Part 3: Tutorial Questions

1. What is in the training set, how big is it?

The training set contains information to train our artificial neuronal network (ANN) for predicting kinase binding and thus kinase activity. It contains data of 179'827 molecules in total.
The training set contains the following data for each molecule:

- Number of the molecule in the list.
- Molecule_chembl_id: Identifier of the specific molecule in the ChEMBL database.
- Standard_value: The IC50 of the molecule. The IC50 is the amount of substance needed to inhibit half of the kinases.
- Standard_units: SI unit of the standard value. Is always nM.
- Target_chembl_id: Identifier of the target molecule in the ChEMBL database.
- Smiles: The chemical structure of the compound. Smiles (Simplified Molecular-Input Line-Entry System) is a notation only using ASCII strings for writing chemical compounds in one row.

2. What modifications do you need to do to the data set to perform the tutorial?

- Calculate the pIC50: From each IC50 value the pIC50 must be calculated. This is done with the formula pIC50 = -log(IC50).
- Convert the SMILES string to numerical data: The code needed to program the function is given in the tutorial.
- Keep necessary columns only: A new data frame is made which only contains the numerical data of the molecule structure and the pIC50.

3. What is a test set? Any other types of set?

A test set is a separate set from the training set, which is used to check how good the learning process of the ANN was after the training set.
This is done by checking the loss function between the ANN and real output values. Another common way is by using a scatter plot, where the ANN values are plotted against the real values. A perfect ANN would result in all values being on the x=y line of the scatter plot.

Other sets:

- Validation set: A validation set is used in case we want to compare different models and choose the best one. Just like the test set, the validation set doesn't train the ANN. The difference between validation and test sets is, that the validation set is used for choosing between multiple trained models, while the test set then gives a final assessment of how good the ANN works.
  The validation and test sets must be done independently to prevent the risk of overfitting.
- Prediction set (external/unlabeled set): The aim of making an ANN is to let it predict certain things based on its training. The prediction set thus has the input, but no previously known output, which can be used to compare how good the ANN is. This prediction set will then allow the ANN to predict real-world problems, such as discovering new potential drugs.

4. Before starting describe with 1-2 sentences, in your own words, what is done in each of the cells.

1. Importing all libraries: These libraries include basic libraries such as Numpy, Pandas, Matplotlib & Seaborn, chemistry-related libraries such as Rdkit and machine learning-related libraries such as Sklearn & Tensorflow.
2. Data preparation: The data is first imported and put into a data frame. The data is further prepared by creating a new data frame without the unnecessary data and converting the SMILES string to numerical data. Furthermore, the available data is split into 70% training set and 30% test set.
3. Defining the neural network: A neural network containing 2 hidden layers is created using keras. The neural network uses ReLU in the hidden layers and a linear function in the output layer. In addition, the neural network uses the mean squared error as a loss function and adam as an optimizer.
4. Training the model: We train the ANN with 3 different mini-batch sizes using the training set. The mini-batch with the smallest loss is chosen as the best ANN.
5. Evaluation via test set: We assess the ANN's performance using the test set and then plot the obtained data vs. the real values in a scatter plot.
6. Prediction of external data: The trained ANN is used to predict the ligand binding (pIC50 values) of new, unseen molecules based on the input (SMILES chemical structure). The top 3 compounds are selected, visualized and can be analyzed further.