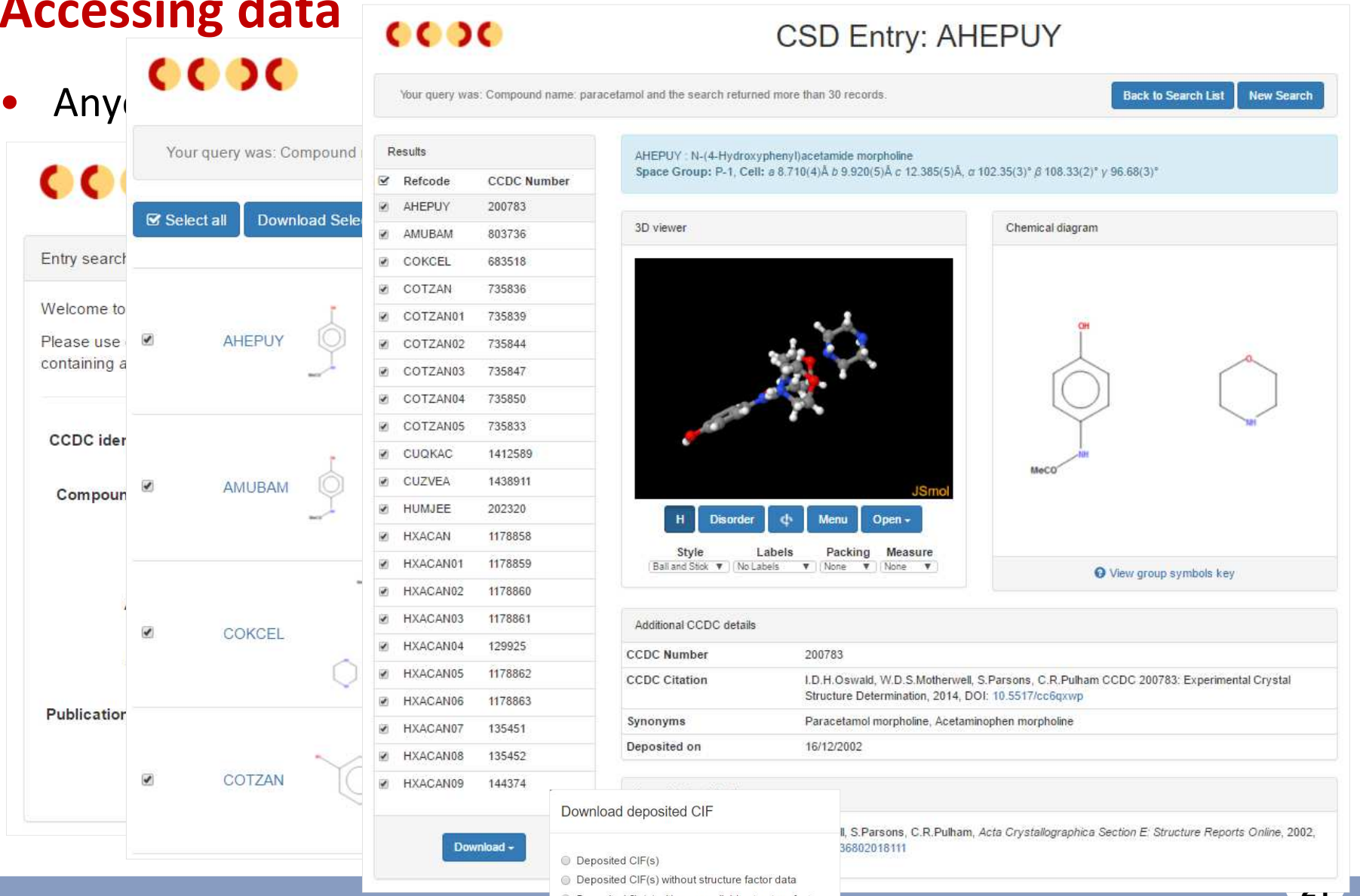


Ensure consistency of early CSD entries

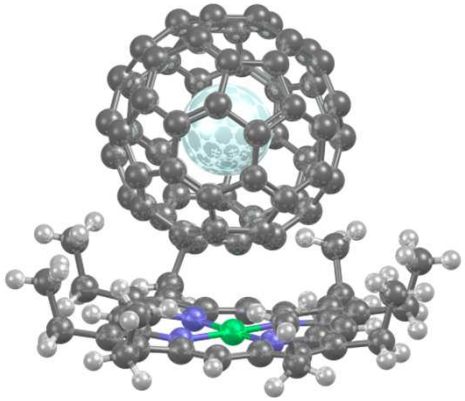


Accessing data

- Any



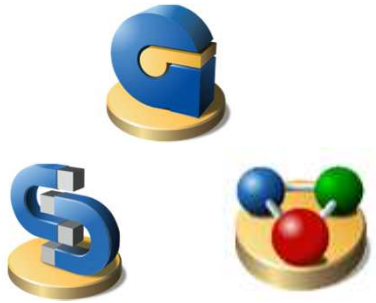
The CCDC Portfolio 2017-2018



- Alongside the CSD, we produce a suite of software enabling research across the whole breadth of structural science

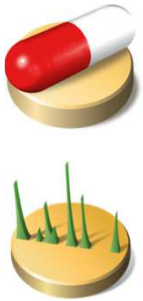
CSD-Enterprise

CSD-Discovery



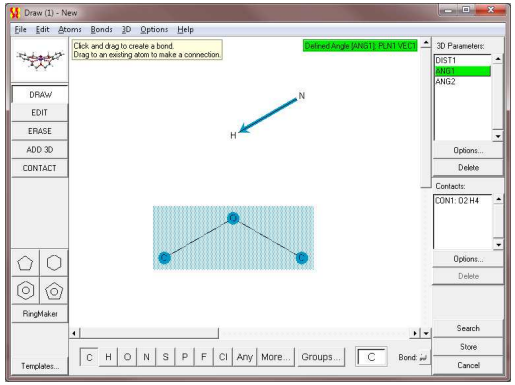
CSD-System

CSD-Materials

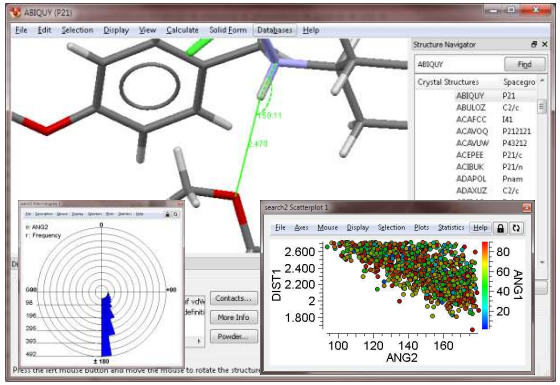


CSD-System

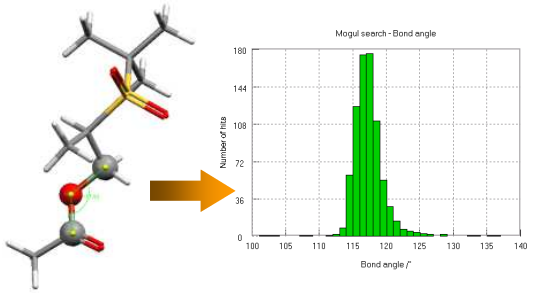
ConQuest: Advanced 3D searching



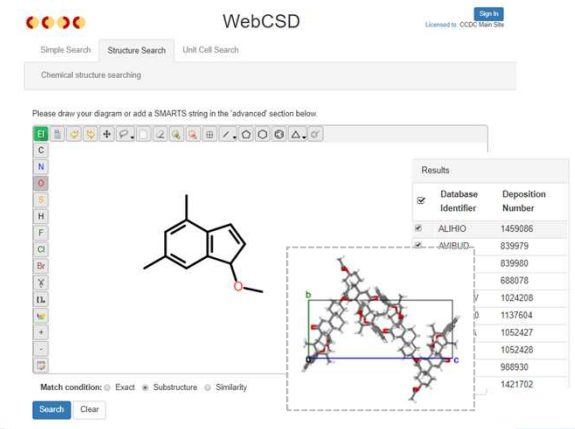
Mercury: Visualisation & data analysis



Mogul: Molecular geometry analysis



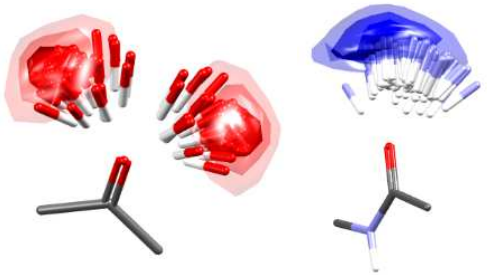
WebCSD: On-line portal to the CSD



CSD Python API: Custom search and analysis

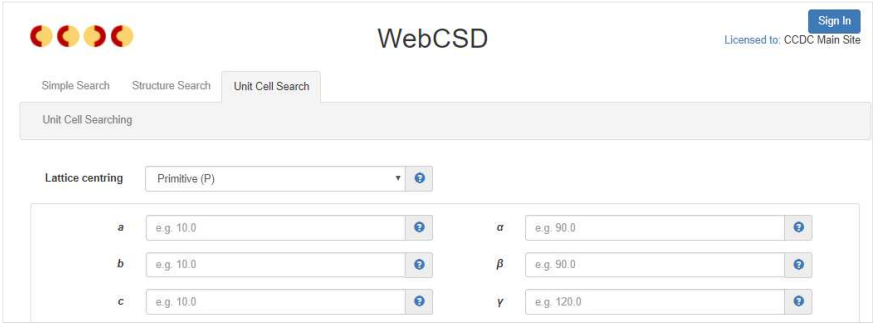
```python
import sys

from ccdc import conformer
from ccdc import io

args = parser.parse_args()

mol_reader = io.MoleculeReader(args.inmolfn)
engine = conformer.GeometryAnalyser()

molecules = []
min_unusual_torsions = sys.maxint
for (idx, molecule) in enumerate(mol_reader):
    molecule.standardise_aromatic_bonds()
    molecule.standardise_delocalised_bonds()

    # Do the analysis
    geometry_analysed_molecule = engine.analyse_molecule(molecule)

    # Count number of unusual torsions
    molecule.unusual_torsions = [
        t for t in geometry_analysed_molecule.analysed_torsions
          if t.unusual and t.enough_hits]
    num_unusual_torsions = len(molecule.unusual_torsions)
    molecule.num_unusual_torsions = num_unusual_torsions
    molecules.append(molecule)

    if num_unusual_torsions < min_unusual_torsions:
        min_unusual_torsions = num_unusual_torsions
```

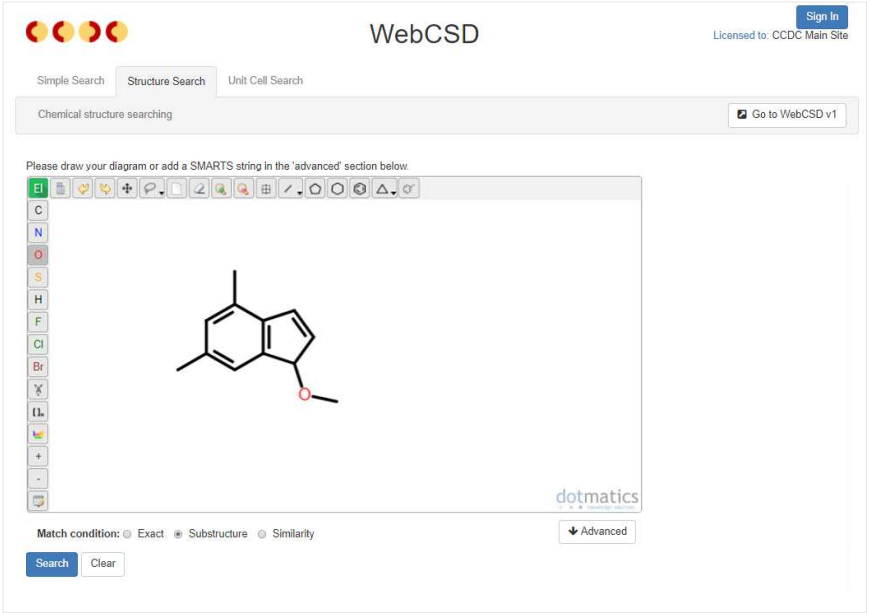
IsoStar: Molecular interaction analysis



New WebCSD launched in 2017

- Unit cell search
- 2D structure search
  - exact
  - substructure
  - similarity
- Additional text/numeric searches
- Display of full CSD entry

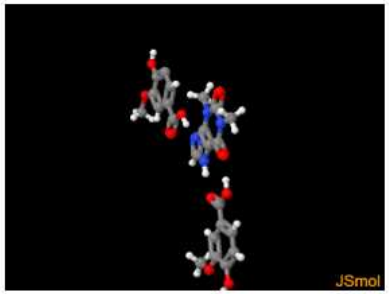




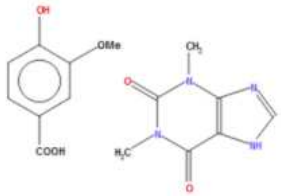
WebCSD - More information visible to users

