

A faint, light gray network pattern of interconnected lines and nodes is visible in the background, resembling a molecular or structural diagram.

ccDC

advancing structural science

What's Up

Customer Update Webinar

24th September 2020



Today's presenters

3



Seth Wiggin

Senior Scientific
Editor



Ilenia Giangreco

Discovery Science Team
Leader



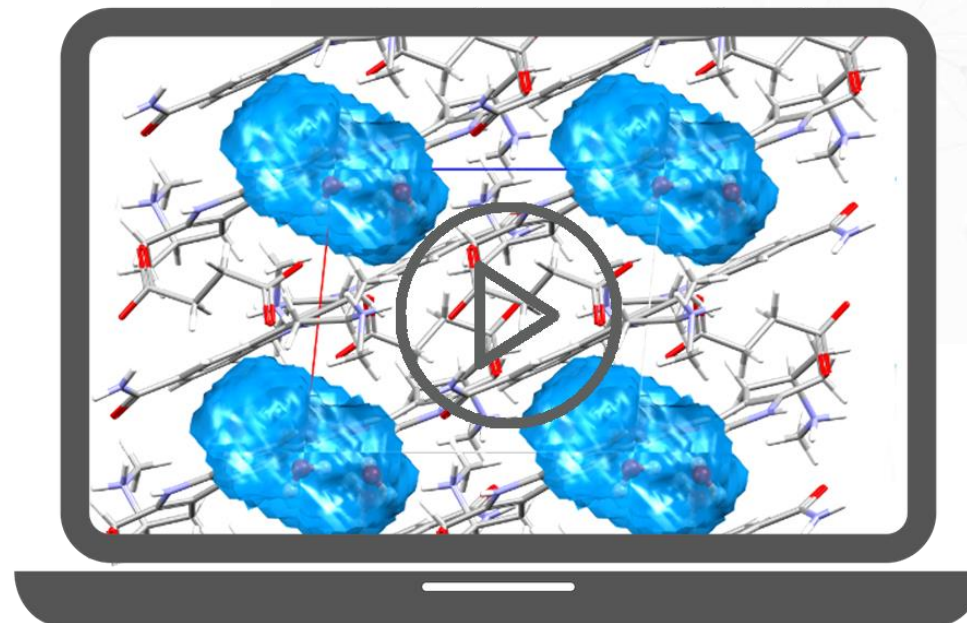
Francis Atkinson

Cheminformatics Data
Scientist

Overview

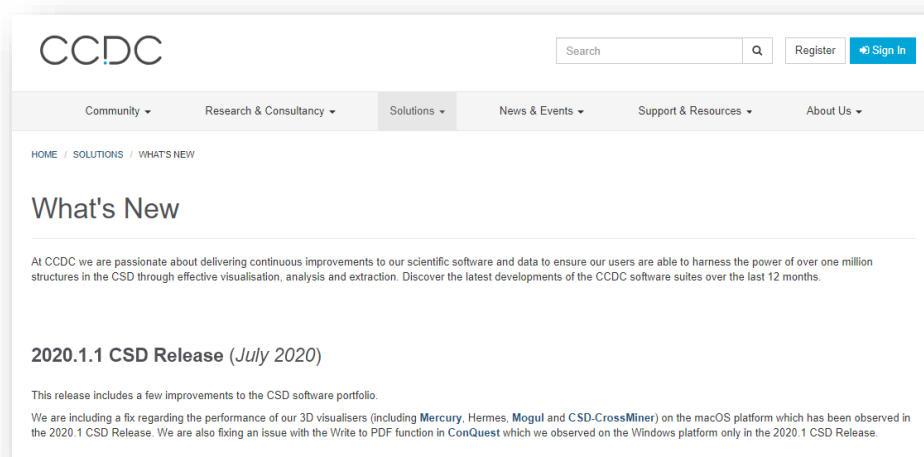
In this webinar we will discuss:

- Latest updates and news
- CSD KNIME Component Collection
- New CSD Subsets (Drugs, Pesticides & COVID-19)
- Q&A: the floor is yours



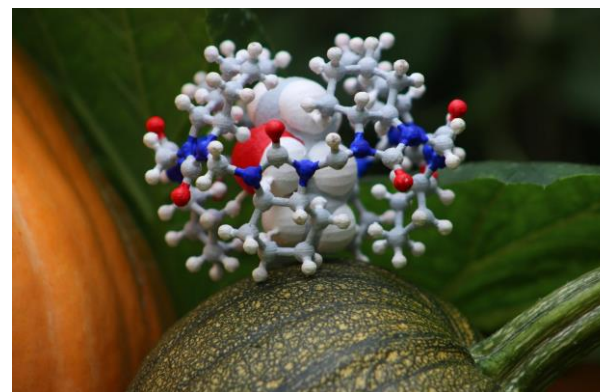
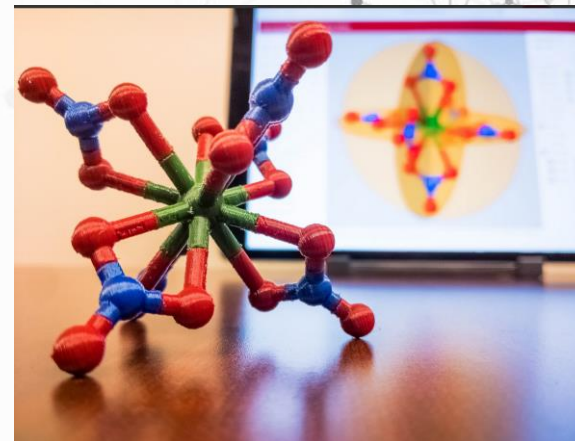
Latest updates and news from CCDC

- **2020.2 CSD Release** is available – install latest updates!
support@ccdc.cam.ac.uk
- CCDC renewed its partnership with **University of Cambridge**
- CSD Licence academic renewals – contact
admin@ccdc.cam.ac.uk by **30th September** for a discounted renewal.



Latest updates and news from CCDC

- CCDC annual 3D print contest is open. To get involved visit our website at ccdc.cam.ac.uk
- CCDC Community survey: we want to hear your feedback about our services, please take the survey by 31st October.



