AI and Biotechnology/Bioinformatics

Assignment 2:
QSAR Data Curation using CHEMBEL

Course: AI and Drug Discovery (2026)
Abeera Iftikhar

QSAR Data Curation Summary

1. Introduction

The objective of this assignment was to perform bioactivity data retrieval, evaluation, and curation from the ChEMBL database for use in Quantitative Structure–Activity Relationship (QSAR) modeling. QSAR modeling relies on high-quality, standardized bioactivity data associated with a well-defined molecular target. This work focused on assessing target suitability, retrieving relevant bioactivity records, and applying essential preprocessing steps required for QSAR analysis.

The molecular target selected for this assignment was **BCL2 (B-cell lymphoma 2)**. BCL2 plays a central role in regulating apoptosis and is of significant clinical relevance in various cancers, particularly hematological malignancies, making it an important target in cancer biology and drug discovery research.

2. Target and Data Selection

Selected Target

- **Target Name:** BCL2 (B-cell lymphoma 2)

Bioactivity data associated with BCL2 were explored using the ChEMBL database to evaluate their suitability for QSAR modeling. The focus was placed on standardized bioactivity measurements, particularly IC50 values, which are commonly used in QSAR studies.

Data Evaluation and Justification

- **Number of available bioactivity records:** 86 IC50 measurements

The analysis revealed that BCL2 has a sufficient number of directly associated small-molecule bioactivity records in ChEMBL for educational QSAR modeling purposes. The dataset provides a representative set of BCL2 inhibitors suitable for demonstrating QSAR data retrieval, cleaning, and preprocessing workflows. The relatively robust data availability highlights BCL2 as a well-characterized target in drug discovery research.

3. Data Curation Workflow

Bioactivity data were retrieved programmatically from the ChEMBL database using Python. The following preprocessing steps were applied to ensure data consistency and quality:

- Retrieval of BCL2-associated bioactivity records using the ChEMBL API
- Selection of standard bioactivity measurements (IC50)
- Removal of records with missing, invalid, or non-numeric values
- Standardization of bioactivity units to nanomolar (nM)
- Elimination of duplicate or inconsistent entries
- Export of the curated dataset into CSV format for downstream analysis

4. Workflow Overview

Step-by-step workflow:

1. Selection of BCL2 as the molecular target
2. Retrieval of bioactivity data from the ChEMBL database
3. Evaluation of data availability and quality
4. Filtering and preprocessing of IC50 bioactivity records
5. Standardization and final dataset preparation

Conceptual Workflow Diagram:

Target Selection → ChEMBL Data Retrieval → Data Evaluation → Data Cleaning & Standardization → Curated Dataset

5. Conclusion

This assignment successfully demonstrates the QSAR data curation workflow using BCL2 as a case study. The curated dataset, consisting of **86 IC50 bioactivity records**, provides sufficient entries to illustrate the complete QSAR data retrieval and preprocessing methodology. The exercise highlights the importance of target selection, data availability assessment, and careful data cleaning in QSAR studies. The resulting dataset serves as a robust example for educational and methodological demonstrations in QSAR modeling.

6. GitHub Repository

GitHub Repository Link:

https://github.com/Abeeraiftikhar/AI_and_Drug_Discovery_Course_2026.git