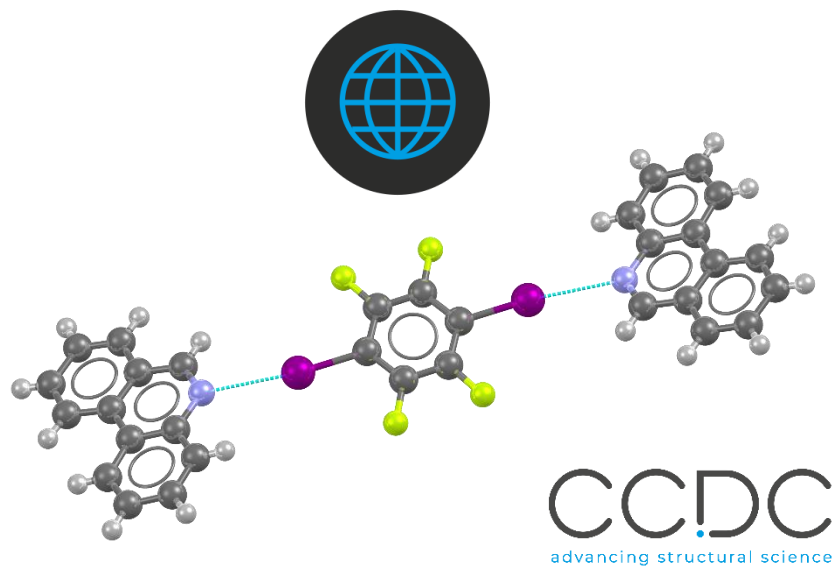


# Similarity Searching with WebCSD (WCSD-003)

Developed using WebCSD version 1.9.46, Feb 2025



## Table of Contents

Introduction.....	2
Learning Outcomes .....	2
Pre-required Skills .....	2
Materials.....	2
Example 1: Conducting a similarity search for pharmaceutically active molecules in the CSD .....	3
Conclusion .....	5
Summary .....	6
Next Steps.....	6
Feedback .....	6
Glossary .....	7
Activating WebCSD.....	9
Basics of the WebCSD Sketcher.....	10
Explanation of 3D parameters.....	11

## Introduction

This workshop will introduce you to similarity searching using WebCSD.

WebCSD allows you to search for a specific structure and to find the context of the structure in relation to its originality and/or comparison to other known molecules. WebCSD supports text/numeric, structure, unit cell, and formula searching across the CSD, with the benefit of returning the most up-to-date possible results as structures are added to the CSD. Original deposited data can be obtained and exported for further analysis.

Before beginning this workshop, please ensure that you are signed in to WebCSD with a user account connected to a valid CSD license (see [here](#)) or that you are on a network configured to access the CSD.

### Learning Outcomes

In this workshop you will explore the structure searching tools available from WebCSD for searching the CSD online. After completing this workshop, you will be able to:

- Conduct a similarity search based on a molecular structure.