ZOYBIA : bis(4-Hydroxy-3-methoxybenzoic acid) 1,3-dimethyl-3,7-dihydro-1H-purine-2,6-dione
Space Group: P21/c, **Cell:** *a* 11.2741(6)Å *b* 15.7712(10)Å *c* 13.3746(8)Å, *α* 90.00° *β* 90.572(2)° *γ* 90.00°

3D viewer



Chemical diagram



View group symbols key

Additional CCDC details

CCDC Number	1022107
CCDC Citation	Ayesha Jacobs, Francoise M. Amombo Noa CCDC 1022107: Experimental Crystal Structure Determination, 2014, DOI: 10.5517/cc139l66
Synonyms	bis(Vanillic acid) theophylline
Deposited on	30/08/2014

Associated publications

Ayesha Jacobs, Francoise M. Amombo Noa, *CrystEngComm*, 2015, 17, 98, DOI: 10.1039/C4CE01795A

Chemical details

Formula	2(C8 H8 O4),C7 H8 N4 O2

Crystal details

Space group	P21/c
Unit cell	*a* 11.2741(6)Å *b* 15.7712(10)Å *c* 13.3746(8)Å *α* 90.00° *β* 90.572(2)° *γ* 90.00°
Reduced cell	*a* 11.2741Å *b* 13.3746Å *c* 15.7712Å *α* 90.0000° *β* 90.0000° *γ* 90.5720°
Z, Z'	4.00, 1.00
Habit	rectangle
Recrystallisation solvent	water
Disorder	C7A,O4A and C9A,O5A disordered over two sites with occupancies 0.812:0.188.
Colour	colorless

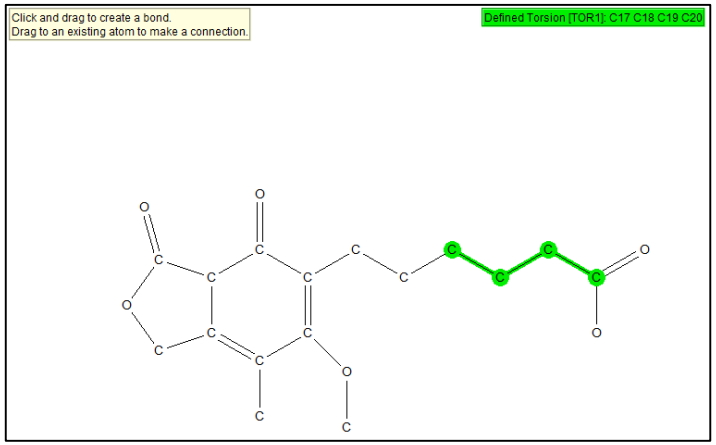
Experimental details

R-factor (%)	4.7
Temperature (K)	173.00
Density (CCDC)	1.44257

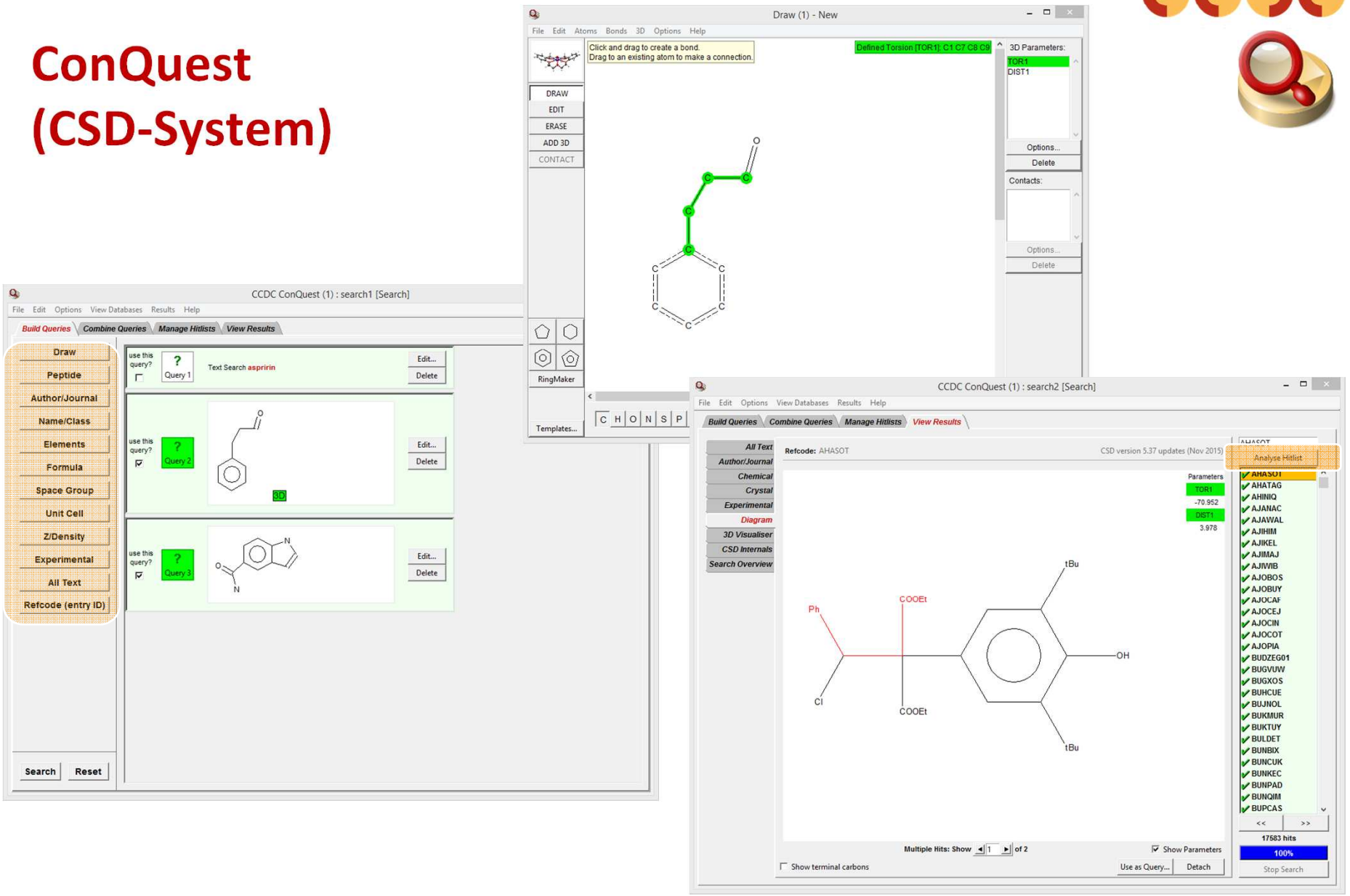
Questions

- *Is my molecule, or something like it, in the CSD?*
- *Are there any structures of aspirin in the CSD?*
- *What values are seen for this type of torsion angle?*
- *What is the typical intermolecular contact distance between these two groups?*

→ Use ConQuest (CSD-System)



ConQuest (CSD-System)

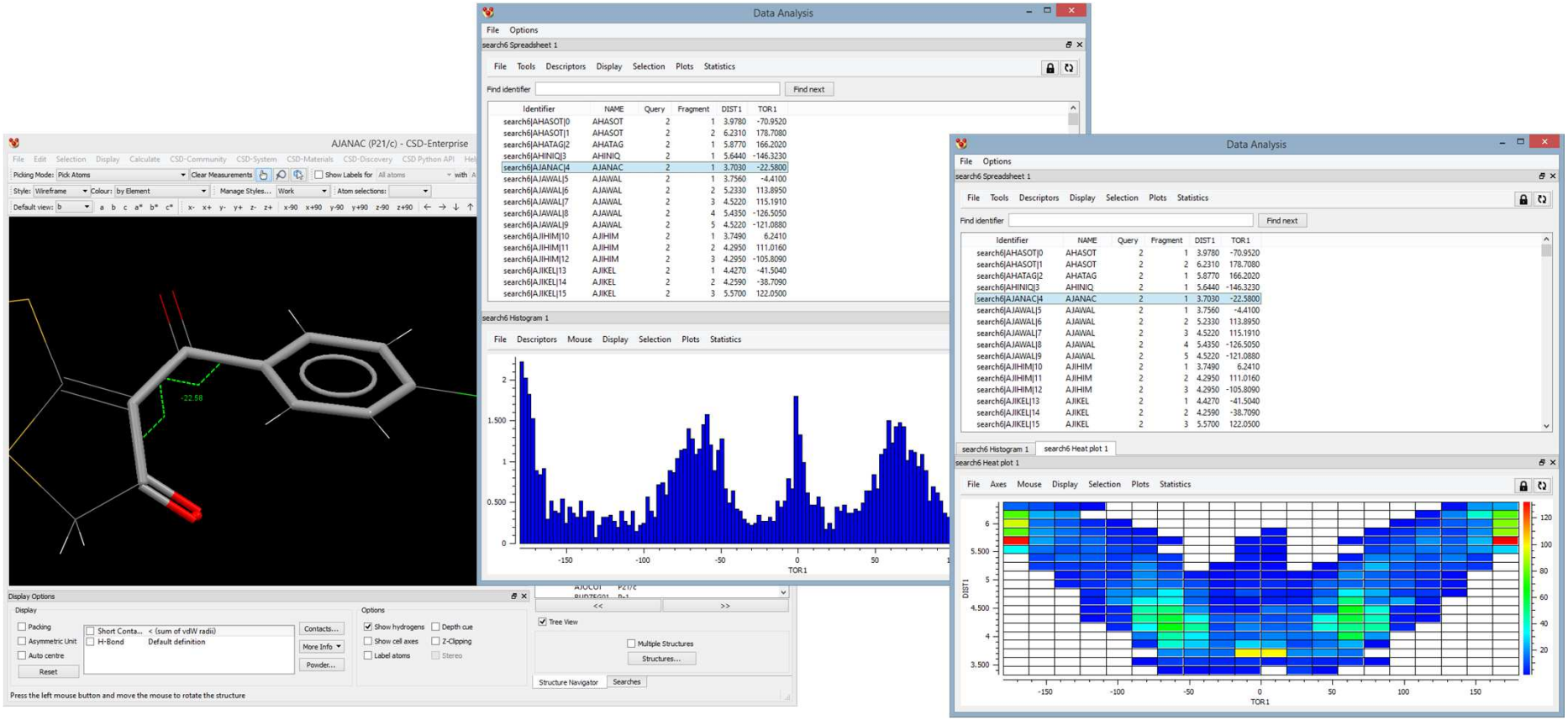


Hitlists can be exported in sdf, mol2, pdb, cif, plus additional formats.

Question

- *How can I analyse data produced by ConQuest searches?*

→ Use Mercury (Data Analysis Module)

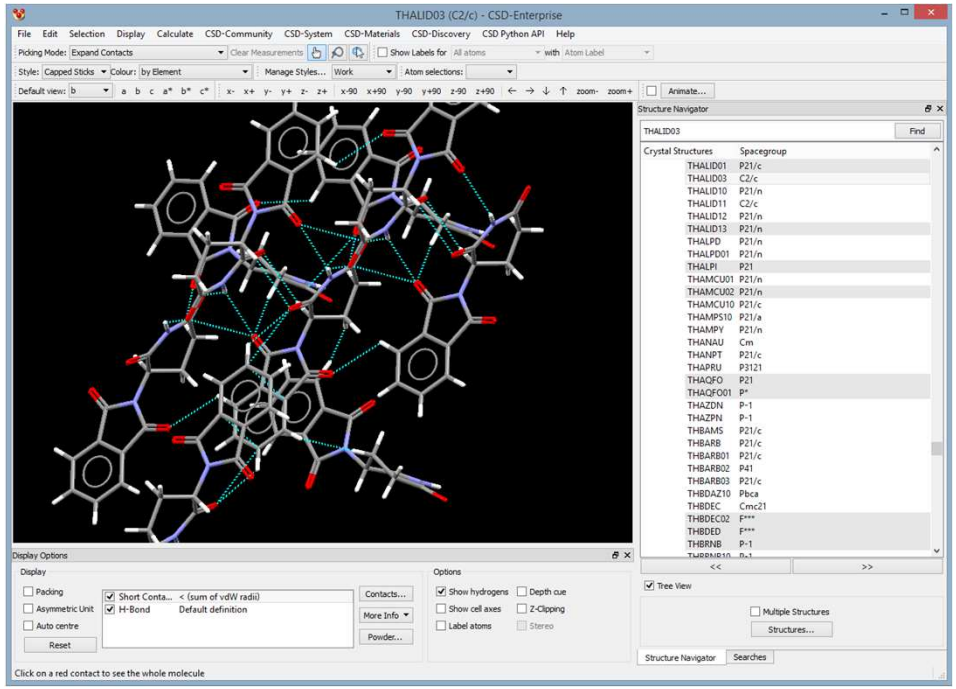


Can save the data to a csv file for analysis in other tools (Excel, Spotfire, StarDrop).