Latest updates and news from CCDC

>Events

- [Advancing the CCDC/FIZ Karlsruhe collaboration](#)
 - Live session on 7th October
 - 3.30 pm – 5 pm (BST)
- [Global User Group Meeting](#) – 14th October
 - Virtual event
 - 9 am to 4 pm (BST)
- [Student Day](#) – 15th October
 - Virtual event
 - 9 am to 3 pm (BST)

→ Register for all CCDC events here <https://www.ccdc.cam.ac.uk/News/Events>

CCDC 2020.2 Release

CSD KNIME Components



Francis Atkinson

Cheminformatics Data Scientist

KNIME

- A workflow tool
 - ‘Programming without code’
 - Many powerful tools packaged up and available out of the box
- Free to download from <https://www.knime.com>
- Fully-functional desktop application
 - Also paid-for [KNIME Server](#), which can be run on [AWS](#) or [Azure](#)
- Good online learning [resources](#)
- Responsive help [forum](#)
- Lively and broadly-based user community

CCDC KNIME Components

- This first release comprises a basic set of functionality
 - Designed to start a conversation with the community
- Implemented as KNIME [Components](#)
 - Uses the [CSD Python API](#) via Python Script Nodes
 - Requires the new ccdc_knime module to be installed
 - Distributed through the [KNIME Hub](#)
- CCDC Example Workflows
 - Illustrate the use of the Components
- No specific license is required to use the Components
 - Licences for the underlying CCDC functionality used will be required

CSD Search

- Text/Numeric Search
 - Chemical Name
 - Author(s)
- Substructure Search
 - 2D only
 - no constraints or measurements
 - MOL-format queries supported
- Similarity Search
 - 2D only
 - Tanimoto similarity/fingerprint based
 - MOL or MOL2 format queries supported

GOLD Docking

- Ligand preparation from SMILES
 - 2D to 3D only
 - No (de)protonation, tautomerization, stereocentre enumeration
- Docking
 - Configured using an uploaded gold.conf file
 - e.g. as prepared in Hermes
 - Can optionally start Hermes to view solutions
- Export of solutions
 - Solutions written to individual MOL2 files
 - Can also optionally start Hermes

User Guide

- Comprehensive setup and usage instructions
- <https://www.ccdc.cam.ac.uk/support-and-resources/ccdcresources/>
 - Resource Type > User Guide

Documentation and Resources

Resource Type

User Guide

Category

Please select a Category

Product

Please select a Product

Sort Order

Most Recent

Search

18 Results Found

CCDC KNIME Components User Guide

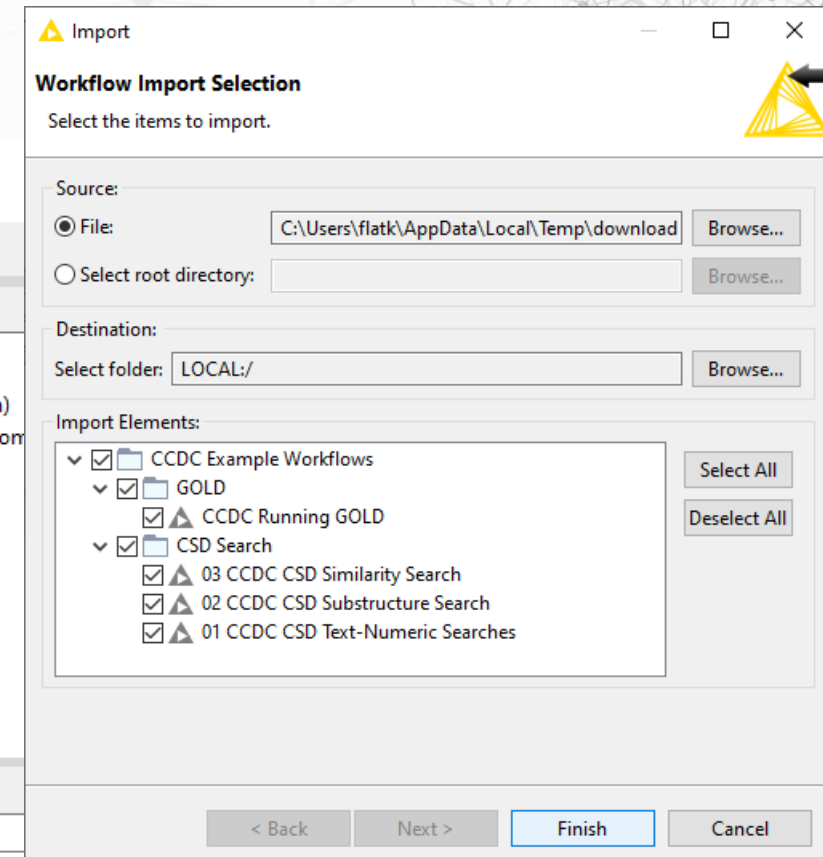
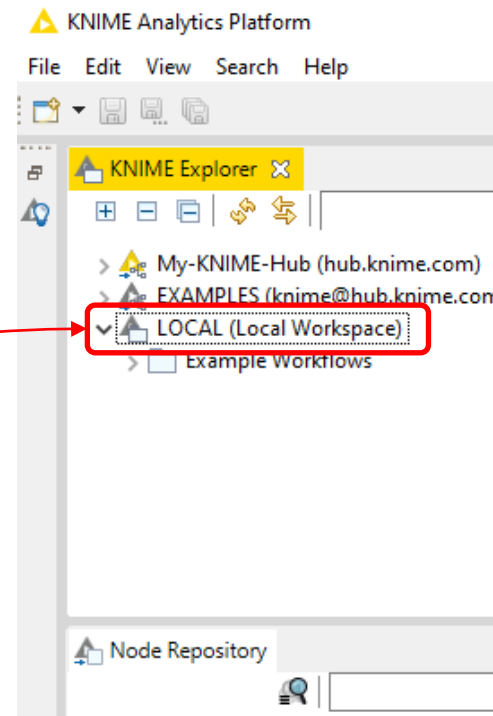
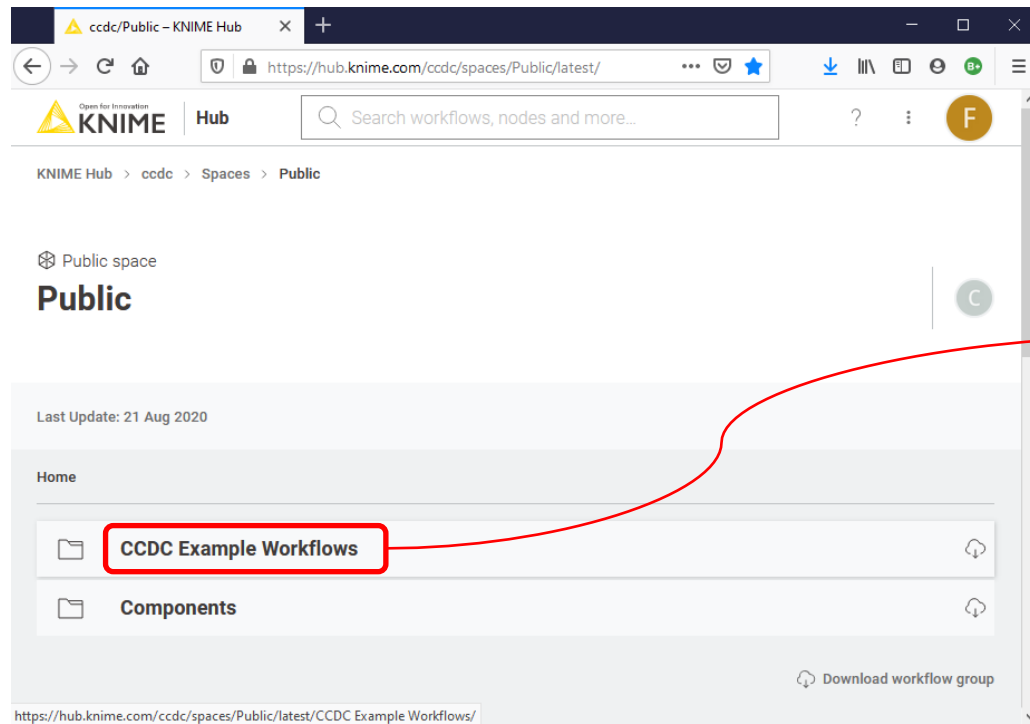
Last Modified: 03/09/2020 PDF

