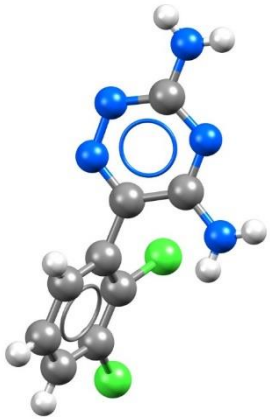
Using Subsets in ConQuest
(CQ-004)

Developed using
2023.1 CSD Release
CSD version 5.44 (April 2023)



CCDC
advancing structural science

Table of Contents

Introduction

ConQuest is the desktop search interface to the Cambridge Structural Database (CSD). All textual, numeric and structural data stored within the CSD can be searched using ConQuest. ConQuest provides an extensive range of flexible search options including searching based on compound name, formula, elemental composition, and literature search to name a few. With over a million entries in the database, individual structures can be missed. Subsets break the database down into more manageable subcategories and are an efficient way to search a certain category of compounds. As of the 2022.1 release, the latest additions to the subsets are the Election diffraction, High pressure, Hydrates, Polymorphs and three *MOF* dimensionality subsets. The teaching subset is also now available in ConQuest.

Before beginning this tutorial, ensure that you have a registered copy of CSD-Core or above installed on your computer. Please contact your site administrator or workshop host for further information.

Learning Outcomes

This tutorial will guide you using curated subsets in the CSD. At the end of this tutorial, you will be able to:

- Locate and load subsets.
- Search structures at the intersection of two subsets.
- Restrict a search to a chosen subset.

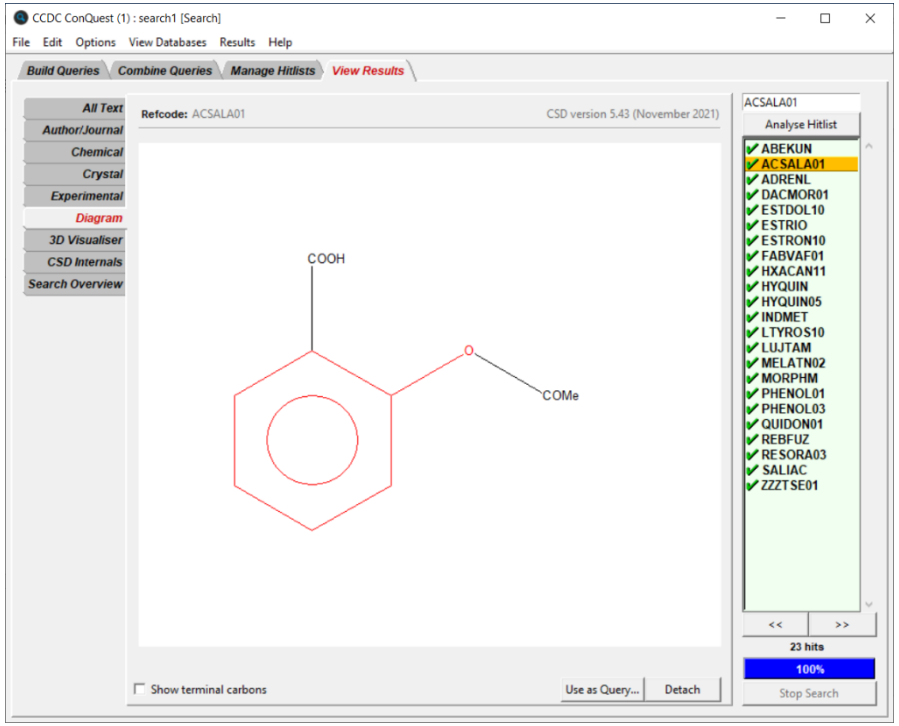
This workshop will take approximately ***45*** minutes to be completed. The words in *Blue Italics* are reported in the Glossary at the end of this handout. A review of the ConQuest interface is also at the end of this handout.

Pre-required Skills

For this tutorial, we recommend working through the introduction to ConQuest, and advanced searching in ConQuest workshop which can be found here (workshop code CQ-001 and CQ-002).

Materials

There are no additional materials required for this workshop.



