**Data pre-processing /Data curation**

* Data collected from CHEMBL contains three columns **–“ COMPOUD\_ID”** **,”PROTEIN\_ID”,”COMPOUND’S SMILE FORMULLA “,”IC-50” value.**
* **Compound’s smile (CNCCOOH etc.)** is needed to be converted in to numeric vector format to make it understandable to the algorithms.
* **Different strategy’s and algorithms are available for molecular fingerprints** (features) **from the Chemical formula of Compound. i.e. –** EstateFingerprinter,

GraphOnlyFingerprinter, MACCSFingerprinter, SubstructureFingerprintCount, AtomPairs2DFingerprintCount etc.

**Pre-processing and data preparation for proteins**

* Since proteins have ligand as well as steroid binding site ,our first step was to only select the sequence responsible for ligand binding .
* Molecular fingerprint of protein sequence was calculated from the uniport web server which creates the descriptors depending on the sequence of bases **(Adenine , Guanine ,Cytosine ,Thiamine).**