Questions

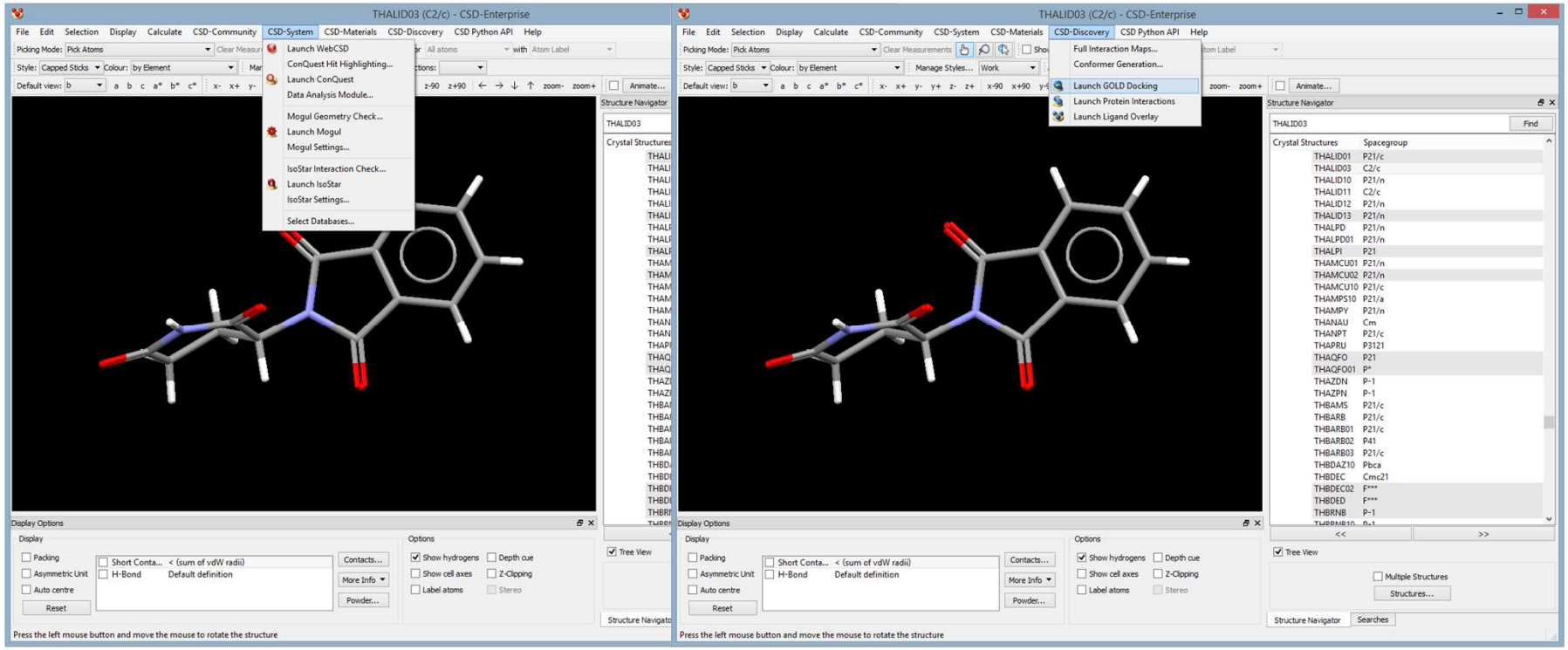
- *What is the 3D structure of my molecule?*
- *What are the crystal contacts of my molecule?*
- *What is the crystal packing of my molecule?*

→ Use Mercury (CSD-System)



More about Mercury

- Overall interface to small molecule visualisation & analysis
- Interface to other CCDC tools
- Can generate high-quality graphics (POV-Ray integration)

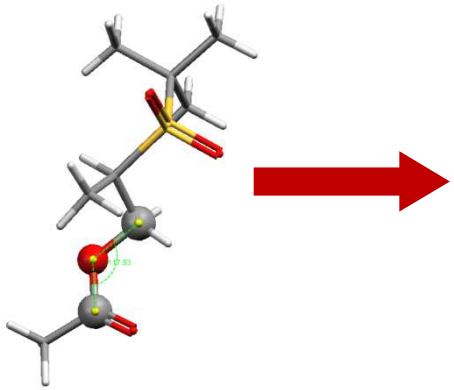


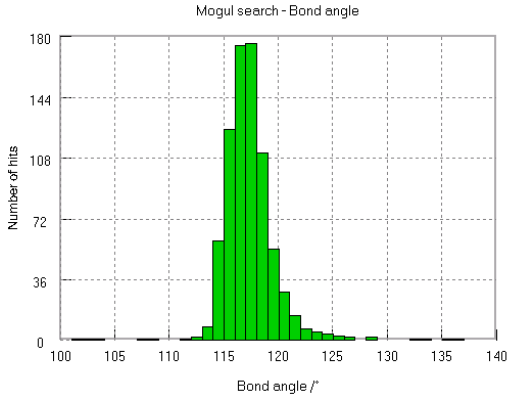
Questions

- *Is the geometry of my molecule typical of the 900,000+ structures in the CSD?*
- *Are the geometries of my modeled compounds realistic?*
- *What is the typical geometry of my molecular idea?*
- *Is the geometry of my docking pose reasonable?*

→ Use Mogul (CSD-System)

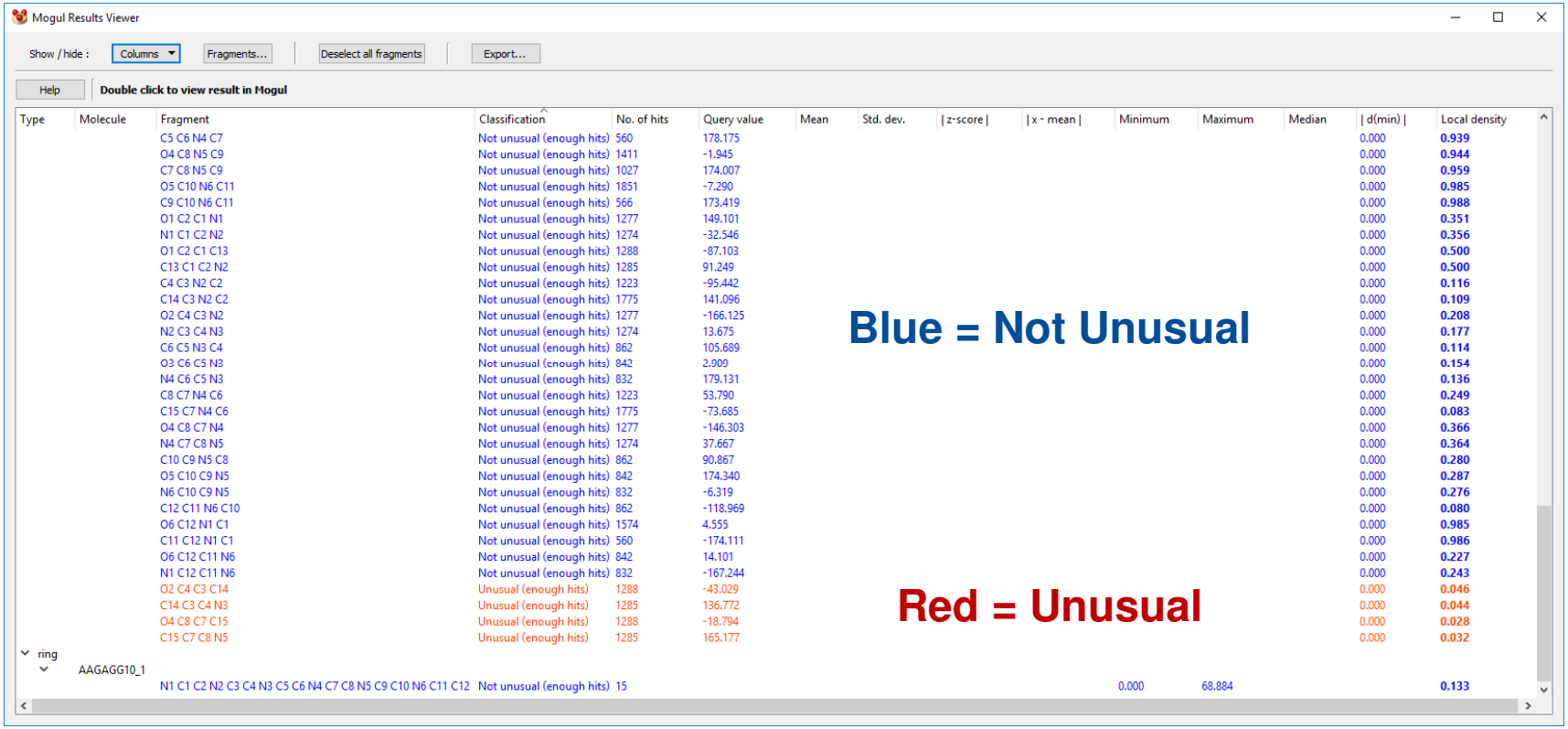
Pre-computed knowledge base of intramolecular geometry information in organics





Mogul (CSD-System)

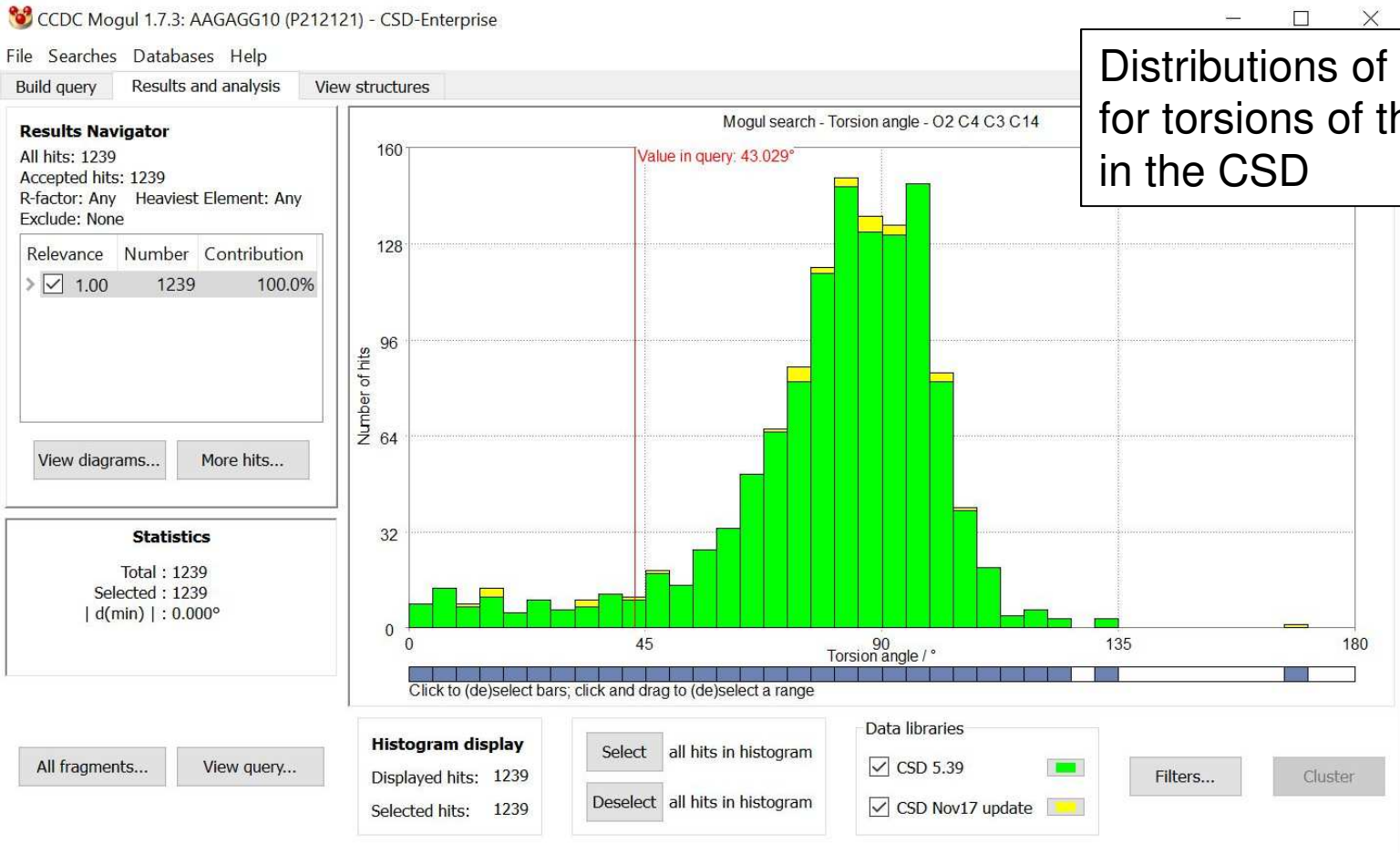
Specify geometric parameter to search on, or validates complete geometry of molecule
Classifies each bond length, valence angle, torsion angle, and ring as usual or “unusual”



