CrossMiner User Guide

CSD-CrossMiner User Guide

A Component of CSD-Discovery Suite

2022.3 CSD Release

To access our new format tutorials please visit the CSD-CrossMiner web page

Conditions of Use

The Cambridge Structural Database Portfolio (CSD Portfolio) including, but not limited to, the following: ConQuest, CSD-Editor, Decifer, Mercury, Mogul, IsoStar, CSD Conformer Generator, Hermes, GOLD, SuperStar, the CSD Python API, web accessible CSD tools and services, WebCSD, CSD sketchers, CSD data files, CSD data updates, the CSD database, sub-files derived from the foregoing

data files, documentation and command procedures, test versions of any existing or new program, code, tool, data files, sub-files, documentation or command procedures which may be available from time to time (each individually a Component) encompasses database and copyright works belonging to the Cambridge Crystallographic Data Centre (CCDC) and its licensors and all rights are protected.

Any use of a Component of the CSD Portfolio, is permitted solely in accordance with a valid Licence of Access Agreement or Products Licence and Support Agreement and all Components included are proprietary. When a Component is supplied independently of the CSD Portfolio its use is subject to the conditions of the separate licence. All persons accessing the CSD Portfolio or its Components should make themselves aware of the conditions contained in the Licence of Access Agreement or Products Licence and Support Agreement or the relevant licence.

In particular:

- The CSD Portfolio and its Components are licensed subject to a time limit for use by a specified organisation at a specified location.
- The CSD Portfolio and its Components are to be treated as confidential and may NOT be disclosed or re-distributed in any form, in whole or in part, to any third party.
- Software or data derived from or developed using the CSD Portfolio may not be distributed without prior written approval of the CCDC. Such prior approval is also needed for joint projects between academic and for-profit organisations involving use of the CSD Portfolio.
- The CSD Portfolio and its Components may be used for scientific research, including the design of novel compounds. Results may be published in the scientific literature, but each such publication must include an appropriate citation as indicated in the Schedule to the Licence of Access Agreement or Products Licence and Support Agreement and on the CCDC website.

- No representations, warranties, or liabilities are expressed or implied in the supply of the CSD Portfolio or its Components by CCDC, its servants or agents, except where such exclusion or limitation is prohibited, void or unenforceable under governing law.

Licences may be obtained from:

CCDC Software Ltd.

12 Union Road

Cambridge CB2 1EZ

United Kingdom

Web: www.ccdc.cam.ac.uk

Telephone: +44-1223-336408

Email: admin@ccdc.cam.

Installation Notes

Minimum system requirements

CSD-CrossMiner can run on 64-bit Windows, Linux and macOS systems. For a list of supported platforms please refer to the Release and Installation Notes. If you choose to use a version other than those listed there, we cannot guarantee that CSD-CrossMiner will work correctly, although we will attempt to assist you with any problems you may encounter. If you do encounter any difficulties, please contact us at support@ccdc.cam.ac.uk to discuss possible solutions.

A complete installation of the 2022.3 CSD Release requires approximately 24 GB of disk space. This includes all software (7 GB) and all data files except for Crossminer (17 GB). The optional CrossMiner data install will occupy another 11 GB.

The minimum recommended RAM is 32GB. The software update mechanism can require more RAM to be available during the update activity. We advise that additional swap space or RAM is available at that moment to accommodate a further 4GB.

On Linux, CSD-Crossminer requires the Network Security Services (NSS) libraries to be present on your system. These can be installed on CentOS/RedHat 7/8 with the command:

```
sudo yum install nss
```

Introduction

CSD-CrossMiner is a novel tool that allows crystal structure databases such as the Cambridge Structural Database (CSD) and the Protein Data Bank (PDB) to be searched in terms of pharmacophore queries.

Intuitive pharmacophore queries describing, among others, protein–ligand interaction patterns, ligand scaffolds, or protein environments can be built and modified interactively. Matching crystal structures are overlaid onto the pharmacophore query and visualised as soon as they are available, enabling the user to quickly modify a hypothesis on the fly.

This delivers an overall interactive search experience with application in the areas of interaction searching, scaffold hopping or the identification of novel fragments for specific protein environments. For example use cases, please see:

Korb O, Kuhn B, Hert J, Taylor N, Cole J, Groom C & Stahl M “Interactive and Versatile Navigation of Structural Databases” J Med Chem, **2016**, 59(9):4257, **DOI:** 10.1021/acs.jmedchem.5b01756.

CSD-CrossMiner Terminology

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

Features

Features are 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 pharmacophore point is a feature that has been selected to be 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.

Excluded volume

An excluded volume is a a special feature that defines the occupational volume, where no solute molecule in a solution can be present. Excluded volume can be set to be a protein and/or small molecule.

Exit vector

An exit vector is a two-point feature that represents a single, non-ring bond between two heavy atoms. In CSD-CrossMiner an exit vector is bi-directional, therefore the directionality of the bond is not accounted for.

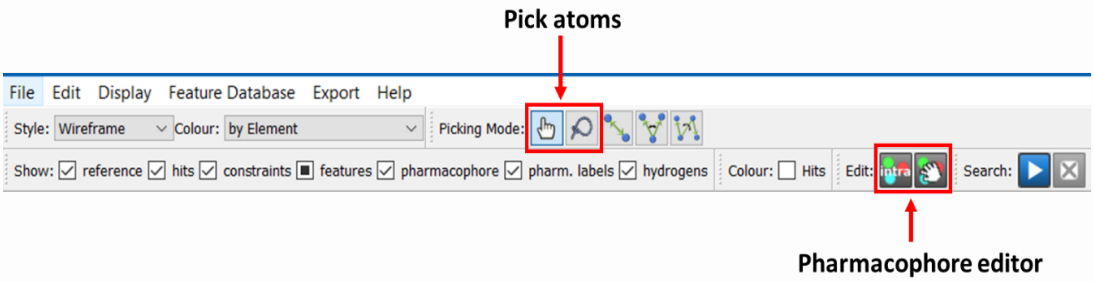
Structure database

A stucture database is a database containing the 3D coordinates of small molecule structures and/or protein-ligand binding sites (see Databases in CSD-CrossMiner).

Feature database

A 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 (see Databases in CSD-CrossMiner).

Overview of the User Interface



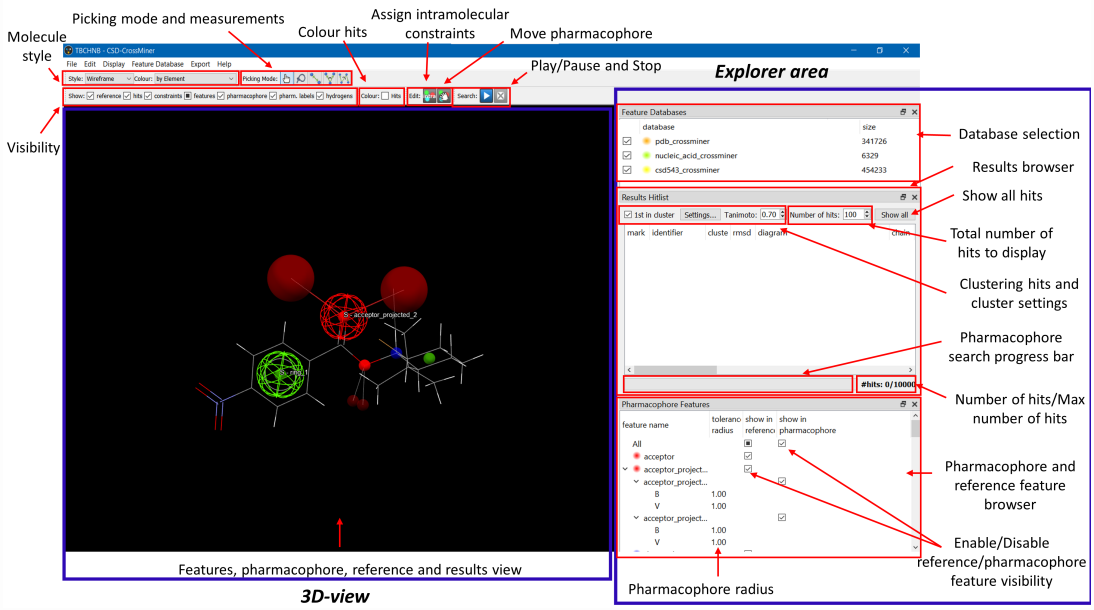
CSD-CrossMiner is a powerful tool with a simple user interface where the mouse function will depend on the selected picking mode: pick atoms and pharmacophore editor.

The CSD-CrossMiner user interface consists of the following:

- Top-level menu (see CSD-CrossMiner Top-Level Menu).
- **Style**, **Colour** and **Picking Mode** toolbar containing common, basic options, e.g., **Style** for setting global display styles; **Colour** for setting the colour mode of the 3D view and **Picking mode** for picking or lassoing atoms and for measuring distances, angles and torsions.

- **Show**, **Colour: Hits**, **Edit** and **Search** toolbar.
    - The **Show** checkboxes are used to select what is displayed in the 3D view.
    - **Colour: Hits** is used to colour matching hits in rainbow.
    - The **Edit** section gives access to some pharmacophore edit options.
    - The **Search** section is used to Start/Pause and Stop the pharmacophore search (see CSD-CrossMiner Toolbars).
- Display area (3D view) for showing 3D structures, features and pharmacophore points.
- Explorer area composed of the following windows:
    - **Feature Databases**, containing the name and the total number of structures contained in the loaded feature database (see Databases in CSD-CrossMiner).
    - **Results Hitlist**, containing information about the hits derived from the pharmacophore search (see Results Hitlist and Results Hitlist Browser).
    - **Pharmacophore Features**, containing all feature definitions assigned to the loaded feature database (see APPENDIX B. Feature Definitions in CSD-CrossMiner).

In CSD-CrossMiner, the different toolbars and windows are dockable; therefore, it is possible to hide and move any of these windows that may obscure the user view, or they can be kept as entirely separate windows. To do this, move the mouse cursor to the top of one of the windows (e.g., Feature Database) then drag the window with the mouse, keeping the left-button depressed, and put it where you want by releasing the mouse button.



All toolbars and windows can also be switched on or off by right-clicking in the toolbar area and then enabling or disabling the desired toolbar, or by clicking on **Display** in the top-level menu, selecting **Toolbars** and then enabling or disabling the desired toolbar or window.

Feature and Pharmacophore Representation in CSD-CrossMiner

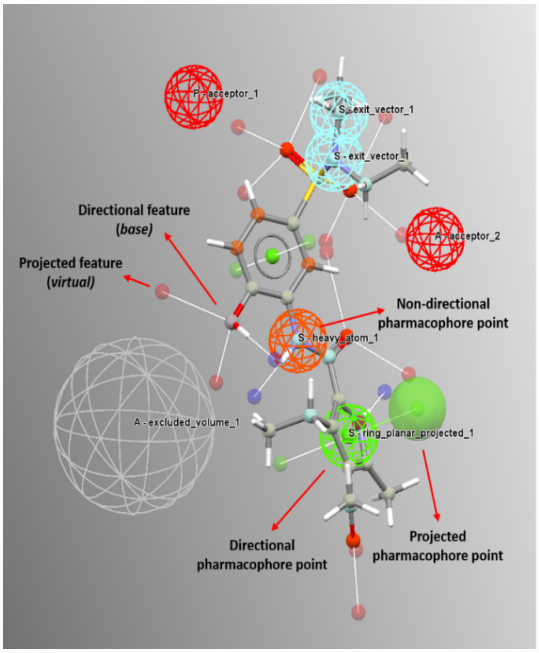
In the CSD-CrossMiner 3D view, a single-point feature (e.g., a heavy atom feature) is represented as a single small translucent sphere coloured as defined in the **Pharmacophore Feature** browser. A directional feature (e.g., a projected donor feature) is represented by two types of sphere, base and virtual, displayed as small translucent spheres. The base feature represents the feature itself while the virtual feature point(s) represent the directionality of the feature. Directional features can have more than one virtual sphere to represent the different directionality.

A pharmacophore point (e.g., a heavy atom pharmacophore point) is represented as a mesh sphere. The sphere radius of each pharmacophore point represents the tolerance radius and reflects the uncertainty in the position of the pharmacophore point. The radius of each pharmacophore point can be varied and thus be used to control the specificity of a pharmacophore query. A

directional pharmacophore point (e.g., a projected donor pharmacophore point) is represented by two types of sphere, base and virtual, displayed as a mesh sphere and a solid sphere, respectively. A directional pharmacophore point is represented by one base pharmacophore point and virtual pharmacophore point(s), where directionality is defined by the virtual pharmacophore point(s). A pharmacophore point can be set to belong to a protein (P), small molecule (S) or to any molecule (A).

An excluded volume pharmacophore point is displayed in the CSD-CrossMiner 3D view as a single mesh sphere.

An exit vector pharmacophore point is displayed as two mesh spheres.



P: Protein pharmacophore point
S: Small molecule pharmacophore point
A: Any molecule pharmacophore point

Translucent sphere: Feature point
Mesh sphere: Pharmacophore point itself and pharmacophore *base* in directional feature
Solid sphere: *Virtual* pharmacophore point

Databases in CSD-CrossMiner

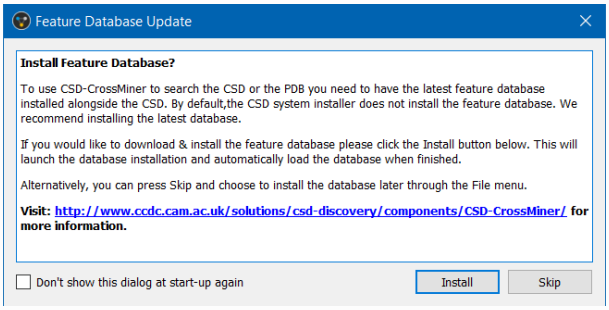
In CSD-CrossMiner there are two distinct types of databases: structure database and feature database, where the structure database contains the 3D coordinates of the molecules, and the feature database contains the structure database indexed with the feature definitions (see CSD-CrossMiner Terminology).

Downloading the CSD-CrossMiner Feature Database

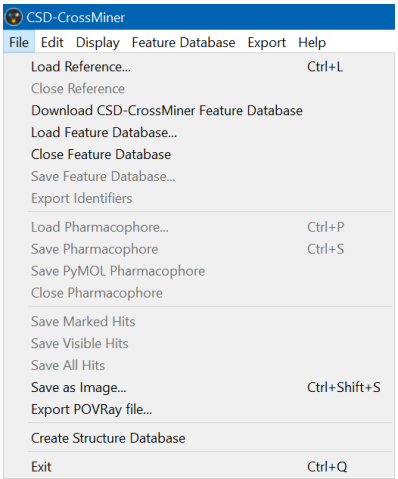
The first step required to use CSD-CrossMiner is to load a feature database and to initialise the associated structure database(s). Feature databases can be created for 3D molecular structures (see Creating Databases).

A feature database, containing the molecular structures of small molecules stored in the CSD and protein-ligand binding sites extracted from the PDB, can be downloaded from CSD-CrossMiner.

The first time CSD-CrossMiner is launched a pop-up window will walk you through the process of downloading the supplied feature database.



Clicking on **Install** starts the downloading and when completed the database will be loaded in the CSD-CrossMiner session. By selecting **Skip**, the feature database won't be downloaded at this stage however, you can download it later by selecting **Download CSD-CrossMiner feature database** option from the **File** menu in the CSD-CrossMiner top-level menu.

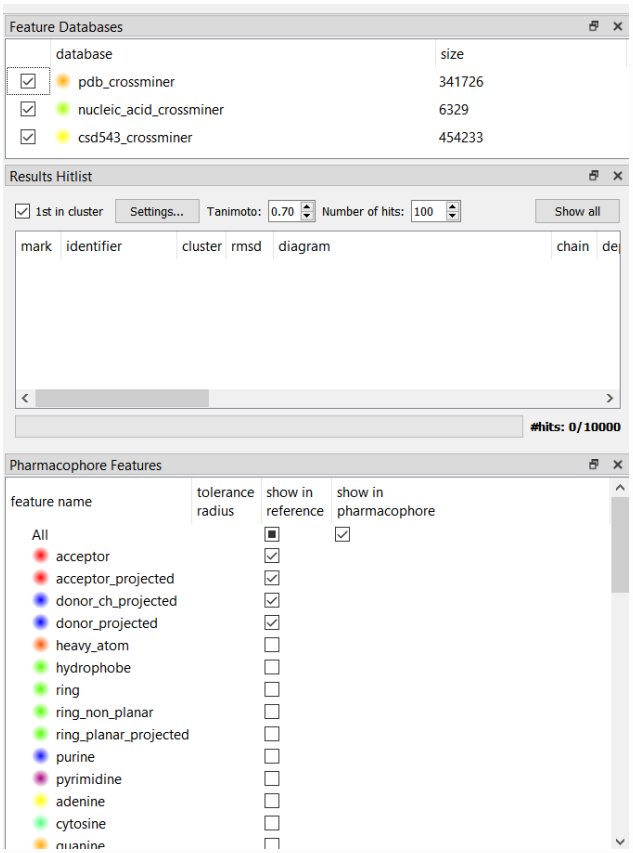


Note that the speed of the download depends on the quality of your network.

It is also possible to download and install the database update manually, by accessing the 'Data & Software Updates' section of our Downloads page.

The downloaded feature database is called csd_pdb_crossminer.feat and will be placed in `crossminer_data` folder in the `CSD_2022` directory together with the `pbd_crossminer.csdsqlx` and `nucleic_acid_crossminer.csdsqlx` structure databases.

After the loading is complete, the name of the structure database(s) and the number of included entries will be displayed in the **Feature Databases** window. The **Pharmacophore Features** window in the bottom right corner displays the feature definitions that were used to create the feature database.

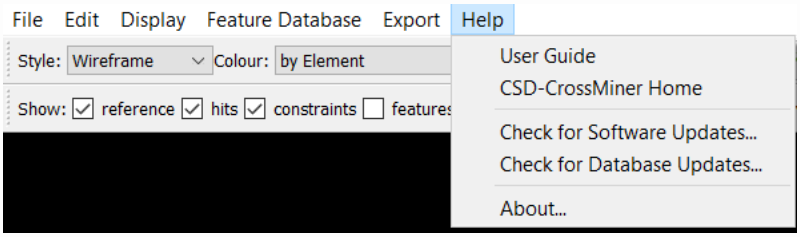


Because the feature database is updated regularly, the number of entries in the latest database could differ from the ones displayed in this user guide.

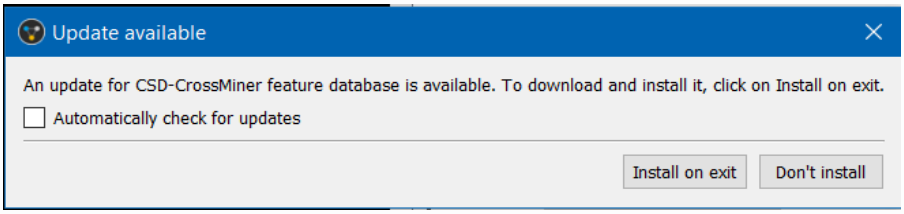
The location of the loaded feature database will be automatically remembered between separate CSD-CrossMiner sessions, where a **Load Feature Database** pop-up window will provide information about the name and path to the database being loaded. You can interrupt the loading process by clicking on the **Cancel** button in the pop-up window and then load a new feature database by clicking on **File** from the top-level menu in CSD-CrossMiner, and **Load Feature Database**. Note that an alternative feature database can be loaded at any point during a CSD-CrossMiner session, but that doing so will clear the current session of any existing pharmacophores or results.

Software and Feature Database Updates

CSD-CrossMiner software and CSD-CrossMiner feature database have a built-in auto-update mechanism that makes updating both the software and the database easy, and will also notify you if a software or data update is available. Each option will check for an update after CSD-CrossMiner session being open for five minutes and if it determines an update is available it will let you know. It is also possible to initiate a manual check by using the **Help** menu in the CSD-CrossMiner top-level menu and selecting either **Check for Software Updates** and/or **Check for Database Updates**.



You can tick the **Automatically check for updates** tick-box so that everytime that a CSD-CrossMiner session is open the software will also check for database updates.

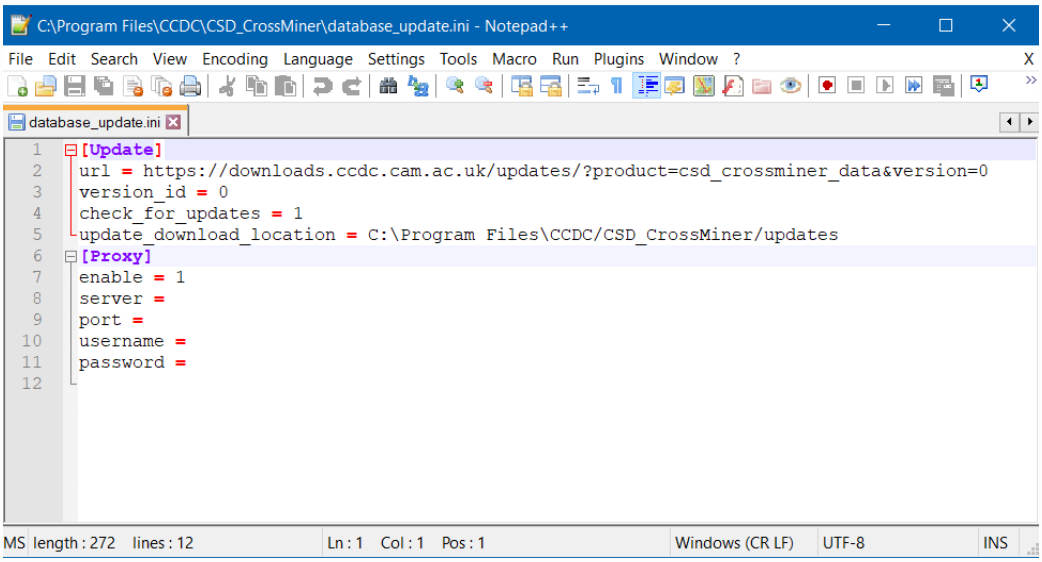


Once the auto-update mechanism has identified that an update is available, you will be given the option to either **Install on exit** (this will start the updated process when you close the CSD-CrossMiner application) or **Don't install** if you don't want to update at this time.

Note: The speed of the download depends on the quality of your network.

It is also possible to download and install the database update manually, by accessing the 'Data & Software Updates' section of our Downloads page.

If for some reason the `crossminer_data` folder in `CSD_2022` directory has been manually moved or deleted the database autoupdate mechanism won't install the updated CSD-CrossMiner feature database. In such cases, it is possible to manually intervene by editing the `database_update.ini` file located in the `CSD-CrossMiner` directory (`<CCDC installation folder>/Discovery_2022/CSD-CrossMiner`) and set the version numbers in the `url` and `version_id` to `0` as displayed below. This will trigger the download of the up-to-date CSD-CrossMiner feature database. You can use any text editor software to open and edit the `database_update.ini` file.



Structure and Feature Databases Supplied with CSD-CrossMiner

CSD-CrossMiner is supplied with the structure and feature databases containing the molecular structures of small molecules stored in the CSD and protein-ligand binding sites extracted from the PDB.

For CSD structures (version 5.43) a subset has been created containing structures that (a) are organic plus a small list of transition metals, i.e., Mn, Fe, Co, Ni, Cu, Zn (b) have an R-factor of at maximum 10%, (c) for which 3D coordinates have been determined, (d) have no disorder, and (e) are not polymeric. This resulted in the csd453_crossminer subset containing over 400 000 structures.

For PDB structures, database entries were generated for each HET group (non-standard protein residues) in protein-ligand complexes, except metals, water molecules and commonly found small ions. Two subsets of the PDB structures have been created, the pdb_crossminer subset consists of PDB structures that a) are not determined by Electron Microscopy (EM) or CryoEM, b) do not contain nucleic acids, c) have a resolution < 3Å, d) have ligands with more than 5 or less than 100 atoms, e) contain only the first model in NMR structures, f) does not contain unknown ligands. A second nucleic_acid _crossminer subset consists of PDB structures that a) are not determined by Electron Microscopy (EM) or CryoEM, b) contain nucleic acids in addition to protein-ligand complexes, c) have a resolution < 3.5Å, d) have ligands with more than 5 or less than 100 atoms, e) contain only the first model in NMR structures, f) does not contain unknown ligands. The Hydrogens were added using the Protein `add_hydrogens` function in the CSD Python API (see CSD Python API Documentation).

All molecules and atoms in the protein-ligand binding site were included in the respective database entry in mol2 format, where the binding site is defined within a cut-off radius of 6 Å around the selected HET group. Note that for nucleic_acids_crossminer subset, the protein-ligand binding site may not contain DNA/RNA. The resulting CSD SQL FastBinary structure databases

(`pbd_crossminer.csdsqlx` and `nucleic_acid_crossminer.csdsqlx`) consist of more than 300 000 binding sites derived from over 60 000 PDB entries.

The supplied feature database holds the structures from the structure database described above, indexed with a set of feature definitions, to define the ensemble of steric and electronic features that characterise a protein, a nucleic acid and/or a small molecule. This feature database is used to perform the actual 3D search. The feature definitions used to create the supplied feature database are stored in `<CCDC Installation Folder>/Discovery_2022/CSD-CrossMiner/feature_definitions` (see APPENDIX B. Feature Definitions in CSD-CrossMiner).

In addition to the provided structure and feature databases, CSD-CrossMiner allows the user to create both structure and feature databases starting with their own structures (see Creating Databases).

Entry Identifiers

All entries in the structure and feature databases need to have an identifier that is unique in that database, that is generally defined during the creation of the structure database. In the provided feature database, the indentifier used for a small molecule structure is the Refcode used in the CSD. For the PDB structures in the supplied feature database, the unique identifier contains detailed information about the entry such as: PDB structure ID, model number, the protein chain(s) involved in the protein-ligand interactions and the ligand ID. For example:

- `1A9U_m1_A_bs_SB2_A_800` corresponds to the binding site from 1A9U, single model structure with only one ligand: SB2 in chain A with residue number 800.
- `1A29_m1_A_bs_TFP_A_153` and `1A29_m1_A_bs_TFP_A_154` correspond to the binding sites for structure 1A29, single model structure with two TFP ligands.
- `3D5Q_m1_A-B_bs_T30_A_293`, `3D5Q_m1_A-B_bs_T30_B_1`, `3D5Q_m1_A_bs_NAP_A_1`, `3D5Q_m1_B_bs_NAP_B_2`, `3D5Q_m1_C-D_bs_T30_C_1_2`, `3D5Q_m1_C-D_bs_T30_D_1`, `3D5Q_m1_C_bs_NAP_C_3`,

`3D5Q_m1_C_bs_NAP_D_4`, and `3D5Q_m2_C-D_bs_T30_C_1_2`, correspond to the binding sites of different ligands and binding sites with more than one protein chain forming the binding site.

The identifiers used in the feature database can be saved by clicking on **File** from the top-level menu in CSD-CrossMiner and then **Export Identifiers**. The saved identifiers can be used to create a csv file containing annotations for the database entries (see Annotating a Feature Database).

Annotations

The annotations in a feature database contain information about each database entry such as: the deposition date, the resolution, the PDB code, the CSD refcode. The data takes the form of key-value pairs of text and is displayed in the **Results Hitlist** browser for all hits matching the pharmacophore query.

The list of all annotations used in the supplied feature database are listed below:

`chain`, `deposition_date`, `ec_number`, `molecule`, `is_covalent`, `molecule_fragment`, `molecule_synonym`, `organism`, `organism_taxid`, `pdb`, `pdb_class`, `pdb_title`, `resolution`, `structure_method`, `CSD Refcode`, `formula` and `r factor`.

These annotations can be used to filter the database from which hits can be found (see Using Annotations as Filter).

Feature database can be annotated during a CSD-CrossMiner session with additional user-defined data (see Annotating a Feature Database). In addition, entries can be annotated during the process of creating the structure database (see APPENDIX D. Create a Feature Database with In-House Data).

Creating, Modifying and Saving Pharmacophore Queries

To start a pharmacophore search, a pharmacophore query must be created or loaded. This activates the pharmacophore search buttons in CSD-CrossMiner. Instructions on how to load, create and modify a pharmacophore query are discussed below. Instructions on how to start, pause and stop a pharmacophore search are discussed later (see Pharmacophore Search).

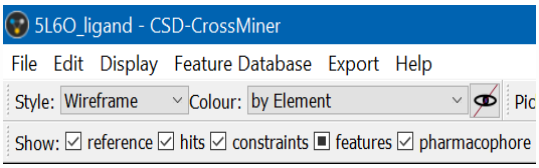
Loading an Existing Pharmacophore Query

An existing pharmacophore query can be loaded by selecting **File** from the top-level menu in CSD-CrossMiner, then **Load Pharmacophore**. Several pharmacophore query examples can be found in the `example_pharmacophores` folder of the `CSD-CrossMiner` directory. Once a pharmacophore is loaded, it is displayed in the 3D view and the corresponding feature definitions are shown in the **Pharmacophore Features** window.

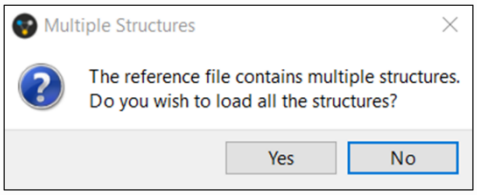
A pharmacophore search can then be started for this pharmacophore query by clicking on ▶️ (see Pharmacophore Search). The pharmacophore query can be cleared via **File** > **Close Pharmacophore**.

Creating a Pharmacophore Query from a Reference Structure

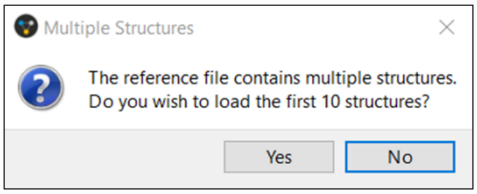
A pharmacophore query can be created from a reference structure that can be loaded via **File** > **Load Reference** (common molecule file formats are supported). The name of the loaded reference structure is then included in the title bar of the CSD-CrossMiner window (5L6O_ligand in the screenshot below).



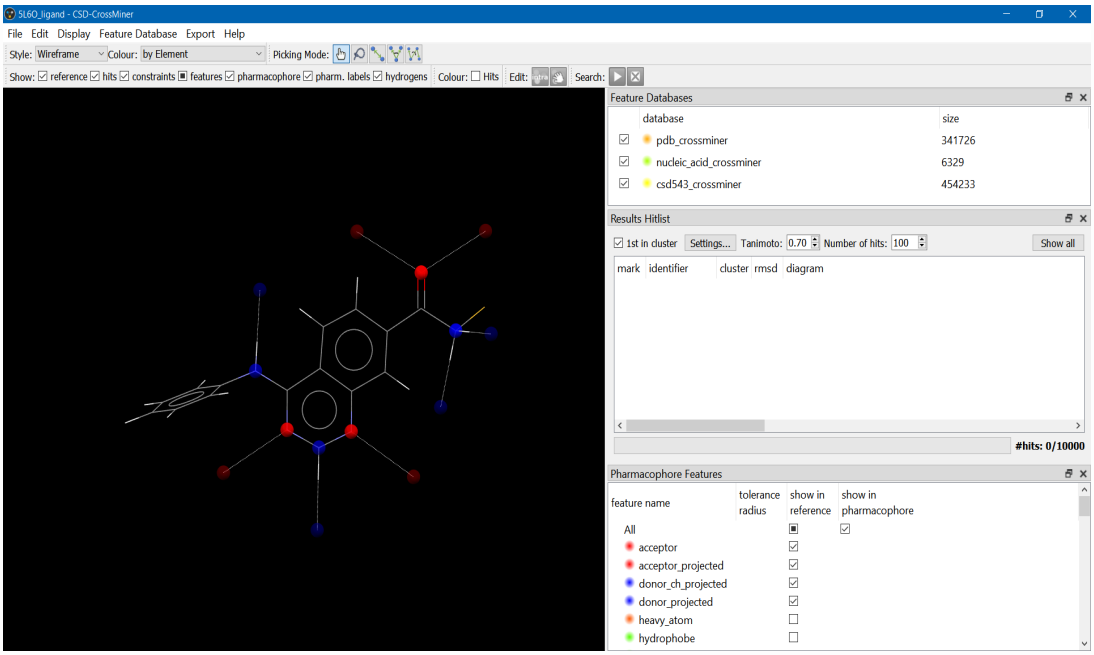
In CSD-CrossMiner it is possible to use a multi structures file as reference (e.g., overlaid ligands). Note that the maximum number of reference structures that can be loaded and displayed in CSD-CrossMiner is ten. If the loaded reference file contains up to ten structures, by clicking **Yes** in the **Multiple Structures** pop-up window, all the structures included in the reference file will be loaded and displayed in the 3D view; by clicking **No** only the first structure in the reference file will be loaded.



If a reference file contains more than ten structures, clicking **Yes** in the **Multiple Structures** pop-up window will load and display only the first ten structures of the reference file in the 3D view. Otherwise, by clicking **No**, only the first structure of the reference file will be loaded.



When a new molecule from a reference structure is loaded in CSD-CrossMiner, by default only donor and acceptor features associated with the reference structure are displayed in the 3D view. The features are represented in the 3D view as small translucent spheres, whose identity and associated colour is shown in the **Pharmacophore Features** window.

