Searching and Subsets in ConQuest

**Developed using
2022.1.0 CSD Release
CSD version 5.43 (November 2021) + 1 update**
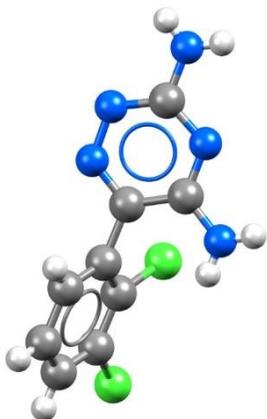
Table of Contents

CCDC
advancing structural science

Introduction

ConQuest is the desktop search interface to the Cambridge Structural Database (CSD). All text, 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 is 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 workshop, ensure that you have a registered copy of CSD-Core or CSD-Enterprise installed on your computer. Please contact your site administrator or workshop host for further information.

Objectives

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

- Draw a 2D structure in ConQuest and conduct a basic search.
- Restrict a search to a chosen subset.
- Locate and load subsets.
- Search structures at the intersection of two subsets.

This workshop will take approximately **25** minutes to be completed.

Note: The words in *Blue Italic* in the text are reported in the Glossary at the end of this handout. A *review* of the ConQuest interface and of the ConQuest Draw window is also at the end of this handout.

Note: Your results could differ based on the CSD version that you have installed. Do not worry if you are on a different version and get a different number of structures in your hitlists. A similar order of magnitude could be expected and don't forget to always approach the results critically!
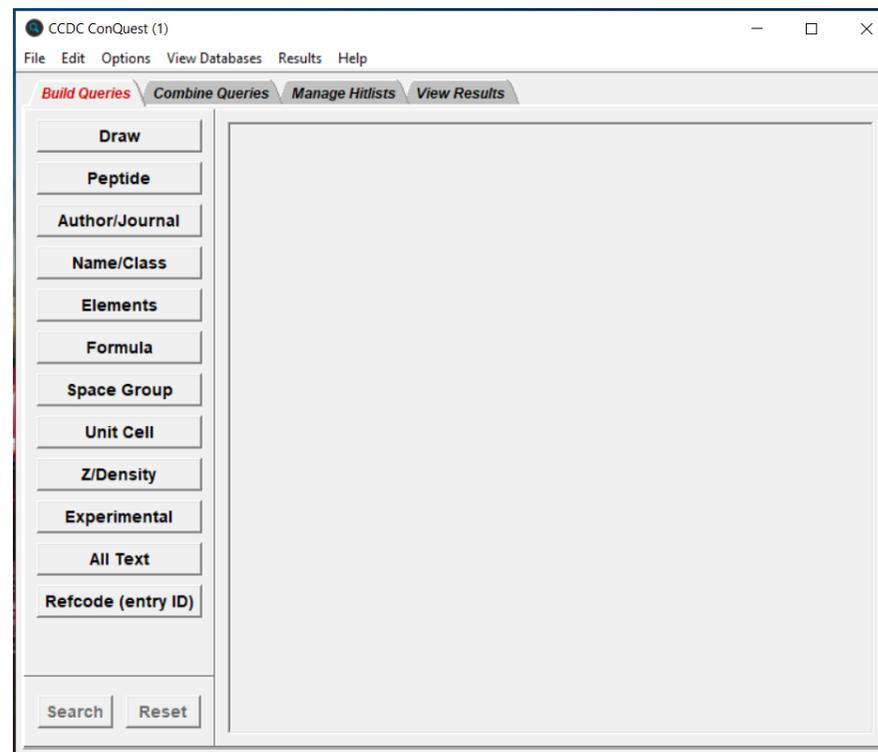
Pre-required skills
This is an entry-level session; no previous experience is required.
Materials
There are no additional materials required for this workshop.

Example 1. Basic Searching in ConQuest

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.

For this example, we will sketch and search for fragments (a) and (b). If you complete the workshop early and are interested, you can draw the remaining three fragments and conduct a search.

This example will demonstrate sketching techniques in ConQuest.

1. Launch ConQuest by clicking the ConQuest Icon on your desktop or launching it from the Start or Applications menu and open the sketcher by clicking the **Draw** button.

2. Sketch a simple alcohol fragment. First, add a carbon atom by clicking anywhere in the drawing area, making sure you have carbon selected ("C" in the white square at the bottom of the draw window). Then, add an oxygen atom to it with a single bond. To do so, select "O", oxygen, from the elements at the bottom of the draw window, make sure the bond type selected is "Single", and click and drag from the carbon atom.

3. To search for fragments with one hydrogen bonded to the oxygen, right click on the O, choose *Hydrogens* and then *1* from the drop-down menu to add the H atom.

4. To ensure that the carbon atom is bonded to at least 2 hydrogen atoms and one other atom in addition to the OH group, first add two H atoms as in **Step 3** above, but this time select 2 hydrogens. Then right-click the C again and choose *Number of Bonded Atoms*, and then *4* from the drop-down menu. This will produce fragment "a" from the list show in the introduction to this exercise.

