1. Download all T1/2 results with keywords “human liver microsomes” and “half-life” from ChEMBL, and remove the data whose experimental notes indicate that they are not under Phase I conditions.

Data downloaded on 12th March, 2021

Activity 🡪 human liver microsomes (359,595) 🡪 t1/2 filter (9,129)

Remove data

1. wrong species (8,579)
2. non-liver (6,337)
3. t1/2 value not determined (5,590)
4. incorrect t1/2 unit label (5,583)
5. phase 2 (Glucuronidation, UDPGA) (5,504)
6. associated with human liver supersomes (5,484)
7. measured without NADPH (5,458)
8. in the presence of inhibitor (5,230)
9. duplicates (same SMILES and value) (4,920)
10. Removed concentration units “mg/mL”
11. Separated “microsomes”, “hepatocytes”, and “S9 fraction”
12. Categorized data based on drug concentrations (uM)
13. Removed duplicates (including isoforms)
14. Identify unlabeled concentration
15. Concentration curation
16. Selected data based on concentration from 0.1 to 3 uM
17. Removed outliers (t1/2 and MW)
18. Removed macroring
19. If the molecule contains ions 🡪 discard ions
20. If the molecule contains metal 🡪 discard the whole molecule

Q: How to find data measured in vivo?

What is the criterion to delete data with same SMILES but different values?

```Retained a single entry if all their entries had an unambiguous (similar) t1/2; otherwise,

all discard

```If the STD was higher than 10% of the average value 🡪 discard

2.Delete entries without numerical values, and keep only one record for duplicates (when both SMILES and values are identical).

3.Delete high-divergence entries (if SMILES are identical but the values differ by more than 15 minutes, all data for that compound are removed).

4.Merge low-divergence entries (if SMILES are identical and the values differ by less than 15 minutes, average all values for that compound and keep the result).

The resulting file is training\_data\_halflife\_20200821.csv.

5.Delete all entries with greater-than or less-than signs, keeping only equal-to entries.

The resulting file is training\_data\_halflife\_equals.csv.

6.Convert the values into minutes or take the natural logarithm (ln).

The resulting files are training\_data\_halflife\_eqmin.csv and training\_data\_halflife\_eqln.csv.