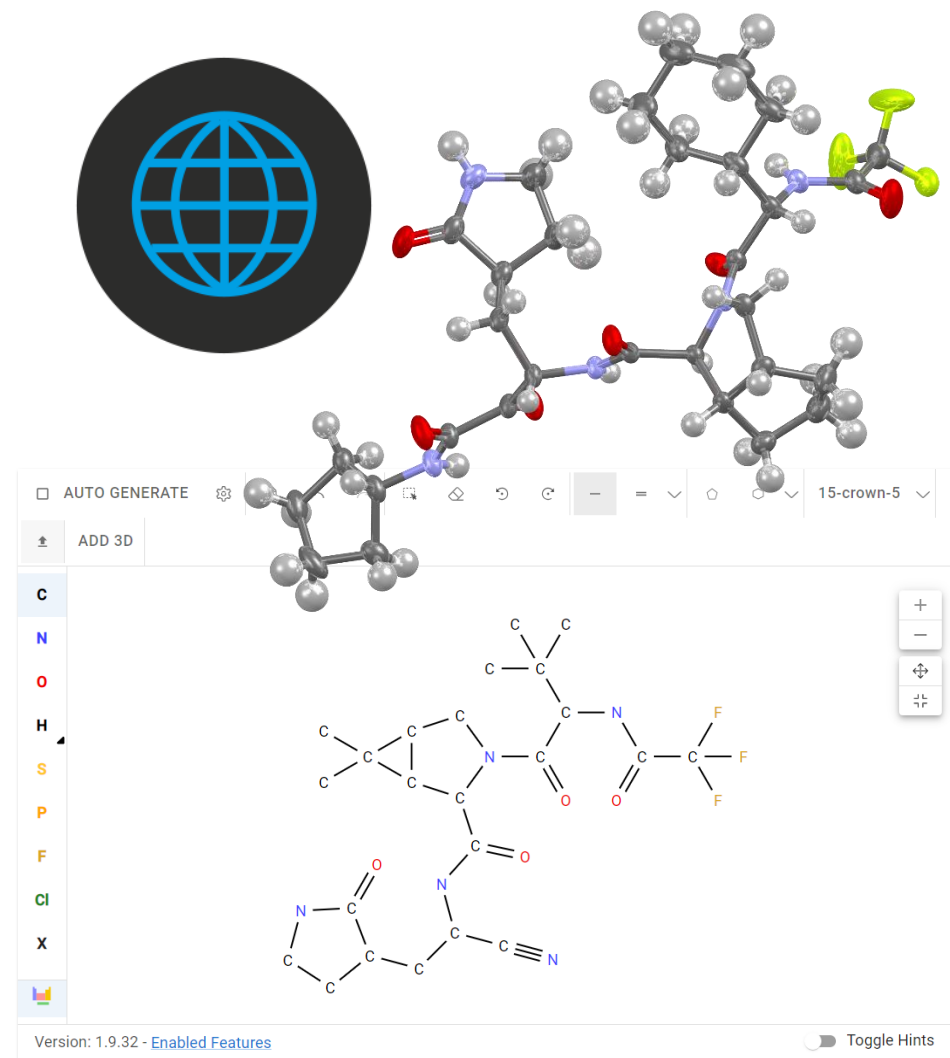
This workshop will take approximately **15** minutes to complete. The words in *Blue Italic* in the text are reported in the [Glossary](#) at the end of this handout.

### Pre-required Skills

There are no pre-required skills for this workshop.

### Materials

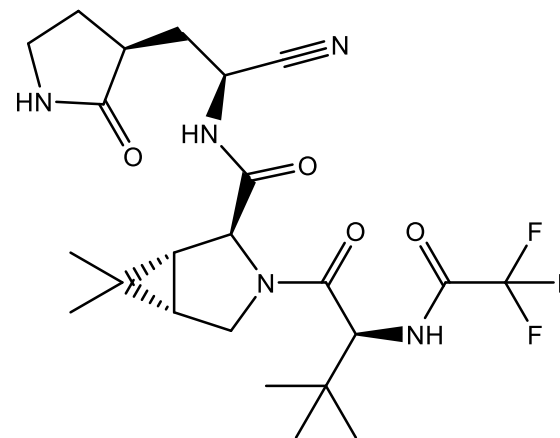
Download the MOL file for nirmatrelvir [here](#).



## Example 1: Conducting a similarity search for pharmaceutically active molecules in the CSD

WebCSD offers the ability to conduct a [similarity search](#) using a molecular fingerprint generated from a structure, with the similarity evaluated based on the [Tanimoto coefficient](#). Hits are returned if the Tanimoto coefficient is 0.7 or above.<sup>1</sup>

The drug nirmatrelvir is an orally-active C3-like protease inhibitor which is a component of Paxlovid, a combination drug developed by Pfizer for the treatment of SARS-CoV-2 infection.<sup>2</sup> The structure of Nirmatrelvir was reported and added to the CSD in 2023 (CSD entry ZIVMEA). In this example, we will uncover structurally similar compounds which may potentially also have biological activity, using a WebCSD Similarity search.