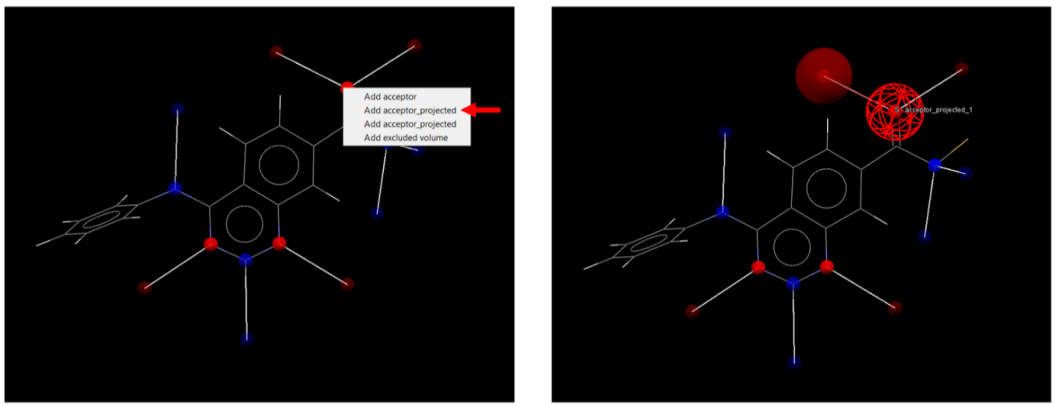


The displayed features of the reference structure are ticked in the **show in reference** column in the **Pharmacophore Features** window. Features can be displayed/hidden in the 3D view by ticking/unticking the relative tick-box in the **Pharmacophore Features** window. When only some of the features of the reference structure are displayed, the **features** tick-box in the CSD-CrossMiner **Show** toolbar and the **All** features tick-box in the **Pharmacophore Features** window are shown as ▣.

All features associated with the reference structure can be displayed by clicking on the **feature** ▣ tick-box in the CSD-CrossMiner **Show** toolbar (which will turn to ☑) or alternatively by ticking the tick-box for **All** features in the **show in reference** column in the **Pharmacophore Features** window. All features can be hidden by clicking on **features** ☑ (which will turn to ☐) or alternatively by unticking the **All** tick-box.

Note that if a different choice of displayed features is made (e.g., all features displayed), the new settings will be remembered if a new reference molecule is loaded (e.g., in this case all features of the new reference molecule would be displayed).

Right-clicking on a feature of the reference structure allows a pharmacophore point of the type(s) available for this feature to be created.



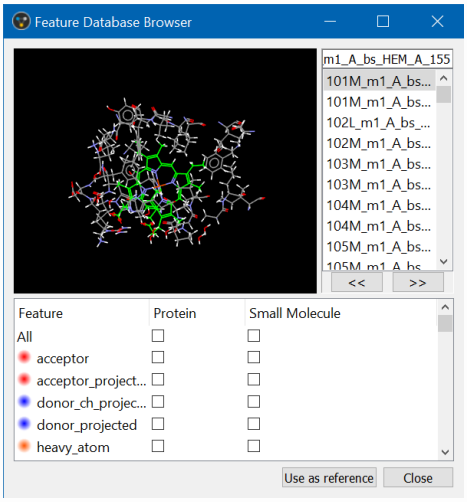
Note that for some features (e.g., acceptor) it is possible to define the directionality of the feature by choosing to create a projected pharmacophore point (e.g., acceptor_projected). In doing so, the pharmacophore search will lead to hits where both the base and the virtual feature(s) of the pharmacophore point have to be satisfied (see Editing and Creating Feature Definitions).

A pharmacophore search can then be started for this pharmacophore query by clicking on ▶️ (see Pharmacophore Search). The entire pharmacophore query and the reference molecule can be cleared via **File** > **Close Pharmacophore** and the **File** > **Close Reference** respectively.

Creating a Pharmacophore Query from a Feature Database Entry

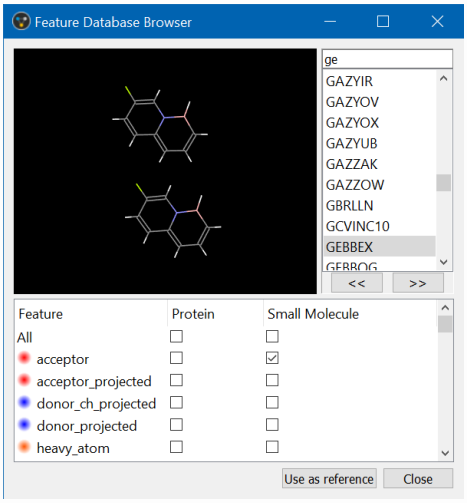
The feature database browser can be accessed by clicking on **Feature Database** in the CSD-CrossMiner top-level menu and then **Browse**. The **Feature Database Browser** pop-up window displays the list of all entries stored in the feature database and the feature definitions divided by **Protein** and **Small Molecule**. By default, the first molecule in the list is shown in the 3D Display of the **Feature Database Browser** window. In the 3D Display, the carbon atoms of the ligand are coloured in green, while the carbon atoms in the protein residues are coloured in grey. It is possible to change these

settings using the **Style Preferences** in the **Display** top-level menu (see Setting Default Style and Colour Preferences for Reference and Hit).



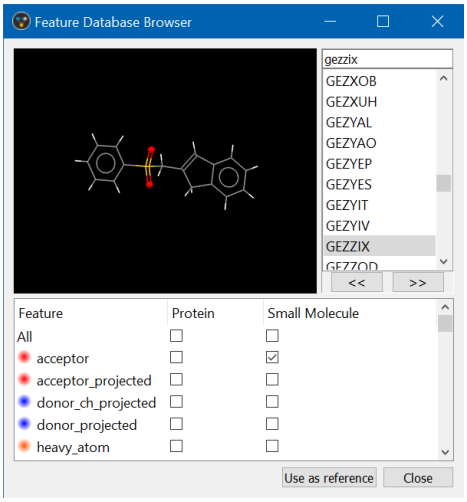
The user can navigate through the different database entries by clicking on a structure name in the right-hand panel, or by clicking on the **<<** and **>>** buttons at the bottom of it, or by using the up and down keyboard arrows to scroll through the list.

It is also possible to search for a specific structure in the database by typing the entry in the top-right text box of the **Feature Database Browser** window. Note that the search is not case sensitive, and the list will be sorted by relevance while typing, with the most relevant result appearing at the top of the list.



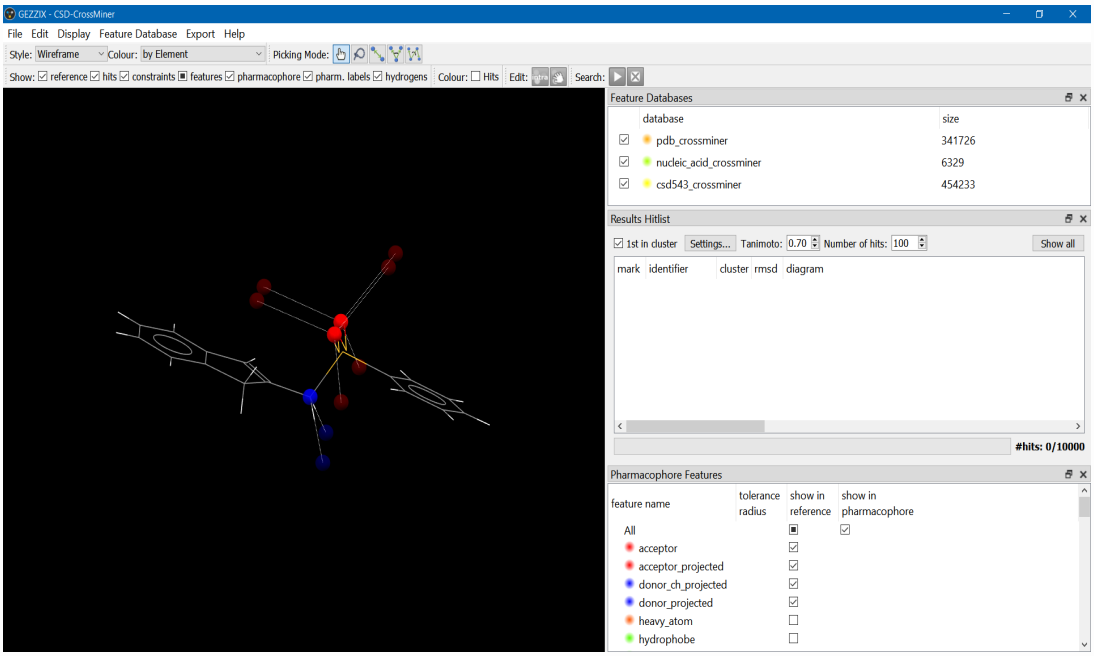
Selecting a feature database entry in the feature database list in the right-hand panel also displays it in the 3D display of the **Feature Database Browser** window (note that multi-selection is not allowed). Additionally, features associated with the selected

structure can be displayed by ticking the **Protein** and/or **Small Molecule** tick-box corresponding to the desired feature types. Note that nucleic acids features are associate with the **Small Molecule** component.



Once a feature database entry is selected, it can then be loaded in the main 3D view of CSD-CrossMiner by clicking on the **Use as reference** button in the **Feature Database Browser** window.

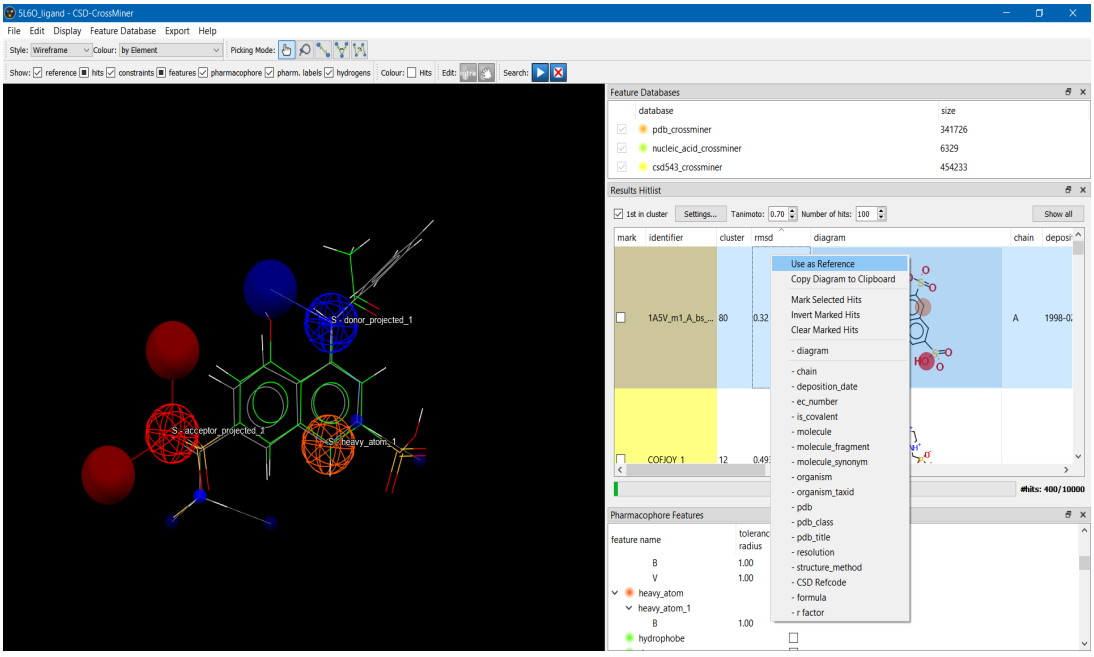
Note that the displayed features of the new reference structure will correspond to the choice of displayed features made during the CSD-CrossMiner session (features with ticked tick-box in **show in reference** column in the **Pharmacophore Features** window).



Features can be displayed, and pharmacophore points can be created from the features of the selected database entry that is now used as a reference structure (see Creating a Pharmacophore Query from a Reference Structure).

Creating a Pharmacophore Query from a Hit

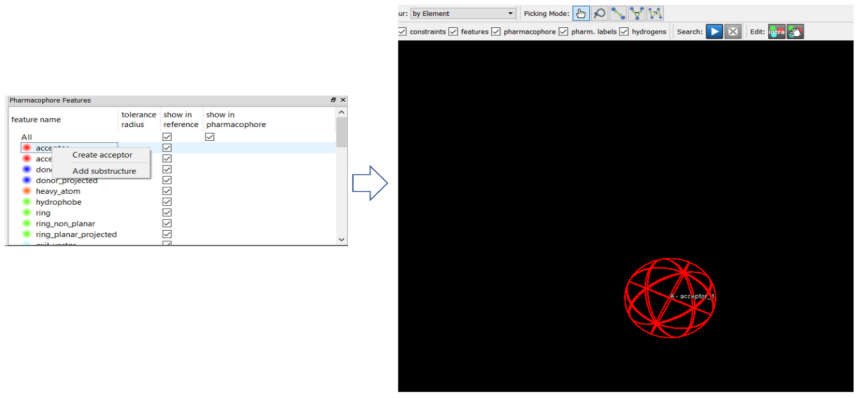
A pharmacophore query can also be created based on any hit identified during a pharmacophore search, by right-clicking on the hit of interest in the **Results Hitlist** browser and selecting **Use as reference** from the resulting context menu. The pharmacophore query can then be defined from the features of this reference hit structure (see Creating a Pharmacophore Query from a Reference Structure).



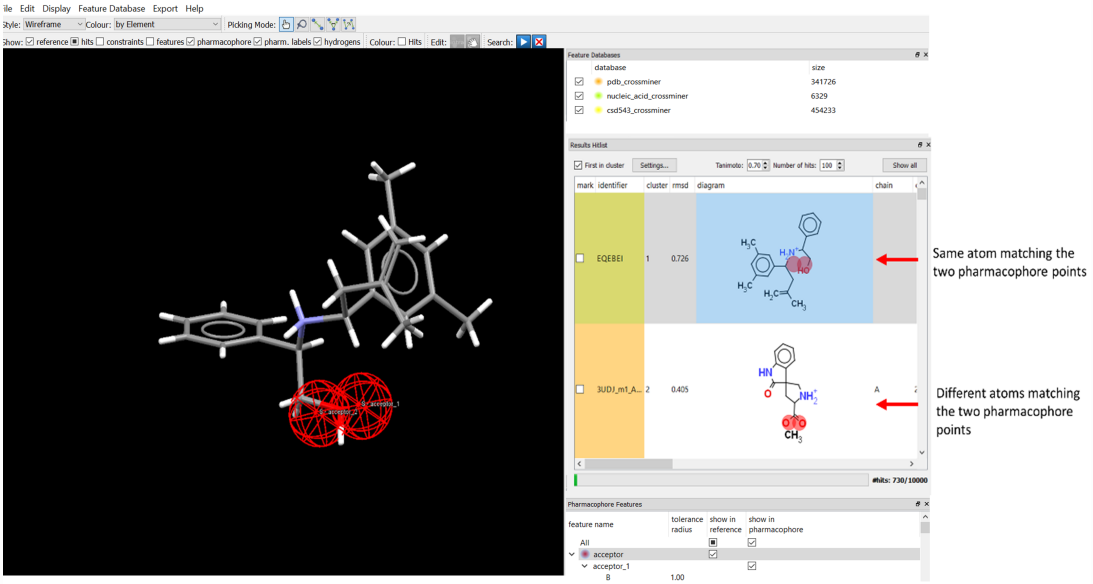
Creating a New Pharmacophore Query

Starting from an empty 3D view (if necessary, delete an existing pharmacophore query using **File** > **Close Pharmacophore** and/or delete any existing reference structure using **File** > **Close Reference**), pharmacophore points can be added from scratch by

right-clicking on a desired feature type in the **Pharmacophore Features** window and then choosing **Create** from the resulting context menu. The corresponding pharmacophore point will appear in the 3D view:



By default, this will create an any molecule (A), pharmacophore point. The molecule type of the pharmacophore point can be modified to be protein (P) or small molecule (S) (see Modifying a Pharmacophore Query). If multiple pharmacophore points are created in this manner, they may be overlaid on top of each other in the 3D view. It is possible to translate each pharmacophore point by clicking on the interactive pharmacophore editing mode  in the **Edit:** toolbar. This will turn the mouse cursor to a small hand, allowing the pharmacophore point to be translated, even during a pharmacophore search. Pressing the left mouse button (LMB) and dragging a pharmacophore sphere to a given position will move it to this position.



Note that if pharmacophore points overlap, it is possible to obtain hits where the same atom in a structure can match multiple pharmacophore points.

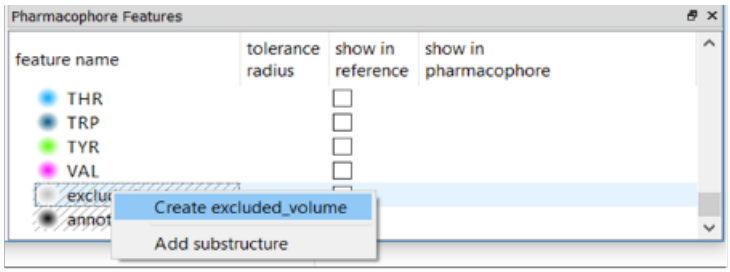
Adding an Excluded Volume to a Pharmacophore Query

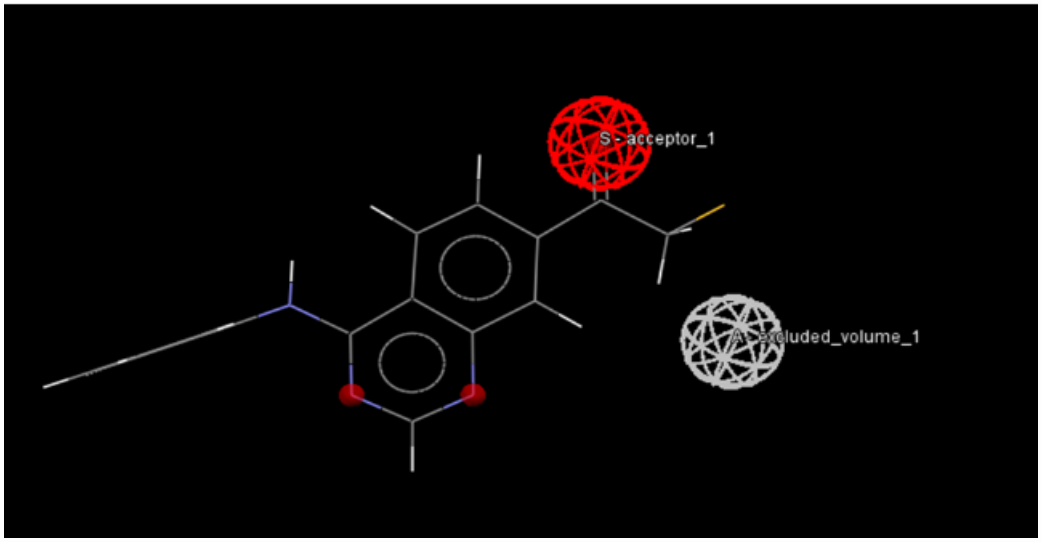
An excluded volume is a special type of pharmacophore point that can be added to a pharmacophore query. This feature can be defined as a volume of occupation that can be set to be any of the three molecule types, i.e., Protein, Small Molecule or Any Molecule. There is no limitation on the number of excluded volume features that can be used in a pharmacophore query.

Because an excluded volume feature is never indexed in the feature database, it is represented with diagonal hatching in the **Pharmacophore Features** window.

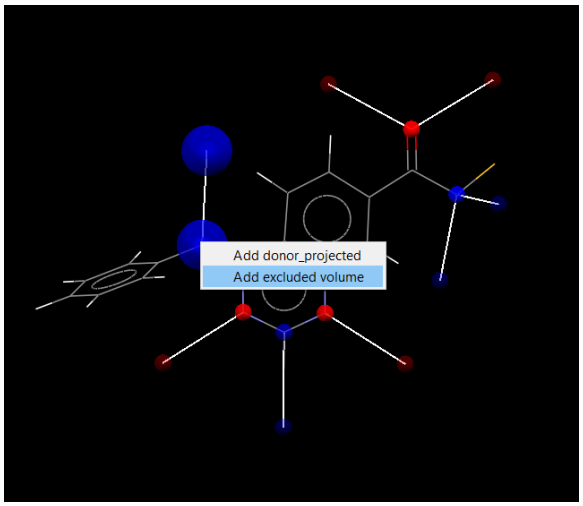
An excluded volume feature can be added:

- From scratch: through the **Pharmacophore Features** window, by right-clicking on **excluded_volume** feature and then selecting **Create excluded_volume.** Note that it is not possible to perform a pharmacophore search with only an excluded volume pharmacophore point defined.

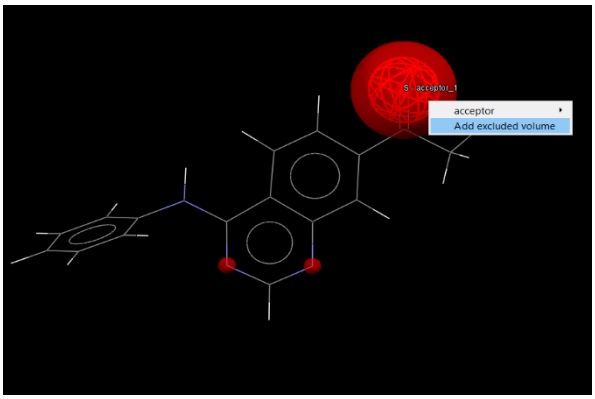




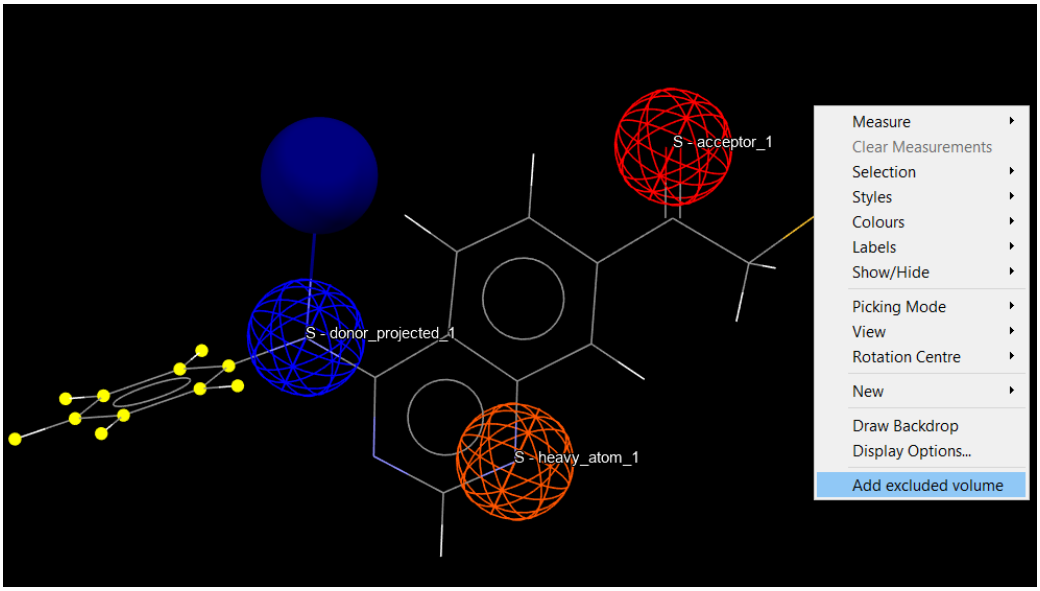
- From the features of a reference structure: by right-clicking on a feature and then selecting **Add excluded volume**.



- From an existing pharmacophore: by right-clicking on a pharmacophore point and then selecting **Add excluded volume**. Note that this will add an excluded volume pharmacophore point on top of the other pharmacophore point.



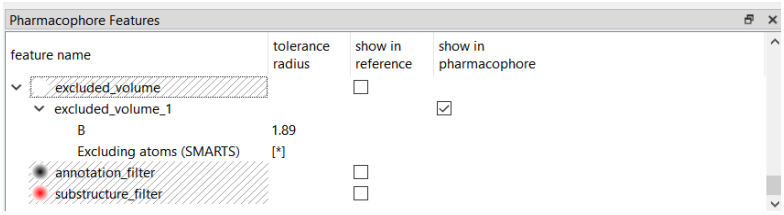
- From the atom(s) of a reference structure: by selecting the atom(s) of a structure loaded in the 3D view (click on each atom or select multiple atoms when in lasso mode), then right-clicking anywhere in the 3D view and selecting **Add excluded volume** from the context menu.



Note: the excluded volume pharmacophore sphere will include all atoms thus selected.

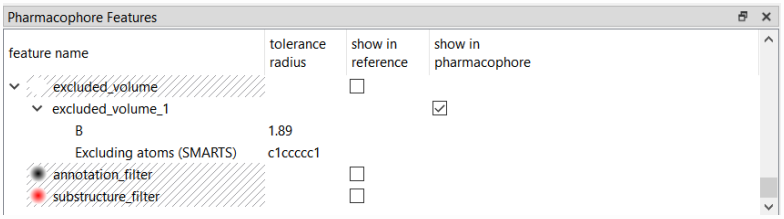
Regardless of how an excluded volume pharmacophore point is created, its tolerance radius can be changed (see Changing the Pharmacophore Tolerance) and its position can be changed (see Creating a New Pharmacophore Query).

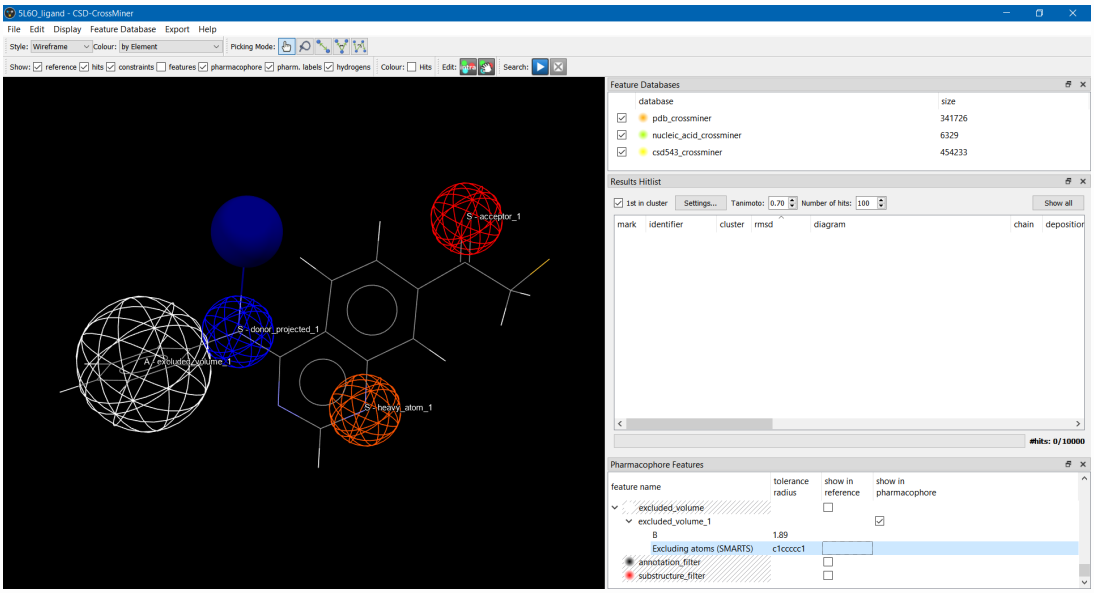
By default, an excluded volume pharmacophore feature point will result in rejection of any hit with any atom located within the specified volume represented as **[*]** in the **Excluding atoms (SMARTS)** section of the **excluded volume** pharmacophore point in the **Pharmacophore Features** window.



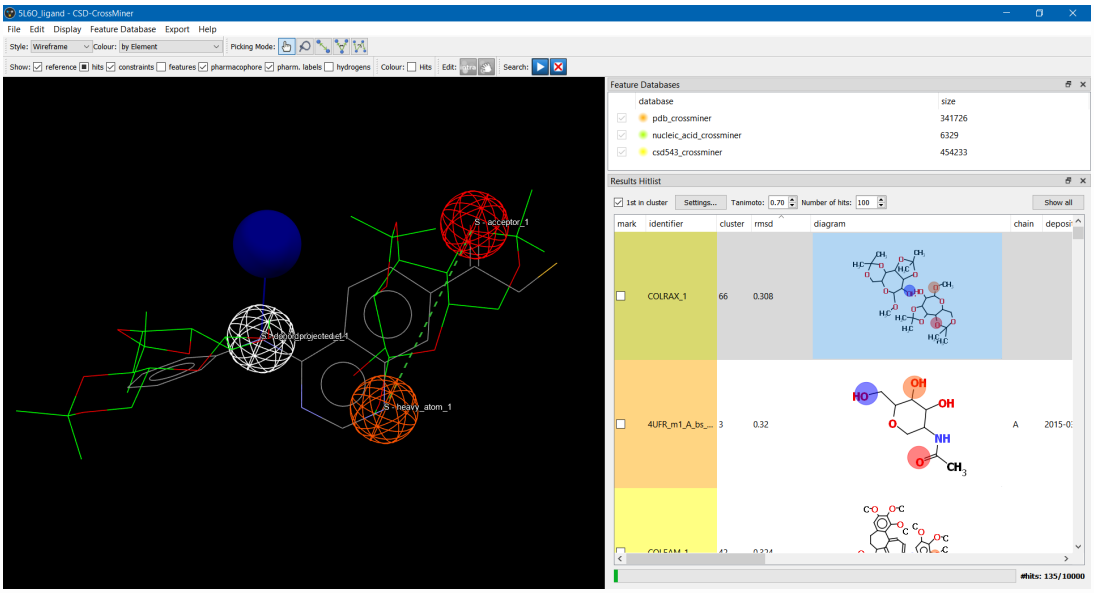
Use the **feature name** column delimiter in the **Pharmacophore Feature** to visualise the full text.

The **[*]** argument is user editable therefore, by double-clicking on it, is possible to define the SMARTS pattern corresponding to a substructure you wish to exclude within the specified volume radius (e.g., **c1ccccc1** in the example below).





It is also possible to combine excluded volume pharmacophore points with other pharmacophore points to exclude specific moieties that otherwise will match the defined pharmacophore feature. In the example below a customised excluded volume with **Excluding atoms (SMARTS) [OH2],** has been added on top of the H-bond donor pharmacophore point, this will result in hits where the H-bond donor feature is not matching waters.



Note that despite it is possible to create an excluded volume any time during the search, it is not possible to customise the **Excluding atoms (SMARTS)** pattern when the pharmacophore search is running and/or has been paused, but only when it has not yet started or has been stopped (see Modifying a Pharmacophore Query).

Modifying a Pharmacophore Query

In CSD-CrossMiner, a pharmacophore point can be modified:

- Before starting the pharmacophore search.
- When the pharmacophore search is complete or stopped (by clicking on the **Stop** button)
- During the pharmacophore search itself.

Note that it is not possible to modify a pharmacophore point when the search is paused (by clicking on **Pause** button).

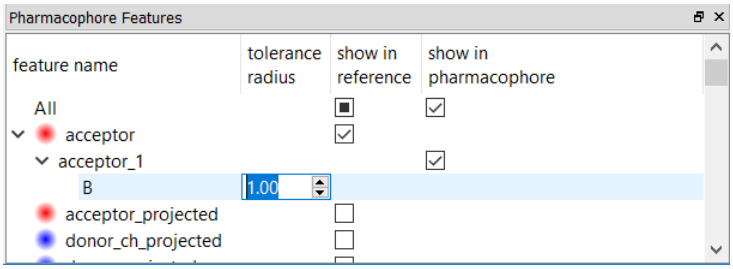
Translating a Pharmacophore Point

A pharmacophore point can be translated by clicking on button in the **Edit** section of CSD-CrossMiner's lower toolbar.

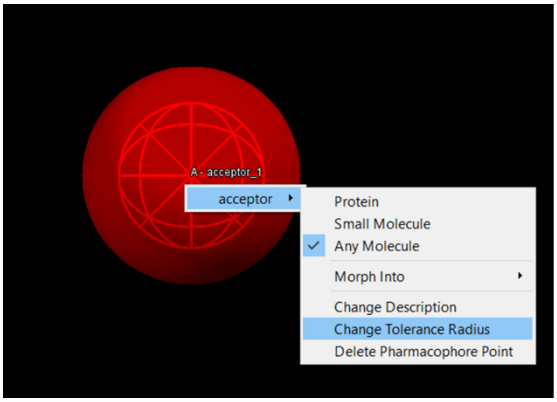
Changing the Pharmacophore Tolerance

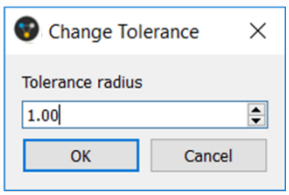
A pharmacophore point by default has a tolerance radius of 1.00 Å. This tolerance radius can be modified in order to increase or decrease the uncertainty in the position of this pharmacophore point in the overall pharmacophore query. The tolerance radius of a pharmacophore point can be changed in three ways:

- By double-clicking on the desired tolerance radius in the **Pharmacophore Features** window to have access to the spin-box and either using the up-down control to change the radius by 0.1 Å increment or entering a desired value in the text box.



