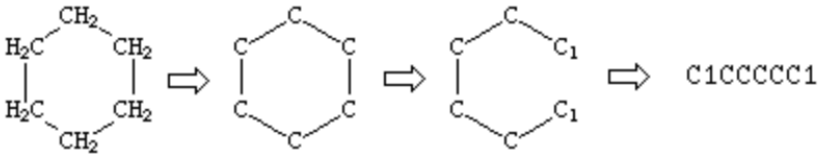
COVID - Drug Discovery for COVID19

Problem Statement- The objective is to prepare a machine learning model that can be used to propose potential novel effective drugs to fight SARS-CoV-2, the virus responsible for COVID-19.
You are provided with a dataset containing drug molecules (encoded as SMILES) and their binding affinities. The task is to use this dataset to make a regression model for binding affinity prediction.

- **SMILES** (Simplified Molecular Input Line Entry System) is the simplest way to reflect a molecule. The idea behind is to use simple line notations for chemical formulas that are based on some rules.

A simple example:



RDKit

- RDKit is a collection of chem-informatics and machine learning tools written in C++ and Python. Which is more important, it allows to work with many representations of chemical data and has a power to extract almost each chemical descriptor from the data you have.
- As you see the only information we have here is SMILES representation of molecular formulas. But RDkit is able to work with MOL representations. And it's actually nice to know RDkit still provides an opportunity to transform SMILES to MOL.
- Since size of a molecule can be approximated by a number of atoms in it, let's extract corresponding values from MOL. RDkit provides GetNumAtoms() and GetNumHeavyAtoms() methods for that task.
- The next obvious step is to count numbers of the most common atoms. RDkit supports subpattern search represented by GetSubstructMatches() method. It takes a MOL of a substructure pattern as an argument. So you can futher extract occurance of each pattern you'd like.
- We can use the above 2 features to train the model.

Mol2vec

- From package description 'Mol2vec is an unsupervised machine learning approach to obtain high dimensional embeddings of chemical substructures. Mol2vec learns substructure embeddings where vectors of chemically related substructures end up close in vector space.
- mol2alt_sentence() constructs a so-called 'molecular sentence' with desired Morgan fingerprints' radius (uses RDkit backend) where 'words' are unique substructure identifiers; MolSentence() is an internal wrapper function; sentences2vec() generates molecular embeddings with the help of the trained model; DfVec() is an internal wrapper for embeddings' generator (attribute .vec yields aggregated vectors).
- This feature extraction works wonderfully alone. But you can also concatenate the features from rdkit and mol2vec for better results.