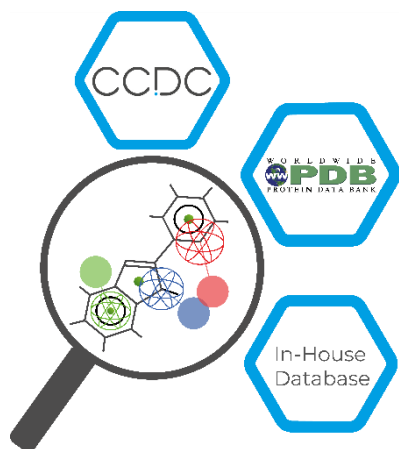


# Performing a Pharmacophore Search using CSD-CrossMiner (CROSS-002)

Developed using 2024.1 CSD Release



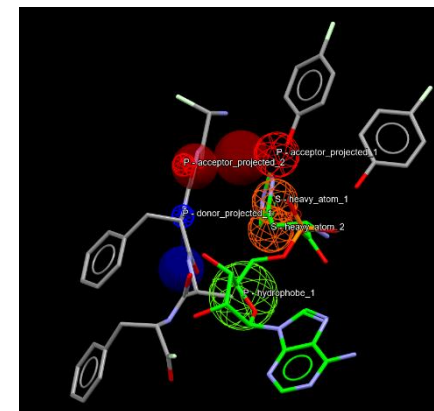
## Table of Contents

Introduction.....	2
Learning outcomes.....	2
Pre-required Skills .....	2
Materials.....	2
Searching with a Pharmacophore .....	5
Summary .....	10
Next Steps.....	10
Feedback .....	10
Glossary .....	11
CSD-CrossMiner terminology .....	3
Overview of CSD-CrossMiner .....	4

## Introduction

CSD-CrossMiner can be thought of as a pharmacophore-based query tool. However, it is much more powerful than traditional pharmacophore query tools as it allows you to query not only databases of ligands, but also proteins and protein-ligand interactions. CSD-CrossMiner includes a preconfigured database of biologically relevant subsets of the Cambridge Structural Database (CSD) and the Protein Data Bank (PDB). The pharmacophore used in the query is interactive, allowing you to easily edit it and in a number of ways through a simple user interface. This delivers an overall interactive search experience with application areas in interaction searching, scaffold hopping or the identification of novel fragments for specific protein environments.

The supplied feature database contains a subset of the CSD and PDB. The CSD subset consists of structures which are organic plus a small list of transition metals i.e., Mn, Fe, Co, Ni, Cu, Zn, have an R-factor of at maximum 10%, have 3D coordinates, have no disorder, and are not polymeric (more than 400 000 structures total). The supplied PDB database is divided in two subsets, one composed by protein-ligand complexes and another subset composed by protein-ligand-nucleic acids complexes. For the PDB subsets only the protein-ligand binding site and protein-ligand-nucleic acid binding site is provided, where the binding site is defined as all molecules with an atom within a 6 Å radius around the ligand (> 300 000 binding sites). For further discussion, please refer to the [CSD-CrossMiner User Guide](#) or the original paper: Korb O *et al.*, "Interactive and Versatile Navigation of Structural Databases" *J Med Chem*, **2016**, 59(9):4257, DOI: [10.1021/acs.jmedchem.5b01756](https://doi.org/10.1021/acs.jmedchem.5b01756).



## Learning outcomes

After completing this tutorial, you will:

- Understand the terminology used in CSD-CrossMiner to describe pharmacophores and understand how these features combine together to form a pharmacophore search.
- Be able to conduct a pharmacophore search and efficiently explore the search results.

This workshop will take approximately **30** minutes to be completed. The words in *Blue Italic* in the text are reported in the [Glossary](#) at the end of this handout.

## Pre-required Skills

This tutorial is geared towards the novice CSD-CrossMiner user who has Life Science experience. It covers the primary features, in searching across ligands, proteins, and ligand-protein interactions. Some of the results may vary depending on your version of CSD-CrossMiner.

## Materials

No additional materials are required to complete this workshop.

## CSD-CrossMiner terminology

### Exit vector

A two-point feature that represents a single, non-ring bond between two heavy atoms features; and it will be represented as two mesh spheres. In the case of CSD-CrossMiner, directionality in an exit vector does not matter.

### Features

An ensemble of steric and electronic features that characterise a protein and/or a small molecule. In CSD-CrossMiner a feature is defined as point(s), centroid or vector which represent a [SMARTS](#) query and, in the case of a vector, this includes geometric rules.

