CCDC

advancing structural science

What's Up

Customer Update Webinar

24th September 2020

CCDC

advancing structural science

Today's presenters

Seth Wiggin

Senior Scientific Editor

Ilenia Giangreco

Discovery Science Team Leader

Francis Atkinson

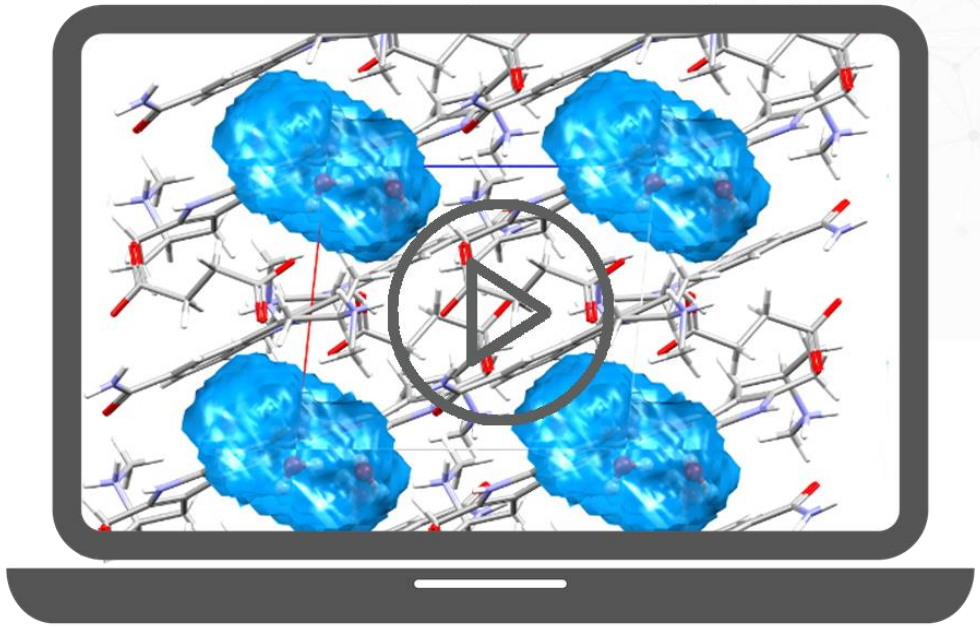
Cheminformatics Data Scientist

CCDC

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



CCDC

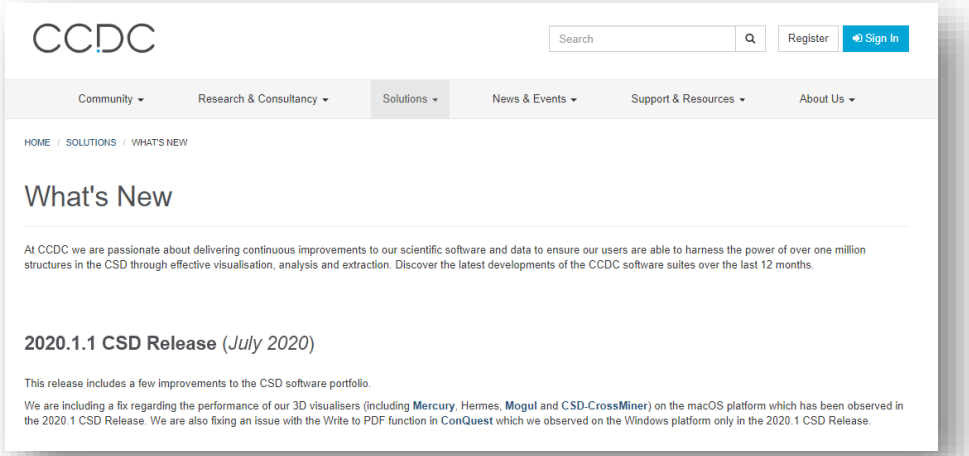
Latest updates and news from CCDC

- **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.

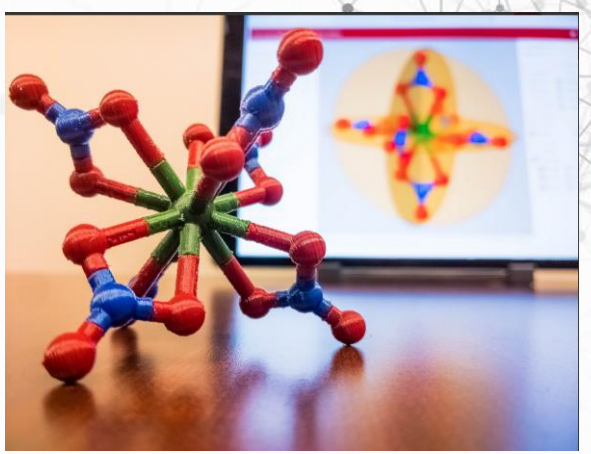


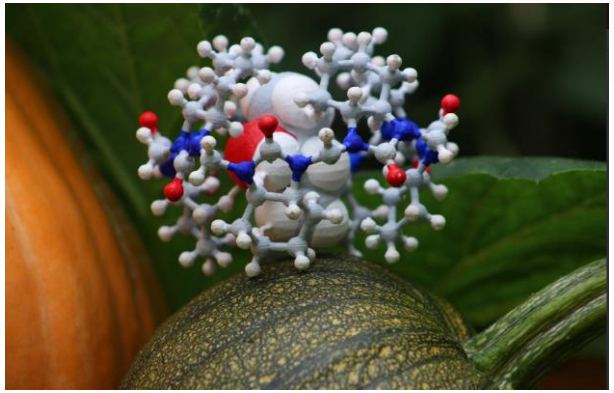
CCDC

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**.





