

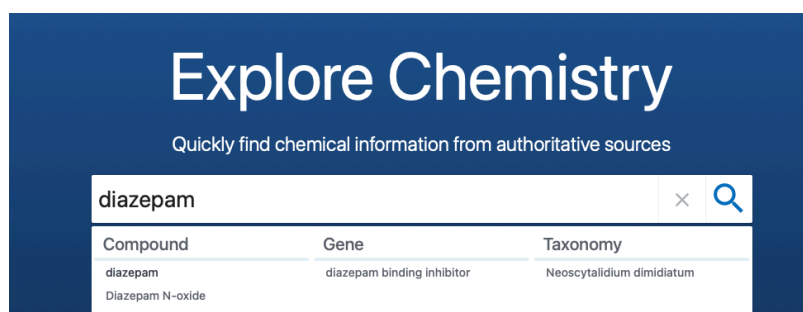
2. Molecular Docking: Preparing Ligands

In this lesson we are preparing ligands for molecular docking as in a virtual screening experiment. For that, we are going to learn a bit about drug databases and then use DataWarrior.

2.1. Drug Databases

2.2.1. PubChem





[PubChem](#) is one of the most used databases, and it is very user-friendly. One can look for compounds in different ways. Draw structure, upload ID list, or directly by name. In this example we search diazepam and select the best match.



Feel free to explore the compound summary.

COMPOUND SUMMARY

Diazepam

PubChem CID	3016
Structure	<div> 2D</div> <div> 3D</div> <div> Crystal</div>
	Find Similar Structures
Chemical Safety	<div> Acute Toxic Laboratory Chemical Safety Summary (LCSS) Datasheet</div>
Molecular Formula	$C_{16}H_{13}ClN_2O$
	diazepam 439-14-5

Then, we can click on find similar structures.


SEARCH FOR
CID3016 structure

Treating this as a structure search for CID 3016. [Edit Structure](#) Search for [CID3016 structure as text instead](#).

Identity (1) **Similarity (984)** Substructure (>1,000) Superstructure (>1,000) 3D Similarity (>818) [Settings](#)


Fingerprint Tanimoto-based 2-dimensional similarity search.

984 results [Filters](#) SORT BY Relevance [Download](#)




Flurazepam; 17617-23-1; Dalmane; Flurazepamum; Flurazepam HCL; ...
 Compound CID: 3393
 MF: C12H13ClFN3O MW: 387.9g/mol
 IUPAC Name: 7-chloro-1-[2-(diethylamino)ethyl]-5-(2-fluorophenyl)-3H-1,4-benzodiazepin-2-one
 Isomeric SMILES: CCN(CC)CCN1C(=O)CNC=C(C2=CC=CC=C2C(F)=C)C3=CC=CC=C3F
 InChIKey: SAQDBWUQADIS-UHFFFAOYSA-N
 InChI: InChI=1S/C21H23ClFN3O/c1-3-25(4-21)-12-26-19-9-15(22)13-17(18)21(24-14-20)(26)27(16-7-5-6-8-18)16(23)15-10,13(13-4,11-12,14)2,1-2(13)
 Create Date: 2005-03-25

[Summary](#) [Similar Structures Search](#) [Related Records](#) [PubMed \(MeSH Keyword\)](#)



Diazepam; 439-14-5; Valium; Anselisina; Diazemulc; ...
 Compound CID: 3916
 MF: C12H13ClN2O MW: 284.76g/mol
 IUPAC Name: 7-chloro-1-methyl-5-phenyl-3H-1,4-benzodiazepin-2-one
 Isomeric SMILES: CN1C(=O)CNC=C(C2=CC=CC=C2)C3=CC=CC=C3C(F)=C
 InChIKey: AACVJMBGZNE-UHFFFAOYSA-N
 InChI: InChI=1S/C15H13ClN2O/c1-19-14-8-7-12(17)9-13(14)16(18-10-15)19(20)11-5-3-2-4-6-11(12-8)10(10H),1(13)
 Create Date: 2005-03-25