### Pharmacophore point

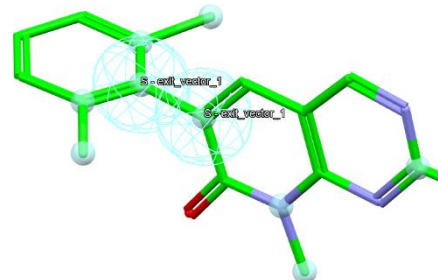
A feature that has been selected to be part of a pharmacophore because its presence is necessary to ensure the optimal supramolecular interactions with a specific biological target and to trigger or block its biological response.

### Structure database

Is a database containing the 3D coordinates of small molecule structures and/or protein-ligand binding sites. This database is used to create a feature database.

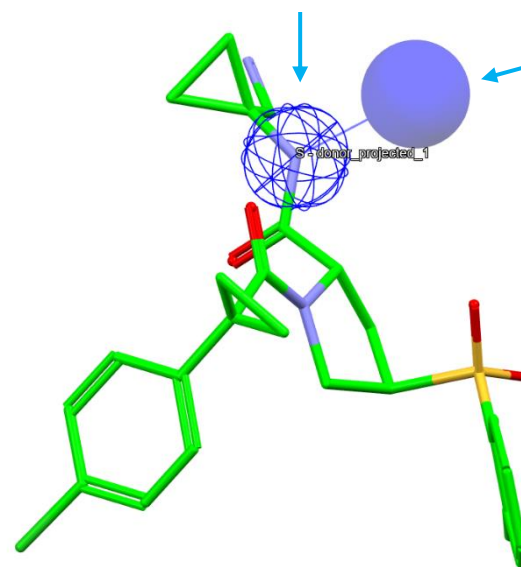
### Feature database

A database containing the structures from the structure database, indexed with a set of feature definitions provided by CSD-CrossMiner and any additional features defined by the user. This is the database that CSD-CrossMiner uses to perform the actual 3D search against a pharmacophore query.



*An exit vector (light blue mesh spheres) defined by the position of two carbon atoms.*

*Base point (the heavy atom of the donor group)*




*Virtual point - defines the direction the X-H group should point (Base point → Virtual point)*

*A molecule with a donor\_projected pharmacophore point defined.*

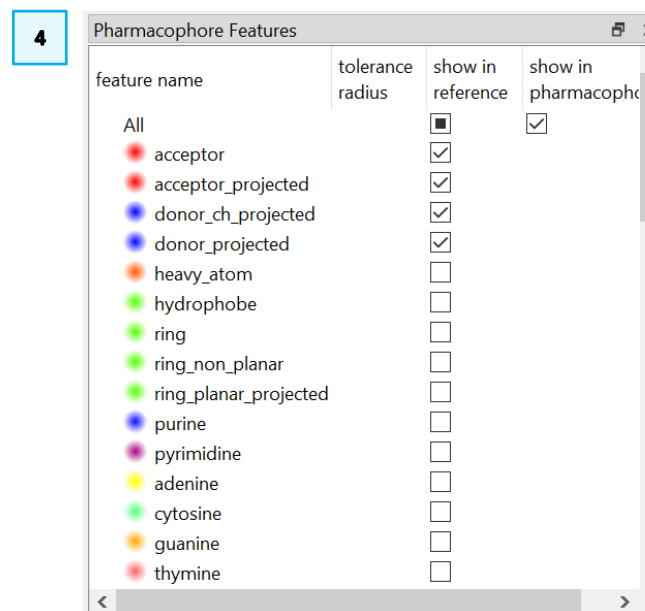
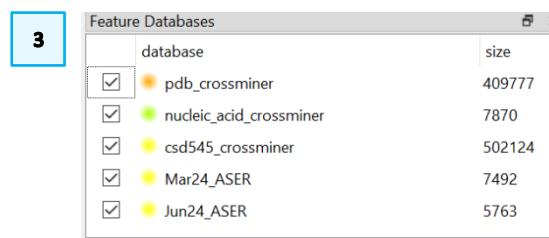
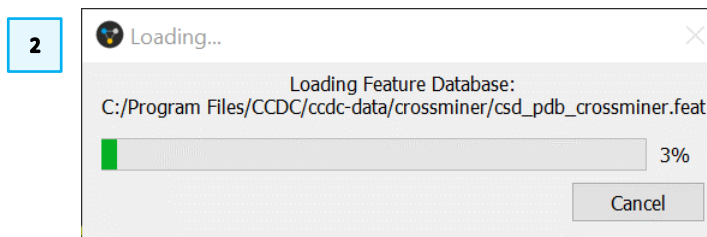
## Overview of CSD-CrossMiner

CSD-CrossMiner is a powerful tool with a simple user interface. This quick section will familiarise you with the basic functions and underlying data components before moving on to exploring some scientific questions.