5. To run the search, click the **Search** button at the bottom right of the Draw window.


(a)   (b)   (c)   (d)   (e)
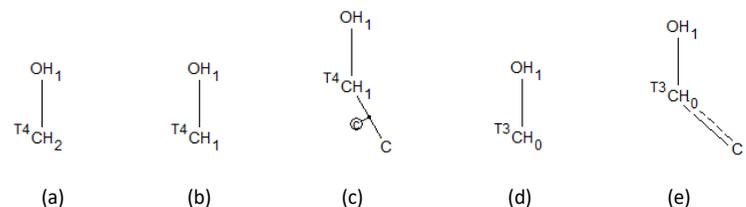
These subtly different definitions would retrieve the following types of functional group:

(a) primary alcohols
(b) cyclic and acylic secondary alcohols
(c) cyclic secondary alcohols only
(d) unsaturated cyclic and acyclic alcohols and carboxylic acids
(e) unsaturated cyclic alcohols only e.g. phenols
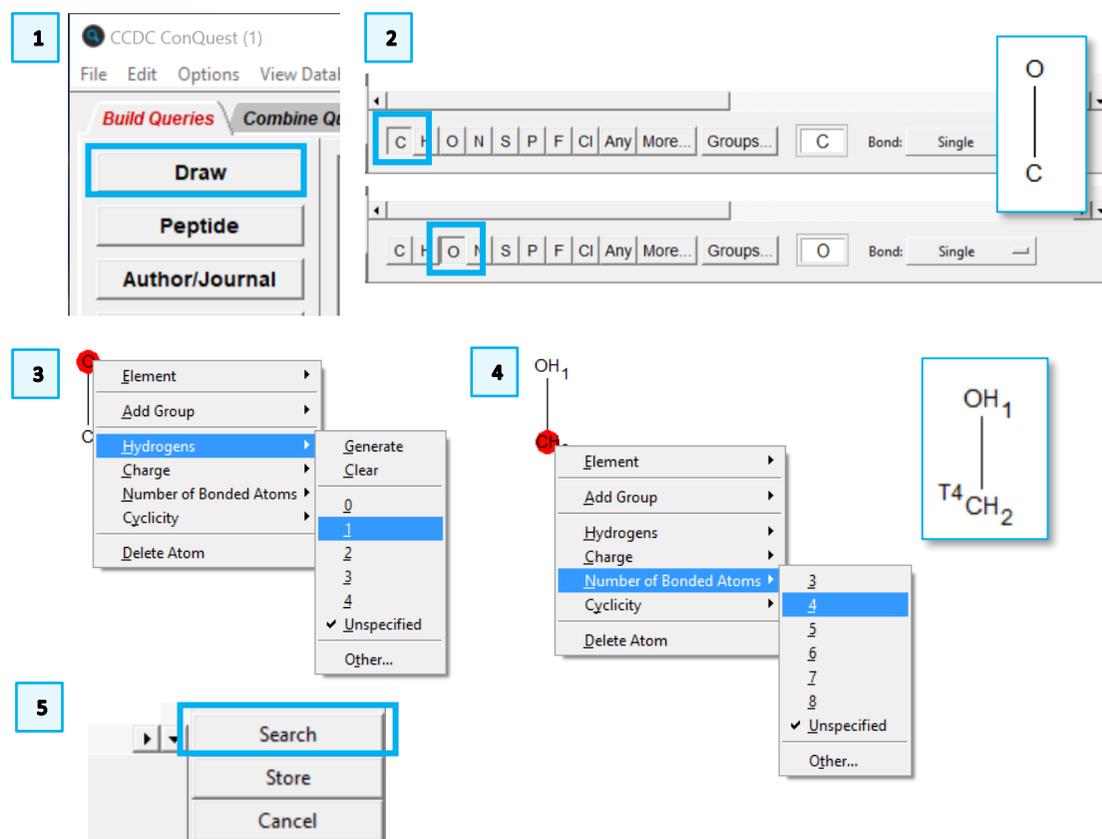
In the **Search Setup** pop-up, tick the boxes for *3D coordinates determined* and *Only Organics* (make sure both the box next to *Only* and the one next to *Organics* are selected). Then click the **Start Search** button.

6. Based on your CSD version, you should get results in the order of or over 16,901 with this query. Use the arrow keys to scroll through the refcode list to view the results. Explore the results using the tabs on the right-hand side, for example to see the 3D structure, or learn more about the crystal structure and the experimental conditions. Observe how the fragment we searched is highlighted in the 2D diagram and on the 3D structure. *What do you find*?
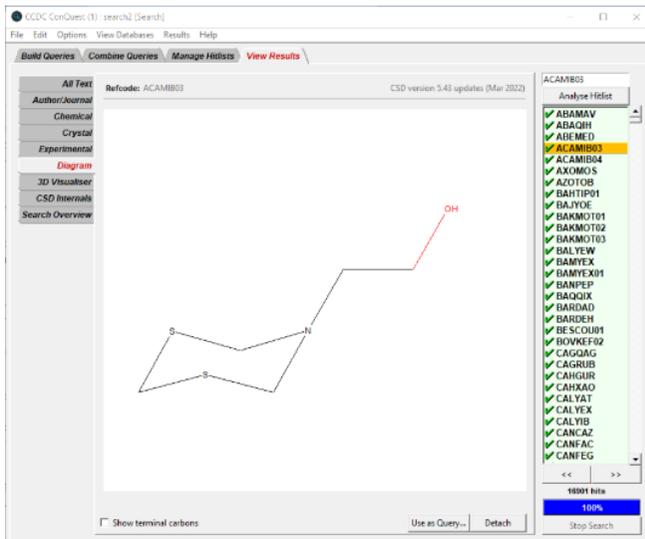
7. When you are done looking at your results, click the *Build Queries* tab to return to the Query window. You should see your first search fragment listed as Query 1.