- By right-clicking on the pharmacophore point, to have access to the pharmacophore context menu, and double clicking on **Change Tolerance Radius.** The **Change Tolerance** pop-up window allows to change the tolerance sphere of the pharmacophore point by 0.1 Å increment using the spin-box and either using the up-down control or entering a desired value in the text box.

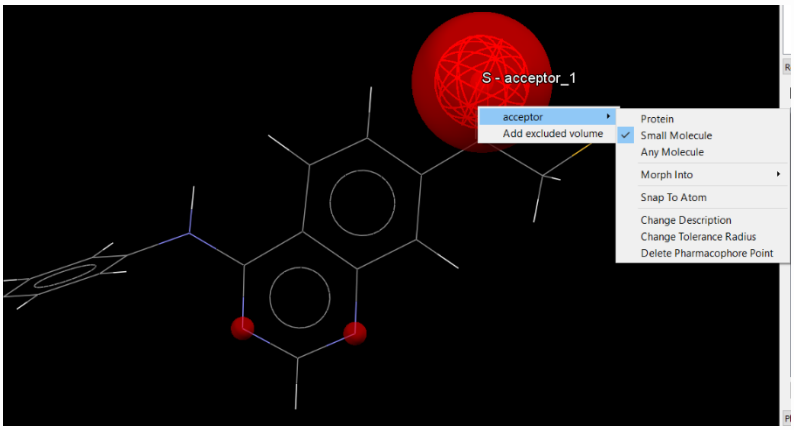




- When in interactive pharmacophore editing mode (having pressed the  button), pressing the middle mouse button (MMB) whilst above a pharmacophore sphere and moving the mouse will increase/decrease the tolerance radius for the pharmacophore point.

Changing the Molecule Type

A pharmacophore point can be modified by accessing the pharmacophore context menu, available by right-clicking on a pharmacophore point. The molecule type of a pharmacophore can be changed to be Protein (P), Small Molecule (S) or Any molecule (A) by selecting **Protein**, **Small Molecule** or **Any Molecule** from the right-click pharmacophore menu for the pharmacophore point. The pharmacophore point labelling in the 3D view is then updated to indicate the chosen molecule type as P, S, or A.

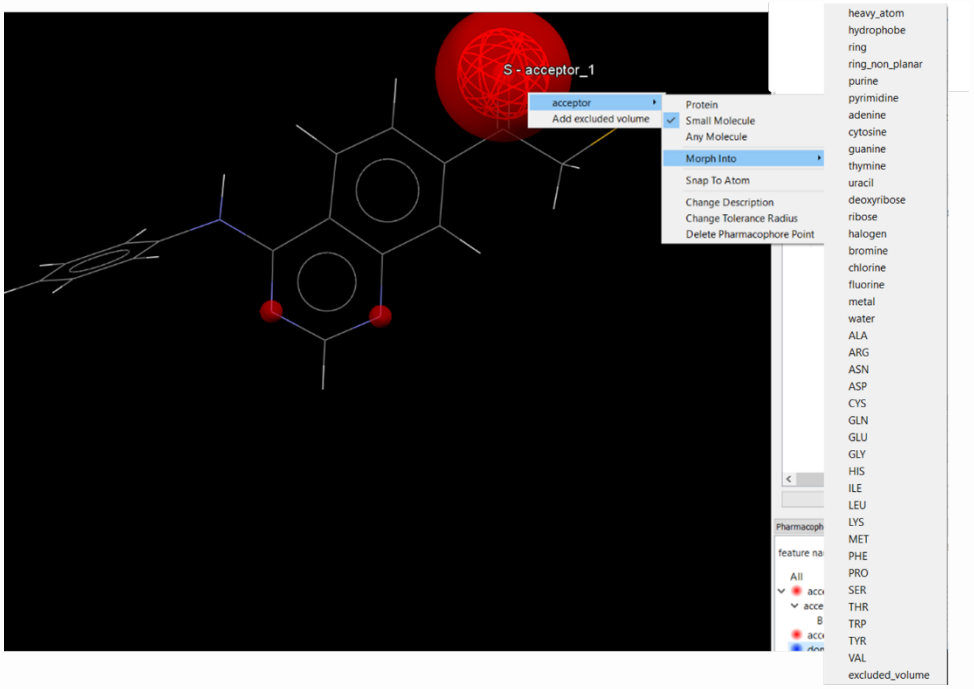


Changing the Pharmacophore Type

A pharmacophore point can be changed into another pharmacophore type by selecting **Morph Into** from the right-click pharmacophore context menu and choosing a new pharmacophore type from the list of available pharmacophore point types.

Directional pharmacophore point(s) can only be morphed into another directional pharmacophore point(s), and equally a one-point pharmacophore point can only be morphed into another

one-point pharmacophore point. (APPENDIX B. Feature Definitions in CSD-CrossMiner for a list of all one-point and directional pharmacophore feature definitions).

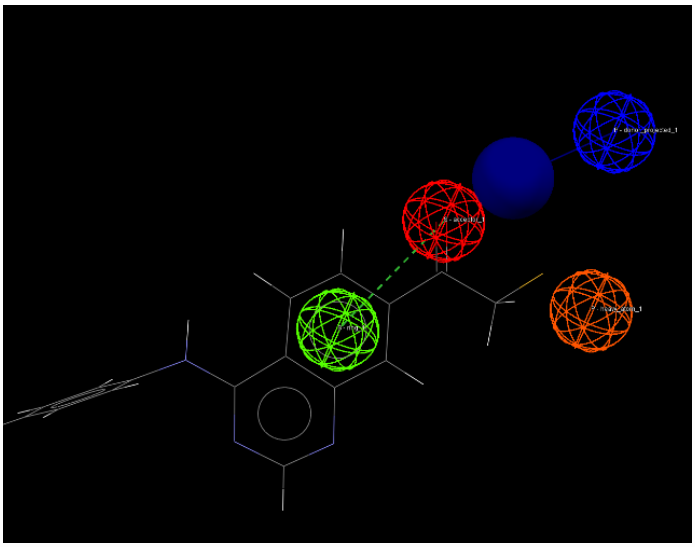


Note that an excluded volume pharmacophore point can be additionally created from an existing one-point pharmacophore: by right-clicking on a pharmacophore point and then selecting **Morph Into** and then **excluded_volume**.

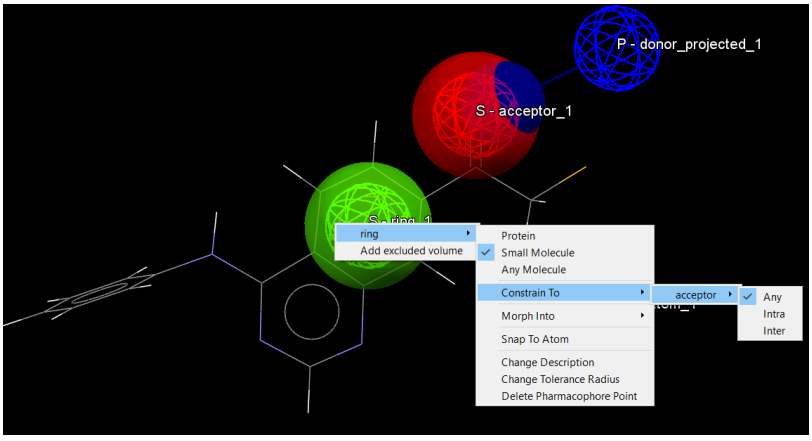
Setting Intramolecular and Intermolecular Constraints

In CSD-CrossMiner it is possible to set intramolecular constraints between all **Small Molecule** and **Any Molecule** pharmacophore points displayed in the 3D view by clicking on 

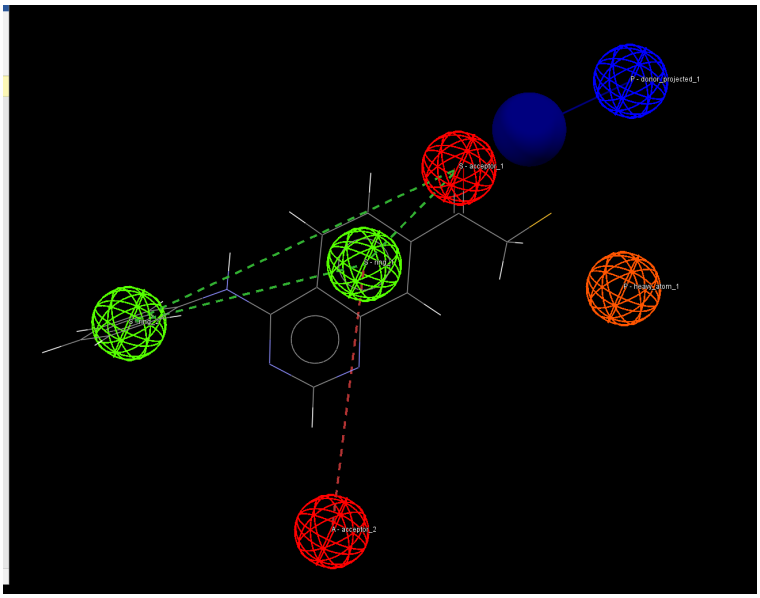
button. The intramolecular constraints are displayed as green dashed lines in the 3D view. It is also possible to create individual intramolecular constraints, intermolecular constraints, as well as ‘Any’ constraints (where the two pharmacophore points can either belong to the same molecule or not).



To access to these constraint options, right-click on a pharmacophore point and then select **Constrain to**. This will list all pharmacophore points present in your pharmacophore query. Pharmacophores points of the same molecule type can be constrained to be part of the same molecule (**intra**) or part of different molecules (**inter**). If **Any** is selected both intra- and intermolecular hits may be found when searching with such a pharmacophore query. Intermolecular constraints are represented as red dashed lines in the 3D view.



Note that if a pharmacophore point is set to **Any Molecule (A)** type, both protein and small molecule molecules (including nucleic acids) are considered as a match; therefore, intramolecular constraints involving an **Any Molecule** pharmacophore point can only be set with any other **Any Molecule** pharmacophore point, and intermolecular constraints can be set with any other **Any Molecule**, **Small Molecule** or **Protein** pharmacophore point.



The multitude of possible combinations of intra- and intermolecular constraints allows for highly tailored interrogations of nonbonded interactions.

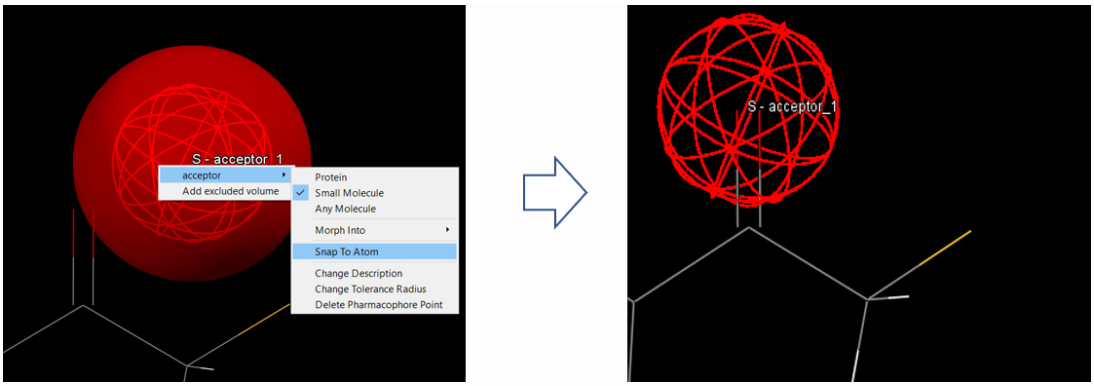
Note that constraints can be enabled/disabled before starting a search, interactively during the search, or when the search is terminated (whether stopped or complete), but not when the pharmacophore search is simply paused.

Changing the molecule type or morphing a pharmacophore point will automatically delete all intra- and intermolecular constraints that this pharmacophore point was previously involved in.

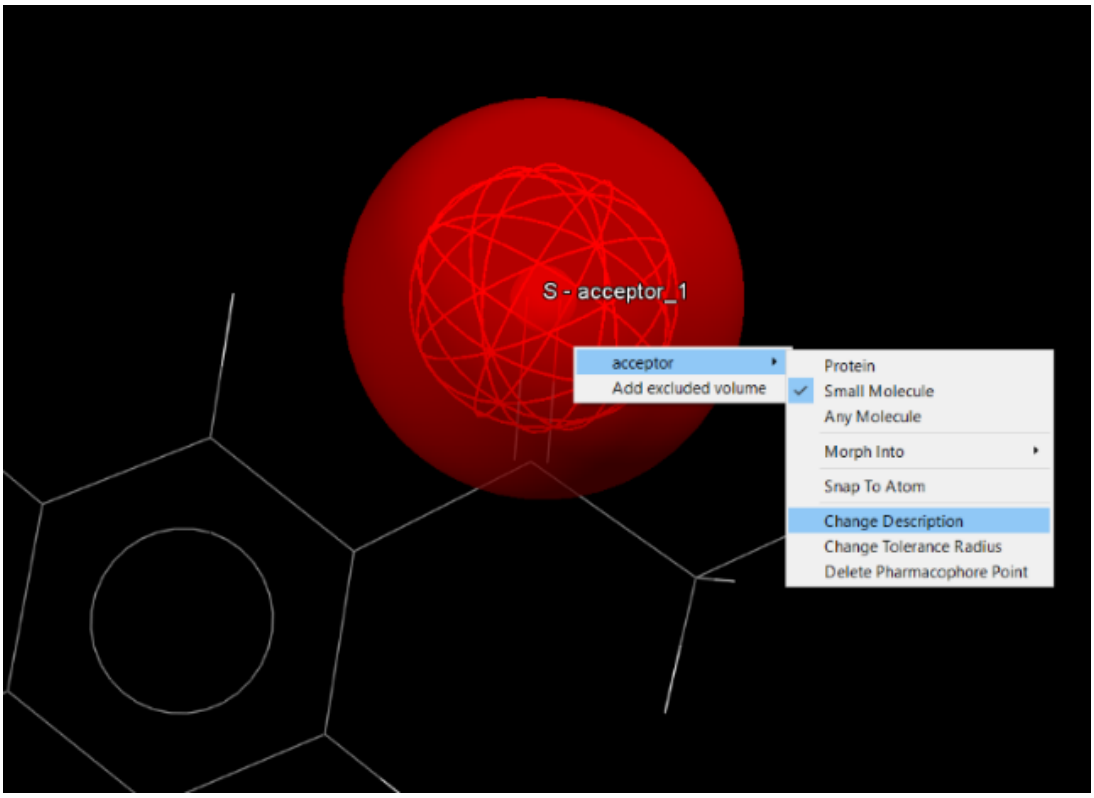
Note that it is not possible to constrain excluded volume pharmacophore points.

Further Editing of the Pharmacophore Point

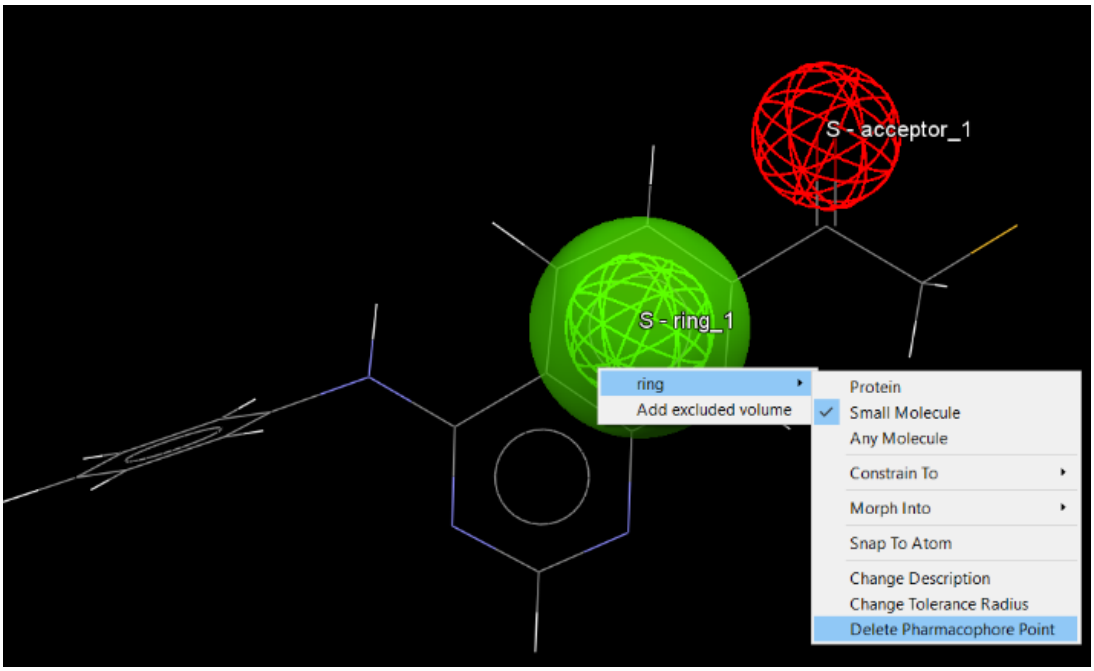
If a pharmacophore point is supposed to coincide with an atom position, it is possible to drag the pharmacophore sphere close to the atom and select **Snap To Atom** from the right-click context menu of the pharmacophore point.



Upon accessing the right-click context menu of a pharmacophore point, it is also possible to change its labelling in the 3D view by selecting **Change Description**. The **Change Feature Description** pop-up window invites the user to enter a new description.



Finally, it is possible to delete a pharmacophore point by selecting **Delete Pharmacophore Point** in the right-click context menu.



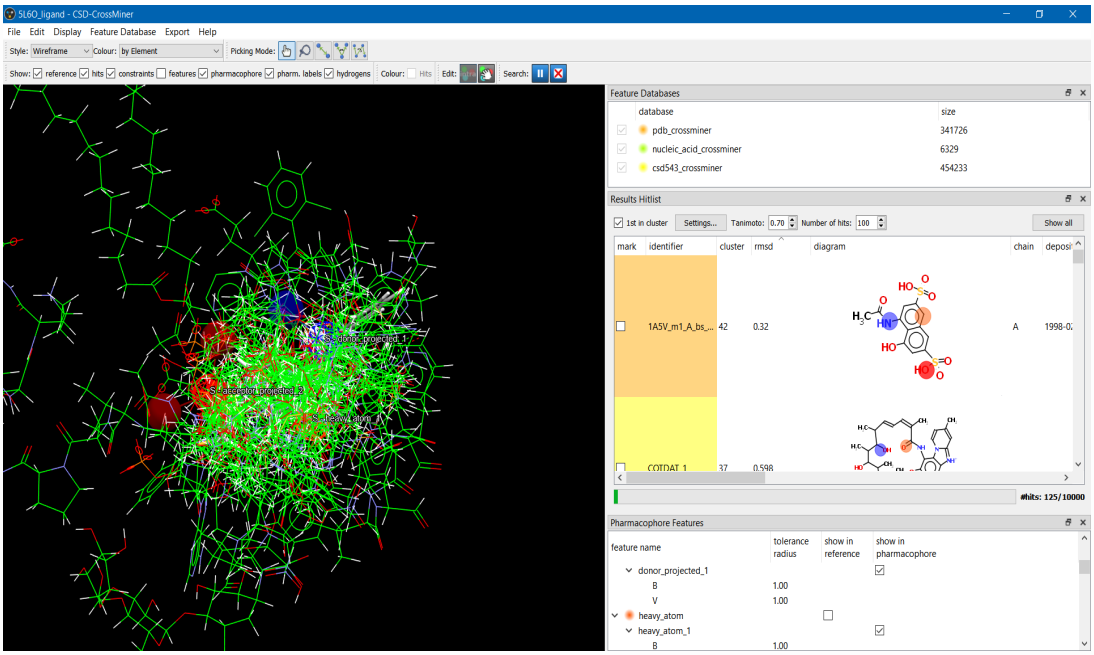
Saving a Pharmacophore Query

After creating and/or editing a pharmacophore, the resulting pharmacophore query can be saved, either in the CSD-CrossMiner pharmacophore format (`.cm`) by clicking on **File** in the CSD-CrossMiner top-level menu and then selecting **Save Pharmacophore**, or in PyMOL pharmacophore format (`.py`) by clicking on **File** and then selecting **Save PyMOL Pharmacophore**.

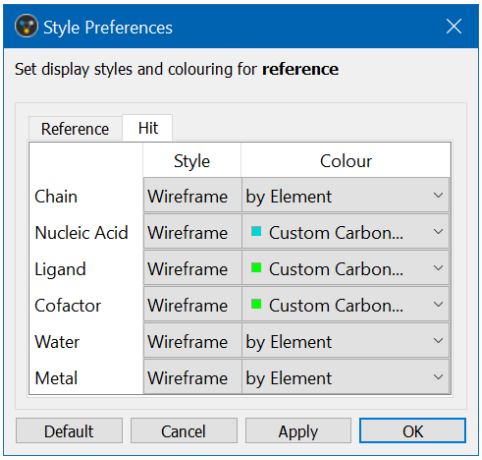
A pharmacophore query in the CSD-CrossMiner pharmacophore format can be loaded into CSD-CrossMiner using **File** > **Load Pharmacophore**. Note that a pharmacophore query in the PyMOL format cannot be loaded into CSD-CrossMiner but is compatible with third party software such as PyMOL.

Pharmacophore Search

Once the pharmacophore query has been created, a pharmacophore search can be started by pressing the **Start** ▶️ button in the CSD-CrossMiner toolbar. The loaded feature database will be investigated to find hits matching the pharmacophore query, and both the 3D view and the **Results Hitlist** browser will be updated with the identified hits:



By default, the small molecule hits will be displayed in wireframe with the carbon atoms coloured in green and the protein residues with the atom coloured by element. However, you can change the way to display the reference and the hits and all the different components using the **Style Preference** in the **Display** menu (see Setting Default Style and Colour Preferences for Reference and Hit).



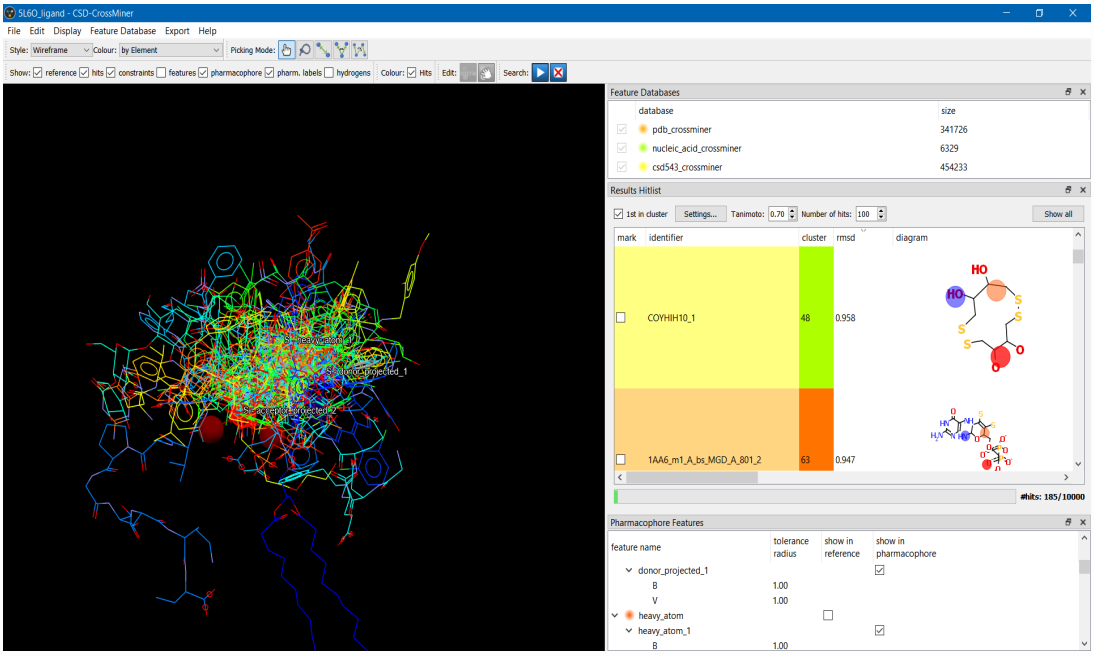
Note that the number of hits in the progress bar will update every 5 hits. During the pharmacophore search, any change in the topology or geometry of the pharmacophore query triggers a new database search and consequently both the 3D view and the **Results Hitlists** browser update, giving the user instant feedback on the kind of hits to be expected from the current query setup. During the search, the **Start** button turns to a **Pause** button that can be used to pause

the pharmacophore search (is then restored and can be used to continue the search). Note that in pause mode, the pharmacophore query cannot be interactively changed.

While the search is in pause, to better distinguish between the different hits matching the pharmacophore query, it is possible to colour the matching hits by ticking the **Colour:** ☐ **Hits** tick-box. Hits will be coloured by rainbow in the 3D view and the associated colour displayed in the **cluster** column in the **Results Hitlist** browser.

Note that the colouring will only apply and update when the search is paused or completed. Additionally, because the colouring is applied to the cluster, the colouring will update if a new cluster is found, and the pharmacophore search is paused or completed.

The hits matching the pharmacophore query can be selected in the **Results Hitlist** browser and visualised in the **3D view**, marked and/or sorted anytime during the pharmacophore search or when the search is paused (see Results Hitlist and Results Hitlist Browser).



The pharmacophore search can be stopped by pressing the **Stop** button; this will clear the **Results Hitlist** browser and restore the **Start** button. When the pharmacophore search is stopped or is complete the pharmacophore query can be edited again and a new search can be started.

In CSD-CrossMiner, by default, all the subsets included in the feature database are searched during the pharmacophore search however, it is possible to limit the search to specific subsets by using the tick-boxes associated to each subset of the feature database (e.g., csd543_crossminer).

The pharmacophore search starts with the conversion of the pharmacophore query, defined in Cartesian space by the tolerance spheres, into a distance space representation (fingerprint). For each pair of tolerance spheres, a distance constraint is derived by measuring the distance between the two sphere centres and adding, as well as subtracting, the sum of the sphere radii to obtain upper and lower bounds, respectively. Additionally, each distance constraint stores whether an intra- or intermolecular constraint has been defined for each pair of pharmacophore points, and whether a pharmacophore point is constrained to be part of a small molecule, a protein and/or any component.

The resulting pharmacophore query fingerprint can then be compared to the respective fingerprint of any database entry. Each bit set in the query fingerprint is required to be present in the database entry fingerprint. Therefore, for a database structure to be a hit against the query, it must contain all pharmacophore points of the pharmacophore query. It is only if the fingerprint comparison passes that a database entry is subjected to a 3D search using the pre-calculated feature points.

It is necessary to perform a final Cartesian space overlay check on full matches, since not all matches in distance space between a pharmacophore query and a database entry correspond to matches in Cartesian space.

Since for a single database entry multiple matches may have been identified, for each hit, CSD-CrossMiner calculates the minimum overlay root-mean-square deviation (rmsd) of the point coordinates in the match with respect to the pharmacophore sphere centres.

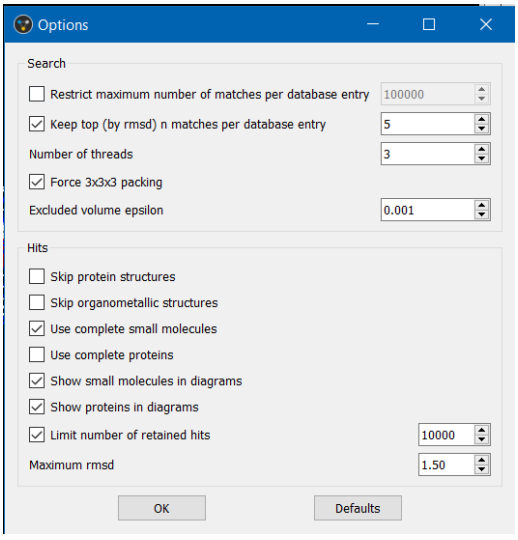
This is used to obtain a unique ranking of the matches. The number of matches per entry can be customised by the user (see Pharmacophore Search Options).

For all feature database entries that match the pharmacophore query, the molecular structures are loaded from the structure database, and matching hits overlaid onto the pharmacophore query are shown in the 3D view.

Pharmacophore Search Options

The pharmacophore search options can be accessed by clicking on **Edit** in the CSD-CrossMiner top-level menu and selecting **Options**. Note that these are not available to be changed (and **Options** is greyed-out) when the pharmacophore search is running or paused.

The **Search** section of the **Options** window contains modifiable settings for the pharmacophore search itself. The **Hits** section of this window contains modifiable settings for the processing of hits found by the pharmacophore search (i.e., clustering of hits and display of hits in the 2D diagram of the **Results Hitlist** browser and in the main 3D view of CSD-CrossMiner).

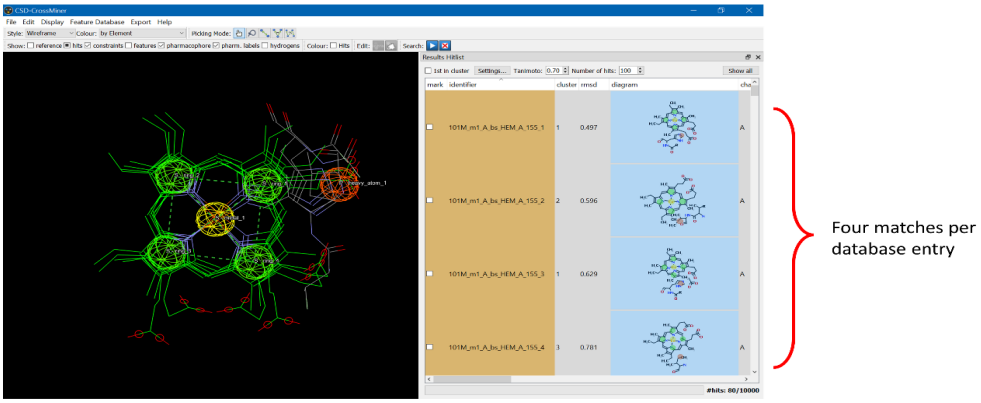


In the **Search** section of the **Options** window the user has access to **Restrict maximum number of matches per database entry** and **Keep top (by rmsd) n matches per database entry**, two options which allow the user control over the number of matches returned per database entry in a search.

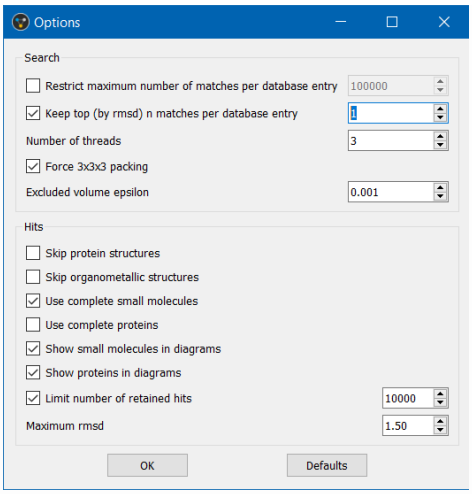
The first option (**Restrict maximum number of matches per database entry**) sets after how many hits per database entry the pharmacophore search is terminated (by default, no restriction is applied, and all possible hits are generated for each matching database entry).

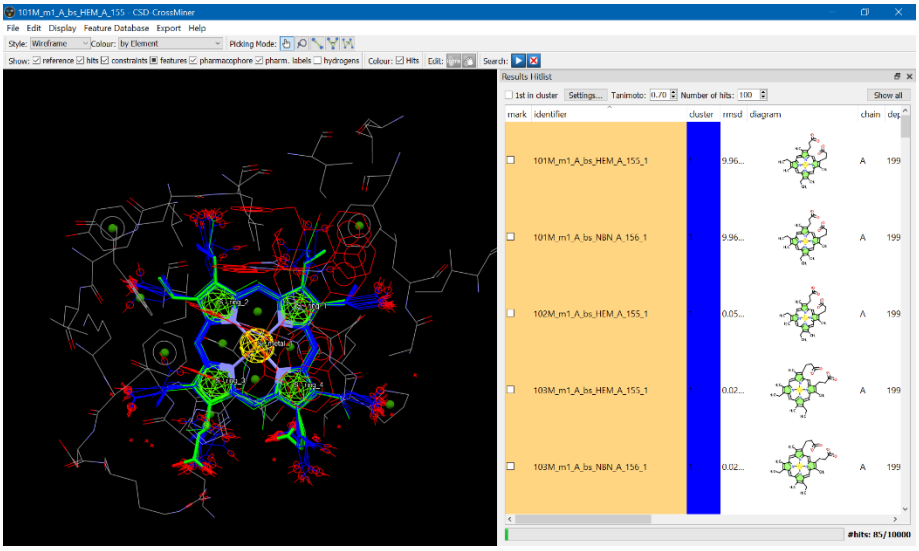
The second option (**Keep top (by rmsd) n matches per database entry**) sorts all hits per database entry according to rmsd and returns the top n.