1. Launch CSD-CrossMiner clicking on the CSD-CrossMiner icon: .
2. A progress window will indicate that the Feature Database is loading. If you do not see this window, it may be necessary to load the feature database from `/path/to/CCDC/ccdc-data/crossminer` or install the CrossMiner Data using the CCDC Maintenance tool.

Loading may take a few minutes.

3. Once loaded, you will see the CSD and PDB (*pdb\_crossminer* and *nucleic\_acid\_crossminer*) databases listed in the *Feature Databases* window. You can load multiple databases and use the tick boxes to indicate which database should be searched.
4. You will also see a list of features in the bottom right *Pharmacophore Features* window. These are the features used to generate these databases. The features with the *show in reference* tick-box toggled are displayed in the 3D view.



## Searching with a Pharmacophore

In this tutorial we will learn how to run a pharmacophore search in CSD-CrossMiner using one of the pharmacophore examples provided in the CSD-CrossMiner installation folder.

Human cathepsin L plays a major role in protein catabolism and is implicated in several pathological processes. As such, it is a common research target and serves as a good example for novel drug discovery. For this example, we will be using a pharmacophore included with CSD-CrossMiner to explore the CSD and PDB for possible hits.

1. Load the cathepsin L pharmacophore by clicking *File > Load Pharmacophore...* and select:

<CCDC installation folder>\ccdc-software\csd-crossminer\example\_pharmacophores\catl\_s3.cm

2. This will load the pharmacophore into the viewing area. Take a moment to rotate the pharmacophore and understand the different pharmacophore feature points:

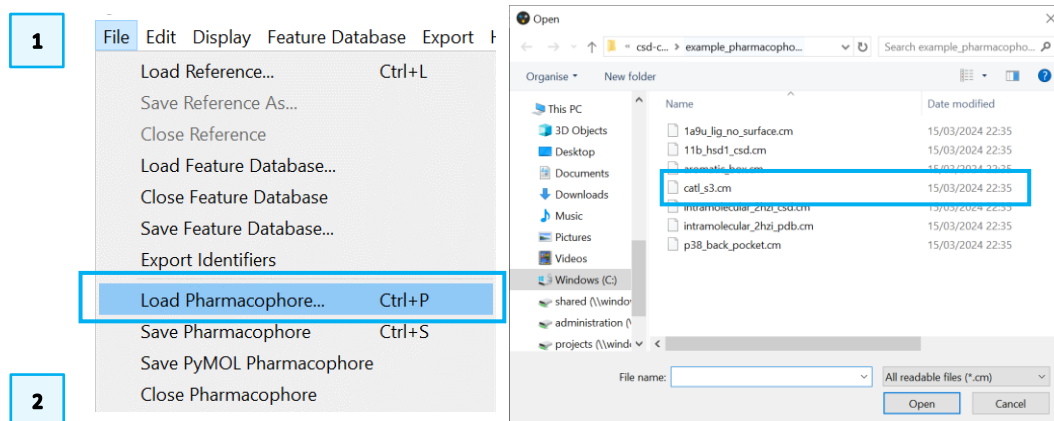
**P:** Protein feature

**S:** Small molecule feature

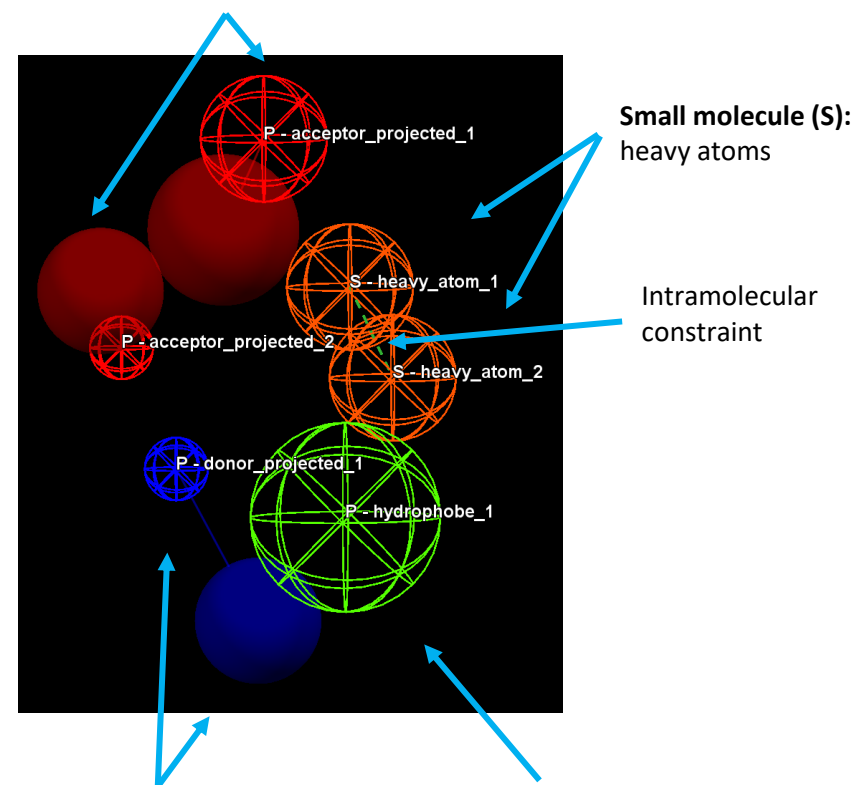
**Dashed line:** intra and intermolecular constraints. Constrained features must belong to either the same molecule as each other (*intra*, dashed green line) or different molecules (*inter*, dashed red line).

