Bioinformatics Assignment 4

Dataset: BACE (β-secretase 1 inhibitors)

📥 **Dataset download:**
https://deepchemdata.s3-us-west-1.amazonaws.com/datasets/bace.csv

Learning Objectives

After completing this assignment, you should be able to:

- Critically analyze a real drug discovery dataset
- Identify chemical issues such as duplicates and missing stereochemistry
- Visualize and reason about chemical space and dataset splits
- Understand molecular flexibility as a 3D ensemble property
- Design a chemically meaningful train/test split
- Train and evaluate a simple predictive model
- Interpret model performance beyond raw accuracy

This is **not a competition for the best model**, but an exercise in **scientific reasoning**.

Dataset Description

The BACE dataset contains small molecules tested for inhibition of **BACE-1**, a protease target relevant to Alzheimer's disease.

Each row contains:

- A SMILES string
- Experimental activity (pIC50)
- Precomputed physicochemical descriptors (MW, ALOGP, HBD, HBA, etc.)
- A predefined dataset split

Part 1 — Dataset Sanity and Chemical Reality

Task 1. Load and Inspect the Dataset

- Load the CSV file
- Inspect column names and basic statistics
- Identify structure column(s), target variable, precomputed descriptors, provided dataset split

Deliverables

- Short description of dataset contents
- Summary table of key descriptors

Task 2. Duplicate Molecules and Label Consistency

- Identify duplicate molecules using InChIKey
- Check whether duplicates have identical pIC50 values?

Questions

- Are duplicates present?
- If yes, how are they labeled?
- What problems could duplicates cause for ML models?

Task 3. Missing or Undefined Stereochemistry

- Identify molecules with:
  - Unspecified stereocenters
  - Partial chirality information
- Count how many molecules are affected

Part 2 — Chemical Space and Dataset Splits

Task 4. Chemical Space Visualization

- Generate Morgan fingerprints (choose radius and length)
- Perform UMAP projection of chemical space
- Color molecules by pIC50 and split

Deliverables

- UMAP plot(s)
- Short interpretation

Task 5. Is the Provided Split Chemically Meaningful?

- Visually inspect train/test overlap in UMAP
- Compare similarity distributions train-train and train-test

Key Question

Does the test set represent genuinely new chemistry?

Task 6. Propose a Better Dataset Split

Choose **one** of the following splitting strategies:

- Scaffold split
- Butina clustering split
- UMAP-based split

Explain

- Why you chose this split
- What problem it solves compared to the original split

Part 3 — Molecular Size, Shape, and Diversity

Task 7. Dataset Diversity Analysis

Plot and analyze distributions of:

- Molecular weight
- Heavy atom count
- Number of rings
- Rotatable bonds
- ALOGP

Compare distributions across train vs test

Question

- Is this dataset chemically broad or narrow?

Part 4 — Molecular Flexibility (3D Structure)

⚠️ **To keep runtime reasonable, perform this step only for the TOP 100 molecules** (e.g., selected randomly or by activity range).

Task 8. Conformer Generation

For each selected molecule:

- Generate **10 conformers**
- Perform energy minimization
- Remove failed or invalid molecules

Task 9. Quantifying Flexibility

Compute at least one **ensemble-based flexibility metric**, such as:

- Mean per-atom RMSF across conformers
- Variance of radius of gyration
- Average atomic displacement

Deliverables

- One scalar flexibility value per molecule

Task 10. Structure–Activity Relationships

Analyze correlations between pIC50 and flexibility, MW, ALOGP, rotatable bonds.

Interpret

- Which properties appear meaningful?
- Which correlations may be misleading?

Part 5 — Modeling and Evaluation

Task 11. Feature Selection

Choose features from:

- Precomputed physicochemical descriptors
- Optional: flexibility metric

Explain:

- Why you selected these features
- What chemical information they encode

Task 12. Model Training

Train **one model** Random Forest **or** XGBoost using your chosen split and your selected features.

Task 13. Model Evaluation

Evaluate using RMSE, Spearman rank correlation.

Create:

- Predicted vs true pIC50 plot
- Rank correlation plot

Optional Bonus (Not Required)

- Visualize conformers of:
  - One rigid molecule
  - One flexible molecule
- Compare their conformational spread in 3D

Final Report

Your submission should include:

1. Dataset issues identified
2. Chemical space analysis
3. Justification of your chosen split
4. Flexibility analysis results
5. Model performance table
6. Critical discussion