CCDC

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

CCDC 2020.2 Release

CSD KNIME Components

Francis Atkinson

Cheminformatics Data Scientist

CCDC

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

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

CCDC

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

CCDC

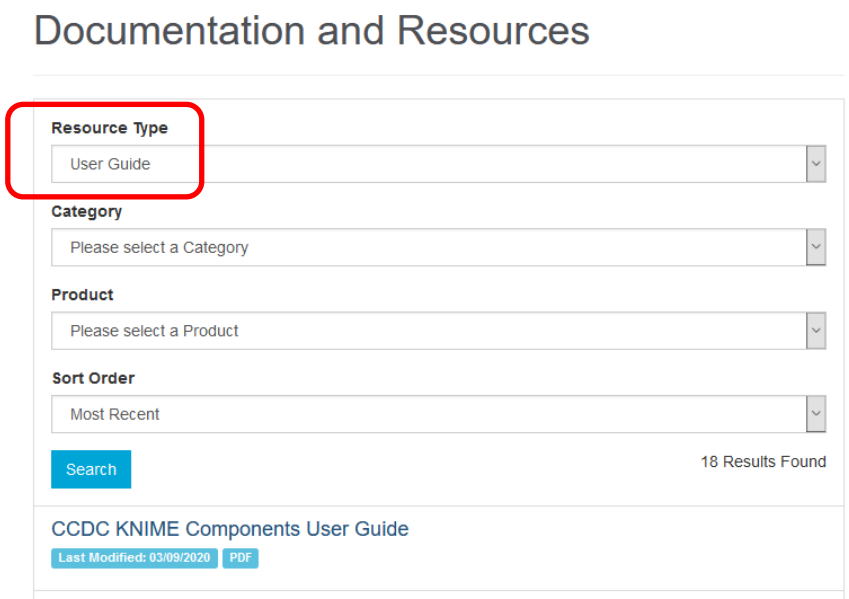
GOLD Docking

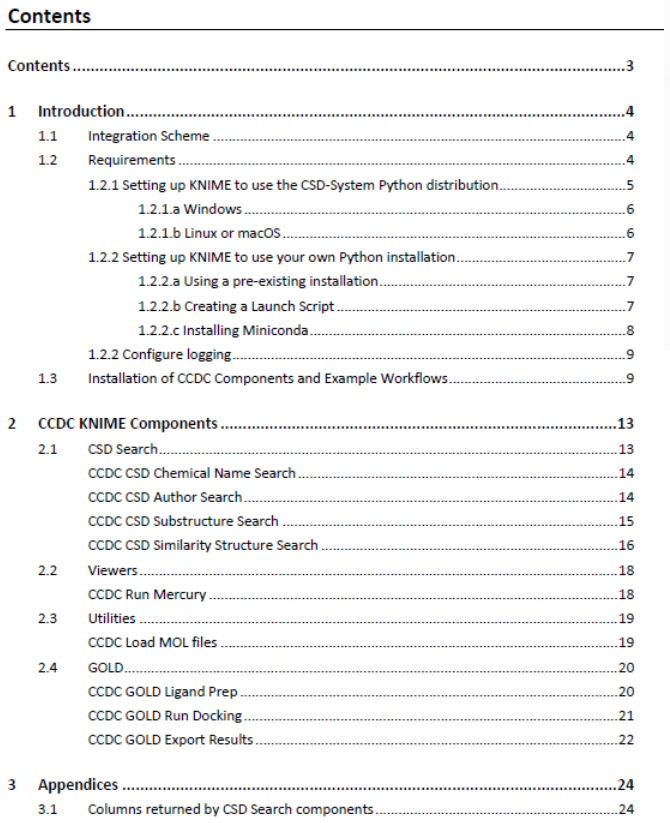
- Ligand preparation from SMILES
  - 2D to 3D only
  - No (de)protonation, tautomerization, stereocentre enumeration
- Docking
  - Configured using an uploaded gold.conf file
    - *e.g.* as prepared in Hermes
  - Can optionally start Hermes to view solutions
- Export of solutions
  - Solutions written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