8. We are now going to search for a secondary alcohol, as shown as fragment b in the list at next to the introduction. We will save time, by editing the drawing that we did for our first search. Click the **Edit…** button in the query window to return to the **Draw** window.

9. Right-click on the C atom and choose *Hydrogens,* and *1* from the drop-down menu as in **Step 3** above. Leave the *Number of Bonded Atoms* set to 4 to match a secondary alcohol (fragment "b").
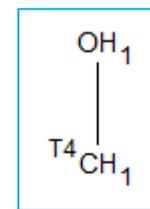
10. Click **Search** as in **Step 5** above. A pop-up warning will ask if you want to overwrite the existing Query. Click **No** to save this as a new Query.

11. In the **Search Setup,** make sure the "3D coordinates determined" and "Only Organics" filters are ticked. Then, start the search.

12. Based on your CSD version, you should get in the order or over 30,650 hits with this new query. Use the arrow keys to scroll through the refcode list to view the results. *Observe: how are the hits different from the first search?*

Conclusion

You should now be familiar with sketching substructures using the ConQuest *Draw* window; how to edit searches; how to change bond and atom types; and how to view the data for each hit in your results list.
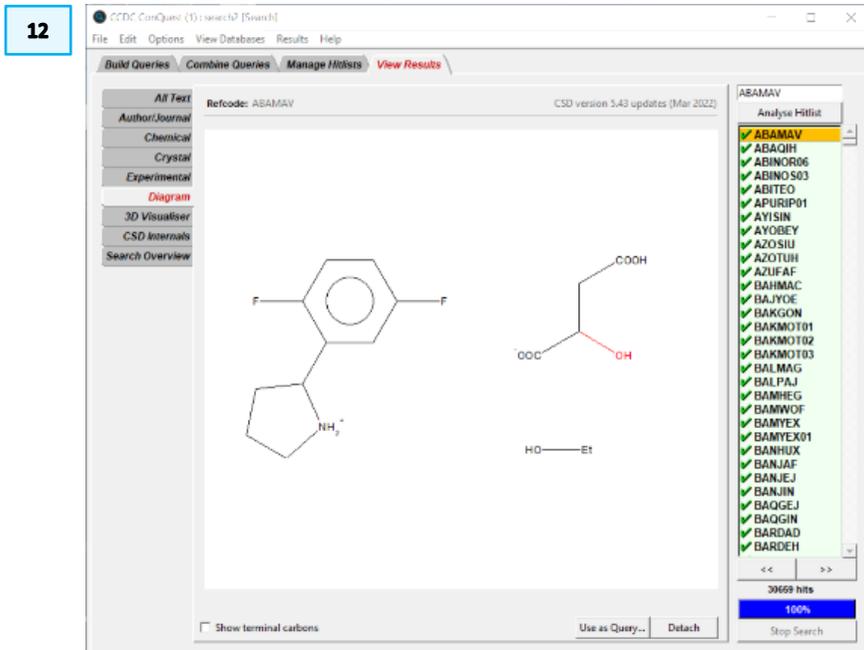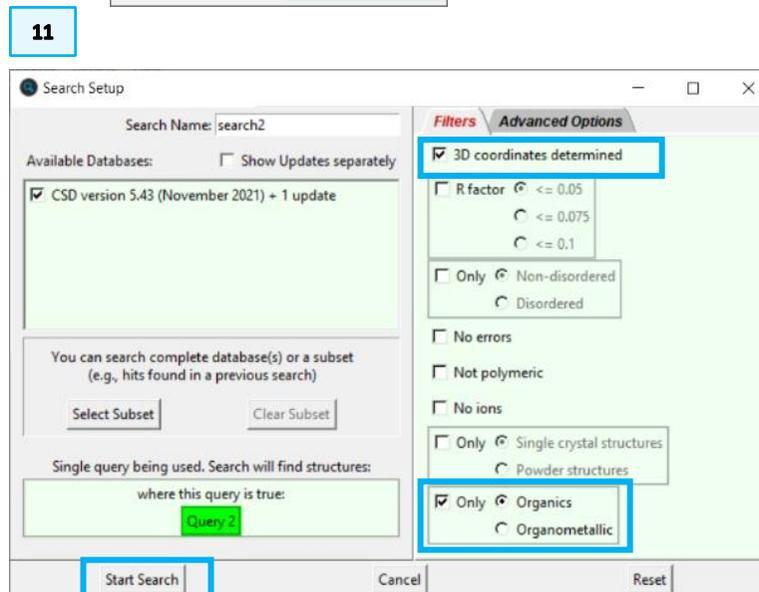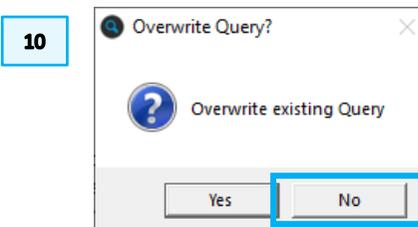
## Example 2. Restricting searches to a subset

