# Database Characterization Report using R

**Lucas Galvão Janot**

June 12, 2023

# 1 Introduction

The chosen database was Chembl, which contains information about bioactive molecules and their properties. The database can be obtained at `https://www.ebi.ac.uk/chembl/`.

Chembl is managed by the European Molecular Biology Laboratory - European Bioinformatics Institute (EMBL-EBI) and consists of over 2 million distinct records. These records correspond to molecules that have been tested in different experiments to evaluate their pharmacological properties.

Each Chembl record contains information about the molecule, such as its chemical structure, names, and identifiers.

Chembl is one of the largest public databases on bioactive molecules and is widely used in research in the field of drug development and medicinal chemistry.

# 2 Database Information

- Database Name: ChEMBL
- Data Release Date: Oct 2009
- Last Version Date: Jan 2023
- Maintainer: European Molecular Biology Laboratory European Bioinformatics Institute (EMBL-EBI).
- Objective: The objective of the ChEMBL database is to provide a comprehensive and freely accessible resource of bioactive molecules with drug-like properties and their associated biological activities, targeting the fields of drug discovery and development.

# 3 Variables used for the analysis

- **ChEMBL ID**: a unique identifier assigned to each compound entry in the database.
- **Name**: Molecule Name
- **SMILES**: ”Simplified Molecular Input Line Entry System” is a string-based representation of a molecule’s structure. It is a compact and human-readable format used to represent chemical structures in a text-based form.
- **Molecular Formula**: represents the actual number and types of atoms present in the molecule. It provides the simplest ratio of the different elements in the compound.
- **Molecular Weight**: is the mass of a molecule expressed in atomic mass units (u) or grams per mole (g/mol). It is calculated by summing the atomic masses of all the atoms in a molecule.
- **ALogP**: ”Aqueous Partition Coefficient” of a molecule. It is a measure of the lipophilicity or hydrophobicity of a compound and indicates how the molecule partitions between a nonpolar (lipid) phase and an aqueous (water) phase.
- **NumHAcceptors**: the number of hydrogen bond acceptor sites in a . Hydrogen bond acceptors are atoms that can accept hydrogen bonds from other molecules or functional groups.

- **NumHDonors**: the number of hydrogen bond donor sites in a molecule. Hydrogen bond donors are atoms or functional groups that can donate hydrogen atoms involved in hydrogen bonding.
- **NumRotatableBonds**: the number of bonds in a molecule that can freely rotate. Rotatable bonds are typically single bonds connecting non-terminal (i.e., not at the ends of the molecule) saturated carbon atoms or other atoms with a relatively low rotational barrier.
- **RingCount**: the number of rings present in a molecule. Rings, also known as cyclic structures, are closed loops of atoms that are connected by alternating single and/or multiple bonds.
- **TPSA**: ”Topological Polar Surface Area.” TPSA is a measure of the total surface area of a molecule that is occupied by polar atoms or polar functional groups.
- **Type**: the classification of a molecule based on its chemical structure and characteristics. It describes the general category or class of the molecule.

# 4 Variables Summary

- **ChEMBL ID:**
  - **Length:** 2,354,965
  - **Class:** Character
  - **Mode:** Character
  - **NA‘s:** 0
- **Name:**
  - **Length:** 2,354,965
  - **Class:** Character
  - **Mode:** Character
  - **NA‘s:** 2,306,892 ($\approx 97.96\%$)
- **SMILES:**
  - **Length:** 2,354,965
  - **Class:** Character
  - **Mode:** Character
  - **NA‘s:** 0
- **Molecular Formula:**
  - **Length:** 2,354,965
  - **Class:** Character
  - **Mode:** Character
  - **NA‘s:** 0

