Using CSD-CrossMiner to Create a Feature Database

2021.2 CSD Release

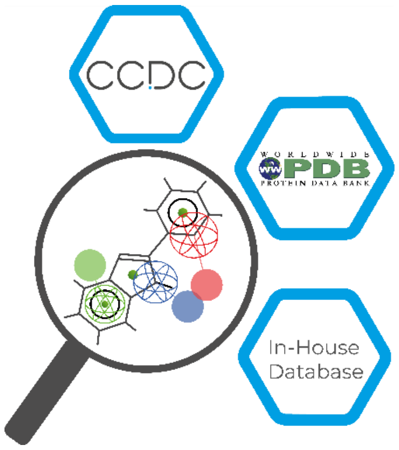


Table of Contents

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.

We provide a feature database that can be downloaded from the CSD-CrossMiner application or by accessing to the 'Data & Software Updates' section of our Downloads page.

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 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 CSD-CrossMiner users who are already familiar with the software and its terminology. It covers the basics on how to use the feature definitions provided in CSD-CrossMiner to create a feature database. When you have completed this tutorial, you should be able to create your own customized feature database and execute a pharmacophore search using a user-created feature database.

The files to perform this tutorial are provided in the `tutorial4` folder here.

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.

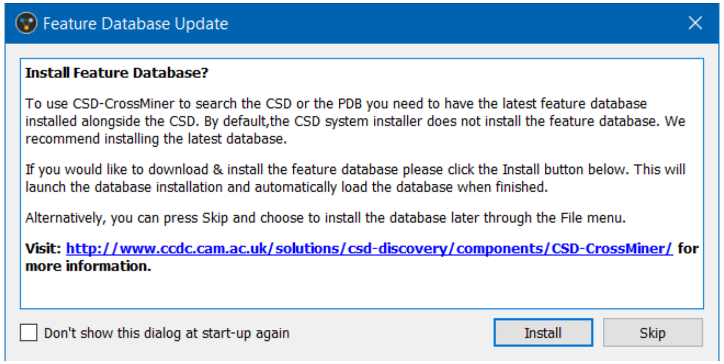
Overview of CSD-CrossMiner

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: .

2. If it is the first time you opened the application, *Feature Database Update* pop-up window will guide you through the process of downloading the feature database. Clicking on **Install** will start the downloading process and when completed the database will be automatically loaded in CSD-CrossMiner.

2



Feature Database Update

Install Feature Database?

To use CSD-CrossMiner to search the CSD or the PDB you need to have the latest feature database installed alongside the CSD. By default,the CSD system installer does not install the feature database. We recommend installing the latest database.

If you would like to download & install the feature database please click the Install button below. This will launch the database installation and automatically load the database when finished.

Alternatively, you can press Skip and choose to install the database later through the File menu.

Visit: http://www.ccdc.cam.ac.uk/solutions/csd-discovery/components/CSD-CrossMiner/ for more information.

☐ Don't show this dialog at start-up again

Note that the speed of the download depends on the quality of your network.

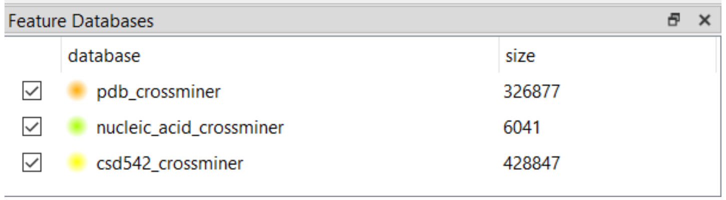
3. The *csd_pdb_crossminer.feat* feature database, will be saved in the crossminer_data folder in CSD_2021 directory. The location of the loaded feature database will be automatically remembered between separate CSD-CrossMiner sessions.

   If you have already downloaded the supplied feature database, it will be automatically loaded when you launch CSD-CrossMiner.

