Assignment 2 - Deadline: March 31, 2026

EECS 6895: Advanced Big Data and AI

Overview

In this assignment, you will work on two molecular property prediction problems using machine learning and deep learning techniques, and read a state-of-the-art research paper on deep learning for genomics. You will gain hands-on experience with:

- **Molecular representations**: SMILES strings, ECFP fingerprints, molecular descriptors, and graph representations
- **Traditional ML models**: Random Forest for molecular property prediction
- **Deep learning models**: Graph Convolutional Networks (GCN), Convolutional Neural Networks (CNN)
- **Evaluation metrics**: ROC-AUC, Accuracy, Precision, Recall, F1-score
- **Ablation studies**: Understanding feature importance
- **Reading scientific literature**: Understanding state-of-the-art deep learning models for genomics

---

Assignment Structure

Problem 1: Predicting Molecular Toxicity from SMILES (50 points)

- Load and preprocess the Tox21 dataset
- Convert SMILES to ECFP fingerprints
- Train a Random Forest classifier
- Build and train a Graph Convolutional Network
- Compare and analyze results

Problem 2: HIV Activity Prediction (50 points)

- Load and explore the HIV dataset
- Convert SMILES to molecular descriptors
- Train CNN and GNN models
- Perform ablation study
- Compare and analyze results

Problem 3: Reading Assignment - AlphaGenome

- Read the AlphaGenome paper (Nature, 2026)
- Answer questions on architecture, training, and evaluation

- Demonstrate understanding of multimodal deep learning for genomics

---

Problem 1: Predicting Molecular Toxicity from SMILES (50 points)

Toxicity prediction is an essential task in drug discovery, as **toxic molecules** can cause harmful effects in biological systems. The **Tox21 dataset** provides molecular toxicity data across **12 biological targets**, with molecules represented as **SMILES strings**.

Please refer to the course slides and DeepChem tutorials, and complete the following tasks:

1. **Load the Tox21 dataset** from MoleculeNet and extract SMILES strings. Implement **scaffold splitting** to create train/validation/test sets.
2. **Convert SMILES strings into ECFP** (Extended Connectivity Fingerprints) using RDKit.
    - *Note: Please do not load the dataset by using* `featurizer='ECFP'`*, convert SMILES into ECFP manually.*
3. **Train a Random Forest classifier** to predict molecular toxicity. Handle class imbalance using oversampling techniques (e.g., `RandomOverSampler`).
4. **Build and train a Graph Convolutional Network** to predict molecular toxicity. Your model should:
    - Convert SMILES to molecular graphs with atom features
    - Use multiple GCN layers with residual connections
    - Apply global pooling for graph-level predictions
5. **Compare the results** using metrics such as ROC-AUC (Area Under Curve), accuracy, precision, recall scores, and training time. Answer the analysis questions in the notebook.

Problem 2: HIV Activity Prediction (50 points)

HIV (Human Immunodeficiency Virus) activity prediction is crucial for drug discovery. The **HIV dataset** contains molecules labeled as **active or inactive** against HIV, making it a binary classification problem with significant class imbalance.

Please refer to DeepChem tutorials and complete the following tasks:

1. **Load the HIV dataset** from DeepChem, and **display 10 molecules** with their activity labels using RDKit visualization.

2. **Convert SMILES into molecular descriptors** using RDKit. Compute at least 10 physicochemical descriptors (e.g., molecular weight, LogP, number of H-bond donors/acceptors, TPSA, etc.).
    - *Note: Similar to Problem 1, please implement the descriptor computation manually.*
3. **Train two custom deep learning models**:
    - **CNN (1D Convolutional Neural Network)** on molecular descriptor features
    - **GNN (Graph Neural Network)** on molecular graph representations
4. **Perform an ablation study**—remove one feature type (e.g., the `IsInRing` atom feature from the GNN) and observe how it affects model performance. Analyze what this tells you about feature importance.
5. **Evaluate and compare** all models using Accuracy, F1-score, and ROC-AUC. Answer the analysis questions in the notebook.

Problem 3: Reading Assignment - AlphaGenome

Read the paper: *"Advancing regulatory variant effect prediction with AlphaGenome"* (Nature, 2026) and answer the following questions.

Paper link: https://www.nature.com/articles/s41586-025-10014-0

Instructions: Answer the following questions and combine your answers with your Jupyter notebook PDF for submission. Each answer should be concise (2-4 sentences per sub-question).

Q1. Architecture & Design

(a) What is the key trade-off that previous sequence-to-function models faced, and how does AlphaGenome address it?

(b) Describe the U-Net-inspired architecture of AlphaGenome. What role do the convolutional layers play versus the transformer blocks?

Q2. Training Methodology

(a) AlphaGenome uses a two-stage training process: pretraining and distillation. Explain the purpose of each stage and why distillation improves variant effect prediction.

(b) During distillation, the input sequences are randomly mutated. Why is this perturbation strategy beneficial?

Q3. Multimodal Learning

AlphaGenome predicts 11 different output modalities (e.g., RNA-seq, splice sites, histone modifications, chromatin contact maps).

(a) What are the advantages of training a single unified model across multiple modalities compared to training separate specialized models?

(b) Based on the ablation studies (Figure 7d), which tasks benefit most from multimodal training versus single-modality training?

Q4. Evaluation & Benchmarking

(a) The paper reports AlphaGenome outperforms existing models on 25 of 26 variant effect prediction benchmarks. Name two specific benchmark tasks and the metrics used to evaluate them.

(b) Why is predicting the *direction* (sign) of an eQTL effect important for practical applications like GWAS interpretation?

Q5. Critical Thinking

(a) What are two limitations of AlphaGenome acknowledged by the authors?

(b) The paper demonstrates AlphaGenome can recapitulate known oncogenic mechanisms in T-ALL. However, the authors note the model predicts *molecular* consequences, not *phenotypic* outcomes. Why is this distinction important for clinical applications?

Instructions

1. **Complete all TODO sections** marked with `# TODO: Your code here`
2. **Do not modify** the provided helper functions unless specified
3. **Answer all questions** in the designated markdown cells
4. **Run all cells** before submission to ensure reproducibility

---

Submission Requirements

Submit the following to CourseWorks:

1. **Completed Jupyter notebook** (`.ipynb` file) with all code cells executed
2. **Combined PDF** containing:
   - PDF version of the Jupyter notebook (Problems 1 & 2)
   - Answers to the reading assignment questions (Problem 3)

   Grading will be done on the PDF version.

Grading

Component	Points
Problem 1: Predicting Molecular Toxicity	50
Problem 2: HIV Activity Prediction	50
Problem 3: Reading Assignment - AlphaGenome	TBD
Total	**TBD**

Resources

- DeepChem Documentation
- RDKit Documentation
- PyTorch Geometric Documentation
- MoleculeNet Benchmark
- Course slides on Graph Neural Networks