By default, CSD-CrossMiner will return a maximum of five matches per database entry. Tweaking this option is useful when multiple matches in the same entry are returned. For example, in the case of symmetrical queries, such as a heme group with a pharmacophore query defined by four rings, a metal atom and one protein heavy atom, leaving the default value in **Keep top (by rmsd) n matches per database entry** would lead up to four matches per entry (one for each ring).



Reducing the number of matches per database entry to one for example would reduce the redundancy of the solutions and would provide only one match per database entry.



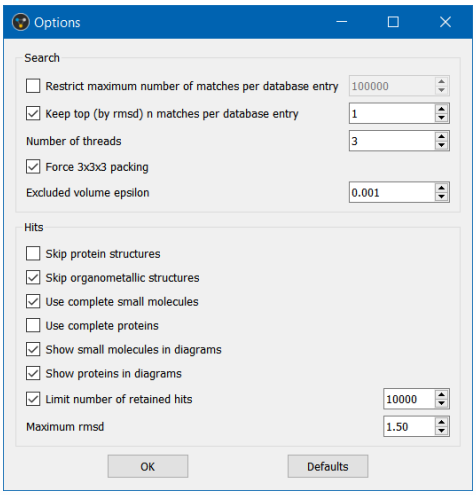


Via **Number of threads** it is possible to specify the computational resources to dedicate to the pharmacophore search, by defining that the database search be distributed to the specified number of CPU cores, if available.

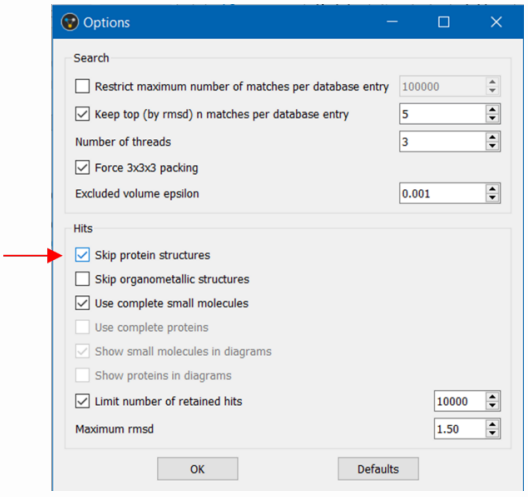
By default, the **Force 3x3x3 packing** tick-box is ticked, which restricts the search to 26 unit cells around the central unit cell. This allows symmetry-related copies of the feature points to be considered for a small molecule crystal structure database entry that matches.

The **Excluded volume epsilon** is the small value added to the excluded volume pharmacophore point radius. The default value is 0.001 however, this can be customised.

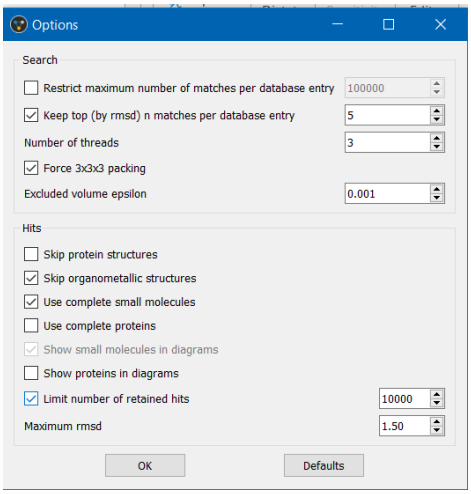
The **Hits** section of the **Options** window contains all settings for the processing and clustering of hits during a pharmacophore search.



Ticking the **Skip protein structures** option (unticked by default) results in the protein components of a matching database entry (although used for the pharmacophore search itself) being omitted in the hit clustering and not being displayed in the 2D diagram or in the 3D view. Only the small molecule components would thus be used for the hit clustering and would be displayed in the 2D diagram and 3D view.



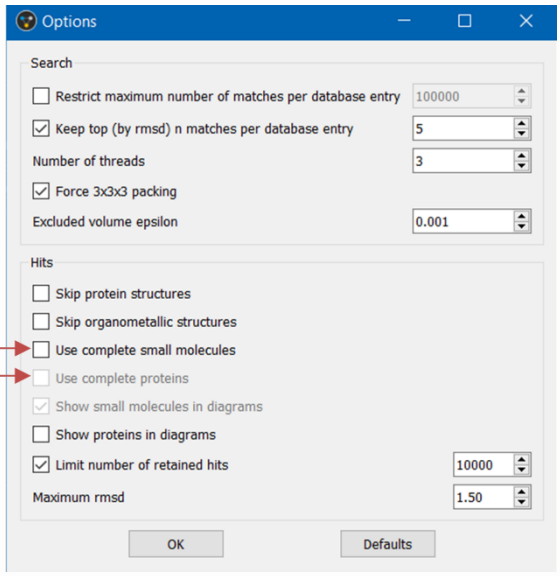
Similarly, ticking the **Skip organometallic structures** option (unticked by default) results in hits being omitted if they contain at least one transition metal, lanthanide, actinide, or any Al, Ga, In, Tl, Ge, Sn, Pb, Sb, Bi, Po.



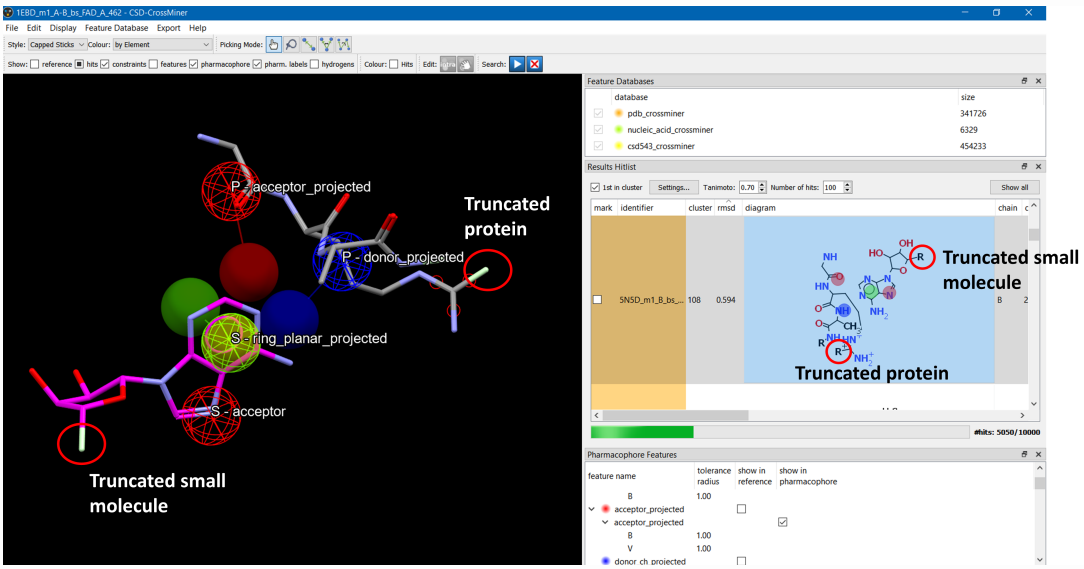
For a given pharmacophore search query, CSD-CrossMiner calculates a bounding sphere around the entire pharmacophore, which is an approximation of the overall smallest sphere that encompasses all pharmacophore points of the query and with a sphere radius at least equal to the largest sphere radius among the pharmacophore points plus 1.5 Å. When matched structures are overlaid onto the pharmacophore search query, there may be protein and/or small molecule atoms of the hit that are outside this bounding sphere. This is relevant for the **Use complete small molecules** and **Use complete proteins** options.

Ticking the **Use complete small molecules** option (ticked by default) results in the entire small molecule being used in clustering and displayed in 2D/3D, even if some small molecule parts are outside the pharmacophore bounding sphere. Similarly, the **Use complete proteins** option (unticked by default) can be ticked so that the entire protein binding site, as obtained from the structure database, is used for clustering and in the 3D display, even if some protein parts are outside the pharmacophore bounding sphere. Note that, for ease of visualisation, by ticking this option the protein component will not be shown in the 2D diagram.

Note that it is not possible to tick only **Use complete proteins** tick-box if the **Use complete small molecules** tick-box is unticked.



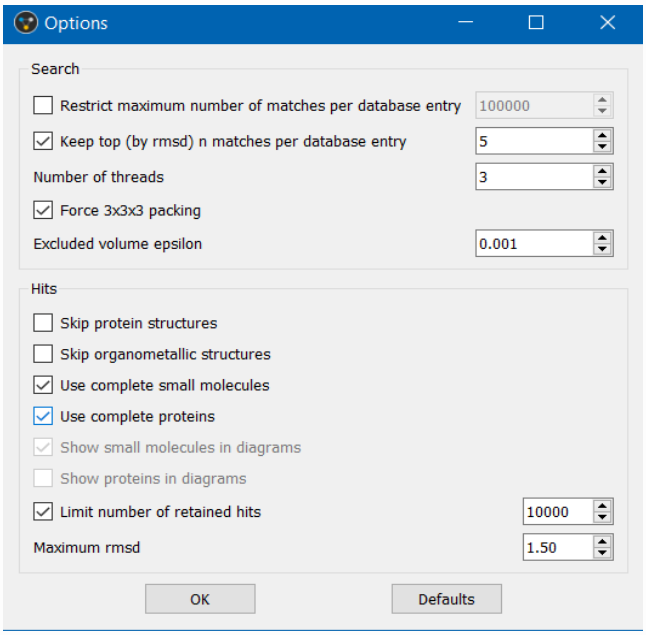
If the **Use complete small molecules** and/or **Use complete small proteins** tick-boxes are unticked, the part of the small molecule and/or protein component that is outside the pharmacophore bounding sphere will be truncated in both the 2D diagram and in the 3D view.



Ring systems (plus attached substituents) are always kept as intact units if at least one ring atom is inside the bounding sphere. For bonds which have one atom located inside and the other one located outside the bounding sphere, the latter atom will be replaced by an R-group symbol in the 2D diagram and light green R atom in the 3D view.

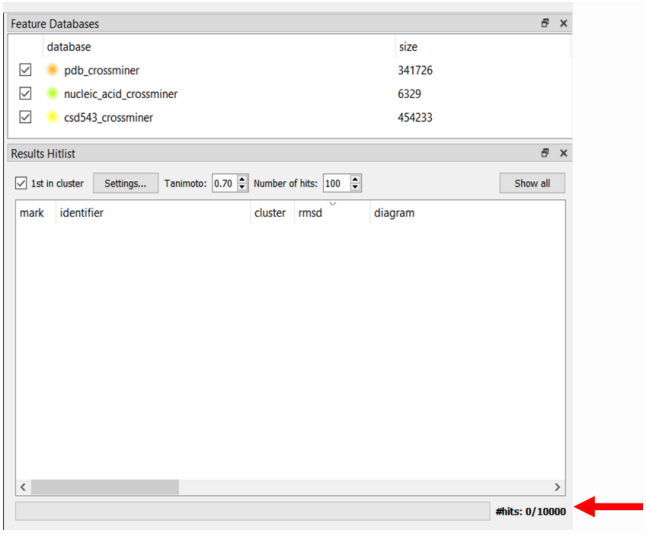
In addition, unticking the **Use complete small molecules** and/or **Use complete proteins** tick-boxes will affect the way CSD-CrossMiner performs clustering (see Clustering Algorithm and Clustering Settings).

The **Show small molecules in diagrams** and **Show proteins in diagrams** options are ticked by default, allowing the user to see both small molecules and proteins in the 2D diagram in the **Results Hitlist** browser. For easy visualisation, when the **Use complete proteins** tick-box is ticked, the **Show proteins in diagrams** option will automatically be disabled.

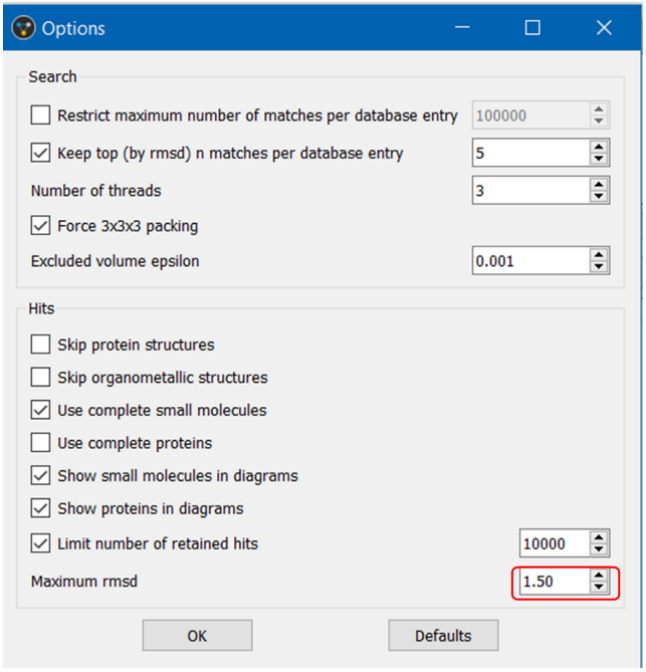


The **Show small molecules in diagrams** and **Show proteins in diagrams** options can be individually disabled by unticking the desired tick-box, such that the 2D diagram can contain only the small molecule component or only the protein component.

The **Limit number of retained hits** option defines how many hits will be retained in the 3D view and in the **Results Hitlist** browser. The default number of retained hits is 10 000 and is shown in the pharmacophore search progress bar. Changes in the number of retained hits will automatically update the upper limit in the pharmacophore search progress bar.



The rmsd upper limit for a pharmacophore match to be added to the hit list can be also edited by modifying the **Maximum rmsd** option. If tolerance spheres with large radii are used, this value may need to be modified accordingly.



By clicking **OK**, the new pharmacophore search option settings are saved and will be retained between separate CSD-CrossMiner sessions. However, it is possible to restore the default settings by clicking on the **Defaults** button in the **Options** window.

Clustering Algorithm and Clustering Settings

Although a pharmacophore search may provide the means for screening many compounds, this may be undesirable because resources may be wasted if this large-scale effort results in the production of redundant information. Clustering the solution space in real time provides a powerful help to remove such redundancy, thus allowing the user to quickly grasp the diversity in ligand topology and protein-ligand interactions found across the searched database.

The clustering algorithm in CSD-CrossMiner generates two similarity fingerprints: a small molecule fingerprint and a protein fingerprint. The small molecule fingerprint enumerates any small molecule substructure features that are present in the hit substructure. Similarly, the protein substructure features present in the hit are hashed in the protein fingerprint.

The hits matching the pharmacophore query are subject to two different clusterings: on-the-fly clustering (during the pharmacophore search itself) and post-search clustering (when the search is complete).

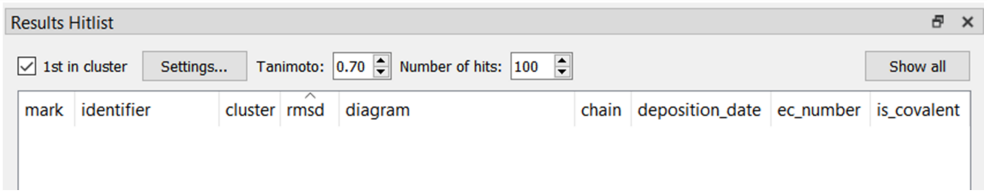
During the on-the-fly clustering, for each new hit the algorithm will loop through all current cluster representatives. If there are no cluster representatives within the user-defined Tanimoto threshold, the new hit will be added to the set of cluster representatives (i.e., this will create a new cluster, with cluster number n+1 if there are already n clusters); otherwise it will be discarded (i.e., it will not create a new cluster and not become a cluster representative; however, it will be saved as a hit). During the search, this clustering method is heuristic, as it is possible that a new hit may have a lower rmsd than its cluster member; this will not be accounted for by on-the-fly clustering.

Note that the 3D view is constantly updated during the search by adding any new hit that has become a cluster representative (i.e., any new hit that has been assigned a new cluster number in the

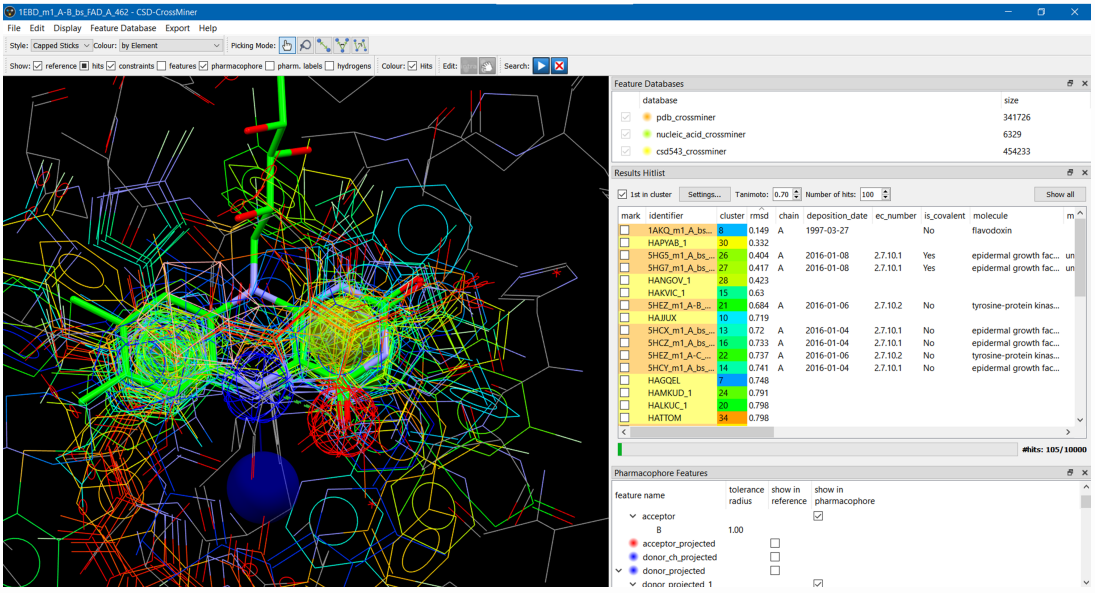
Results Hitlist browser). The result is that if selected hit changes its cluster representative status (i.e., discarded as a cluster representative), then all hits will be displayed in the 3D view.

The post-search clustering will instead sort all hits by rmsd once the search is complete. The algorithm will search in all hits matching the pharmacophore query for the cluster representatives, within the user-defined Tanimoto threshold. This time the cluster representative with the best rmsd will be selected.

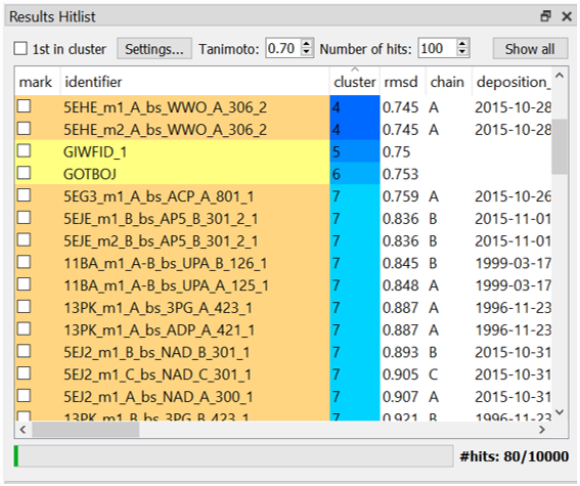
The clustering settings are displayed in the **Result Hitlist** window and can be edited at any time during a pharmacophore search.



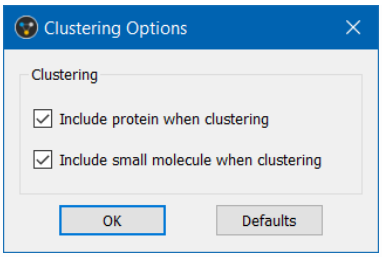
By default, the **1st in cluster** tick-box is ticked, such that it is only the cluster representative of each cluster that is shown in the **Results Hitlist** browser and in the 3D view. If the **Colour: Hits** tick-box is ticked, molecules of each cluster are coloured by rainbow in the 3D view and the colour associated to each cluster is displayed in the **cluster** column in the **Results Hitlist** browser. Each cluster will be represented with a different colour.



All cluster members can be visualised by unticking the **1st in cluster** tick-box any time during and after the pharmacophore search. This will increase the number of hits listed in the **Results Hitlist** browser and displayed in the 3D view and if **Colour: Hits** is selected (**Colour:** ☑**Hits**), all cluster members will have the same colour.

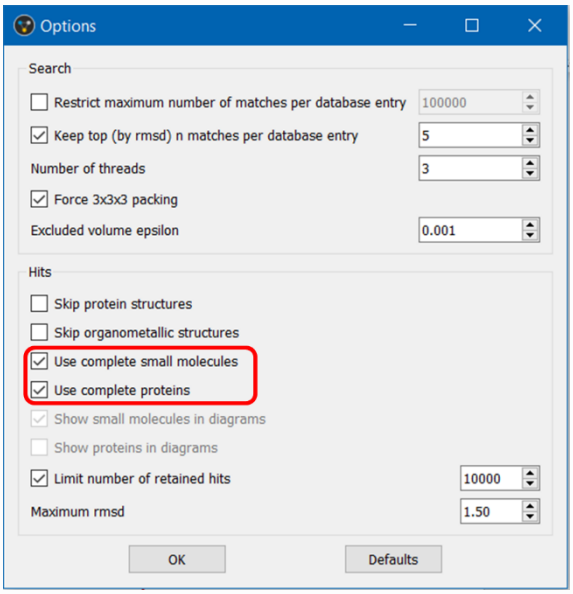


Note that unticking and then re-ticking this **1st in cluster** tick-box whilst the search is running, will activate the post-search clustering and the cluster representative with the best rmsd out of all hits found by this point of the search in progress will be selected. This may result in a different cluster representative being listed in the **Results Hitlist** browser and displayed in the 3D view. By default, two structures are deemed 'similar' if both the small molecule and protein fingerprints have a Tanimoto threshold within 0.7 Å. The user can tailor this threshold by entering a new value in the **Tanimoto** spin-box in the header of the **Results Hitlist** window (see Results Hitlist and Results Hitlist Browser). Please note that the default Tanimoto threshold of 0.7 Å will be restored between separate CSD-CrossMiner sessions. Additionally, the **Settings...** button in the header of the **Results Hitlist** window allows the user to configure the clustering settings by choosing whether to include the protein and/or the small molecule fingerprint in the **Cluster Options** pop-up window. Please note that any modification of these clustering settings will be retained between separate CSD-CrossMiner sessions.



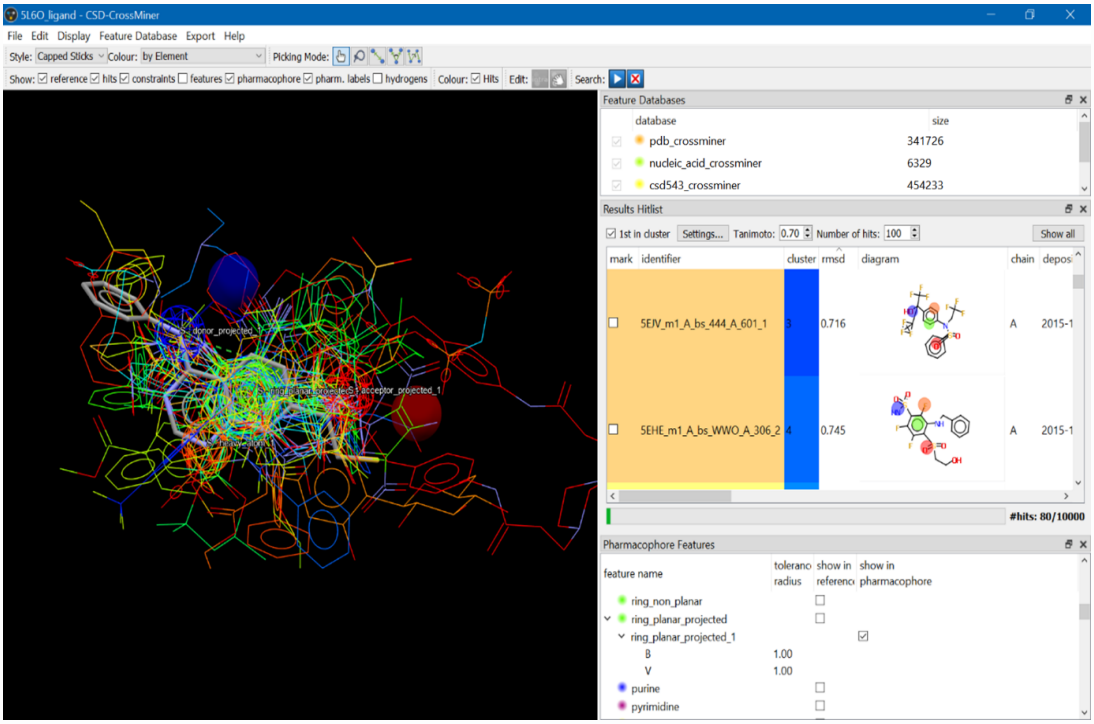
If the pharmacophore search options are set to the Defaults (see Pharmacophore Search Options), then the small molecule fingerprint and the protein fingerprint will be created by enumerating the pharmacophore features for all small molecule atoms and for those protein atoms within the bounding sphere of the pharmacophore, respectively. However, the user can change this clustering behaviour using the **Options** dialogue available through the CSD-CrossMiner top-level **Edit** menu.

By additionally ticking the **Use complete protein** tick-box, the clustering algorithm will include all atoms in the protein binding site into the protein fingerprint (see Pharmacophore Search Options), instead of only those within the pharmacophore bounding sphere if this tick-box remained unticked as per default.



Results Hitlist and Results Hitlist Browser

In CSD-CrossMiner, all hits matching the pharmacophore query (up to a user-defined rmsd limit, see Pharmacophore Search Options) are collected and (up to a user-defined number of hits) are displayed in the 3D view as well as listed in the **Results Hitlist** browser. The first five columns in the **Results Hitlist** window contain: the ability to mark an entry (by putting a tick in the column **mark**), the entry name of each database structure (e.g., `5EJV_m1_A_bs_444_A_601` in the example below) followed by the number of time the entry matches the pharmacophore query as underscore (**identifier**), (up to the top number of matches per database entry defined in the **Options**, see Pharmacophore Search Options) (e.g., `5EJV_m1_A_bs_444_A_601_1` in the example below), the number of the cluster (**cluster**), the rmsd overlay (**rmsd**), and the 2D diagram of the pharmacophore overlay match (**diagram)** with the pharmacophore matches indicated.



Further columns in the **Results Hitlist** browser contain any other information stored in the feature database as annotations (see Annotating a Feature Database).

The hits can be sorted according to any text column in the **Results Hitlist** browser at any point during or after the pharmacophore search. For example, hits can be sorted according to rmsd values or according to the number of the cluster by clicking on the **rmsd** or **cluster** column label in the **Results Hitlist** browser. A hit can be selected by clicking on it (multi-selection is enabled, with **Ctrl + left mouse button (LMB)** combination to select individual hits or with **Shift + LMB** combination to select a continuous list), which will automatically display the selected hit(s) in the 3D view. Note that when selecting hits during the pharmacophore search, it is possible that selected hits (even if retained as matched hit) will disappear from the 3D view and from the list in **Results Hitlist** browser. This can happen because **Results Hitlist** shows only a limited number of hits and it updates during the pharmacophore search.

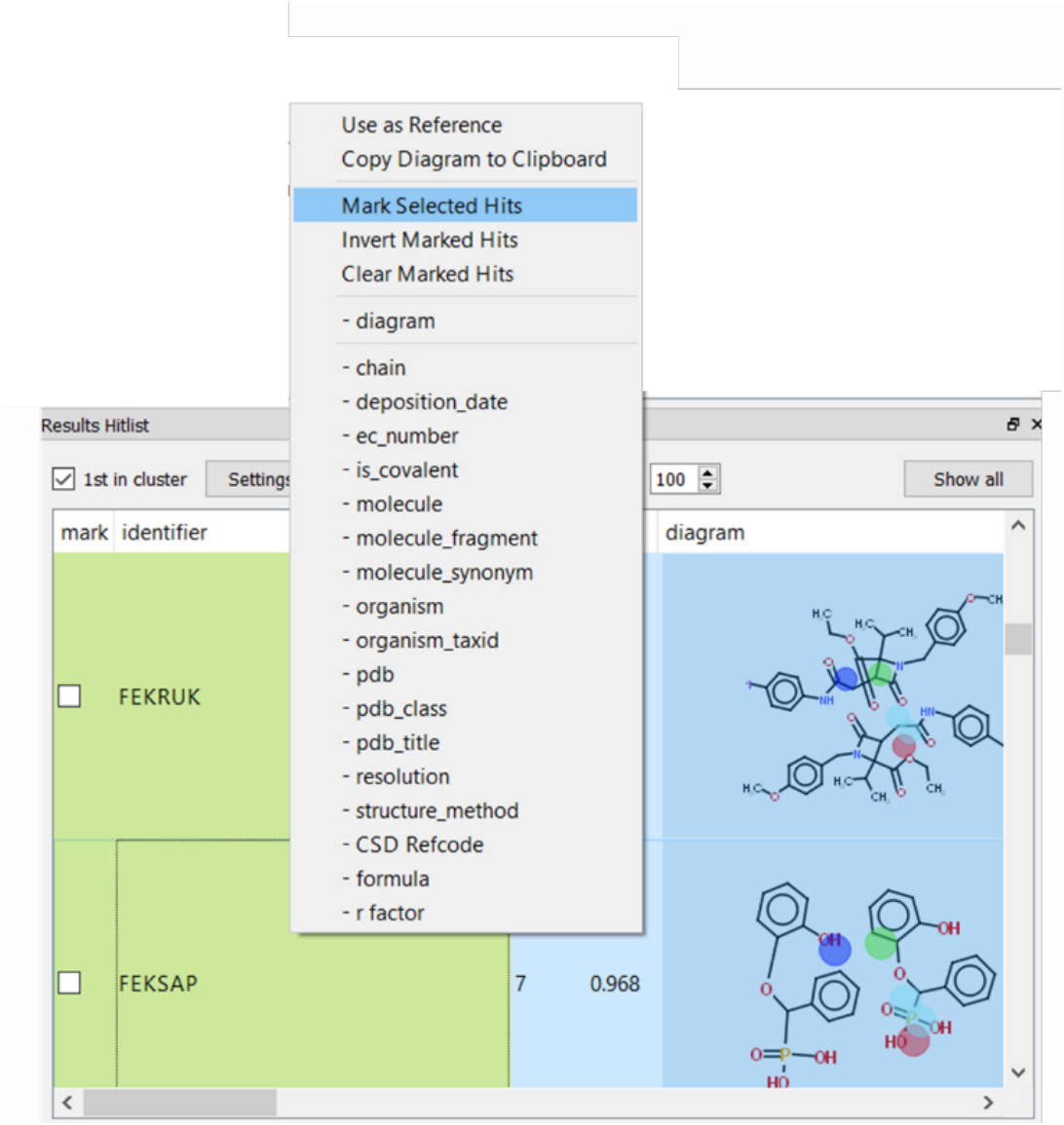
The default number of displayed hits is set to 100; however, this number can be changed (up to 1000) at any point before, during or after the search by using the **Number of hits** spin-box in the **Results Hitlist** browser. The edited **Number of hits** value will then be retained between separate CSD-CrossMiner sessions.

Note that whilst the maximum number of hits displayed in the **Results Hitlists** browser can be varied an overall list of all hits (up to the number of matches per database entry defined in **Edit** > **Options**) is maintained at all times.

Once a hit is selected in the **Results Hitlist** browser, the **up** and **down arrow** keys, **Page Up** and **Page Down** keys and **Home** and **End** keys can be used to browse through the list.

Interesting hits (i.e., clusters of particular interest) can be marked by ticking the respective tick-boxes in the **mark** column and saved for later inspection using **Save Marked Hits** in the CSD-CrossMiner top-level **File** Menu or can be further explored by other CSD-Discovery tools by using **Send marked hits to Hermes** in the CSD-CrossMiner top-level **Export** menu (see Exporting Hits).

Selected hits can be also marked by right-clicking in the **Results Hitlist** browser and select **Mark Selected Hits.** Note that, the new marked hits will be added to the previously marked hits, if present.



Through the right-click menu it is also possible to invert the marked hits using the **Invert Marked Hits** option and clear the marked hits by using the **Clear Marked Hits** option.

All hits listed in the **Results Hitlist** browser can be displayed in the 3D view by clicking the **Show all** button in the header of the **Results Hitlist** window.

All visible hits can be saved for later inspections using **Save Visibile Hits** in the CSD-CrossMiner top-level **File** Menu or can be exported to Hermes using the **Send visible hits to Hermes** in the CSD-CrossMiner top-level **Export** menu (see Exporting Hits).

When multiple hits are displayed in the 3D view (but not selected in the **Results Hitlist** browser), it is possible to select a specific hit by clicking on any of its atoms in the 3D view. This will hide all other hits from the 3D view.

