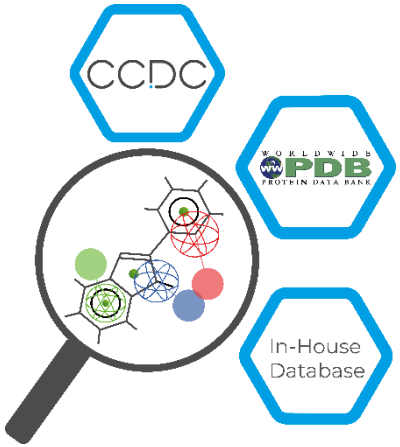
Creating Substructure Features in CSD-CrossMiner (CROSS-001)

Developed using 2024.1 CSD Release



CCDC
advancing structural science

Table of Contents

Introduction

Pharmacophore searching is a key component in many drug discovery efforts and represents an effective mechanism of virtual screening.

In this approach, a pharmacophore query is created to describe features that are essential for the molecule to carry out its function. The query is then used to identify new possible lead compounds by searching a three-dimensional (3D) structural database (Figure 1).

CSD-CrossMiner provides the possibility to search crystal structure databases such as the Cambridge Structural Database (CSD) and the Protein Data Bank (PDB) in terms of pharmacophore queries (Figure 2).

In addition to the common pharmacophore features including hydrophobic, hydrogen bond donor, and hydrogen bond acceptor, with CSD-CrossMiner it is possible to create customised features that allow the inclusion of more specific chemistries in the pharmacophore search.

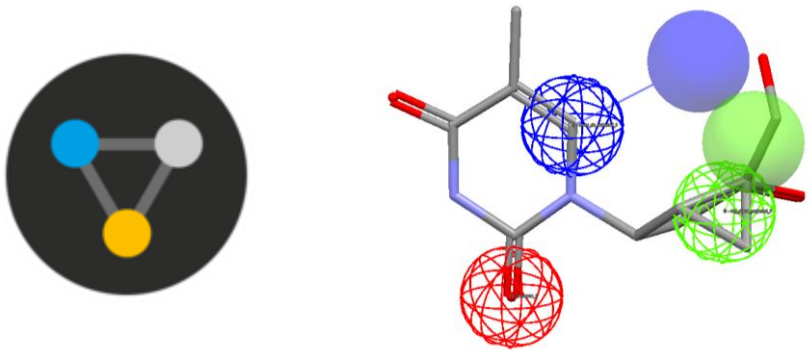


Figure 1. *Pharmacophore example*

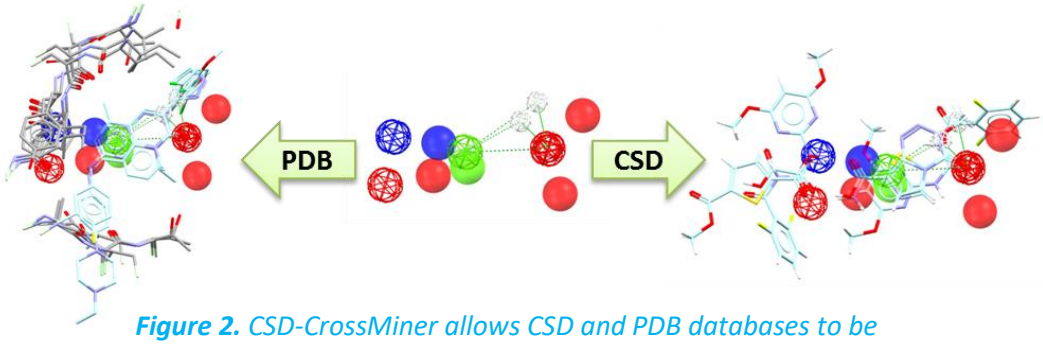


Figure 2. *CSD-CrossMiner allows CSD and PDB databases to be searched in terms of pharmacophore queries.*

Learning Outcomes

After completing this workshop you will be able to:

- Create a new substructure feature definition for a specific chemistry.
- Create and run a pharmacophore search that includes the new feature.

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

Pre-required Skills

There are no pre-required skills for this workshop, however some knowledge of the representation of pharmacophores and features in CSD-CrossMiner is expected. A summary can be found at the end of this handout.

Material

The files to perform this workshop are provided in the `CROSS-001` folder here.

Example 1. Exploring SGLT2-selective inhibitors in CSD-CrossMiner using a 3D pharmacophore search

Sodium glucose cotransporters (SGLTs) have recently attracted considerable attention as new drug targets for the treatment of diabetes. In particular, the selective inhibition of the SGLT subtype 2 (SGLT2) could provide a highly effective method of glycemic control. By targeting renal tubular glucose reabsorption, SGLT2-selective inhibitors exhibit a novel mechanism of action resulting in excretion of glucose into the urine.

A study published by Yoshihito Ohtake *et al.* (DOI: 10.1021/jm300884k) proposes a new class of highly potent and SGLT2-selective inhibitors incorporating a unique *spiroketal* structure. By performing a structural database search using a 3D pharmacophore model based on the superimposition of known inhibitors, Ohtake *et al.* discovered a new potent *O*-spiroketal *C*-arylglucoside scaffold (Figure 3).

In this workshop, we will replicate this work, demonstrating how CSD-CrossMiner can be used to efficiently identify interesting hits that can suggest possible chemistries for use in new lead compounds. The example used here mimics the situation where a researcher wants to search for specific chemistries or features (common to reference molecules in a molecule overlay, Figure 4) in a crystal database that has not been indexed with the desired chemistry or feature.

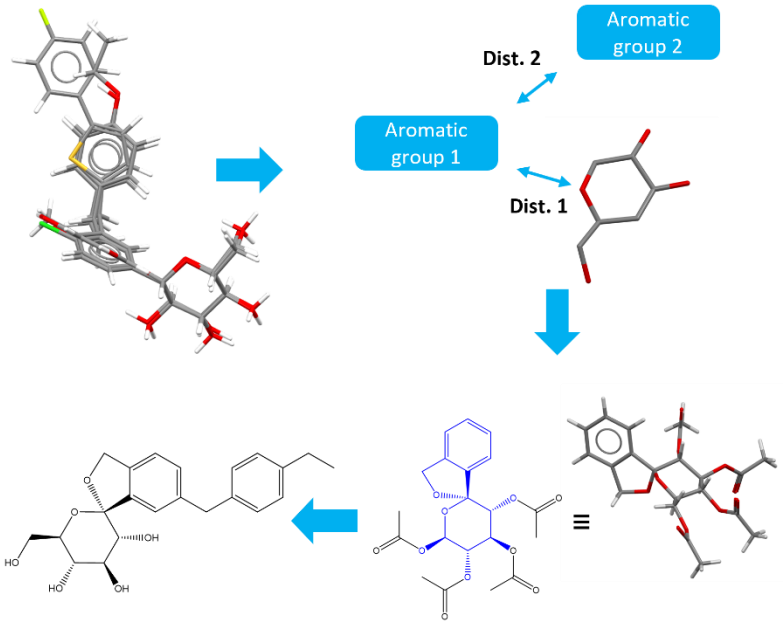


Figure 3. *From the superimposition model of the SGLT2 inhibitors to CSG452, tofogliflozin, SGLT2-selective inhibitor (DOI: 10.1021/jm300884k).*