CCDC

User Guide

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

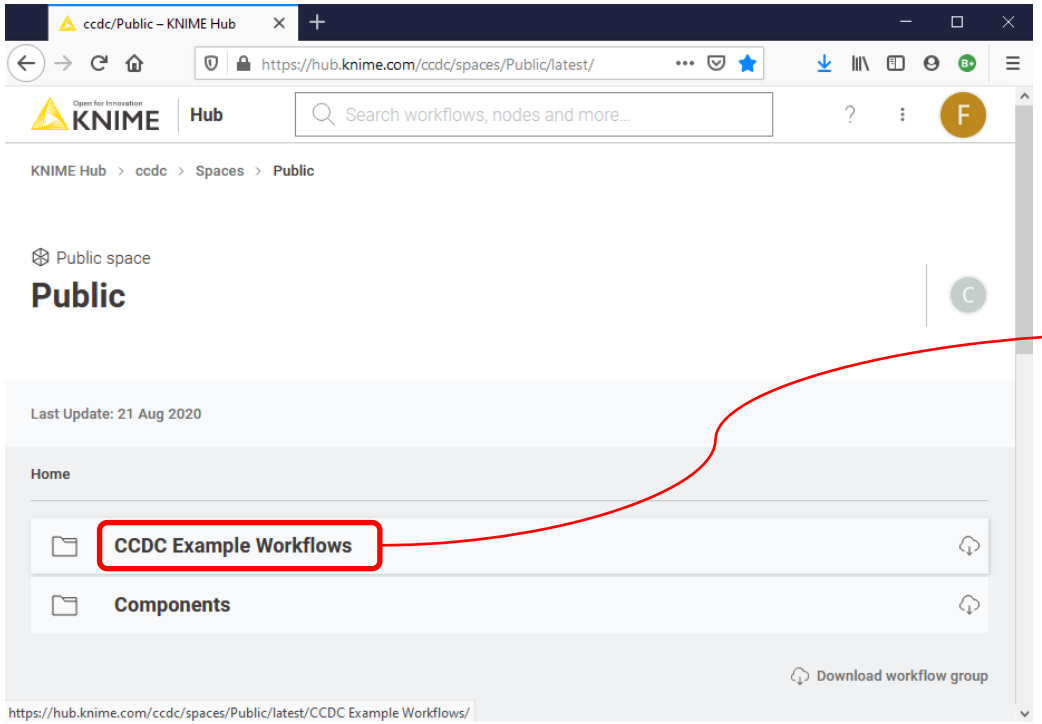


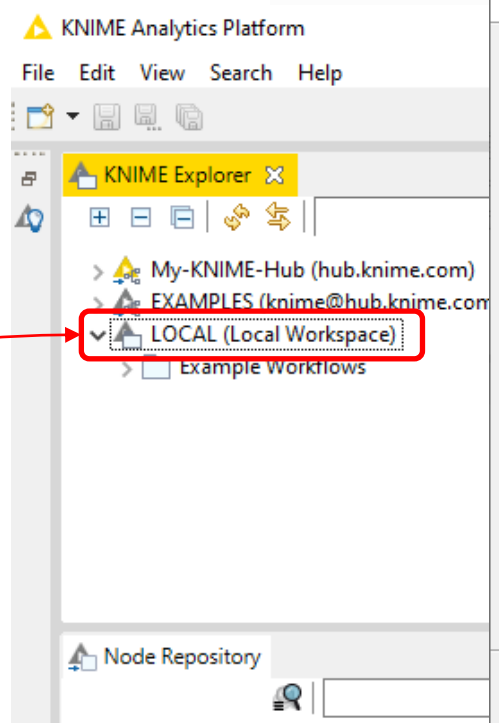


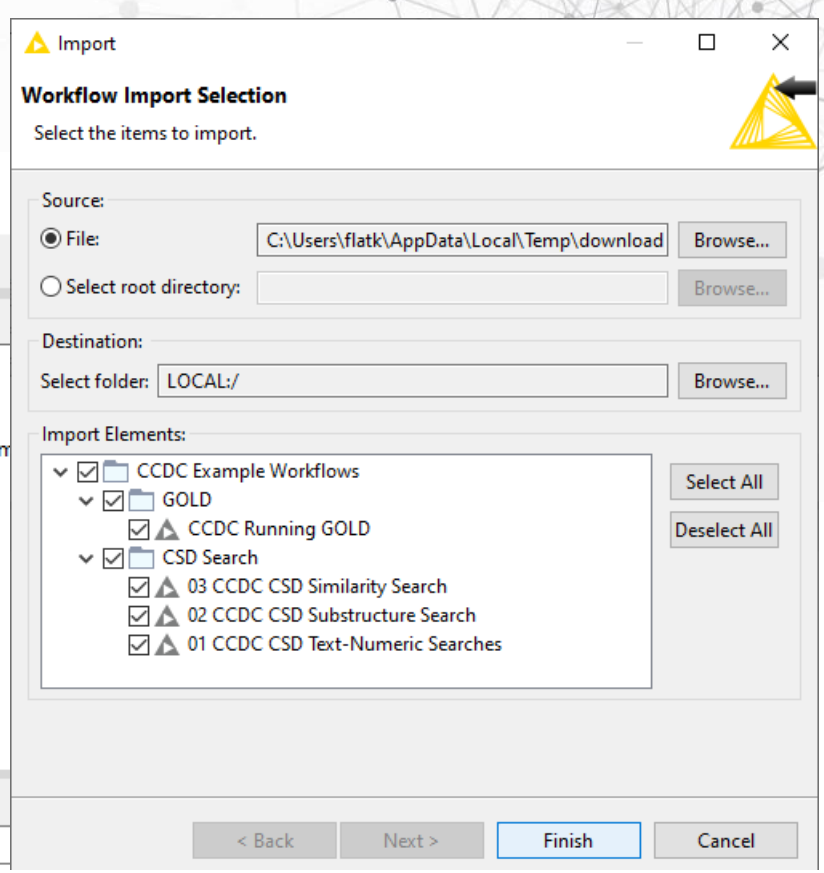
Contents

CCDC Example Workflows

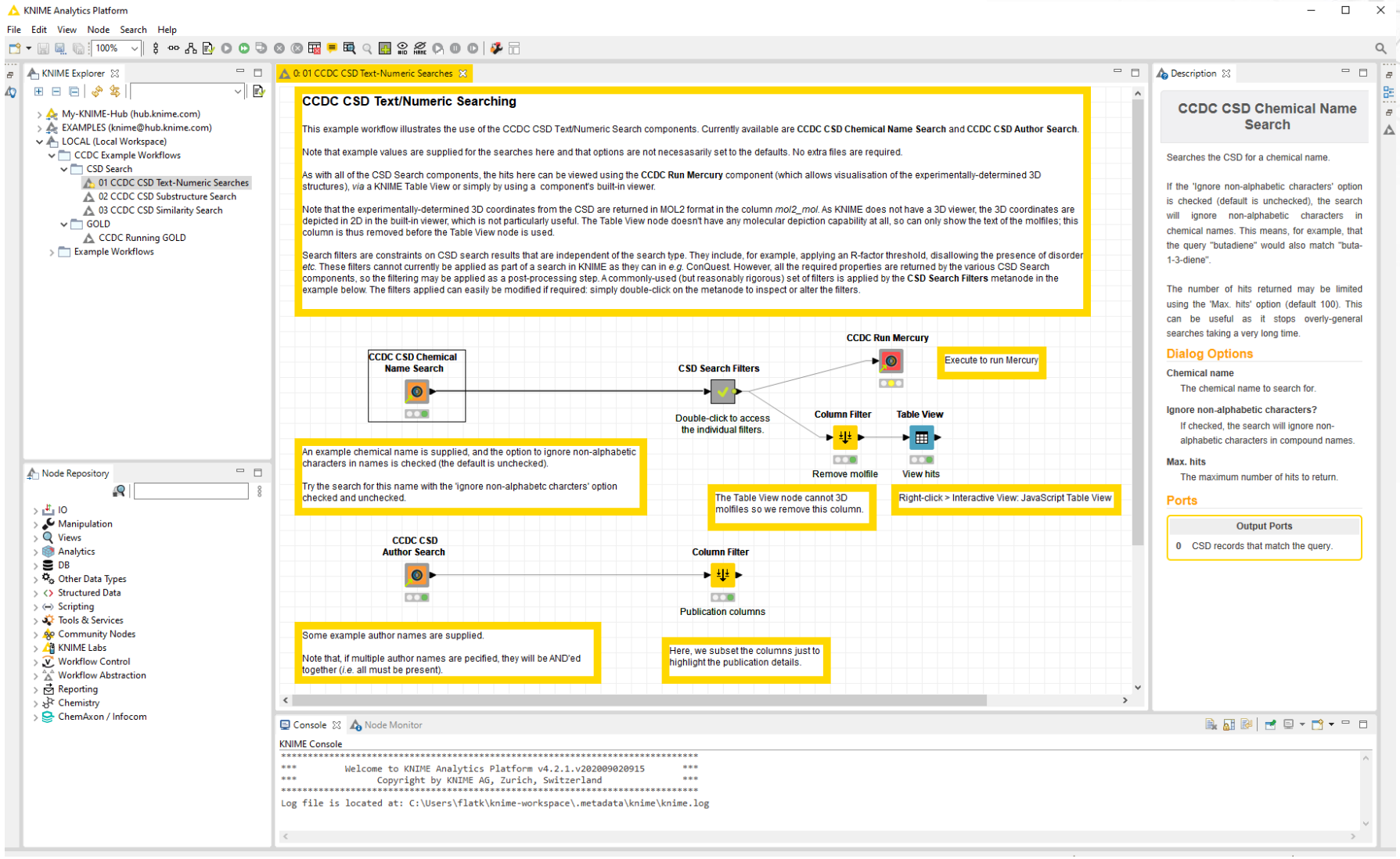
- Drag from KNIME Hub into ‘Local Workspace’ folder in KNIME Explorer
- https://hub.knime.com/ccdc/spaces/Public/latest/







On to the demo...