[Summary](#) [Similar Structures Search](#) [Related Records](#) [PubMed \(MeSH Keyword\)](#)



Halazepam; Paxipam; Halazepamum [INN-Latin]; 23092-17-3; Sch 12041; ...
 Compound CID: 31640
 MF: C12H13ClF3N2O MW: 352.7g/mol
 IUPAC Name: 7-chloro-5-phenyl-1-[2,2,2-trifluoroethyl]-3H-1,4-benzodiazepin-2-one
 Isomeric SMILES: C1C(=O)CNC(=C(C2=CC=CC=C2)C3=CC=CC=C3)C4=CC(=C(C(F)(F)F)C(F)(F)F)C5=CC=CC=C5
 InChIKey: WYCLVQLVUQNZ-UHFFFAOYSA-N
 InChI: InChI=1S/C21H17ClF3N2O/c18-12-6-7-14-13(8-12)16(11-4-2-1-3-5-11)22-9-15(24)23(14)10-17(18,20)21(11-8)9-10(12)

[Summary](#) [Similar Structures Search](#) [Related Records](#) [PubMed \(MeSH Keyword\)](#)

ACTIONS ON RESULTS WITH ID TYPE: Compounds

[Push to Entrez](#) [Save for Later](#) [Linked Data Sets](#)

As it can be seen, it retrieves 984 similar compounds, other benzodiazepines. Once the selection is made, structure data can be downloaded in SDF format clicking in the right menu “Download” > “SDF” (coordinate type 3D).

[Settings](#)

DOWNLOAD [X](#)

Summary (Search Results)

[CSV](#) [JSON](#) [XML](#)

COMPRESSION:
☒ None ☐ GZip

Chemical Structure Records

[SDF](#) [JSON](#) [XML](#) [ASNT](#)

COORDINATE TYPE:
☒ 2D ☐ 3D

COMPRESSION:
☒ None ☐ GZip

Chemical Structure Images

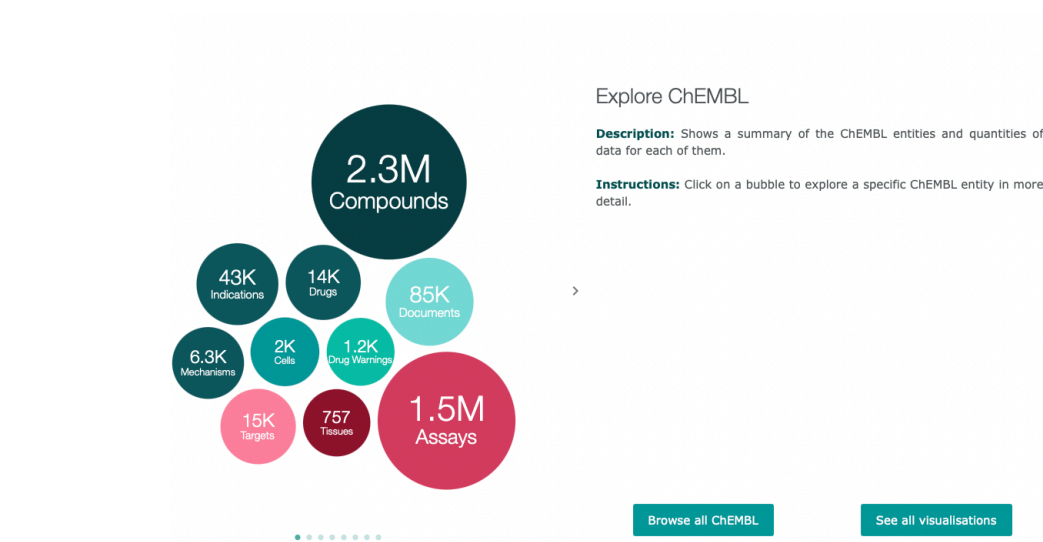
[PNG](#)

IMAGE SIZE:
☒ Small ☐ Large

COMPRESSION: ZIP ONLY

2.2.2. ChEMBL

ChEMBL is an open database of considerable amount of bioactivity data which comes from scientific literature, public databases, patents, etc.