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. To begin with, we will focus on how to restrict the search to a subset. You can find how to load subsets in the bonus exercise. If instead you are interested in accessing subsets through the API, we recommend looking at the API workshops on this page and user guide.

For the purposes of this workshop, we will assume that you are interested in CSD entries with "amide" in the compound name in the CSD COVID-19 subset.

1. 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.

2. 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.
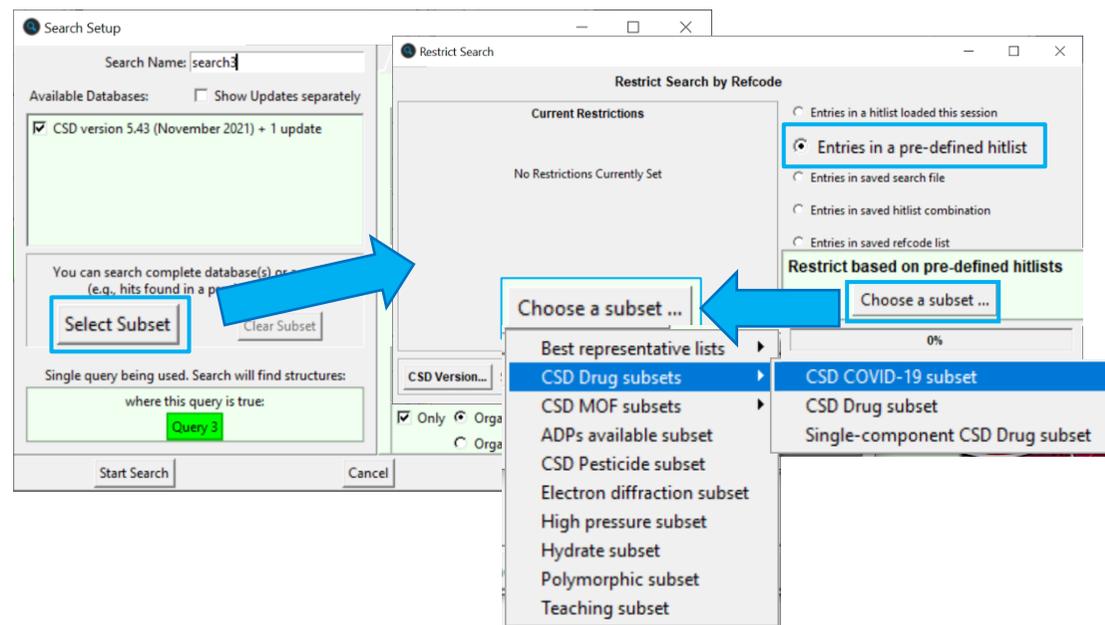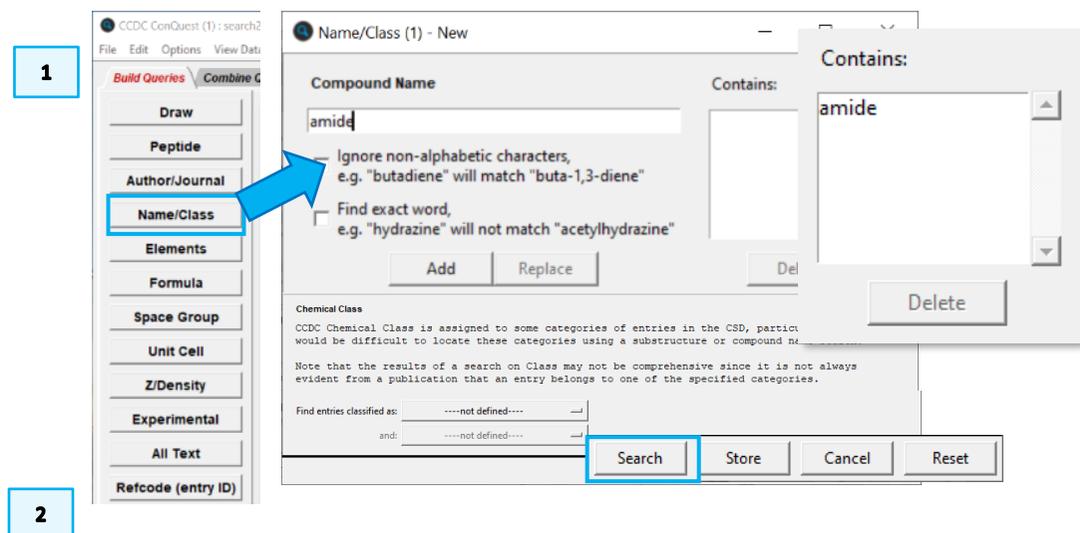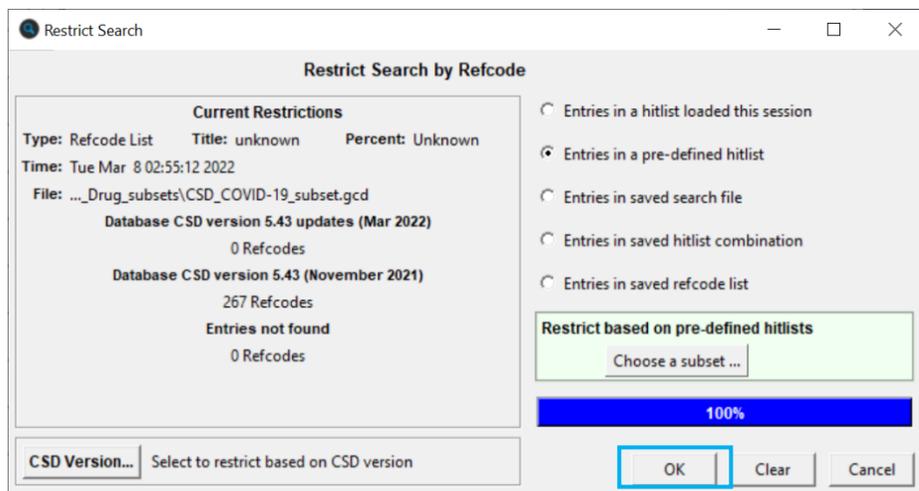
3. In the *Search Setup* window, note that under *Available Databases*, there is an additional line below the CSD version that indicates the search is restricted to 267 refcodes, which is the number of structures in the CSD COVID-19 subset (note: this number might change based on your CSD version). Make sure to tick the boxes for "3D coordinates determined" and "Only Organics" to apply these filters. Then click **Start Search** to begin the search.

