Data description

* **erbB1\_train.csv:** This file includes the molecular descriptors of all molecules intended for training. It comprises two targets: one for regression, specifically the 'pIC50\_nM' value, and another for classification, indicating whether the molecule is 'active' or 'nonactive.' It's important to note that the classification target is derived solely from the regression target. In cases where we emphasize binary classification, we omit the regression target.
* **erbB1\_test.csv:** Similar to "erbB1\_train.csv," this file is designated for molecules intended for testing the model. It contains the molecular descriptors and corresponding targets, allowing us to evaluate the model's performance on unseen data.
* **erB1-prepared.sd:** This file contains comprehensive information about each molecule, allowing you to generate the adjacency matrix based on the molecular descriptions provided. Each molecule is assigned a unique name, which serves as a reference to the corresponding target in the CSV files. Additionally, you can determine whether a molecule is active or nonactive by examining the IC50\_nM value given in this file. Molecules with IC50\_nM values less than or equal to 200 are considered active, while those with higher values are classified as nonactive.

Moreover, the adjacency matrix can also be derived from the connection table, as it provides detailed information about the connections between atoms. For instance, taking molecule 1 as an example, the connection table begins at line number 49 and concludes at line number 94. By analysing this connection table, you can identify that atoms 23 and 24 share a single bond, while atoms 21 and 22 have a double bond. This valuable information aids in constructing the adjacency matrix for each molecule.

We should notice here that the relation between **pIC50\_nM = -log10(IC50\_nM).**