**Mesh sphere:** the actual feature itself, where the sphere size represents the radius of tolerance.

**Solid sphere:** the projected virtual point to represent the directionality of a [hydrogen bond acceptor/donor](#). A feature can have more than one projected point. For example, a H bond acceptor can have multiple potential lone pair preferred projections.




**Protein (P):** H bond acceptor feature (mesh) with projected directionality (solid)





**Protein (P):** H bond donor feature (mesh) with projected directionality (solid)

**Protein (P):** hydrophobe feature

Note that the colour coding is defined in the **Pharmacophore Features** browser. E.g., [hydrophobe](#) features are green, hydrogen bond acceptors are red, and so on. The pharmacophore in your viewer has only one projected H bond donor. These points correlate to the feature browser: B indicates the **B**ase feature, and V indicates the accompanying **V**irtual point.

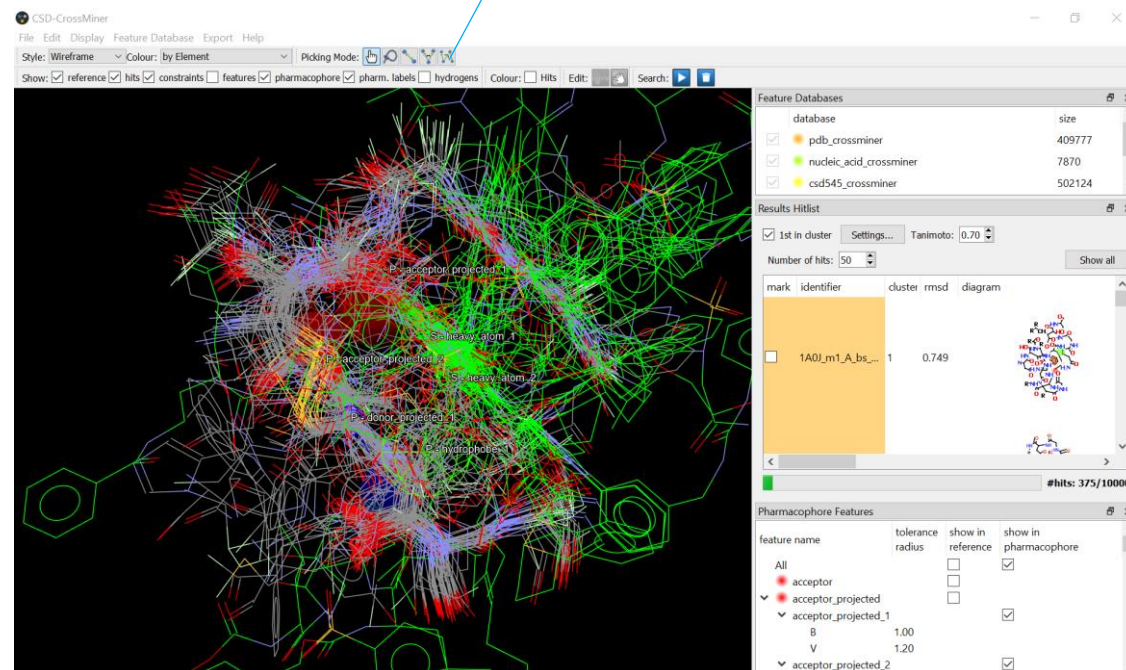
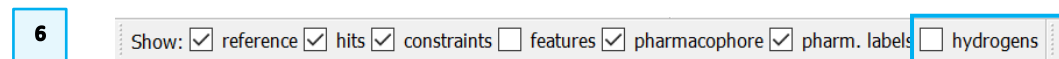
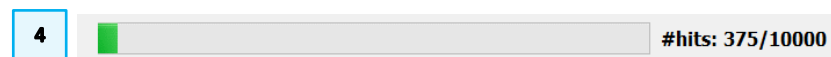
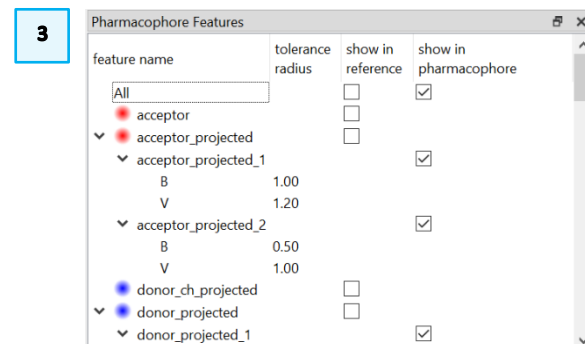
- Sphere size is directly correlated to tolerance – the smaller the sphere the lower the tolerance for geometric matching. The radius of each feature is expressed in Å is shown in the **Pharmacophore Features** window.
- Note that the pharmacophore features can be hidden by unticking the *All* tick-box in the *show in pharmacophore* column.
- Click the **Play** button  to begin searching across the databases for matches. As the search runs, you will see results populating the *Results Hitlist* window, as well as the 3D view and the progress bar with total number of hits listed at the bottom of the *Results Hitlist* window.