4. There should be at least 31 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.
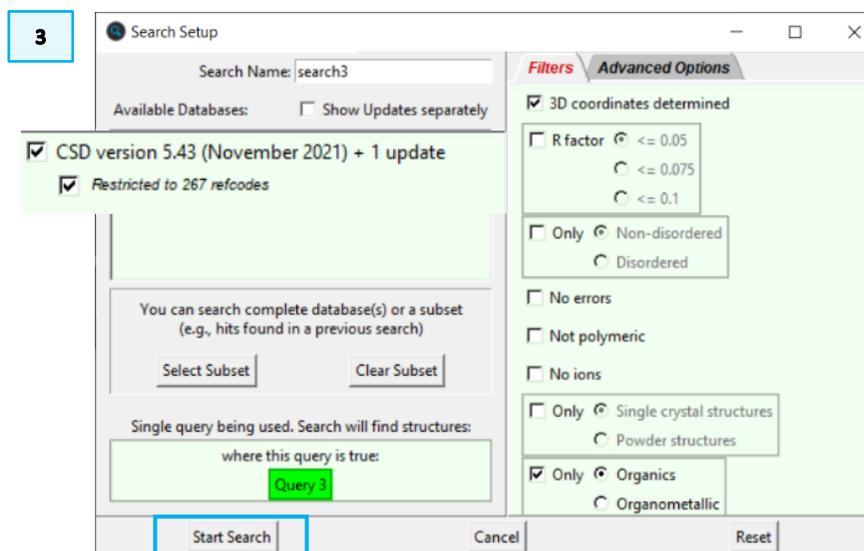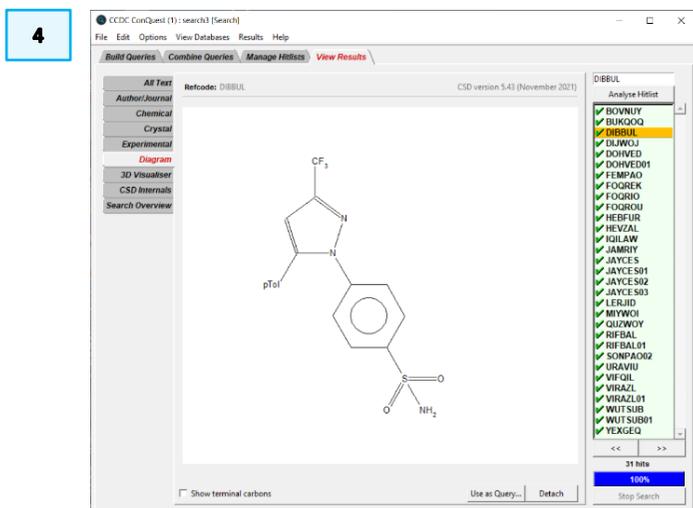
Conclusion

