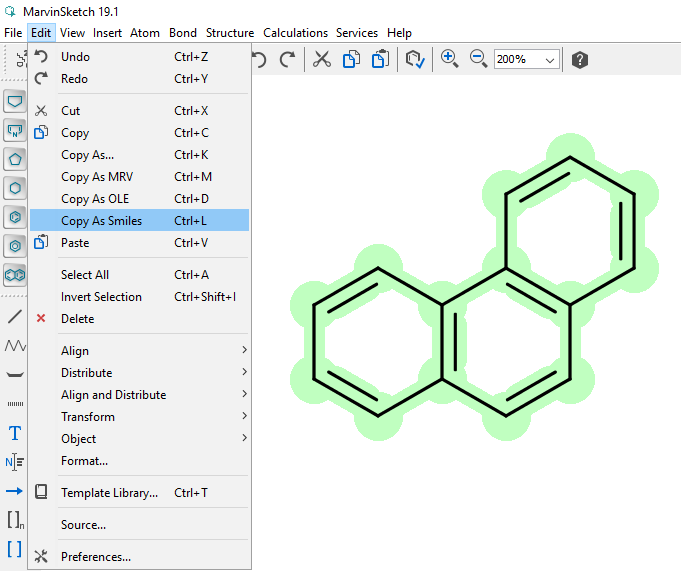
USER MANUAL TO CALCULATE PC1:PC2 COORDINATES

Compound Processing

**Step 1**

Obtain the SMILES string for the compound of interest. This will usually be supplied by the database from which the compound is sourced, OR, if the compound is novel, most chemical structure drawing programs have the option to generate the compound SMILES string. See an example using MarvinSketch to copy the SMILES string below.



**Step 2**

Download and open Notepad ++ (https://notepad-plus.en.softonic.com/download), paste the SMILES string here and save as a .smiles file.

**Step 3**

This step is used to remove any molecular fragments such as counter ions or solvent molecules that are included in the SMILES string. To do this, the ChemAxon standardizer is suggested. First, create a free academic ChemAxon account and download the Standardizer toolkit (<https://chemaxon.com/products/chemical-structure-representation-toolkit>). Thereafter, open the Standardizer and select the input file i.e. the .smiles file obtained from OpenBabel ( step 1). In step 2, select the ‘remove fragment’ option followed by ‘add’. This will automatically keep the largest fragment depending on the number of atoms. In step 3, specify the output file. This can be saved in a .smi format. Thereafter, select ‘run’ (step 4). The removal of fragments can also be carried out using Open Babel, however, since other Chemaxon software have previously been used this was convenient in our case.

**Step 4**

For each compound the SMILES strings are canonicalized. This is carried out with OpenBabel (<https://sourceforge.net/projects/openbabel/files/latest/download>). First, open the OpenBabel software and load the saved .smi file as the ‘input format’ (a). De-select ‘output below only’ and specify an output file location (b). Select ‘convert’ (b). This will standardize the SMILES strings. Ensure that the ‘output format’ is .can.

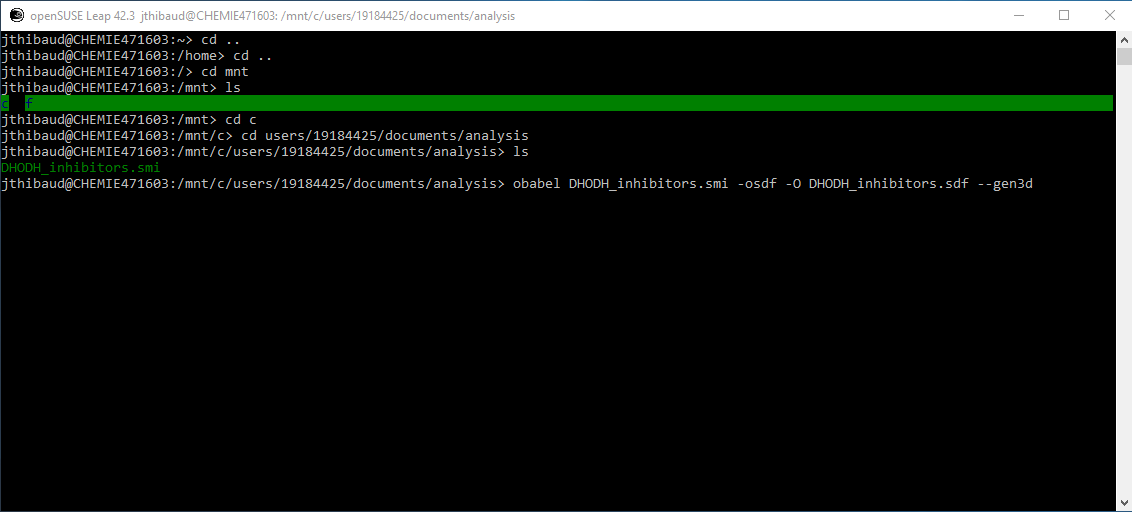


**Step 5**

Open Babel can be used for generating a structure data file which is required as an input for PaDEL‑Descriptors. This can be done in different ways either using gen3d in the command line version of Open Babel or if you are not interested in 3D structures by using gen2d or just the GUI.

