**Цель:** Проверка навыков end-to-end разработки ML-модуля в контексте биофармы: от обработки структурных данных до построения и интерпретации модели.

**Сценарий:** вам нужно разработать прототип ML-модуля для **предсказания токсичности** малых молекул (binary classification: toxic / non-toxic).

### **Входные данные:**

Файл tox\_dataset.csv, содержащий:

* SMILES (структура молекулы),
* toxicity (0/1 — токсична или нет),

### **Задание**

### **1. Подготовка данных**

* Очистить и стандартизовать молекулы (RDKit): убрать соли, дубликаты, некорректные SMILES,
* Рассчитать молекулярные дескрипторы:
  + ECFP4 fingerprints (обязательно),
  + 1–2 других вида (например, Mordred, MACCS, GraphConv эмбеддинги — на выбор).

### **2. Моделирование**

* Построить минимум **2 классификационные модели**:
  + классическую (например, RandomForest или XGBoost),
  + нейросетевую (например, MLP, GNN или 3D Transformer — упрощённо),
* Реализовать пайплайн кросс-валидации, измерить ROC-AUC, PR-AUC, confusion matrix.

### **3. Интерпретация и анализ**

* Оценить важность признаков (например, SHAP для дерева, attention или Grad-CAM — если применимо),
* Провести error analysis: проанализировать false positives/false negatives, найти закономерности (например, группы соединений или подструктуры, вызывающие ошибки).

### **4. Документирование и формат сдачи**

* Все этапы должны быть оформлены в виде чистого Jupyter Notebook или Python-скриптов с комментариями,
* Добавьте requirements.txt или environment.yml.

### **Дополнительно:**

* Использовать MLflow для трекинга экспериментов,
* Обернуть пайплайн в CLI-интерфейс (argparse, click),
* Визуализировать embedding пространство (UMAP/t-SNE) для анализа классов.