You have seen one way to use subsets in your ConQuest searching by restricting a search to the CSD COVID-19 subset. You have also seen how to create a query based on the compound name. There are many subsets that you can explore. For example, the MOF subset can be extremely useful if you want to search specifically for MOFs.
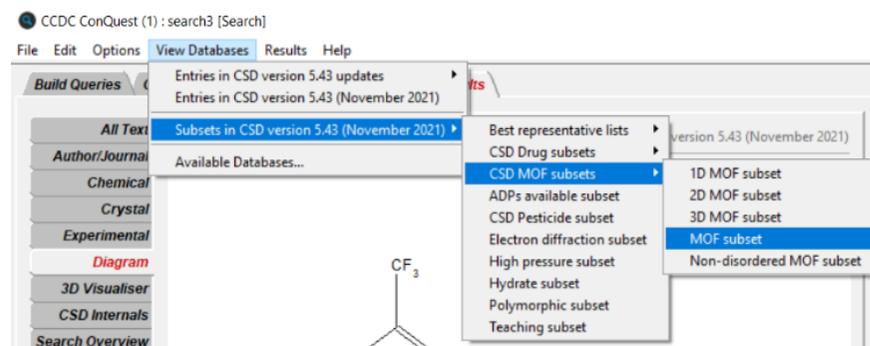
Further Exercises

- Can you sketch the remaining alcohol substructure from Example 1 and search the entire database?
- Can you find structures that are common to the Hydrate and MOF subsets? (*Hint*: see bonus exercise)
- How would you conduct a name search for other functional groups in the CSD COVID-19 subset?

(c)　　　　(d)　　　　(e)

Conclusions

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

- Sketching substructures using the ConQuest *Draw* window.
- Restricting searches to a specific subset.

In the bonus exercise you will see in more detail:

- Loading pre-defined subsets.
- Finding structures in multiple subsets.

Next steps

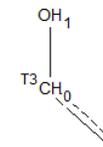After this workshop, you can explore more ConQuest exercises in the self-guided workshops available in the CSD-Core workshops area on our website: https://www.ccdc.cam.ac.uk/Community/educationalresources/workshop-materials/csd-core-workshops/. Please see the ConQuest documentation and other available ConQuest documents here .
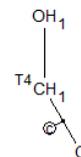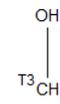
Bonus Exercise. Loading Subsets

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 bonus exercise, we will focus on the first method. If you are interested in accessing subsets through the API, we recommend looking at the API workshop and user guide.
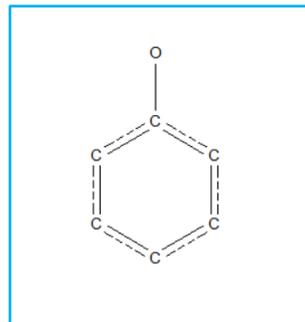
For the purposes of this exercise, we will assume that you are interested in CSD entries with a specific functional group, which are in the CSD Drug subset (containing the approved Drug list by the DrugBank) and in the CSD teaching subset.

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 5.43 (November 2021)*. A list of available subsets will be displayed.

2. You will first load the drug subset by clicking on *CSD Drug subsets* and then on *CSD Drug subset*. It will take a few seconds to load and there should be at least 13,700 structures in the subset. Repeat the procedure to load the *Teaching subset*. There should be at least 837 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 hitlists function from the Manage Hitlists tab. Click on the *Manage Hitlists* tab. We want to find out how many structures are in both subsets. Under the *Combine Hitlists* menu, List A should be CSD_Drug_subset. To select the teaching subset as List B, click on List B and in the drop-down box, choose Teaching subset. Under *Generate a List of Entries*, the radio next to "common to List A and List B" should be selected. Click **OK**, a new hitlist named "combination1" should be generated with 107 hits.

4. To view the structures in "combination1" in the main ConQuest window, make sure combination1 is selected (it will be highlighted in red), then click the **View** button.