Contents

Contents	3
1 Introduction	4
1.1 Integration Scheme	4
1.2 Requirements	4
1.2.1 Setting up KNIME to use the CSD-System Python distribution	5
1.2.1.a Windows	6
1.2.1.b Linux or macOS	6
1.2.2 Setting up KNIME to use your own Python installation	7
1.2.2.a Using a pre-existing installation	7
1.2.2.b Creating a Launch Script	7
1.2.2.c Installing Miniconda	8
1.2.2 Configure logging	9
1.3 Installation of CCDC Components and Example Workflows	9
2 CCDC KNIME Components	13
2.1 CSD Search	13
CCDC CSD Chemical Name Search	14
CCDC CSD Author Search	14
CCDC CSD Substructure Search	15
CCDC CSD Similarity Structure Search	16
2.2 Viewers	18
CCDC Run Mercury	18
2.3 Utilities	19
CCDC Load MOL files	19
2.4 GOLD	20
CCDC GOLD Ligand Prep	20
CCDC GOLD Run Docking	21
CCDC GOLD Export Results	22
3 Appendices	24
3.1 Columns returned by CSD Search components	24

CCDC Example Workflows

- Drag from KNIME Hub into 'Local Workspace' folder in KNIME Explorer
- <https://hub.knime.com/ccdc/spaces/Public/latest/>



On to the demo...

KNIME Analytics Platform

File Edit View Node Search Help

100%

KNIME Explorer

- My-KNIME-Hub (hub.knime.com)
- EXAMPLES (knime@hub.knime.com)
- LOCAL (Local Workspace)
 - CCDC Example Workflows
 - CCDC Search
 - 01 CCDC CSD Text-Numeric Searches
 - 02 CCDC CSD Substructure Search
 - 03 CCDC CSD Similarity Search
 - GOLD
 - CCDC Running GOLD
 - Example Workflows

01 CCDC CSD Text-Numeric Searches

CCDC CSD Text/Numeric Searching

This example workflow illustrates the use of the CCDC CSD Text/Numeric Search components. Currently available are **CCDC CSD Chemical Name Search** and **CCDC CSD Author Search**.

Note that example values are supplied for the searches here and that options are not necessarily set to the defaults. No extra files are required.

As with all of the CSD Search components, the hits here can be viewed using the **CCDC Run Mercury** component (which allows visualisation of the experimentally-determined 3D structures), via a KNIME Table View or simply by using a component's built-in viewer.

Note that the experimentally-determined 3D coordinates from the CSD are returned in MOL2 format in the column *mol2_mol*. As KNIME does not have a 3D viewer, the 3D coordinates are depicted in 2D in the built-in viewer, which is not particularly useful. The Table View node doesn't have any molecular depiction capability at all, so can only show the text of the molfiles; this column is thus removed before the Table View node is used.

Search filters are constraints on CSD search results that are independent of the search type. They include, for example, applying an R-factor threshold, disallowing the presence of disorder etc. These filters cannot currently be applied as part of a search in KNIME as they can in e.g. ConQuest. However, all the required properties are returned by the various CSD Search components, so the filtering may be applied as a post-processing step. A commonly-used (but reasonably rigorous) set of filters is applied by the **CSD Search Filters** metanode in the example below. The filters applied can easily be modified if required: simply double-click on the metanode to inspect or alter the filters.

CCDC CSD Chemical Name Search

An example chemical name is supplied, and the option to ignore non-alphabetic characters in names is checked (the default is unchecked).

Try the search for this name with the 'ignore non-alphabetic characters' option checked and unchecked.

CCDC CSD Author Search

Some example author names are supplied.

Note that, if multiple author names are specified, they will be AND'ed together (i.e. all must be present).

CSD Search Filters

Double-click to access the individual filters.