This particular search returns a high number of hits and may require several minutes.

- Pause the search by clicking the **Pause**  button when you have a few hundreds of hits. Do not stop the search (pressing ), as that stops the entire process and removes all hits. By default, ligands are displayed with green carbons and protein residues with grey carbons.

All results are overlaid in the 3D view, which provides an easy appreciation of the common motifs matching the pharmacophores. However, you can view one hit at the time by clicking on each result in the **Results Hitlist** window or, hold down the **Shift** or **Ctrl** keys to select multiple.

- Disable the **hydrogens** tick-box in the *Show:* toolbar, this will hide the hydrogen atoms of the matched hits in the 3D view.






8. Locate the result with the lowest [RMSD](#) by clicking on the *rmsd* column in the **Results Hitlist** window, to show ascending order. The lowest RMSD result in this example is 1ADY\_m1\_A\_bs\_HAM\_A\_423\_1 – note that yours may be different, depending on when you stopped your search. The results come back with a 2D diagram, with pharmacophore matches indicated. Select your smallest RMSD result by clicking on that row in the *Results Hitlist* window.
9. For ease of viewing, change the style in the *Style:* toolbar to **Capped Sticks**. Note this will only apply to what is currently displayed in the 3D view.

Take a few moments to explore how the returned result matches the pharmacophore query. In particular note that:


- Feature matching is based on size of sphere (and thus the tolerance level). Tiny spheres with tight tolerance will result in hits with very close alignment to the centre of the sphere, while larger spheres have a wider area for alignment.

The pharmacophore search performed has provided a high number of hits, likely due to the large radius of the features. Lowering the tolerance of the features will get a smaller result set, with more precise alignment to the pharmacophore centres.

To do this, you will need to edit the pharmacophore query, which cannot be done in pause mode.

10. Click the **Stop**  button to clear the results and enable editing.

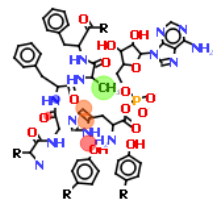
11. Double click on the radius sizes in the *Pharmacophore Features* window to change them. Change the radius of every feature to 0.5 to lower the tolerance. This will result in fewer hits, which are all more closely aligned with the centre of the pharmacophore features.

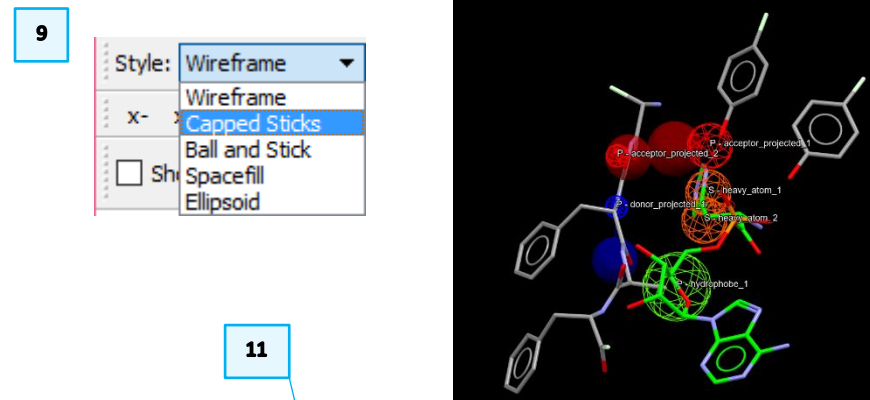
Then click on  to start the search on the new, tighter, pharmacophore.