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. 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 107 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 23 structures should be returned. Shown in the diagram is a structure of Aspirin (ACSALA01).
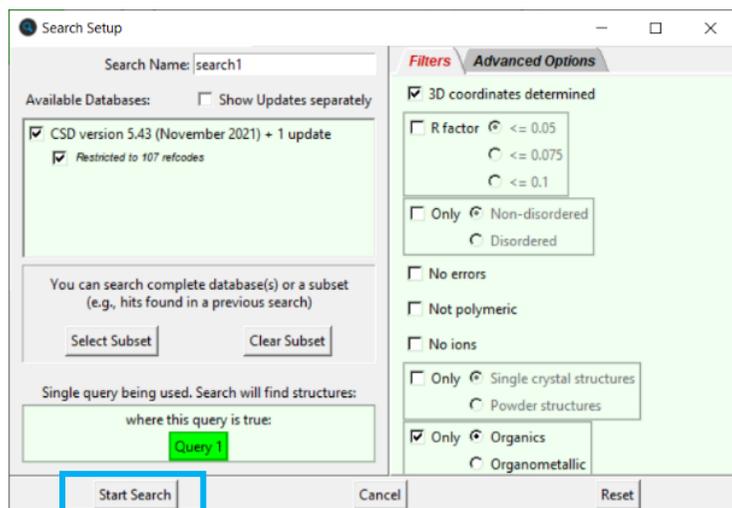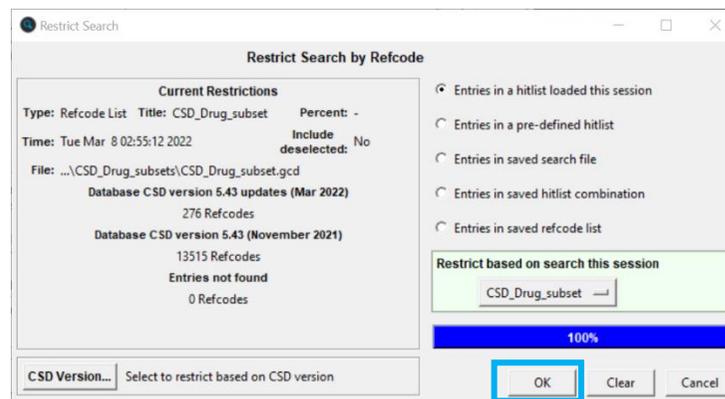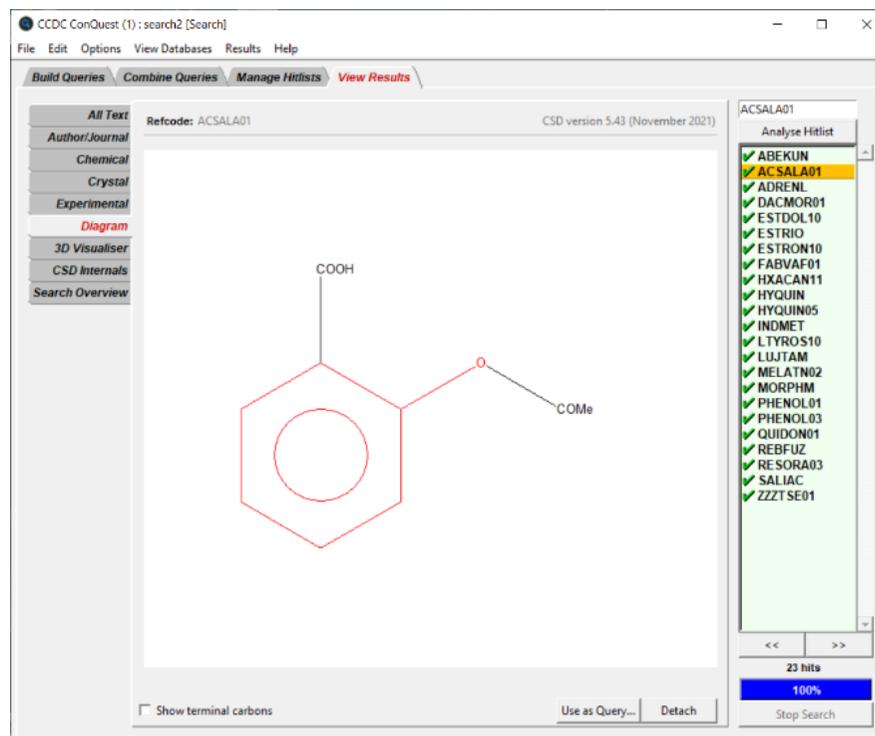
Glossary

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 Complex

A chemical compound that contains a metal and organic ligand.
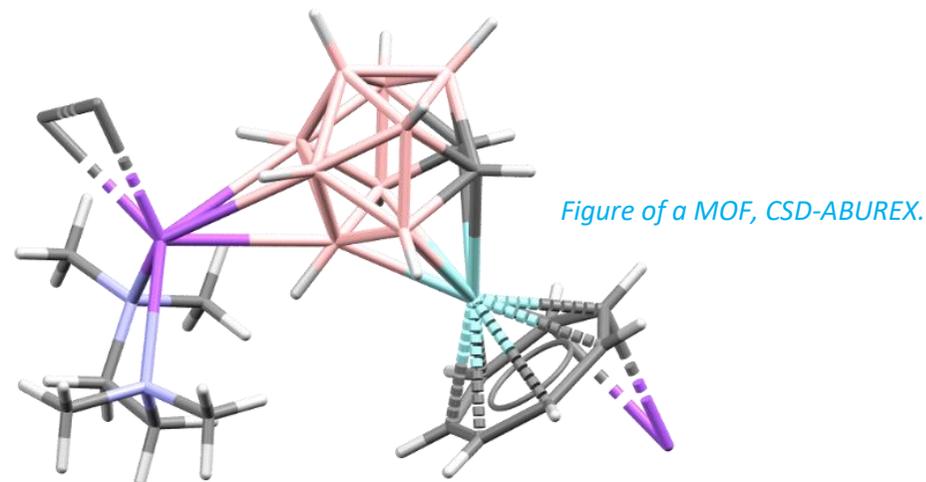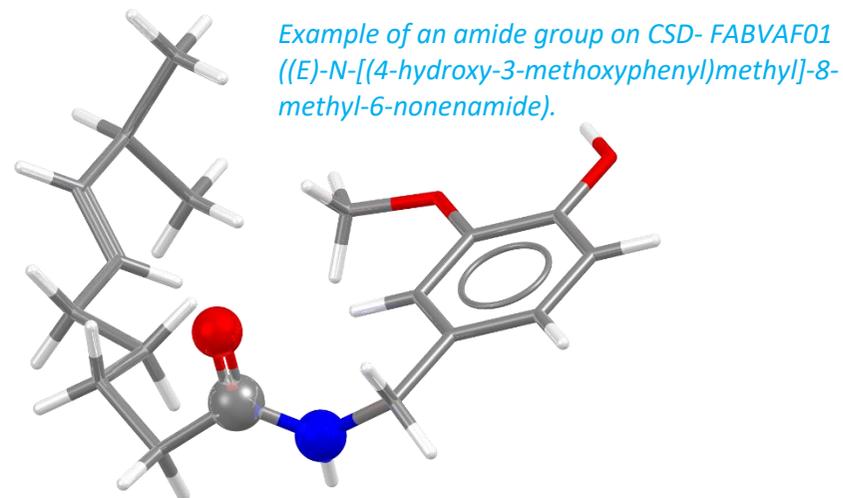
Metal-Organic Frameworks (MOFs)