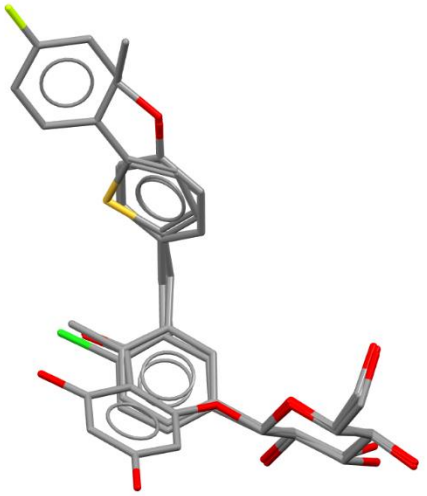


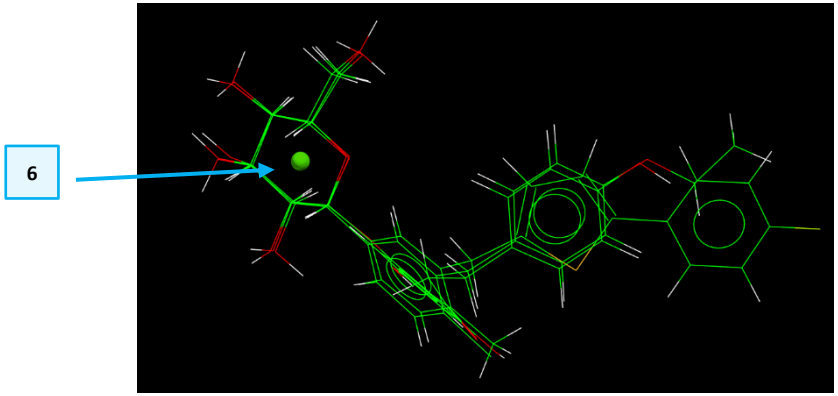
Figure 4. *Molecular overlay of three SGLT2 inhibitors obtained using CSD Ligand Overlay.*

Viewing the query overlay

1. Start CSD-CrossMiner by clicking on the CSD-CrossMiner icon:
2. Download and unzip *CROSS-001.zip*, which contains the file "molecular_overlay.mol2". Open this file by selecting *File > Load Reference* from the main menu. This loads the overlay of three known SGLT2 inhibitors generated using the CSD Ligand Overlay tool available from the CCDC (See CSD Ligand Overlay).
3. By default, only the *donor* (blue) and *acceptor* (red) features associated with the loaded reference molecules are shown in the 3D view. The displayed features are represented in the 3D view as small translucent spheres and are ticked in the *show in reference* column in the ***Pharmacophore Features*** window.

Note that, if a different choice of displayed features was made during previous CSD-CrossMiner sessions, those features (if present in the reference molecule) will be displayed instead.

4. In the **Pharmacophore Features** window, uncheck the default displayed features by unticking the corresponding check-boxes in the *show in reference* column.
5. Check the **ring_non_planar** check-box. This will show the non-planar ring features present in the overlay (a solid green sphere).
6. Here, the non-planar ring feature corresponds to a glucose ring that is essentially conserved in the three SGLT2 inhibitors used in the overlay.
7. Disable the **hydrogens** tick-box in the *Show:* toolbar, this will hide the hydrogen atoms of the matched hits in the 3D view.



Creating a new substructure to a feature definition

Using the *ring_non_planar* feature to create the pharmacophore query will result in hits with different non-planar rings that match the glucose location.

However, from the overlaid structures, we know that among the non-planar rings, it is the glucose that is conserved in the SGLT2 inhibitors.

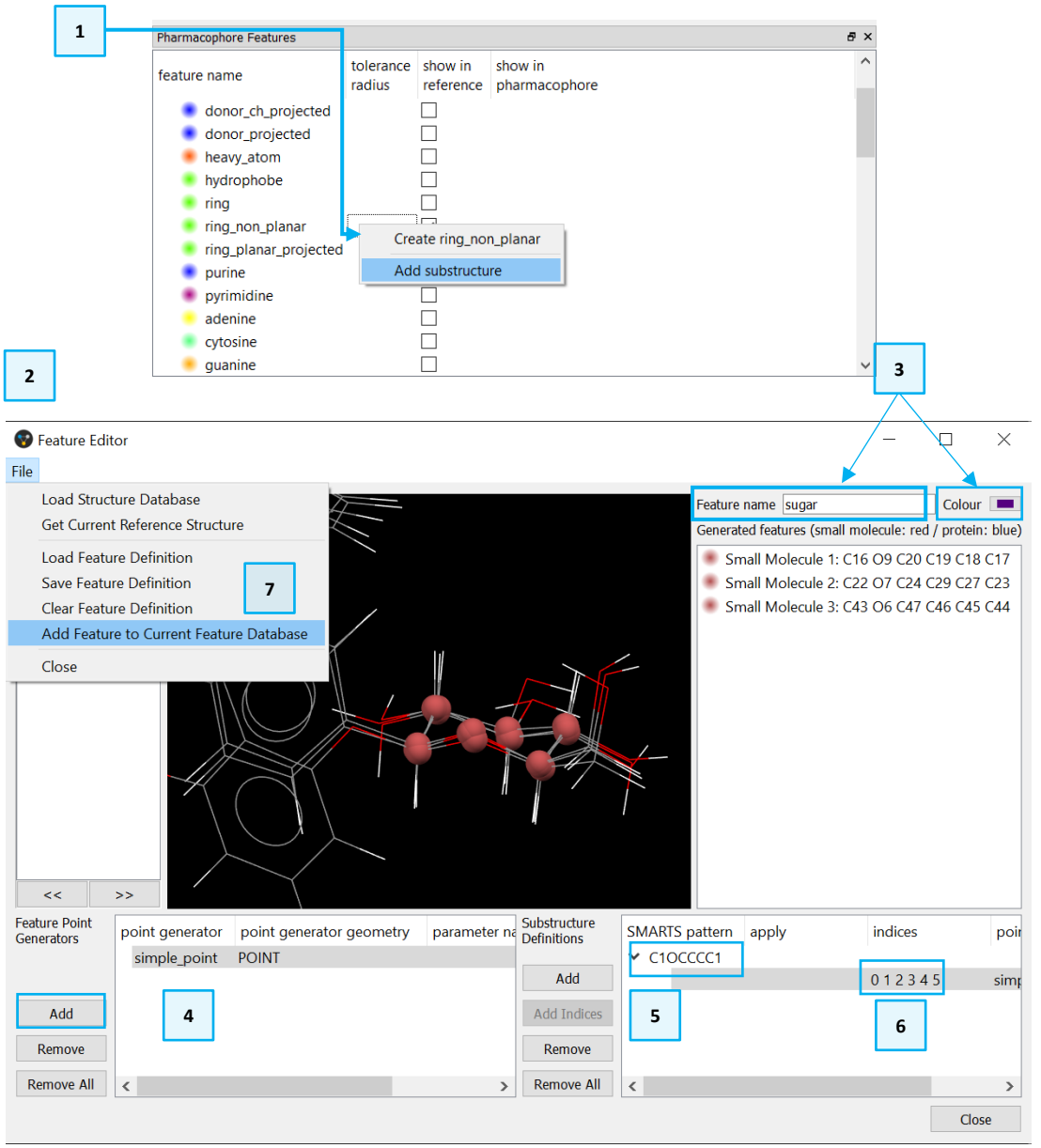
Therefore, we want to discriminate between all non-planar rings, and find hits with only the glucose ring. To do so, we can use CSD-CrossMiner to easily create a glucose feature on-the-fly and then use it to investigate the loaded database.

