DEPARTMENT OF ENGINEERING MATHEMATICS

Integrating multi-omics datasets to build a regulatory network and identify molecular drivers of micronutrient deficiency in pregnancy

Modeling Performance Across Feature Selection Strategies and Learning Algorithms

Weilin He

A dissertation submitted to the University of Bristol in accordance with the requirements of the degree of Master of Science in the Faculty of Engineering.

Wednesday 3rd June, 2026

Supervisor: Daniel D'Andrea

Declaration

This dissertation is submitted to the University of Bristol in accordance with the requirements of the degree of MSc in the Faculty of Engineering. It has not been submitted for any other degree or diploma of any examining body. Except where specifically acknowledged, it is all the work of the Author.

Weilin He, Wednesday 3rd June, 2026

Contents

List of Figures

List of Tables

List of Algorithms

List of Listings

Abstract

Vitamin B12 is a key micronutrient that supports metabolic balance and fetal development during pregnancy. A deficiency in B12 is linked to various metabolic disorders, adverse pregnancy outcomes, and restricted fetal growth. Despite its clinical importance, the molecular mechanisms underlying B12 deficiency remain poorly understood. Predictive modelling efforts in this area are also limited.

To address this gap, we developed a multi-omics framework by integrating DNA methylation (CpG), mRNA, and miRNA data with clinical and pathological features. Based on this dataset, we proposed two CpG-centred modelling strategies and evaluated the performance of three machine learning models.

Route A (statistical-driven) selects CpG sites from the genome that show strong statistical associations and large effect sizes with B12 deficiency status. **Route B (regulation-driven)** follows a biological framework. It identifies CpG–mRNA pairs in promoter or proximal regions that exhibit significant negative correlations, and prioritizes CpGs with stronger regulatory effects for modelling.

We then built and compared three predictive models: Logistic Regression (LR), Random Forest (RF), and Artificial Neural Network (ANN). Our hypothesis was that the regulation-driven approach would provide better interpretability and stability, and that non-linear models (RF and ANN) would outperform LR in predictive accuracy.

Main contributions and findings:

- Developed over 700,000 lines of code for data preprocessing, feature construction, modelling, and validation. The full implementation is publicly available at `https://github.com/WeilinHe00619/P056_Integrating_Msc.git`.
- Integrated multi-omics data (DNA methylation, mRNA, miRNA) with clinical and pathology-related variables into a unified dataset. Proposed two modelling routes: statistical-driven (Route A) and regulation-driven (Route B).
- Built and evaluated three models. The regulation-driven Random Forest (RF-Route B) achieved the most balanced and stable performance (mean AUC = 0.774). The regulation-driven ANN showed the highest peak performance (mean AUC = 0.784, recall = 74%), though with reduced stability.
- Identified key CpG–mRNA regulatory pairs enriched in metabolic, inflammatory, and developmental pathways.

Conclusion: This study presents a comprehensive framework of "two CpG-centred routes + three predictive models". Among them, the regulation-driven Random Forest model achieved the best balance of accuracy and interpretability. The Artificial Neural Network showed strong potential in capturing complex patterns. These results contribute both methodological and biological insights for understanding and predicting B12 deficiency during pregnancy.

Keywords: Vitamin B12 deficiency; pregnancy; pathology-related features; DNA methylation; CpG–mRNA regulation; multi-omics predictive modelling

Supporting Technologies

This section summarises the third-party resources (software tools, libraries, and hardware) that supported the project. These technologies were essential for data processing, modelling, visualisation, and documentation.

- I used the **BioRender** online platform to generate biological illustrations, including molecular mechanism diagrams, regulatory network schematics, and workflow figures.
- I used the **R** programming language, mainly with libraries such as `ggplot2` and `ComplexHeatmap`, for statistical visualisation of omics data. Typical applications included differential feature plots, PCA analysis, heatmaps, and pathway enrichment visualisations.
- I used **Python** as the primary programming language for data preprocessing, modelling, and analysis. Development was conducted in Jupyter Notebook format, with modular scripts covering clinical data integration, methylation expression extraction, standardisation, regulatory network construction, and predictive modelling. The following open-source libraries were employed:
  - `pandas`, `numpy` – for data cleaning and matrix operations;
  - `scikit-learn` – for Logistic Regression and Random Forest models;
  - `PyTorch` – for the Artificial Neural Network (ANN) model;
  - `matplotlib` – for results visualisation.
- I used **GitHub** (https://github.com/WeilinHe00619/P056_Integrating_Msc.git) for version control, collaboration, and to ensure reproducibility of the analysis pipeline.
- I used the **Overleaf** online platform to prepare and typeset my dissertation using LaTeX.
- The project was executed on a local high-performance machine: a MacBook Pro (14-inch, **Apple M3 Pro chip**, **18 GB RAM**, **macOS Sonoma 14.4**), which supported the processing of large-scale multi-omics data and training of machine learning models.

Notation and Acronyms

This section provides a consolidated list of acronyms and notational symbols used throughout the thesis. Acronyms are listed alphabetically, followed by mathematical and algorithmic notation. Defining these elements upfront ensures consistency and clarity across all chapters.

Acronyms

ANN	Artificial Neural Network
AUC	Area Under the Curve
BMI	Body Mass Index
CpG	Cytosine–phosphate–Guanine site
CV	Cross-Validation
DEG	Differentially Expressed Gene
DEmiR	Differentially Expressed microRNA
DMR	Differentially Methylated Region
FDR	False Discovery Rate
GO	Gene Ontology
KEGG	Kyoto Encyclopedia of Genes and Genomes
LASSO	Least Absolute Shrinkage and Selection Operator
LR	Logistic Regression
PCA	Principal Component Analysis
RF	Random Forest
ROC	Receiver Operating Characteristic
RRBS	Reduced Representation Bisulfite Sequencing
RNA-seq	RNA Sequencing
sRNA-seq	Small RNA Sequencing
U-test	Mann–Whitney U Test

Mathematical Symbols and Notation

x_i	The i-th input feature vector
y_i	Binary class label of sample i ($y_i \in \{0, 1\}$)
$\mathbf{x}$, $\mathbf{w}$	Feature and weight vectors (dimension p)
b	Bias term in linear models
$\hat{y}_i$	Predicted probability or output for sample i
$\sigma(\cdot)$	Sigmoid activation function, $\sigma(z) = \frac{1}{1+e^{-z}}$
$\phi(\cdot)$	Activation function in hidden layers (e.g., tanh)
L	Loss function (e.g., cross-entropy)
$\mathcal{L}_{\text{BCE}}$	Binary cross-entropy loss
λ	Regularization coefficient
C	Inverse regularization parameter ($C = 1/\lambda$)
$\|\mathbf{w}\|_1$	L1 norm (sum of absolute weights)
$\|\mathbf{w}\|_2^2$	L2 norm (sum of squared weights)
R^2	Coefficient of determination (model fit in linear regression)
β	Regression coefficient (e.g., in CpG–mRNA analysis)
d	Cohen's d effect size
p-value	Statistical significance in hypothesis testing
F1-score	Harmonic mean of precision and recall
n	Total number of samples
p	Number of features
T	Number of trees in a random forest
$h_t(\cdot)$	Prediction from the t-th decision tree
majority_vote$(\cdot)$	Aggregated class prediction in random forest