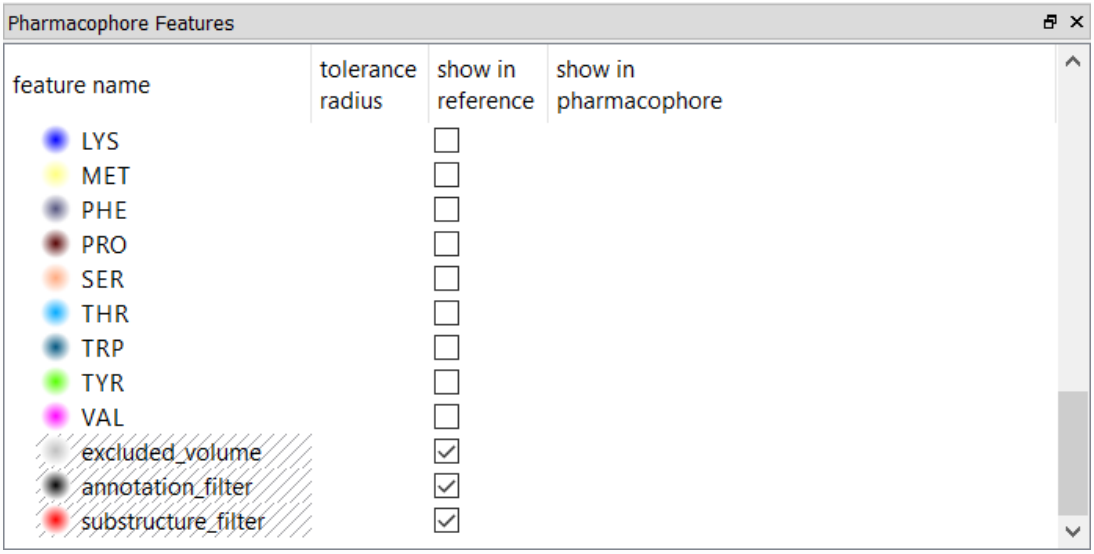
By default, the 2D diagram of any hit matching the pharmacophore query is shown in the **Results Hitlist** browser; however, it can be hidden by right-clicking in the **Results Hitlist** browser and selecting the **- diagram** option of the right-click menu. The size of the 2D diagram can be also changed by increasing or decreasing the **diagram** column while left-clicking on the **diagram** column delimiter in the **Results Hitlist** window.

In addition, by accessing the **Options** window from the CSD-CrossMiner top-level **Edit** menu (see Pharmacophore Search Options) it is possible to control what is contained in the 2D diagram: the small molecule, protein or both components by ticking/unticking the **Show small molecules in diagrams** and/or **Show proteins in diagrams** tick-boxes.

By default, all information (defined as annotations) stored in the loaded feature database are shown in the **Results Hitlist** browser. As for the 2D diagram, these annotations can be disabled individually by right-clicking in the **Results Hitlist** browser and selecting the desired annotation. Furthermore, the annotations can be used to filter the matched hits (see Using Annotations as Filter).

Filtering in CSD-CrossMiner

In CSD-CrossMiner is possible to filter the matching hits based on the presence/absence of specific substructure(s) and/or by annotations. The **substructure_filter** and **annotation_filter** features are available at the bottom of the **Pharmacophore Features** window. Because both **substructure_filter** and **annotation_filter** are not indexed in the feature database, they are represented with diagonal hatching in the **Pharmacophore Features** window.



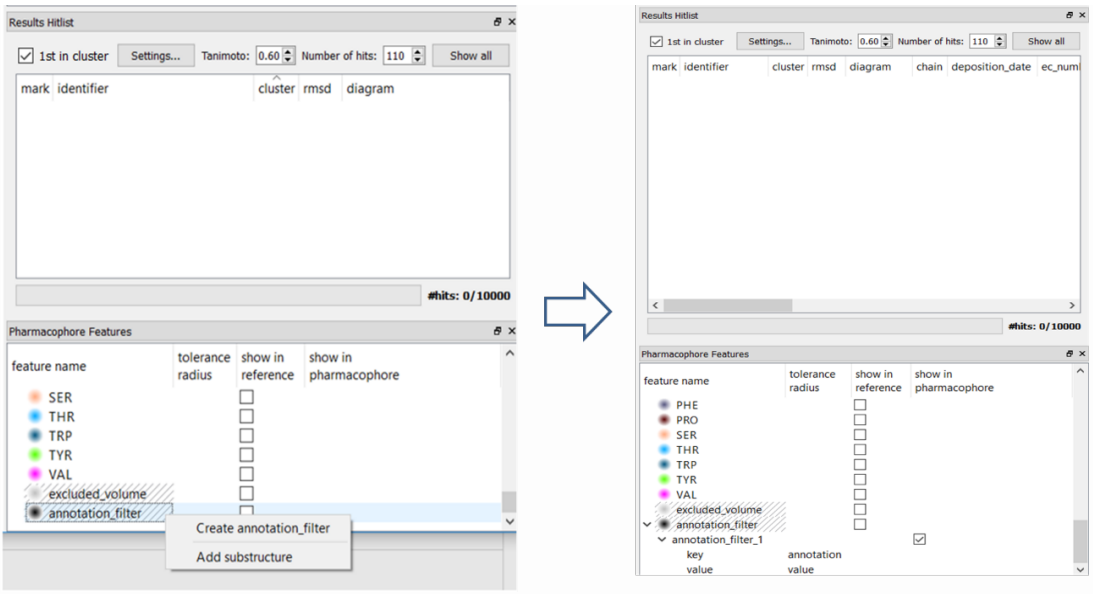
Note that the **substructure_filter** and **annotation_filter** can only be added before starting a pharmacophore search and/or after a pharmacophore search is stopped or completed.

Using Annotations as Filter

If the searched feature database contains annotations (as does the supplied feature database), these can be used to filter the database from which hits can be found using the **annotation_filter** listed in the **Pharmacophore Features** window.

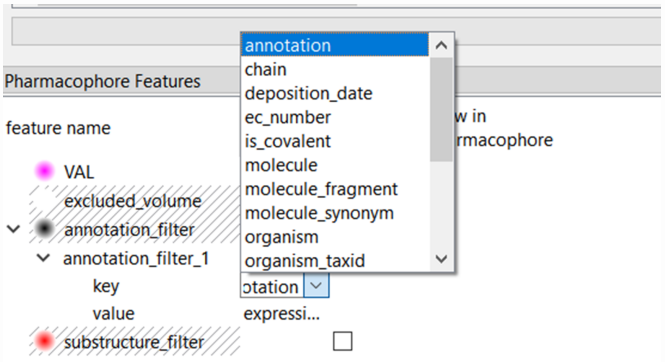
An **annotation_filter** is a specialised feature type that defines a textual filtering rule instead of pharmacophore feature spheres and thus is represented with diagonal hatching to differentiate it from indexed pharmacophore feature types.

An annotation filter can be created by right-clicking on the **annotation_filter** feature listed in the **Pharmacophore Features** window and then selecting on **Create annotation_filter**. Note that it is not possible to create and/or edit an annotation filter when the pharmacophore search has been paused, but only when it has not yet started or has been stopped.



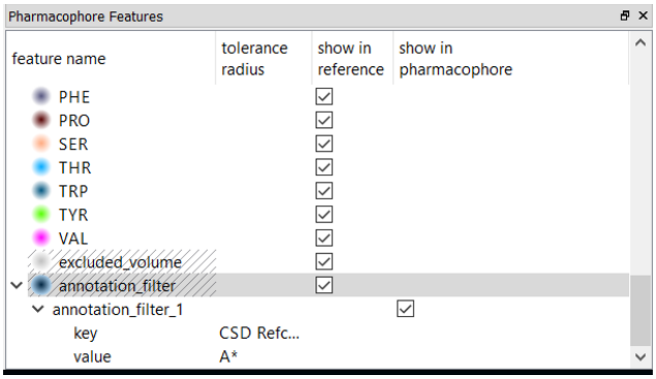
An **annotation_filter** is composed of two parts: a key and a value, where the key corresponds to one of the column headers listed in the **Results Hitlist** window and the value corresponds to the content in that specific column.

The list of all annotation keys for the loaded feature database is accessible by double clicking on the **annotation** key.

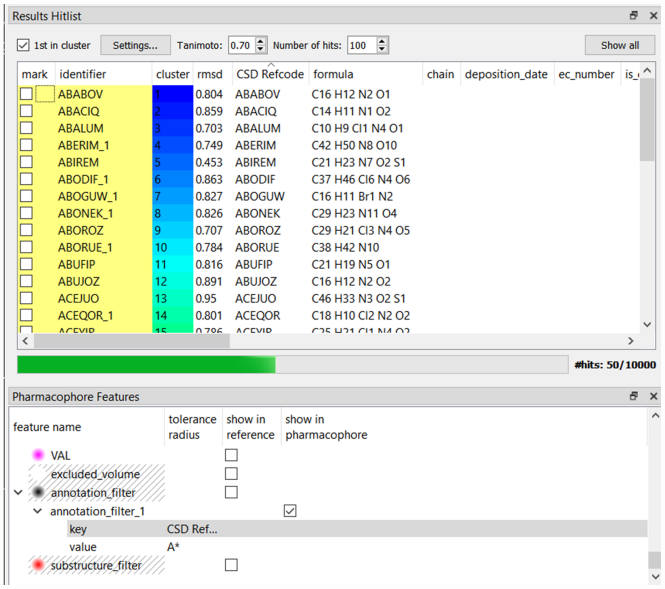


Note that this list will change if further or different annotations are added to the feature database (see Annotating a Feature Database).

A mismatch between a specified **annotation_filter** and the annotation associated with an entry will result in rejection of any putative hits from that entry.



In the above example, the key labelled **CSD Refcode** will be compared with the value rule **A***. This would result in showing only hits that belong to entries with a CSD REFCODE that begins with the letter A (matching A*) and all other hits being rejected:



If an annotation is numeric, e.g., resolution for a PDB entry, the annotation is filtered as numeric value. Numeric operators such as <, <=, >=, = can be additionally used. Numeric annotations can be filtered by ranges using the numeric operator - (e.g., 1.5-2.5).

Note that multiple annotation filters can be added in a pharmacophore search query.

Filtering Matching Rules

Filtering matching understands a few wildcard rules. These rules resemble the UNIX shell wildcards:

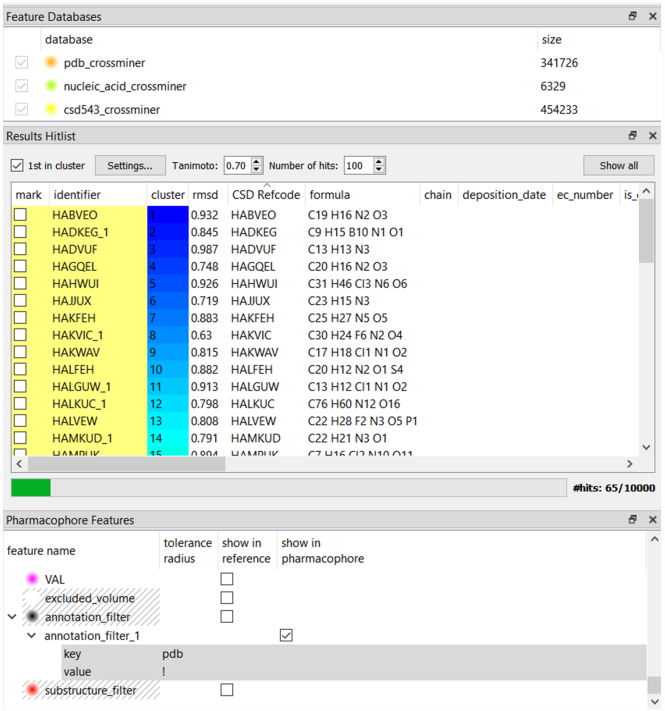
- **?** - Matches any one single character.

- ***** - Matches zero or more of any characters.
- **[...]** - Matches any one of the set of characters listed within the square brackets; e.g., “[0123456789]” will match a numeric character only.
- **** - Escape character to treat the special character that follows as a literal character; e.g., “?” will match a “?” character only.

Any other character represents itself apart from those described above; e.g., “a” matches the character “a”.

In addition to UNIX shell wildcards, an initial character of “!” will negate the comparison. For instance, “A*” will select every value beginning with “A” while “!A*” will select every entry not beginning with “A”.

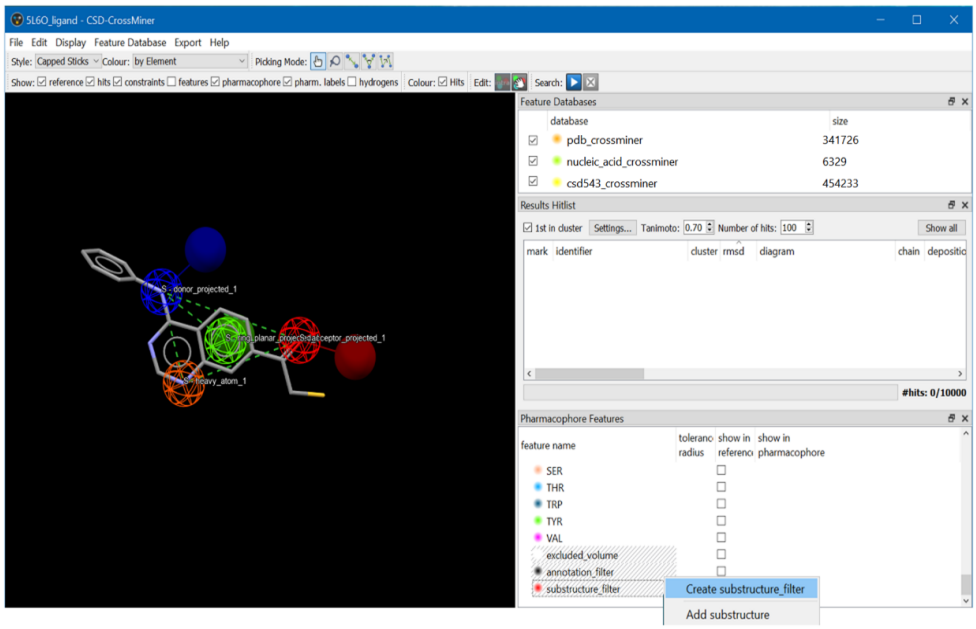
An empty value in the **value** text-box is treated as a check for existence of the annotation. This can also be negated, so the rule “!” will select every entry for which the specified annotation does not exist. In the example below, the key labelled **pdb** will be compared with the negation value rule **!**. This would result in showing only hits that do not have a PDB code.



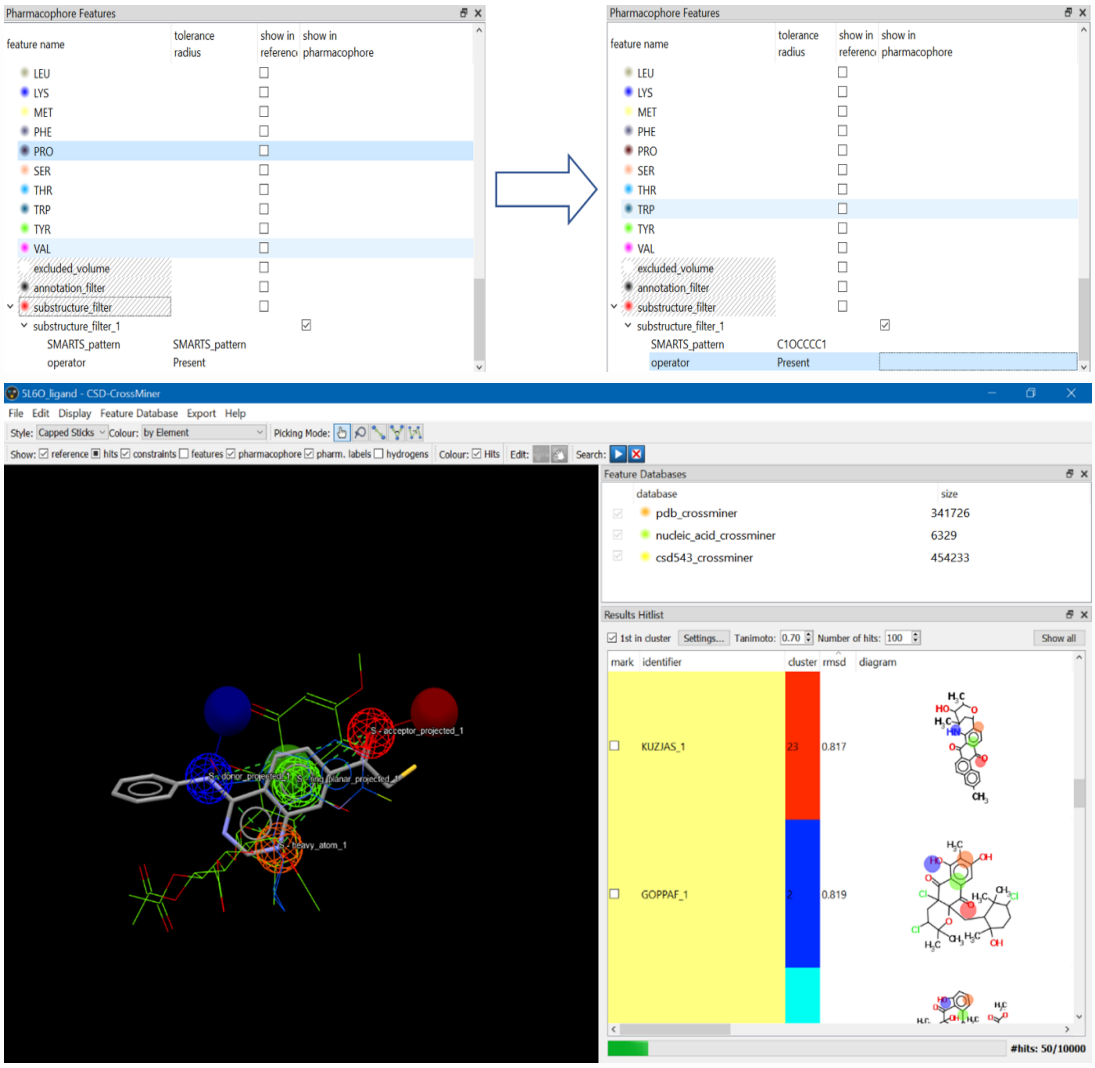
Substructure Filter

A **substructure_filter** is a specialised feature type that defines a SMARTS format filtering rule instead of pharmacophore feature spheres and thus is represented with diagonal hatching to differentiate it from indexed pharmacophore feature types. This filter can be used to constrain the pharmacophore search by the presence or absence of a specific substructure. The SMARTS format is simply a language for describing substructure patterns in molecules as a simple string of ASCII characters.

A substructure filter can be created by right-clicking on the **substructure_filter** feature listed in the **Pharmacophore Features** window and then selecting on **Create substructure_filter**. Note that it is not possible to create and/or edit a substructure filter when the pharmacophore search has been paused, but only when it has not yet started or has been stopped.



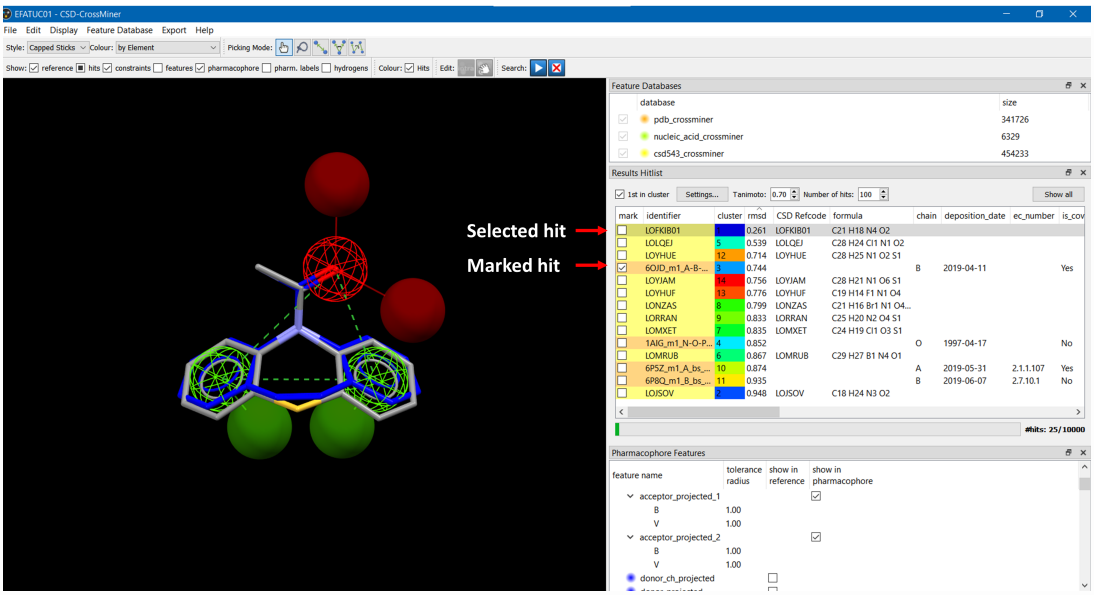
A **substructure_filter** is composed of two parts: a **SMARTS_pattern** and an **operator**, where the SMARTS pattern corresponds to the desired substructure (C1OCCCC1 in the example below) and the operator can be set to **Present** or **Not Present,** this will result in matched hits that contain or not contain the specified SMARTS pattern, respectively. In the example below, the set substructure filter will retrieve only hits matching the pharmacophore query that contain a C1OCCCC1 substructure.



By changing the operator to **Not Present**, the pharmacophore search will retrieve only hits matching the pharmacophore query that do not contain the C1OCCCC1 substructure. Note that multiple substructure filters can be added in a pharmacophore search query.

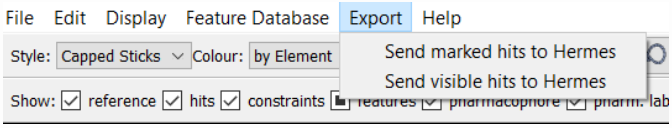
Exporting Hits

Interesting hits in the **Results Hitlist** browser can be marked (by ticking the **mark** tick-box for each hit of interest) and/or selected to be visualised in the 3D view (by clicking in the row for one hit of interest, multi selection is available by using **Ctrl + LMB** and **Shift + LMB** combination) see Results Hitlist and Results Hitlist Browser for details.



By default the max number of visible hits is 100 however, this can be changed by using **Number of hits** spin-box in the **Results Hitlist** window.

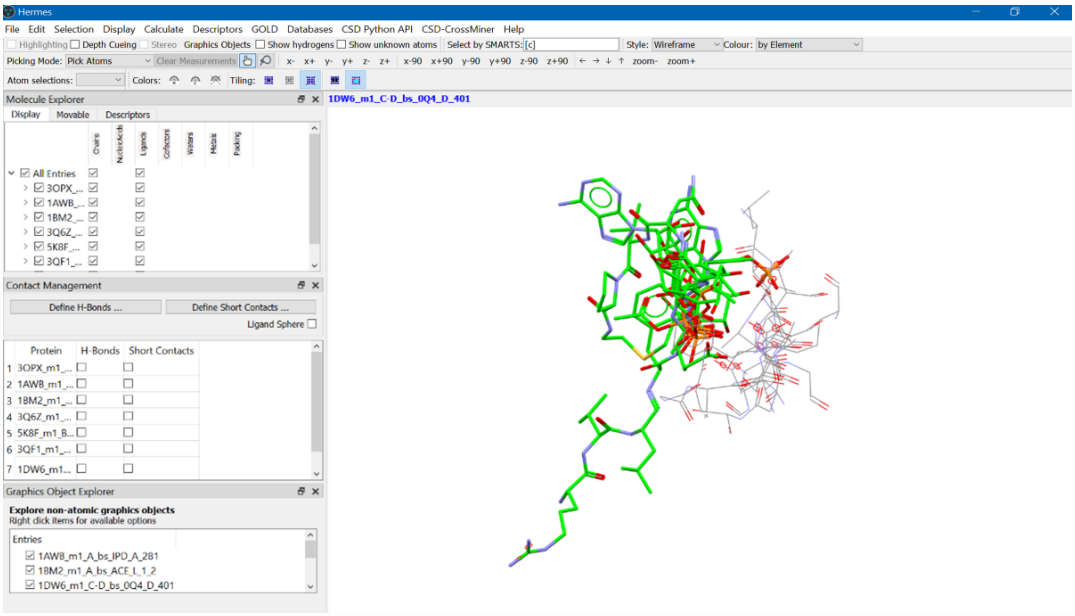
Marked and visible hits can be exported to Hermes, the CCDC protein visualiser and interface for several tools such as GOLD, Mogul, SuperStar etc., see https://www.ccdc.cam.ac.uk/solutions/csd-discovery/. Hits can be exported to Hermes using the **Export** menu in the CSD-CrossMiner top-level menu.



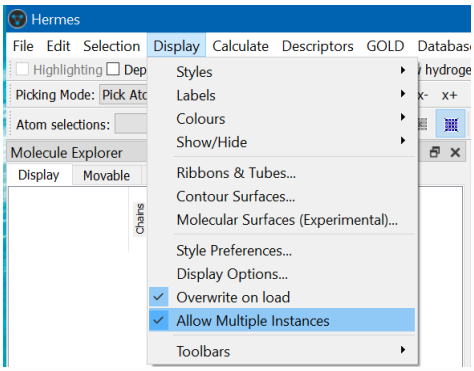
Note that the options in the **Export** menu are only available during the search or when the search is paused/completed.

When selecting **Send marked hits to Hermes** or **Send visible hits to Hermes**, a Hermes session will open with the exported hits displayed in the Hermes 3D visualiser and listed in the Hermes **Molecule Explorer window.**

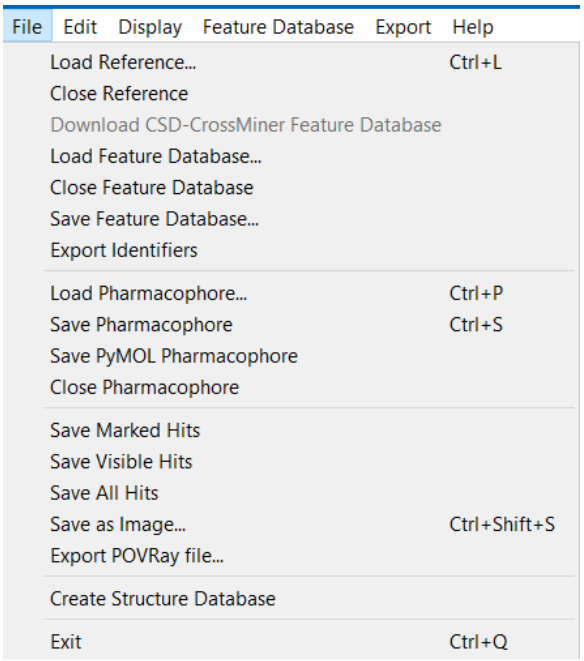
Note that we do not recommend sending more than 200 hits to Hermes as the application may become unresponsive.



Every time you send hits to Hermes a new Hermes session will open; therefore, you will have several instances of the Hermes visualiser open at the same time. You can switch off this option by un-ticking the **Allow Multiple Instances** option in the Hermes **Display** main menu. On doing so, when new hits are sent to an already opened Hermes session, these hits will append to the other structure loaded in that Hermes session.



In CSD-CrossMiner, 3D coordinates of marked and/or visible hits can be also saved to disk as mol2 or sdf files by selecting **Save Marked Hits** and/or **Save Visible Hits** from the CSD-CrossMiner **File** top-level menu and saving in mol2 or sdf file format. Otherwise, all hits can be saved by clicking on **File** and then selecting **Save All Hits.**



As for the marked and visible hits, the 3D coordinates of all the hits matching the pharmacophore query can be saved as mol2 or sdf files. The rmsd and cluster number are stored in the mol2 or sdf files (in mol2 under `@<TRIPOS>COMMENT` and in sdf under `> <cluster_number>`, `> <rmsd>`). Moreover, if the searched feature database has annotations (see Annotating a Feature Database), then all annotations will be stored as well in the mol2 or sdf files.

In addition to mol2 and sdf format, all the hits and/or the marked and/or the visible hits, can be saved in csv file format. This results in a table of the saved hits that include the SMILES of the small molecule matching the pharmacophore query in addition to the rmsd, cluster number and other annotations stored with the searched feature database.

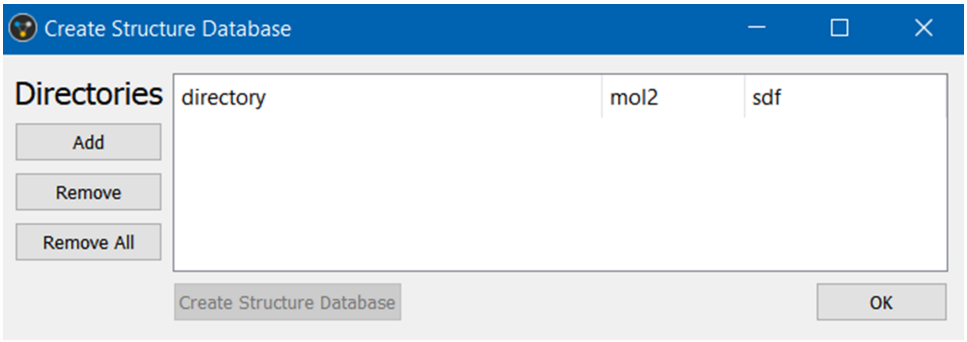
Note: Hits cannot be saved during a pharmacophore search, only when the search is paused or complete.

Creating Databases

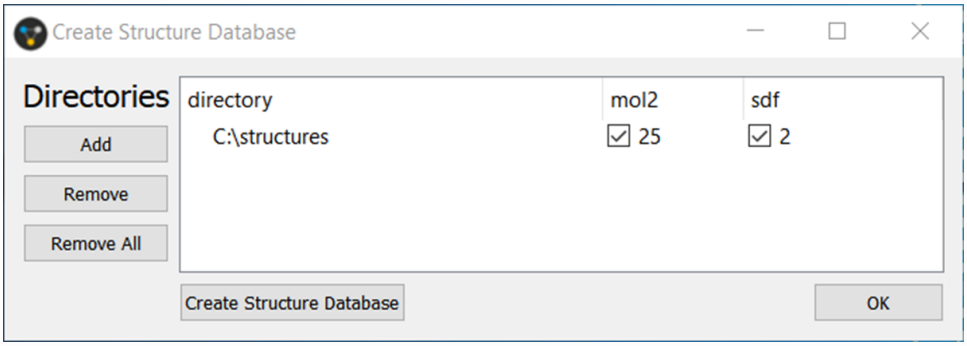
In CSD-CrossMiner there are two different types of databases: structure database and feature database (see Databases in CSD-CrossMiner). Both databases can be created through the CSD-CrossMiner interface in the manner described below.

Creating a Structure Database

A new structure database can be created via the CSD-CrossMiner top-level menu **File** > **Create Structure Database**.



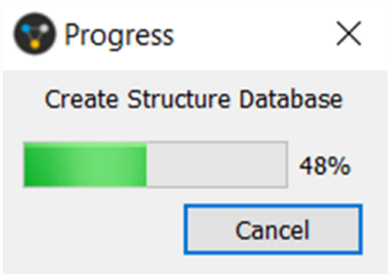
Single or multiple directories containing the files of the structures (in mol2 and/or sdf format) can be included in the new structure database by clicking on **Add** in the **Create Structure Database** pop-up window. The respective number of files with both file types (mol2 and sdf) in the directories will be displayed in the respective mol2 and sdf column with a tick-box. Any of these tick-boxes can be unticked to ignore those structures.



CSD-CrossMiner uses the residue information stored in mol2 files to distinguish between protein and small molecule components (note that nucleic acids are treated as small molecule component). Thus, mol2 is the preferred format if protein-ligand binding sites are to be added into a new structure database. For the associated collaborators, a python script in CSD Python API is available to convert protein-ligand and protein-ligand nucleic acids structures in PDB format into protein-ligand binding sites in mol2 (see APPENDIX F: Example Scripts Available for Associated Collaborators).

Once the directories have been added, clicking on the **Create Structure Database** button will prompt the user to specify an output filename and will start the creation process of the structure database in CSD SQL FastBinary database file format (`.csdsqlx`).

A **Progress** pop-up window indicates the progress in creating the structure database and gives the ability to cancel the process.