- **Molecular Weight:**
  - **Min.:** 4.0 (CHEMBL1796997, Name: Undefined, He , Small Molecule)
  - **1st Qu.:** 324.4
  - **Median:** 392.4
  - **Mean:** 433.9
  - **3st Qu.:** 474.5
  - **Max.:** 12546.3 (CHEMBL2179464, Name: Undefined, $C_{396}H_{390}F_{252}N_{66}O_{24}P_{42}$, Small Molecule)
  - **NA's:** 23,442 ($\approx 1\%$)
- **ALogP:**
  - **Min.:** -14.26 (CHEMBL1797815, Name: CELLOHEXOSE, $C_{36}H_{62}O_{31}$ , Small Molecule)
  - **1st Qu.:** 2.31
  - **Median:** 3.42
  - **Mean:** 3.45
  - **3st Qu.:** 4.57
  - **Max.:** 22.57 (CHEMBL1206998, Name: Undefined $C_{64}H_{127}NO_3S$, Small Molecule)
  - **NA's:** 84,994 ($\approx 3.61\%$)
- **NumHAcceptors:**
  - **Min.:** 0.0 (CHEMBL425822, Name: Undefined, $C_{36}H_{62}O_{31}$ , Small Molecule)
  - **1st Qu.:** 4.0
  - **Median:** 5.0
  - **Mean:** 5.5
  - **3st Qu.:** 6.0
  - **Max.:** 32 (CHEMBL604420, Name: Undefined, $C_{24}H_{16}F_6$, Small Molecule)
  - **NA's:** 84,994 ($\approx 3.61\%$)
- **NumHDonors:**
  - **Min.:** 0.0 (CHEMBL149936, Name: Undefined, $C_{21}H_{18}N_2O_4S_2$ , Small Molecule)
  - **1st Qu.:** 1.0
  - **Median:** 1.0
  - **Mean:** 1.59
  - **3st Qu.:** 2.0
  - **Max.:** 25 (CHEMBL1162334, Name: GUANIDINONEOMYCIN, $C_{29}H_{58}N_{18}O_{13}$, Small Molecule)
  - **NA's:** 84,994 ($\approx 3.61\%$)
- **NumRotatableBonds:**
  - **Min.:** 0.0 (CHEMBL373674, Name: Undefined, $C_{12}H_{10}N_2O$ , Small Molecule)
  - **1st Qu.:** 3.0

  - **Median:** 5.0
  - **Mean:** 5.73
  - **3st Qu.:** 7.0
  - **Max.:** 67 (CHEMBL541380, Name: Undefined, $C_{56}H_{129}ClN_{14}$, Small Molecule)
  - **NA's:** 84,994 ($\approx 3.61\%$)
- **RingCount:**
  - **Min.:** 0.0 (CHEMBL409633, Name: Undefined, $C_{74}H_{12}O$ , Small Molecule)
  - **1st Qu.:** 2.0
  - **Median:** 2.0
  - **Mean:** 2.46
  - **3st Qu.:** 3.0
  - **Max.:** 30 (CHEMBL409633, Name: Undefined, $C_{74}H_{12}O$, Small Molecule)
  - **NA's:** 84,994 ($\approx 3.61\%$)
- **TPSA:**
  - **Min.:** 0.0 (CHEMBL425822, Name: Undefined, $C_{24}H_{16}F_6$ , Small Molecule)
  - **1st Qu.:** 54.88
  - **Median:** 75.21
  - **Mean:** 81.87
  - **3st Qu.:** 99.27
  - **Max.:** 595.22 (CHEMBL32709, Name: Undefined, $C_{36}H_{74}N_{24}O_7$, Protein)
  - **NA's:** 84,994 ($\approx 3.61\%$)
- **Type:**
  - **Length:** 2,354,965
  - **Class:** Character
  - **Mode:** Character
  - **NA's:** 0
  - **Proportions:**
    - **Small Molecules:** 1,920,599 ($\approx 81.555\%$)
    - **Unknown:** 409,991 ($\approx 17.41\%$)
    - **Protein:** 22,750 ($\approx 0.966\%$)
    - **Oligosaccharide:** 95 ($\approx 0.004\%$)
    - **Oligonucleotide:** 201 ($\approx 0.009\%$)
    - **Gene:** 107 ($\approx 0.005\%$)
    - **Enzyme:** 121 ($\approx 0.005\%$)
    - **Cell:** 55 ($\approx 0.002\%$)
    - **Antibody:** 1046 ($\approx 0.044\%$)