Algorithmic Terms

$\mathcal{D}$	Full dataset used for modeling
$\mathcal{D}_{\text{train}}$, $\mathcal{D}_{\text{test}}$	Training and testing subsets
$\mathcal{C}$	Candidate set of C values for tuning
$\mathcal{H}$	Hyperparameter search space
E	Maximum number of training epochs
B	Batch size during ANN training
K	Number of repetitions or independent runs
h, h^*	Hyperparameter configuration, and the optimal one selected
f_h	Model trained under hyperparameter configuration h

Acknowledgements

I would like to express my deepest gratitude to my dissertation supervisor and personal tutor, Dr. Daniel D'Andrea. From the very beginning of my Master's year, he has been a constant source of support and guidance, both academically and personally. As my personal tutor, he always offered advice and encouragement whenever I faced difficulties, and his kindness and patience helped me to navigate the challenges of postgraduate study.

Later, during the Data Science mini-project, I was both surprised and honored to be assigned to the small project group under his supervision. This experience not only allowed me to explore a research topic that I found genuinely interesting, but also gave me the opportunity to work more closely with him. It was through this project that I realized how professional, knowledgeable, and supportive he was as a supervisor. The positive experience and invaluable guidance I received during the mini-project were major reasons why I chose to pursue my Master's dissertation in his area of expertise. It was a decision driven by both my academic interest and my deep appreciation for his mentorship.

Throughout my Master's year, Dr. D'Andrea has always been incredibly professional and genuinely caring. He consistently answered my questions in a timely manner, encouraged me when I doubted myself, and showed understanding and tolerance of my shortcomings. His constructive feedback helped me improve both my research and my academic writing. When I later encountered personal difficulties due to family reasons, he went out of his way to help me coordinate and always considered my situation from my perspective. For this, I am profoundly grateful. I believe this kind of care and empathy is one of the greatest gifts I have received during my Master's journey.

Dr. D'Andrea's excellence lies not only in his expertise but also in his approachability and warmth. His professional ability is outstanding—he could immediately understand my code, pinpoint problems within seconds, and explain concepts with great clarity. Even when I struggled to express my thoughts clearly in English, he was able to grasp what I meant and guide me with patience. He often sketched out detailed derivations on paper, introduced me to useful online tools, and offered practical advice that significantly improved my work.

Having the opportunity to be mentored by Dr. D'Andrea has been the most valuable part of my Master's experience. His combination of strong academic expertise, dedication, and genuine kindness has made a lasting impact on me. For all his guidance, encouragement, and support, I am truly and deeply thankful.

Chapter 1

Introduction

1.1 Research Background and Significance

Vitamin B12 deficiency is one of the most common micronutrient problems during pregnancy. It has been linked to many health risks. These include insulin resistance, gestational diabetes, low birth weight, and impaired neurodevelopment in the fetus [4, 2]. The long-term effects may also include metabolic disorders in the child.

B12 is a key cofactor in one-carbon metabolism. It supports the synthesis of S-adenosylmethionine (SAM), a major methyl donor. SAM plays a central role in DNA methylation, which regulates gene expression at the epigenetic level [10, 19]. Therefore, B12 deficiency is more than a nutritional issue. It may cause molecular disruptions that affect both the mother and the developing fetus.

Global studies confirm the high prevalence of B12 deficiency during pregnancy. A meta-analysis reported rates ranging from 19% to 29% across trimesters, with the highest levels in South Asia and the Eastern Mediterranean [21]. Several large cohorts have linked low maternal B12 to metabolic conditions and poor pregnancy outcomes [7, 1]. For example, the Generation R study found connections between maternal B12 levels and brain development in offspring [5]. Other research highlighted higher risks among South Asian women in Canada [12].

Emerging evidence suggests a layered biological pathway. B12 deficiency may affect gene expression through a DNA methylation–miRNA–mRNA regulatory axis [11, 9]. However, most existing research focuses on a single omics layer. There is limited understanding of how multiple regulatory systems interact.

This study addresses that gap. We integrate DNA methylation (RRBS), miRNA expression (sRNA-seq), and mRNA expression (RNA-seq) data. Our aim is twofold: (1) to uncover the regulatory network behind B12 deficiency, and (2) to build predictive models using multi-omics features and clinical data. This dual approach combines mechanistic insight with real-world application.

The study is significant in three ways. First, it uncovers how B12 deficiency affects biological systems across multiple layers. Second, it blends machine learning and causal inference to improve both accuracy and interpretability. Third, it proposes new biomarkers and modeling tools that can support early diagnosis and personalized nutrition in pregnancy.

1.2 Literature Review

1.2.1 Molecular Mechanisms

B12 deficiency changes gene expression patterns. Mosca et al. showed that B12 deficiency lowers SAM and affects mRNA methylation [19]. Other studies found that deficiency alters fat metabolism and increases inflammation in fat cells [20]. The PRIYA trial revealed that B12 supplements before pregnancy can change gene expression in the baby's cord blood [13].

miRNAs also play a role. Research shows that B12 deficiency can change the levels of circulating miRNAs. These miRNAs may target genes such as $PPAR\gamma$, which are involved in fat storage and insulin sensitivity [1]. Animal studies further showed that B12 levels can influence miRNA expression across generations [14, 15].

DNA methylation is another key mechanism. Studies have found links between maternal B12 levels and methylation at specific genes such as *IGF2* and *H19* [6, 17]. Placental methylation studies support

the role of one-carbon metabolism in regulating fetal development [22]. EWAS studies found that B12-associated CpGs can persist into childhood and affect cognitive and metabolic traits [18, 16].

1.2.2 Systems Biology and Modeling

Some recent studies use multi-omics data to study complex conditions. In cancer research, multi-omics models have revealed ethnic differences in gene regulation [23]. Others have used network-based methods to explore lipid metabolism and B12 function [7]. Tools like the MORE model provide ways to identify key regulatory genes through network inference [3].

Machine learning has become popular in nutritional research. Random forests, XGBoost, and artificial neural networks (ANNs) help predict outcomes and identify biomarkers [8]. In maternal B12 studies, researchers have used these tools to predict risk of insulin resistance and gestational diabetes [2].

Still, most studies do not combine these methods with multi-omics data. Integrative approaches remain rare in pregnancy-related research.

