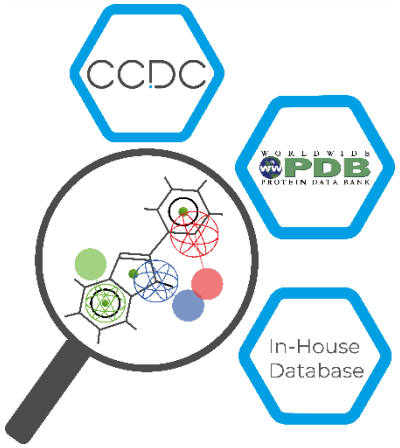
Performing a Pharmacophore Search using CSD-CrossMiner

2020.0 CSD Release



CCDC
advancing structural science

Table of Contents

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 subset of the CSD included with CSD-CrossMiner 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 (about 400 000 structures total). The included PDB database is divided in two subsets, one composed by of all 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 6Å 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.

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. When you have completed this tutorial, you should be able to build queries, execute pharmacophore search and save hits.

CSD-CrossMiner Terminology

CSD-CrossMiner uses several terms, some common to the field of drug discovery, and some not. For reference, these terms are defined as below:

Features: can be defined as 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: is 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: is 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.

Exit vector: is 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.

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: . The *csd_pdb_crossminer.feat* feature database, provided in the CSD package (`$CSD_2020/crossminer_data`), is loaded as default feature database.

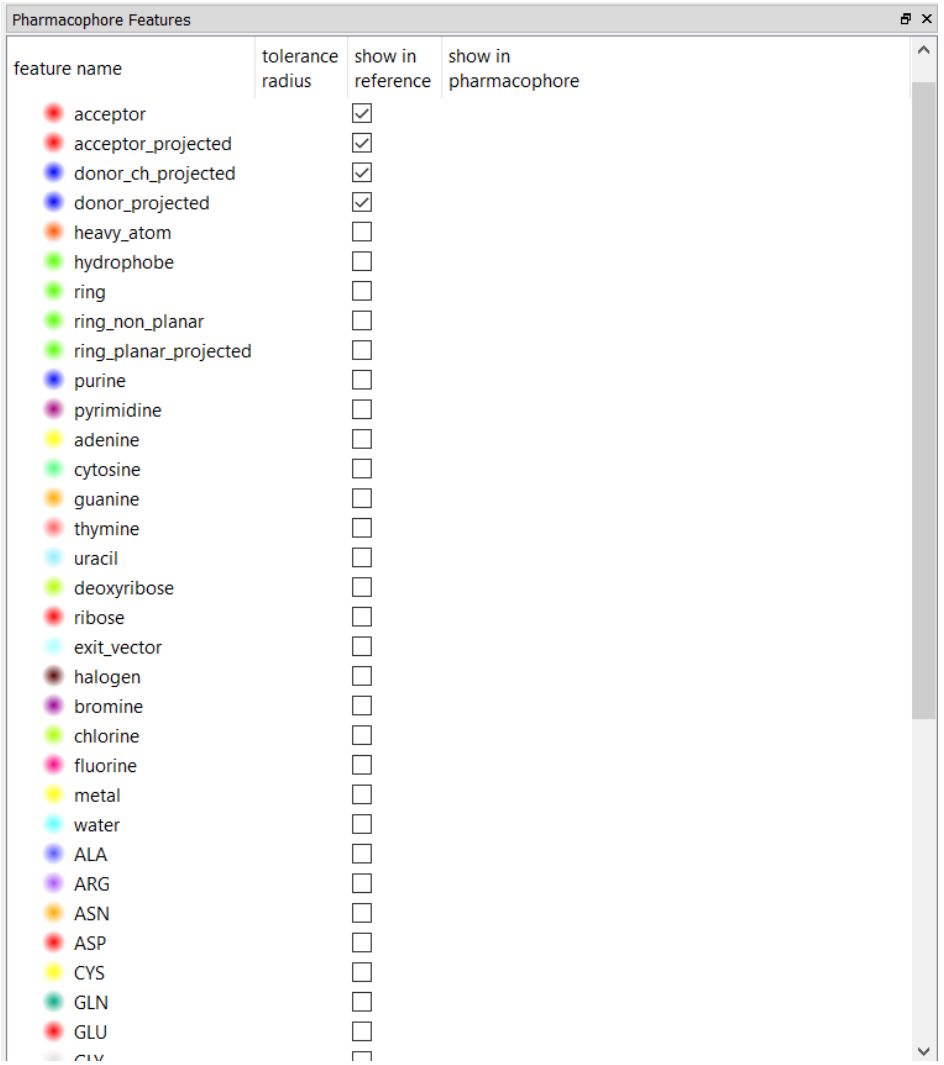
Loading will take a few minutes, but even once the bar hits 100%, it will need a moment to initialise the structures.