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

4. 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

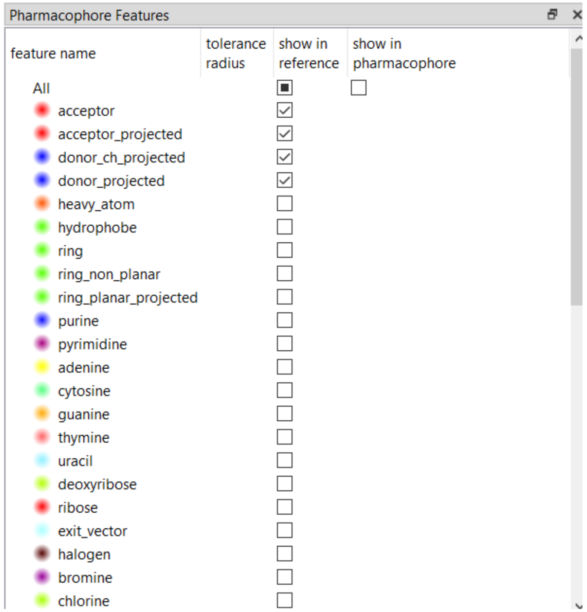


Feature Databases

	database	size
☑	● pdb_crossminer	326877
☑	● nucleic_acid_crossminer	6041
☑	● csd542_crossminer	428847

5. 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.

5



Pharmacophore Features

feature name	tolerance radius	show in reference	show in pharmacophore
All		■	☐
● acceptor		☑	
● acceptor_projected		☑	
● donor_ch_projected		☑	
● donor_projected		☑	
● heavy_atom		☐	
● hydrophobe		☐	
● ring		☐	
● ring_non_planar		☐	
● ring_planar_projected		☐	
● purine		☐	
● pyrimidine		☐	
● adenine		☐	
● cytosine		☐	
● guanine		☐	
● thymine		☐	
● uracil		☐	
● deoxyribose		☐	
● ribose		☐	
● exit_vector		☐	
● halogen		☐	
● bromine		☐	
● chlorine		☐	

Features and Pharmacophore Representation

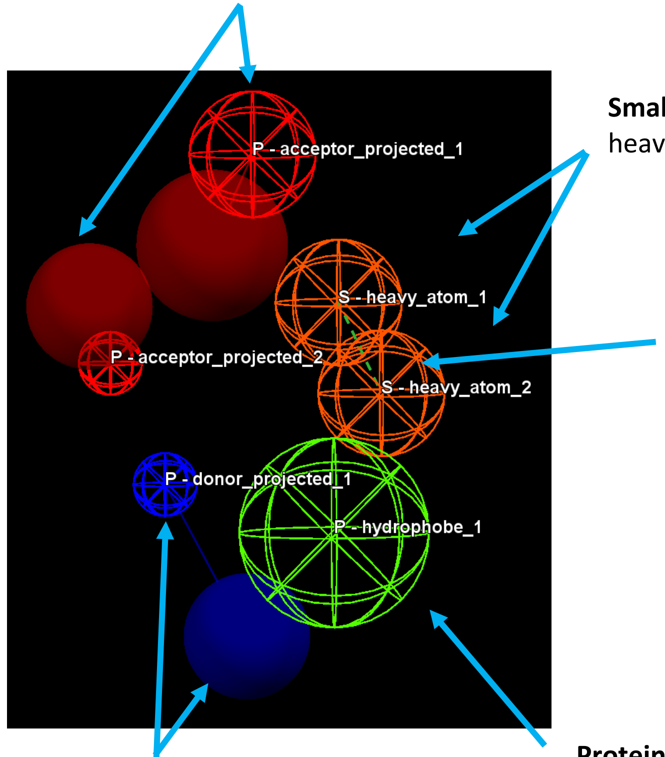
In the CSD-CrossMiner 3D view, features are represented as small translucent spheres coloured as defined in the *Pharmacophore Features* window. A pharmacophore point is represented as a mesh sphere which reflects the uncertainty in the position of the pharmacophore point. In the 3D view:

- **P**: Protein pharmacophore point
- **S**: Small molecule pharmacophore point
- **A:** Either protein or small molecule pharmacophore point
- **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 of the pharmacophore points is defined in the *Pharmacophore Features* browser; e.g. hydrophobe features are green, hydrogen bond acceptors are red, and so on.

In the directional pharmacophore, the mesh sphere (the actual feature itself) is defined as *B* in the *Pharmacophore Features* window (**B**ase feature), and the projected virtual point representing the directionality of the feature is defined as *V* (**V**irtual point).