**Results Hitlist**

☒ 1st in cluster   Settings...   Tanimoto: 0.70

Number of hits: 50   Show all

mark	identifier	cluster	rmsd	diagram
<input type="checkbox"/>	1ADY_m1_A_bs_...	16	0.48	



**Pharmacophore Features**

feature name	tolerance radius	show in reference	show in pharmacophore
▼ acceptor_projected_1			<input checked="" type="checkbox"/>
B	0.50		
V	0.50		
▼ acceptor_projected_2			<input checked="" type="checkbox"/>
B	0.50		
V	0.50		
● donor_ch_projected		<input type="checkbox"/>	
▼ donor_projected		<input type="checkbox"/>	
▼ donor_projected_1			<input checked="" type="checkbox"/>
B	0.50		
V	0.50		

It will take a bit more time to pick up hits, as there are far fewer. Let the search go to completion this time, resulting in 34 hits. Note: your search may return a different number of hits depending on your CrossMiner Data version.

12. Activate the **Colour: Hits** check-box in the CSD-CrossMiner toolbar to colour visible hits with rainbow colour. The colour applied to the hits will be also displayed in the **cluster** tab in the **Results Hitlist** browser.

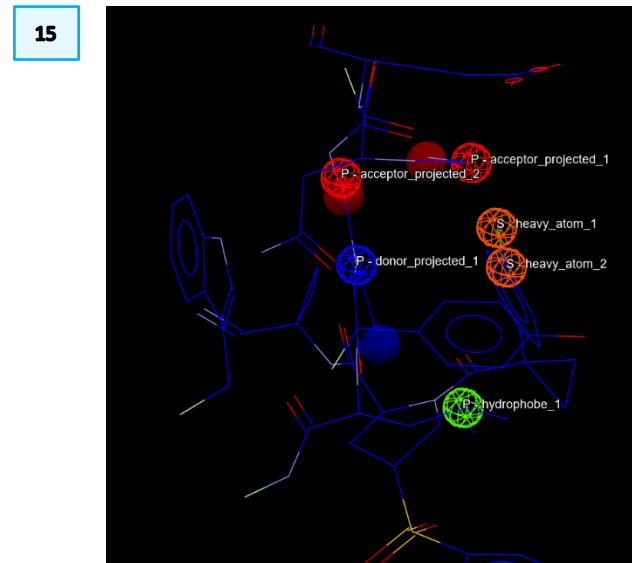
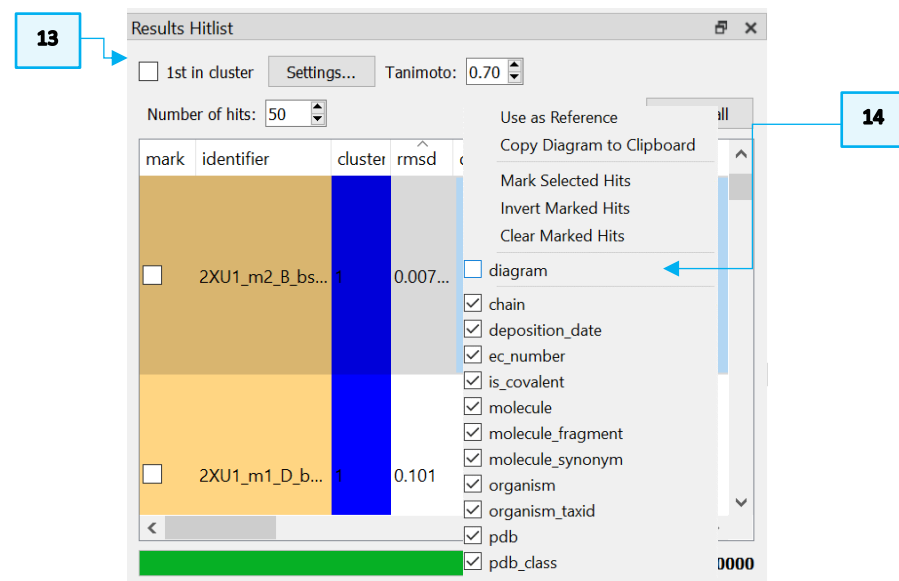
13. If **1st in cluster** is selected in the *Results Hitlist*, then the results are clustered based on the adjacent [Tanimoto](#) value. If clustered, you will see representatives of those similar groups in the results viewer (and in this case only seven hits are returned). Uncheck the **1st in cluster** tick-box to see all hits, including those which are very similar to each other. Hits belonging to the same cluster will have the same colour.