If 3D structures are needed, the lowest-energy conformation can be generated for each compound. To do this, download OpenSUSE Leap 15.2 from the Microsoft Apps Online Store. Once the application has been opened, locate the folder in which the required .can or smi file is located. This can be done using the cd command (change directory). Commonly, cd mnt/c is used to go to the ‘This PC’ domain or the C Drive. From there, one can access various folders such as downloads or documents. Once the folder has been located, the ls command can be used to double-check its contents. In the example below, the folder name is ‘analysis’ and the file of interest within this folder is called DHODH\_inhibitors.smi. Thereafter, the following command line must be entered.

obabel filename.smi -osdf -O filename.sdf - -gen3d

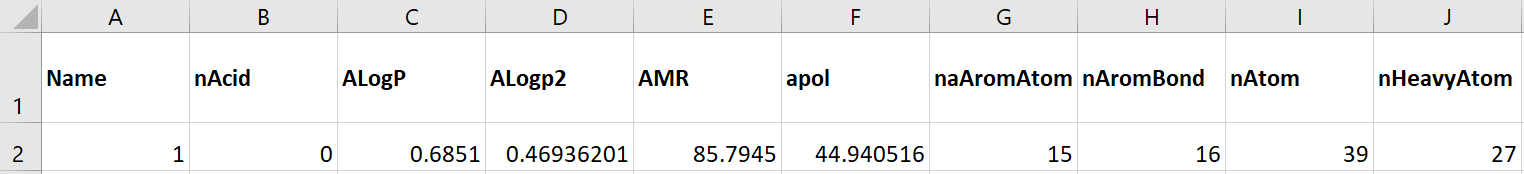
Once the calculation is complete, an .sdf file with the same filename can be found in the same location as the original .smi file.

This step is not strictly necessary if only 2D descriptors will be calculated by PaDEL, however, PaDEL does require a structure data file (.sdf) and so this process is one way of generating structure data from smiles data.

**Step 6**

To calculate one-, two- and if needed three-dimensional molecular descriptor with PaDEL, download and open the software (<http://www.yapcwsoft.com/dd/padeldescriptor/>). Thereafter, under the ‘general’ tab load the input .sdf file obtained from the gen3D calculation. It is often easier to just copy and paste the file directory into this line. Thereafter, paste the exact same line into the output file directory, however, it is essential to change the file format from .sdf or .smi to .csv. Thereafter, select the options to calculate 3D descriptors, and to retain 3D coordinates. Lastly, de-select the option to retain molecule order and select the option to use filename as molecule name.



PaDEL will calculate a total of 1444 1&2D and 429 3D molecular descriptors per compound. An example of an output file for a single compound showing only nine descriptors is given below. The compound identifier or ‘Name’ is not a descriptor.

Calculating PC1: PC2 Coordinates

Open the Excel spreadsheet for the appropriate descriptor set titled Descriptor set x PC1 and PC2 Coordinate calculation and mapping.xlsx, as well as the compound descriptor .csv output file from PaDEL.

**Sheet 1 – Given Information**

This includes all data used to generate the original T-SID map that will also be used in the PC1:PC2 coordinate calculations. The means and standard deviations (Columns D and E) for each of the 20 descriptors are given and used for Standardizing the data. The coefficient for the first two PCs are specified in columns G and H.

**Sheet 2 – Sorting Descriptors**

Here, the required subset of 20 descriptors is sorted from the total list of 1875.

* In column E, paste transpose all 1875 descriptor numerical values calculated by PaDEL. Compound Name must be placed in row 1. Ensure that all the data are selected for filtering. If not, select from cell B1 all the way to then end of the first row and select filter under the Data tab.
* Click on the drop-down arrow in cell B1 and select the option to sort from largest to smallest. This will sort all compound data and filter the required subset of 20 descriptors to the top of each column. Each of these 20 descriptors corresponds to a ‘1’ in column B.
* Copy these top 20 descriptors for each compound, including compound name which is in the first row starting from columns E.

**Sheet 3 – Descriptor Subset**

* Paste this data starting in cell B1. For all remaining sheets, it is suggested that compound name (row 1) is copied and pasted throughout in order to keep track of compounds.

**Sheet 4 to 6 – Standardization, PC1 Multiplication, PC2 Multiplication**

* Ensure there are enough columns for processing all of the data. If not, copy the last column and paste to right enough times to cover all the compounds

**Sheet 7 – PC1;PC2 Coordinates**

* Once again ensure there are enough columns for processing all of the data. If not, copy the last column to the right enough times to cover all of the data.
* The scores data appears on this page and will be plotted in the following sheets.

**Sheet 8 to 11 – Combined, TCAMS, DHODH and PKG maps**

* The PC scores will be plotted onto the respective enrichment maps.