Type	Molecule	Fragment	Classification	No. of hits	Query value	Mean	Std. dev.	| z-score |	| x - mean |	Minimum	Maximum	Median	| d(min) |	Local density
		C5 C6 N4 C7	Not unusual (enough hits)	560	178.175								0.000	0.939
		O4 C8 N5 C9	Not unusual (enough hits)	1411	-1.945								0.000	0.944
		C7 C8 N5 C9	Not unusual (enough hits)	1027	174.007								0.000	0.959
		O5 C10 N6 C11	Not unusual (enough hits)	1851	-7.290								0.000	0.985
		C9 C10 N6 C11	Not unusual (enough hits)	566	173.419								0.000	0.988
		O1 C2 C1 N1	Not unusual (enough hits)	1277	149.101								0.000	0.351
		N1 C1 C2 N2	Not unusual (enough hits)	1274	-32.546								0.000	0.356
		O1 C2 C1 C13	Not unusual (enough hits)	1288	-87.103								0.000	0.500
		C13 C1 C2 N2	Not unusual (enough hits)	1285	91.249								0.000	0.500
		C4 C3 N2 C2	Not unusual (enough hits)	1223	-95.442								0.000	0.116
		C14 C3 N2 C2	Not unusual (enough hits)	1775	141.096								0.000	0.109
		O2 C4 C3 N2	Not unusual (enough hits)	1277	-166.125								0.000	0.208
		N2 C3 C4 N3	Not unusual (enough hits)	1274	13.675								0.000	0.177
		C6 C5 N3 C4	Not unusual (enough hits)	862	105.689								0.000	0.114
		O3 C6 C5 N3	Not unusual (enough hits)	842	2.909								0.000	0.154
		N4 C6 C5 N3	Not unusual (enough hits)	832	179.131								0.000	0.136
		C8 C7 N4 C6	Not unusual (enough hits)	1223	53.790								0.000	0.249
		C15 C7 N4 C6	Not unusual (enough hits)	1775	-73.685								0.000	0.083
		O4 C8 C7 N4	Not unusual (enough hits)	1277	-146.303								0.000	0.366
		N4 C7 C8 N5	Not unusual (enough hits)	1274	37.667								0.000	0.364
		C10 C9 N5 C8	Not unusual (enough hits)	862	90.867								0.000	0.280
		O5 C10 C9 N5	Not unusual (enough hits)	842	174.340								0.000	0.287
		N6 C10 C9 N5	Not unusual (enough hits)	832	-6.319								0.000	0.276
		C12 C11 N6 C10	Not unusual (enough hits)	862	-118.969								0.000	0.080
		O6 C12 N1 C1	Not unusual (enough hits)	1574	4.555								0.000	0.985
		C11 C12 N1 C1	Not unusual (enough hits)	560	-174.111								0.000	0.986
		O6 C12 C11 N6	Not unusual (enough hits)	842	14.101								0.000	0.227
		N1 C12 C11 N6	Not unusual (enough hits)	832	-167.244								0.000	0.243
		O2 C4 C3 C14	Unusual (enough hits)	1288	-43.029								0.000	0.046
		C14 C3 C4 N3	Unusual (enough hits)	1285	136.772								0.000	0.044
		O4 C8 C7 C15	Unusual (enough hits)	1288	-18.794								0.000	0.028
		C15 C7 C8 N5	Unusual (enough hits)	1285	165.177								0.000	0.032
ring														
	AAGAGG10_1													
		N1 C1 C2 N2 C3 C4 N3 C5 C6 N4 C7 C8 N5 C9 C10 N6 C11 C12	Not unusual (enough hits)	15						0.000	68.884			0.133

Double-click unusual torsion to examine in detail

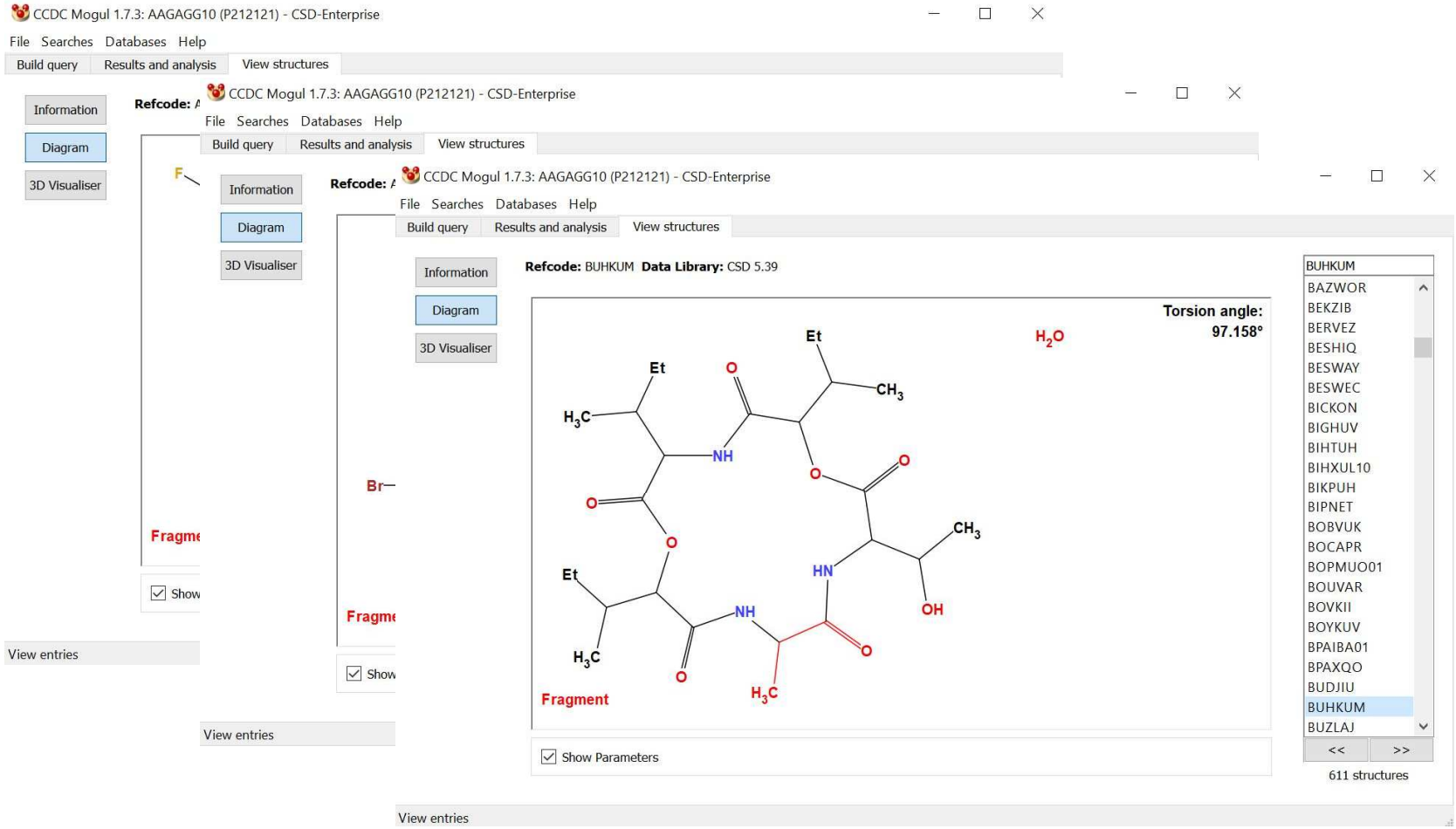
Mogul (CSD-System)



Distributions of angles for torsions of this type in the CSD

Mogul (CSD-System)

You can examine the matched torsions

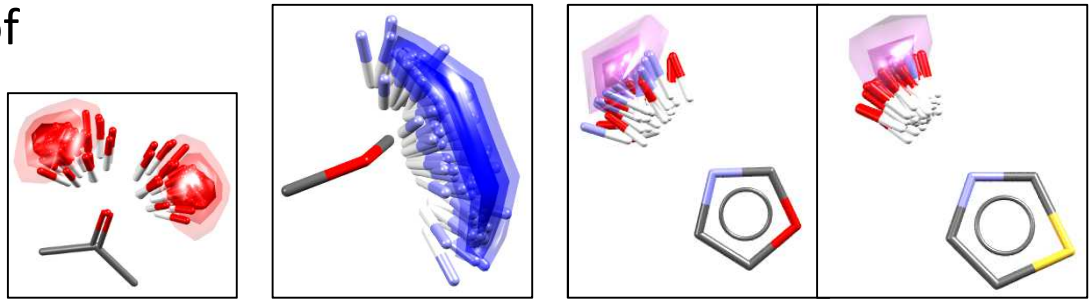


Questions

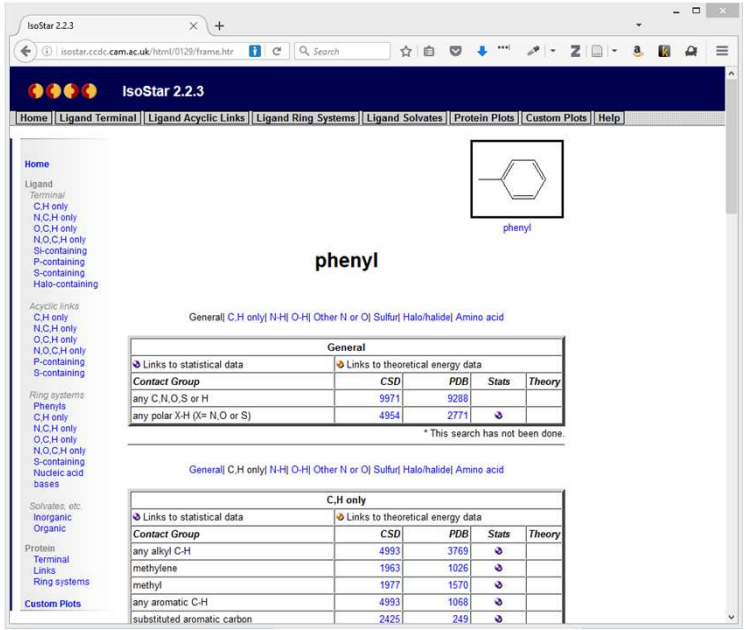
- *How do **particular** groups typically interact in structures in the CSD?* (e.g. is the N or the O of oxazole more likely to H-bond to an OH group?)
- *How do **particular** ligand groups interact with particular protein residues?* (e.g. how can we bind to a tryptophan?)
- *What particular group can be a bioisostere for another particular group?* (e.g. do thiazole and oxazole form similar H-bond interactions?)