14. Hide the 2D diagram of the matched hits by right-clicking in the **Results Hitlist** window and then uncheck the **diagram** tick box.

15. Click and scroll the horizontal scrollbar in the *Results Hitlist* to have access to the annotations of the matched hits. Stop at the *molecule* annotation.

You will see that nearly all matched structures correspond to the cathepsin L PDB entries. This is not surprising because of the low radius of tolerance.

If you explore the top result, with the lowest rmsd (2XU1\_m1\_B\_bs\_424\_B\_1221\_2), you'll see that the hit matches the pharmacophore exactly, it was indeed the original structure used to derive this pharmacophore.





16. To save interesting hits for further work outside CSD-CrossMiner, you can mark them by ticking the respective tick-boxes in the *mark* column.

Note that marking them does not display them in the 3D view, and similarly, displaying them in the 3D view (multiple selection available via **Shift** or **Ctrl**) does not mark them.

17. The 3D coordinates of the marked hits can be then exported to the CCDC protein visualiser Hermes (using the *Export* top-level menu) or saved to disk by clicking on *File* in the CSD-CrossMiner top-level menu and then *Save Marked Hits*.

18. Save your hits as *cathepsin\_hits.mol2*. Alternatively, in the *File* menu other two different save options are available: *Save Visible Hits*, which will save the visible hits (in the 3D view), *Save All Hits* that will save the 3D coordinates of all matched hits.

19. Note that the hits in the *Results Hitlist* browser (visible, marked and all hits) can be additionally saved as a table by choosing *csv* format in the *Save Hits* window. Now click on *File > Close Pharmacophore* to clear the 3D view without saving the edited pharmacophore.

The figure consists of three screenshots from the CSD-CrossMiner software interface, illustrating the workflow for saving hits.

**16** The first screenshot shows the 'Results Hitlist' window. It displays a table of hits with columns: mark, identifier, cluster, rmsd, chain, and deposi. The 'mark' column has checkboxes. The hit '2YJ8\_m2\_A\_bs\_YJ8\_A\_1221\_2' is selected (checked). The 'Number of hits' is set to 50, and the 'Tanimoto' score is 0.70.

**17** The second screenshot shows the 'File' menu. The 'Save Marked Hits' option is highlighted. Other options include 'Load Reference...', 'Close Reference', 'Load Feature Database...', 'Close Feature Database', 'Save Feature Database...', 'Export Identifiers', 'Load Pharmacophore...', 'Save Pharmacophore', 'Save PyMOL Pharmacophore', 'Close Pharmacophore', 'Save Visible Hits', 'Save All Hits', 'Save as Image...', 'Export POVray file...', 'Create Structure Database', and 'Exit'.

**18** The third screenshot shows the 'Save Hits' dialog box. The 'File name' is 'cathepsin\_hits'. The 'Save as type' is set to 'SDF (\*.sdf)'. Other options include 'MOL2 (\*.mol2)', 'SDF (\*.sdf)', and 'CSV (\*.csv)'.

## Summary

In this workshop we have seen how to run a pharmacophore search in CSD-CrossMiner using a pre-prepared pharmacophore query. We have seen how the tolerances can be adjusted to tailor the similarity of result returned to the pharmacophore query.

For your reference, you can find the user manual at this [link](#).

## Next Steps

After this workshop, you can continue learning about CSD-CrossMiner with more exercises available in the self-guided workshops available in the [CSD-Discovery workshops area](#) on our website.

<https://www.ccdc.cam.ac.uk/Community/educationalresources/workshop-materials/csd-discovery-workshops/>

## Feedback

We hope this workshop improved your understanding of CSD-CrossMiner 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 (link also available from workshops webpage), 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 CROSS-002. Thank you!

## Glossary

### Hydrogen Bonds

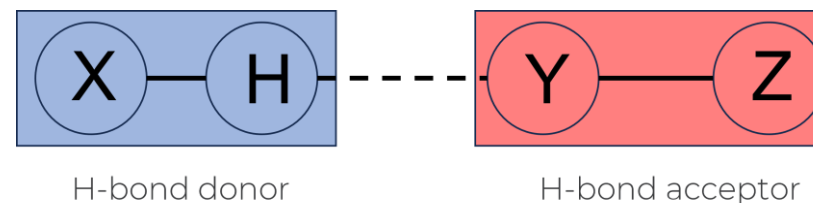
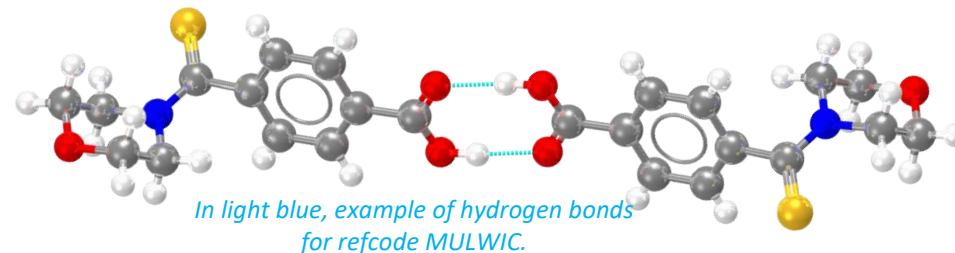
Hydrogen bonding occurs between donor-acceptor interactions precisely involving hydrogen atoms. The H-bonds interactions are classified as: strong (mostly covalent), moderate (mostly electrostatic) and weak (electrostatic). Their strength is observed to be between 12 and 30 kJ/mol.