CCDC Run Mercury

Execute to run Mercury

Column Filter

Remove molfile

Table View

View hits

Right-click > Interactive View: JavaScript Table View

The Table View node cannot 3D molfiles so we remove this column.

Column Filter

Publication columns

Here, we subset the columns just to highlight the publication details.

CCDC CSD Chemical Name Search

Searches the CSD for a chemical name.

If the 'Ignore non-alphabetic characters' option is checked (default is unchecked), the search will ignore non-alphabetic characters in chemical names. This means, for example, that the query "butadiene" would also match "buta-1-3-diene".

The number of hits returned may be limited using the 'Max. hits' option (default 100). This can be useful as it stops overly-general searches taking a very long time.

Dialog Options

Chemical name
The chemical name to search for.

Ignore non-alphabetic characters?
If checked, the search will ignore non-alphabetic characters in compound names.

Max. hits
The maximum number of hits to return.

Ports

Output Ports

0 CSD records that match the query.

KNIME Console

```

*****
*** Welcome to KNIME Analytics Platform v4.2.1.v20200920915 ***
*** Copyright by KNIME AG, Zurich, Switzerland ***
*****
Log file is located at: C:\Users\flatk\knime-workspace\.metadata\knime\knime.log
  
```

CCDC 2020.2 Release

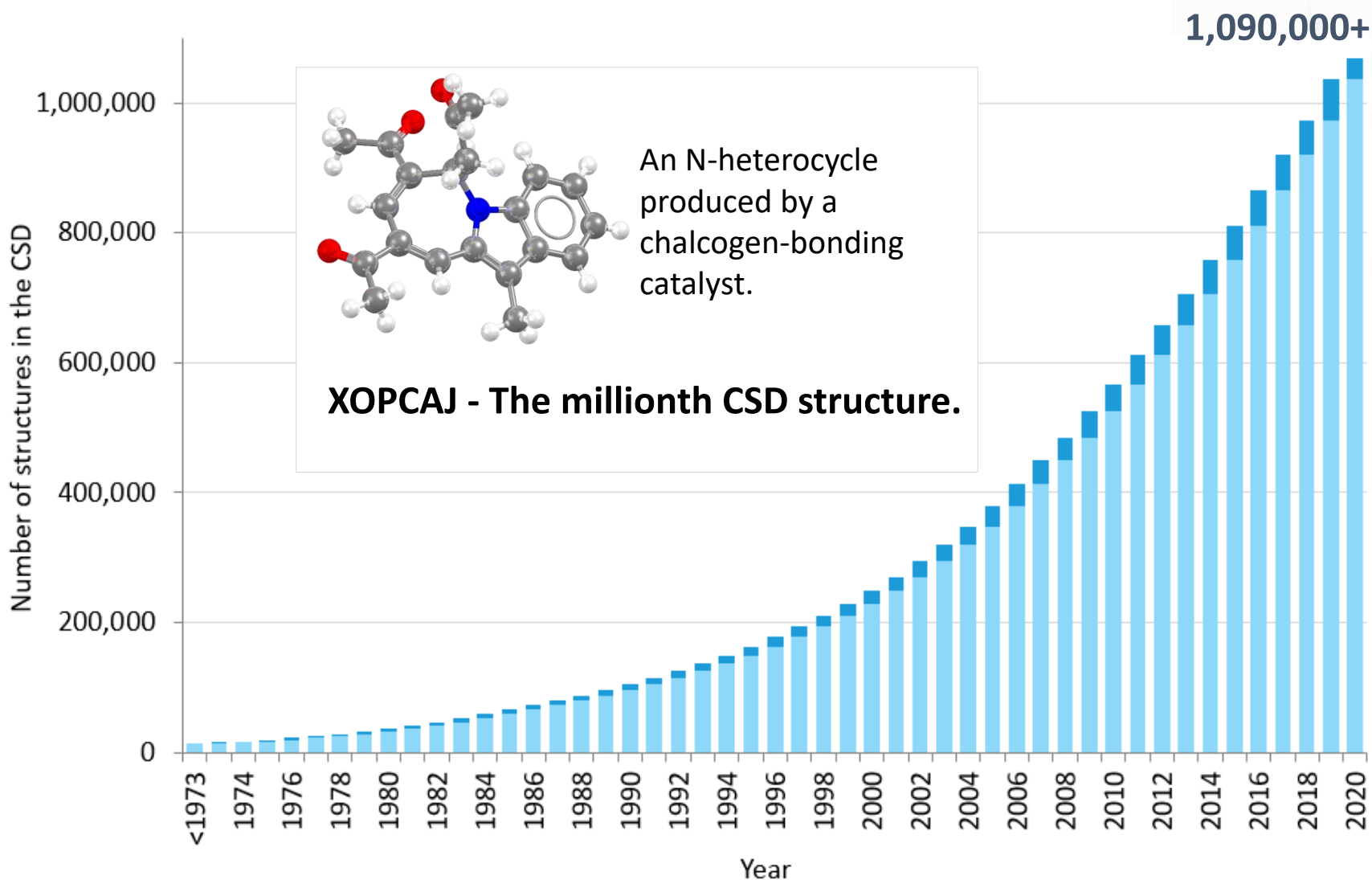
New CSD Subsets (Drugs, Pesticides & COVID-19)



Seth Wiggin

Senior Scientific Editor

The Cambridge Structural Database



- Every published structure
 - Inc. ASAP & early view
 - *CSD Communications*
 - Patents
 - University repositories
- 60,000+ new entries added a year
- Every entry enriched and annotated by experts

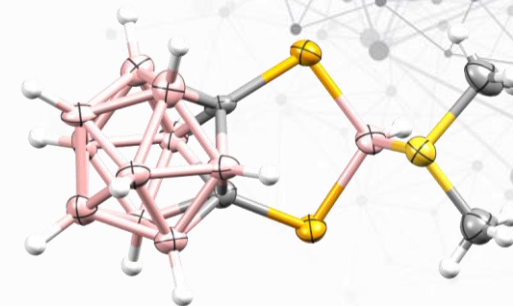
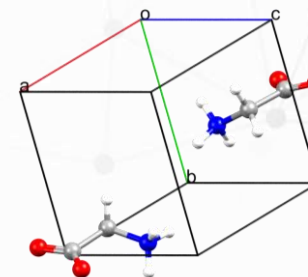
What are CSD Subsets?