→ Use IsoStar (CSD-System)

Pre-computed knowledge base of intermolecular interactions



IsoStar: A library of information about nonbonded interactions
Bruno *et al.*, *J. Comput.-Aided Mol. Des.*, (1997), **11**, 525-537, DOI: 10.1023/A:1007934413448



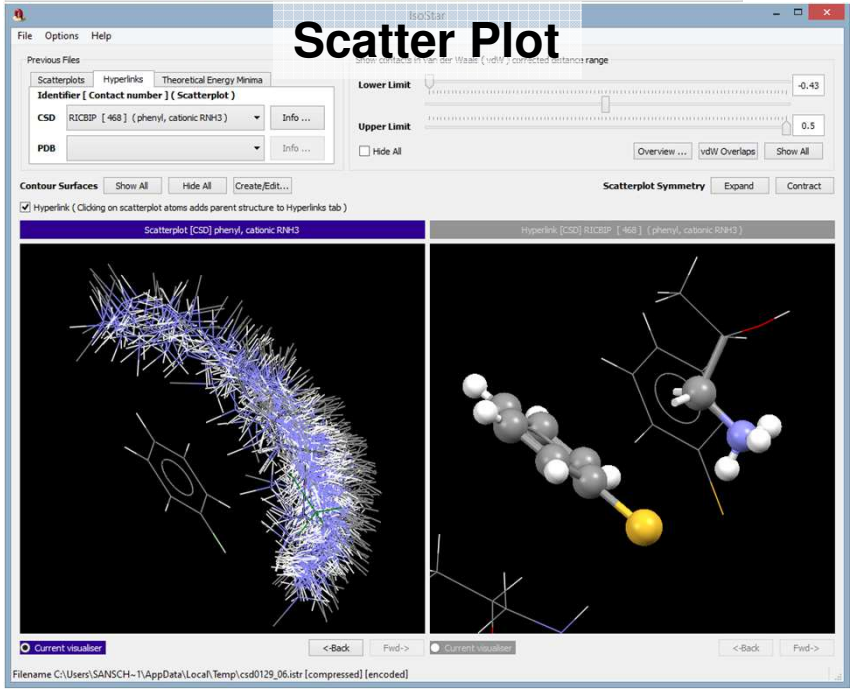
IsoStar (CSD-System)

Client-Server system: http://isostar.ccdc.cam.ac.uk
Select Central Group (Phenyl)
Select Contact Group (R-NH3+)
Select CSD or PDB data

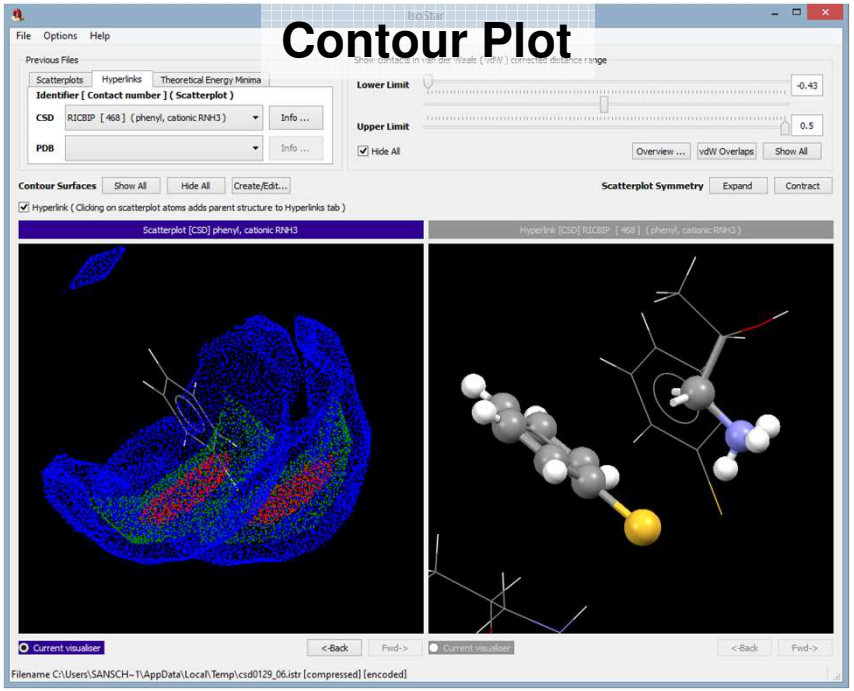
View Scatter plot or Contour plot

- In both cases, adjusted for symmetry

Scatter Plot



Contour Plot

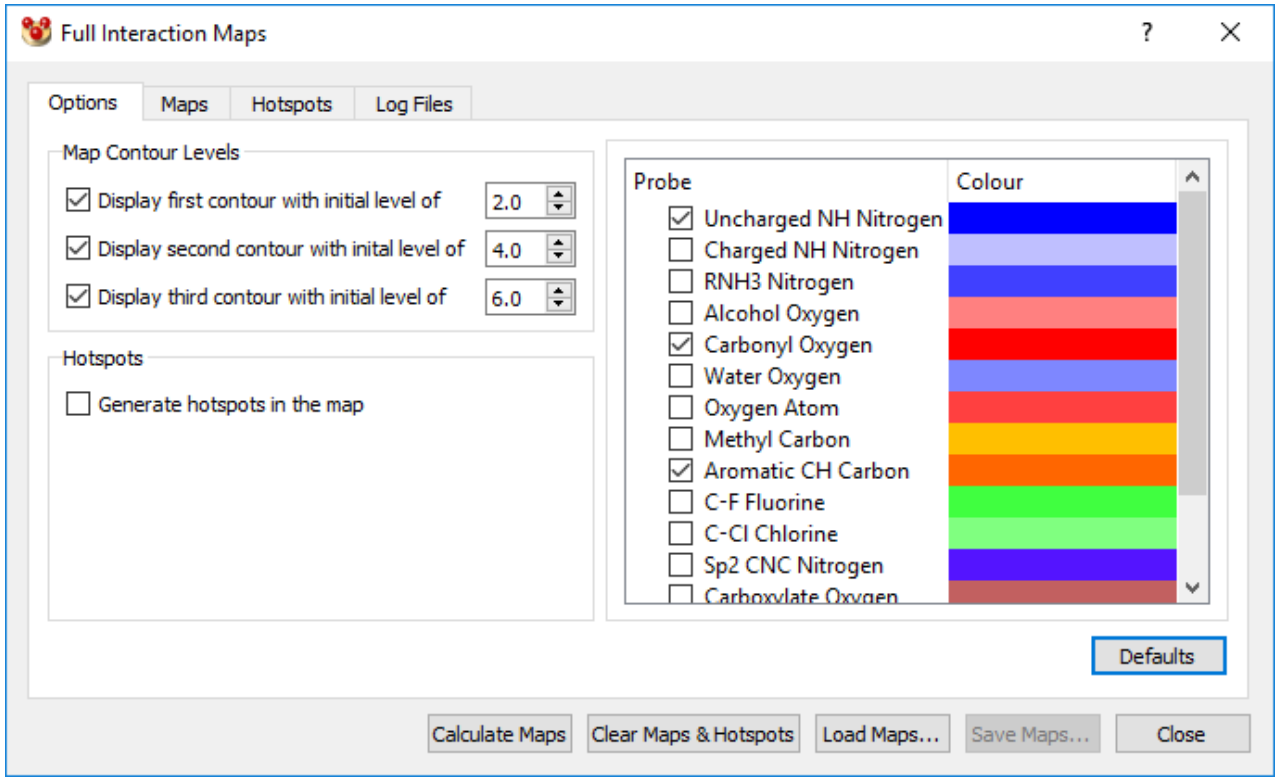


Also: 1,550 theoretical energy minima (IMPT); IsoGen to create custom plots

Question

- *What are the common interactions around my molecule **taken as a whole**?*

Full Interaction Maps (CSD-Discovery/Materials)

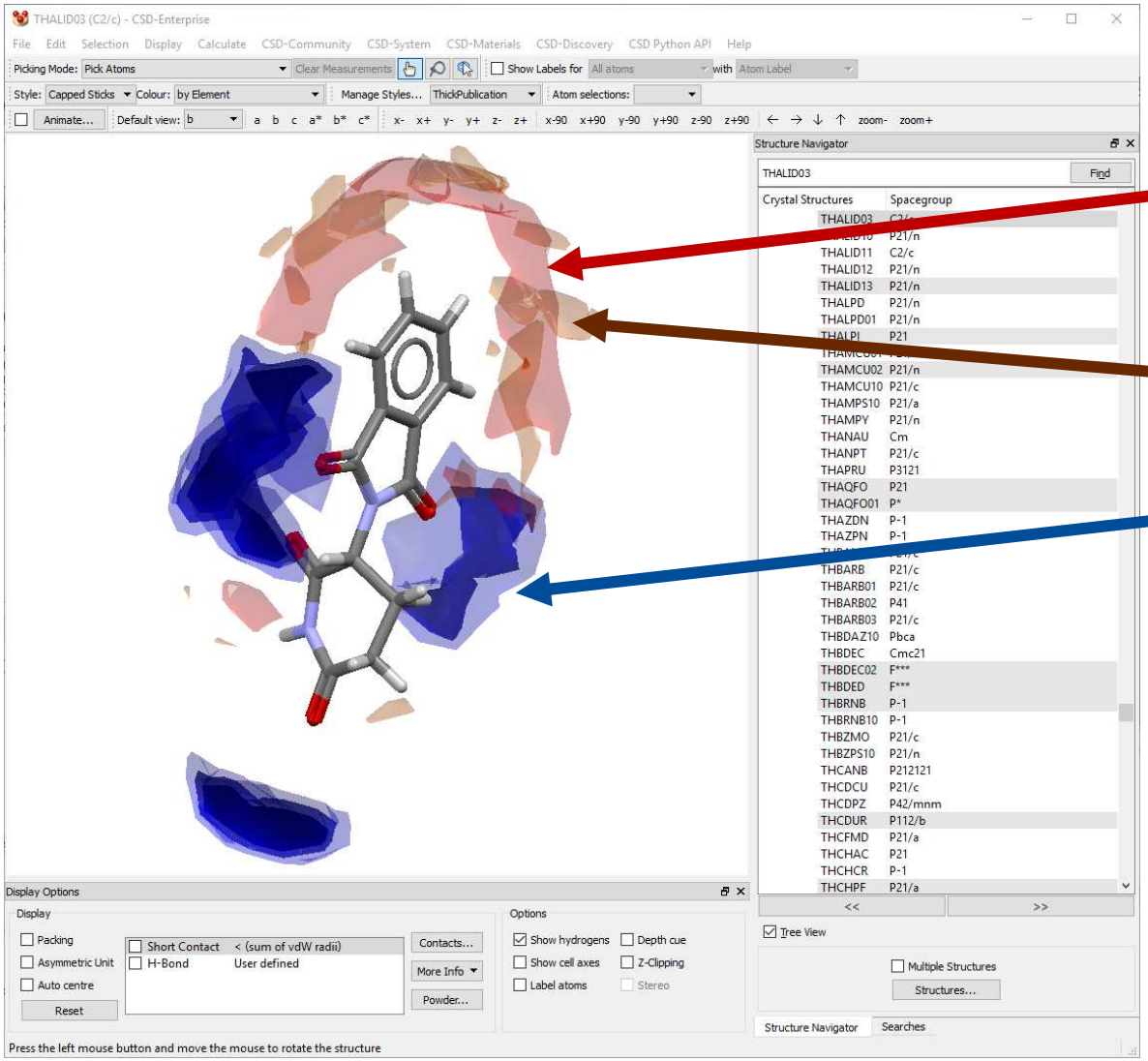


Uses CSD data from IsoStar Accessed in Mercury.

For currently loaded entry:

- Select Full Interaction Maps in CSD-Discovery or CSD-Materials menu.
- Retain defaults
- Click Calculate Maps.

Full Interaction Maps (CSD-Discovery/Materials)

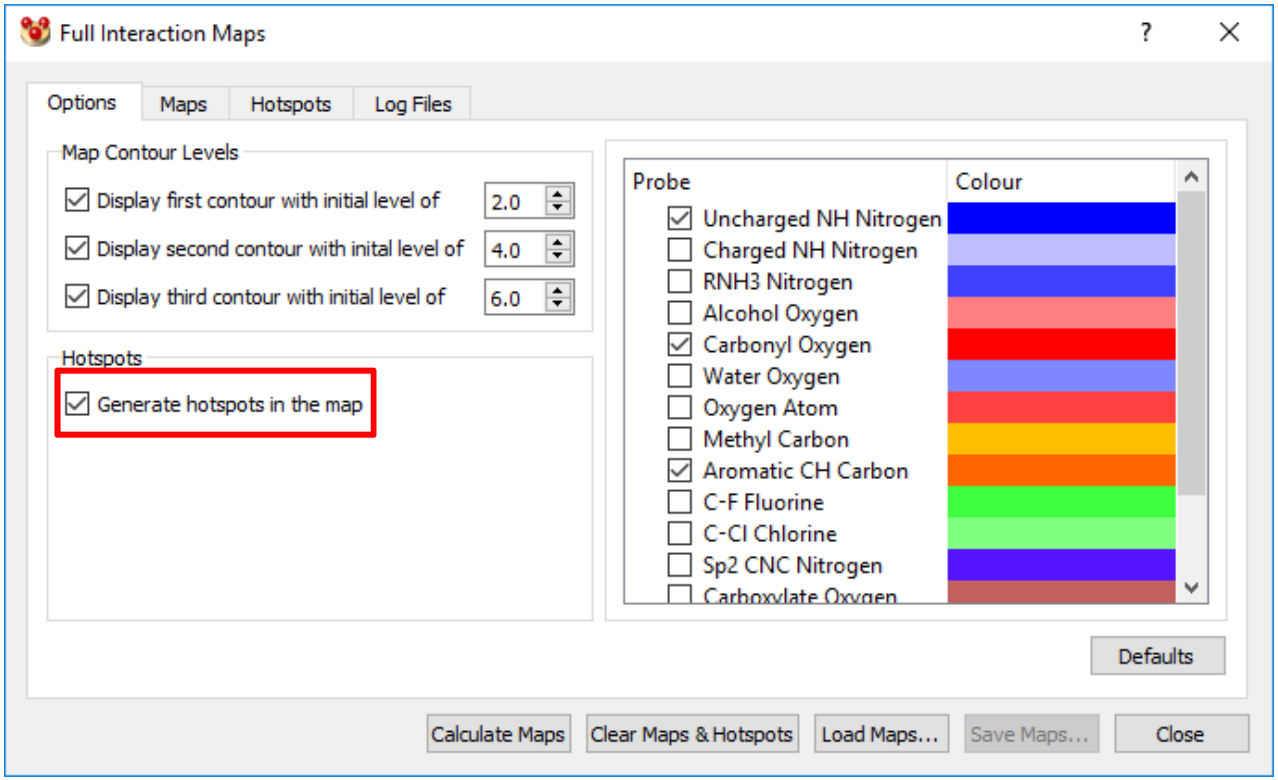


Red = frequently observed H-bonding acceptors

Brown = frequently observed hydrophobic (bit hard to see)