CCDC 2020.2 Release

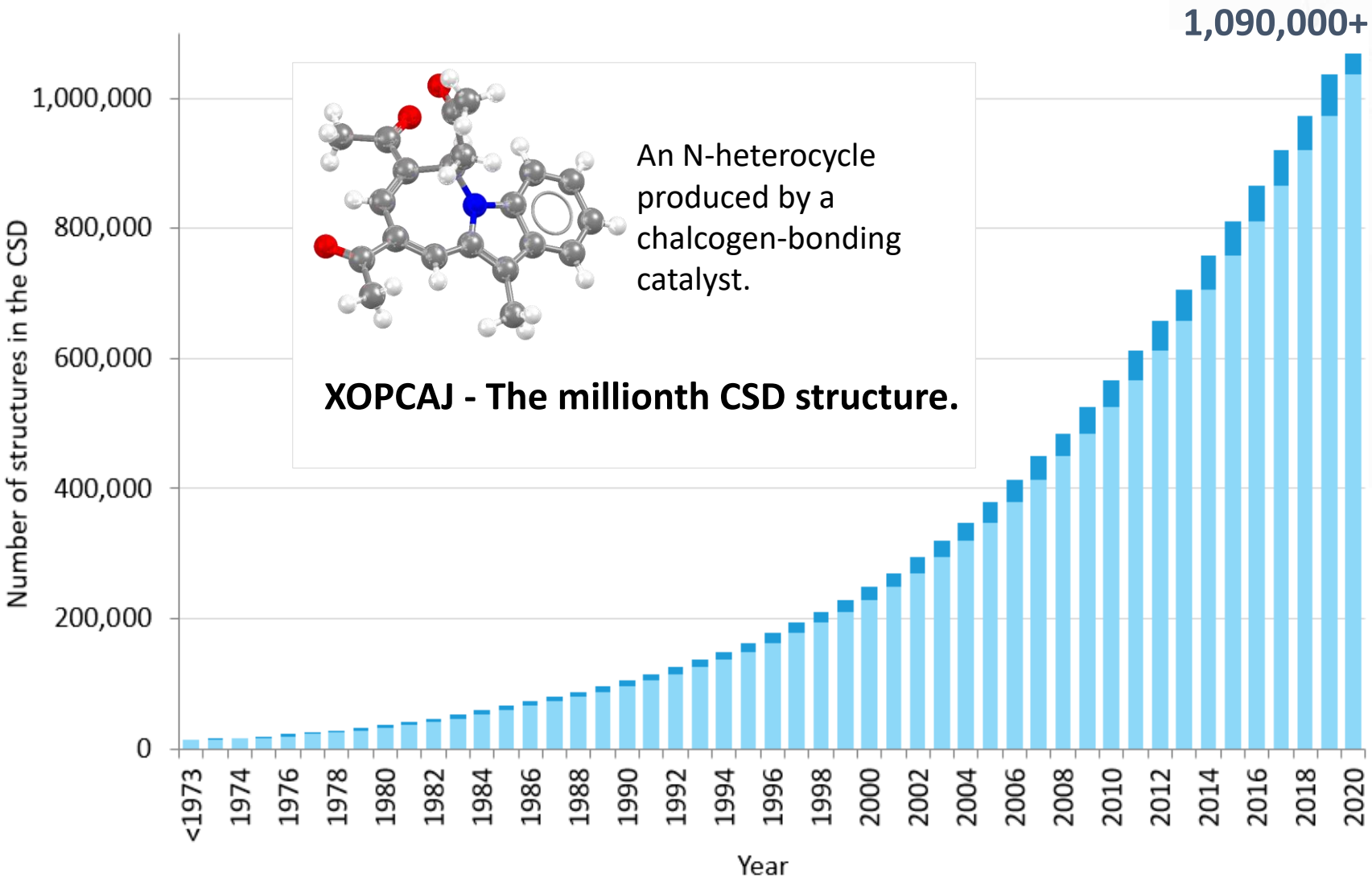
New CSD Subsets (Drugs, Pesticides & COVID-19)

Seth Wiggin

Senior Scientific Editor

CCDC

The Cambridge Structural Database



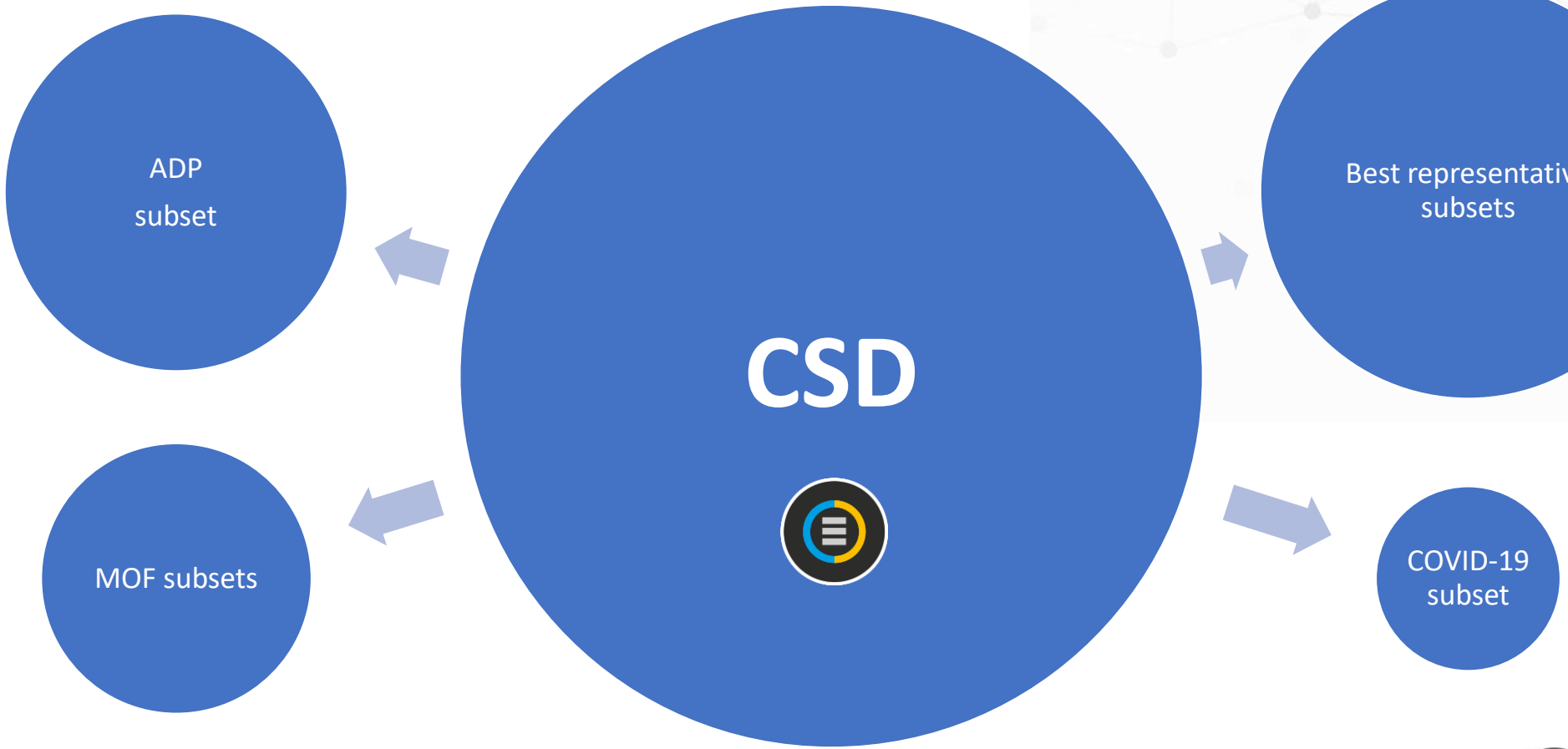
An N-heterocycle produced by a chalcogen-bonding catalyst.