CSD Subsets: Helping find one in a million+

Why provide subsets?

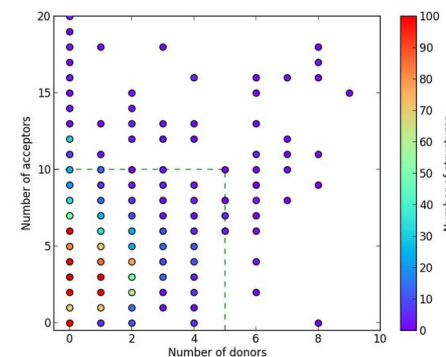
- The CSD contains a huge range of compounds; subsets allow easy access to the most relevant structures of interest
- Gives CSD users the benefit of analysis from in-house and external experts
- Convenient starting point for analysis using CSD or 3rd-party tools



 DRUGBANK

 PPDB

Plot of H-bond donors and acceptors in organic molecule subset



CCDC

CSD Subsets

- 'Best representative' lists first produced in 2006
 - Aim to give one single example of every structure (incl. polymorphs) in the CSD
 - 4 lists: Hydrogens, low temp, room temp, R-factor
 - J. van de Streek, *Acta Crystallogr., Sect. B: Struct. Sci.*, 2006, **B62** 567-579, DOI: [10.1107/S0108768106019677](https://doi.org/10.1107/S0108768106019677)
- MOF subsets added in 2017
 - 'MOF subset' allows users to find *all* MOF-like structures
 - Non-disordered MOF subset is designed for high-throughput calculations
 - P. Z. Moghadam, A. Li, S. B. Wiggin, A. Tao, A. G. P. Maloney, P. A. Wood, S. C. Ward, D. Fairen-Jimenez, *Chem. Mater.*, 2017, **29** 2618-2625, DOI: [10.1021/acs.chemmater.7b00441](https://doi.org/10.1021/acs.chemmater.7b00441)
- ADPs available subset added in 2018 release
 - Allows users to find only structures with thermal ellipsoids
 - <https://www.ccdc.cam.ac.uk/Community/blog/2018-06-jack-csd-adps/>

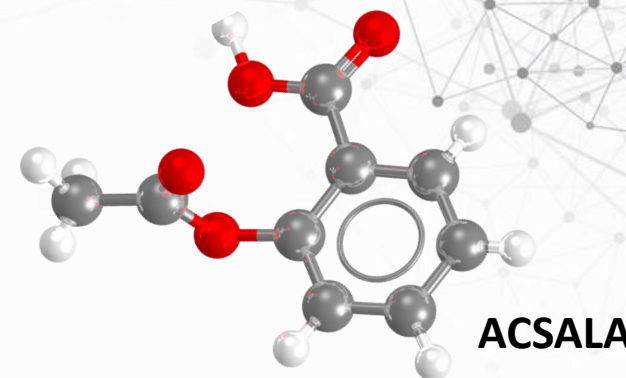
New Subsets for 2020.2

CSD Drug subset; Single-component CSD Drug subset

- Based on the publication

The CSD Drug Subset: The changing chemistry and crystallography of small molecule pharmaceuticals

Mathew J. Bryant, Simon N. Black, Helen Blade, Robert Docherty, Andrew G.P. Maloney, Stefan C. Taylor, *J. Pharm. Sci.*, 2019, **108** 1655-1662, DOI: [10.1016/j.xphs.2018.12.011](https://doi.org/10.1016/j.xphs.2018.12.011)



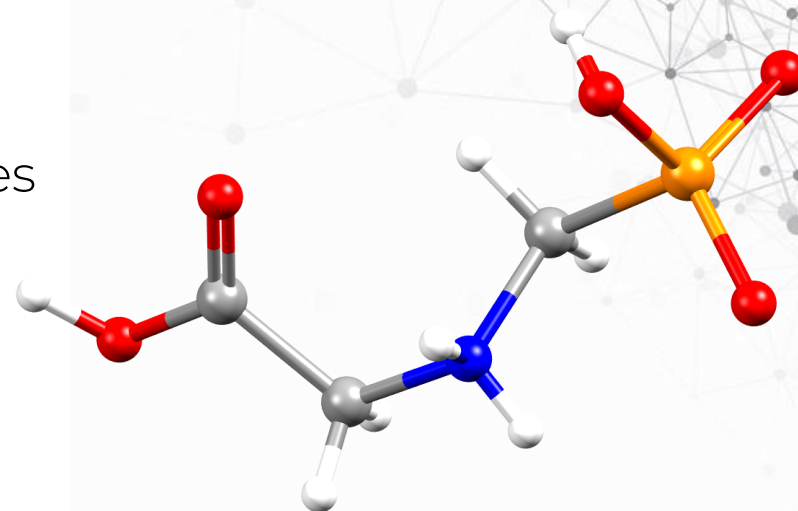
- Uses InChI matching to the DrugBank list of approved drugs
 - CSD Drug subset has all CSD entries containing a drug molecule (incl. hydrates, solvates etc.)
 - Single-component CSD Drug subset has entries where the drug molecule is the only one modelled



New Subsets for 2020.2

CSD Pesticide subset

- Uses InChI matching to the Pesticide Properties Database (PPDB)
- <https://sitem.herts.ac.uk/aeru/ppdb/en/>
- This subset contains all CSD entries (including hydrates, solvates, salts and metal complexes) with a match to a PPDB entry
- The matches do not take into account stereochemistry, meaning some CSD entries in the list may correspond to enantiomers or stereoisomers of the PPDB entry.