Let's take a quick look.
We go for Drugs,

Browse Drugs

[Edit Querystring](#) [Show Full Query](#)

Table Cards Graph Heatmap

14,293 Drugs
0 Selected - [Select All](#)
[Browse Activities](#)

Records per page: 24 ☐ Select All

Showing 1-24 out of 14,293 records

Filters

Type

- Antibody 973
- Cell 47
- Enzyme 118
- Gene 77
- Oligonucleotide 115
- Oligosaccharide 75
- Protein 717
- Small molecule 10987
- Unknown 1184

Max Phase

0	6282
1	1401
2	2170
3	1502
4	2938

#ROS Violations

0	7423
1	1466
2	1466

CHEMBL2105777
Synonyms: Zorbenmycin (USAN)
Research Codes: U-30604, U-30,604
Max Phase: 0

CHEMBL2104207
Synonyms: Caffeine (ATC, BAN, INN)
Research Codes:
Max Phase: 0

CHEMBL4297446
Synonyms:
Research Codes: NSC-73938
Max Phase: 0

And select the second example,

Compound Report Card

Name And Classification



ID: CHEMBL2104207

Name: CAFEDRINE

Max Phase: 0 [Research](#)

Molecular Formula: C₁₈H₂₃N₅O₃

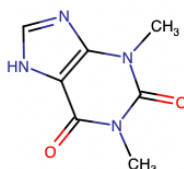
Molecular Weight: 357.41

ChEMBL Synonyms: [CAFEDRINE](#) [\(-\)-CAFEDRINE](#) [NOREPHENDRINETHEOPHYLLINE](#)

Molecule Type: Small molecule

Structure Search

If we click on structure search, we can also perform a substructure search. It also allows for substructure search.




[Connectivity](#) [Similarity](#) [Substructure](#)

> = 95%

We can click on substructure,

Substructure Search Results



Query: Cn1c(=O)c2[nH]cnc2n(C)c1=O

Status: Results Ready These results will expire on 2022-10-26T10:59:04.159889+00:00. [Learn More](#)

[Edit Search](#)

Show Full Query

610 Compounds
0 Selected - [Select All](#)
[Browse Activities](#)

[Info](#) [Cards](#) [Graph](#) [Heatmap](#)

Filters

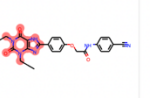
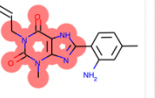

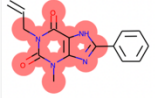
- Type
 - N/A - 23
 - Small molecule - 587
- Max Phase
 - 0 - 504
 - 1 - 1
 - 2 - 0
 - 3 - 0
 - 4 - 0
 - 5 - 0
- #ROS Violations
 - 0 - 469
 - 1 - 135

Records per page: 24

Select All

Showing 1-24 out of 610 records

[1](#) [2](#) [3](#) [4](#) [5](#) ...



As we have a list of 610 records that match that substructure. It can be filtered here by molecular weight for example (DataWarrior also allows that). If you click on SDF, you will have an SDF file available for DataWarrior.

2.2.3. Zinc

[ZINC15](#), database where you can search for compounds to be used in virtual screening. They are in a ready-to-dock format with their 3D conformation details.