1.2.3 Research Gaps

Vitamin B12 deficiency during pregnancy is common and linked to adverse maternal and fetal outcomes. Previous research has revealed that B12 influences molecular pathways through changes in gene expression, miRNA activity, and DNA methylation. However, most existing studies have focused on single omics layers and have not explored how these molecular features interact across regulatory levels.

There remains a significant gap in understanding the integrated multi-omics regulation underlying B12 deficiency. Few studies have modeled the full CpG–miRNA–mRNA cascade or systematically analyzed how epigenetic and post-transcriptional mechanisms jointly shape gene expression in the context of micronutrient imbalance.

From a predictive perspective, machine learning tools have been applied to classify metabolic outcomes, but rarely for B12 status prediction. More importantly, current models seldom integrate both molecular and clinical variables or consider the biological relevance of features.

This study addresses both mechanistic and predictive gaps. First, it constructs a multi-omics regulatory network to reveal key B12-related molecular interactions. Second, it develops predictive models of B12 deficiency using two complementary strategies: a data-driven approach that selects features based on statistical significance, and a mechanism-guided approach that leverages regulatory network hubs with biological interpretability. This dual modeling strategy balances predictive performance with mechanistic insight, offering translational potential for early screening and personalized nutrition interventions.

1.3 Overview of the Thesis and Chapter Structure

The structure of this thesis reflects a logical progression from data acquisition to biological interpretation and predictive modeling. As summarized in Table 1.1, each chapter is designed to build upon the previous one, leading to a comprehensive understanding of the molecular basis and clinical implications of B12 deficiency.

Table 1.1: Thesis Structure and Content Overview

Chapter	**Content Overview**
Introduction	Introduces the research background, literature gaps, and study goals.
Data Sources and Preprocessing	Describes the study population, omics data, differential analysis, and quality control.
Analytical Methods and Modeling Workflow	Details the analysis pipeline, including enrichment, network modeling, and prediction.
Results and Findings	Summarizes the main results from enrichment and modeling.
Conclusion	Discusses the study's contributions and outlines future directions.

Chapter 2

Data Sources and Preprocessing

2.1 Study Population and Sample Information

2.1.1 Basic Characteristics of the Cohort

This study enrolled 50 pregnant women, each of whom provided peripheral blood samples for multi-omics profiling, including RNA sequencing (mRNA), small RNA sequencing (miRNA), and reduced representation bisulfite sequencing (RRBS). The basic characteristics of the study cohort are summarized in Table 2.1.

The average age of participants was 30.7 ± 5.2 years, ranging from 20.5 to 42.2 years. BMI values varied widely (19.3–46.6 kg/m^2, mean = 29.6 kg/m^2), with 48% classified as obese (BMI $\geq$ 30 kg/m^2). In total, 56% were either overweight or obese, suggesting a high prevalence of excessive weight in this cohort. The majority of participants were Caucasian (68%), followed by South Asian (20%) and other ethnic groups (12%). In terms of parity, most were primiparous (76%) and 24% were multiparous.

Table 2.1: Basic Characteristics of the Study Cohort

Characteristic	**N (%)**
Age	
Mean ± SD (years)	30.7 ± 5.2
Range (years)	20.5–42.2
BMI Categories	
Normal weight (< 25 kg/m^2)	22 (44%)
Overweight (25–29.9 kg/m^2)	4 (8%)
Obesity Class I (30–34.9 kg/m^2)	5 (10%)
Obesity Class II (35–39.9 kg/m^2)	15 (30%)
Obesity Class III (≥ 40 kg/m^2)	4 (8%)
Ethnicity	
Caucasian	34 (68%)
South Asian	10 (20%)
Other	6 (12%)
Parity	
Primiparous	38 (76%)
Multiparous	12 (24%)
B12 Supplementation	
Yes	31 (62%)
No / Missing	19 (38%)

2.1.2 B12 Grouping Criteria and Concentration Distribution

Participants were divided into a normal B12 group (NB) and a low B12 group (LB), with 25 samples in each, resulting in a balanced 50:50 study design. Group assignment was based on a natural cutoff observed in serum vitamin B12 concentrations, ensuring clear phenotypic separation without numerical overlap.

Table 2.2: B12 Group Characteristics

Group	Sample Size	Concentration Range (pmol/L)	Mean ± SD (pmol/L)
Low B12 (LB)	25 (50%)	85.17–147.97	128.16 ± 14.68
Normal B12 (NB)	25 (50%)	196.16–558.74	361.39 ± 109.98
Overall	50 (100%)	85.17–558.74	244.77 ± 126.55

As shown in Table 2.2, the boundary between LB and NB groups was distinct (147.97 vs. 196.16 pmol/L), with no overlap in concentration values. The threshold of approximately 175 pmol/L used in this study lies within the WHO-recommended diagnostic range for vitamin B12 deficiency during pregnancy (150–200 pmol/L), enhancing the clinical interpretability of the grouping strategy.

2.1.3 Clinical Feature Comparison Between Groups

Further analysis revealed significant differences in several clinical features between the LB and NB groups, suggesting potential associations with B12 metabolic status. These differences were considered as key covariates in subsequent modeling analyses.

Table 2.3: Comparison of Clinical Features Between LB and NB Groups

Feature	LB (n=25)	NB (n=25)	Description
Age (years)	29.3 ± 5.2	32.0 ± 5.0	NB group slightly older
BMI (kg/m^2)	34.6 ± 6.3	24.6 ± 5.6	LB group significantly higher BMI
B12 Supplementation	10 (40%)	21 (84%)	Higher supplement use in NB group
Caucasian	22 (88%)	12 (48%)	More Caucasians in LB group
South Asian	3 (12%)	7 (28%)	Higher proportion in NB group
Primiparous	18 (72%)	20 (80%)	Similar parity distribution

As shown in Table 2.3, although 62% of all participants reported using B12 supplements, low B12 status remained prevalent. Notably, only 40% of participants in the LB group reported supplement use, compared to 84% in the NB group, suggesting a potential association between supplementation and improved B12 status.

BMI was significantly higher in the LB group (34.6 vs. 24.6 kg/m^2), indicating a possible link between obesity and B12 deficiency. This observation aligns with previous studies suggesting that obesity may impair B12 absorption and metabolism.

Differences in ethnic composition were also observed: 88% of the LB group were Caucasian, compared to 48% in the NB group. This may reflect underlying genetic or dietary differences in B12 metabolism. All these clinical variables were explicitly included as covariates in downstream differential expression and regression analyses.

2.1.4 Batch Distribution and Imbalance Analysis

Due to platform constraints, sequencing for all three omics layers was performed across multiple batches, resulting in imbalanced group distributions that could potentially confound downstream analyses. Therefore, batch effects were systematically evaluated and adjusted for during data processing.