1. Right-click in the **Pharmacophore Features** window and click on **Add substructure**.
2. This will present the *Feature Editor* window.
3. Type “sugar” in the *Feature name* box and click the **Colour** button to specify a colour other than white (here we use dark purple).
4. Under *Feature Point Generators*, click **Add** to create a *simple point* feature.
5. Under *Substructure Definitions*, click the **Add** button. This creates a placeholder *SMARTS* string. Edit the *SMARTS pattern* by double clicking on **[*]**, replacing it with “C1OCCCC1” and then pressing Enter. The inserted pattern represents the SMARTS code of the sugar rings in the reference molecular overlay.
6. Double click on the *0* in the *indices* column and type “0 1 2 3 4 5” then press Enter to define a point based feature that will ensure that all six atoms in the glucose ring are included in the substructure.

The new feature definition is shown in the 3D display of the *Feature Editor* window as red (Small Molecule) feature points and listed in the right-hand panel under *Generated features* (small molecule: red) of the *Feature Editor* window.

Note that this list is associated with the three overlaid structures displayed in the 3D display.

7. In the **Feature Editor** window menu select *File > Add Feature to Current Feature Database* and then click **Close** to close the **Feature Editor** window. This will make the newly created feature available in this session.



Creating the pharmacophore search

1. The *sugar* feature is now displayed in the 3D view as translucent coloured sphere and ticked in the **Pharmacophore Features** window (scroll to the bottom of the **Pharmacophore Features** window to find it).

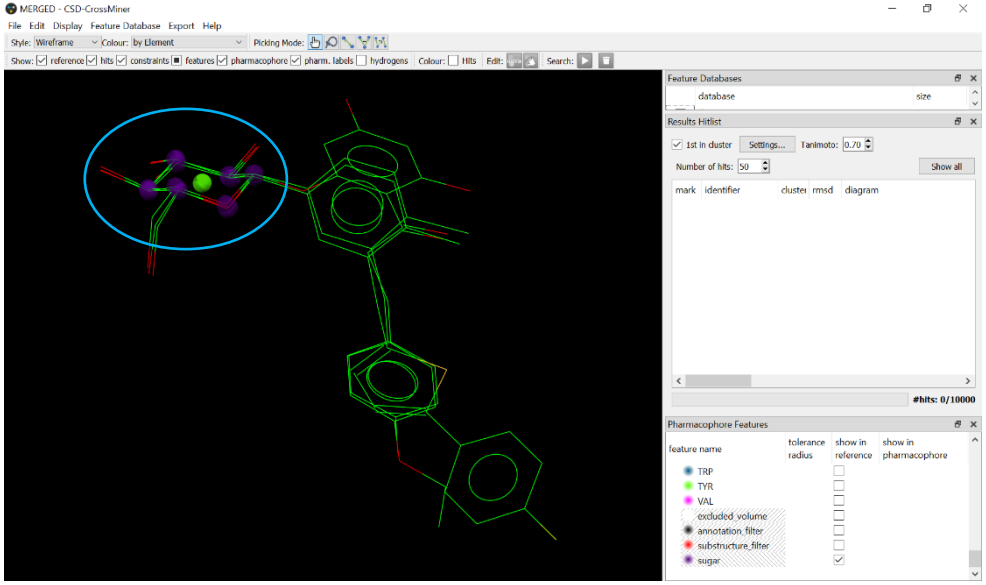
In the **Pharmacophore Features** window, you will notice a diagonal hatching on the *sugar* feature name; this indicates that the feature has not been pre-calculated and therefore, the loaded feature database has not been indexed with this feature definition.

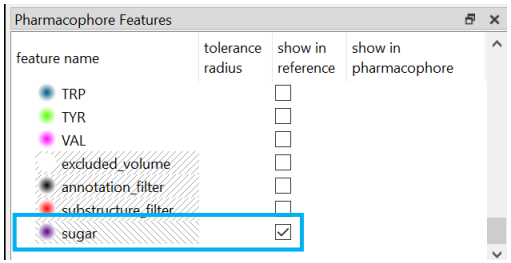
2. Right-click on one of the sugar spheres in the 3D view and select one of the **Add sugar** menu items. Note that three choices are available, as the reference file contains three sugar rings overlaid. Click on one of the options to create a sugar pharmacophore point.

3. Change the radius of each sugar pharmacophore base point to reduce the uncertainty in the position of the ring; this will ensure that the sugar ring in the search is very localised.

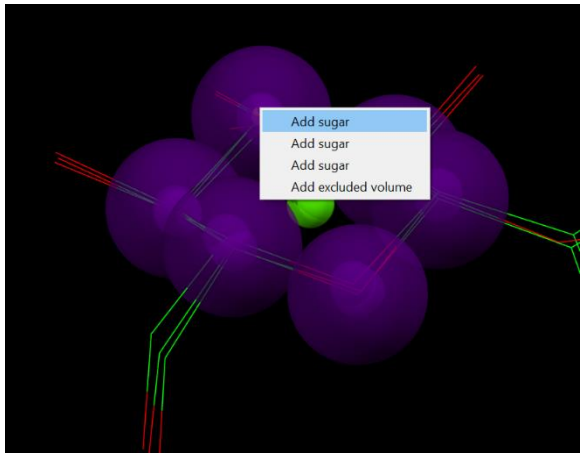
   By default, the sphere radii of the pharmacophore points are set to 1.0 Å; change them to 0.4 Å by double-clicking on the *tolerance radius* of each pharmacophore sugar point in the *Pharmacophore Features* window.