2. 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.

3. 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.

2

3



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:

```
[CSD-CrossMiner
installation]\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:

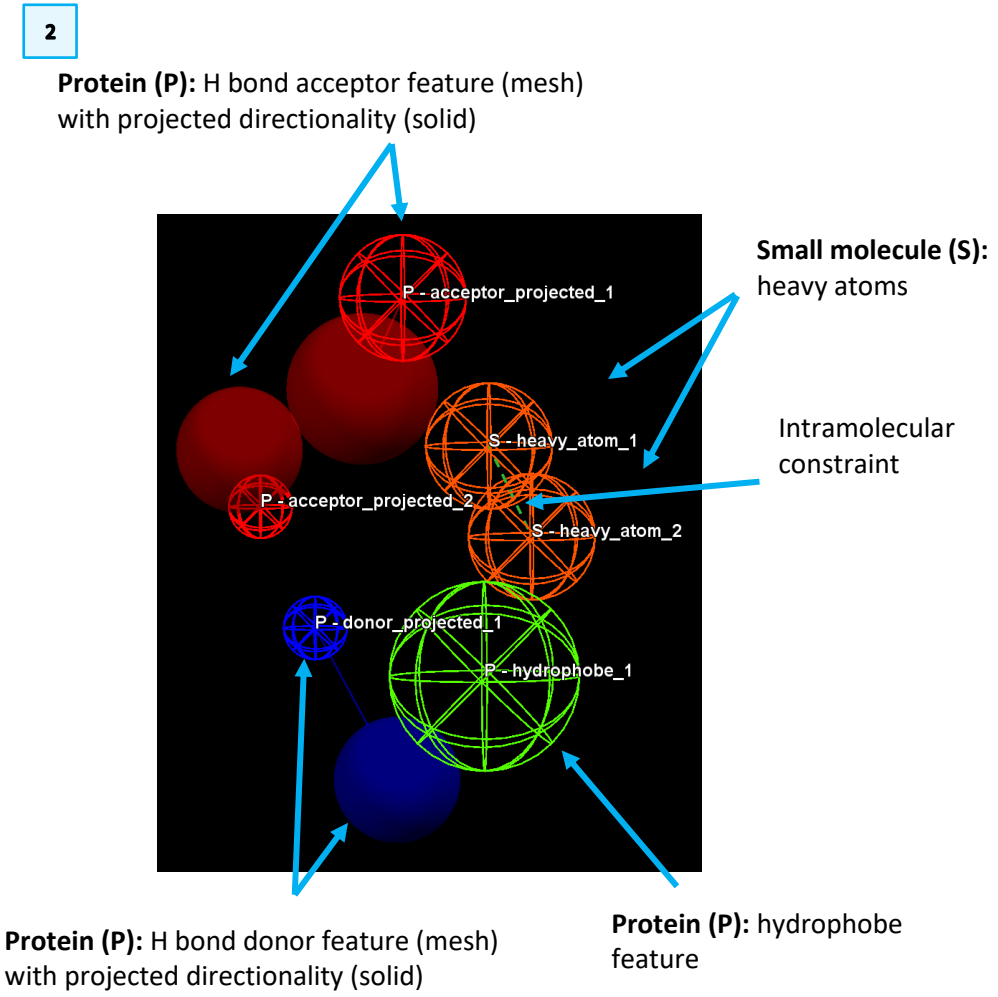
P: Protein feature

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.



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.

3. 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.
4. Note that the pharmacophore features can be hidden by unticking the *All* tick-box in the *show in pharmacophore* column.
5. 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.

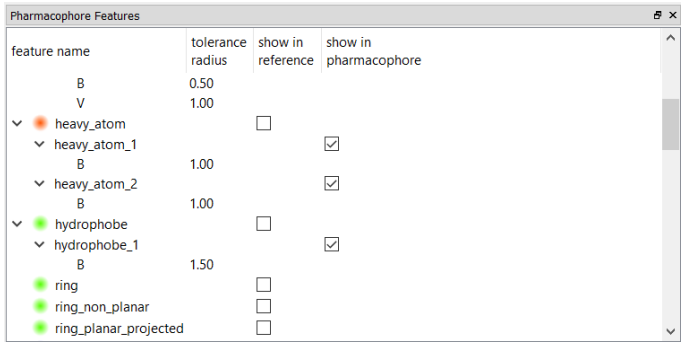
This particular search returns a high number of hits and may require several minutes.