Loading subsets

There are three ways to use subsets; you can load subsets in ConQuest to view the structures, restrict your search to a specific subset, or through the CSD Python API. For this workshop, we will focus on the first two methods. If you are interested in accessing subsets through the API, we recommend looking at the API workshops and user guide.

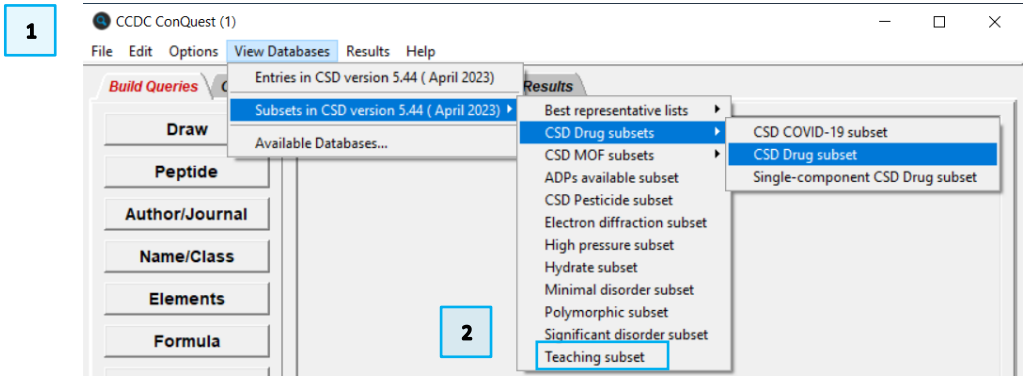
For the purposes of this exercise, we will assume that you are interested in CSD entries with a specific functional group in the *CSD Drug subset* (containing the approved Drug list by the DrugBank) and in the *CSD teaching subset*.

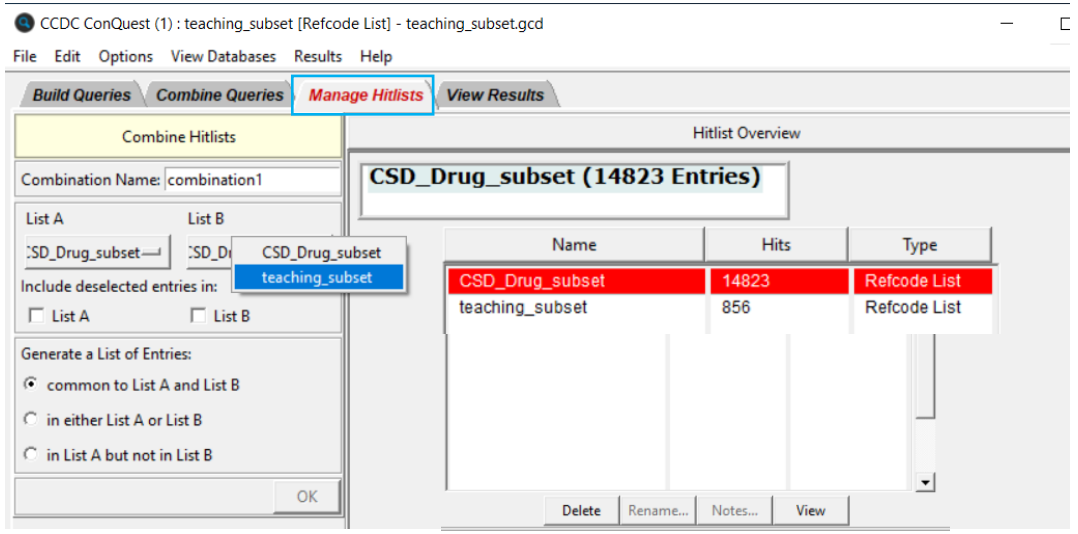
1. Launch ConQuest by clicking the ConQuest Icon on your desktop or launching it from the Start or Applications menu. In the main ConQuest window, click on **View Databases** then click on *Subsets in CSD version 5.44 (April 2023)*. A list of available subsets will be displayed.

