YUSUF MUNIR ALIYU

Introduction

The task retrieved, curated, and pre-processed bioactivity data from the ChEMBL database using Python for QSAR modelling.

The selected biological target is Dihydrofolate Reductase (DHFR), a well-characterized enzyme essential for nucleotide biosynthesis and cell proliferation. DHFR is a validated therapeutic target, particularly in antimicrobial and anticancer drug development, making it an appropriate choice for demonstrating systematic bioactivity data handling and curation.

Target and Data Selection

- Selected Target Name: Dihydrofolate Reductase (DHFR)
- Number of Available Bioactivity Records: 5

The target was selected due to:

a. Availability of a large number of experimentally validated bioactivity measurements
b. Presence of standardized activity endpoints (IC_{50} values)
c. Relevance to real-world therapeutic research

Data Curation Workflow

Target Selection
↓
Bioactivity Data Retrieval
↓
Filtering & Cleaning
↓
Unit Normalization
↓
Duplicate Removal
↓
Curated Dataset for Analysis

Workflow Overview

1. Data Retrieval: Bioactivity records linked to DHFR were extracted using target-specific queries.
2. Filtering: Only records with valid quantitative bioactivity measurements (IC_{50} in nM) were retained.
3. Unit Standardization: Activity values were converted to a uniform unit system where necessary.
4. Duplicate Removal: Redundant compound entries were identified and removed.
5. Data Cleaning: Missing values, inconsistent annotations, and outliers were excluded.
6. Feature Preparation: Cleaned data were formatted for computational analysis.

Conclusion

This work presents a comprehensive and systematic approach to selecting, curating, and preprocessing bioactivity data, underscoring the importance of data quality in computational drug discovery workflows. The curated dataset provides a strong foundation for downstream modeling and predictive analysis. https://github.com/aymunir1/AI-and-Drug-Discovery-Course.git