XOPCAJ - The millionth CSD structure.

- 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

CCDC

What are CSD Subsets?

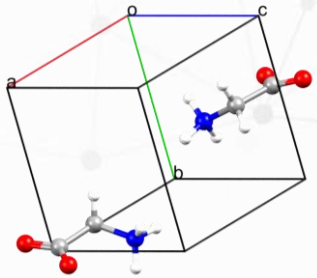


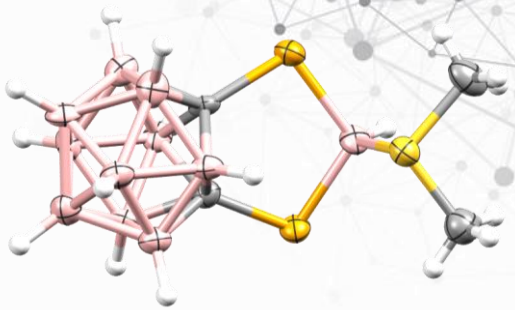
CCDC

CSD Subsets: Helping find one in a million+

Why provide subsets?

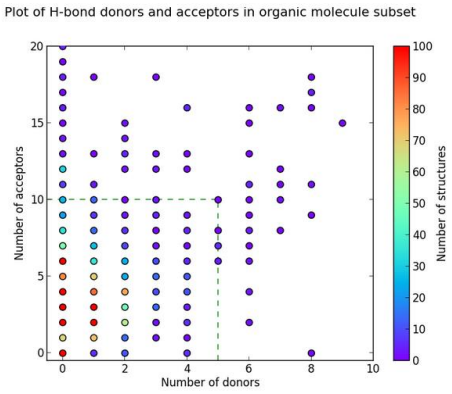
- The CSD contains a huge range of compounds; subsets allow easy access to the most relevant structures of interest
- Gives CSD users the benefit of analysis from in-house and external experts
- Convenient starting point for analysis using CSD or 3rd-party tools





DRUGBANK

PPDB



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

- 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

- 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/

CCDC

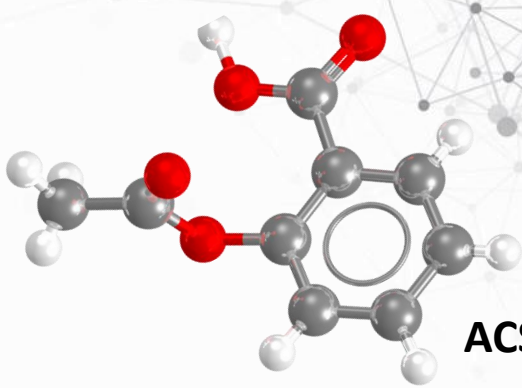
New Subsets for 2020.2

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



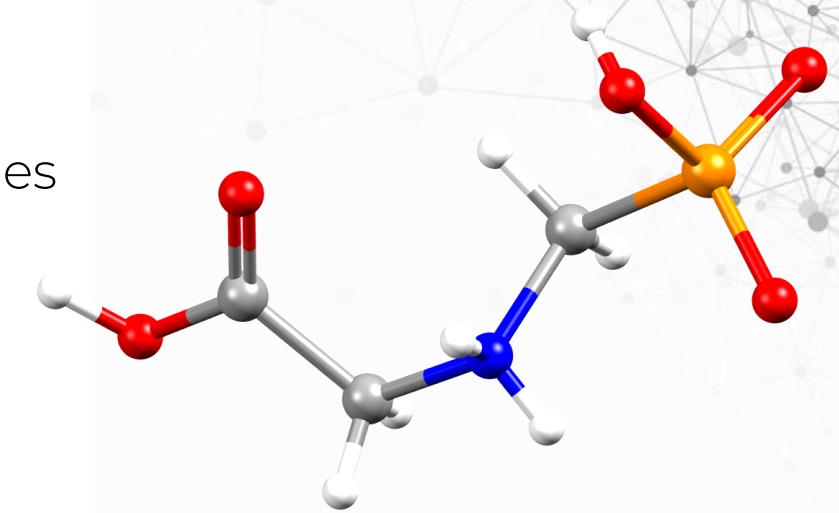
- Uses InChI matching to the DrugBank list of approved drugs
  - **CSD Drug subset** has all CSD entries containing a drug molecule (incl. hydrates, solvates etc.)
  - **Single-component CSD Drug subset** has entries where the drug molecule is the only one *modelled*

CCDC

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/



PHOGLY

- 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.