PHOGLY

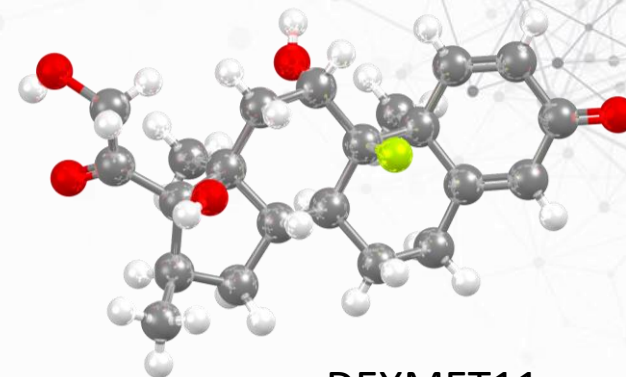


CCDC

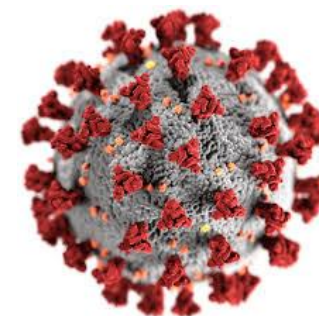
New Subsets for 2020.2

CSD COVID-19 subset

- Based on our recent blog post:
<https://www.ccdc.cam.ac.uk/Community/blog/2020-03-26-molecules-of-interest-in-the-fight-against-covid-1/>
- Manually-curated list from a variety of sources



DEXMET11

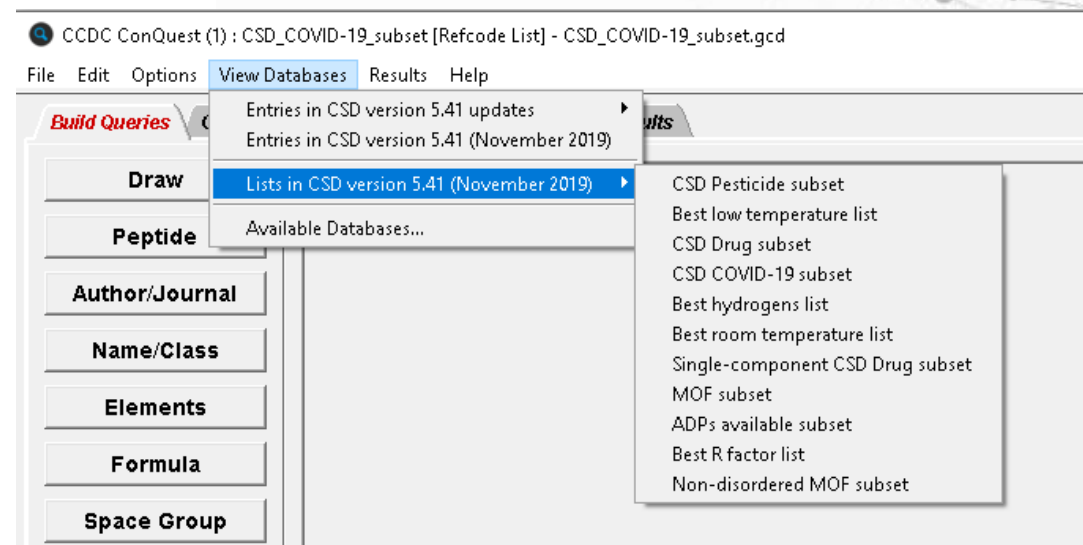


CCDC

Accessing the Subsets

Three ways to use the subsets:

1. Browse the lists in ConQuest



CCDC ConQuest (1) : search1 [Search]

File Edit Options View Databases Results Help

Build Queries **Combine Queries** **Manage Hitlists** **View Results**

Combine Hitlists

Combination Name: combination1

List A: search1

List B: CSD_Pesticide_subset

Include deselected entries in:

☐ List A ☐ List B

Generate a List of Entries:

☒ common to List A and List B

☐ in either List A or List B

☐ in List A but not in List B

OK

Hitlist Overview

CSD_Pesticide_subset (972 Entries)

