

Initial dataset (ChEMBL33)

273,552

1. Grouped by protein and ligand

272,922

2. Remove extreme values

272,265

3. Keep only highest activities

237,265

4. Remove previously published measurements

224,112

5. Remove measurements with author overlap

211,607