6. 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, 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.

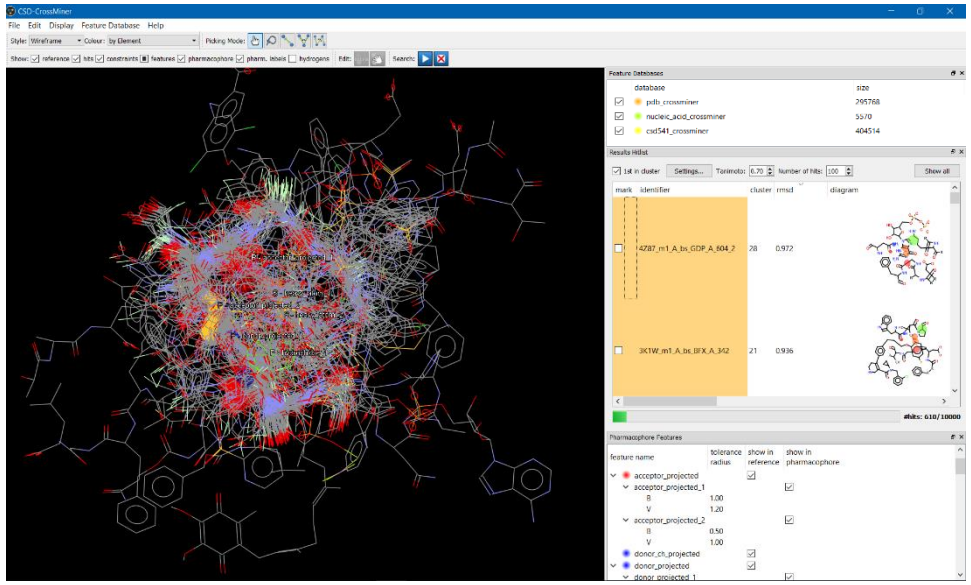
7. Disable the **hydrogens** tick-box in the *Show:* toolbar, this will hide the hydrogen atoms of the matched hits in the 3D view.

3

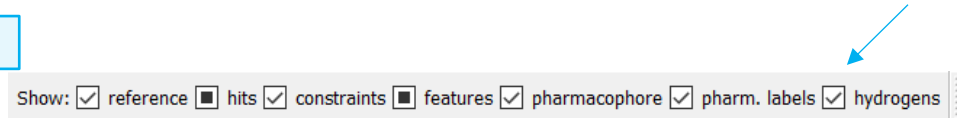


4

6



7



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 `3JU9_m1_A_bs_DBB_A_238_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**.

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, which cannot be done in pause mode.

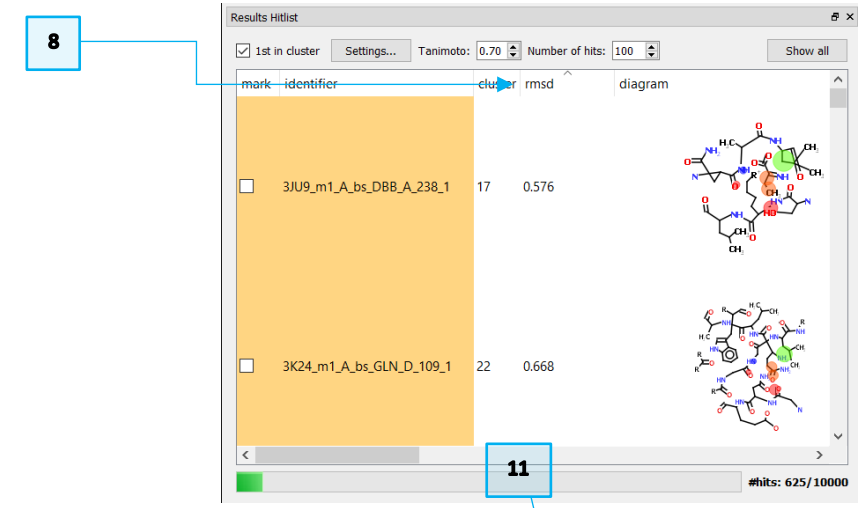
10. Click the **Stop**  button to clear the results and enable editing.

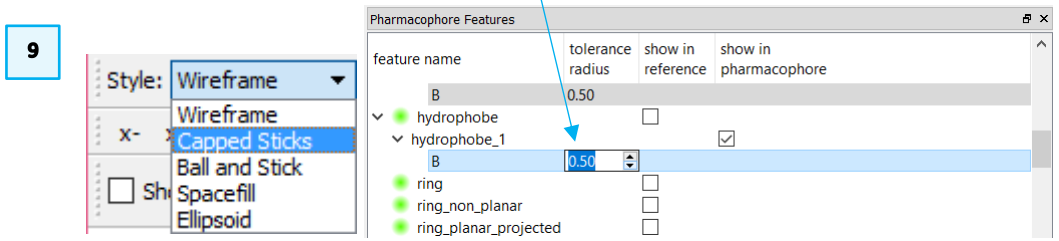
11. Double click on the radius sizes in the *Pharmacophore Features* window to change them. Change the radius of every feature to *0.5* to lower the tolerance. This will result in fewer hits, which are all more closely aligned with the centre of the pharmacophore features.

