**Dataset Analysis for InterDIA Project**

**Dataset Composition**

The dataset consists of 597 drugs total:

* 148 DIA-positive drugs (drugs that induce autoimmunity)
* 449 DIA-negative drugs (drugs that don't induce autoimmunity)

This is split into:

* Training set: 477 drugs (approximately 80%)
* External validation/test set: 120 drugs (approximately 20%)

**File Structure**

The dataset appears to be organized in multiple Excel files that contain different descriptor sets:

1. **Training Set Files**:
   * DIA\_trainingset\_DS\_descriptors.xlsx: Discovery Studio descriptors
   * DIA\_trainingset\_MOE\_descriptors.xlsx: MOE software descriptors
   * DIA\_trainingset\_Mold2\_descriptors.xlsx: Mold2 descriptors
   * DIA\_trainingset\_RDKit\_descriptors.xlsx: RDKit descriptors
2. **Test Set Files**:
   * DIA\_testset\_DS\_descriptors.xlsx: Discovery Studio descriptors
   * DIA\_testset\_MOE\_descriptors.xlsx: MOE software descriptors
   * DIA\_testset\_Mold2\_descriptors.xlsx: Mold2 descriptors
   * DIA\_testset\_RDKit\_descriptors.xlsx: RDKit descriptors
3. **Structure Files**:
   * DIA-TestSet.sdf: Chemical structures for test set compounds
   * DIA-TrainingSet.sdf: Chemical structures for training set compounds

**Data Characteristics**

1. **Chemical Representation**:
   * All compounds are represented using canonical SMILES notation
   * Molecular structures are available in SDF format
2. **Feature Space**:
   * Total of 1622 molecular descriptors across all platforms
   * These descriptors characterize various physicochemical properties and structural features
   * Features likely include properties like molecular weight, logP, topological indices, surface area calculations, etc.
3. **Class Distribution**:
   * Highly imbalanced dataset (approximately 3:1 ratio of negative to positive samples)
   * This imbalance is preserved in both training and test sets
4. **Data Sources**:
   * DIA-positive drugs compiled from:
     + SIDER database (with autoimmune adverse reactions ≥0.1% incidence)
     + Wu et al.'s study
   * DIA-negative drugs from Wu et al.'s study

**Data Processing Requirements**

1. **Initial Processing**:
   * Merging the four descriptor sets for each compound
   * Handling missing values (as mentioned in the workflow)
   * Normalizing/scaling features as needed
2. **Feature Engineering Challenges**:
   * High dimensionality (1622 features)
   * Potential redundancy between descriptors
   * Need for substantial feature selection
3. **Class Imbalance Handling**:
   * Will require specialized resampling techniques as shown in the workflow