2.2.4. DrugBank

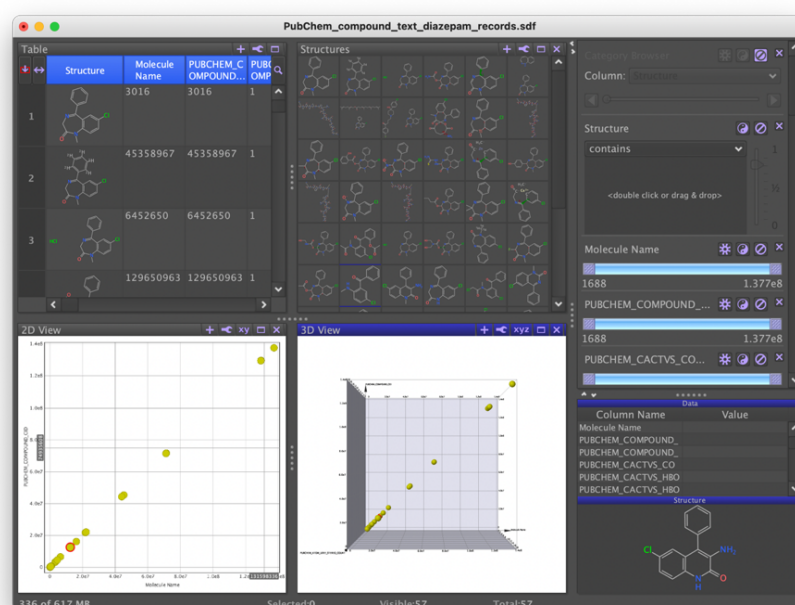
DrugBank is a comprehensive online database that provides detailed information about drugs and their targets. It contains data on both FDA-approved drugs and experimental drugs that are undergoing clinical trials or preclinical testing <https://go.drugbank.com>

2.2. Creating a Chemical Library in DataWarrior

DataWarrior is a free open-source program. You can use it to calculate and examine the properties of your compounds, filter datasets, generate minimized 3D conformations of molecules. Download and Install DataWarrior: <https://openmolecules.org/datawarrior/>

2.2.1. Open file and explore interface.

Once you have downloaded the SDF file, you can open it in DataWarrior.



The DataWarrior interface is composed of several display areas, with the lower ones being non-essential and can be closed if needed. In the upper-left section, there is a table that lists all compounds in your dataset, with each row representing a new compound and each column containing the properties defined in the sdf file. The upper-center window displays the 2D structure of every compound in the dataset, arranged in a left-to-right order. It is possible to click and select any of the compounds shown in this window, just like in the upper-left table. On the upper-right side, there are sliders that allow you to adjust various parameters and customize the view of the compounds. In the lower-left section, there is a

2D representation of the characteristics of the compounds, while the lower-center window displays a 3D plot of the same characteristics. These areas provide a visual representation of the data, enabling the user to easily identify patterns and relationships between the compounds.

2.2.1. Examine properties and filters.

Chemical properties do not come in the SDF file as default, we need to calculate them using DataWarrior. With these properties, we can filter the dataset and select only those that are suitable for docking.

Click. “Chemistry” > “Chemical Structure” > “Calculate Properties”

This will open a dialog box, where you can select the properties.

Most important: *molecular weight*, *LogP*, total polar surface area (*TPSA*), but also the number of H-donors and H-acceptors can be useful to display. When you have everything that you want, click select. And all these properties will be added as new columns in table, also new sliders to filter data have been added for each property in the filter area.

This way, you can now examine the dataset and decide on the ranges you would like to set for each parameter.

As you alter the sliders and select different filters you will see the “visible...” number in the status area change to show how many compounds out of your full dataset you are now viewing (here the full dataset contains 100 compounds, and we have no filters active currently hence all 100 are visible).

If we were to change the position of the sliders as you can see in the screenshot bellow where we have altered the LogP value range, the status area now shows “visible: 83” whilst the “total: 100” remains the same. This is because when you filter out compounds in DataWarrior you do not remove these compounds from the dataset. If you want to save the dataset (see section 6.4) with only the desired compounds, then you will need to remove the other undesirable compounds from your dataset (see later).

We can try with filter ranges:

- MW probably want to keep all below 500
- TPSA over 110 can cause problems with oral absorption (drug bioavailability of swallowed)
- Good LogP could range between 0-4, but not necessarily need to discount compounds just because they have slightly below 0 logP values.