Blue = frequently observed H-bonding donors

Full Interaction Maps (CSD-Discovery/Materials)

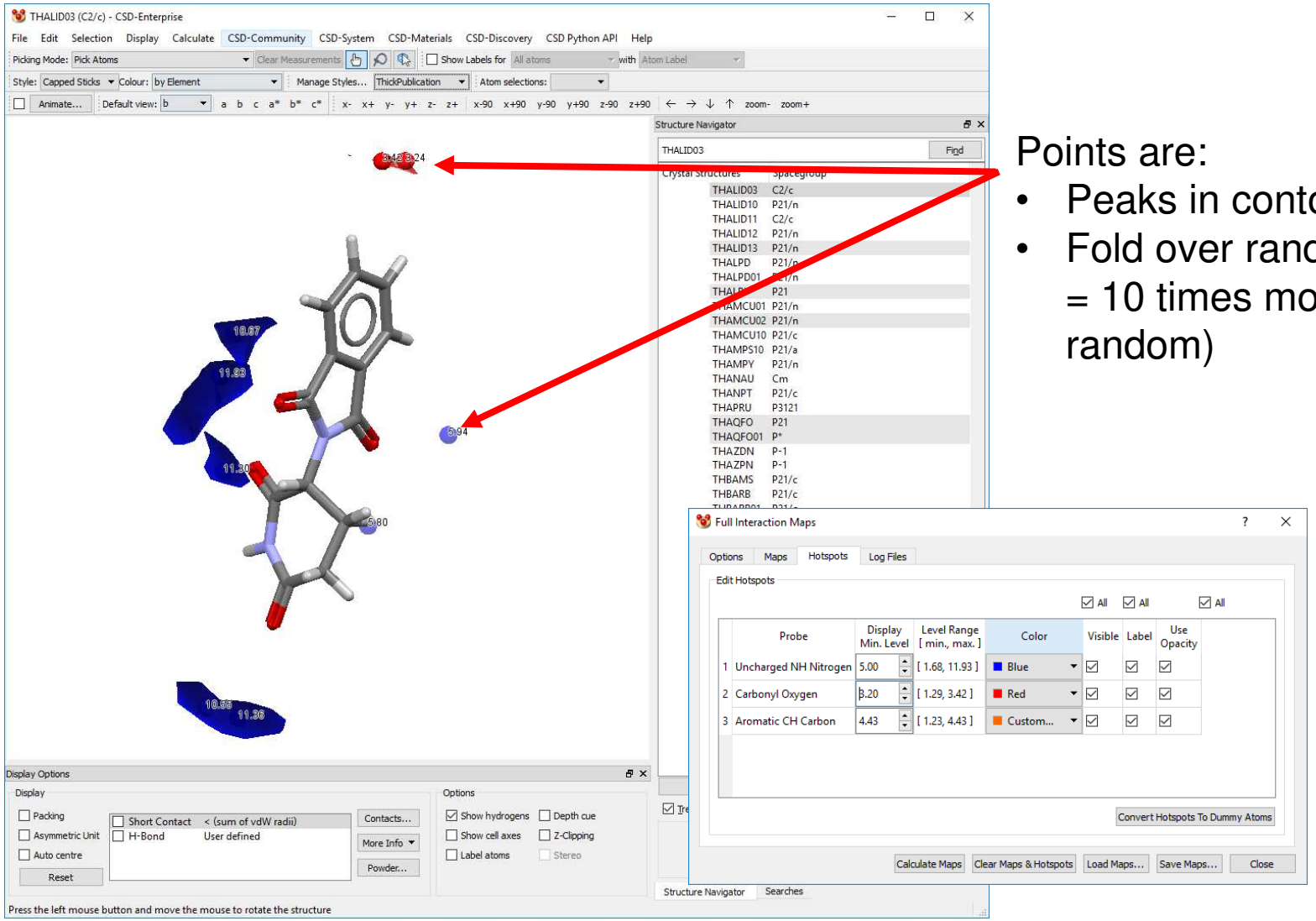


Uses CSD data from IsoStar.

For currently loaded entry:

- Select Full Interaction Maps in CSD-Discovery or CSD-Materials menu.
- Turn on “Generate hotspots”.
- Click Calculate Maps.

Full Interaction Maps (CSD-Discovery/Materials)



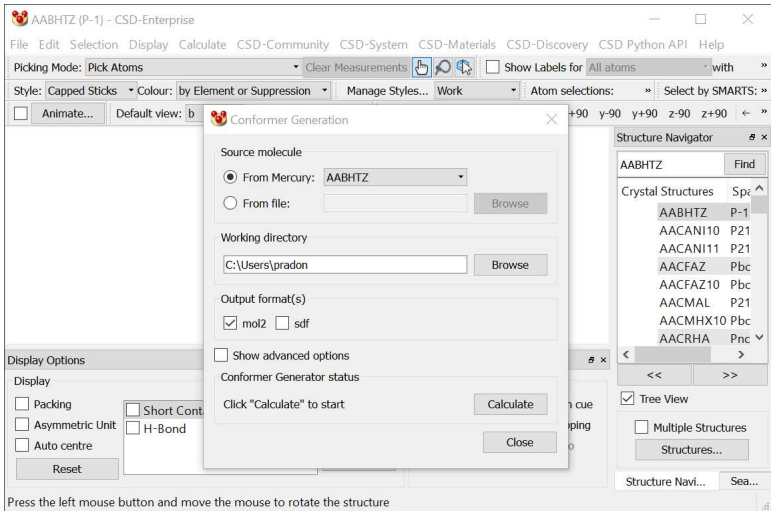
Points are:

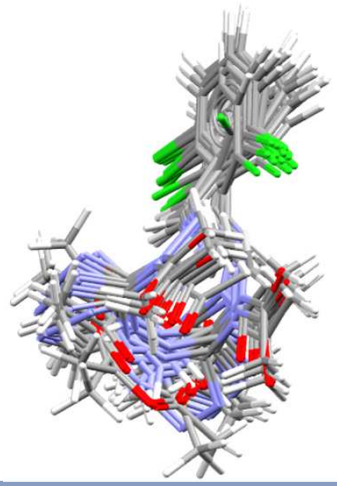
- Peaks in contours
- Fold over random (e.g., 10 = 10 times more likely than random)

Question

- *What are the most likely diverse conformations of my molecule based on the CSD?*

CSD-Conformer Generator (CSD-Discovery/Materials)





Uses CSD data from Mogul Accessed in Mercury or Hermes, or through the command line, or through the CSD Python API.

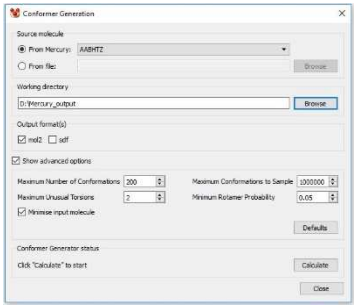
In Mercury, for currently loaded entry:

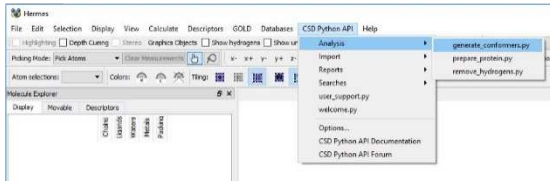
- Select Conformer Generation in CSD-Discovery or CSD-Materials menu.
- Retain defaults
- Click Calculate.

R. Taylor *et al.* (2014) *J. Chem. Inf. Model.* **54**, 2500-2514
J.C. Cole *et al.* (2018) *J. Chem. Inf. Model.* **58**, 615-629

CSD-Conformer Generator

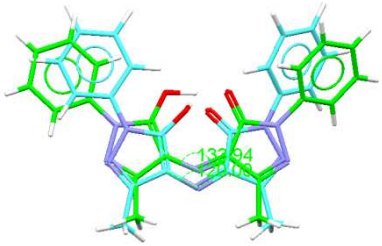
- Part of CSD-Discovery, CSD-Materials and CSD-Enterprise suites

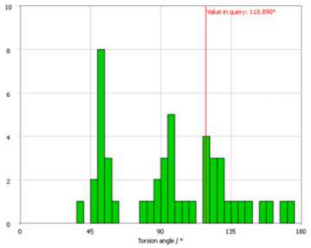


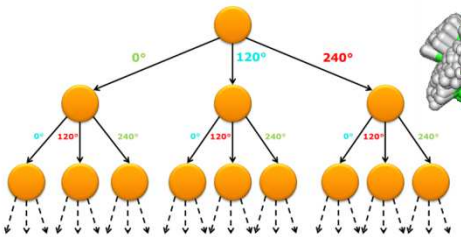


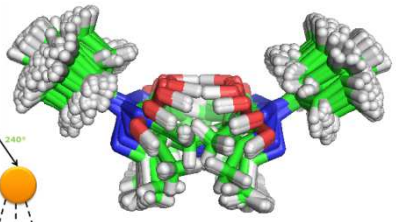
- Program based entirely on Mogul data for conformer generation
- New version: improved handling of flexible ring systems
  - using Mogul templates for isolated, fused, spiro-linked and bridged rings
  - if no template available, then generated on the fly with cyclic bond rotamer distributions
- Fast & chemically plausible conformations
- Initial optimisation step also useful to prepare compound libraries before virtual screening

CSD-Conformer Generator: Methodology

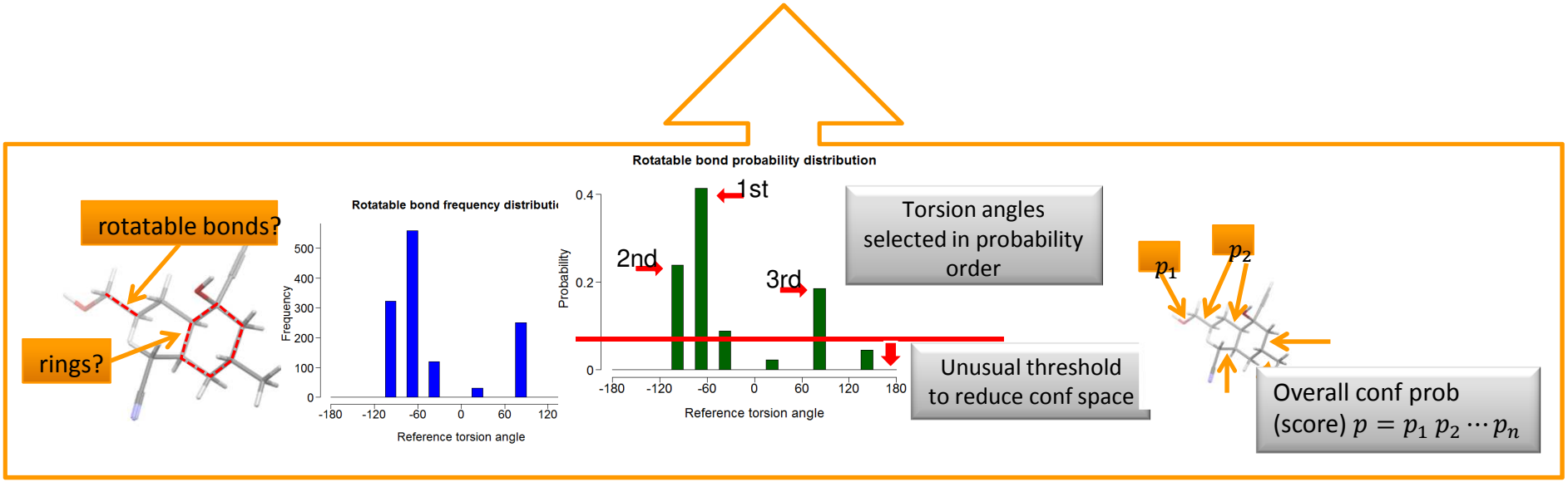








molecule → (optimise) → get rotamer and ring distributions → sample and score conformers → select diverse subset

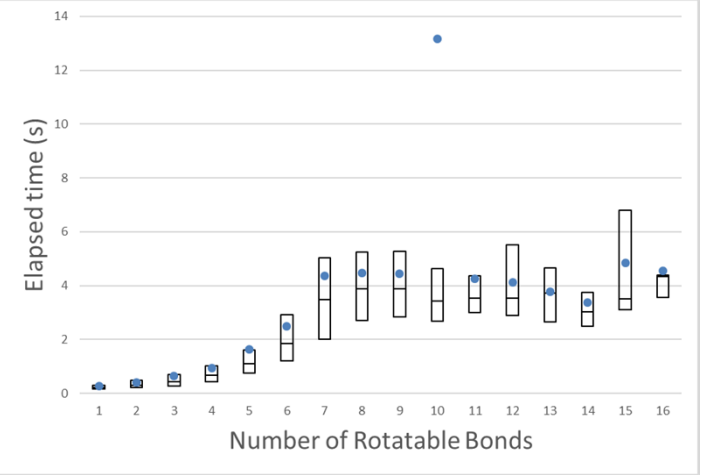


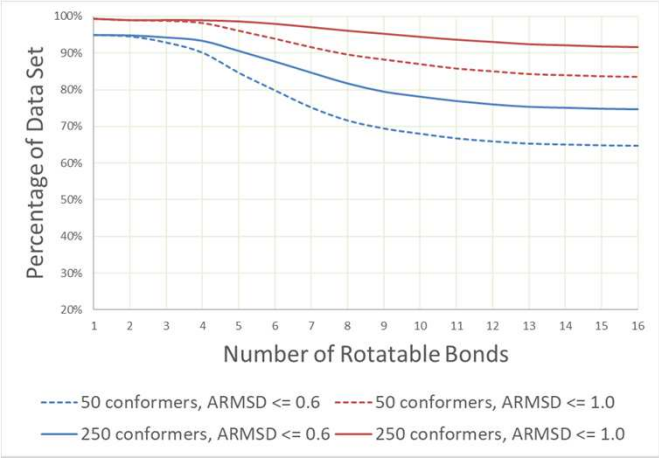
R. Taylor *et al.* (2014) *J. Chem. Inf. Model.* **54**, 2500-2514.

