Steps followed to build the classification dataset from NCI-60, based on the methodology of CDRUG paper:

- Download One Dose and Dose Response Data
  - **One Dose data** (date: June, 2020)
  - **Dose Response data** (date: June, 2020)

- Remove unnecessary spaces on both data.

- Preprocessing on One Dose and Dose Response Data:
  - From column unit, “CONCUNIT”, filter only the values having “M”, i.e., Molar.
  - From column concentration, “LCONC”, filter only “-5”, i.e., the concentration 10^{-5}.
  - From column percentage, “GIPRCNT”, filter NaN values.
  - From the column “GIPRCNT”, remove cases equal to “ . ”. In other words, null values.
  - From column percentage, “GIPRCNT”, get only values in the interval [0, 100].
    - Values below 0 do not indicate inhibition, but lethality.
    - Values above 100 are not defined.
    - For more details, see the methodology at: **NCI-60 Screening Methodology.**
    - **GIPRCNT meaning:** 100 is control growth, 0 is complete inhibition of growth (cytostasis), and -100 is complete cell kill.

- Average of the **GIPRCNT** based on each NSC id.

- Following **CDRUG supplementary material** (Figure S1):
  - From One Dose data, we defined as inactive molecules, those that have the average of the inhibition rate of growth lower than 5% at the dose of 10^{-5} M (i.e., GIPRCNT > 95).
  - From Response data, we defined as active molecules, those that have the average of the inhibition rate of growth higher than 50% at the dose of 10^{-5} M (i.e., GIPRCNT < 50).

- Merge this dataset with previous CDRUG activity dataset.

- Remove those SMILES that are invalid for RDKit:
  - For more details, check **RDKit documentation.**

- We have a sanity check to verify if molecules present in the active dataset are not into the inactive dataset; and vice-versa.
  - If these molecules are found, we just remove them from each dataset.