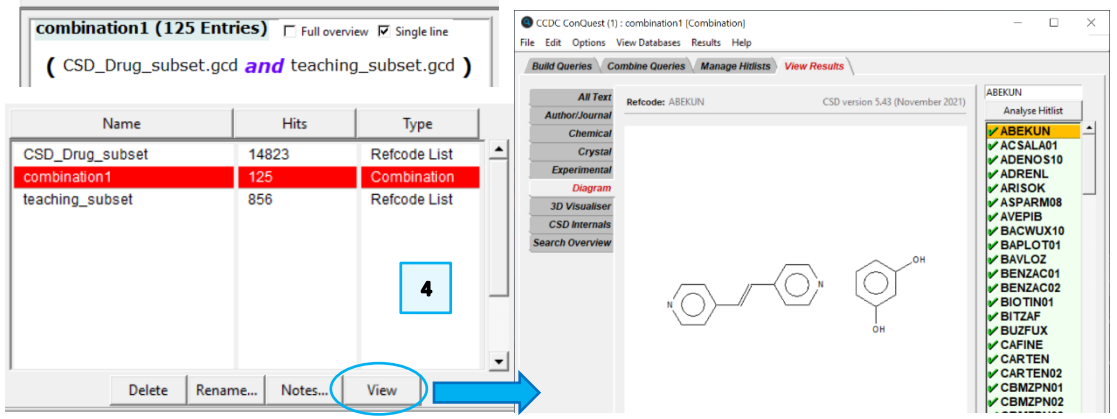
2. You will first load the drug subset by clicking on *CSD Drug subset*. It will take a couple of seconds to load and there should be at least 14,823 structures in the subset. Repeat the procedure to load the *Teaching subset*. There should be at least 856 entries. With both subsets loaded, *how can you determine how many structures are in both subsets*?

3. You can answer this question by using the combine *hitlist* function from the Manage Hitlists tab. Click on the *Manage Hitlists* tab. Under the *Combine Hitlists* menu, List A should be CSD_Drug_subset. In the drop-down box under List B, choose Teaching_subset. To find out how many structures are in both subsets, select "common to List A and List B" under *Generate a List of Entries*. Click **OK**, and a new hitlist named "combination1" should be generated with at least 125 hits. Tick the single line box to see the components of the combination.

4. To view the structures in "combination1" in the main ConQuest window, Click the **View** button.







Restricting searches to a subset

We are interested in looking for a structure with a phenol functional group that belongs in both the drug and teaching subset. *How can we conduct such search, limiting the search to the results from the combination in the previous step*?