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

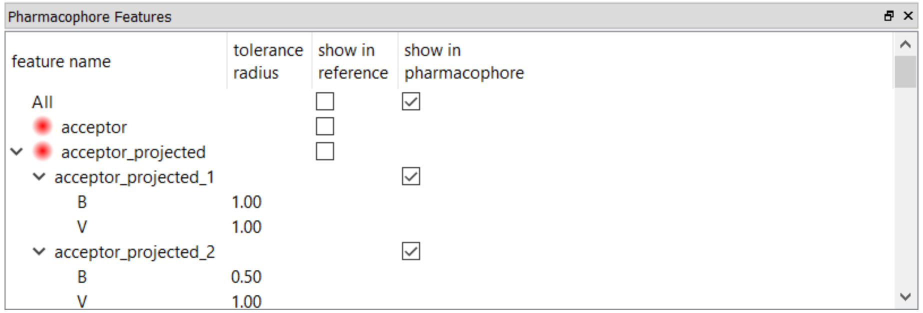


Small molecule (S): heavy atoms

Intramolecular constraint

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

Protein (P): hydrophobe feature



feature name	tolerance radius	show in reference	show in pharmacophore
All		☐	☑
● acceptor		☐	
● acceptor_projected		☐	
acceptor_projected_1			☑
B	1.00		
V	1.00		
acceptor_projected_2			☑
B	0.50		
V	1.00		

Building your Own Feature Database

For this exercise you will be building your own feature database from a set of 3D coordinates of several ligands. When creating a feature database, feature definitions are applied to the structure database. All structure databases must be converted simultaneously into a single feature database in order to ensure homogeneity in the feature definitions.

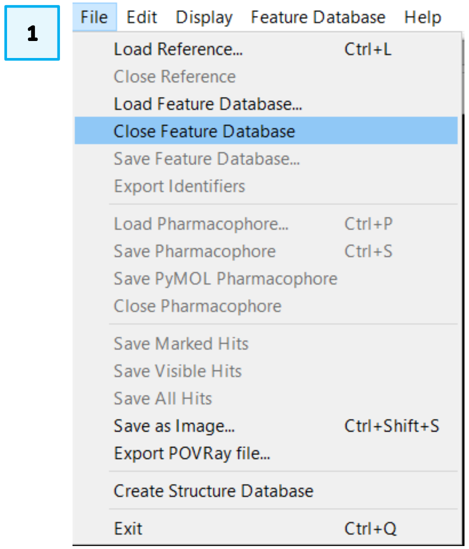
1. If you have a CSD-CrossMiner session open with a feature database already loaded, select **File** > **Close Feature Database** to close the feature database. Then click **OK** in the *Close Feature Database* pop-up window. CSD-CrossMiner may become unresponsive for a short time whilst the feature database is being closed.

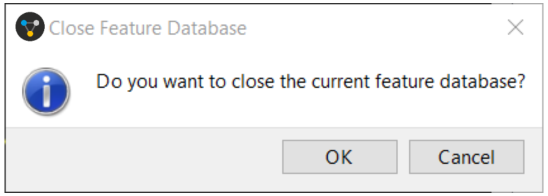
   Alternatively, if you are starting a new CSD-CrossMiner session, you can stop the default loading of the feature database by clicking on **Cancel** in the *Loading…*pop-up window, and then **OK**. This will open CSD-CrossMiner without any feature database loaded.

2. Select **Feature Database** > **Create** from the CSD-CrossMiner top-level menu**.**

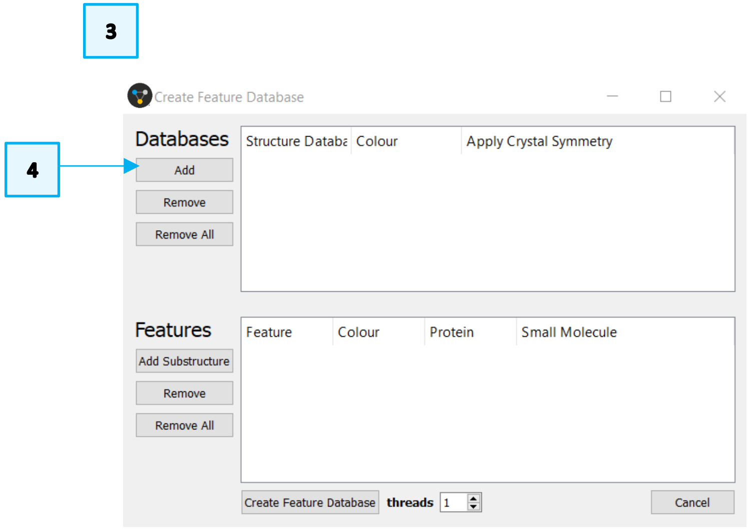
3. In the *Create Feature Database* dialog, structure databases (*Databases*) and feature definitions (*Features*) can be specified to create a new feature database.