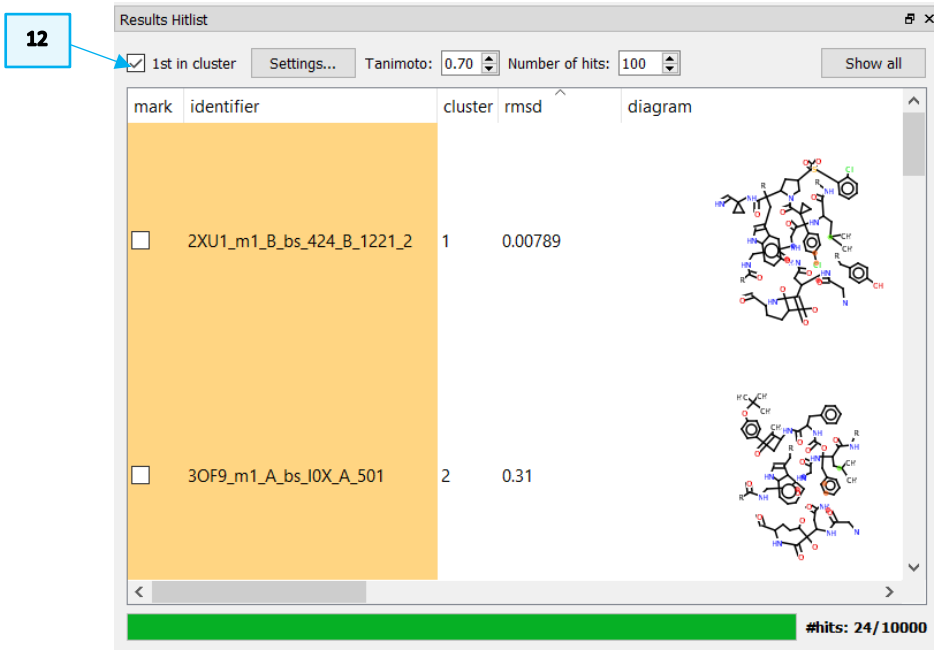
    Then click on  to start the search on the new, tighter, pharmacophore.

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 24 hits.

12. 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 one hit is returned). Uncheck the **1st in cluster** tick-box to see all hits, including those which are very similar to each other.

13. Hide the 2D diagram of the matched hits by right-clicking in the *Results Hitlist* window and then click on **– diagram**.

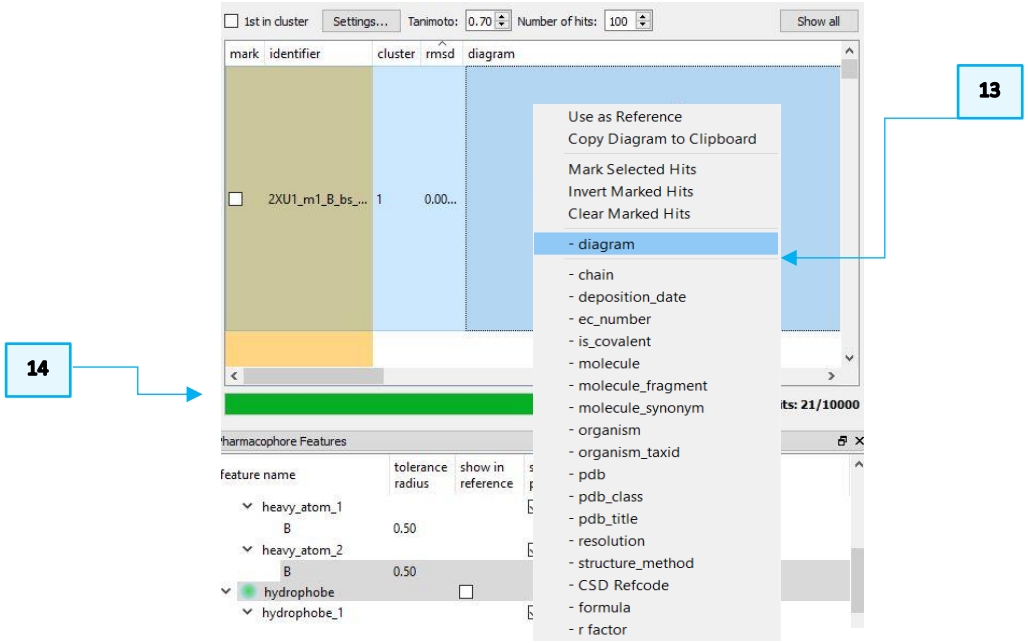
14. Click and scroll the horizontal scrollbar in the *Results Hitlist* to have access to the annotations of the matched hits. Stop to the *molecule* annotation.