1

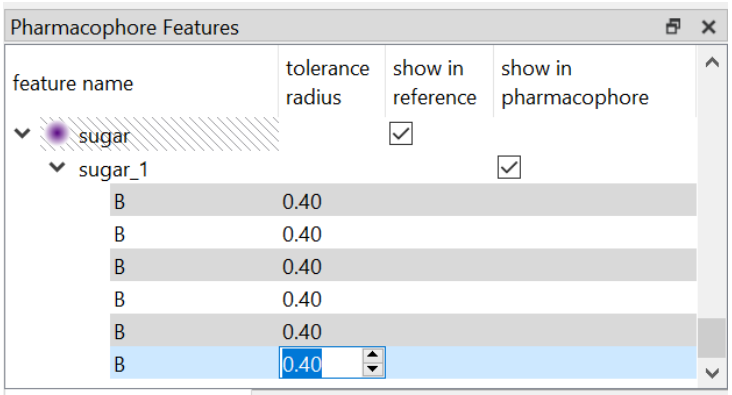




2



3



The overlay model shows that having a properly aligned sugar ring, two aromatic moieties, and special distances between two pairs of pharmacophore points are important in order to achieve high activity. We can make the pharmacophore search more selective by including the phenyl rings in the pharmacophore query.

For this workshop, we will include only the phenyl ring adjacent to the sugar ring into the pharmacophore query.

4. Tick the *ring_planar_projected* feature tick-box in the **Pharmacophore Features** window to display all the planar ring features present in the reference overlay in the 3D view.

Note that there are two ring_planar_projected features in the 3D view adjacent to the glucose ring (because each ring_planar_projected can have two projections, you will have four options in the feature right-click menu).

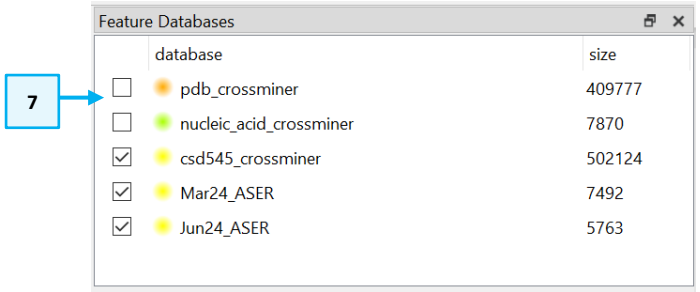
Right-click on the *ring_planar_projected* green sphere pointed in the figure on the right and click on **Add ring_planar_projected**.

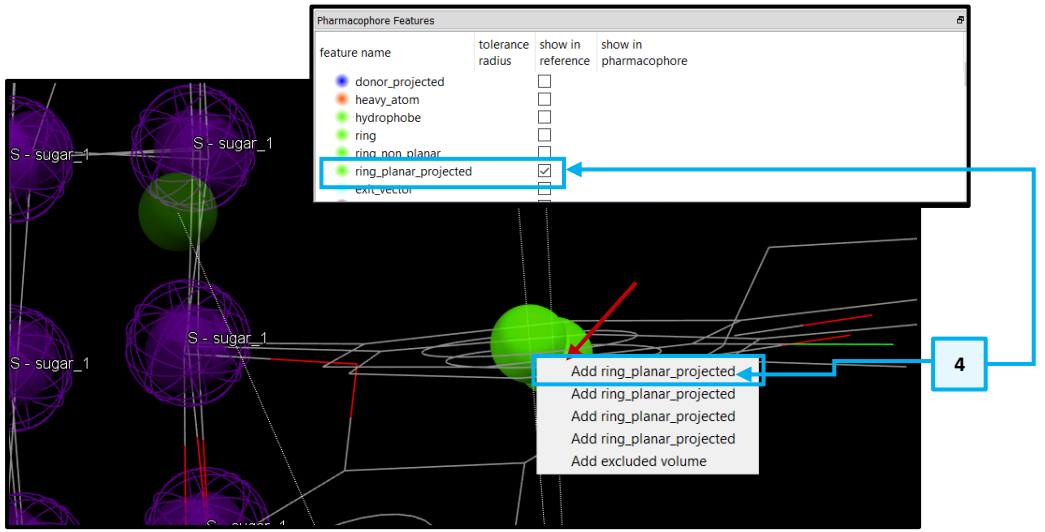
5. Although the aromatic moiety adjacent to the sugar ring is conserved in the SGLT2 inhibitors, its location is not certain. To make the pharmacophore query less restricted on the location of the aromatic ring, change the sphere radius of the *ring_planar_projected* pharmacophore point in the *Pharmacophore Features* window from 1.0 Å to 1.3 Å, by double-clicking on '*B'* and use the spin-box under *tolerance radius.*

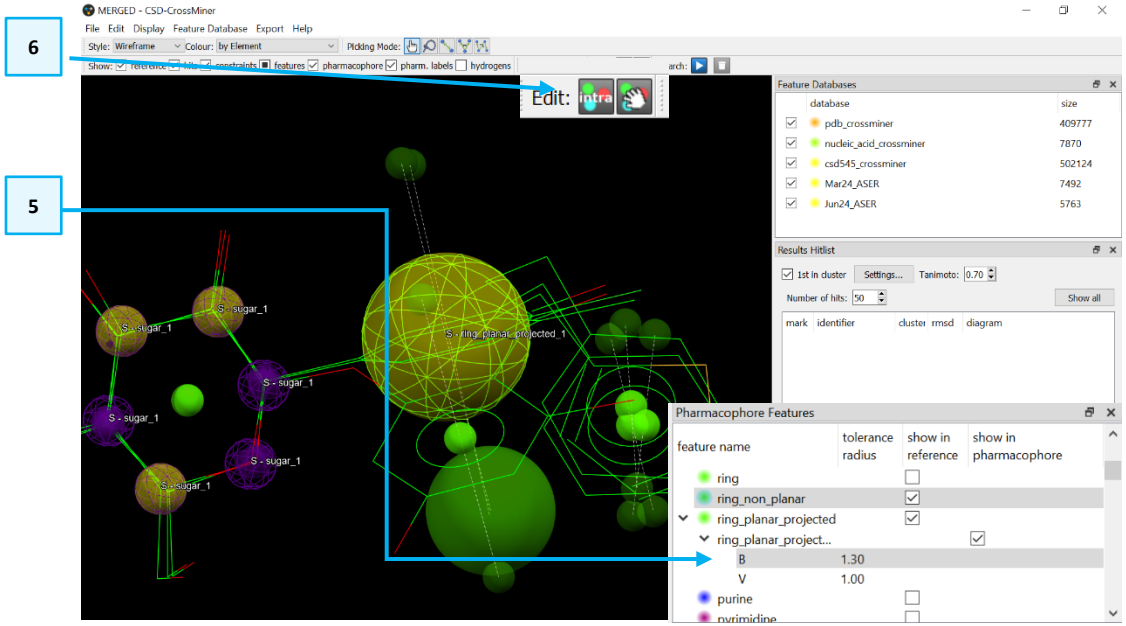
6. Finally, we want to find structures where all pharmacophore points belong to the same molecule. To add all intramolecular constraints, click the **Intra** button 

of the *Edit:* toolbar.