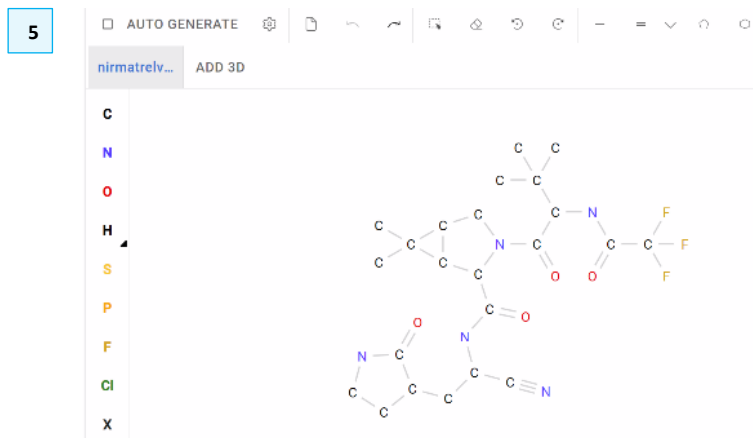
The molecular structure of nirmatrelvir (CSD refcode ZIVMEA).

1. Open a web browser and navigate to the [Access Structures](https://www.ccdc.cam.ac.uk/structures/) webpage (<https://www.ccdc.cam.ac.uk/structures/>).
2. Click on the **Structure Search** tab to bring up the sketcher. You can drag the periodic table to reposition it in a convenient area.
3. Rather than drawing the structure of Nirmatrelvir, we will upload a MOL file which has been prepared for this workshop using the program ChemDraw. Download the MOL file from [here](#) if you have not already done so. *Tip: If you are exporting your own file from ChemDraw, it is not necessary to explicitly draw hydrogen atoms, they will be automatically added during a similarity search.*

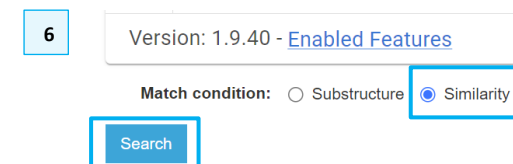
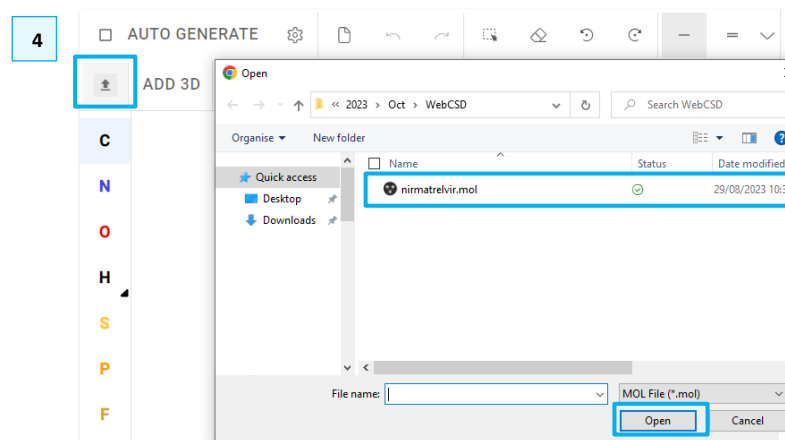
<sup>1</sup> I. R. Thomas, I. J. Bruno, J. Cole, C. F. Macrae, E. Pidcock and P. A. Wood, *J. Appl. Crystallog.*, 2010, **43**, 362 – 366.

<sup>2</sup> S. M. R. Hashemian, A. Sheida, M. Taghizadieh, M. Y. Memar, M. R. Hamblin, H. B. Baghi, J. S. Nahand, Z. Asemi and H. Mirzaei, *Biomed. Pharmacother.*, 2023, **162**, 11467 – 11475.

- In the top toolbar in the sketcher window, click the upwards arrow to launch the file explorer and select *nirmatrelevir.mol* and click **Open**.
- The cursor will now show the outline of the molecule. Move to a suitable position in the sketcher and click once to add the molecule.