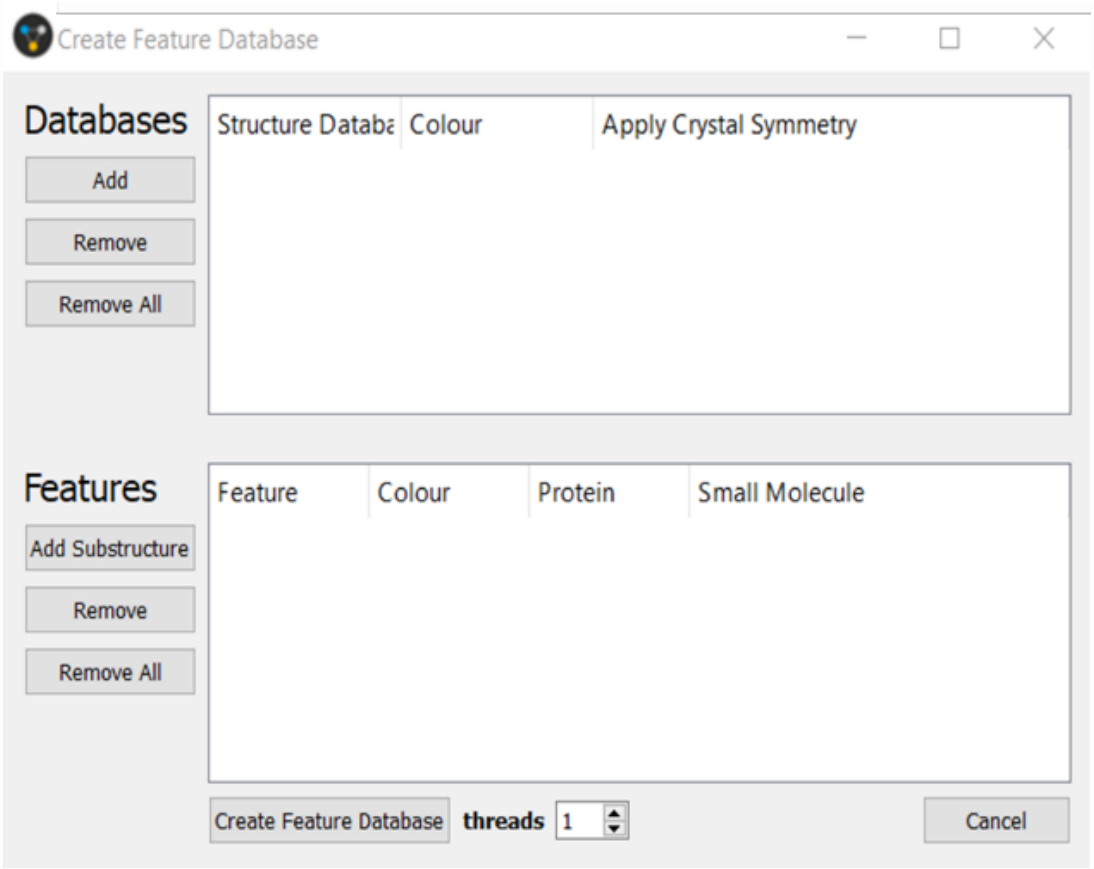


The creation of the database will be complete when the **Progress** pop-up window disappears. Clicking **OK** in the **Create Structure Database** window will then close the window.

Note that, it is important that the structure database does not contain duplicate identifiers. All entries in the structure databases need to have an identifier that is unique in that database.

Creating a Feature Database

A new feature database can be created by choosing **Feature Database** from the CSD-CrossMiner top-level menu and then **Create**. In the **Create Feature Database** dialog, structure databases (**Databases)** and feature definitions (**Features**) can be specified to create a new feature database.



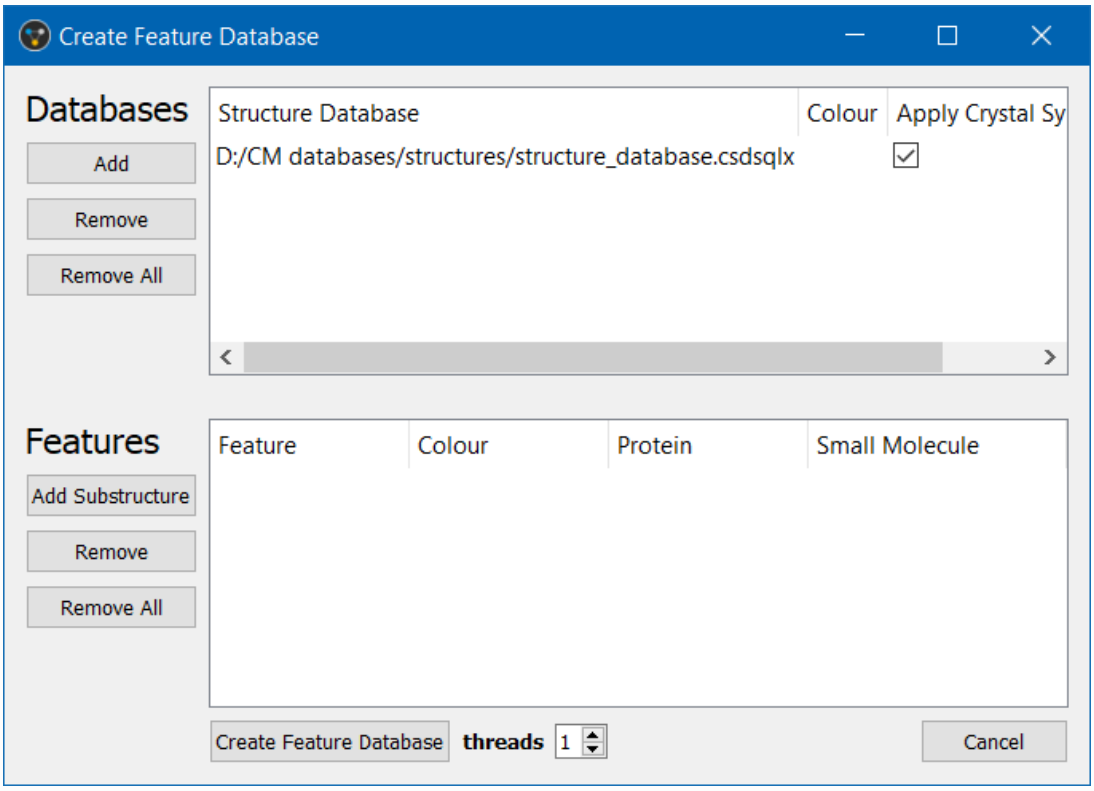
The following structure database input file formats are currently supported:

Format	File extension(s)	Comment
ASER	`*.ind`, `*.dbl`	Recommended for small molecule crystal structures (CSD-format databases).
CSD SQL FastBinary	`*.csdsqlx`	Recommended for protein-ligand and for protein-ligand-nucleic acids binding sites (can be created from mol2 and/or sdf files via CSD-CrossMiner, see Creating a Structure Database).
MOL2 SQLite	`*.sqlmol2`	It is supported but it is not the recommended file format protein-ligand and protein-ligand-nucleic acids binding site. Please use `*.csdsqlx` instead.
	`*.mol2`	

Format	File extension(s)	Comment
Tripos MOL2		It is not recommended to use **large** multi-mol2 files; please convert to csdsqlx first in these situations.
SDF	`*.sdf`	It is not recommended to use **large** multi-sdf files; please convert to csdsqlx first in these situations.

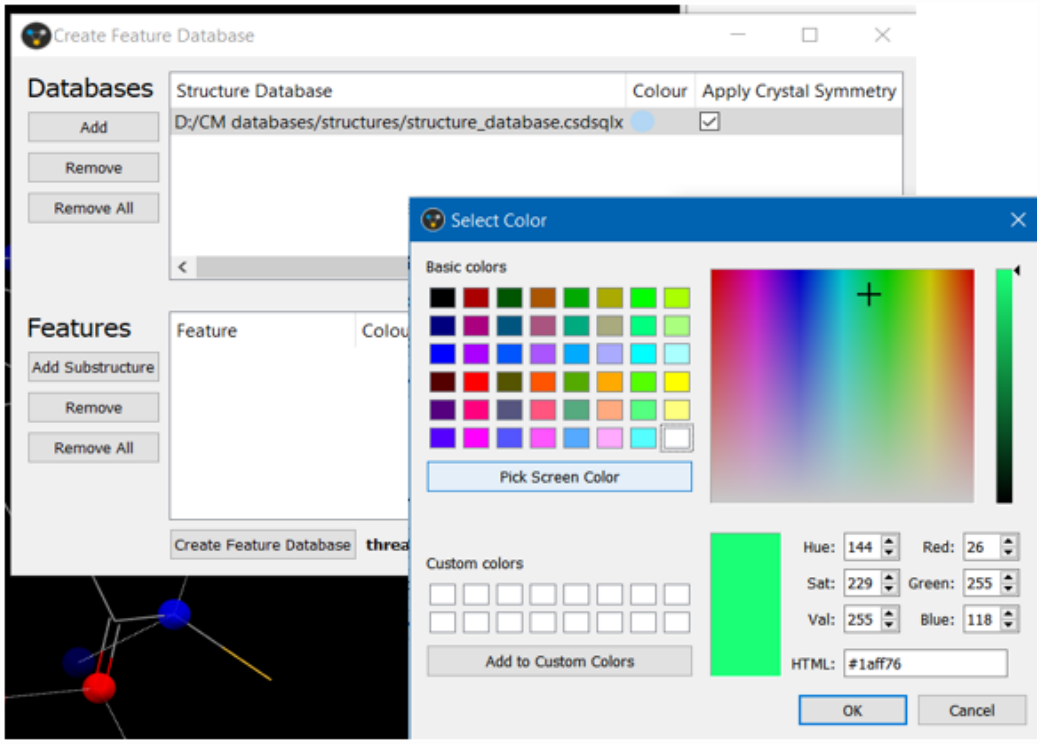
When creating a feature database, feature definitions are applied to the structure database(s); therefore, a feature database can contain different structure databases (e.g., CSD and PDB structures as well as in-house small molecule and/or in-house protein-ligand binding site structures). All structure databases must thus be converted simultaneously into a single feature database in order to ensure homogeneity in the feature definitions (APPENDIX B. Feature Definitions in CSD-CrossMiner).

Structure database(s) containing the molecular structures for which the features will be created can be added via the **Add** button in the **Databases** section of the **Create Feature Database** dialog. For each database two settings can be specified: **Colour** and **Apply Crystal Symmetry**.



The user can control these options individually for each database by highlighting each database one at a time and setting these options, which correspond to:

- **Colour:** The default colour is white; however, it can be customised by clicking on the sphere adjacent to the selected structure database, located in the **Colour** column and selecting the desired colour. When the created feature database is loaded in CSD-CrossMiner, this colour will be shown in the **Feature Databases** window, and it will be used for the hit colouring in the **Results Hitlist** browser when a search is performed against this feature database. This is useful if multiple structure databases have been used to generate a single feature database.



- **Apply Crystal Symmetry:** This tick-box controls the symmetry generation behaviour; when it is ticked, CSD-CrossMiner will create symmetry-related copies based on the spacegroup information supplied for each structure. This option, therefore, should only be ticked for small molecule crystal structures (not all small molecule structures) and not for protein-ligand binding sites.

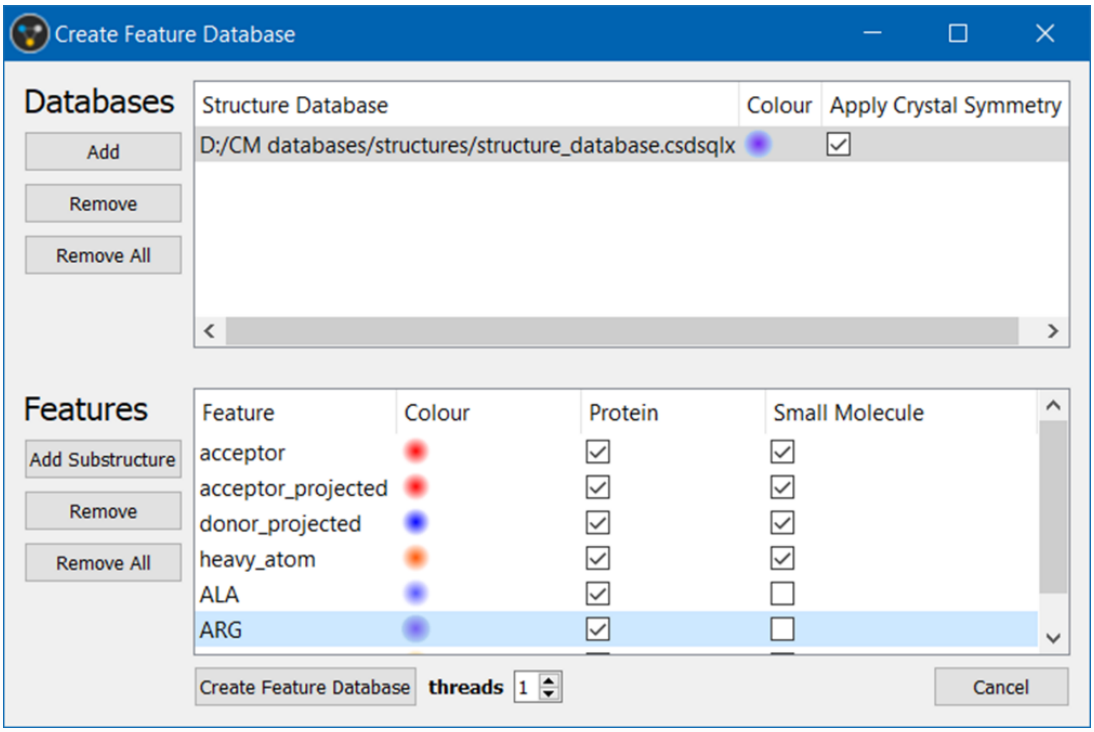
Once the structure databases have been specified, the features to be assigned to all structures contained in the database need to be loaded in the **Features** section of the **Create Feature Database** dialog.

The feature definitions used to create the feature database are derived by applying point generation rules to sets of atomic coordinates extracted from molecular structures by substructure definitions (see Editing and Creating Feature Definitions). The substructure-based features used to generate the supplied feature database are available for the user to use in the `feature_definitions` folder of the `CSD-CrossMiner` directory (APPENDIX B. Feature Definitions in CSD-CrossMiner).

Note that excluded volume features are not assigned in the feature database (see Adding an Excluded Volume to a Pharmacophore Query).

These feature definition files (and any additional/updated feature created by the user) can be selected by clicking on the **Add Substructure** button in the **Features** section of the **Create Feature Database** dialog. Multiple feature definition files can be selected at the same time (using **Shift + LMB** combination to select a continuous list, or **Ctrl + LMB** combination to select individual files). Pressing **Open** results in the selected features appearing in the **Features** list in the **Create Feature Database** window, along with their name and colour.

Since not all features may need to be created for protein and small molecule components, the feature creation should be turned **on** or **off** for individual components using the **Protein** and **Small Molecule** tick-boxes for each feature type that is loaded in the **Features** list.



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.

The creation process can be started by clicking the **Create Feature Database** button and specifying a filename for the new feature database, which will be saved in the `.feat` file format.

The general workflow on how to create a feature database, including how to generate the input files is shown in APPENDIX D. Create a Feature Database with In-House Data. Additionally, it is possible to fully automate the creation of the structure and feature by using the CSD Python API (see APPENDIX E: Pharmacophore search through the CSD Python API and CSD Python API Documentation).

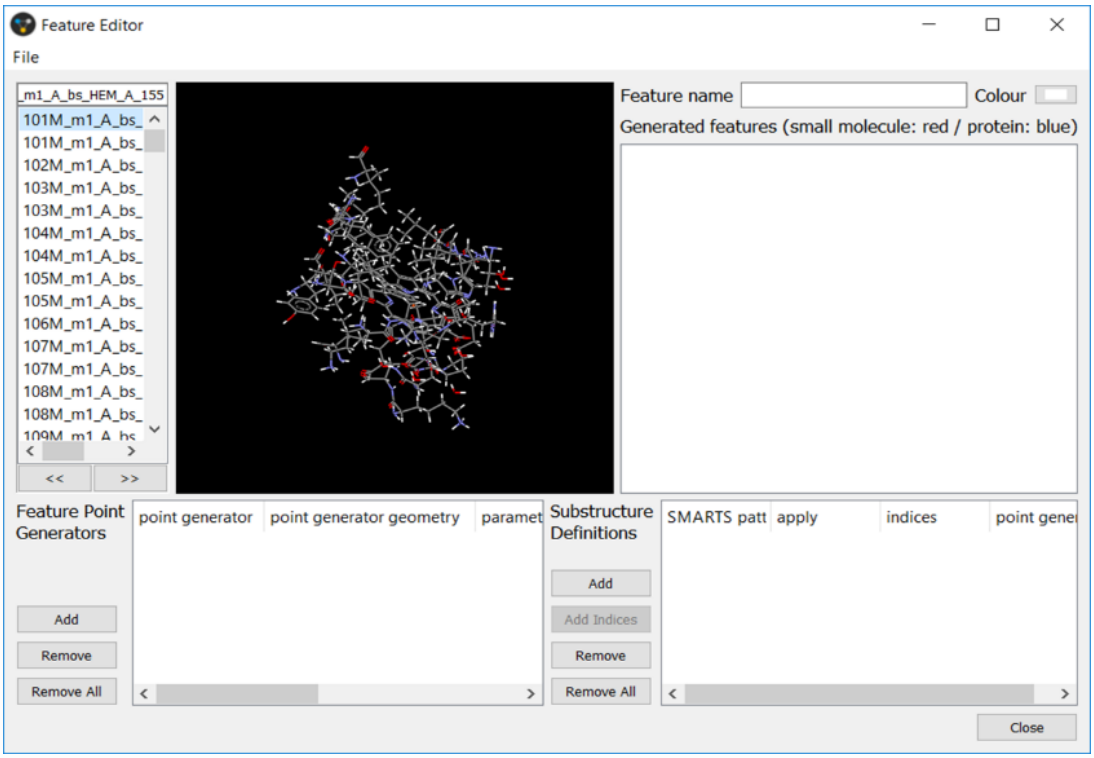
Editing and Creating Feature Definitions

Feature definitions can be created and edited using the **Feature Editor** accessible by clicking on **Feature Database** in the CSD-CrossMiner top-level menu and then selecting **Edit Features**.

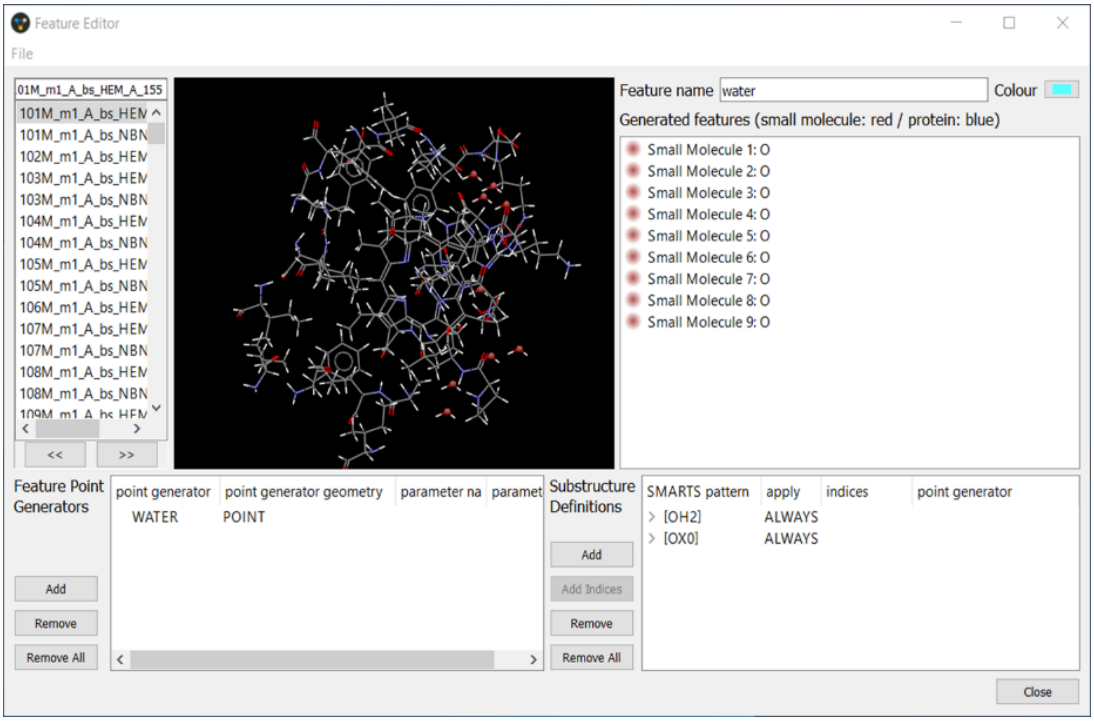
In general, the **Feature Editor** allows the generation of new features by applying point generators to sets of atomic coordinates extracted from molecular structures by substructure definitions.

It is recommended to first load a structure database by clicking on **File** in the **Feature Editor** window and then **Load Structure Database** (for example, choosing the structure database downloaded supplied in the `crossminer_data` folder in the `CSD_2022` directory).

Once a structure database has been loaded, the contained structure(s) will be displayed in the 3D display and listed in the upper left panel of the **Feature Editor** window. If the database contains multiple structures, these can be browsed by using the slider, by clicking on the **<<** and **>>** buttons at the bottom of the list, by entering an identifier in the text box, or by using the **up** and **down arrow** keys, **Page Up** and **Page Down** keys and **Home** and **End** keys.



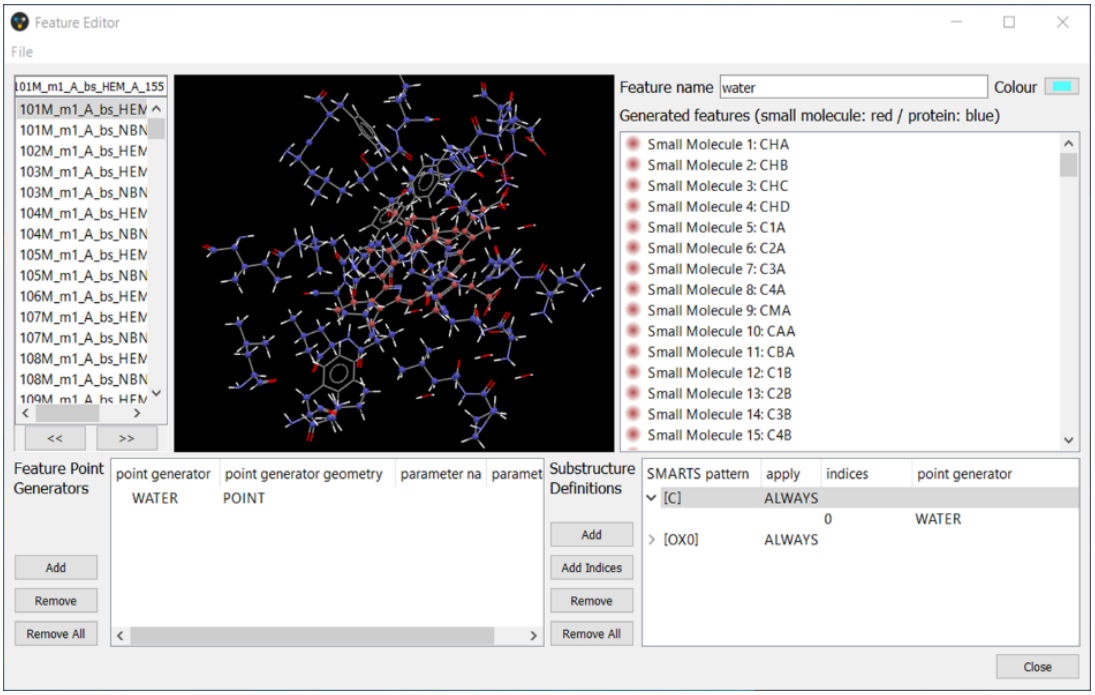
Existing feature definitions can be loaded by clicking on **File** in the **Feature Editor** window and then choosing **Load Feature Definition**. Note that only one feature definition can be loaded. In the example below, the water feature definition used to create the provided feature database is loaded. You can find the water feature together with all the other substructure-based features used to generate the supplied feature database in the `feature_definitions` folder of the `CSD-CrossMiner` directory.



The **Feature Point Generators** section includes the point generators associated with the loaded feature, while the **Substructure Definitions** section stores the substructure definitions corresponding to the loaded feature definition.

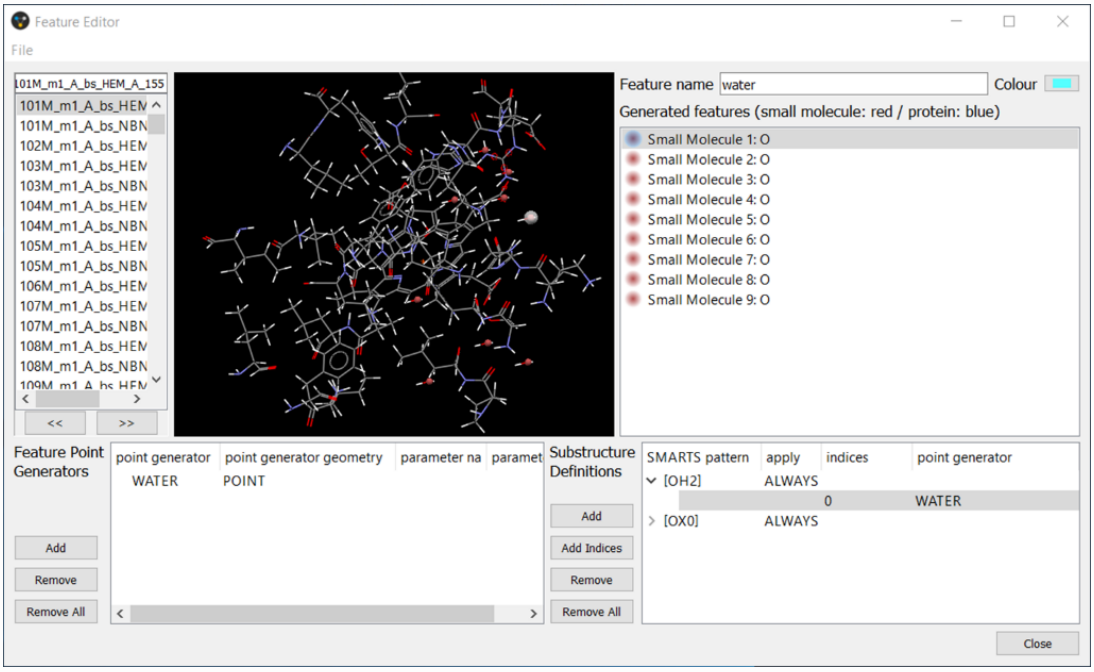
The loaded feature definition is displayed in the 3D display of the **Feature Editor** window as red (Small Molecule) and blue (Protein) feature points and listed in the right-hand panel under **Generated features (small molecule: red / protein: blue)** of the **Feature Editor** window. Note that this list is associated with the structure displayed in the 3D display. Clicking on any of the features in the list will highlight the respective feature in the 3D display and will select the SMARTS substructure definition that has been used to create this particular feature point (note that it can be multiple points for more complex feature definitions).

The **Feature Editor** will respond with immediate feedback on any changes to the feature definitions; e.g., selecting the **[OH2]** SMARTS under **SMARTS pattern** in the **Substructure Definitions** list, changing it to **[C]** in the text box and pressing the enter key will immediately update the newly matched feature points (in this case, only aliphatic carbons) in the 3D display and the **Generated features** list.

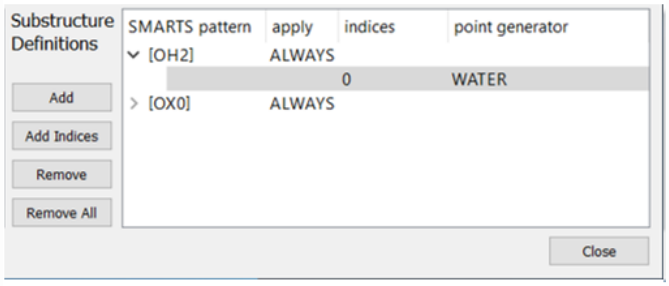


It is also possible to add a new SMARTS pattern to the list of those provided for the loaded feature or to remove one from the list by clicking on the **Add** or **Remove** button, respectively, in the **Substructure Definitions** panel.

Below the SMARTS pattern definition, there is another row in which the **indices** column is populated with a **0** index. In the example below, the **0** index selects the first atom in the SMARTS substructure it is assigned to (which is **[OH2]**); thus, it selects the oxygen atom and applies the **WATER** point generator, where the **WATER** point generator is defined in the **Feature Point Generators** panel. This point generator is simply used to create a feature point at the respective atom position.



It is also possible to add new indices in a selected SMARTS pattern by clicking on **Add Indices** in the **Substructure Definitions** panel.



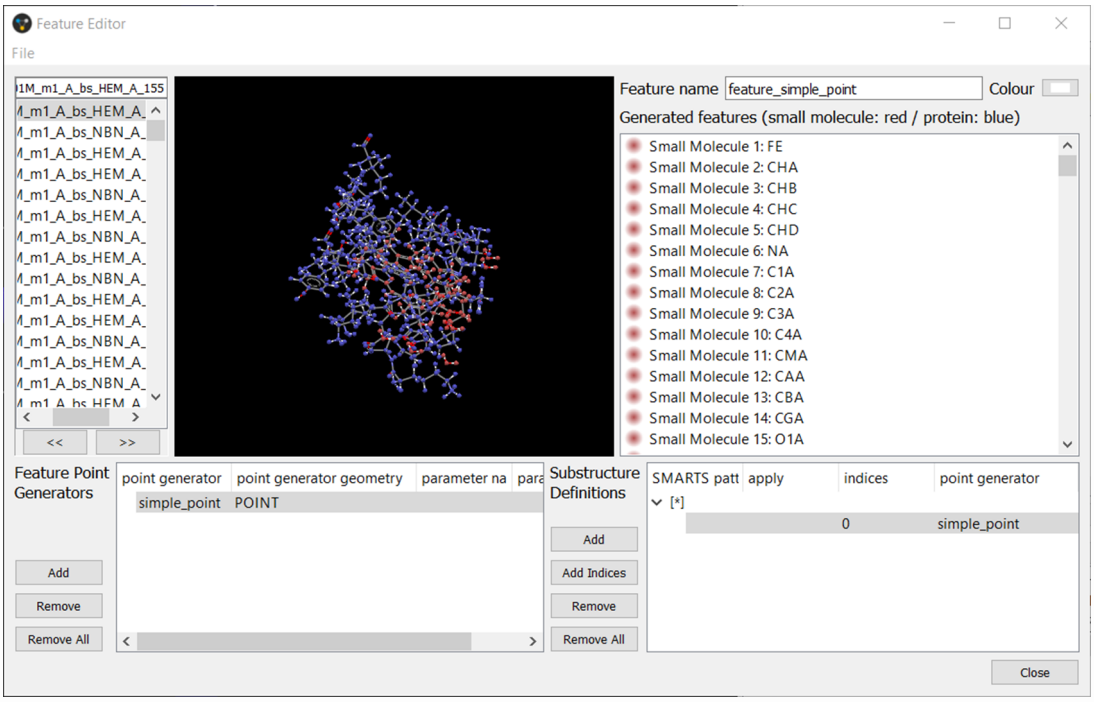
Multiple **SMARTS pattern** (along with their respective indices and point generator definitions) can be arranged in a hierarchy in which the substructures are matched in the specified order (substructure definitions can be dragged within the list box to change their priority). The substructure with the highest priority will be matched first and all selected indices will be used to mark the respective atoms of the structure as used. If a second substructure in the hierarchy with a lower priority is matched and all atoms in the structure selected by the indices have already been marked as used, then no feature points will be generated in this case (APPENDIX C. SMARTS Implementation and SMARTS Description).

The name and the colour associated with the loaded feature is specified in upper right corner of the **Feature Editor** window. These can be edited by typing in the **Feature name** text box the new name associated with the feature and by clicking on the **Colour** box to select the desired colour.

The edited feature definition can be saved to disk in `.cpf` file format by clicking on **File** and then **Save Feature Definition** from the **Feature Editor** window and specifying a filename. This will make the new feature definition available to be used to create a new feature database (see Creating a Feature Database). It is possible to clear the **Feature Editor** window by clicking on **File** and then **Clear Feature Definition**.

A feature definition can also be created from scratch, by clicking on **Add** in the **Feature Point Generators** panel of the **Feature Editor** window.

By default, a point generator of type POINT is created, named simple_point, and then a SMARTS substructure definition can be created by clicking **Add** in the **Substructure Definitions** panel. By default, a SMARTS pattern of **[*]** that selects every atom is created.



Both the point generator geometry type and SMARTS pattern can be changed; the list of all feature point generators available is provided in the following table. Details of SMARTS definition are provided in APPENDIX C. SMARTS Implementation and SMARTS Description.