As summarized in Table 2.4, several batch-specific imbalances were observed that required correction. First, **mRNA Batch 2** included only NB samples, entirely lacking LB representation, which could introduce strong confounding effects. Second, **miRNA Batch 5** showed a severe imbalance, with 8 LB samples and only 1 NB sample. Third, **DNA Batch B** consisted exclusively of LB samples, with no NB counterparts.

Table 2.4: Sequencing Batch Distribution and B12 Group Balance

Data Type	Batch ID	Total Samples	LB	NB	Balance Notes
mRNA	Batch 1	37 (74%)	19	18	Relatively balanced
	Batch 2	2 (4%)	0	2	**No LB samples**
	Batch 3	7 (14%)	4	3	Mild imbalance
	Batch 4	4 (8%)	2	2	Balanced
miRNA	Batch 1	8 (16%)	2	6	Moderate imbalance
	Batch 2	4 (8%)	2	2	Balanced
	Batch 3	6 (12%)	4	2	Moderate imbalance
	Batch 4	19 (38%)	7	12	Mild imbalance
	Batch 5	9 (18%)	8	1	**Severe imbalance**
	Batch 6	2 (4%)	1	1	Balanced
	Batch 7	2 (4%)	1	1	Balanced
DNA	Batch A	37 (74%)	21	16	Mild imbalance
	Batch B	2 (4%)	2	0	**No NB samples**
	Batch C	11 (22%)	2	9	Severe imbalance

To mitigate these issues, batch variables were explicitly modeled in downstream statistical analyses. Specifically, DESeq2 was used for both mRNA and miRNA differential expression analyses, and MethylKit for DNA methylation analysis, with batch included as a covariate. This ensured proper control for non-biological variability and minimized confounding effects during regulatory

2.2 Acquisition of Multi-Omics Data and Differential Analysis

2.2.1 Omics Profiling and Differential Expression/Methylation Analysis

This study integrated three layers of omics data: mRNA expression (RNA-seq), miRNA expression (sRNA-seq), and DNA methylation (RRBS), together with complete clinical information.

For mRNA expression, RNA-seq was performed on whole-blood samples from all 50 participants. Transcript-level quantification was normalized using TPM values, yielding 58,735 transcripts, of which 19,853 protein-coding genes were retained for downstream analysis. Differential expression analysis was conducted using the `DESeq2` package (v1.30.0), based on the negative binomial distribution, with sequencing batch and B12 group included as covariates. Significance thresholds were defined as adjusted $p < 0.05$ and $|\log_2 \mathrm{FC}| > 0.25$.

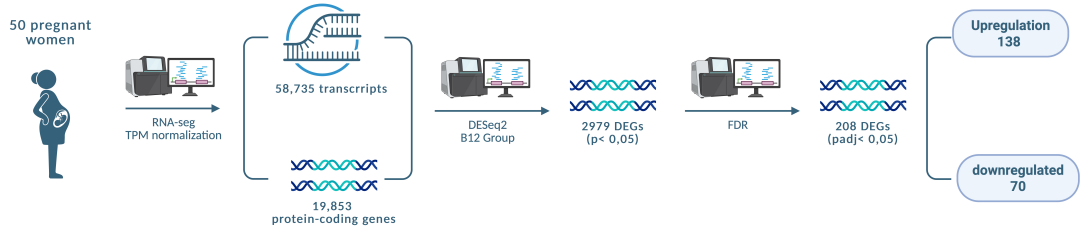

Figure 2.1: Workflow of mRNA expression profiling and differential expression analysis.

For miRNA expression, matched sRNA-seq profiling detected 2,201 miRNA species. TPM normalization was applied to generate the expression matrix. Differential analysis was also conducted using `DESeq2`, including batch and group as covariates. The threshold for statistical significance was set at adjusted $p < 0.05$ and $|\log_2 \mathrm{FC}| > 0.5$.

Figure 2.2: Workflow of miRNA expression profiling and differential expression analysis.

For DNA methylation, reduced representation bisulfite sequencing (RRBS) quantified CpG methylation as beta values (0 = unmethylated, 1 = fully methylated). Quality control yielded an average of approximately 26,650 effective CpG sites per sample with 100% completeness. Differentially methylated regions (DMRs) were identified using `methylKit` (R v3.16) and `DMRcate`, applying thresholds of $|\Delta\text{meth}| > 25\%$ and FDR < 0.05. DMRs were annotated to promoters, exons, introns, CpG islands, and intergenic regions.

Table 2.5: Summary statistics of differentially methylated regions (DMRs)

Metric	**Value**
Total significant DMRs	493,648
Hypermethylated DMRs	297,021 (60.2%)
Hypomethylated DMRs	196,627 (39.8%)
Average DMR length	999 bp
Associated unique genes	16,928
Mean CpG sites per sample	∼26,650
Coverage completeness	100%

As summarized in Table 2.5, the methylation landscape revealed a predominance of hypermethylated regions. In total, differential analyses identified 208 differentially expressed genes (DEGs; 138 upregulated, 70 downregulated), 46 differentially expressed miRNAs (2 FDR-adjusted), and 493,648 significant DMRs across the genome. These results laid the foundation for multi-omics regulatory network construction and phenotype association analyses.

2.2.2 Visualization and Annotation of Differential Features

Exploratory analyses were performed to visualize and interpret differential features across the three omics layers.

Volcano plots were used to illustrate the overall significance and effect size distribution of differentially expressed genes (DEGs) and differentially expressed miRNAs (DEmiRs) between low-B12 and normal-B12 groups. As shown in Figure 2.3, a total of 208 DEGs (adjusted $p < 0.05$, $|\log_2 \text{FC}| > 0.25$) and 46 DEmiRs (nominal $p < 0.05$; 2 FDR-adjusted) were identified.

To assess the distribution of effect sizes, fold change histograms were generated for both DEGs and DEmiRs (Figure 2.4). Most genes and miRNAs showed moderate $\log_2$ fold changes, supporting the presence of subtle but coordinated expression differences across B12 status groups.

For DNA methylation, differentially methylated regions (DMRs) were annotated to genomic features, including promoters, exons, introns, upstream regions (1–5 kb), and CpG islands. Figure 2.5 summarizes the functional distribution and methylation directionality of DMRs. Approximately 60.2% of DMRs were hypermethylated and 39.8% hypomethylated, with promoters and CpG islands being the most enriched regions.