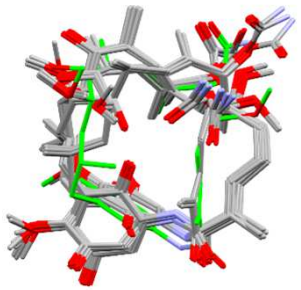
CSD-Conformer Generator: Latest validation

- Performance evaluated against the Platinum [2] diverse data set → get fast & chemically plausible conformations

	Top n conformers		
n	1	50	250
least ARMSD < 0.5 Å (%)	18	57	66
least ARMSD < 1.0 Å (%)	39	84	92
mean ARMSD		0.59	0.48







Geldanamycin, 19 atoms in ring (PDB 3C11): top 50 conformers, least AMRSD = 0.95 Å

[1] "Knowledge-Based Conformer Generation Using the Cambridge Structural Database" J.C. Cole *et al.* (2018) *J. Chem. Inf. Model.* **58**, 615-629, DOI: 10.1021/acs.jcim.7b00697

[2] J. Chem. Inf. Model. 2017, 57, 2719–2728.

CSD Python API: Contents

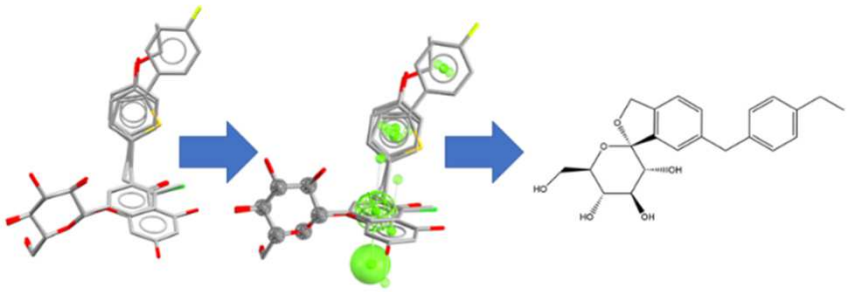
- CSD Python API components:
  - Reading and writing of molecules, crystals and entries
  - Entry, crystal and molecule functions (incl. PXRD output & comparison)
  - Search capabilities (substructure, similarity, text/numeric, reduced cell)
  - Molecular geometry analysis
  - Intermolecular interaction analysis
  - Geometric, molecular & crystallographic descriptor calculation
  - Crystal packing similarity calculation
  - 2D diagram generation
  - 3D conformer generation
  - … much more to come, exposing CSD-Materials and CSD-Discovery features

Questions

- *What other molecules exhibit a similar short-contact or hydrogen-bond pattern?* (easy non-bonded contact search in the CSD: how can we use pharmacophore features to search for a given small molecule-small molecule interaction pattern?)
- *What other scaffolds would fit in my molecule?* (easy scaffold hopping in the CSD: how can we use exit vectors to retrieve diverse ligand topologies that could be used as scaffold?)

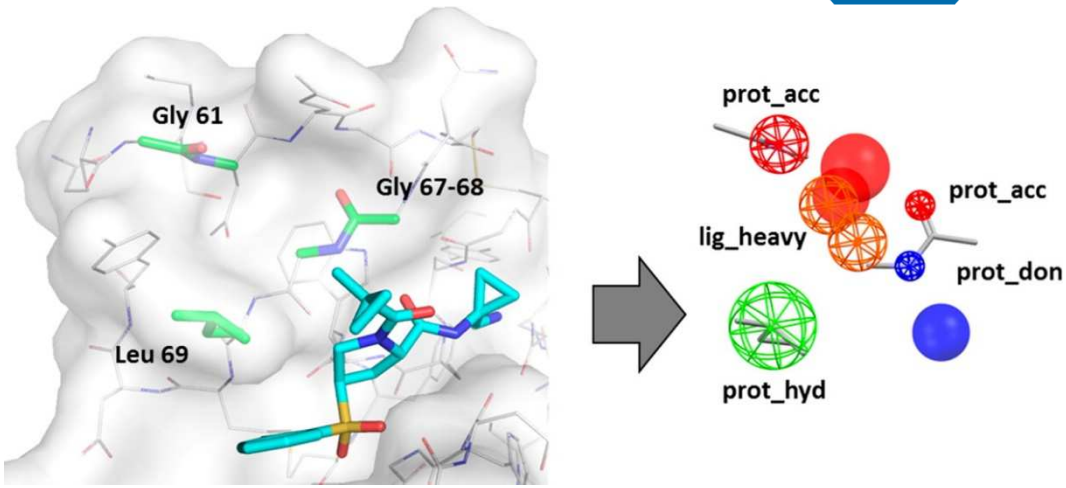
→ Use CSD-CrossMiner (CSD-Discovery)

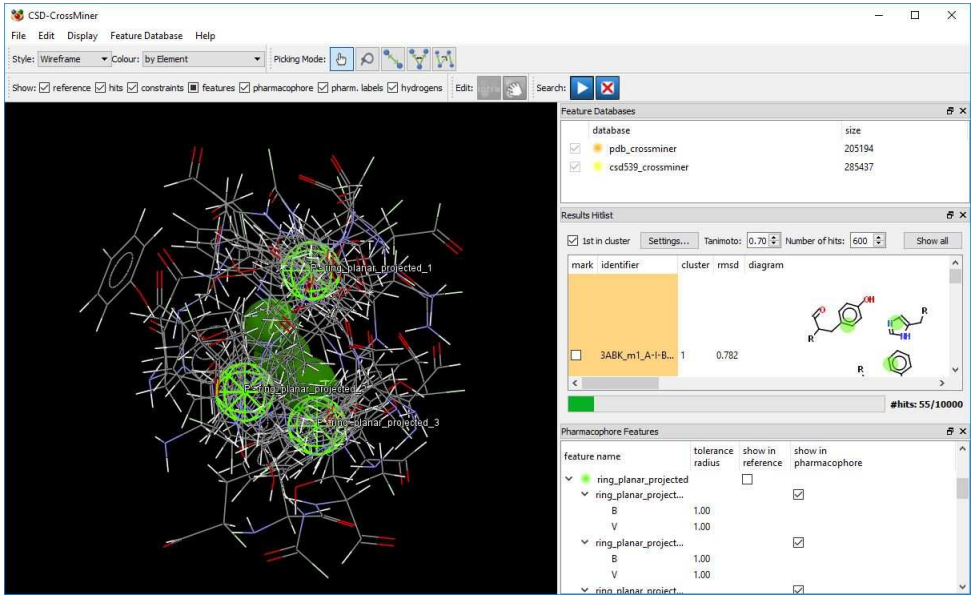
Search structural databases (i.e. the CSD and/or the PDB), concurrently and interactively, in terms of pharmacophore queries



CSD-CrossMiner overview

- Pharmacophore-based searches of structural databases (CSD & PDB, simultaneously)
- Interactive tool: modify a hypothesis/results on the fly
- Applications in drug design:
  - Which structural motifs bind in a similar environment?
  - Which ligand motifs have similar protein interaction patterns?

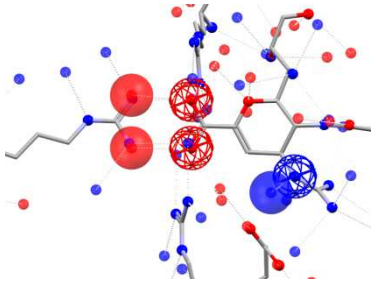




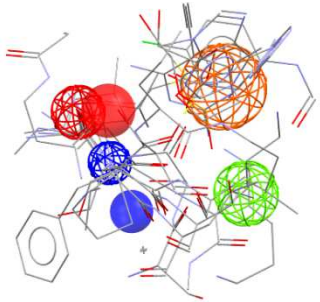
Interactive and versatile navigation of structural databases
O. Korb, *et al.* (2016) *J. Med. Chem.*, **59**, 4257-4266. DOI: 10.1021/acs.jmedchem.5b01756

Our motivation for using pharmacophores

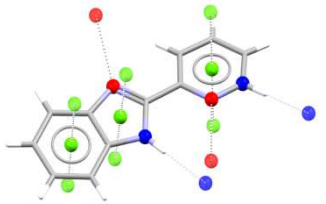
- Intuitive as the functional groups can be specified relative to a 3D structure
  - Sphere positions and radii can easily be refined



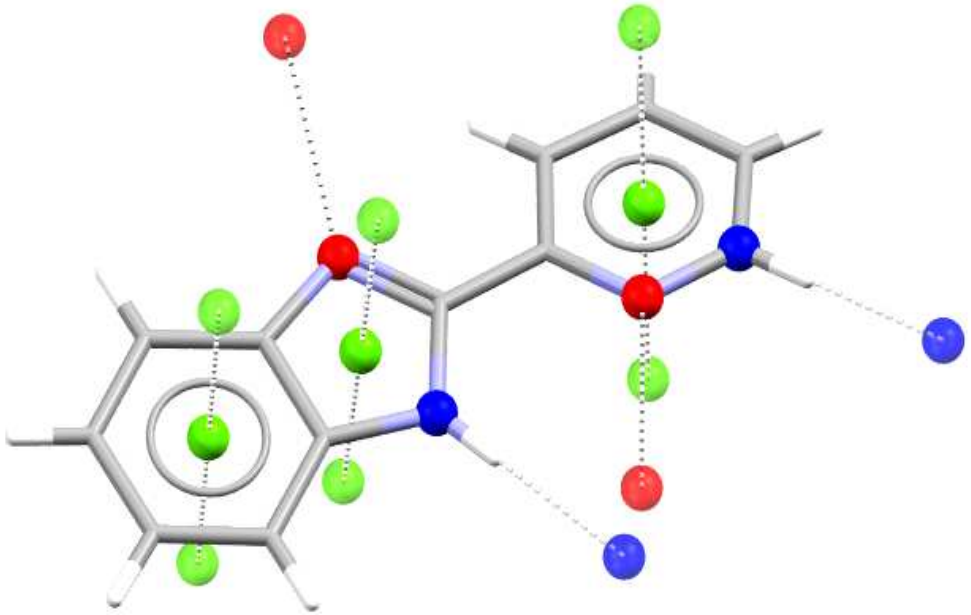
- *"What you see is what you get"*
  - Hits will be overlaid onto the query and can be analysed in the same context



- Concept well-understood by medicinal chemists
- Flexible feature definitions
  - User can control the level of chemical abstraction