7. Deselect the PDB and nucleic acid subsets in the database, by unticking the *pdb_crossminer* and *nucleic_acid_crossminer* tick-boxes in the *Feature Databases* window) before searching. This will speed up the search, as the hits of most relevance in this search are in the CSD.





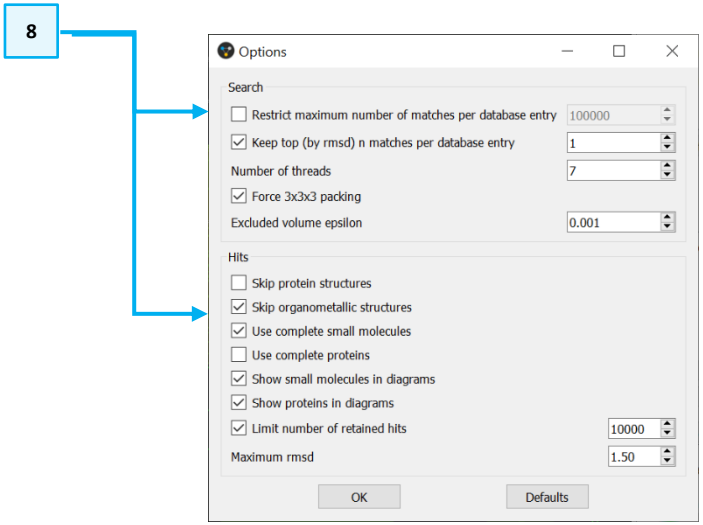


8. To reduce redundancy of the solutions we will reduce the number of matches per database entry in the pharmacophore search option menu. The pharmacophore search options can be accessed from *Edit > Options* in the CSD-CrossMiner top-level menu. From here reduce the **Keep top (by rmsd) n matches per database entry** from 5 to 1 and then activate the **Skip organometallic structures** to exclude hits that contain at least one transition metal, lanthanoid, actinoid, or any Al, Ga, In, Tl, Ge, Sn, Pb, Sb, Bi, Po.

9. You can now run the search by clicking the **Play** button  in the *Search:* toolbar.

<u>Note that:</u>

- A different choice of sugar atoms and/or phenyl ring can provide different results from the ones discussed below.
- Because the pharmacophore query includes non-indexed feature definitions, the pharmacophore search can require several minutes to complete.



Inspecting the results

1. To inspect the results, ensure the **Results Hitlist** browser is shown. The *Display > Toolbars* from the top-level menu will show all displayed windows and toolbars. From here, hide the **Pharmacophore Features** and **Feature Databases** windows by unticking the tick-boxes. This will provide additional space for the **Results Hitlist** browser. *Tip*: you may wish to untick *show in reference* and *show in pharmacophore* before hiding the Pharmacophore Features toolbar.

2. Edit the *Number of hits* in the **Results Hitlist** to show more than 100 (default) number of hits. To do so, double-click on the *Number of hits*: spin-box and write the number of hits you want to have displayed (400 in this case). This will ensure that all matched hits are displayed.

3. When **1st in cluster** check-box is selected in the **Results Hitlist**, the matching hits are clustered based on the *Tanimoto* value showed in the *Tanimoto:* spin-box. In this case, both the **Results Hitlist** and the 3D view show only the cluster representatives of those similar groups. In this workshop we are interested in similar molecular hits, so it is useful to uncheck the **1st in cluster** check-box in the **Results Hitlist** window.

4. If you haven't done before, disable the *hydrogens* tick-box in the *Show:* toolbar, this will hide the hydrogen atoms of the reference overlay and those of the matched hits.

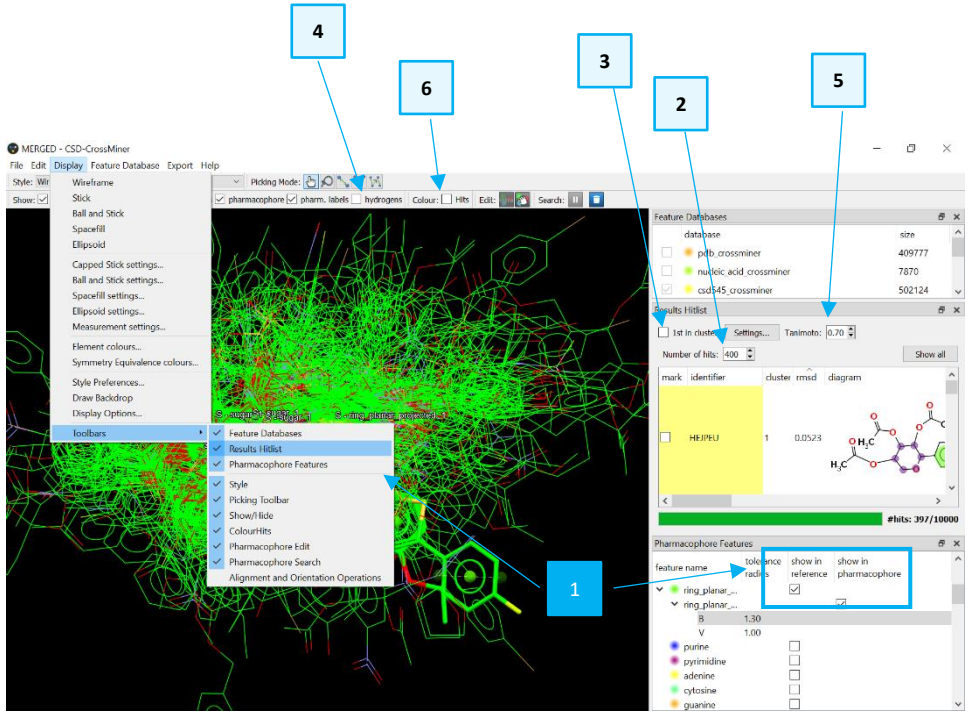
5. Sort the hits by their *RMSD* by clicking on the *rmsd* column in the *Results Hitlist* window.