Always discuss with your supervisor or an available expert if you are unsure on which values to discount, especially for your own designed compounds as these properties are all highly tunable with changes to structure.

As a rule of thumb, any extreme values of these key properties (both high and low) are likely to cause issues for drug development. You should research these properties and how they relate to drug development as part of your computational research project.

Another important property to display is the “Nasty Functions” option, this can be found on the “LE, Tox, Shape” tab of the calculate properties dialog box. Nasty functions are reactive functional groups or moieties with known toxicities contradicting their use in drug development. In this dataset there are no compounds containing nasty functions, but it is important to check this and remove these compounds from your list prior to docking.

There are other reasons you may wish to remove a compound from your dataset besides the property filters, such as their conformation being particularly strained (perhaps a 3 or 4 membered ring motif).

To delete a compound from your dataset, click on the row belonging to that compound (here clicking on the number 34 at the side will select the whole 34th row of table 1). Next click data, then delete rows, then selected rows. This will delete this entry from your dataset and the status area will change accordingly, showing “total: 99” at the bottom of the page here in this example.

You can remove columns in a similar fashion, if you have calculated a property which you no longer wish to display, or if the original dataset contained information, you do not need in your sdf file for docking (such as zinc ID in this example) then you can delete the column by clicking data, delete columns, and then selecting the column you want to delete from the menu in the dialog box that opens. Alternatively, you can right click on the column heading in table 1 and select delete column.

If you have filters active on your dataset and you wish to delete all compounds which have been filtered out then you can easily do this by clicking data, delete rows, then click invisible rows...

This will delete all of the compounds that were being filtered out (i.e. all the invisible compounds). For the above screenshot we altered the parameters of the LogP values, by reducing the range to only show compounds which had a LogP value between 1 and 3. This reduced the visible compounds from 100 to 83. By deleting the invisible rows this removed the 17 other compounds from the dataset.

2.2.3. Save the refined dataset in a single sdf file.

Once you have filtered down your dataset and removed any undesirable compounds you need to save the new dataset as a single sdf file. To do this click “file” > “save special” > “SD-file”.

6.4 Saving your Refined Dataset as a Single sdf File

After choosing where to save your sdf (and naming the file), this will bring up a small dialog box detailing how exactly you are saving the sdf, it is very important that you pay attention to these details as this will define the way the compounds information is saved. For molecular docking experiments you will need to have the 3D conformations of your compounds and sdf files should be saved based on these 3D conformations. To do this click on the drop-down menu next to “Atom coordinates:” and select “3D if available”.

It is also very important that you know which compound is which when looking at your docking results! If your compounds were not named in the original dataset, then they will not have any assigned numbers such as the “structure number” column in this dataset. An easy way to get around this is to save the sdf file based on the row number, this will name each compound as the row number it was listed as before you clicked save. It is useful to always order your compounds in the same way in table 1 **before you save**, you can do this by clicking on the heading of the column you wish to order results by (for example LogP) and DataWarrior will order your compounds in ascending/descending value for this property.

Once you have saved your refined dataset as a single sdf file based on the 3D atom coordinates of your compounds you now have a sdf file suitable for use in docking experiments.

If you open your saved sdf file, you can see a new “molecule name” column in table 1 has appeared. All the properties you calculated using DataWarrior when examining the original dataset have also remained and been saved in the sdf file.

6.5 Generating 3D Conformations of Compounds in DataWarrior

If you have designed your own compounds or have retrieved compounds from an online source which did not contain 3D structure conformation data, then you will need to generate these in order to use your compounds in docking experiments.

To generate the minimized energy 3D conformation of your compounds in DataWarrior, click Chemistry, then generate conformers...

Now you have generated the minimized 3D conformation of all compounds in your dataset, you can view these 3D structures in the detail area (bottom right) and can go ahead and save the dataset as a single sdf for docking (see section 6.4).

Finally, open with Pymol and visualize how do don't have any plain molecule anymore. You can compare both and show it in your assignment. Remember the name you select for your final sdf file, you will use it now for the molecular docking.