Refcode List : CSD_Pesticide_subset.gcd

Location : C:\Program Files\CCDC\CSD_2020\CSD_541\subsets

Created : Fri Aug 14 10:51:36 2020

Name	Hits	Type
CSD_Pesticide_subset	972	Refcode List
search1	135	Search

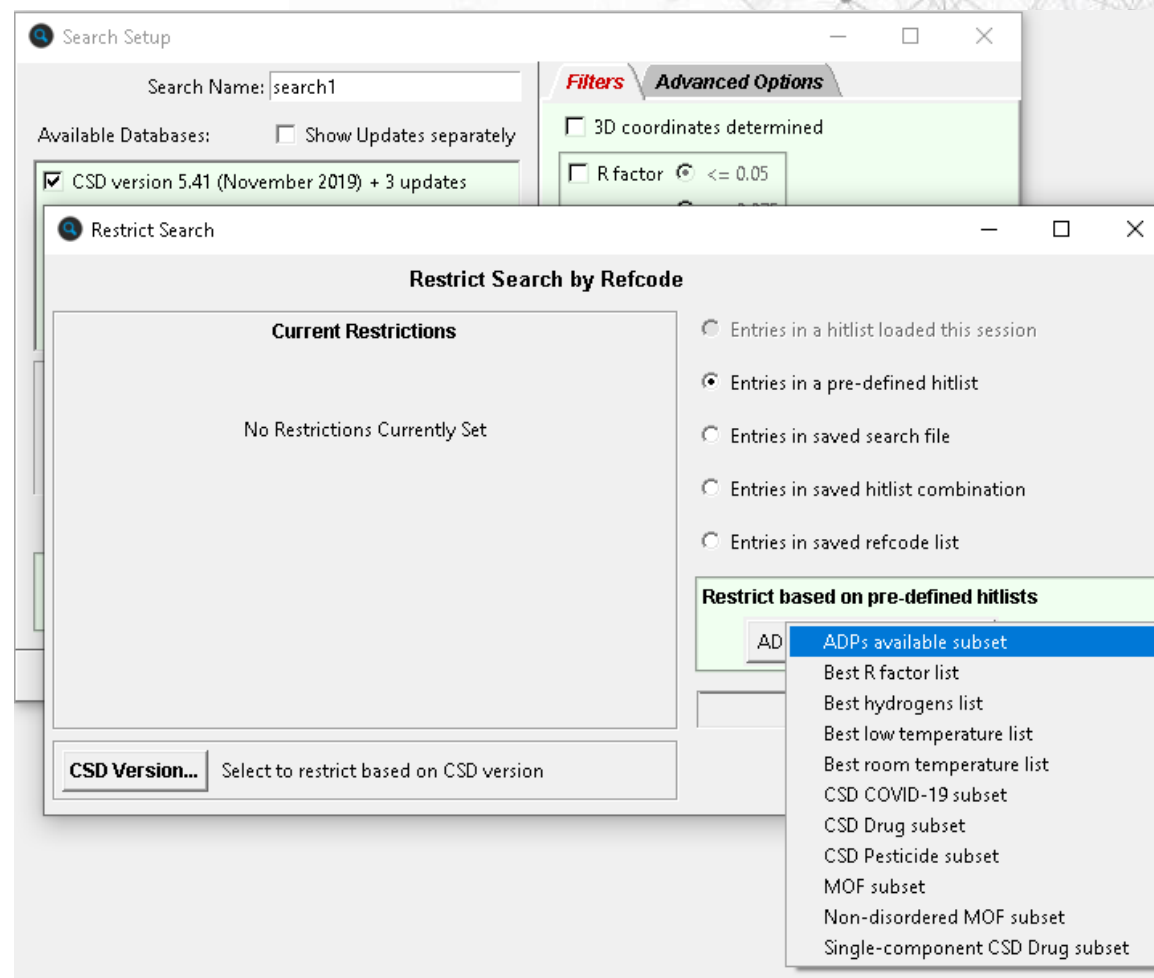
Delete Rename... Notes... View

- Once a subset has been loaded into ConQuest, it can be combined with other searches

Accessing the Subsets

2. Restrict your search in ConQuest

Search setup > Select subset > Entries in a pre-defined hitlist



Accessing the Subsets

3. Use subsets with the CSD Python API

Subsets stored in a single subsets folder in your CSDS installation:

C:\Program Files\CCDC\CSD_2020\CSD_541\subsets

Entry examples

Create indexes of useful information for subsets of CSD entries

Note that this script makes use of functionality from the [cookbook utility module](#).

```
#!/usr/bin/env python
#
# This script can be used for any purpose without limitation subject to the
# conditions at http://www.ccdc.cam.ac.uk/Community/Pages/Licences/v2.aspx
#
# This permission notice and the following statement of attribution must be
# included in all copies or substantial portions of this script.
#
# 2015-06-17: created by the Cambridge Crystallographic Data Centre
#
'''
Provide information on a set of structures in the CSD.

This script takes as input a gcd file (a text file with CSD refcodes) and
writes out the identifier, author(s), literature reference, formula, compound
name and compound synonym(s). The output can be formatted as csv or html.
'''
from __future__ import division, absolute_import, print_function
import six
import sys
import os
import csv
import html
import argparse
import codecs

from ccdc.io import EntryReader

class Writer(object):
    def __init__(self, infile, out, format='csv'):
        try:
            self.rdr = EntryReader(infile, format='identifiers')
        except RuntimeError:
            print('Failed to read input file %s!' % infile)
            exit(1)

        self.out = out
        getattr(self, format + '_header')()
        for e in self.rdr:
            getattr(self, format + '_line')(e)
        getattr(self, format + '_footer')()

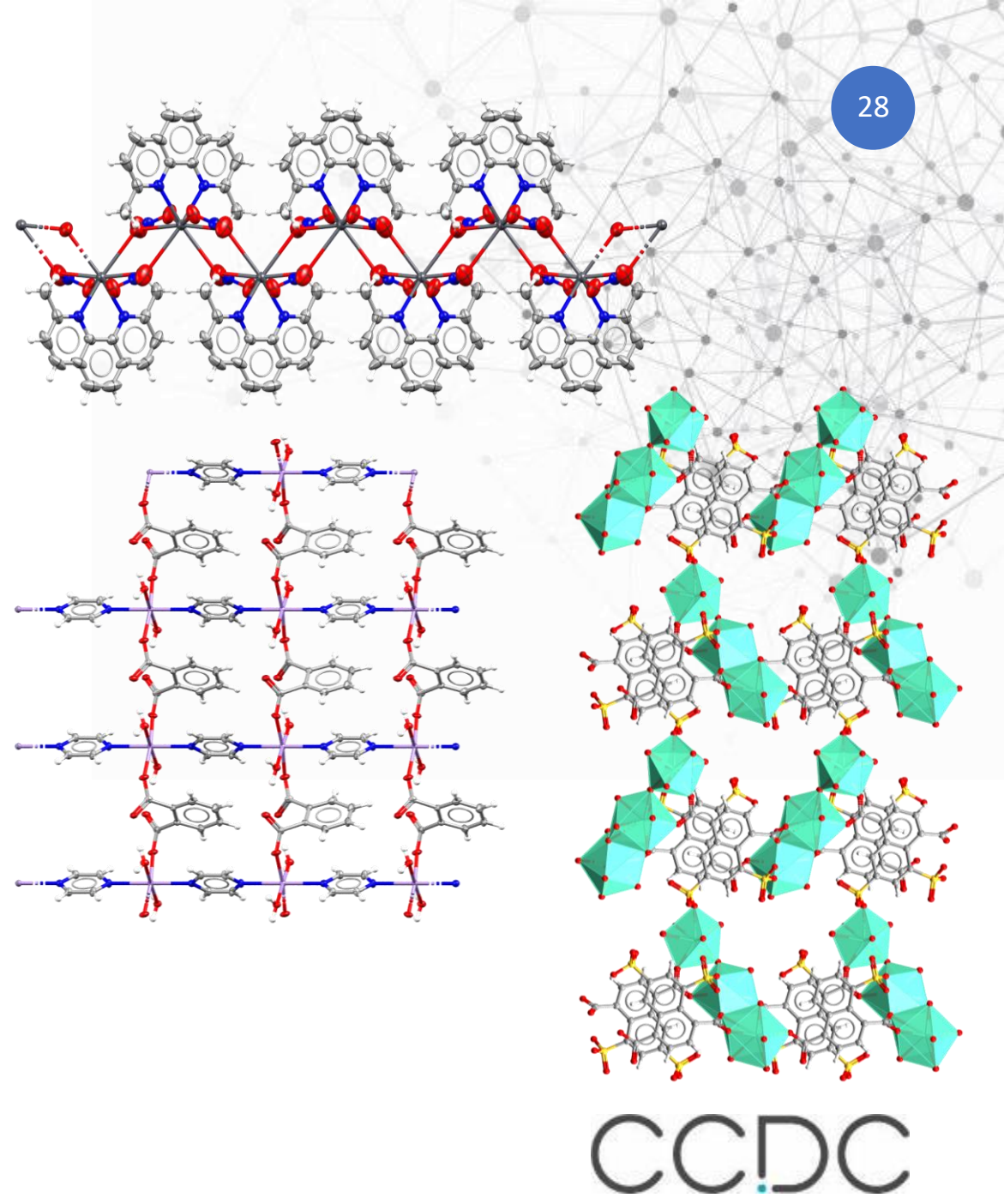
    def csv_header(self):
        data = ','.join([
            'Identifier',
            'Author',
            'Literature Ref',
            'Formula',
            'Compound Name',
            'Synonym'
        ])
```