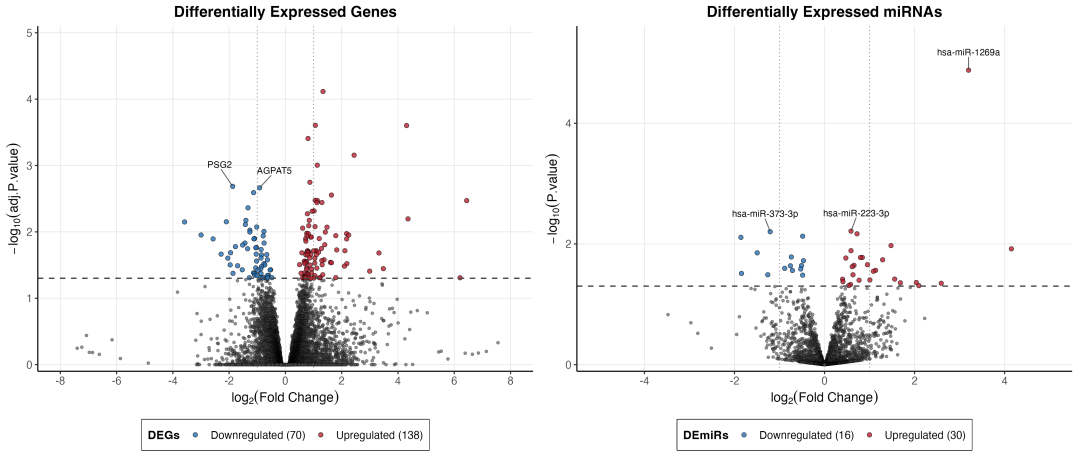

Figure 2.3: Volcano plots of differential features. Left: DEGs ($n = 208$) with adjusted $p < 0.05$, $|\log_2 \mathrm{FC}| > 0.25$. Right: DEmiRs ($n = 46$ nominal, 2 FDR-adjusted).

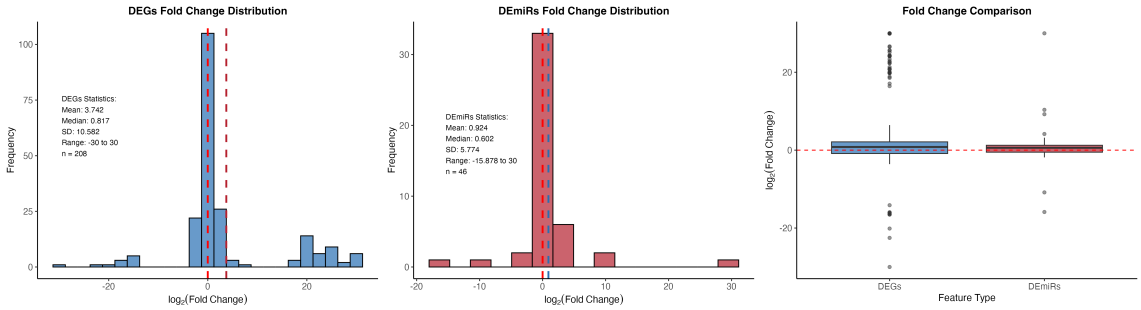

Figure 2.4: Fold change distributions of DEGs and DEmiRs. Left: histogram of $\log_2$ fold changes for DEGs. Middle: histogram for DEmiRs. Right: comparison of fold change distributions across both omics layers.

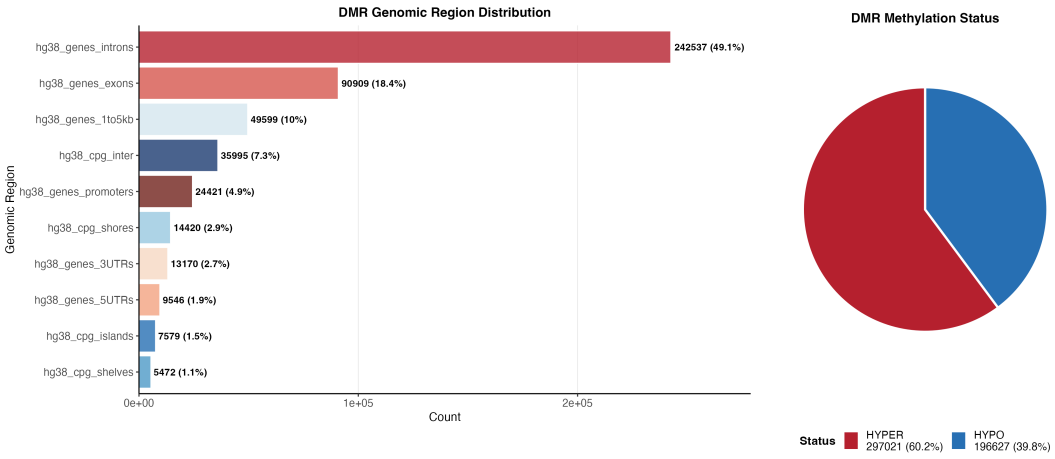

Figure 2.5: Distribution of differentially methylated regions (DMRs). Left: genomic region distribution across introns, exons, promoters, upstream 1–5 kb, and CpG islands. Right: proportions of hyper- vs. hypomethylated DMRs.

2.3 Data Cleaning and Quality Control

2.3.1 Sample Matching and Missing Data Handling

To ensure consistency, multi-omics datasets (mRNA, miRNA, and DNA methylation) were aligned with clinical metadata. Sample identifiers were standardized using a unified NTUID system by removing platform-specific suffixes. Only samples with complete data across all omics layers and clinical variables (age, BMI, parity, supplementation status, and batch) were retained. No imputation was performed; instead, samples with missing covariates were excluded to avoid introducing bias in downstream regression models. After filtering, a total of 49 matched samples remained for subsequent analyses.

2.3.2 Expression Normalization and Batch Correction

All omics data were normalized to reduce technical variation. For mRNA and miRNA, $\log_2(x+1)$ transformation and Z-score normalization were applied. CpG methylation values (0–1) were standardized directly using Z-scores.

Figure 2.6 shows the correlation heatmap across omics layers after normalization. Strong intra-layer correlations were observed for mRNA and miRNA, with moderate cross-layer associations.

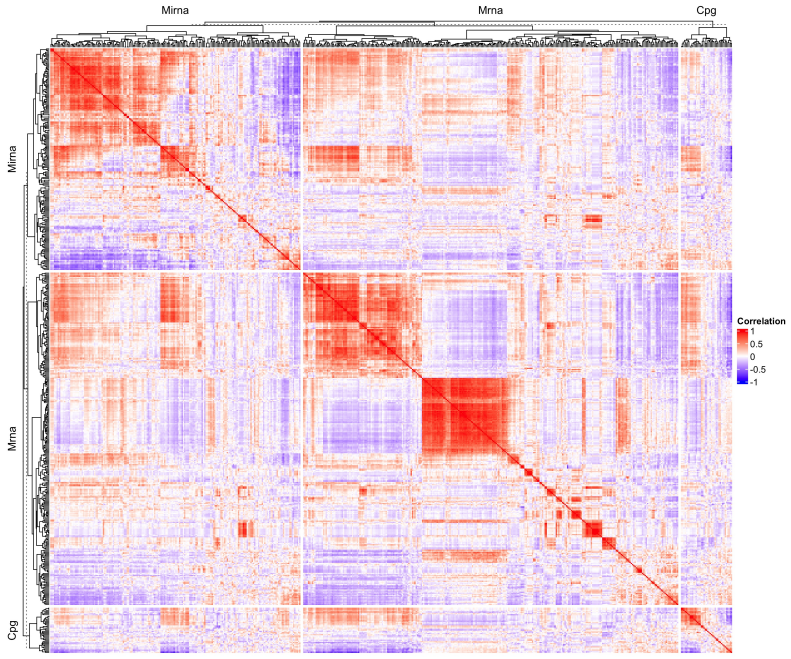

Figure 2.6: Cross-omics correlation heatmap after normalization, showing intra- and inter-layer correlation structures across CpG methylation, miRNA, and mRNA expression.