4. Click on **Add** in the *Databases* section of the *Create Feature Database* dialog and navigate to the tutorial4 folder you downloaded to load the *fviia_ligands01.sdf* structure database.





Do you want to close the current feature database?



5. Select the loaded structure database by clicking on it in the *Create Feature Database* dialog. Click on the sphere under the *Colour* column to associate a specific colour to the new feature database. Leave the *Apply Crystal Symmetry* unticked.

Now we must load the feature definitions that will be assigned to all the structures contained in the loaded structure database.

6. Click on the **Add Substructure** button in the *Features* section of the *Create Feature Database* dialog. From the *Select Feature Definitions* window go to the

   `<CCDC installation folder>\Discovery_2021\CSD-CrossMiner\feature_definitions`

   folder and select every entry in the `any` folder (you can use the **Shift** key and select the first and last file of the folder or **Ctrl-A** to select all the content of the folder) and then click **Open.** This will load the features applied to protein, nucleic acids and small molecule structures in the feature database provided with CSD-CrossMiner.

   The loaded features will be listed in the *Features* section of the *Create Feature Database* window, along with their name and colour.

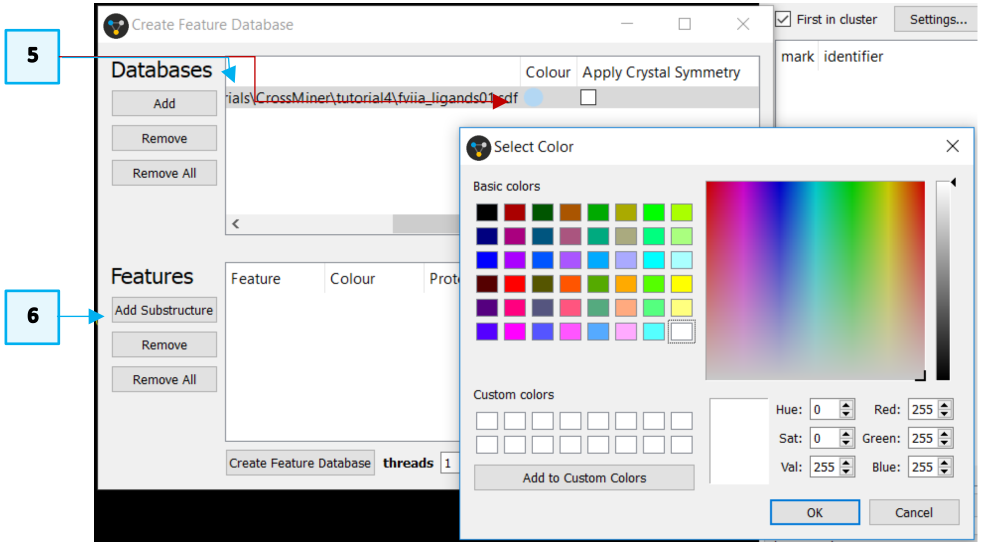
7. Repeat the procedure described in the point 6. for the features contained in the `small_molecule` folder.

Because the structure database provided in the `tutorial4` folder contains only small molecules, we can ignore the features associated exclusively to proteins contained in the protein folder.

8. Click **Create Feature Database**, to create the database. It will ask you for a save location, please indicate the tutorial data directory and name the database *fviaa_ligands01.feat*.

Note that, depending on the number of structures in the structure database(s) and the number of feature definitions loaded, the creation of a feature database can be computationally expensive. Therefore, the feature database creation can be distributed across multiple CPU cores by specifying the desired number of cores in the **threads** spin-box.

