**Project Goals**:

* Data Integration: To successfully integrate scRNA-seq and scATAC-seq datasets to create a unified view of gene expression and chromatin accessibility.
* Data Analysis: To apply the most popular algorithms for analyzing the integrated data, focusing on identifying cell populations, gene regulatory networks, and key regulatory elements.
* Benchmarking ML Models: To benchmark machine learning models suitable for predicting gene expression levels and regulatory mechanisms based on the integrated dataset.
* Insight Generation: To provide biological insights that can guide further experimental designs and validate existing hypotheses about cellular processes.

**Milestones:**

1. Data Preprocessing and Normalization: Completion of data quality checks, normalization, and scaling.
2. Integration of scRNA-seq and scATAC-seq Data: Successful integration of datasets using a state-of-the-art integration tool.
3. Data Analysis and Visualization: Identification of cell populations and regulatory networks, with visualizations created for key findings.
4. Benchmarking and Evaluation: Completion of benchmarking machine learning models with performance metrics reported.
5. Final Reporting: Compilation of findings, insights, and recommendations based on the integrated analysis.

**Modules and Scope:**

**Module 1: Data Preprocessing**

Scope: This module will handle data quality checks, normalization, scaling, and variance stabilization for both scRNA-seq and scATAC-seq datasets.

Content: Implementation of scripts or pipelines for data cleaning, normalization (e.g., log-normalization for scRNA-seq, TF-IDF for scATAC-seq), and identification of highly variable features.

Connections: Outputs will be formatted datasets ready for integration and analysis.

**Module 2: Data Integration**

Scope: Integration of scRNA-seq and scATAC-seq data to create a unified analysis framework.

Content: Use of integration tools (e.g., Seurat's integration methods, Harmony) to harmonize datasets based on shared cell populations or features.

Connections: This module connects processed datasets from Module 1 and feeds integrated data into downstream analysis modules.

**Module 3: Data Analysis and Visualization**

Scope: Analysis of integrated data to identify cell populations, differential expression, and accessibility, as well as the inference of gene regulatory networks.

Content: Implementation of clustering algorithms, differential expression and accessibility analysis, and regulatory network inference tools. Visualization includes plots (e.g., UMAP, heatmaps) to represent findings.

Connections: Utilizes integrated data and provides insights for biological interpretation and validation.

**Module 4: Benchmarking ML Models**

Scope: Benchmarking of machine learning models to predict gene expression or regulatory mechanisms from the integrated dataset.

Content: Selection of machine learning models (e.g., regression models, neural networks), training and testing on the integrated dataset, and evaluation of model performance.

Connections: This module takes input from the integrated dataset and connects to the final reporting by providing performance metrics and insights.

**Module 5: Reporting and Documentation**

Scope: Compilation of the project findings, methodologies, and insights into a comprehensive report or publication.

Content: Detailed documentation of data processing steps, analysis pipelines, code, results, visualizations, and benchmarking results. Includes writing of the manuscript or report sections.

Connections: Synthesizes information and results from all previous modules into a cohesive output.

**Modules and their dependencies diagram**