Figure 2.7 presents sample-level expression distributions. Histograms and boxplots indicate balanced values across omics layers, with no major outliers.

2.3.3 Covariate Selection and Batch Adjustment

Next, we examined clinical variables as potential confounders. As shown in Figure 2.8, BMI was significantly higher in the low-B12 (LB) group ($p < 0.001$, Cohen's $d = -1.68$). B12 supplementation ($p = 0.009$) and ethnicity ($p = 0.005$) were also significantly associated with B12 status. In addition, sequencing batches for mRNA and miRNA showed uneven distribution between groups, indicating batch effects.

To correct for these effects, we applied ComBat batch correction to all three omics layers. The model included BMI, age, parity, and supplementation status as covariates. This ensured that batch effects and clinical confounders were properly adjusted. The corrected expression matrices were then used for regulatory modeling and prediction.

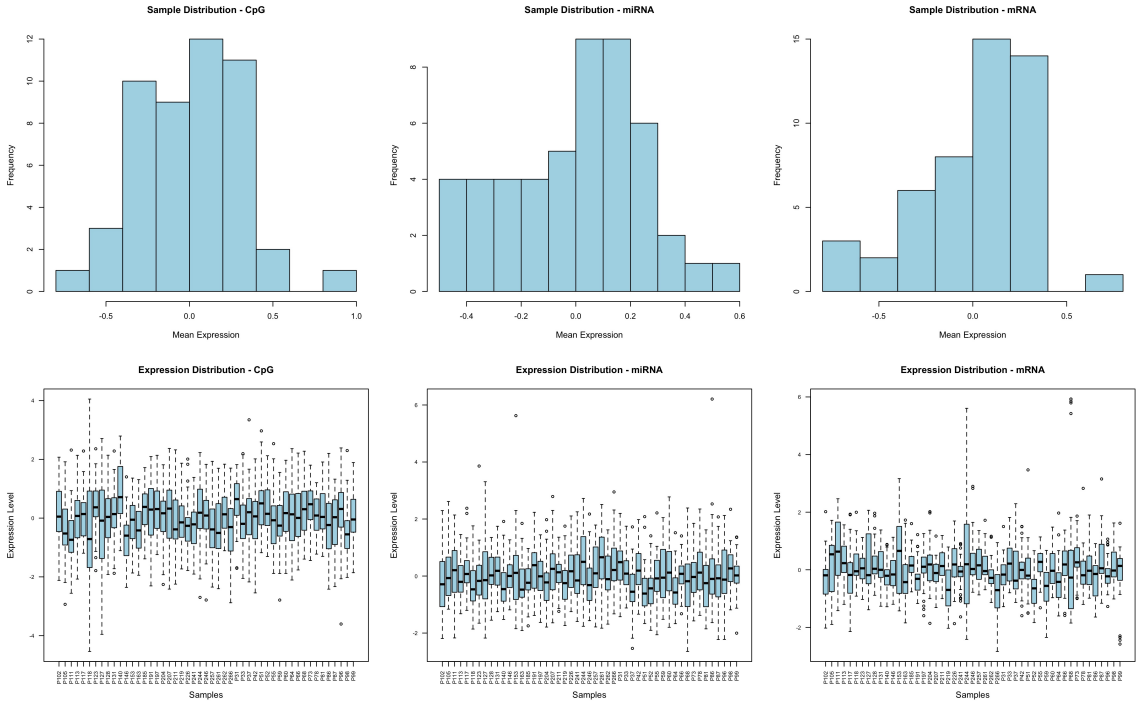

Figure 2.7: Sample-level distribution of normalized expression values. Top: mean expression histograms for CpG, miRNA, and mRNA. Bottom: boxplots of expression values across all samples.