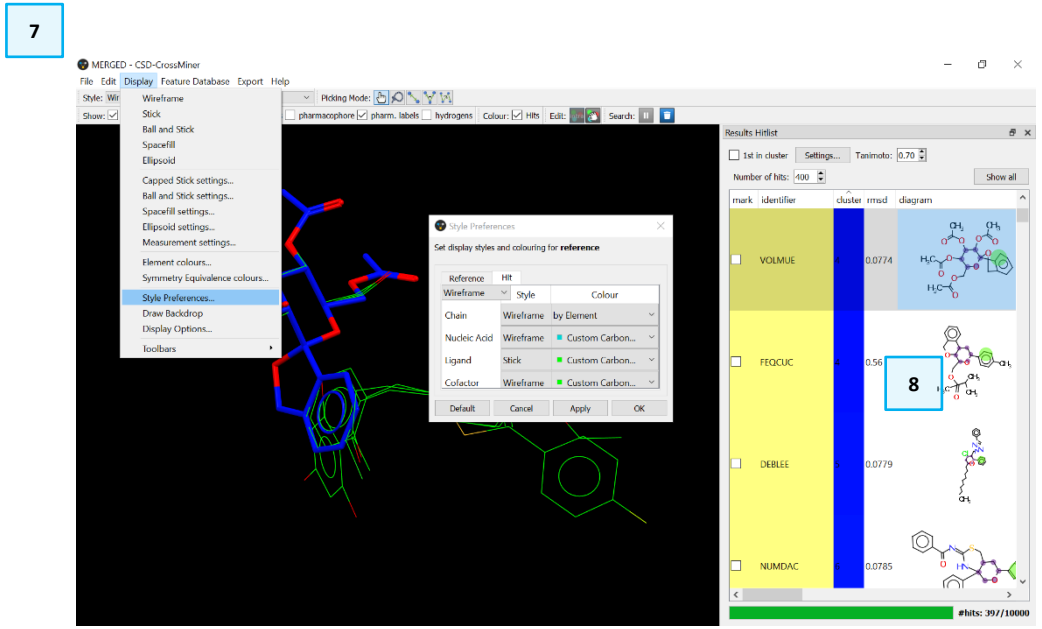
6. Tick the ***Colour: Hits*** check box to colour clusters by rainbow.

7. You can change the display style of the reference and hit by going to *Display > Style Preferences…*. In the *Ligand* row of the **Reference** and **Hit** tabs, choose *Wireframe* and **Stick**, respectively.

Note that the results could be slightly different depending on the choice of the planar ring pharmacophore point and on the choice of the sugar ring pharmacophore point.

8. One of the first 10 hits in the **Results Hitlist** window (hits with low rmsd) is VOLMUE (shown here in yellow) which includes a spiroketal moiety.



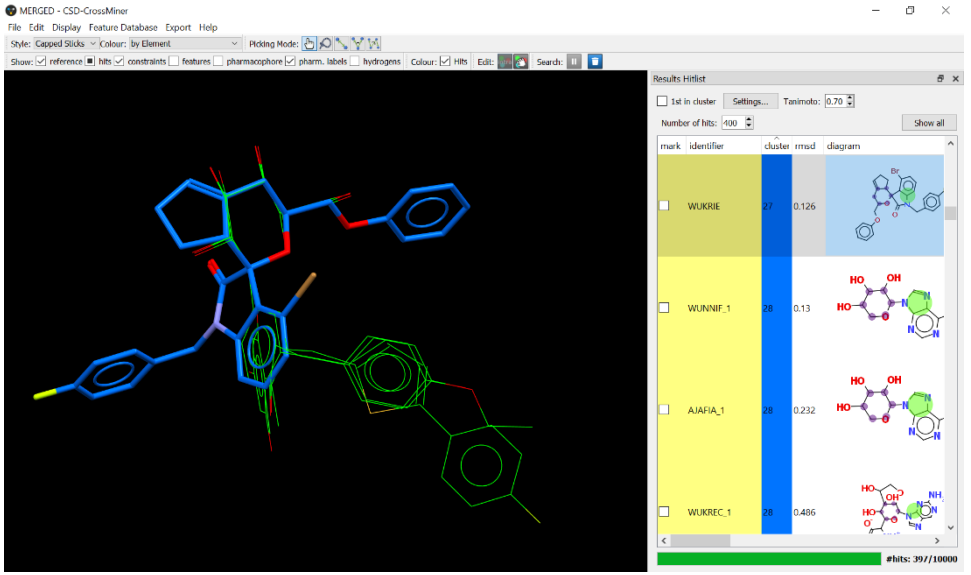


Interestingly, the VOLMUE molecule is the CSD entry that inspired the use of a spiroketal chemistry in the original paper.

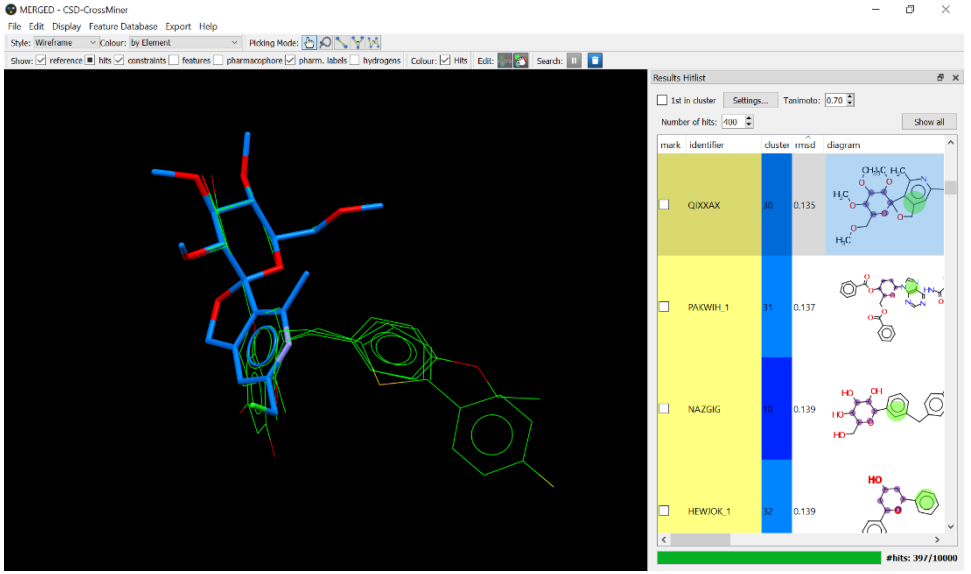
9. Navigating through the **Results Hitlist** browser you will find some other interesting hits:
    a. WUKRIE (shown here in light blue) shows another example of a spiro chemistry.
    b. QIXXAX (shown here in purple) shows a similar spiro motif but with a pyridine ring.
10. Later hits include BAGDIW which was also identified as a hit compound in the original paper.

Inspect potential alternative scaffolds that might allow patent breaking or enhancement of affinity and selectivity.