- Ensure that the **Similarity** radio button is turned on and click **Search**.
- A new page will load showing the results, listed with their refcode and similarity score on the left-hand side, and information about the currently selected entry. Scroll through the results and notice how the information and viewer is updated as you do so.



7

The screenshot shows the search results page. The search is complete with 25 results found. The results table is as follows:

Database Identifier	Deposition Number	Similarity Score	More Info
<input checked="" type="checkbox"/> ZIVMEA	2234935	1.000	
<input checked="" type="checkbox"/> ZIVMEA01	2234936	1.000	
<input checked="" type="checkbox"/> ZIVMCK	2249784	0.927	
<input checked="" type="checkbox"/> QAXHEP	1518587	0.783	
<input checked="" type="checkbox"/> NOTJOV	1222796	0.774	
<input checked="" type="checkbox"/> NOTJOV01	1222787	0.774	
<input checked="" type="checkbox"/> DIDVEV	2251675	0.748	
<input checked="" type="checkbox"/> LOFRAA	669816	0.747	
<input checked="" type="checkbox"/> MIHDEN	635857	0.728	
<input checked="" type="checkbox"/> PAWJON	272987	0.728	
<input checked="" type="checkbox"/> PAWJUT	274268	0.728	
<input checked="" type="checkbox"/> IDEZEA	2219883	0.717	
<input checked="" type="checkbox"/> IVOXOG	235208	0.710	
<input checked="" type="checkbox"/> TUPJIZ	1407311	0.705	
<input checked="" type="checkbox"/> CAZKEH	265525	0.701	

The detailed view for the selected entry (ZIVMEA) shows the following information:


ZIVMEA: N-[1-cyano-2-(2-oxopyrrolidin-3-yl)ethyl]-6,6-dimethyl-3-[3-methyl-N-(trifluoroacetyl)valyl]-3-azabicyclo[3.1.0]hexane-2-carboxamide  
 Space Group: P 2<sub>1</sub> 2<sub>1</sub> 2<sub>1</sub> (19), Cell: a 9.2108(6)Å b 15.0939(11)Å c 18.0815(12)Å, α 90° β 90° γ 90°

The 3D viewer shows a ball-and-stick model of the molecule. The chemical diagram shows a 2D representation of the molecule.

8. There should be at least 29 hits retrieved (your value may be significantly higher as WebCSD is updated every minute). Note that hits are only returned when the similarity score is 0.7 or above. CSD entries in the ZIVMEA [refcode family](#) have perfect scores of 1.000 because they are identical to the query. Scroll through the results and study the *Chemical details* section to look for *Bioactivity* information. You should find that a couple of results are flagged as bioactive; one of them, CSD entry DIDVEV, is in fact another SARS-CoV-2 protease inhibitor. You can click the DOI link in the *Associated publications* section to explore this result further.

## Conclusion

Similarity searching is a useful way to search the CSD for compounds which are likely to have related properties. In this example we have uncovered an example of a compound in the CSD with similar pharmacological activity to a currently marketed SARS-CoV-2 treatment drug by conducting a similarity search.

Associated publications	
	Xiaoxin Chen, Xiaodong Huang, Qinhai Ma, Petr Kuzmič, Biao Zhou, Jinxin Xu, Bin Liu, Haiming Jiang, Wenjie Zhang, Chunguang Yang, Shiguan Wu, Jianzhou Huang, Haijun Li, Chaofeng Long, Xin Zhao, Hongrui Xu, Yanan Sheng, Yaoting Guo, Chuanying Niu, Lu Xue, Yong Xu, Jinsong Liu, Tianyu Zhang, James Spencer, Wenbin Deng, Shu-Hui Chen, Xiaoli Xiong, Zifeng Yang, Nanshan Zhong, <i>bioRxiv</i> , 2023, DOI: <a href="https://doi.org/10.1101/2023.03.09.531862">10.1101/2023.03.09.531862</a>
	<b>8</b>
Chemical details	
<b>Formula</b>	C <sub>31</sub> H <sub>44</sub> F <sub>3</sub> N <sub>5</sub> O <sub>9</sub>
<b>Bioactivity</b>	alpha-ketoamide-based peptidomimetic inhibitor of SARS-CoV-2 main protease

## Summary

In this workshop, you have explored the similarity search feature in WebCSD, starting from a known molecule. You should now be confident in:

- Importing a mol file in WebCSD
- Setting up a Similarity search in WebCSD.