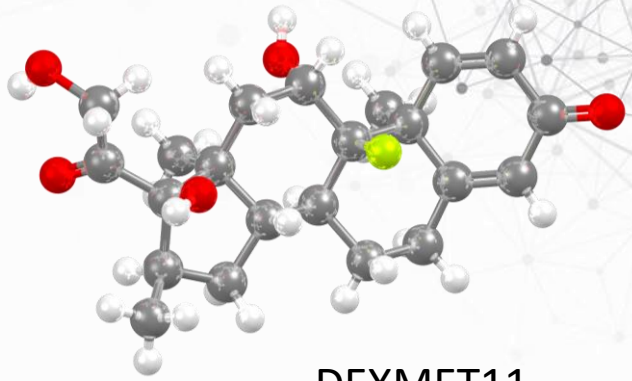
CCDC

New Subsets for 2020.2

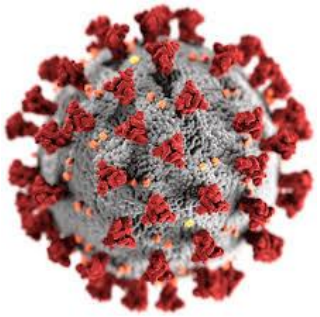
CSD COVID-19 subset

- 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

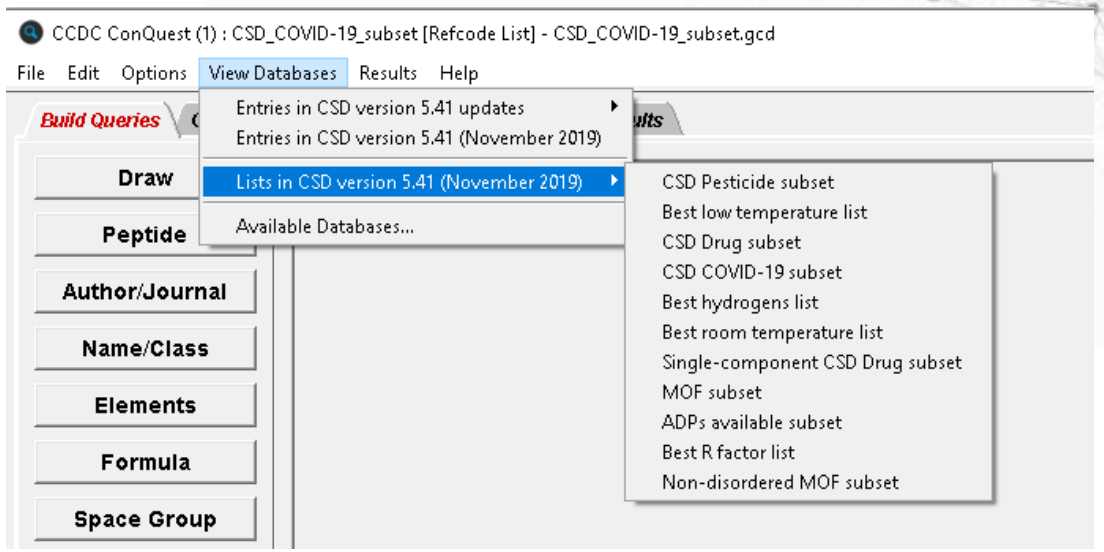


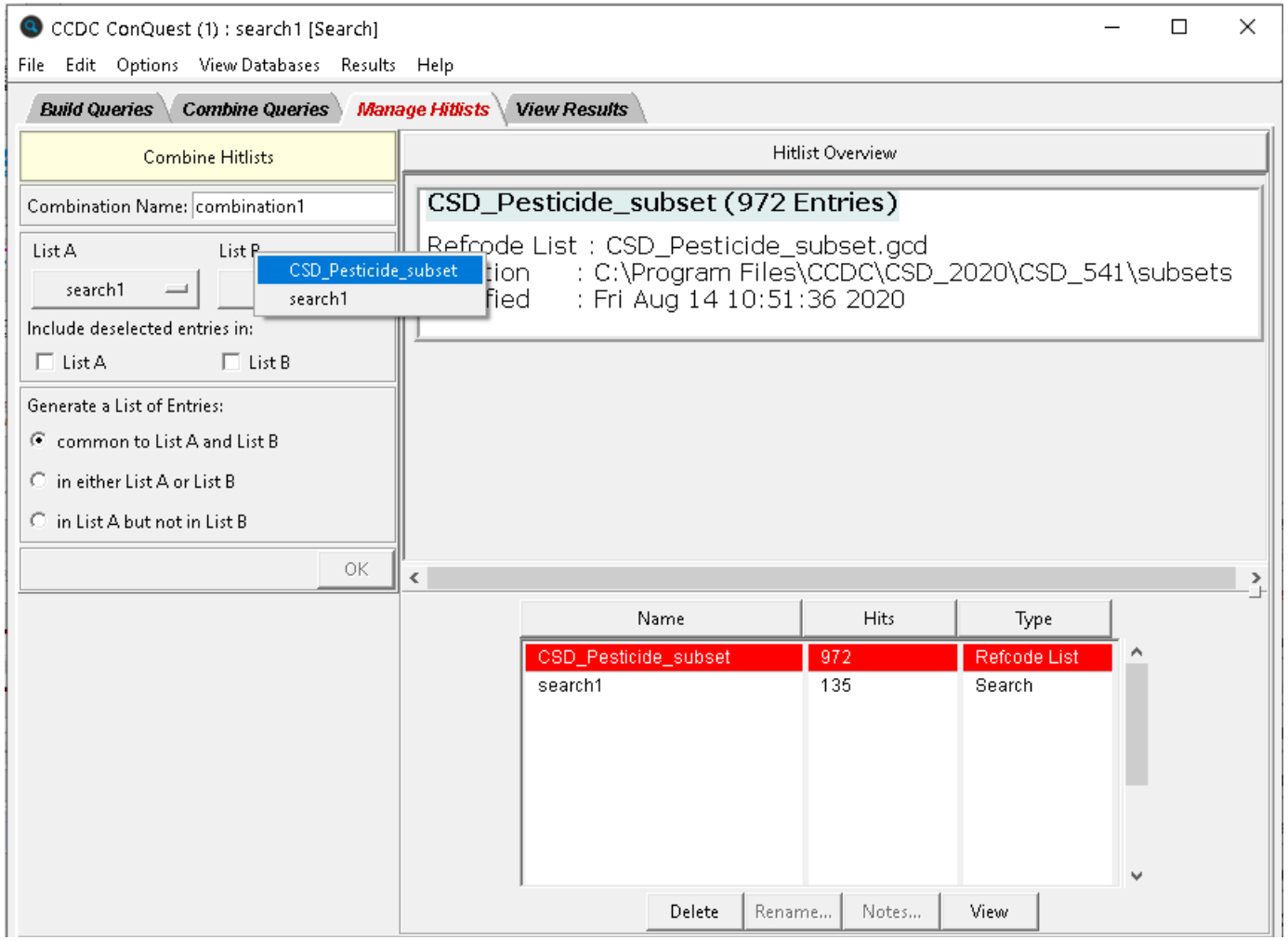
CCDC

Accessing the Subsets

Three ways to use the subsets:

1. Browse the lists in ConQuest



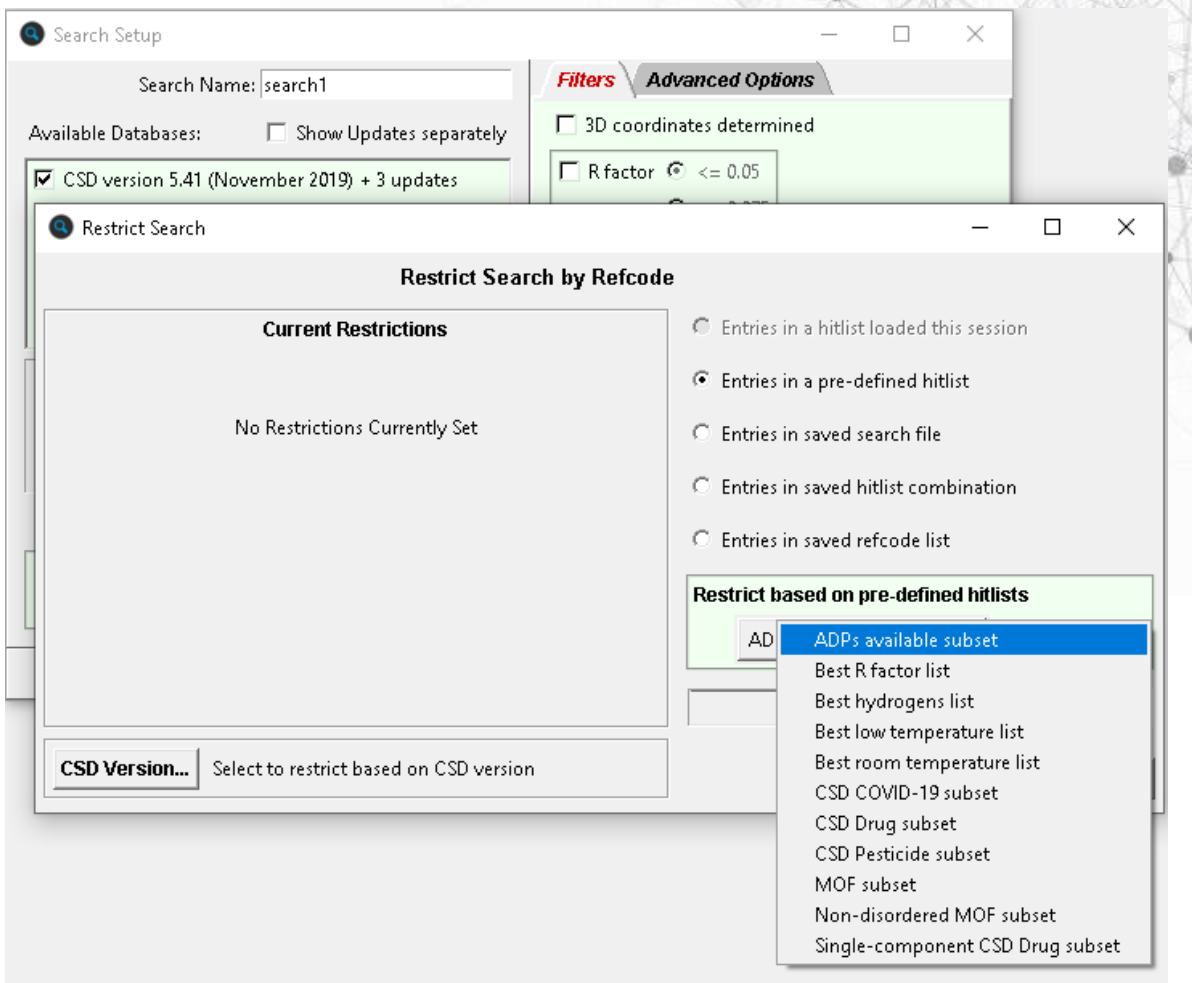


- Once a subset has been loaded into ConQuest, it can be combined with other searches