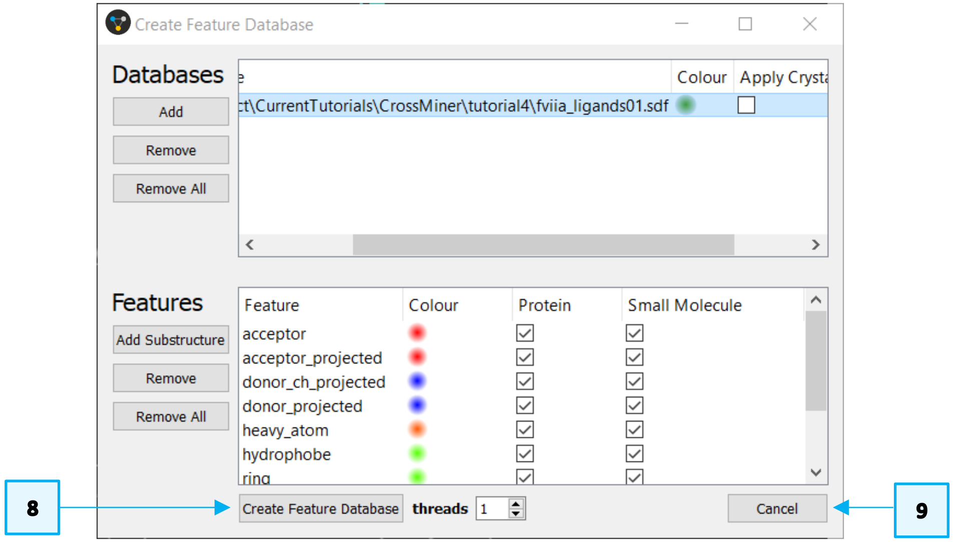
9. Once the indexing has completed, click **Cancel** to close the *Create Feature*



ials\CrossMiner\tutorial4\fviia_ligands01.sdf ● □

threads 1

Hue: 0 Red: 255
Sat: 0 Green: 255
Val: 255 Blue: 255



ct\CurrentTutorials\CrossMiner\tutorial4\fviia_ligands01.sdf ● □

Feature	Colour	Protein	Small Molecule
acceptor	●	☑	☑
acceptor_projected	●	☑	☑
donor_ch_projected	●	☑	☑
donor_projected	●	☑	☑
heavy_atom	●	☑	☑
hydrophobe	●	☑	☑
ring	●	☑	☑

threads 1

Database window.

10. Try out the new database by clicking on **File** > **Load Feature Database** from CSD-CrossMiner top-level menu to load it. Now in the *Feature Databases* window, only your database is listed.

Note that the new feature database will be then loaded by default the next time you start a CSD-CrossMiner session.

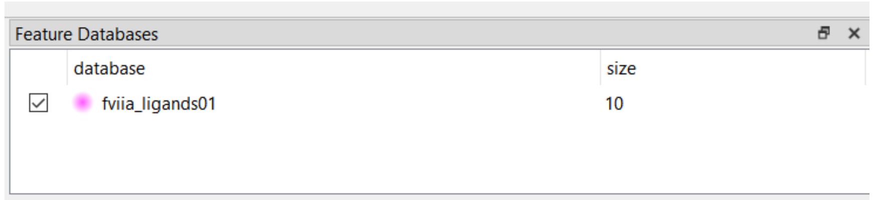
11. The simplest test is to see if you can pull back a molecule from the original data set. Load the reference molecule *4YT6_4JY.sdf* from the tutorial4 folder clicking on **File** from CSD-CrossMiner top-level menu and then **Load Reference** from the pull-down menu.

12. The displayed features of the reference structure are ticked in the *show in reference* column in the *Pharmacophore Features* window.

    Tick the *ring_planar_projected* feature tick-box to show all the molecule's planar ring in the 3D view.

13. Define two *ring_planar_projected* pharmacophore points as indicated in the image on the right by right-clicking on the *ring_planar_projected* features in the 3D view. Make sure they have *intra* constraints by clicking on intra and the click on ▶ to start the search.