For further information, and tips on how to make the most of WebCSD, see the [FAQs relating to WebCSD](#) on our website.

## Next Steps

You might like to explore some of the search results further using the structure visualisation software Mercury. You can find self-guided workshops on Mercury [here](#) and an on-demand training course [here](#).

## Feedback

We hope this workshop improved your understanding of *Similarity searching in WebCSD* and you found it useful for your work. As we aim to continuously improve our training materials, we would love to hear your feedback. Follow [the link](#) on the workshop homepage and insert the workshop code, which for this self-guided workshop is WCSD-003. It will only take 5 minutes and your feedback is anonymous. Thank you!

## Glossary

### MOL

A MDL Molfile (.mol) is a file format for holding information about the atoms, bonds, connectivity and coordinates of a molecule.

### Refcode family

The same substances are assigned the same 6 letter code plus additional 2 numbers. These substances are typically polymorphs, new determinations or refinements of the same substance, determinations at different temperatures or pressures. Stereoisomers, different solvates or co-crystals are assigned to different refcode families.

### Similarity search

A search based on the calculation of similarity between molecules in two dimensions determined by chemical features including atom types, bond types, and bonded paths through the molecule referred to as a 'fingerprint'. The fingerprint is compared with all the unique connectivities in the CSD and the similarity is evaluated with a similarity coefficient. See also [Tanimoto coefficient](#).

### SMARTS

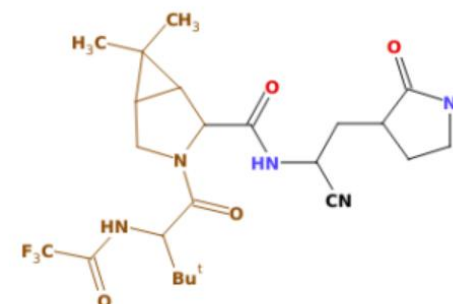
A way of describing a chemical substructure using letters, numbers and symbols. If you are unfamiliar with SMARTS strings, you can visualise them and learn more about the format with SMARTSviewer (<http://smartsview.zbh.uni-hamburg.de/>).

### SMILES

Simplified Molecular Input Line Entry System; a chemical notation for describing the structure of chemical species using short strings.

### Substructure

A substructure is a part or section of a whole molecule.



Nirmatrelvir, CSD Entry ZIVMAE, with a substructure highlighted in brown.

**Tanimoto coefficient**

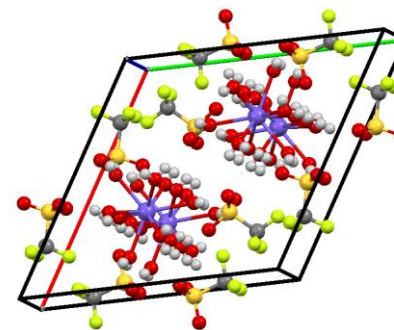
In a [similarity search](#), the Tanimoto coefficient is the ratio of the number of features common to both molecules to the total number of features, i.e.

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

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

**Unit cell**

The unit cell is the basic building block of a crystal, repeated infinitely in three dimensions.



The unit cell of CSD entry YEVROH (with contents).



## Activating WebCSD

To access the advanced WebCSD searching options, such as structure search and unit cell search, within the CSD web interface you will need to sign in to the site and connect a valid CSD licence. You can connect a CSD licence by either of the following methods:

- Accessing the CSD web interface from an IPv4 address registered to a CSD campus licence (we are unable to support IPv6).
- By entering a valid CCDc customer number and activation key once signed in. To do this you will need to:
  1. Go to My Account.
  2. Sign In or Register if necessary.
  3. Select "Activate WebCSD" under Licences on the right-hand side of the page.
  4. Enter your CCDc customer number and activation key.