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)

Example of an amide group on CSD- FABVAF01 ((E)-N-[(4-hydroxy-3-methoxyphenyl)methyl]-8-methyl-6-nonenamide).

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.
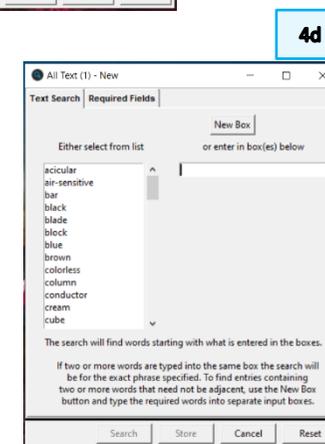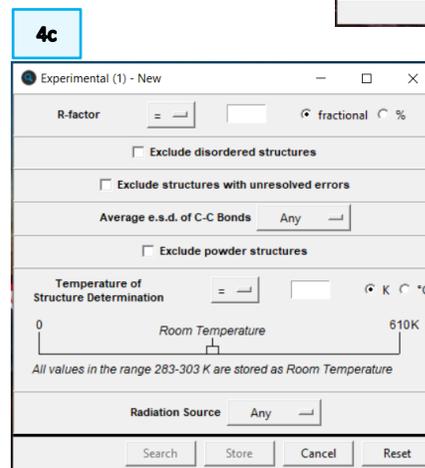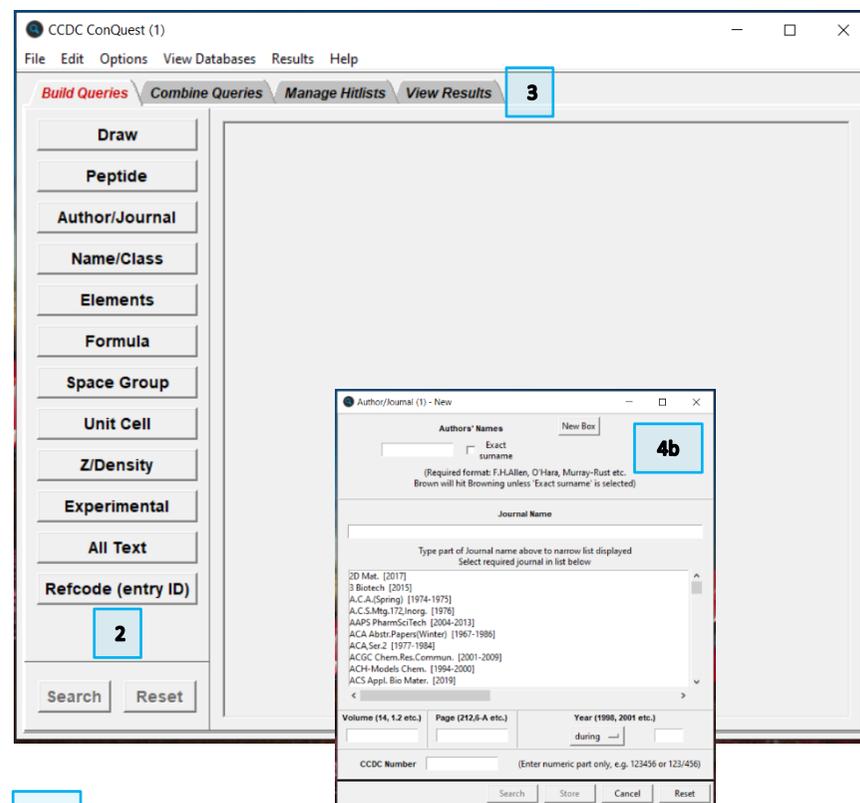
Review. Draw Window

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).