Table 1: List of point generators

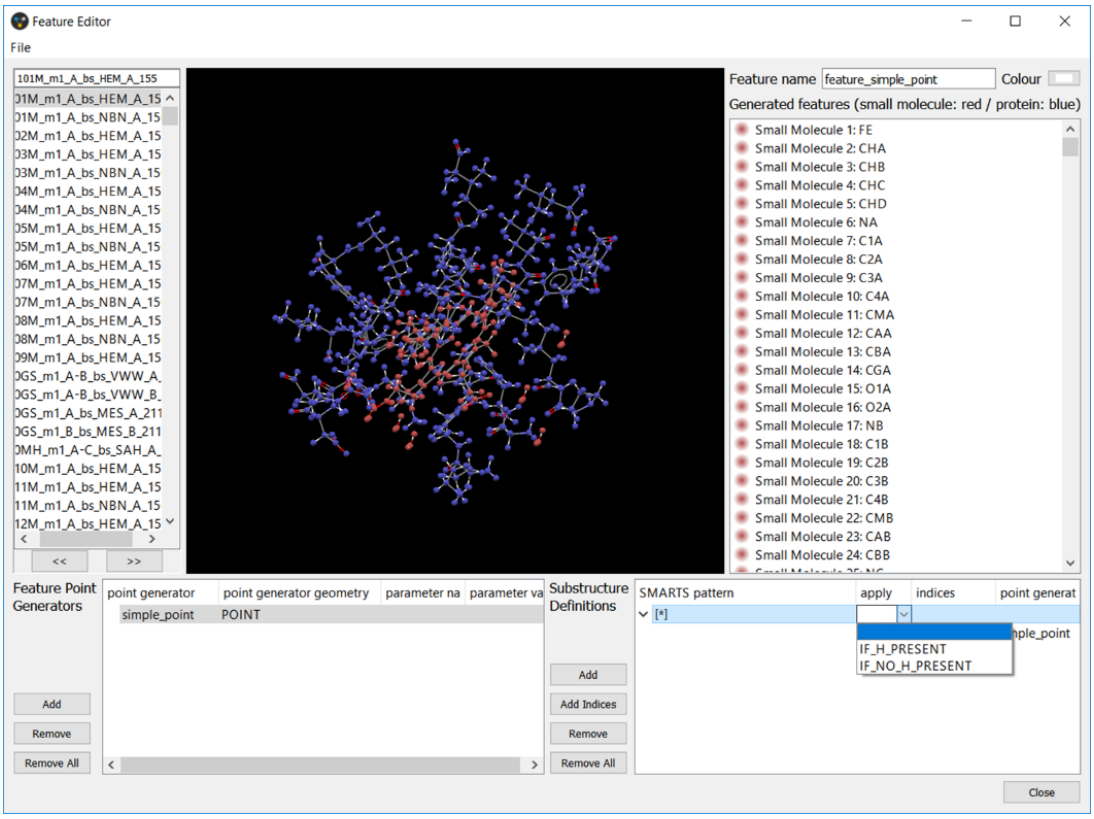
Name	Parameters	Description
DUMMY		Marks the selected atoms as match although they are not used to generate any feature points. Practically, it works as **NOT**; if an atom matches DUMMY, will not generate a feature point.
POINT		Creates a base feature point at the selected atom position .
CENTROID		Creates a base feature point at the centroid

Name	Parameters	Description
		of all selected atom positions.
CENTROID_PLANAR	planarity threshold: 0.1 Å	Creates a base feature point at the centroid of all selected atom positions if the maximum distance of any selected atom positions to the least-squares fitted plane of the whole atom set is less than the planarity threshold.
CENTROID_NONPLANAR	planarity threshold: 0.1 Å	Creates a base feature point at the centroid of all selected atom positions if the maximum distance of at least one selected atom position to the least-squares fitted plane of the whole atom set is higher than the planarity threshold.
LINEAR	virtual point distance: 2.8 Å	creates a base feature point at the selected atom position and a virtual one along the negative direction vector formed with the neighbouring atom (distance specified as a parameter).
LINEAR_NB		creates a base feature point at the neighbour of the

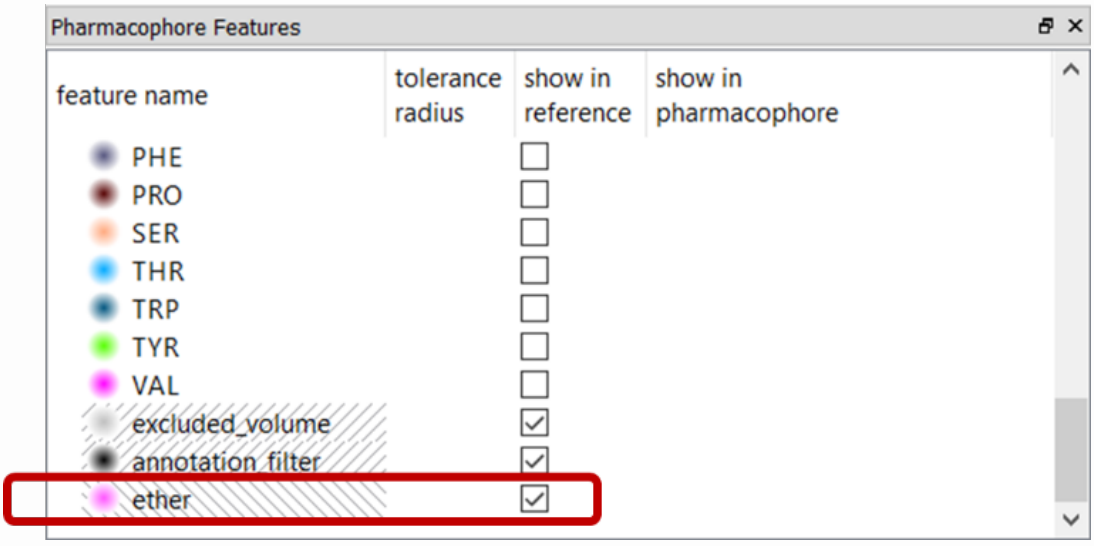
Name	Parameters	Description
	virtual point distance: 2.8 Å	selected atom and a virtual one along the direction vector formed with the selected atom (distance specified as a parameter).
NORMAL	virtual point distance: 2.8 Å	creates a base feature point at the centroid of all selected atom positions and the virtual ones along the normal (N and -N) of the least-squares fitted plane of the whole atom (distance specified as a parameter).
NORMAL_PLANAR	virtual point distance: 2.8 Å planarity threshold: 0.1 Å	creates a base feature point at the centroid of all selected atom positions and the virtual ones along the normal (N and -N) of the least-squares fitted plane of the whole atom (distance specified as a parameter) if the selected atom set is classified as planar with respect to the specified planarity threshold.
TRIGONAL	virtual point distance: 2.8 Å	places a base feature point at the selected trigonal planar acceptor atom and a

Name	Parameters	Description
		virtual one at the virtual lone pair position (distance specified as a parameter). Multiple pairs may be generated; e.g., two for a carbonyl oxygen.
TETRAHEDRAL	virtual point distance: 2.8 Å	places a base feature point at the selected tetrahedral acceptor atom and a virtual one at the virtual lone pair position (distance specified as a parameter). Multiple pairs may be generated; e.g., two for a hydroxyl oxygen.

In addition, it is possible to constrain when to apply the created definition in dependence of the protonation state of the structure.



The new feature definition can be saved to disk by clicking on **File** > **Save Feature Definition** in the **Feature Editor** window. In addition, it is possible to search for this feature on-the-fly by clicking on **File** and then **Add Feature to Current Feature Database** in the **Feature Editor** window. This will make the newly created features available in the **Pharmacophore Features** browser of CSD-CrossMiner with diagonal hatching to indicate that this feature has not been pre-calculated and therefore that the loaded database has not been indexed with this feature definition (e.g., ether in the example below).



Note that performing a pharmacophore search with a pharmacophore query generated with a non-indexed feature definition can affect the time performance of the pharmacophore search.

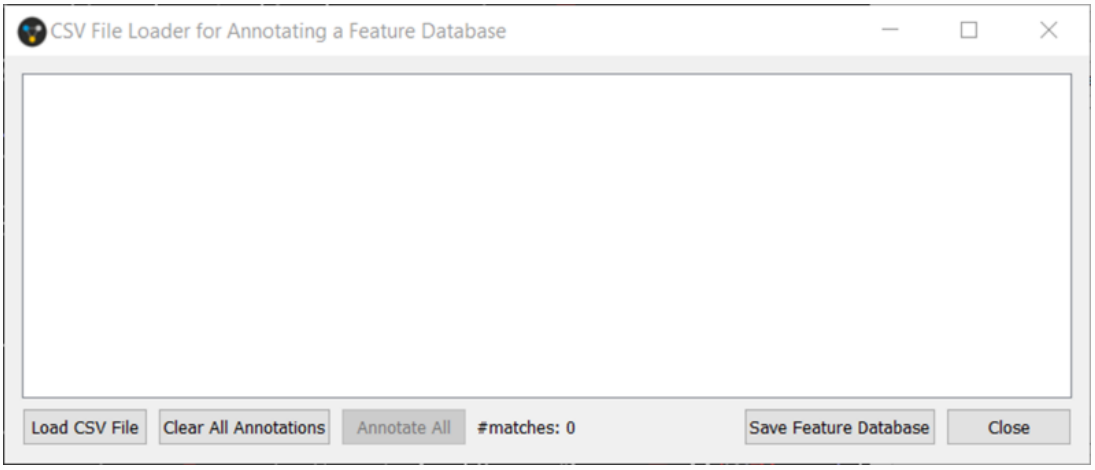
It is possible to permanently index the database with the new feature by recreating the feature database with all features including the new feature (see Creating a Feature Database).

More details on the feature definitions and SMARTS implementation are provided in the APPENDICES.

Annotating a Feature Database

Feature database entries in an existing feature database can be annotated with additional user-defined data. This data takes the form of key-value pairs of text, and they will be displayed in the **Results Hitlist** window.

The annotations can be added to a loaded feature database by clicking on **Feature Database** from CSD-CrossMiner top-level menu and then choosing **Annotate** from the pull-down menu.

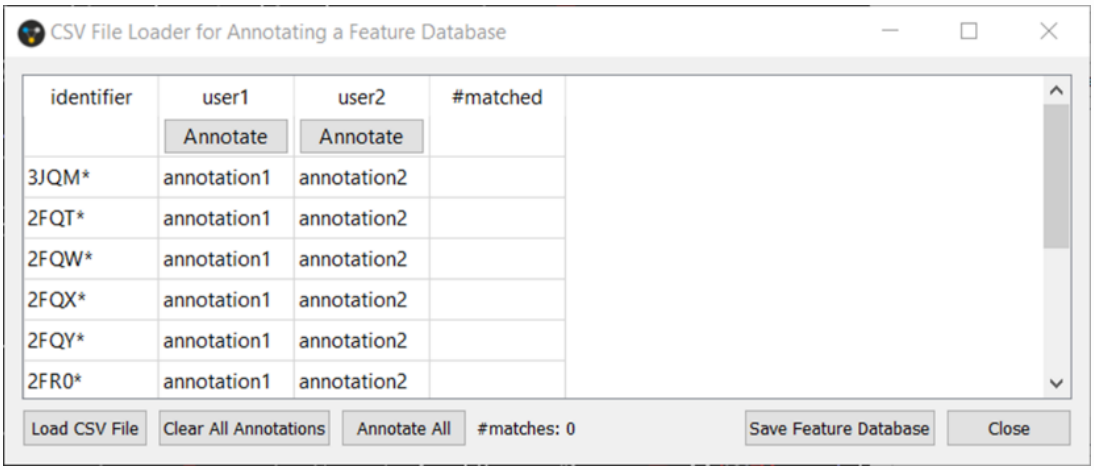


If the feature database already contains annotations (as is the case for the supplied feature database), these can be cleared by clicking **Clear All Annotations** in the **CSV File Loader for Annotating a Feature Database** pop-up window.

A csv file containing annotations for the database entries can be loaded by clicking **Load CSV File**. This csv file should list the annotation names in the first row, and the identifiers should be given in the first column; e.g.

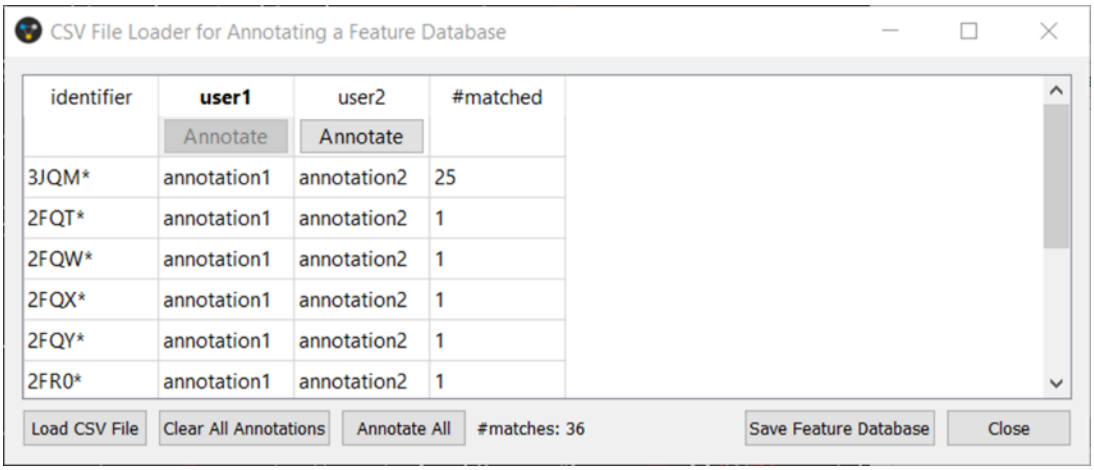
```
identifier, user1, user2, …

AABHTZ, one, two, …
```



The first column of the csv file will be used for matching the identifiers stored in the feature database (see Identifier Matching Rules). Note that only the first 10 lines of the csv file will be displayed in the **CSV File Loader for Annotating a Feature Database**.

A single annotation can be loaded by clicking on the **Annotate** button in the respective annotation table. After the **Annotate** button is clicked, the **#matched** column will display the count of database entries from the feature database that has matched the corresponding identifier. Additionally, the number of overall matches produced will be displayed next to **#matches** at the bottom of the **CSV File Loader for Annotating a Feature Database** window.

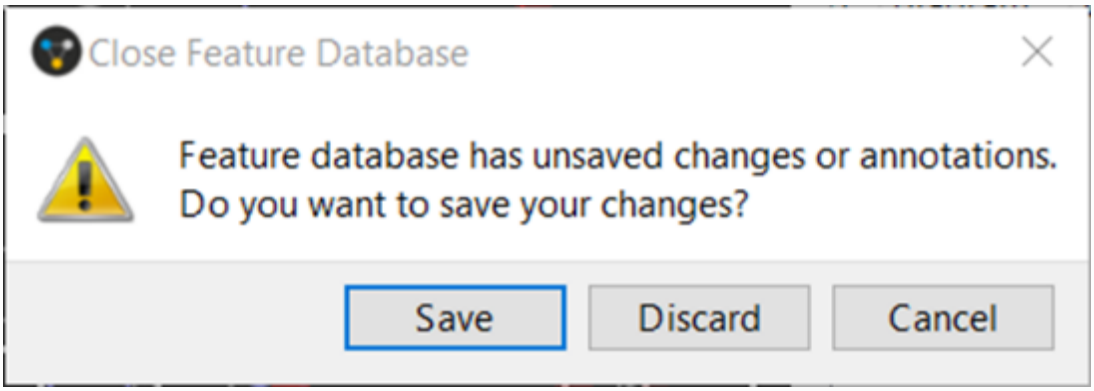


Note that the added annotation will then be immediately available in the **Results Hitlist** window.

All columns from the csv file may be annotated and matched by clicking the **Annotate All** button.

Please note that to permanently store the annotation in the feature database, it is necessary to save the database to disk. This can be done by clicking on the **Save Feature Database** button in the **CSV File Loader for Annotating a Feature Database** window or via **File** > **Save Feature Database** in the CSD-CrossMiner top-level menu.

If the current database is closed before annotations are saved (via **File** > **Close feature database** or **File** > **Exit** from the top-level menu), a **Close Feature Database** pop-up window will allow to **Save** or **Discard** the annotation before closing the database, or to **Cancel** the close feature database action.



Identifier Matching Rules

Identifier matching rules are similar to UNIX shell wildcards used to match the key and value during the annotation filtering process (see Filtering Matching Rules).

- `?` - Matches any one single character.
- `*` - Matches zero or more of any characters.
- `[...]` - Matches any one of the set of characters listed within the square brackets; e.g., “[0123456789]” will match a numeric character only.
- `\` - Escape character to treat the following special character as a literal character; e.g., `\?` will match a `?` character only.

Any other character represents itself apart from those described above; e.g., `a` matches the character `a`.

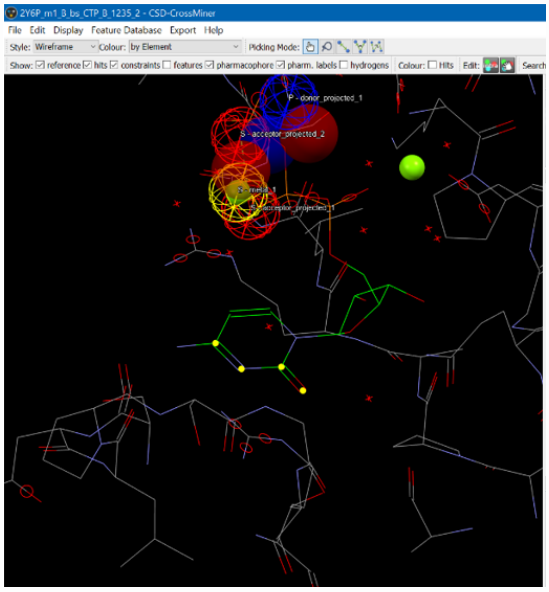
Each identifier in the csv file may match multiple entries in the feature database. For example, the identifier `AACANI*` would match both `AACANI10` and `AACANI11` from the CSD. Subsequent identifier matches to the same entry will overwrite any previously defined annotations for that entry.

Note that using UNIX shell wildcards for applying many annotations to a large database (i.e., a database with many entries) can be very slow. To speed up this process, it is recommended that simple exact string matching is used when the number of identifiers specified for matching is greater than a few tens of thousands.

Selecting molecules

There are several ways to select molecules or atoms in the 3D view:

- **Picking Mode:** toolbar
  - **Pick Atoms** mode  allows the selection of atoms by clicking on them in the display area (atoms which are selected are indicated by a network of yellow lines that wrap around the atom). It is possible to deselect an atom by clicking on it.

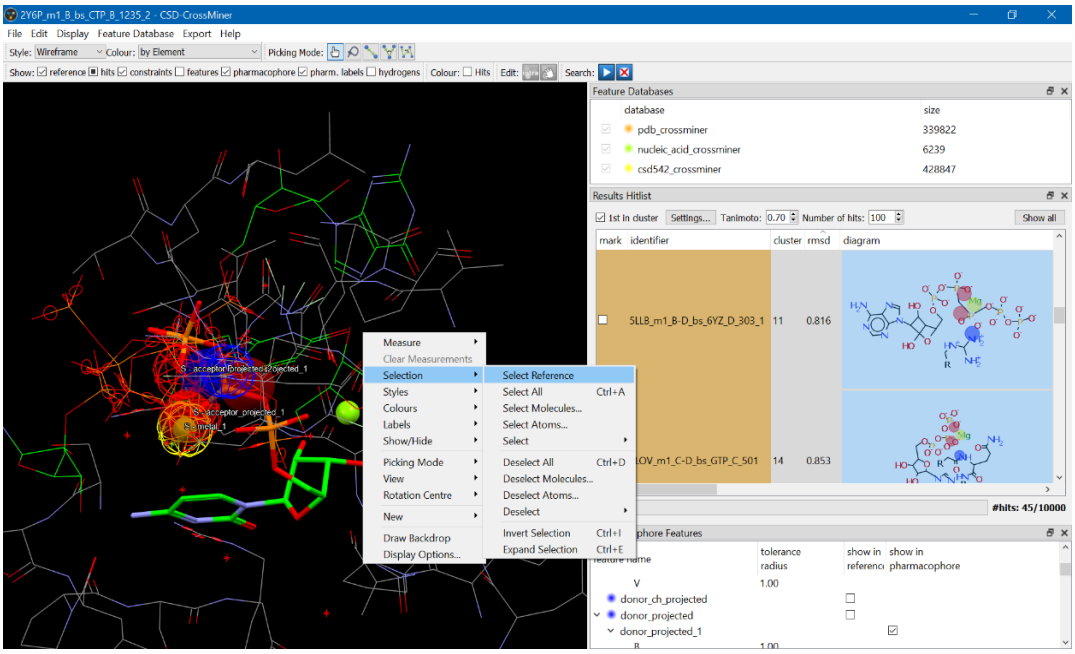


- The **Lasso Atoms** mode  allows the selection of a range of atoms by drawing a perimeter around those atoms.
- 3D view right-click menu; right-click anywhere in the 3D view area (atom, bond or background) and choose **Selection** from the resulting menu. This gives you access to:
  - **Select All**, **Deselect All** and **Invert Selection** options, to select or deselect all atoms in the 3D view or invert the selection.

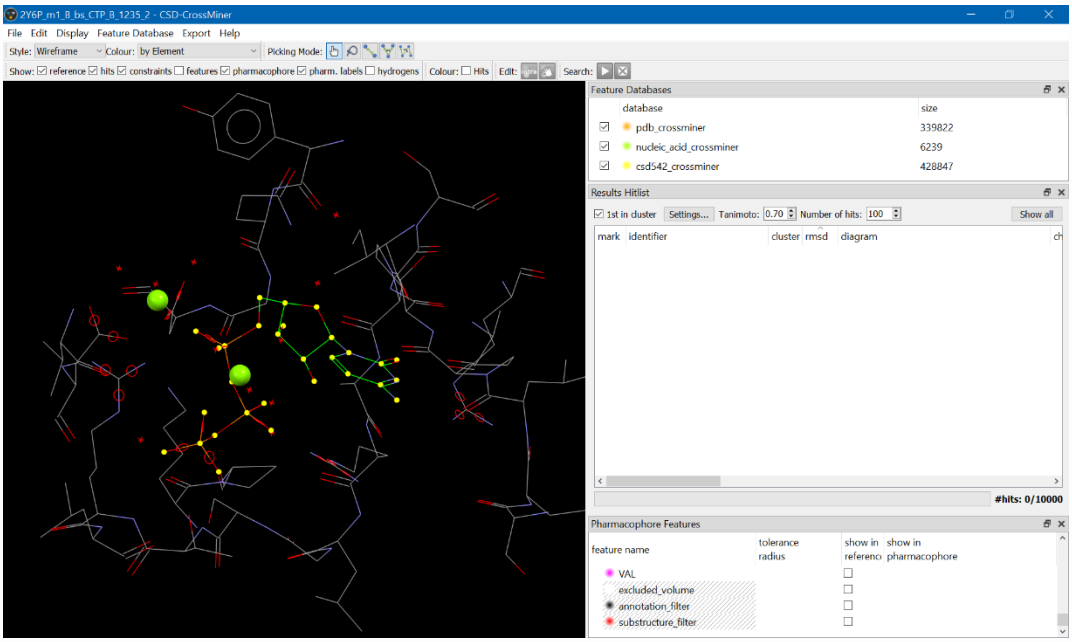
- Select or Deselect to select or deselect all atoms of a particular type, e.g., **Non-Hydrogen**, **Protein Backbone Atoms**.
- **Select Molecules** or **Select Atoms** to select individual molecules or atoms, respectively by left-clicking on them.
- **Expand Selection** to select all atoms that are directly bonded to atoms that have already been selected.

To select only the reference molecule:

- Right-click in the 3D view background, pick **Selection** from the pull-down menu and select **Select Reference.** The atoms in the reference molecule will be highlighted in the 3D view.



To select a single molecule, you can also use **Shift + LMB** combination on one of the atoms of the molecule you want to select. This will select the entire molecule and it will be highlighted in the 3D view.

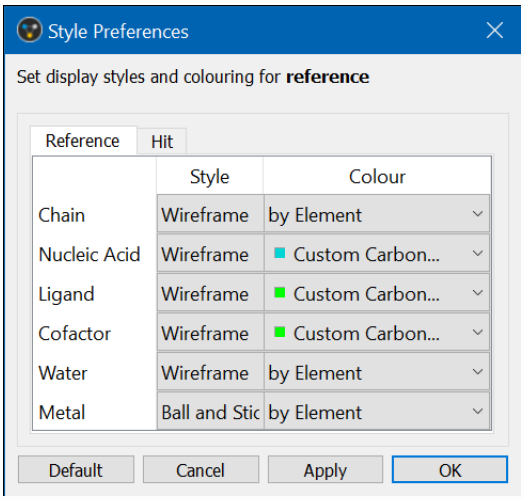


Once selected, a set of atoms can be subjected to an operation, e.g., changing the style and/or colour (see Altering the Style and Colour settings).

Customising the Display

Setting Default Style and Colour Preferences for Reference and Hit

To control the colours and styles that are applied by default to the different components of the reference and hit structures (protein chains, nucleic acids, ligands, cofactors, water molecules, metal ions), hit **Display** in the CSD-CrossMiner top-level menu followed by **Style Preferences...** in the resulting pull-down menu.



The resulting dialogue box has two tabs, **Reference** and **Hit**.

- Select the **Reference** tab to specify the style/colour settings for the reference structure(s).
- Select the **Hit** tab to specify the style/colour settings of the matching hit structures.

The dialogue box may be used to set default styles and colouring schemes for the various parts of a protein structure. The **Custom Carbon** option allows a colour to be specified for carbon atoms via the colour palette, leaving the colours of remaining atoms unchanged.

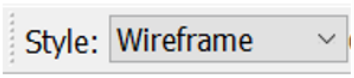
Preferences will be retained between sessions.

Altering the Style and Colour settings

To set all the atoms visible in the 3D view to a new display style (wireframe, stick, ball and stick, spacefill or ellipsoid), you must first ensure that no atoms are selected (atoms which are selected are indicated by a network of yellow lines that wrap around the atom).

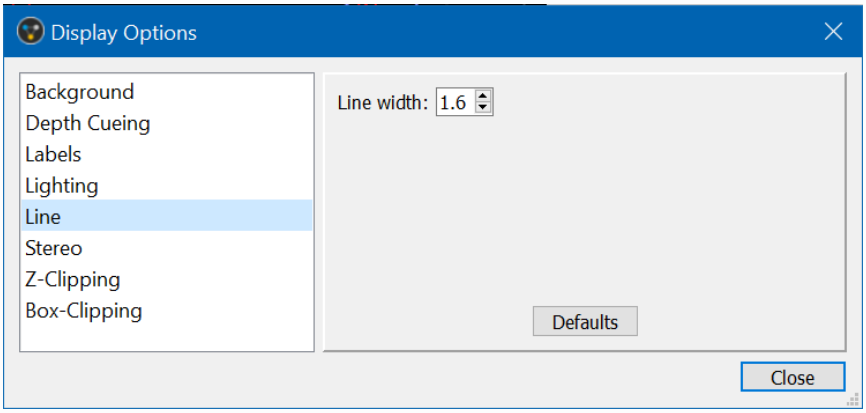
You can then use any of the following methods:

- Set the required style in the **Style** box in the style toolbar, e.g.



- Right-click in the 3D view background, pick **Styles** from the pull-down menu, and select the required style (**Wireframe, Capped Sticks, Ball and Stick, Spacefill or Ellipsoid**).
- Hit **Display** in the top-level menu, **Styles** in the resulting pull-down menu, then pick the required display style.

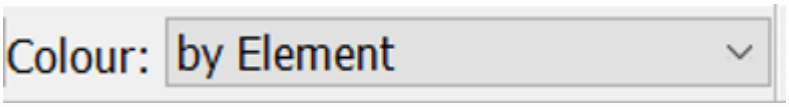
If using wireframe display it is possible to change the line width by selecting **Line** under **Display options...** (under **Display** on the top menu bar). The line width must fall in the range 1.0 to 2.0.



To set all the atoms visible in the 3D view to a new display colour, you must first ensure that no atoms are selected. Atoms which are selected are indicated by a network of yellow lines that wrap around the atom.

You can then use any of the following methods:

- Right-click in the 3D view background, pick **Colours** from the pull-down menu, and select the required colour style.
- Set the required colour in the **Colour** box in the colour toolbar, e.g.



Colour: controls the global colouring scheme of molecules visible in the 3D view. The following global colouring schemes are available:

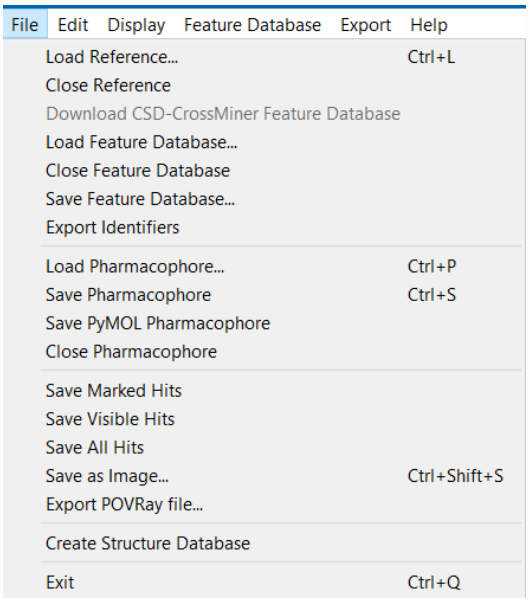
- Colour by Element.
- Colour by Symmetry equivalence.

- Colour by Atomic displacement.
- Colour by Symmetry operation.
- Colour by Gasteiger charge.
- Colour by Partial charge.
- Colour by Element or Suppression.

Descriptive Menu Documentation

CSD-CrossMiner Top-Level Menu

File Menu

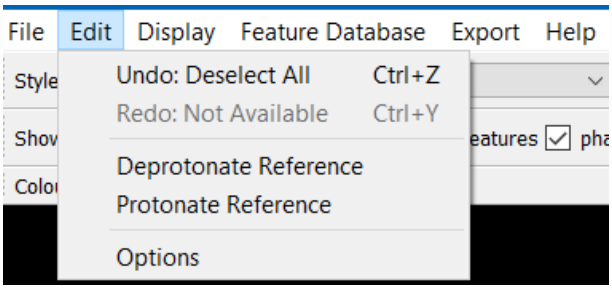


A description of each file option is given below:

- **Load Reference**: Load reference structure(s) which can be used as a template to define a pharmacophore.
- **Close Reference**: Close the loaded reference structure(s).

- **Download CSD-CrossMiner Feature database**: Download the CSD-CrossMiner feature database if it has not been downloaded the first time the CSD-CrossMiner session is started. If the database has been downloaded already, this option won't be available.
- **Load Feature Database**: Loads a feature database.
- **Close Feature Database**: Close the loaded feature database.
- **Save Feature Database**: Saves the current feature database.
- **Export Identifiers**: Save the identifier of the loaded feature database. These can be used to create a cvs file.
- **Load Pharmacophore**: Load an existing pharmacophore.
- **Save Pharmacophore**: Save the current pharmacophore.
- **Save PyMOL Pharmacophore**: Save the pharmacophore in a ".py" file which can be loaded into PyMOL.
- **Close Pharmacophore**: Clear a pharmacophore query.
- **Save Marked Hits**: Save all hits marked in the Results Hitlist browser.
- **Save Visible Hits**: Save all hits currently visible in the 3D view.
- **Save All Hits**: Save all hits (the number of hits will correspond to the hit count shown in the progress toolbar in the Results Hitlist).
- **Save as Image**: Save the 3D view as an image.
- **Export POVRay file**: Store the scene in a POV-Ray file.
- **Create Structure Database**: Create a structure database from mol2/sdf files (see Creating a Structure Database).
- **Exit**: Exit the program.

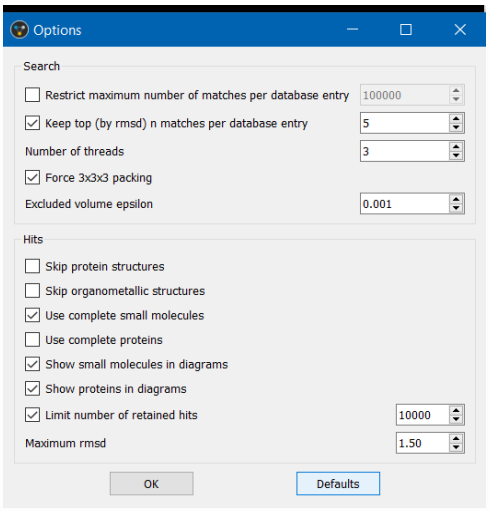
Edit Menu



Note that for macOS users the **Options** dialogue is located in CSD-CrossMiner main task bar **>> Preferences**. A description of the **Edit** menu is given below:

- **Undo**: Undo the last command.
- **Redo**: Redo the last command.
- **Deprotonate Reference**: Remove all hydrogens from the current reference structure. This can have an impact on features that rely on hydrogen absence/presence.
- **Protonate Reference**: Add a default protonation to the current reference structure.
- **Options**: Show the Options dialog. The dialog can only be shown if no pharmacophore search is running (including paused searches).

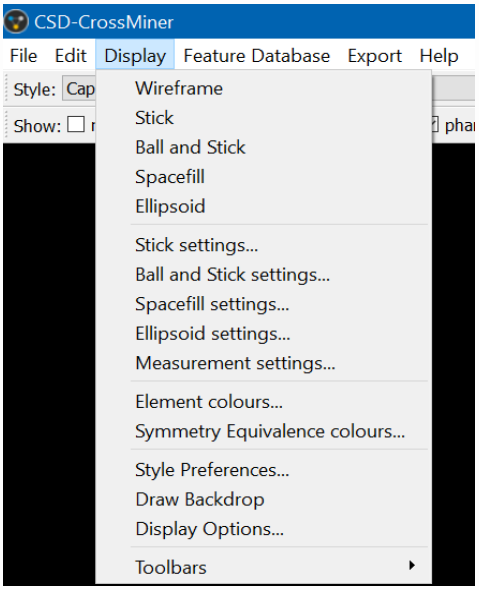
Selecting options brings up an options menu with the following selections:



- **Restrict maximum number of matches per database entry**: Restrict the maximum number of hits per structure to the specified value. If this option is unticked no restriction is applied.
- **Keep top n solutions (by rmsd) n matches per database entry**: Only keep the top n hits with respect to the Kabsch overlay rmsd (relative to the set of hits returned which has at most as many hits as specified in the previous option). If this option is unticked no restriction is applied.
- **Number of threads**: To specify the computational resources to dedicate to the pharmacophore search
- **Force 3x3x3 packing**: Restrict the packing to 3x3x3 unit cells.
- **Excluded volume epsilon**: Epsilon value applied to the excluded volume pharmacophore point radius.
- **Skip protein structures**: Don't display protein structures (only applies to features which have been set to be part of a protein component).
- **Skip organometallic structures**: Don't display hits that contain at least one transition metal, lanthanide, actinide, or any Al, Ga, In, Tl, Ge, Sn, Pb, Sb, Bi, Po.
- **Use complete small molecules**: Don't restrict a small molecule fingerprint to the bounding sphere (only applies to features which have been set to be part of a small molecule component).