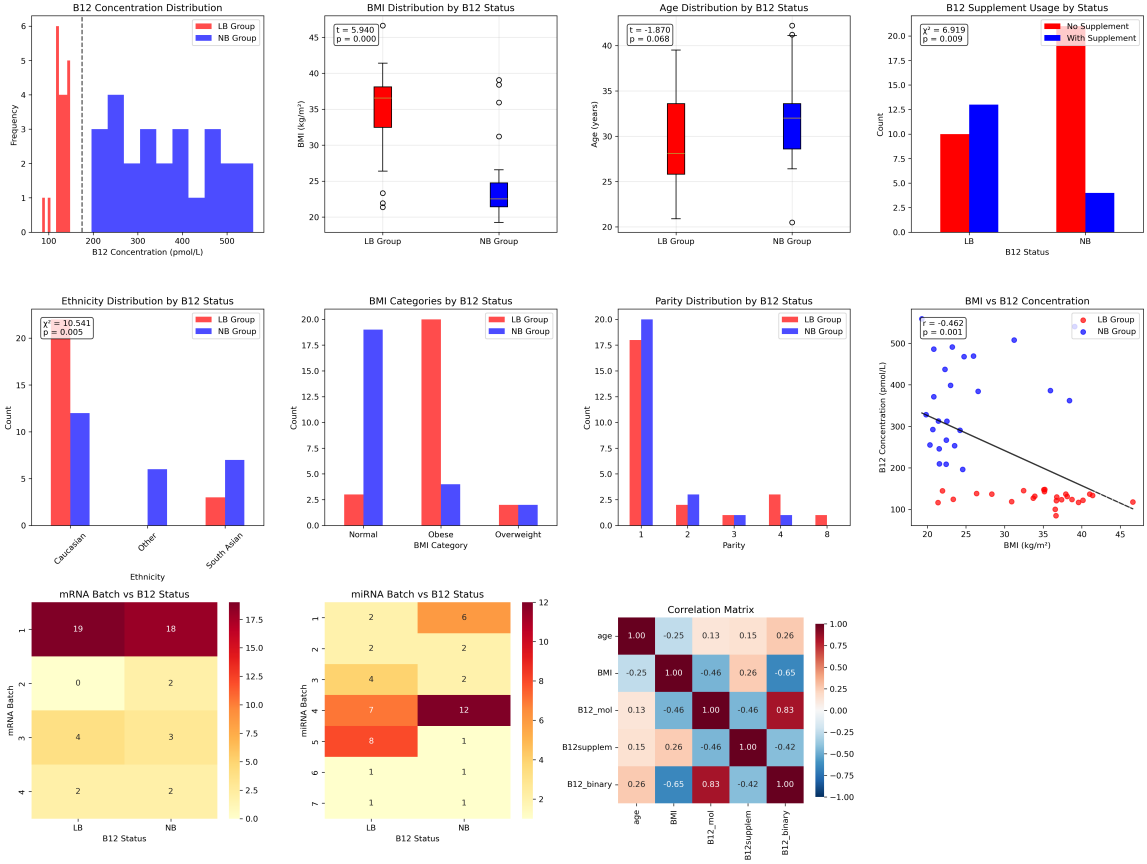

Figure 2.8: Statistical analysis of clinical variables and B12 status, showing significant associations for BMI, supplementation usage, and ethnicity distribution, as well as batch imbalance across sequencing platforms.

Chapter 3

Analytical Methods and Modeling Workflow

This chapter outlines the analytical strategies and workflow used in the study. Figure 3.1 presents an overview of the multi-omics integration process, from data preprocessing to regulatory modeling and prediction.

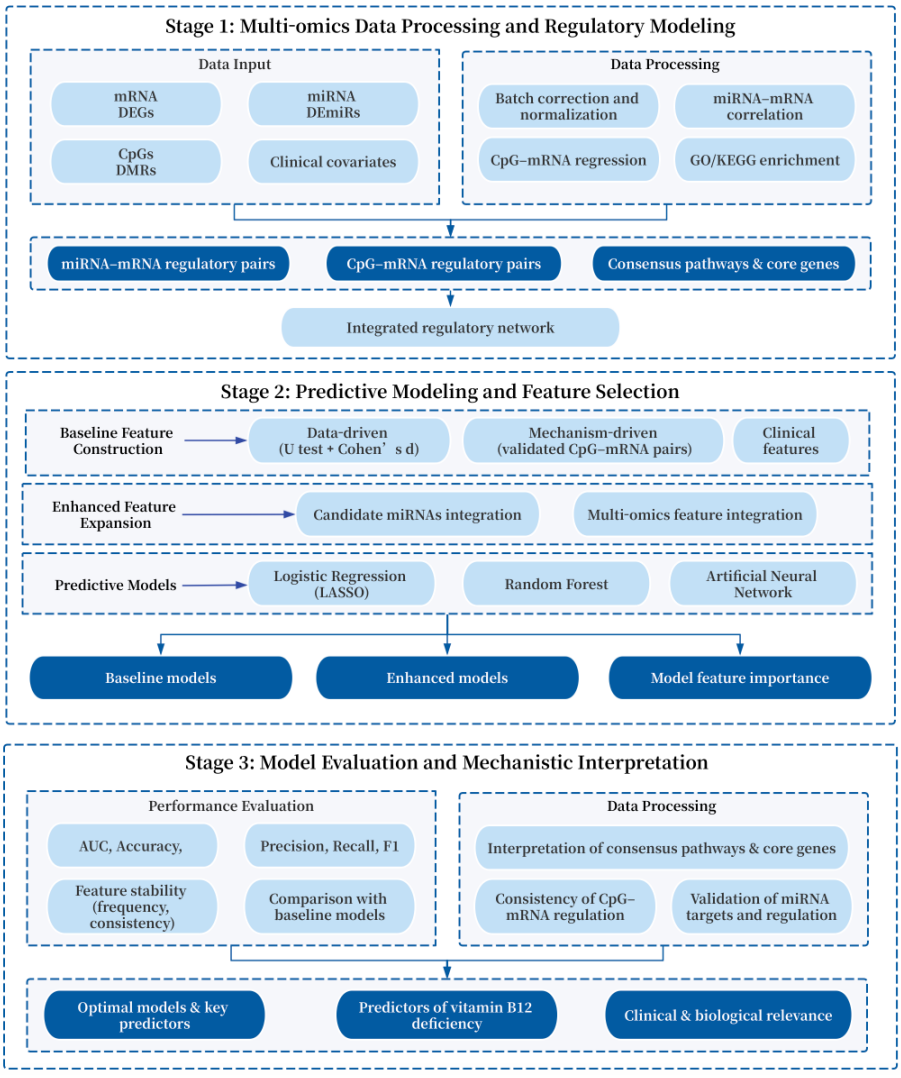

Figure 3.1: Overall workflow of multi-omics data processing, predictive modeling, and mechanistic interpretation.

3.1 Functional Enrichment and Pathway Interpretation

We performed Gene Ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) enrichment analyses to interpret the biological functions of key genes. Three gene sets were analyzed: (1) differentially expressed genes (DEGs), (2) CpG-regulated genes identified through methylation-expression analysis, and (3) genes jointly regulated by both CpG methylation and miRNA expression.

Enrichment workflow and background settings. All analyses were performed using the `clusterProfiler` package (v4.10.0). GO annotations included biological processes (BP), molecular functions (MF), and cellular components (CC). KEGG analysis focused on immune signaling, metabolism, and development-related pathways. All gene symbols were converted to Entrez IDs prior to analysis. Background sets were defined based on the total gene list of each omics type.

Pathway integration and classification. Pathways significantly enriched ($p.adjust < 0.05$) in at least two gene sets were defined as consensus pathways. Enriched genes from each pathway were mapped to their original source sets, allowing comparison across regulation layers. We further grouped consensus pathways into broader functional modules using GO hierarchies and KEGG topologies. Each output included pathway name, p-values, gene members, and regulatory sources. These results formed the basis for interpreting biological mechanisms and building downstream regulatory networks.

3.2 miRNA–mRNA Regulatory Modeling

Correlation analysis and regulatory direction. We investigated canonical miRNA–mRNA negative regulation patterns. Pearson and Spearman correlation tests were applied alongside multivariate linear regression, adjusting for batch and key covariates. A pair was considered negatively regulated if an upregulated miRNA corresponded to a downregulated target, or vice versa. Only significant expression pairs ($p < 0.05$ or FDR-adjusted) were included.

Target prediction and validation. To ensure biological credibility, predicted miRNA–mRNA interactions were cross-referenced using TargetScan, miRDB, and miRWalk. Experimentally validated interactions were collected from miRTarBase. Pairs supported by at least one database or experiment were retained. Most targets were protein-coding genes, which enhanced interpretability.

Hub miRNA identification. Core regulatory miRNAs were selected based on three criteria: number of targets, regulatory strength, and network centrality. Regulatory strength was calculated as a composite of fold-change and p-value. Strong regulators showed higher scores. Network topology was used to identify hub miRNAs with central influence. This process yielded a robust set of miRNA–mRNA relationships for pathway enrichment and integration.