Future work

- Additional MOF subsets
 - 1D, 2D and 3D frameworks
- Based on the publication

Targeted classification of metal-organic frameworks in the Cambridge Structural Database (CSD)

Peyman Z. Moghadam, Aurelia Li *et al* , *Chemical Science*, 2020, 11 8373, DOI: [10.1039/D0SC01297A](https://doi.org/10.1039/D0SC01297A)
- Please give us your feedback and suggestions!

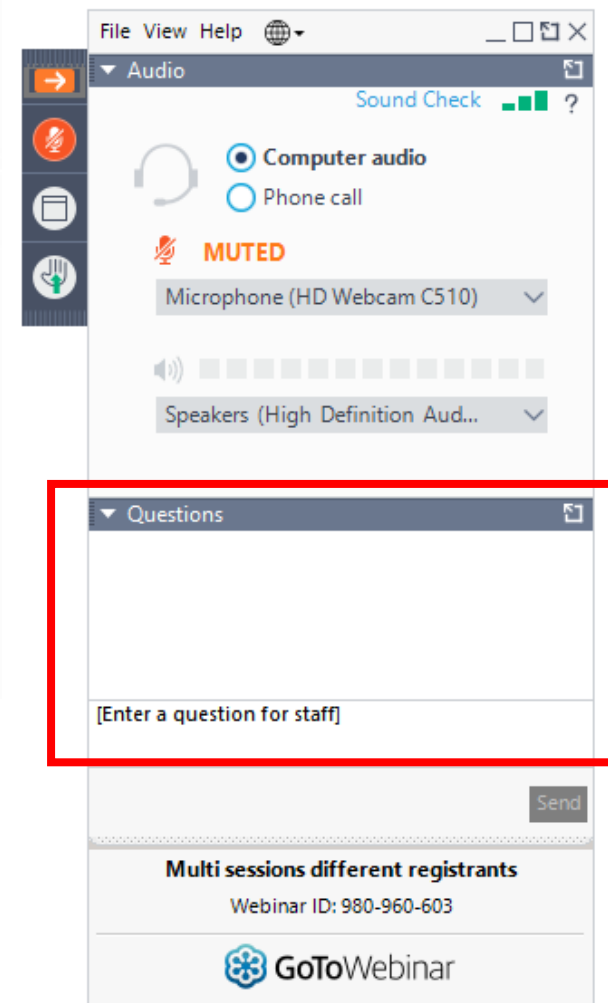


How to use CSD Subsets in ConQuests?

- We recently published a new tutorial video on our YouTube channel to help you find specific entries in the CSD such as drugs, pesticides, MOFs, or best representative structures by showing you how to use subsets in ConQuest. The demonstration includes how to load a subset, how to search structures at the intersection of two subsets (or two hit lists), and how to restrict a search to a chosen subset.
- Watch the video here https://youtu.be/4_yPmc6ssiQ

Q&A

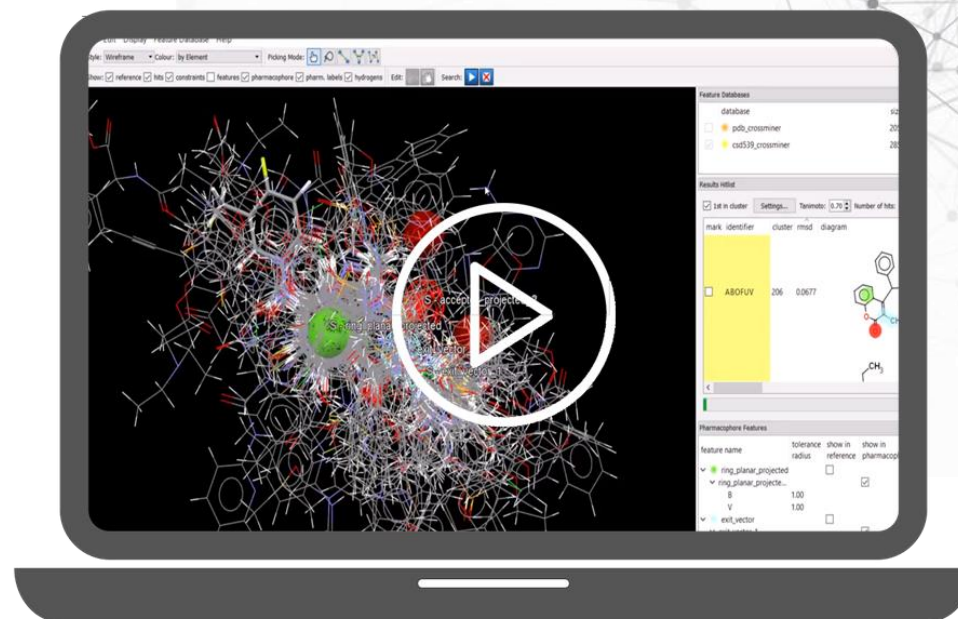
- Type your questions in the box as shown



Next What's Up Webinar

- Next webinar: November 19th
- Send us your ideas and news

hello@ccdc.cam.ac.uk



Thank you

hello@ccdc.cam.ac.uk

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

CCDC