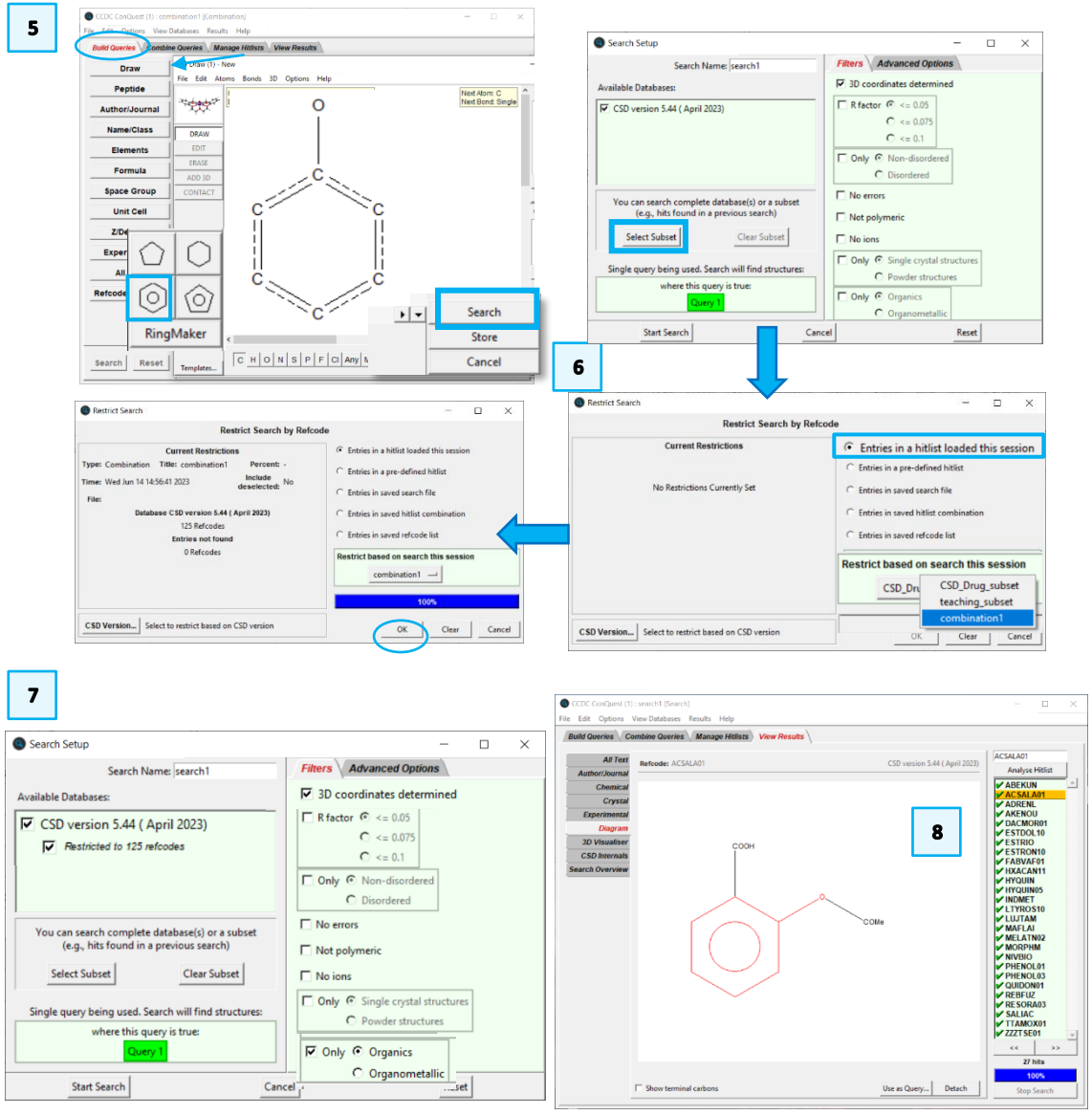
5. In the main ConQuest window, click on the *Build Queries* tab, click on the **Draw** button, and sketch the motif in the figure to the right, which is an aromatic ring with an oxygen attached with a single bond to one of the carbons. To sketch the aromatic ring, you can use the pre-designed one on the left-hand side of the Draw window. If you need a refresher on the Draw window, We recommend revisiting the CQ-001 workshop. Once you are done sketching, click **Search** to proceed to the *Search Setup* window.

6. The *Search Setup* window is where we can select the subset to be used for the search. Click the **Select Subset** button to open the *Restrict Search* window. The subset we want to load is the "combination1" hitlist that was previously generated. Since the hitlist was generated this session, select *Entries in a hitlist loaded this session*. There should be a drop-down menu to select which of the hitlists generated this session you want to use. Click on the menu and select ***combination1***. Once the hitlist is loaded, click **OK** to close the window.

7. Now in the *Search Setup* dialogue box, note that under *Available Databases*, there is an additional line below the CSD version that indicates the search is restricted to 125 refcodes. Make sure both boxes are ticked. Be sure to also tick the boxes for "3D coordinates determined" and "*Only Organics*" to apply these filters. Then click **Start Search** to begin the search.

8. At least 27 structures should be returned[1]. Shown in the diagram is a structure of Aspirin (ACSALA01).

We have drawn a phenol substructure, but we can look at structures with a different functional group in the drug subset. We might want to probe within a



[1] As of version 2023.1 of the CSD. Additional structures might be included in future versions.

subset of the CSD drug subset: the ***CSD COVID-19 subset*** for structures with an *amide* functional group.

9. We will conduct a new search for structures with amide groups in the CSD COVID-19 subset. These are molecules of interest in the fight against COVID-19. Click on the *Build Queries* tab and click on the **Name/Class** button to start a text-based search. Type the word *amide* in the "Compound Name" box. Click **Add** to include the word amide in the "Contains:" box. Click **Search** to begin the search.

10. In the *Search Setup* dialog box, we will select the subset we want to search from. Click on the **Select Subset** button to open the *Restrict Search* window. In this window, select *Entries in a pre-defined hitlist* to load the CSD defined subsets. Click on Choose a subset…, from the drop-down menu, select **CSD Drug subsets > CSD COVID-19 subset**. Once the subset is loaded, click **OK** to close the window.

