Data Science Mini-Project 2024/25
Individual Reflective Report Stone Shi

· Overview

During week 15 to 24, my work primarily focused on data preprocessing, statistical modeling, and network analysis, alongside experiments with graph neural networks (GNNs). Throughout these relative analyses, I helped filter and extract key relationships within the dataset, supporting my collators to explore MIR100HG's regulatory mechanism. I regularly participated group discussions and collaborated on integrating findings into the final report.

· Key Activities and Contributions

Data Preparation and Analysis

I extracted and organized MIR100HG-related methylation datasets from TCGA for multiple cancer types, implemented filtering pipelines, and converted raw beta-values into structured .csv files. EDA visualization was performed though boxplots, scatter plot and distribution curves to identify skewed or outlier-prone features.

Regulatory Network Analysis

I parsed the TF-target dataset (gene_attribute_edges.txt.gz) to find potential upstream regulators of MIR100HG. I developed a Python parser to extract all rows where MIR100HG was annotated as a target gene, enabling the isolation of its direct upstream TFs. These candidate TFs were then matched against TCGA gene expression profiles across multiple cancer types to ensure data availability and expression variability. Subnetworks were generated focusing on MIR100HG, and node features were constructed using methylation and expression data.

Data Stratification

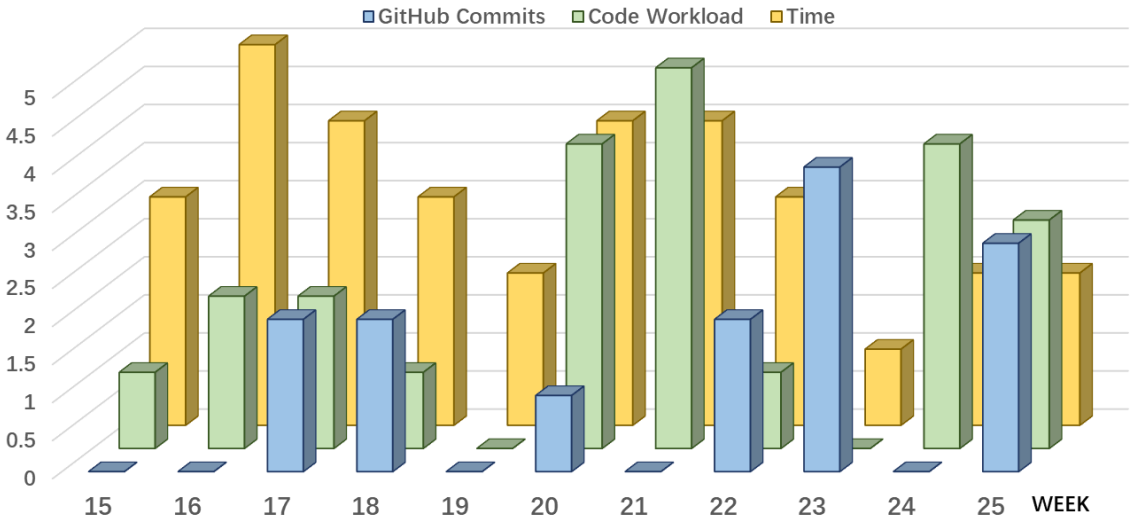
In collaboration with Mason, we formulated data stratification strategies to better interpret and validate the regulatory roles of MIR100HG across samples. We tried to scientifically partition dataset into biologically meaningful groups

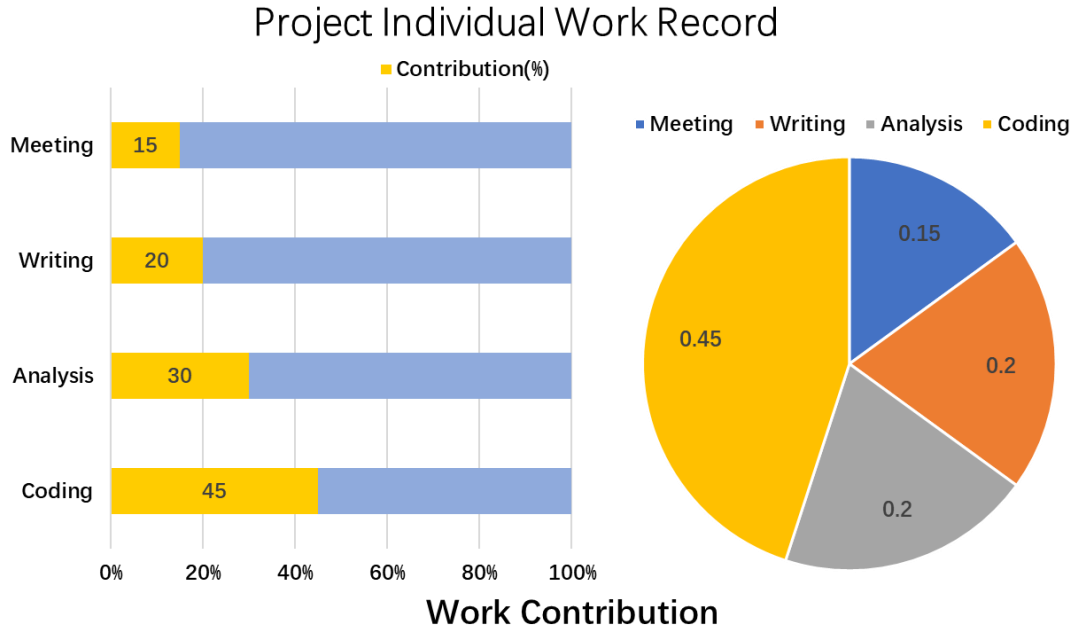
For each TCGA cancer datasets, we computed sample-wise distributions of MIR100HG expression and CpG probe methylation beta-values associated with MIR100HG. After inspecting the distribution curves, we applied quantile-based binning to divide samples into high and low strata. For Gene expression, we defined "high expression" as samples above the 75th percentile and "low expression" below the 25th percentile.