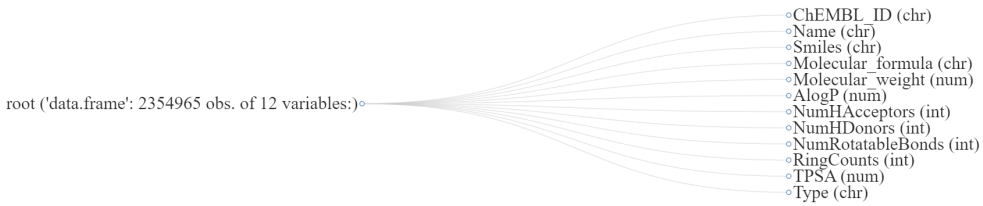


Figure 1: Studied Variables.

# 5 Histograms of the quantitative variables

## 5.1 Histograms

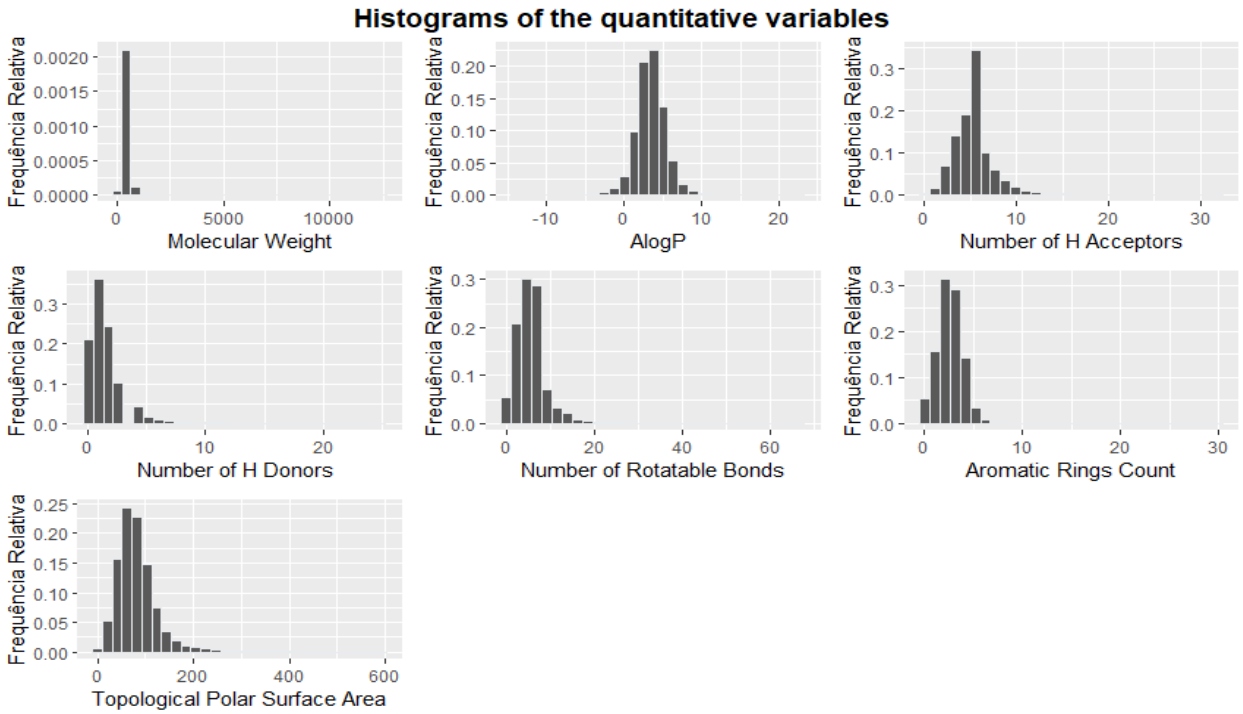


Figure 2: Histograms