- **Use complete proteins**: Don't restrict the protein fingerprint to the bounding sphere.
- **Show small molecules in diagram**: Display the small molecule 2D diagram pharmacophore overlay match.
- **Show proteins in diagram**: Display the protein 2D diagram pharmacophore overlay match.
- **Limit number of retained hits**: The maximum number of hits displayed in the 3D view and in the Results Hitlist browser.
- **Maximum rmsd**: The maximum Kabsch overlay rmsd allowed for any hit.

Display Menu

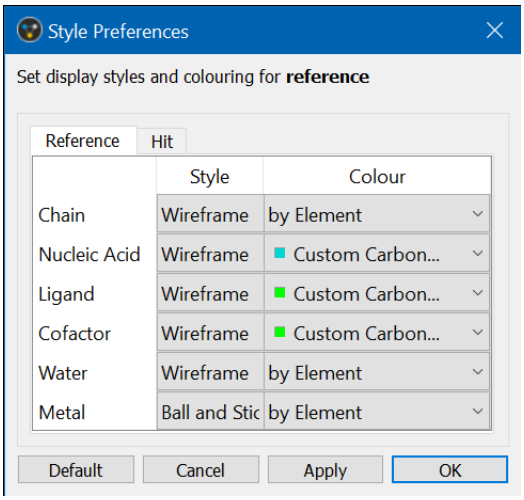


The display menu has the following options:

- **Wireframe, Stick, Ball and Stick, Spacefill, Ellipsoid**: Change the representation style of the molecule displayed in the 3D view.
- **Wireframe settings, Stick settings, Ball and Stick settings, Spacefill settings, Ellipsoid settings, Measurement settings**: Modify the current settings of the representation and measurement style.
- **Element colours**: Change the colour of the chemical element.

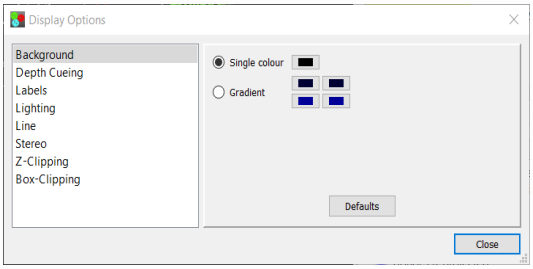
- **Symmetry Equivalence colours**: Change the colour used for Colour by Symmetry equivalence.

- **Style Preferences**: Change the style and colour to apply to the different components of the reference and hit molecules. The style preferences will be remembered between sessions.

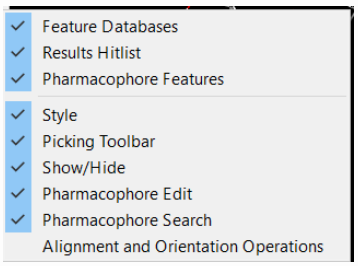


- **Draw Backdrop**: Switch between the default black background and an alternative colour by right-clicking in the background area and hitting **Draw Backdrop**. The alternative colour will be a blue gradient.

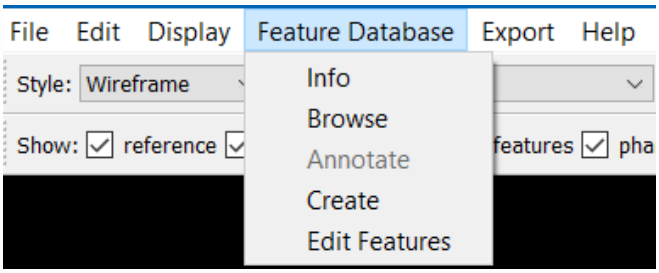
- **Display Options**: Create and modify sets of CSD-CrossMiner display styles.



- **Toolbars**: Switch on and off CSD-CrossMiner toolbars and windows.



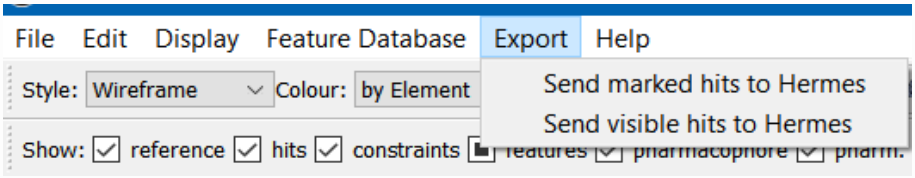
Feature Database Menu



The **Feature Database** menu has the following options:

- **Info**: Display some information on the type of features and the number of structures stored in the database.
- **Browse**: Open the feature database browser (see Creating a Pharmacophore Query from a Reference Structure).
- **Annotate**: Open the annotation dialog (see Annotating a Feature Database).
- **Create**: Open the feature database creator (see Creating a Feature Database).
- **Edit Features**: Open the substructure feature editor (see Editing and Creating Feature Definitions).

Export

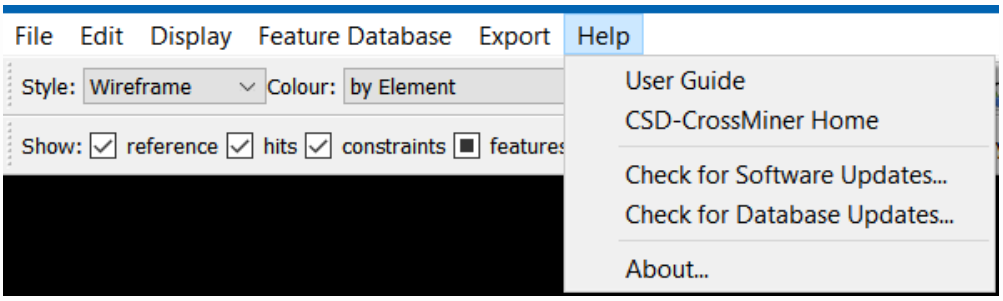


The **Export** menu consists of the following:

- **Send marked hits to Hermes**: Send all hits marked in the Results Hitlist browser to Hermes.
- **Send visible hits to Hermes**: Send all the visible hits listed in the Results Hitlist browser to Hermes.

Note that the options in the **Export** menu are not available before starting the pharmacophore search.

Help Menu



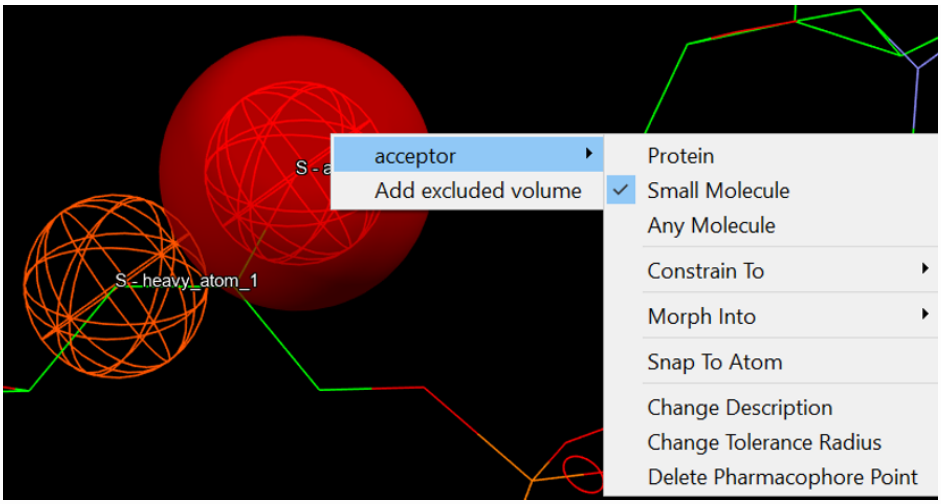
The help menu consists of the following selections:

- **User Guide**: Provide access to CSD-CrossMiner User Guide.
- **CSD-CrossMiner Home**: Provide access to the CSD-CrossMiner web page. From there you can access to User Guide, tutorials and workshop examples.
- **Check for Software Updates**: Download available updates of CSD-CrossMiner.
- **Check for Database Updates**: Download available updates of the CSD-CrossMiner feature database.
- **About**: Display details about your version of CSD-CrossMiner including the current licensing status.

Context Right-Click Menu

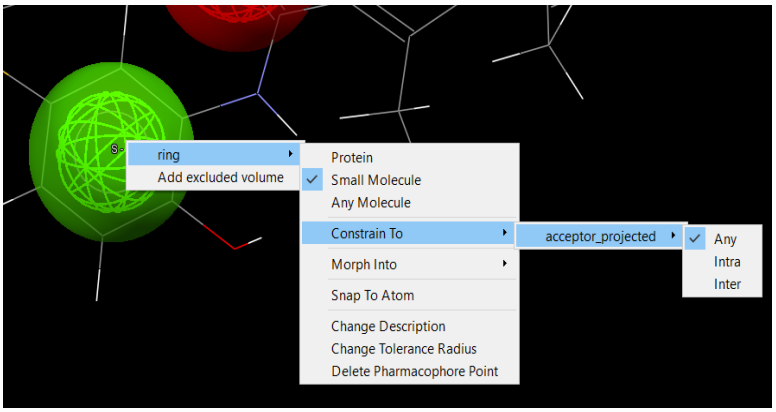
Pharmacophore Context Right-Click Menu

Right-click on a pharmacophore feature (see Modifying a Pharmacophore Query).



This will bring up the following selections:

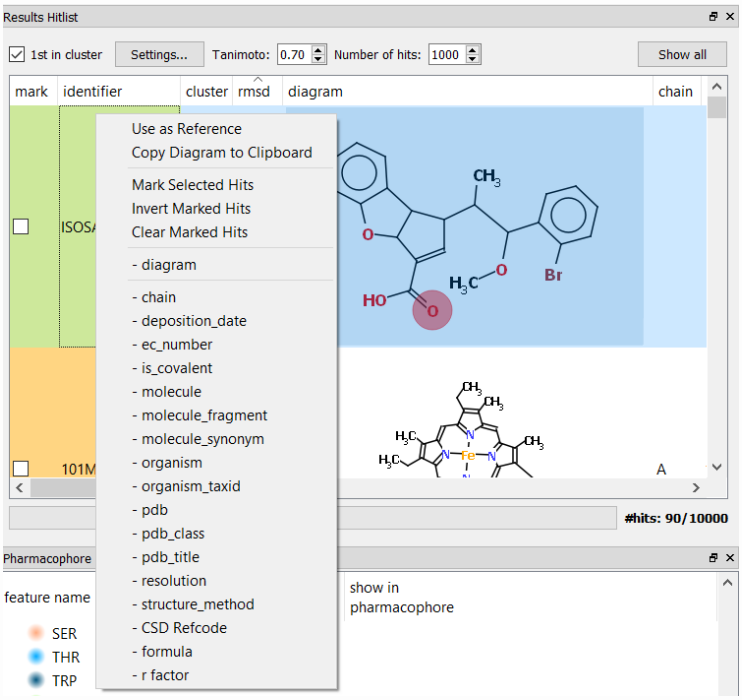
- **Protein**: Require pharmacophore point to be part of a protein. A “P” label will be displayed for the pharmacophore point.
- **Small Molecule**: Require pharmacophore point to be part of a small molecule (including nucleic acids). An “S” label will be displayed for the pharmacophore point.
- **Any Molecule**: Pharmacophore point is part of a protein or a small molecule (including nucleic acids). An “A” label will be displayed for the pharmacophore point.
- **Constrain To**: Allow the specification of a constraint to another pharmacophore point.



- **Any**: Both features can be part of the same or different molecules. No constraint will be visualised.

- **Intra**: Pharmacophore points must be part of the same molecule. A dashed green line will be drawn between the pharmacophore points.
- **Inter**: Pharmacophore points must be part of different molecules. A dashed red line will be drawn between the pharmacophore points.
- **Morph Into**: Change pharmacophore type into another one that has the same number of feature points and feature point types.
- **Snap To Atom**: Move the pharmacophore point to the nearest atom.
- **Change Description**: Change the description displayed next to the molecule type label. If no description is supplied by the user, the feature type will be displayed instead.
- **Change Tolerance Radius**: Change the tolerance radius of a selected pharmacophore point.
- **Delete Pharmacophore Point**: Delete the pharmacophore point.

Results Hitlist Context Right-Click Menu

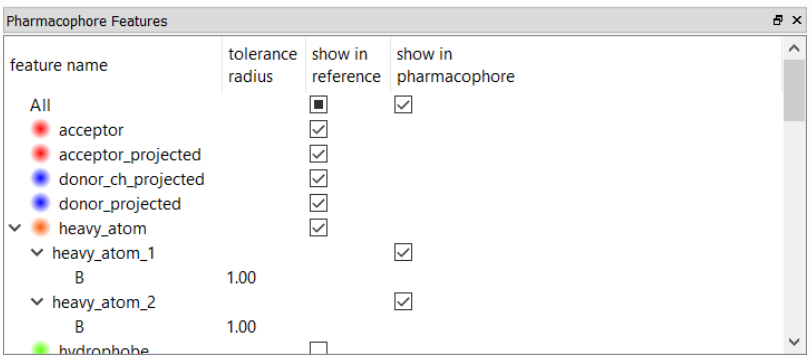


This menu contains the following selections:

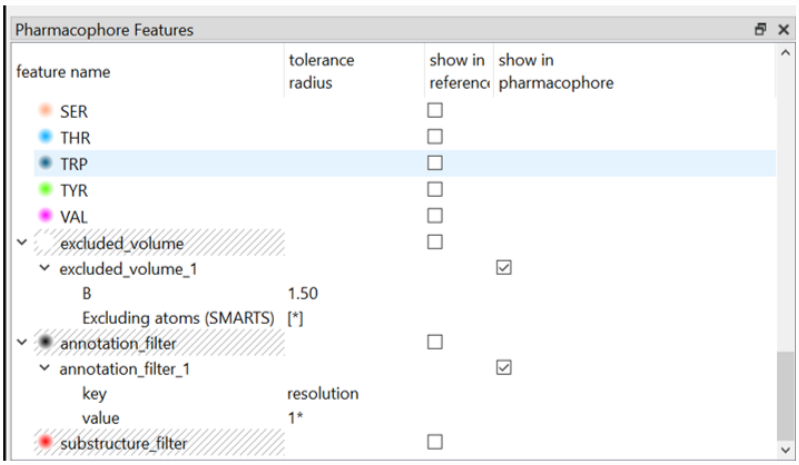
- **Use as reference**: The selected hit will be shown in the 3D view and annotated with the features available in the current feature database (see Creating a Pharmacophore Query from a Hit).
- **Copy Diagram to Clipboard**: Copy the diagram to the clipboard.
- **Mark Selected Hits**: Mark all the hits selected in the **Results Hitlist** browser.
- **Invert Marked Hits**: Invert the previously marked hits.
- **Clear Marked Hits**: Unmark the marked hits.
- **"diagram"**: Hide (-)/show (+) diagram in the **Results Hitlist** window.
- **"chain, deposition_date, ec_number, r_factor etc."**: List of annotations registered in the feature database. It is possible to hide (-)/display (+) an annotation in the **Results Hitlist** window by clicking on it (see Results Hitlist and Results Hitlist Browser).

Feature and Pharmacophore Window Context Right-Click Menu

The feature types available in the feature database as well as the feature spheres used in the current pharmacophore are shown in this window. If a reference structure is shown in the 3D view, the feature points of the respective types can be shown/hidden by ticking/unticking the corresponding tick-boxes, or the **All** tick-box can be used to show/hide all feature types. The same holds for the pharmacophore points. The radii of individual pharmacophore points can be modified using the respective spin-boxes.

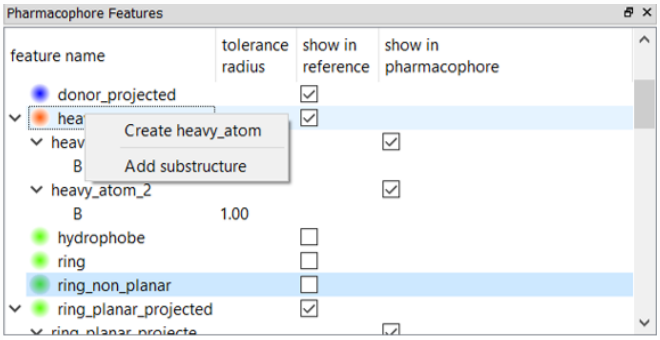


The **excluded volume**, **annotation_filter** and **substructure_filter** features are never indexed in the feature database. This is represented with diagonal hatching in the **Pharmacophore Features** window.



Right-clicking on a feature type will show this (or a similar) context menu:

- **Create**: Create a pharmacophore point of this type.
- **Add substructure**: Add a new substructure feature type to the feature database. This will open the Feature Editor (see Editing and Creating Feature Definitions).



CSD-CrossMiner Toolbars

Style & Colour and Picking Mode Toolbars

This toolbar contains common, basic options; e.g., **Style** for setting global display styles; **Colours** for setting the colour mode; **Picking Mode** for picking or lassoing atoms and for measuring distances, angles and torsions.

Style controls the display style of molecules visible in the 3D view (wireframe, capped sticks, ball and stick, spacefill or ellipsoid).

Colour controls the global colouring scheme of molecules visible in the 3D view. The following global colouring schemes are available:

- Colour by Element.
- Colour by Symmetry equivalence.
- Colour by Atomic displacement.
- Colour by Symmetry operation.
- Colour by Gasteiger charge.
- Colour by Partial charge.
- Colour by Element or Suppression.

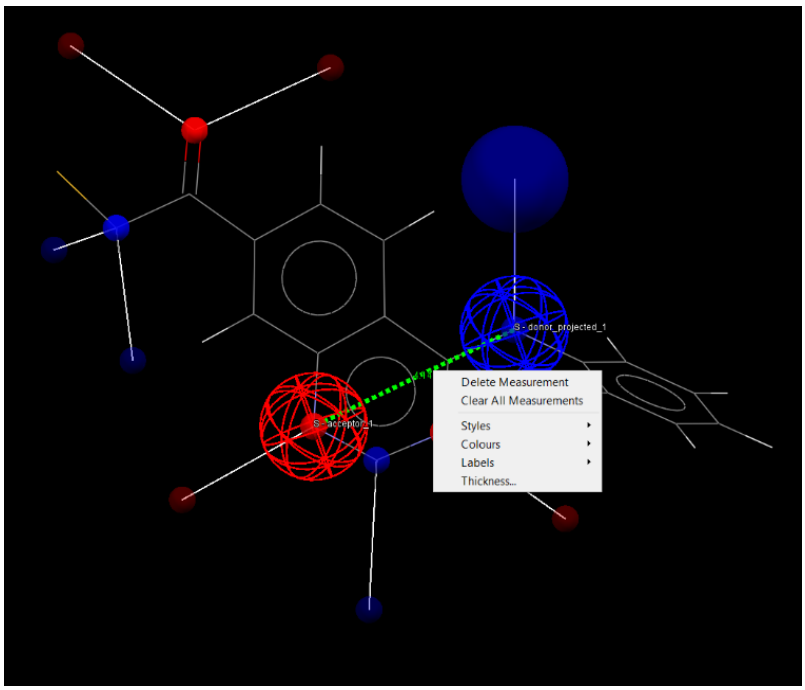
Picking Mode controls what happens when you left-click on items in the display area:

- Use of the **Pick Atoms** mode  allows the selection of atoms by clicking on them in the display area (the selection is represented as a small yellow sphere). It is possible to deselect an atom by clicking on it.
- The **Lasso Atoms** mode  allows the selection of a range of atoms by drawing a perimeter around those atoms.

Once selected, a set of atoms can be subjected to an operation; e.g., changing the style and/or colour (see Altering the Style and Colour settings).

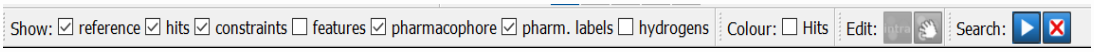
Use of **Measure Distance**, **Measure Angle** or **Measure Torsion** modes  permits the measurement of geometrical parameters by picking two, three or four objects (e.g., atoms,

centroids), respectively. To remove all distance, angle and torsion angle measurements select **Clear Measurements** when right-clicking in the 3D view. It is also possible to remove a single bond, angle or torsion angle measurement by right-clicking on it and selecting **Delete Measurement** from the resulting pull-down menu.



The colour and style of a single measurement can be changed by right-clicking on the measurement to have access to the **Style** and **Colours** from the pull-down menu. The settings of all the distances, angles and torsions can be edited through **Measurement settings** from the **Display** top-level menu.

Show, Edit, Colour: Hits and Search Toolbars



Show: Show/hide the reference structure, the hits, the pharmacophore constraints, the features, the pharmacophore, the pharmacophore labels and the hydrogen atoms.

Colour: Hits: Allow hits to be coloured by rainbow. Colour is assigned to the cluster therefore hits belonging to the same cluster have the same colour.

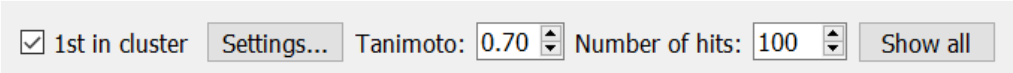
Edit: control two pharmacophore edit options, which are

 Assign intramolecular constraints between all pharmacophore points of the same type.

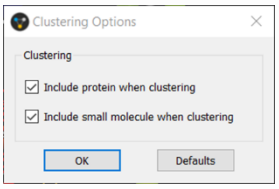
 Enables interactive editing mode. It allows an individual pharmacophore point to be translated. Any change in the position of a pharmacophore point will trigger a new pharmacophore search.

Search: Play/Pause and Stop the pharmacophore search (see Pharmacophore Search).

Results Hitlist Toolbar



- **1st in cluster**: Enable/disable clustering of hits using the Tanimoto threshold specified in the spin box.
- **Settings**: Allow to include/ exclude the protein and/or the small molecule fingerprint when clustering.



- **Number of hits**: Number of displayed hits in the 3D view and in the **Results Hitlist** browser.
- **Show all**: Visualise all hits in the 3D view.

APPENDICES

APPENDIX A. Command Line Interface

When CSD-CrossMiner is launched from a command-line interface, such as a Linux shell, there are some command-line options. The usage details for the executable will be reported if the `-help` argument is added to the command-line. Some commands are listed below:

- `-help`: Display the usage message and quit.
- `-feature_db`: The feature database to load automatically on start-up. This will override reloading of the previously used feature database.
- `-pharmacophore`: The pharmacophore model to load automatically on start-up.
- `-reference`: The reference structure to load automatically on start-up.

For example, you can use (all on one line):

- Windows:

```
"<CrossMiner installation folder>crossminer.exe" -feature_db
        test.feat -reference test.mol2
```

- Linux:

```
<CrossMiner installation folder>/bin/crossminer -feature_db
        test.feat -reference test.mol2
```

- macOS:

```
<CrossMiner installation folder>/CSD-CrossMiner.app/Contents/
        MacOS/crossminer -feature_db test.feat -reference
        test.mol2
```

To open a CSD-CrossMiner GUI with `test.feat` as feature database and `test.mol2` as reference molecule.

APPENDIX B. Feature Definitions in CSD-CrossMiner

Feature definitions are used to create a feature database. They are derived from the substructures by applying point generation rules, which define how the feature points are determined from a set of atomic coordinates. The feature definitions are divided into one-point, directional and non-indexed features.

List of Feature Definitions

- **One-point**: Acceptor, heavy atom, hydrophobe, ring, ring non planar, adenine, cytosine, guanine, thymine, uracil, purine, pyrimidine, deoxyribose, ribose, halogen, bromine, chlorine, fluorine, metal, water, ALA, ARG, ASN, APS, CYS, GLN, GLU, GLY, HIS, ILE, LEU, LYS, MET, PHE, PRO, SER, THR, TRP, TYR, VAL.
- **Directional**: Acceptor projected, donor projected, donor ch projected, ring planar projected, exit vector.
- **Non-indexed**: Excluded volume, annotation_filter and substructure_filter.

The feature definitions used to create the supplied feature database are included in the `feature_definitions` folder of the `CSD-CrossMiner` directory. Here, the features definitions are grouped in three different folders based on the molecule type: `small_molecule`, `protein` and `any`. Note that the `any` folder also includes feature definitions for nucleic acids.

The molecule types and their corresponding features are listed below:

- **Small molecule**: Exit vector, halogen, bromine, chlorine, fluorine, metal, water.
- **Protein**: ALA, ARG, ASN, APS, CYS, GLN, GLU, GLY, HIS, ILE, LEU, LYS, MET, PHE, PRO, SER, THR, TRP, TYR, VAL.

- **Any molecule**: Acceptor, acceptor projected, donor ch projected, donor projected, heavy atom, hydrophobe, ring, ring non-planar, ring planar projected, adenine, cytosine, guanine, thymine, uracil, purine, pyrimidine, deoxyribose, ribose.

The substructure-based features are defined by a hierarchy of SMARTS patterns. The list of SMARTS patterns defined in CSD-CrossMiner and used to generate the supplied feature database (created from CSD and PDB structures) is provided in APPENDIX C. SMARTS Implementation and SMARTS Description. These SMARTS patterns can be tailored and/or extended by the user, and new substructure-based features can be created and saved to disk (see Editing and Creating Feature Definitions).

APPENDIX C. SMARTS Implementation and SMARTS Description

The SMARTS language allows you to specify substructures using rules that are extensions of SMILES (Simplified Molecular Input Line Entry System). The current CSD-CrossMiner implementation of SMARTS is a subset of the SMARTS functionality described at http://www.daylight.com/dayhtml/doc/theory/theory.smarts.html.

substructure_filter feature and the **Feature Editor** give immediate feedback on the validity the SMARTS string; if the SMARTS pattern defined in the substructure_filter is invalid, an **Invalid Search** pop-up window will invite to correct the SMARTS pattern or delete the filter. Likewise, the **Feature Editor** will give immediate feedback on the SMARTS used to define a new feature by displaying matching patterns of the loaded feature database in the 3D display. The following should be taken into consideration when using the CSD-CrossMiner implementation of SMARTS:

Unsupported features (general)

- Dot for "not necessarily connected" fragments or atoms; e.g., C.C
- Recursive SMARTS; e.g., `[$(CC);$(CCC)]`
- Reaction SMARTS; e.g., `CC>>CC`

Unsupported features (atom properties)

- Some atom constraints (where n is an integer):
  - `h<n>`: implicit hydrogens
  - `R<n>`: ring membership
  - `<n>`: atomic mass
- Stereochemical descriptors
- Constraints of different types combined with OR operator; e.g., `[#7X1,#7D2]`. However, a single feature definition can contain several SMARTS strings that are technically **OR**.
- High precedence **AND** in **OR** subexpression; e.g., `[C,N&H1]`. However, these SMARTS can be defined as separate `[C]` and `[N&H1]` in a single feature definition that will consider them as **OR**.

Unsupported features (bond properties)

- Stereochemical descriptors for double bonds: these are treated as single bonds with unspecified stereochemistry
- High-precedence **AND** in **OR** subexpression; e.g., `=&@,-` (cyclic double or single and unspecified cyclicity)
- The following constructs are not supported:
  - **NOT** any bond; e.g., `!\~`
  - different bond types combined with **AND** operator; e.g., `-&=` (single and double)
  - different **NOT** bond types combined with **OR** operator; e.g., `!-,!=` (not single or not double)

APPENDIX D. Create a Feature Database with In-House Data

Input Files

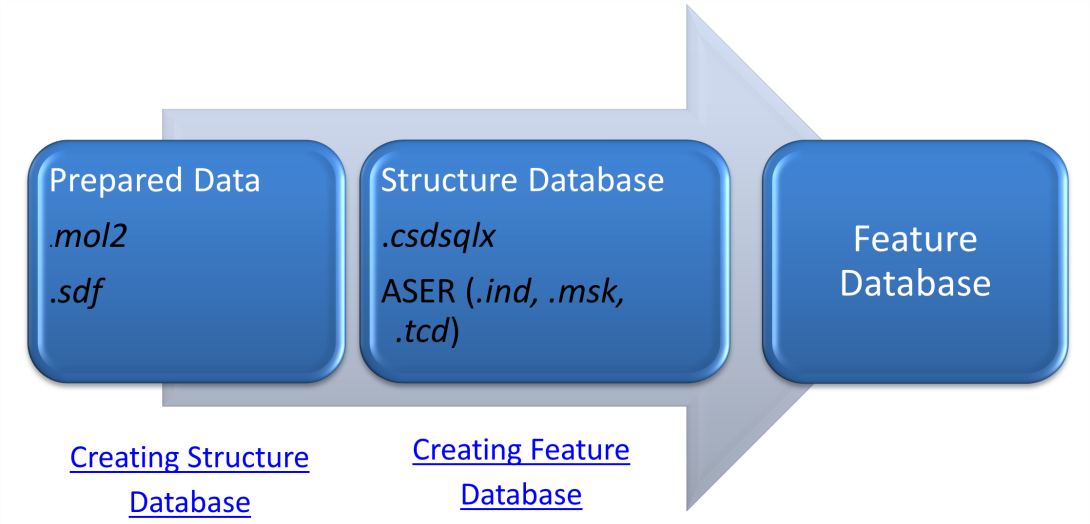
Due to the interactive quality of CSD-CrossMiner, it is recommended that protein files are truncated to the region of interest, such as the binding site(s) for protein-ligand and protein-ligand-nucleic acids complexes. It is possible to include full-length proteins and search across these with CSD-CrossMiner, but there will be an impact on the performance due to holding whole proteins with their associated features in memory.

General Workflow

1. Prepare the data
    - 3D molecules with hydrogen atoms and appropriate atom and bond types are needed. CSD-CrossMiner feature definitions can assign HBD and HBA with no hydrogens based on generic rules about predominant protonation and tautomeric states; however, a more accurate description of the molecules in the database will also result in more accurate results.
    - Recommended format is `.mol2` for protein-ligand complexes and small molecules. For small molecules `.sdf` format can also be used.
    - If only binding sites of ligands in proteins are needed, reduce the protein files to the region surrounding the ligand only.
2. Convert the in-house data into structure database(s) using CSD-CrossMiner
    - See Creating a Structure Database for details.
    - Subsets of the CSD database or in-house small molecule crystal structure datasets are readily structure databases.

3. Convert these structure databases into the feature database using CSD-CrossMiner and the feature definitions
    - See Creating a Feature Database for details.
    - Please note that if one wants several structure databases included in a feature database, these structure databases need to be converted simultaneously into a single feature database following instructions in Creating a Feature Database. This is to ensure homogeneity in the feature definitions.

Note that, structure and/or feature databases can be created using the CSD-CrossMiner interface (see Creating Databases) or using the CSD Python API (see CSD Python API documentation).



APPENDIX E: Pharmacophore search through the CSD Python API

CSD-CrossMiner is fully implemented in the CSD Python API allowing to automate the feature database generation, the pharmacophore query generation and the pharmacophore search workflow. Several cookbook examples are provided in the `examples` folder of CSD Python API downloadable from CCDC website. For a descriptive documentation and available cookbook examples see the CSD Python API documentation.

APPENDIX F: Example Scripts Available for Associated Collaborators

Prepare Input Files for the Structure Database

If only binding sites of ligands in proteins are needed in the feature database, associated collaborators can use `extract_binding_sites_to_mol2.py` example script to reduce the protein files to the region surrounding the ligand only. `extract_binding_sites_to_mol2.py` is available in the `utilities` folder of the `ccdc_rp` CSD Python API package, downloadable from CCDC website, see CCDC RP Utilities section of the CSD Python RP API documentation for more details.

This script can be used to prepare the input data for the creation of the feature database (APPENDIX D. Create a Feature Database with In-House Data).

The `extract_binding_sites_to_mol2.py` script uses the CSD Python API and Biopython[1] packages to extract the protein-ligand binding sites from protein-ligand complexes in PDB format and convert them to mol2 format.

For each ligand with more than 5 atoms and fewer than 100 atoms, included in PDB protein-ligand and protein-ligand-nucleic acids complexes, the `extract_binding_sites_to_mol2.py` script will:

1. Define the protein-ligand binding site as all residues within 6Å of the ligand.
2. Generate a csv file containing information about each generated protein-ligand binding site. This file can be used to annotate the feature database.
3. Add hydrogens using the add_hydrogens function in Python API (see CSD Python API Documentation).
4. Write the result out to a mol2 file.

These mol2 files can then be used to create the structure and feature database as described in APPENDIX D. Create a Feature Database with In-House Data.

1. Biopython is a set of freely available tools for biological computation written in Python. See http://biopython.org/↩