The campus licence holder at your institution should be able to provide you with the CCDc customer number and activation key for your institution or they can provide us with the relevant IP addresses. If you are unsure of who this is or you are unsure if your institution has a licence then please contact us using our [enquiry pages \(https://www.ccdc.cam.ac.uk/contact-us/\)](https://www.ccdc.cam.ac.uk/contact-us/).

**1** My Account

Support and Resources About

**2** Sign in with your CCDc account

To access this additional functionality, please sign in here with a user account connected to a valid CSD licence

Username or Email

Password

Remember me?

Sign In Forgotten Username or Password

**3**

Profile

My Structures

My Subsets

Deposit

My licence portal

Security

Sign Out

Change Password

Change Email

Licences

Activate WebCSD

**4** Activate WebCSD Licence

Licence Customer Number \*

Licence Activation Key \*

Captcha

I'm not a robot

reCAPTCHA Privacy Terms

Activate

## Basics of the WebCSD Sketcher

The WebCSD sketcher is the interface for creating substructure and similarity searches in WebCSD.

In the following we will see some of the basics of using the sketcher to create search queries.

In the WebCSD sketcher, we find:

- Top left:** tools for automatic hydrogen and connection generation.
  - Top middle:** tools to create new sketches, to edit and select parts of existing sketches and options to undo/redo operations.
  - Top right:** tools to select bond types, simple ring drawing templates, and advanced templates.
- Second line from the top:** button to upload a structure from a MOL file directly into the sketcher, and button to access the 3D parameters menu.
- Left hand side toolbar:** quick access to elements and options for hydrogen placement, button to show/hide periodic the table. *Note: the periodic table is floating and can be repositioned using ||*
- Main window:** click here to add structure components (atoms and bonds) and shift + click on atoms to select them to define 3D parameters.
- Below sketcher:** options to select substructure or similarity search.
  - Below sketcher:** options: advanced options – input SMARTS here.

The screenshot shows the WebCSD Sketcher interface with several key components labeled:

- Auto generate H-atoms and bonds:** Points to the 'AUTO GENERATE' checkbox.
- Editing features:** Points to the undo, redo, and other editing icons.
- Structure drawing tools:** Points to the bond type and ring drawing templates.
- Upload MOL file:** Points to the 'ADD 3D' button.
- 3D parameters menu:** Points to the 'ADD 3D' button.
- Quick access to elements:** Points to the vertical toolbar on the left containing elements like C, N, O, H, S, P, F, Cl, X.
- Periodic table:** A floating periodic table is shown with element C highlighted. A note says 'Hold shift to select multiple elements or groups'.
- Select search type:** Points to the 'Match condition' section with radio buttons for 'Substructure' and 'Similarity'.
- Input SMARTS here:** Points to the text input field for SMARTS strings.

At the bottom, there is a 'Search' button and a 'Toggle Hints' option.

## Explanation of 3D parameters

3D parameter options are accessed from **ADD 3D**. To define a 3D parameter, you must select atoms. Shift + left mouse click to select multiple atoms.

- Distances require exactly two atoms.
- Angles require exactly three atoms, in the correct sequence.
- Torsions require exactly four atoms, in the correct sequence.
- Planes require a minimum of three atoms (the sequence does not matter).
- Centroids require a minimum of two atoms.
- Vectors and points on lines require two atoms (and the point on the line requires a distance of extension from the tip of the vector defining the line).

It is possible to define 3D parameters involving other defined geometric objects. For example, with a vector and a plane defined, the angle between them may also be defined.

The Add 3D menu consists of:

1. Currently selected atoms.
2. Choice of 3D parameters to define. Valid parameters for the atoms selected are in black; non-valid ones are greyed out.
3. List of defined parameters and geometric objects, which may be edited by selecting the pen tool (geometric objects may be selected using the tick box).

To add a parameter, click the + icon next to a valid option.

Certain 3D parameters may be constrained:

- Distances may be constrained to be intramolecular or intermolecular, or any type, within a specified range.
- Angles and torsions may be constrained within a specified range.

To add a constraint, move the slider to constrained and enter the desired values.