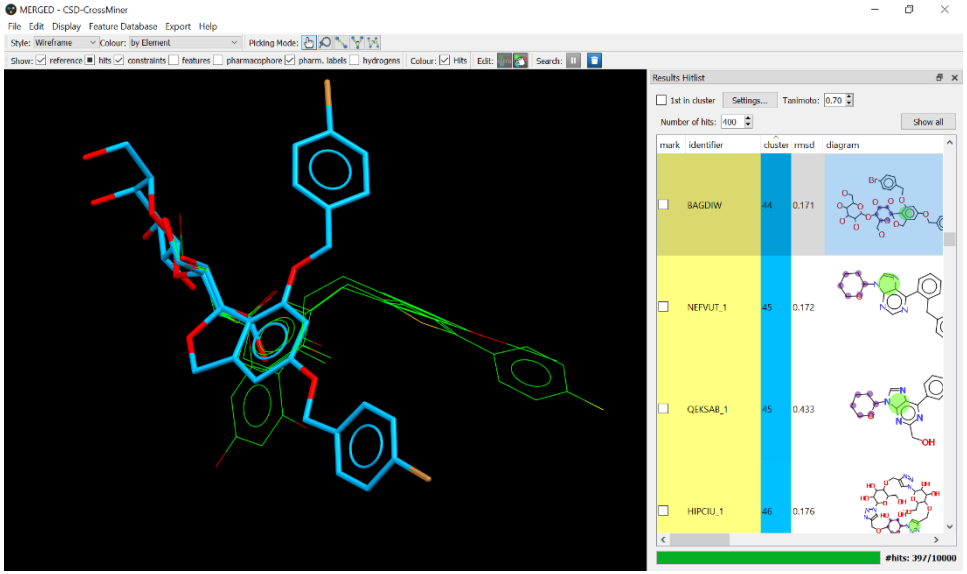
9a.



9b.



10

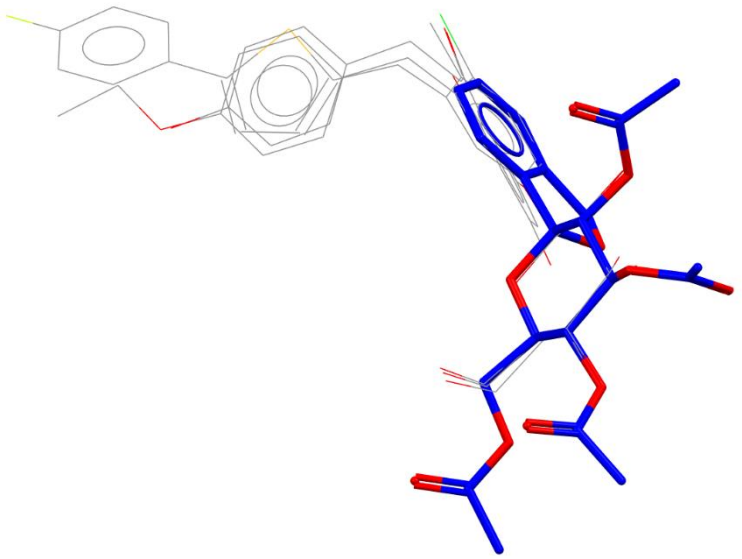


Conclusion

This example has demonstrated that creating a custom pharmacophore in CSD-CrossMiner from a reference molecule(s) can help to retrieve more relevant results when a specific moiety is thought to be important to pharmacological activity. In this case, the search successfully recovered molecules discussed in a publication where spiroketal compounds were investigated, by adopting the same search protocol.

Exercises

- Experiment with more strongly defined sugar rings by using more elaborate SMARTS substructure definitions, or by adding exocyclic acceptor features to the pharmacophore.
- Explore more extensively the effect of the radii used in the pharmacophore query on the hits generated.
- Try the effect of changing the location of the planar ring projected pharmacophore.
- Try adding the additional *hydrophobic* planar ring in the overlay to see if you can find any SGLT2 inhibitors in the CSD.



A spiroketal molecule, CSD Entry VOLMUE, overlaid with a selection of SGLT2 inhibitors.

Summary

This workshop shows a way of using a new substructure feature in a CSD-CrossMiner pharmacophore search to focus on closely related structures.

It is relatively simple to mine similar compounds in this way and quickly assess the match of the hits generated. While such a search is possible in other CSD applications, such as ConQuest, the query in these tools is more challenging to create; thus CSD-CrossMiner provides a more convenient method for interrogating possibilities in the CSD.

You should now be able to:

- Create a custom substructure using a SMARTS pattern
- Use a custom substructure as a pharmacophore feature in a CSD-CrossMiner search

For your reference, you can find the user manual at this link.

Next steps

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

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

Feedback

We hope this workshop improved your understanding of CSD-CrossMiner and you found it useful for your work. As we aim to continuously improve our training materials, we would love to get your feedback. Click on this link to a survey (link also available from workshops webpage), it will take less than 5 minutes to complete. The feedback is anonymous. You will be asked to insert the workshop code, which for this self-guided workshop is CROSS-001. Thank you!

Glossary

Hydrogen Bonds

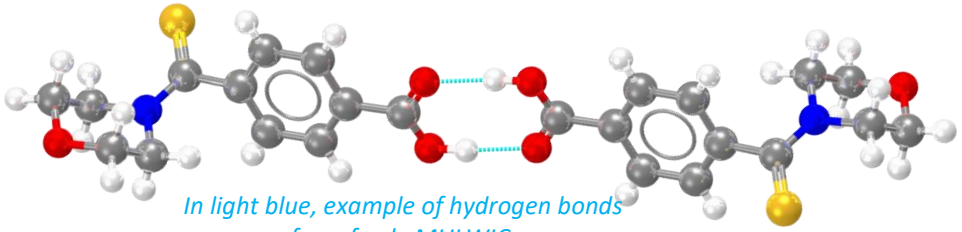
Hydrogen bonding occurs between donor-acceptor interactions precisely involving hydrogen atoms. The H-bonds interactions are classified as: strong (mostly covalent), moderate (mostly electrostatic) and weak (electrostatic). Their strength is observed to be between 12 and 30 kJ/mol.

Hydrogen Bond Donor/Acceptor

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

Hydrophobic/hydrophobe

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



In light blue, example of hydrogen bonds for refcode MULWIC.