### Hydrogen Bond Donor/Acceptor

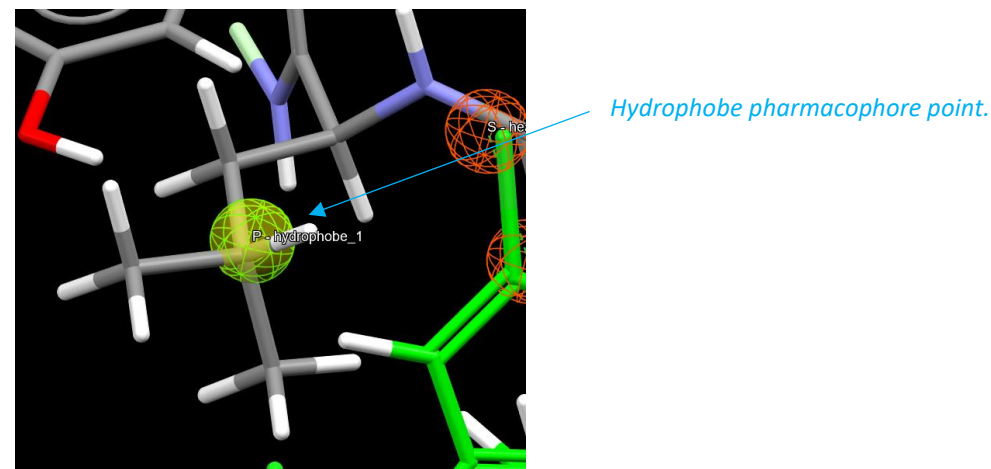
If a typical hydrogen bond is depicted as  $X-H\cdots Y-Z$ , where the dots denote the bond,  $X-H$  represents the hydrogen bond *donor*. The *acceptor* may be an atom or anion  $Y$ , or a fragment of a molecule,  $Y-Z$ , where  $Y$  is bonded to  $Z$ . The acceptor is an electron-rich region such as, but not limited to, a lone pair on  $Y$  or a  $\pi$ -bonded pair of  $Y-Z$ . [Source: E. Arunan, G. R. Desiraju, R. A. Klein, J. Sadlej, S. Scheiner, I. Alkorta, D. C. Clary, R. H. Crabtree, J. Dannenberg, P. Hobza, H. G. Kjaergaard, A. C. Legon, B. Mennucci and D. J. Nesbitt, *Pure Appl. Chem.*, 2011, **83**, 1637 – 1641.]

### Hydrophobic/hydrophobe

Hydrophobic molecules effectively “repel” water and thus tend to self-aggregate in aqueous media, excluding water in so doing. On a structural level, these are non-polar groups such as alkyl or aryl moieties. If these functional groups or molecular fragments are also pharmacophore features, then they are called *hydrophobes* in CSD-CrossMiner.



*Illustration of a hydrogen bond interaction with between hydrogen bond donor  $X-H$  and hydrogen bond acceptor  $Y-Z$ .*



*An isobutyl group is hydrophobic. The green mesh sphere indicates the position at which such a feature (functionally a hydrophobe) must be found.*

**Root Mean Square Deviation (RMSD)**

The root mean square deviation (RMSD) is a commonly used measure of the difference between two sets of values (usually comparing observed data to estimated data). The RMSD is defined as the square root of the mean squared error.

**SMARTS string**

A way of describing a chemical substructure using letters, numbers and symbols. If you are unfamiliar with SMARTS strings, you can visualise them and learn more about the format with SMARTSviewer (<http://smartsview.zbh.uni-hamburg.de/>).

**Tanimoto coefficient**

Tanimoto coefficient is the ratio of the number of features common to both molecules to the total number of features, i.e.

$$T(A, B) = (A \cap B) / (A + B - (A \cap B))$$

where  $A$  and  $B$  are the number of attributes of object  $a$  and  $b$ , respectively.