**Materials and Methods**

***Data set***

A data set comprising the bioactivity data of compounds tested against the nuclear receptor protein family was compiled from the ChEMBL database, release number 23 (updated May 2017). Particularly, this encompassed the compilation of a total of 23 member proteins belonging to the nuclear receptor protein family.

**Data Collection**

Ligand binding and steroid binding domain sequences of each protein are retrieved from Uniprot web server release 2017\_07. The binding site can be confirmed by structure visualization in Pymol.

**2. Data was curated using biocurator server for each set of target compounds**

**After data curation**

1. Descriptor generation for each compound

12 sets of different descriptors were generated using PADEL software .

To calculate the descriptor we need to create an xml file **specifying the descriptor type as true.**

E.g- For **ExtendedFingerprinter**

<Descriptor name="Fingerprinter" value="false"/>

**<Descriptor name="ExtendedFingerprinter" value="true"/>**

<Descriptor name="EStateFingerprinter" value="false"/>

<Descriptor name="GraphOnlyFingerprinter" value="false"/>

<Descriptor name="MACCSFingerprinter" value="false"/>

<Descriptor name="PubchemFingerprinter" value="false"/>

<Descriptor name="SubstructureFingerprinter" value="false"/>

<Descriptor name="SubstructureFingerprintCount" value="true"/>

<Descriptor name="KlekotaRothFingerprinter" value="false"/>

<Descriptor name="KlekotaRothFingerprintCount" value="false"/>

<Descriptor name="AtomPairs2DFingerprinter" value="false"/>

<Descriptor name="AtomPairs2DFingerprintCount" value="false"/>

**After creating the xml file - Code to generate the Descriptor**

! java -jar -Xmx2G PaDEL-Descriptor/PaDEL-Descriptor.jar \

-removesalt -standardizenitro -standardizetautomers -fingerprints \

-descriptortypes SubstructureFingerprintCount.xml -dir smiles\_data\_.smi \

-file /home/adarsh/big\_data\_features/substructurecount\_111.csv

**Note-**