AI and Drug Discovery

Assignment # 2

QSAR Data Curation Using ChEMBL

Submitted By: Syeda Rabia Gillani

Introduction – Objective and Selected Target Background

The objective of this assignment was to collect and curate high-quality bioactivity data for building a QSAR-ready dataset against a selected protein target. The focus was to retrieve experimentally reported small-molecule inhibition data, clean and standardize the records, and prepare a final dataset containing valid chemical structures and corresponding activity values that can be used for downstream QSAR modelling.

The selected target for this work was Fibroblast Growth Factor Receptor 2 (FGFR2). FGFR2 is a receptor tyrosine kinase that regulates key cellular signaling processes related to growth, proliferation, and survival. Because FGFR2 is a well-studied drug target in cancer-related research, it has a substantial amount of publicly available inhibitor data, making it suitable for QSAR-based modelling and prediction tasks.

Target and Data Selection – Data Evaluation and Justification

Target identification and data retrieval were performed using the ChEMBL database through the ChEMBL web resource client in Google Colab. The target was searched using the keyword "FGFR2," and the most relevant human FGFR2 entry was selected as the modelling target. The selected target entry was FGFR2 (ChEMBL Target ID: CHEMBL4142), which represents the single-protein target appropriate for inhibitor bioactivity analysis.

Bioactivity data were evaluated with a preference for consistent and comparable endpoints. Therefore, the dataset was restricted to records with standard_type = IC50, since IC50 is a widely used and interpretable measure of inhibitory potency and is commonly used in QSAR studies. This endpoint-based filtering reduced heterogeneity and improved suitability for modelling.

Selected target name: Fibroblast growth factor receptor 2 (FGFR2)

Number of available bioactivity records: Reported as the number of retrieved IC50 activity rows from ChEMBL for CHEMBL4142 (**raw IC50 records =4654**).

Data Curation Workflow – Retrieval and Preprocessing Steps

Data collection began by setting up Google Colab, mounting Google Drive, and organizing an output folder to ensure files could be saved and reused across sessions. Required libraries were installed, particularly the chembl_webresource_client, and core Python libraries such as pandas were imported for data handling.

After target selection, bioactivity records were retrieved from ChEMBL for the chosen FGFR2 target (CHEMBL4142). The records were filtered to retain only those with IC50 as the standardized activity type. The retrieved dataset was saved as a raw CSV file for traceability and reproducibility.

Preprocessing was then applied to produce a QSAR-ready dataset. Records with missing or invalid potency values were removed by filtering out rows without standard_value. Next, the dataset was reduced to only the most relevant columns required for modelling, specifically the molecule identifier, canonical SMILES, and potency values. Compounds with missing, empty, or invalid canonical_smiles were removed to ensure that every record had a usable chemical structure. Finally, a categorical bioactivity label (active/intermediate/inactive) was assigned based on IC50 thresholds as implemented in the notebook, and the cleaned dataset was exported as the final preprocessed CSV file.

Workflow Overview – Step-by-Step Explanation

1. Set up Google Colab, mount Google Drive, and create a project data folder for saving outputs.
2. Install and import required libraries, especially `chembl_webresource_client` and pandas.
3. Search ChEMBL for “FGFR2” and select the most relevant human single-protein target (CHEMBL4142).
4. Retrieve bioactivity records for the selected target from ChEMBL.
5. Filter the retrieved records to keep only IC50 entries (standard_type = IC50).
6. Save the filtered output as a raw CSV file for traceability.

7. Remove records with missing potency values by keeping only rows with a valid `standard_value`.
8. Keep only essential fields for QSAR (molecule ID, canonical SMILES, IC50 values).
9. Remove entries with missing/invalid SMILES so all remaining compounds have valid structures.
10. Assign bioactivity classes based on IC50 thresholds (active/intermediate/inactive).
11. Save the final cleaned dataset as the preprocessed CSV for QSAR modelling.

GitHub repository link:

AI_and_Drug_Discovery_Course_2026