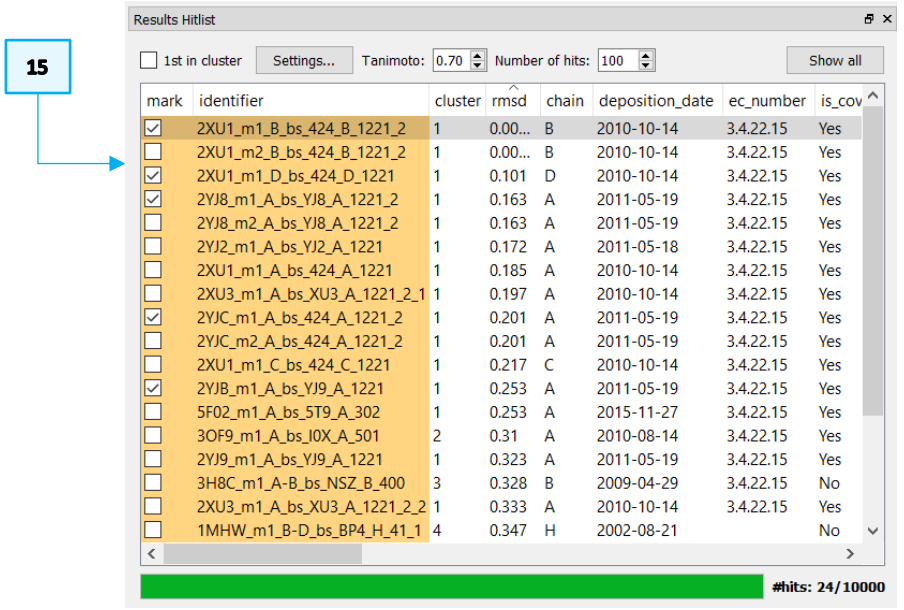
You will see that nearly all matched structures correspond to the cathepsin L PDB entries. This is not surprising because of the low radius of tolerance.

If you explore the top result (`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.

15. To save interesting hits for further work outside CSD-CrossMiner (the three lowest RMSD hits in this case), 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.





16. The 3D coordinates of the marked hits can be then saved to disk by clicking on **File** in the CSD-CrossMiner top-level menu and then **Save Marked Hits**. 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.

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.

17. Now click on **File** > **Close Pharmacophore** to clear the 3D view without saving the edited pharmacophore

This ends the tutorial.