Feature assignment



planar ring

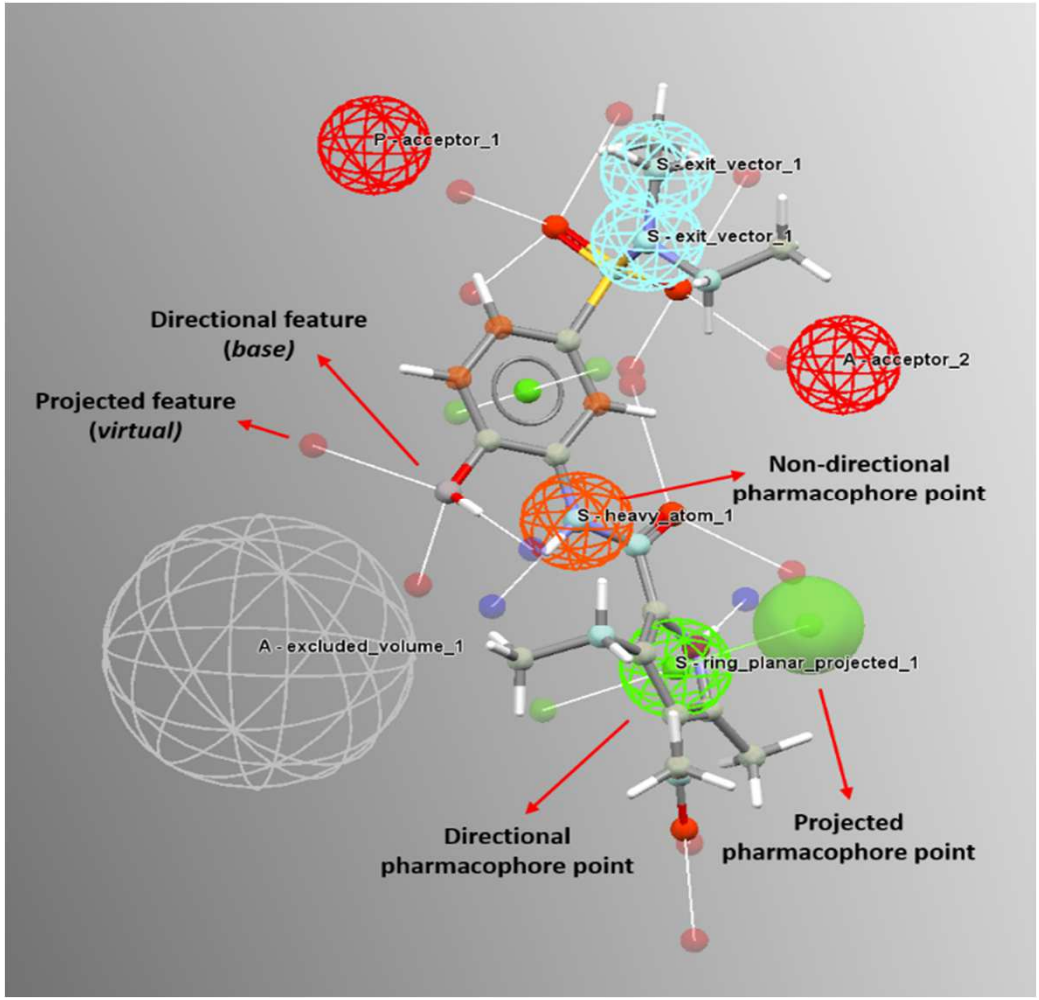
donor

acceptor

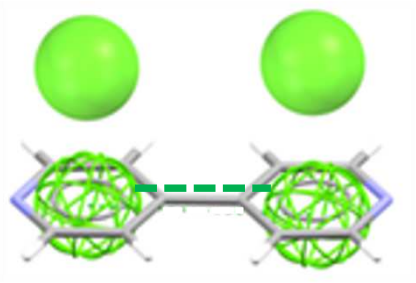
- Features are derived from the substructures by applying a point generation rules.
- Features are based on a series of modifiable SMARTS patterns.
- Each entry in a structural database is indexed with a set of feature definitions which are used to perform the actual 3D search.

Pharmacophores in CSD-CrossMiner

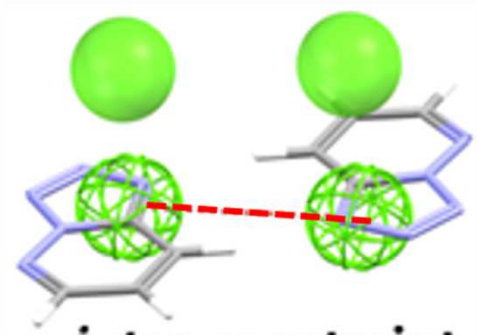
- Molecular structures are annotated with *features*
- Pharmacophore query is based on *tolerance spheres*
  - Sphere radii reflect uncertainty in the position of the features
- Pharmacophore points can be set to be protein (**P**), small molecule (**S**) or any molecule (**A**)



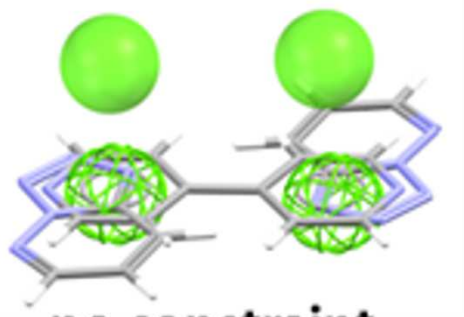
Constraints



intra-constraint



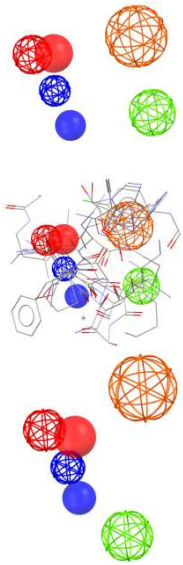
inter-constraint



no constraint

- ***Intra*-constraint** → pharmacophore points located in the same structure
- ***Inter*-constraint** → pharmacophore points located in different structures
- ***No* constraint** → pharmacophore points located either in the same or in different structures

Pharmacophore search



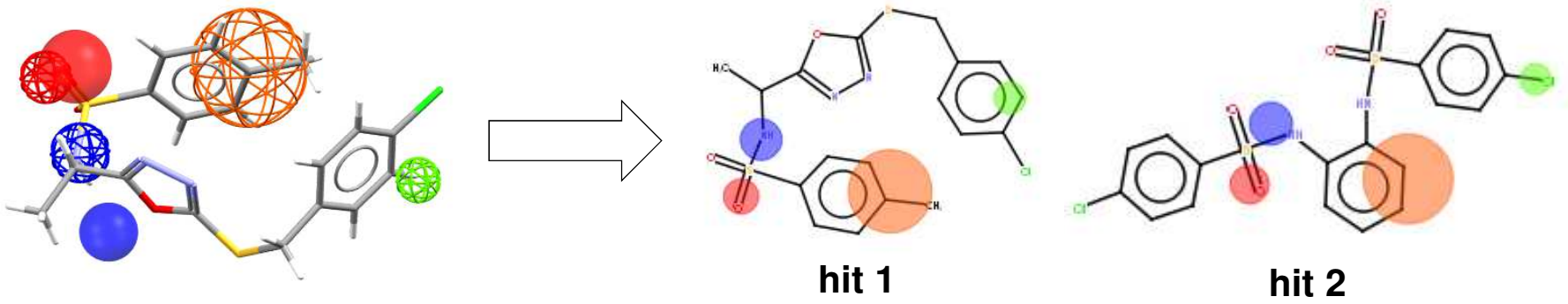
1. generate initial query
2. run search
3. analyse results
4. modify query

can be time-consuming!

Wouldn't it be nice to *interactively* modify the query and analyse the results?

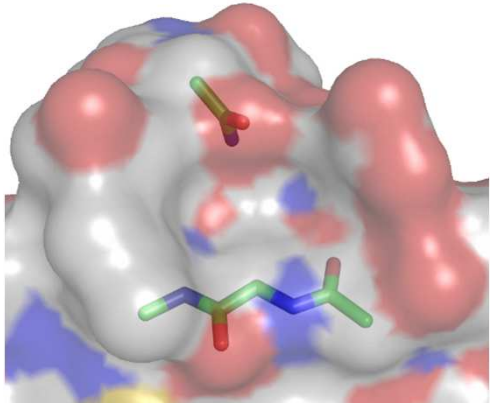
Hit processing

- Rank hits according to Kabsch[1] overlay *rmsd*
  - Overlay feature point to centre of the query sphere
- Hits are clustered according to molecular similarity
  - Only display diverse set of hits with respect to *Tanimoto* threshold (ligand, protein, or ligand+protein)
- 2D diagrams
  - 3D pharmacophore is projected into a 2D representation

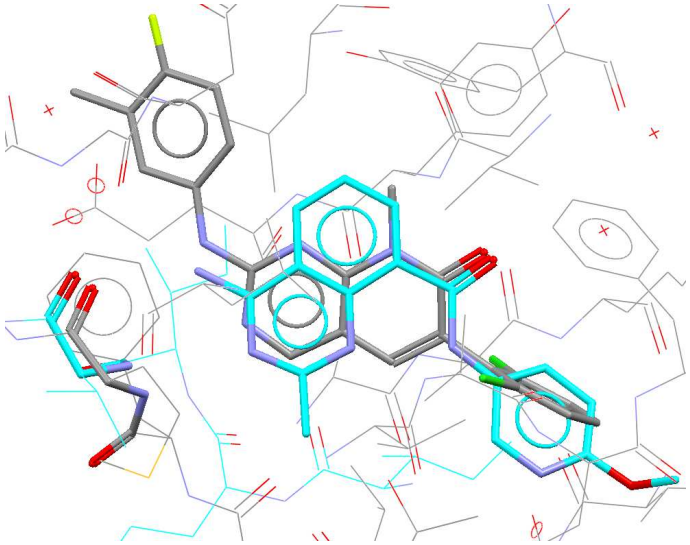


[1]W. Kabsch (1976) *Acta Cryst. A*, ***32***, 922-923. DOI: 10.1107/S0567739476001873

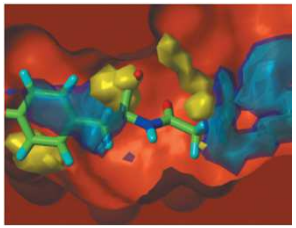
Applications of CSD-CrossMiner



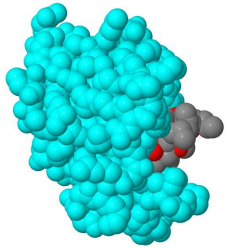
- Determine common protein binding sites in PDB structures
- Determine structural motifs that bind in similar environments
- Inform cross-pharmacology between protein targets
- Find new ideas:
  - Design novel motifs that molecular mimic extablished ligands
  - Scaffold-hopping: retrieve a diversity of ligand topologies that can be used as scaffolds



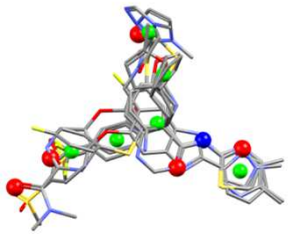
CSD-Discovery: Other applications not covered here



SuperStar: Knowledge-based prediction of intermolecular interactions in a protein binding site based on data from CSD or PDB crystal structures



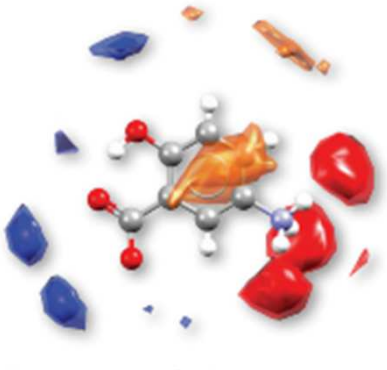
GOLD Suite: Protein-ligand docking - virtual screening, lead optimisation and binding mode prediction, with the **Hermes** protein visualiser

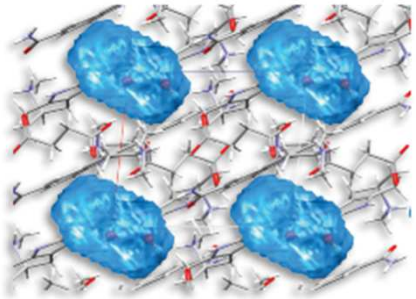


Ligand-Based Virtual Screening Workflow: identify plausible common binding modes of structurally diverse ligands using the **Ligand Overlay Program**, and virtually screen molecules against such a binding mode hypothesis using the **Field-Based Ligand Screener**

CSD-Materials: Other applications not covered here

Informatics-based solutions that aid in the understanding and prediction of solid form stability and properties





- Interpret crystal packing and compare with CSD data using powerful **Packing Feature, Similarity** and **Motif searches**, and **Hydrogen Bond propensity Analysis**
- Understand the effects of hydration on your lattice with the **Hydrate Analyser**
- Explore the structures of potential co-crystals using the **Molecular Complementarity Tool**
- Use CSD data to help solve crystal structures from powder diffraction data using **DASH**

CSD-Enterprise: Summary of what was covered here

- **ConQuest**: CSD search & geometric data generation
- **Mercury**: Visualisation, analysis, overall small molecule interface
- **Mogul**: Geometric analysis
- **IsoStar**: Precomputed functional group-to- functional group interactions
- **Full Interaction Maps**: Interactive whole molecule interactions
- **CSD-Conformer Generator**: Quick generation of chemically plausible conformations
- **CSD Python API**: Scripting, handling of many molecules, integration, conformer generator and much more
- **CSD-CrossMiner**: interactive and simultaneous search of CSD & PDB in terms of pharmacophore queries