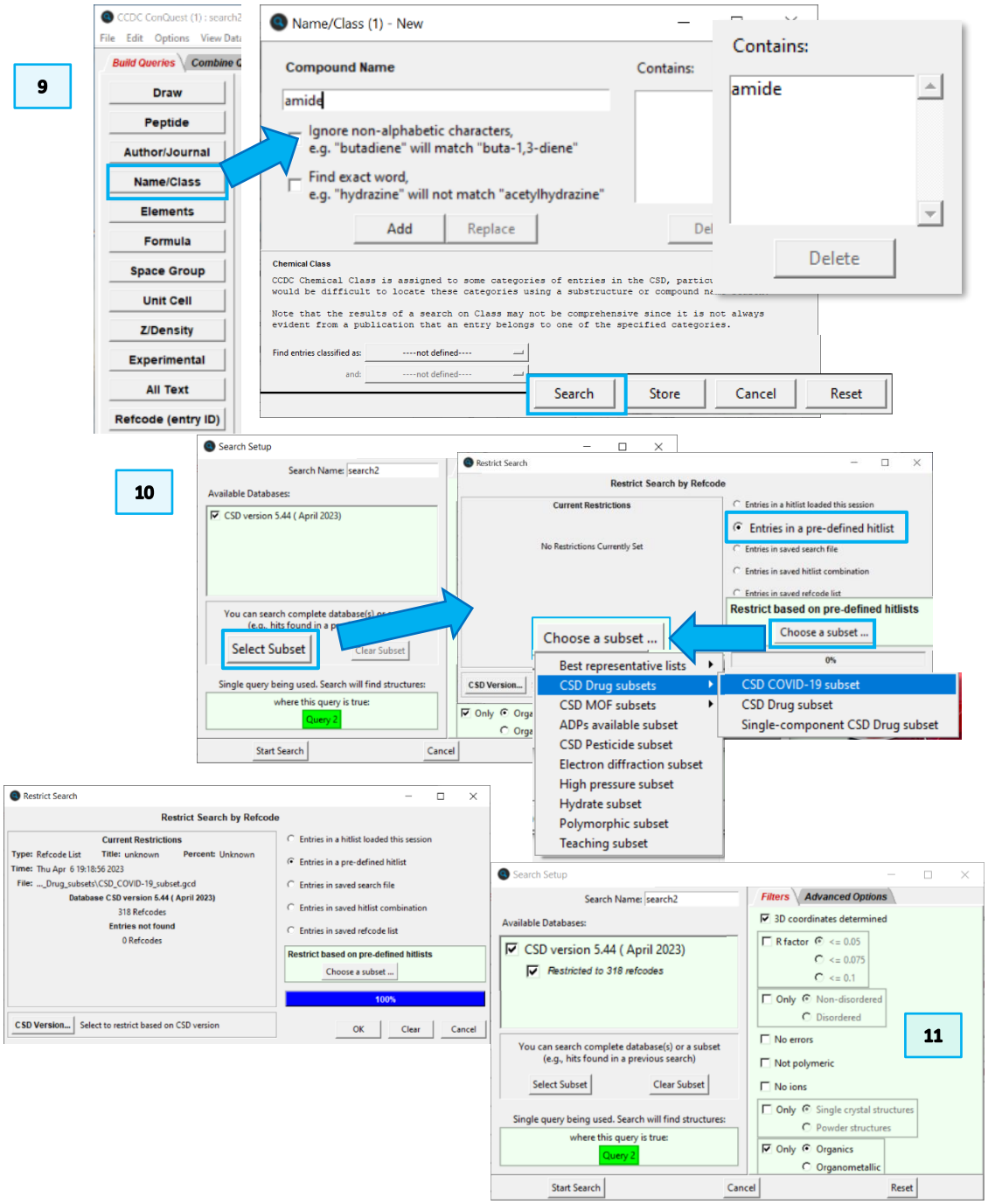
11. In the *Search Setup* window, note that under *Available Databases*, there is an additional line below the CSD version that indicates the search is restricted to 318 refcodes, which is the number of structures in the CSD COVID-19 subset (note: this number might change based on your CSD version). Make sure to tick the boxes for "3D coordinates determined" and "Only Organics" to apply these filters. Then click **Start Search** to begin the search.