GNN-Based Modeling

I designed and trained a GNN model to infer regulatory relationships, experimenting with MSE loss for regression and cross-entropy for classification. Dropout regularization and graph-centered feature engineering were explored to improve performance.

- **Project Activity Record**





DSMP progress report

Student Name: Pangbo Shi

UoB Username: Stone Shi

15	Individual tasks:	Conducted a literature survey on MIR100HG's role in cancer epigenetics and gene regulation. Studied TCGA data structure and available methylation datasets.
	Collaborative tasks:	Team meeting for warming up
	Challenges:	Integrating fragmented biological literature to define a focused and feasible computational scope.
16	Individual tasks:	Implemented automated extraction of MIR100HG-related CpG methylation data from TCGA for multiple cancer types. Performed initial preprocessing, including handling missing values and unifying sample formats.
	Collaborative tasks:	Team meeting and clarified task requirement and allocated tasks.
	Challenges:	Ensuring consistent formatting and sample alignment across multi-cancer TCGA datasets.
17	Individual tasks:	Generated boxplots and distribution curves for each cancer-specific dataset. Identified cancer types with irregular distributions or potential outliers.
	Collaborative tasks:	Explain details about dataset and create GitHub Branches
	Challenges:	Efficiently parsing and filtering large-scale TF-target data while preserving biological relevance.
18	Individual tasks:	Annotated CpG probes showing highly skewed methylation patterns and filtered upstream TFs potentially regulating MIR100HG across cancers.
	Collaborative tasks:	Code sharing and initial results discuss
	Challenges:	Managing noisy and skewed distributions that hinder straightforward interpretation.
19	Individual tasks:	Built subgraphs centered on MIR100HG based on TF-target relations with defined feature vectors for each gene node.
	Collaborative tasks:	Collaborate with Mason on data stratification and implement some algorithms.
	Challenges:	Balancing network complexity and interpretability when selecting TF-target interactions and Filter conditions.
	Individual tasks:	Used NetworkX/Plotly to visualize regulatory subgraphs.

20		Provided visual insights into MIR100HG-centered regulation across cancers.
	Collaborative tasks:	Held meetings to prepare presentation. Refined content and rehearsed.
	Challenges:	Designing meaningful layouts that capture key regulatory structures in an interactive way.
21	Individual tasks:	Conceived a feasible GNN architecture Chose supervised tasks: regression (predicting MIR100HG expression) and classification (abnormal regulation detection).
	Collaborative tasks:	Met Dr Daniel D' Andrea to justify relation between TF / MIR100HG/ Gene targets. Refined subgraph network logic.
	Challenges:	Formulating biologically meaningful node features that support effective learning in GNNs.
22	Individual tasks:	Implemented the GNN training pipeline using PyTorch Geometric, in specific applied MSE and cross-entropy losses depending on task setup.
	Collaborative tasks:	Hold regular team meetings before last class to discuss methodologies and keep task completion
	Challenges:	Preventing overfitting in a high-dimensional graph with limited labeled data.
23	Individual tasks:	Evaluated GNN performance on held-out samples. Interpreted results to highlight key TF predictors and regulation patterns.
	Collaborative tasks:	Discussed overall report structure and allocated composing tasks
	Challenges:	Interpreting GNN outputs in context of heterogeneous molecular features and sparse annotations.
24	Individual tasks:	Summarized workflow and document methods adopted in data extraction; data stratification and network modeling.
	Collaborative tasks:	Collaborated with team to finalize a draft of the MIR100HG regulatory network and discuss detail about Task 3
	Challenges:	Summarizing complex workflows in a reproducible and collaborative manner for proper assessment.