Task 2: QSAR Data Curation – Summary

Internship: AI and Omics Research Internship (2025)
Module: Drug Discovery & QSAR Modeling
Student: Aysha Meharin

Objective

The objective of Task 2 was to curate high-quality bioactivity data suitable for QSAR modeling by extracting, cleaning, and standardizing compound–target interaction data from ChEMBL.

Data Source

• Database: ChEMBL
• Data type: Bioactivity records (IC50 / Ki / EC50)
• Target: Protein target selected in Task 1

Methodology

1. Retrieved bioactivity data using the ChEMBL web resource client.
2. Converted raw JSON responses into structured Pandas DataFrames.
3. Selected QSAR-relevant columns such as molecule_chembl_id, canonical_smiles, standard_value, and standard_units.
4. Removed missing, duplicate, and non-numeric activity values.
5. Standardized activity units and exported the curated dataset as CSV.

Output Files

• chembl_activities.csv – raw extracted activity data
• qsar_curated_data.csv – cleaned and QSAR-ready dataset

Conclusion

The curated dataset is clean, standardized, and suitable for downstream QSAR modeling and machine learning applications in drug discovery.