## 5.2 Trimmed Histograms

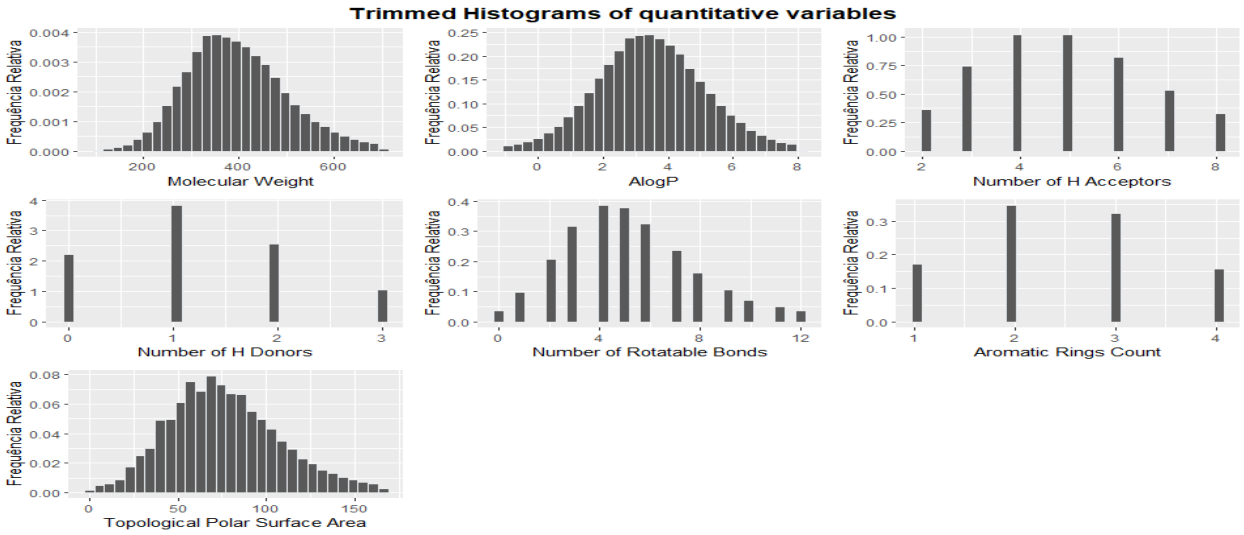


Figure 3: Trimmed Histograms

# 6 Boxplot of the quantitative variables grouped by type of molecule

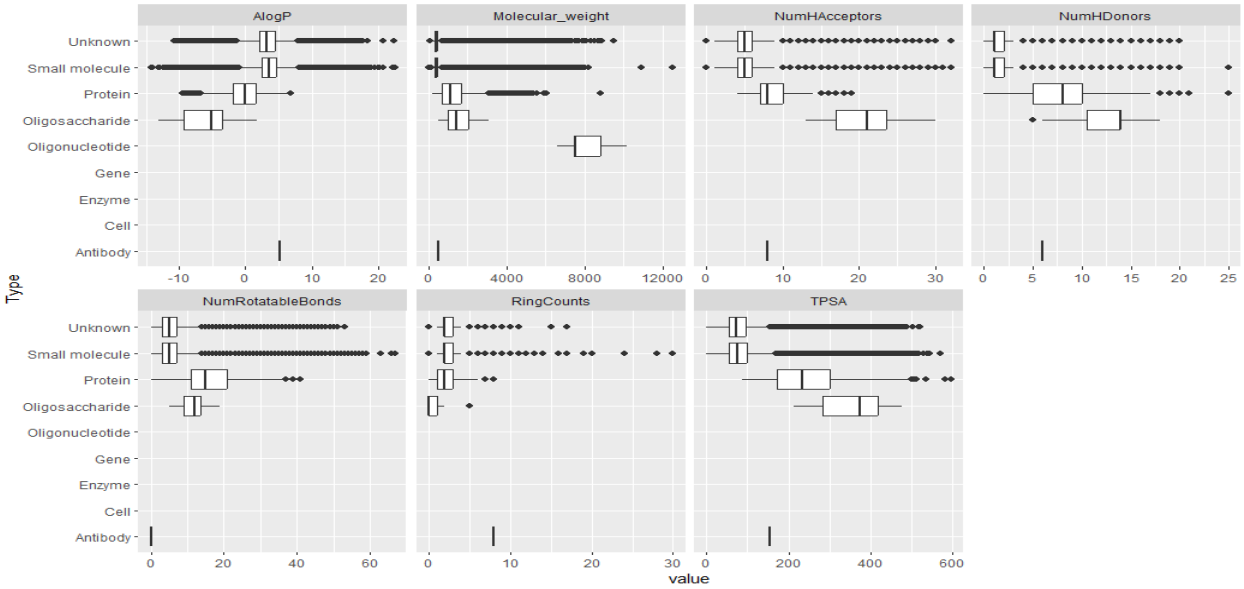