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'
```

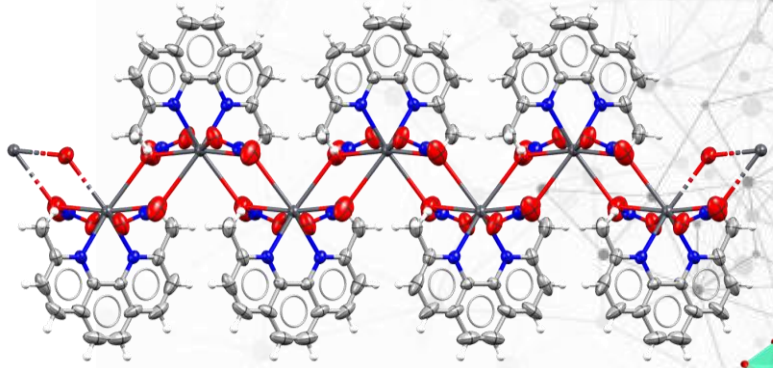
Future work

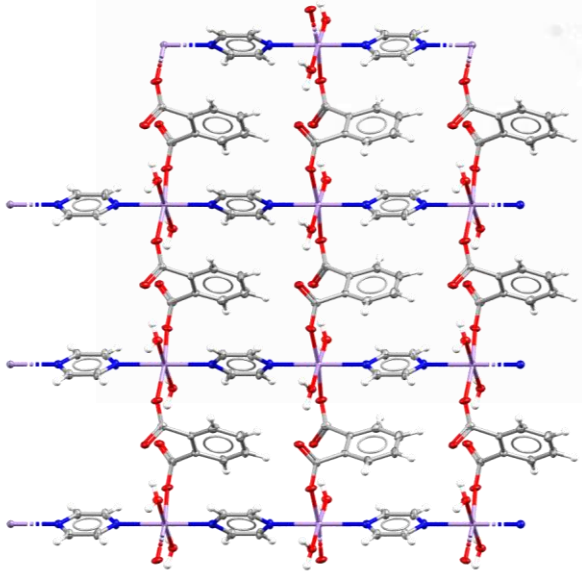
- Additional MOF subsets
  - 1D, 2D and 3D frameworks
- Based on the publication

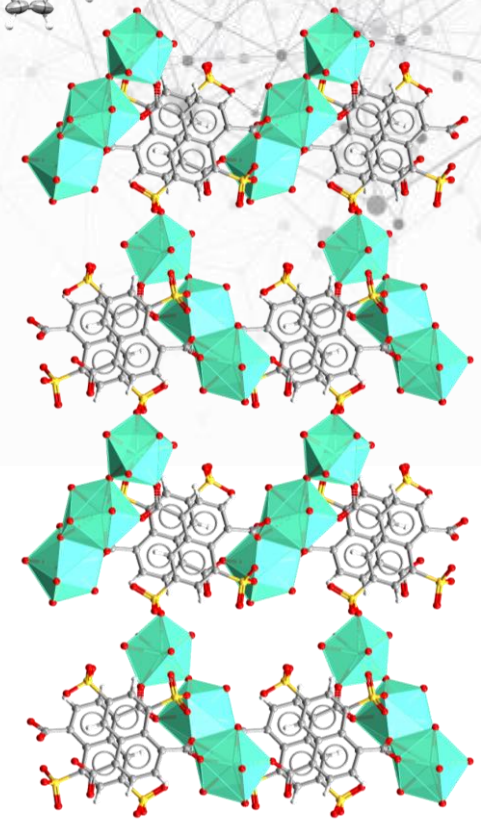
  Targeted classification of metal-organic frameworks in the Cambridge Structural Database (CSD)

  Peyman Z. Moghadam, Aurelia Li *et al* , *Chemical Science*, 2020, **11** 8373, **DOI:** 10.1039/D0SC01297A

- Please give us your feedback and suggestions!







CCDC

How to use CSD Subsets in ConQuests?

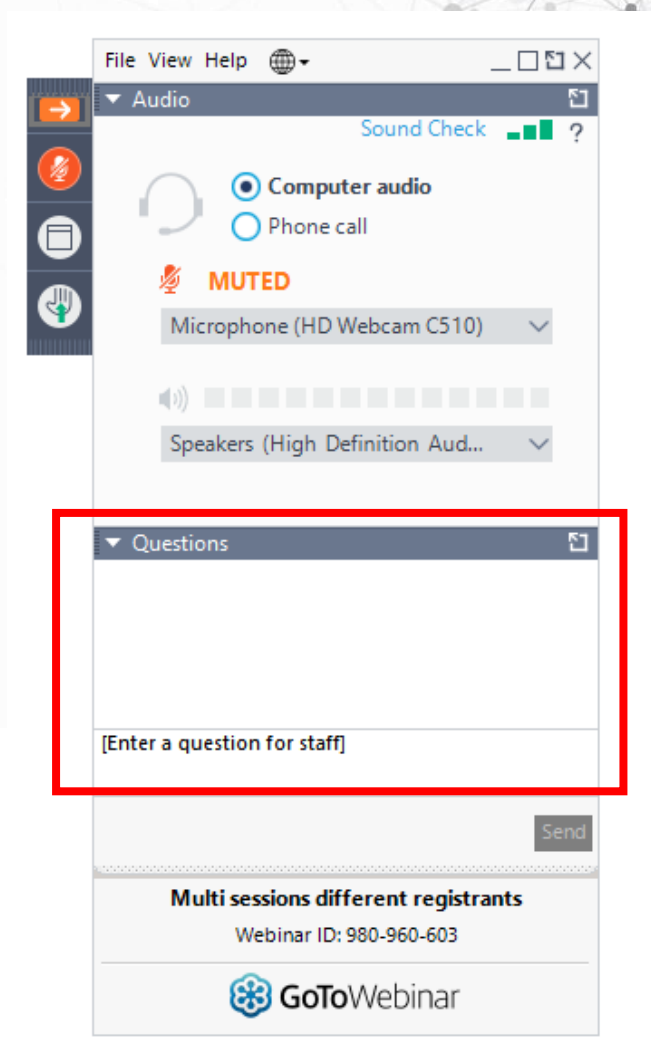
- We recently published a new tutorial video on our YouTube channel to help you find specific entries in the CSD such as drugs, pesticides, MOFs, or best representative structures by showing you how to use subsets in ConQuest. The demonstration includes how to load a subset, how to search structures 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

CCDC

Q&A

- Type your questions in the box as shown

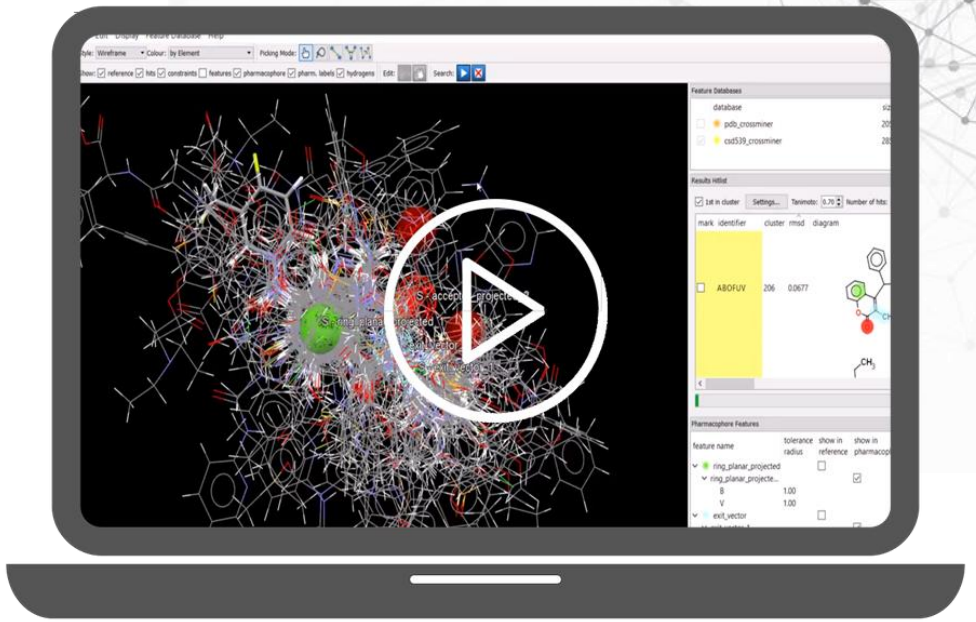


CCDC

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