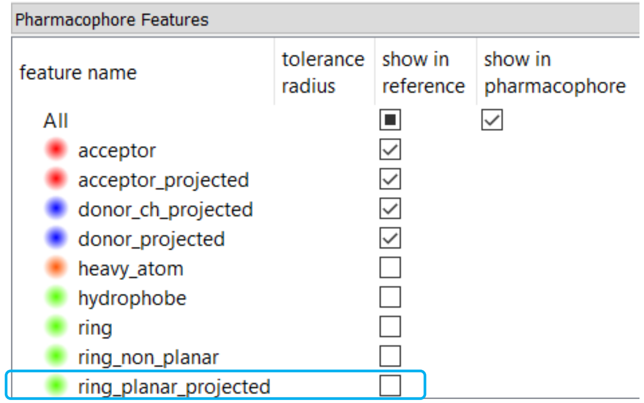
10



Feature Databases

	database	size
☑	● fviia_ligands01	10

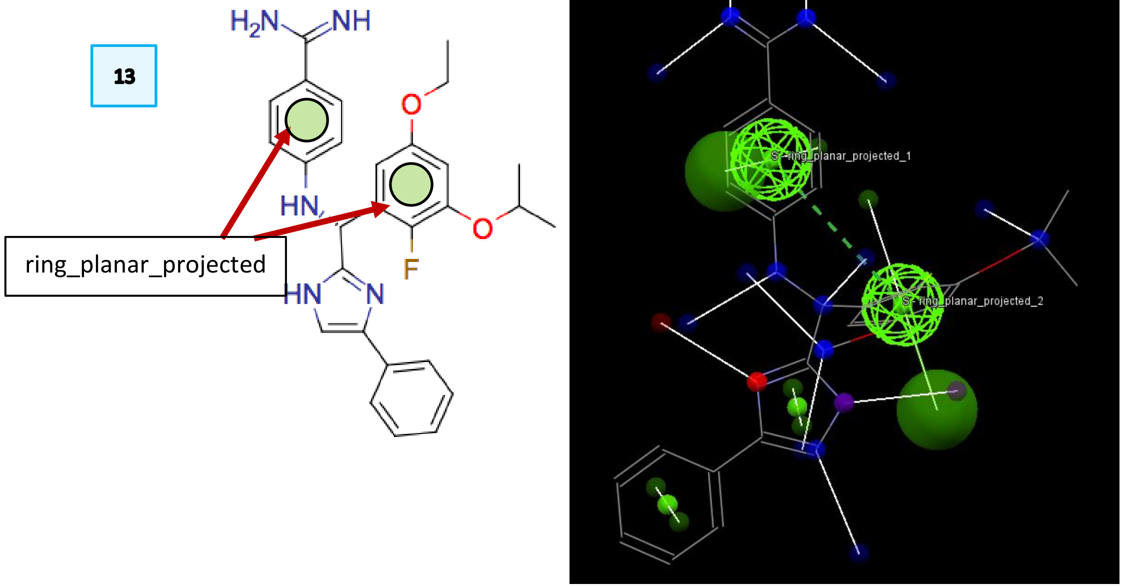
12



Pharmacophore Features

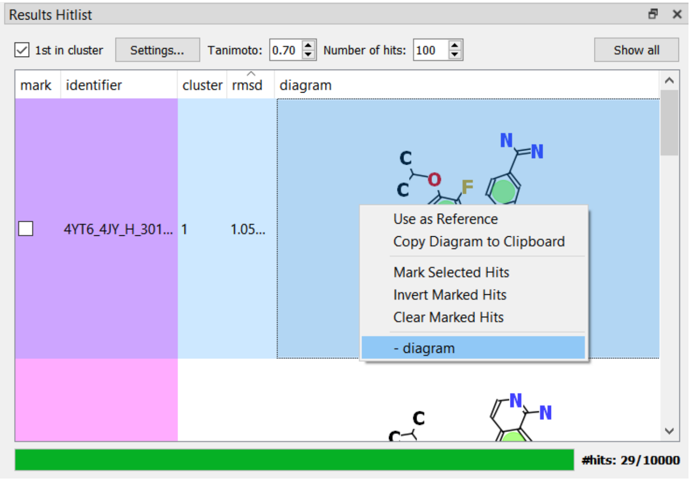
feature name	tolerance radius	show in reference	show in pharmacophore
All		■	☑
● acceptor		☑	
● acceptor_projected		☑	
● donor_ch_projected		☑	
● donor_projected		☑	
● heavy_atom		□	
● hydrophobe		□	
● ring		□	
● ring_non_planar		□	
● ring_planar_projected		□	

13



Note that the two options in the **Add ring_planar_projected** define a different placement of the projected ring indicating the potential position of the coupling feature.

- If you pick the **Add ring_planar_projected** options as shown in the 3D image, you should get 29 hits.
- You can hide the 2D diagram in the *Results Hitlist* window by right-clicking on one of the hits and select **– diagram** from the options available.
- By doing so you will see that with the 29 structures are clustered in 5 results. To see all 29 results untick the **1st cluster** tick-box.


Results Hitlist

☑ 1st in cluster Settings... Tanimoto: 0.70 Number of hits: 100 Show all

mark	identifier	cluster	rmsd	diagram
□	4YT6_4JY_H_301...	1	1.05...	

#hits: 29/10000

<u>If you play with different combinations of the **Add ring_planar_projected** options, you will have different number of matching hits.</u>

When you don’t have any information about the location of the coupling feature, any of these combinations is equally valid and all possible combinations should be explored.

This ends the tutorial.

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-005. Thank you!