12. There should be at least 35 structures returned. You can look through the results to assess where the amide groups are located and use such information to either refine or expand future searches.

Conclusions

Now you have seen two ways to use subsets in your ConQuest searching. We have looked at three subsets in the CSD for this workshop, however, there are 11 subsets in total[2] that you can explore. For example, the MOF subset can be extremely useful if you want to search for MOFs but are unsure of how to draw a substructure that will find the right results.

[2] As of 2023.1 release.



Exercises

- Can you find structures that are common to the CSD Drug and *MOF* subset?
- How would you conduct a name search for other functional groups in the CSD COVID-19 subset?
- Can you export your results to Mercury and conduct data analysis on the structures? *For this exercise you can refer to the workshop CQ-003.*

Summary

This workshop introduced how to use subsets in ConQuest. You should now be familiar with:

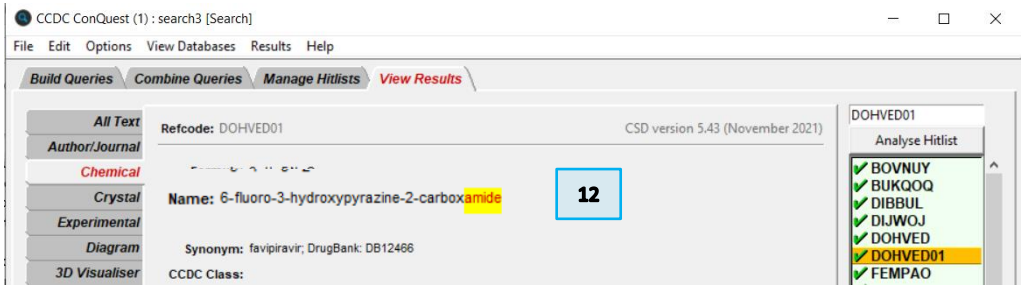
- Loading pre-defined subsets.
- Combining subsets.
- Restricting searches to a specific subset or to a combination generated during the session.

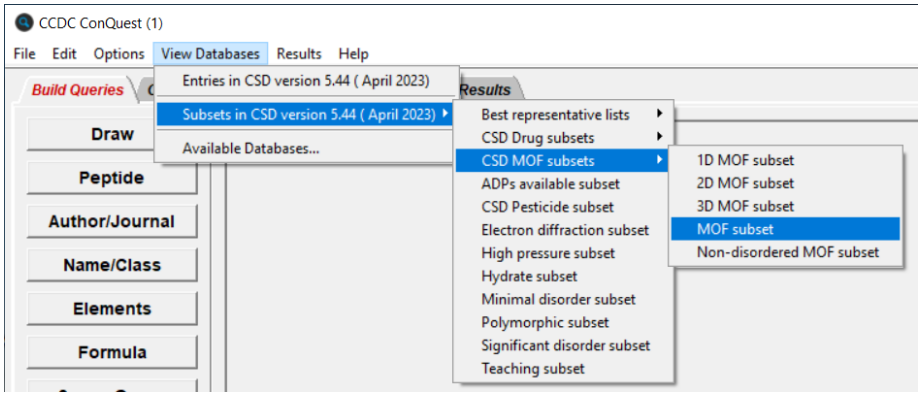
Next Steps

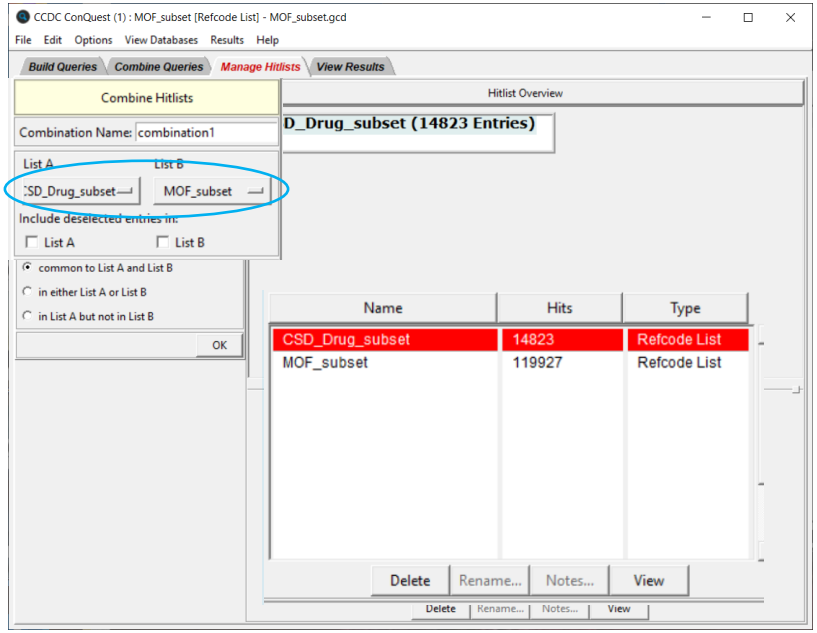
Advanced ConQuest workshops can be found here (https://www.ccdc.cam.ac.uk/community/training-and-learning/workshop-materials/csd-core-workshops/). The ConQuest user guide and other available ConQuest documents can be found here.