Figure 4: Boxplot

# 7 Correlation Matrix of the quantitative variables


Figure 5: Correlation Matrix

# 8 Scatter plots with regression line

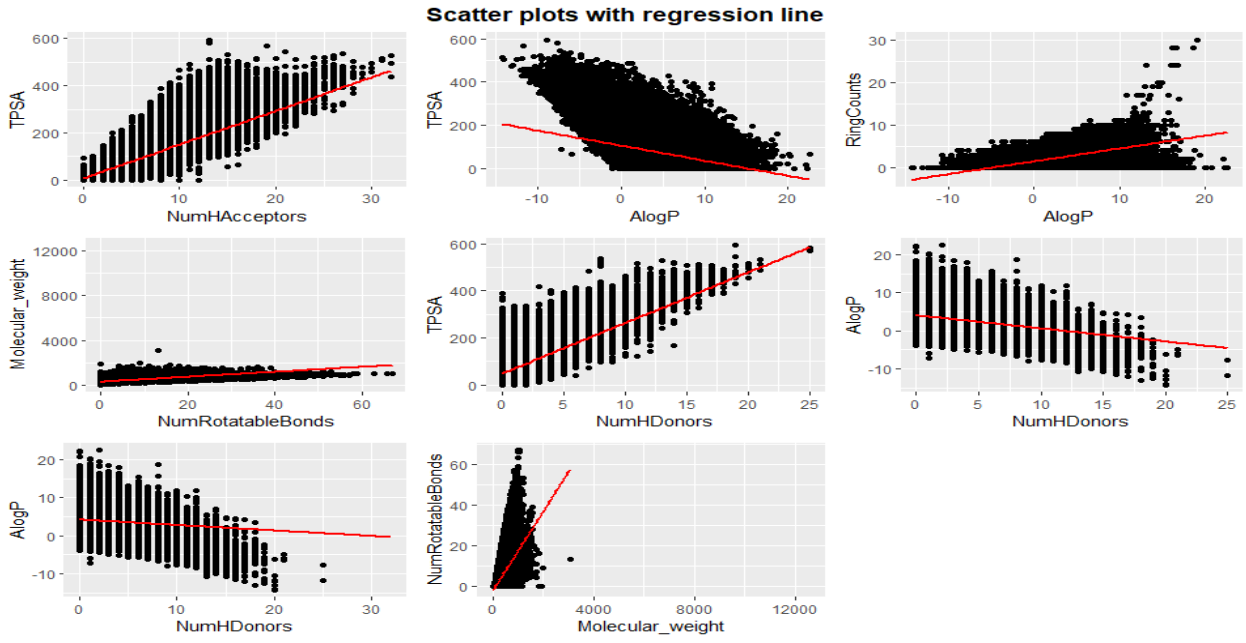


Figure 6: Correlation Matrix