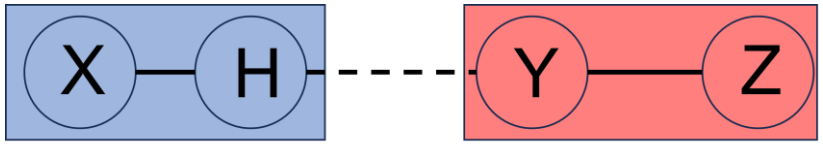
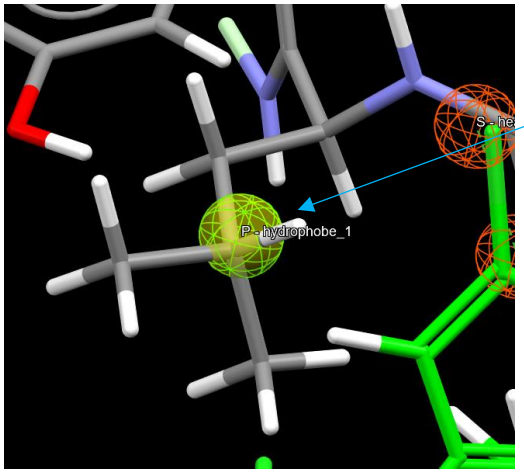


Illustration of a hydrogen bond interaction with between hydrogen bond donor X–H and hydrogen bond acceptor Y–Z.



Hydrophobe pharmacophore point.

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

Root Mean Square Deviation (RMSD)

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

SMARTS string

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

Spiroketal

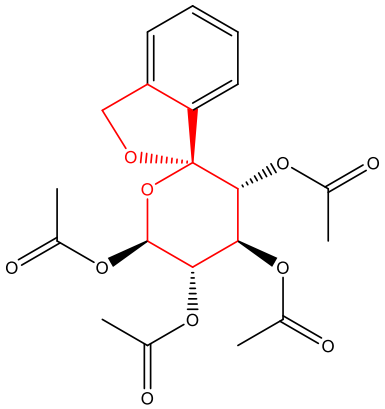
A subclass of spiro compounds. Two heterocyclic rings are fused through a common carbon, the spiro carbon. Spiroketals have an oxygen atom alpha to the spiro carbon in each ring (an acetal group).

Tanimoto coefficient

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

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

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



CSD entry VOLMUE with the spiroketal substructure in red.

CSD-CrossMiner Terminology

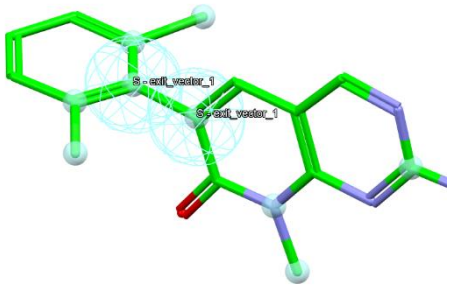
Exit vector
A two-point feature that represents a single, non-ring bond between two heavy atoms features; and it will be represented as two mesh spheres. In the case of CSD-CrossMiner, directionality in an exit vector does not matter.

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

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

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

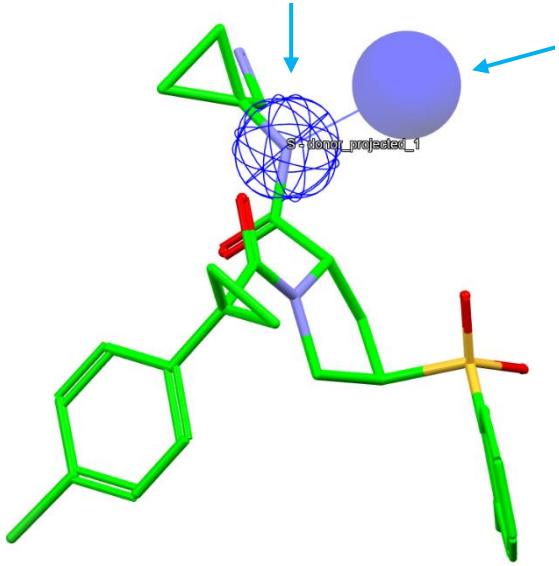
Feature database
A database containing the structures from the structure database, indexed with a set of feature definitions provided by CSD-CrossMiner and any additional features defined by the user. This is the database that CSD-CrossMiner uses to perform the actual 3D search against a pharmacophore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



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

Base point (the heavy atom of the donor group)

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



A molecule with a donor_projected pharmacophore point defined.

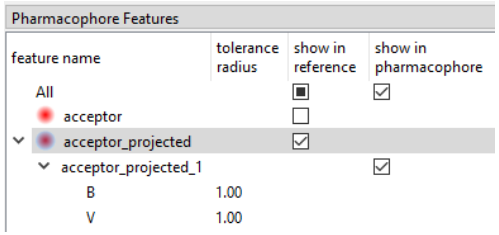
Features and Pharmacophore Representation

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

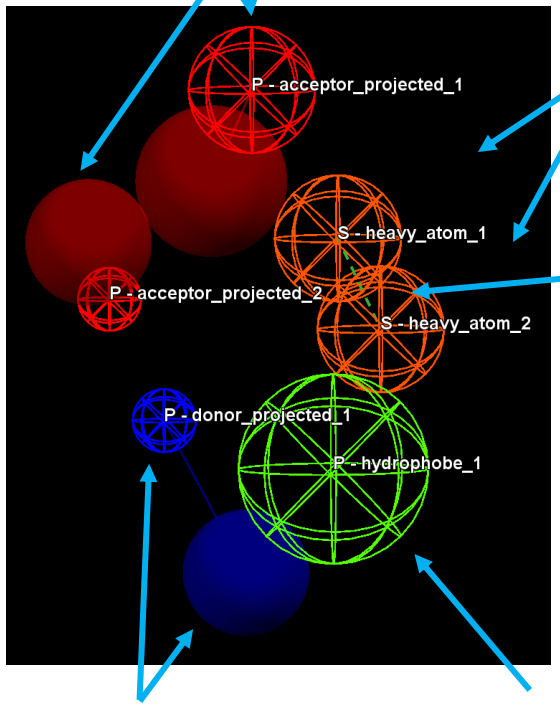
- **P**: Protein pharmacophore point
- **S**: Small molecule pharmacophore point
- **A:** Either a small molecule or protein pharmacophore point
- **Dashed line**: intra and intermolecular constraints. Constrained features must belong to either the same molecule as each other (*intra*, dashed green line) or different molecules (*inter*, dashed red line).
- **Mesh sphere**: the actual feature itself, where the sphere size represents the radius of tolerance of the pharmacophore point.
- **Solid sphere**: the projected virtual point to represent the directionality of e.g. a hydrogen bond acceptor/donor. A feature can have more than one projected point. For example, a H bond acceptor can have multiple potential lone pair preferred projections.

Note that the colour coding of the pharmacophore points is defined in the *Pharmacophore Features* browser; e.g. hydrophobe features are green, hydrogen bond acceptors are red, and so on.

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



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



Small molecule (S): heavy atoms

Intramolecular constraint

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

Protein (P): hydrophobe feature