Feedback

We hope this workshop improved your understanding of ConQuest and you found it useful for your work. As we aim to continuously improve our training materials, we would love to get your feedback. Click on this link to a survey, it will take less than **5** minutes to complete. The feedback is anonymous. You will be asked to insert the workshop code, which for this self-guided workshop is CQ-004. Thank you!







Glossary

Amide
An amide is a compound with the group RCONR'R''.

CSD COVID-19 subset
The subset contains CSD entries that have been reported as small molecule drug candidates for the treatment of COVID-19.

CSD Drug subset
Subsets that contain approved drug compounds as defined by the DrugBank database (https://www.drugbank.ca/). Further information can be found here: 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., 108 1655-1662, 2019 [DOI: 10.1016/j.xphs.2018.12.011] CSD COVID-19 subset

Hitlist
A hitlist is a subset of CSD entries which can include search results, refcode lists, or the results of combining these.

Metal-Organic Frameworks (MOFs)
A MOF, or metal organic framework, is a material formed of metal clusters or nodes, linked by organic molecules.

Only Organics
Selecting this filter eliminates from the search any structure that contains a transition metal, lanthanide, actinide or any of Al, Ga, In, Tl, Ge, Sn, Pb, Sb, Bi, Po.

Teaching subset
The CSD Educational Collection of structures containing 700+ structures carefully selected to enhance chemistry learning is now available to view in ConQuest (having previously been available in Mercury)

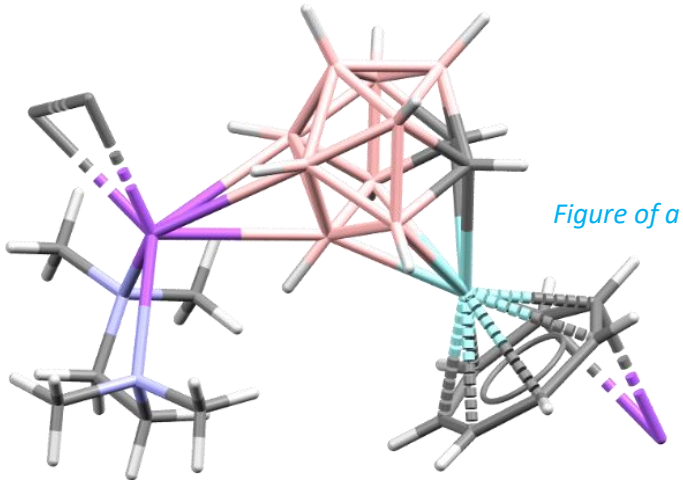
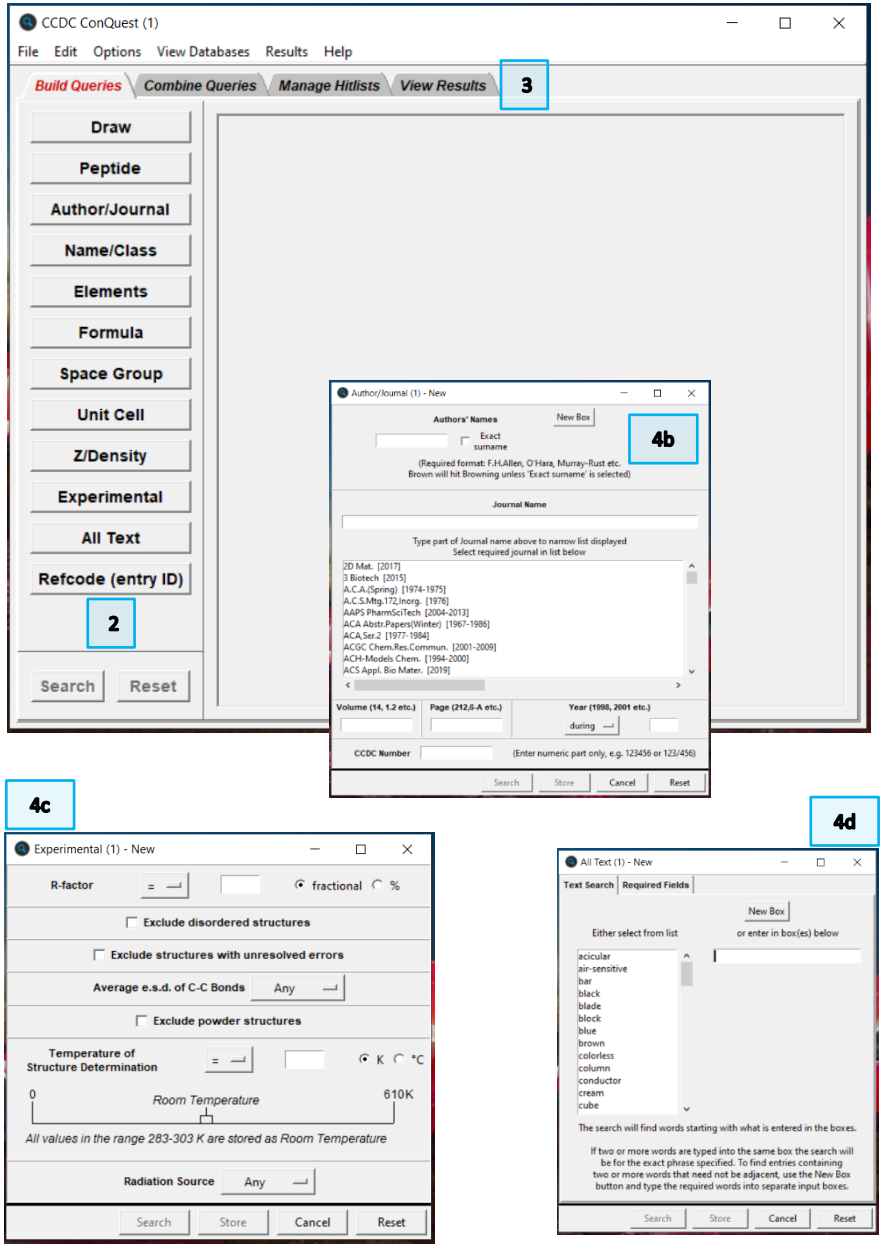


Figure of a MOF, CSD-ABUREX.

Review. ConQuest Interface

1. Launch ConQuest by clicking the ConQuest Icon on your desktop or launching it from the Start or Applications menu.
2. The ConQuest main window shows all the search routines you can perform on the left-hand side of the window.
3. The row of tabs across the top of the window will guide you through the steps of the search process.
4. Some example searches are
    a. Draw – substructure and 3D information searching
    b. Author/Journal – bibliographic searching
    c. Experimental – experimental set up searching
    d. All Text – generic text-based searching
5. The majority of the searching we will do in these tutorials will be substructure searching, so we will focus on the Draw tab here.



Review. Draw Window

All drawing takes place in the central white area of the *Draw* window. In addition to creating 2D chemical structure sketches, the *Draw* window allows for the inclusion of 3D parameters for searching or for filtering.

ConQuest sketching conventions

- Left click in the sketcher to insert the selected atom type
- Left click and drag to sketch two bonded atoms
- Use the **Edit** button to modify properties of or delete atoms, bonds or entire substructures
- Right-click on atoms or bonds to modify their properties
- Use the **Templates…** button to pick from a list of CSD editor devised and drawn substructures
- Use the **More…** button to find less frequently used element types, or generic atom type groups (e.g. halogens), or define custom element combinations (e.g. C or N or O).