3.3 CpG–mRNA Regulatory Modeling

Regression-based negative regulation. We evaluated DNA methylation–gene expression relationships using linear regression. CpG–mRNA pairs were mapped using Ensembl genome annotations and filtered for negative associations ($\beta < 0$). Models adjusted for age, BMI, parity, and smoking. Batch effects were corrected during preprocessing and not included in the regression.

Mapping and annotation. Each CpG site was annotated with its corresponding gene using HGNC-compliant symbols. Functional and genomic contexts were included (e.g., promoter, 5' UTR, exon). This mapping yielded over 16,000 CpG–mRNA candidate pairs for testing.

Statistical filtering and significance. All regression models were evaluated for significance and corrected using the Benjamini–Hochberg FDR method. Only CpG–mRNA pairs with significant negative associations (FDR ¡ 0.05) were retained. These pairs were later used in downstream regulatory network construction and mechanism-driven modeling.

3.4 Modeling Workflow

This chapter focuses on the predictive modeling of vitamin B12 deficiency during pregnancy, integrating multi-omics features and clinical covariates into machine learning frameworks. At the baseline stage, three models were implemented: logistic regression (LR), random forest (RF), and artificial neural network (ANN). Each model was tested under two feature selection strategies: a data-driven strategy and a mechanism-driven strategy. The data-driven strategy selected significant CpG sites in promoter regions

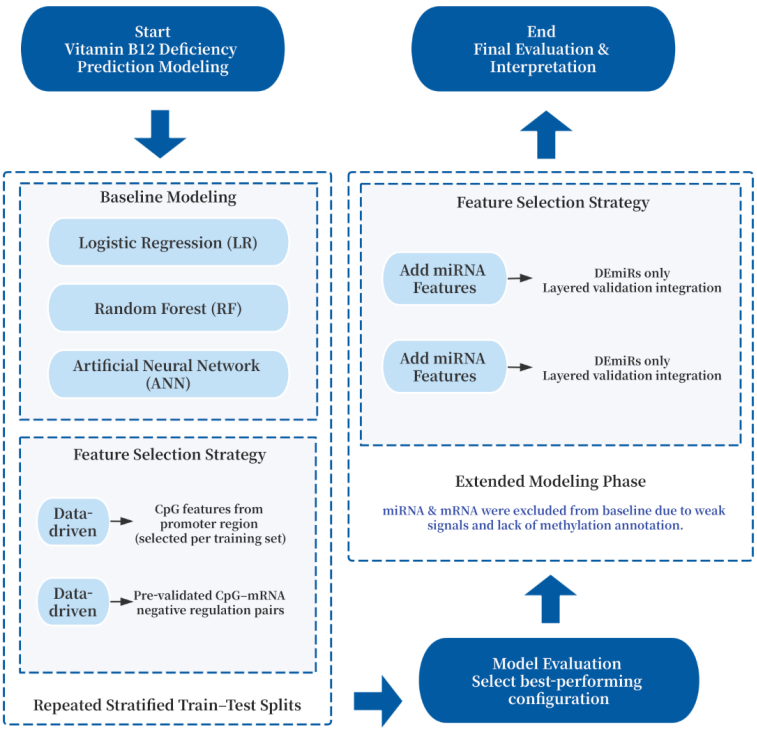

Figure 3.2: Overview of modeling workflow for vitamin B12 deficiency prediction.

within each training set, while the mechanism-driven strategy employed pre-validated CpG–mRNA negative regulatory pairs as epigenetic features. Both strategies incorporated clinical covariates and were trained under strict train–test separation with repeated stratified random splits to avoid information leakage.

As illustrated in Figure 3.2, following the baseline comparison, the best-performing configuration was selected as the core framework for the extended stage. In this phase, transcriptional and post-transcriptional signals were introduced in two steps. First, differentially expressed miRNAs were incorporated into the baseline feature set through a layered validation approach. Second, differentially expressed mRNAs were further added on top of the miRNA-enhanced model to evaluate the contribution of transcript-level features. It is important to note that miRNA and mRNA were not included in the baseline models because their differential signals were limited in number and lacked direct annotations with methylation features, making it difficult to support robust modeling. Therefore, a stepwise strategy was adopted to gradually integrate them in the extended phase.

3.4.1 Feature Selection Strategies and Validation Framework

A stratified feature selection strategy was applied, together with a rigorous validation framework, to balance statistical robustness and biological interpretability.

At baseline, the feature set included selected clinical covariates and DNA methylation features. Core clinical variables—such as age, BMI, parity, B12 supplementation, folic acid and multivitamin use, and smoking—were retained based on biological relevance and observed group differences. Covariate inclusion and preprocessing strategies are detailed in Table 3.1. DNA methylation features were selected using two complementary approaches. The data-driven strategy applied the Mann–Whitney U test to CpG sites located in promoter regions and their upstream 1–5 kb, combined with Cohen's d effect size to compute a composite ranking score. CpG sites with $p < 0.05$ and $d > 0.3$ were considered significant, and the top 10–20 sites were retained per training set. The mechanism-driven strategy used 10 predefined CpG–mRNA negative regulatory pairs identified in prior analyses, which ensured feature stability in small-sample settings. This dual approach allowed both adaptive, data-driven selection and biologically grounded feature inclusion.

At the extended stage, the baseline feature set was further expanded. First, miRNA features were added. Among 2,201 detected miRNAs, only 46 showed nominal differential expression ($p < 0.05$),

Table 3.1: Summary of covariate handling strategy

Covariate	Keep in Model	Handling Strategy and Rationale
Age	Yes	Used as a continuous variable; avoids information loss from arbitrary grouping.
BMI	Yes	Keep as a continuous variable; avoid redundancy with BMI_cat.
Parity	Yes	Converted to binary dummy variable: 0 = nulliparous, 1 = multiparous.
B12 Supplementation	Yes	Binary variable (Yes/No); significant group difference (p ¡ 0.001); major confounder.
Folic Acid Use	Yes	Binary variable (Yes/No); functionally related to B12 in one-carbon metabolism.
Multivitamin Use	Yes	Binary variable (Yes/No); may indirectly affect B12/Folic acid intake.
Smoking	Yes	Binary variable (Yes/No); potential confounder via oxidative stress pathways.
BMI_cat	No	Redundant with BMI; removed.
Ethnicity	No	Imbalanced sample size across groups; potential bias.
Batch	No	Already adjusted during omics preprocessing (e.g., ComBat); redundant to include.

and just 2 remained significant after FDR correction. To address this limitation, a stratified confidence strategy was applied: high-confidence (FDR < 0.05), medium-confidence (uncorrected $p < 0.01$), and exploratory (uncorrected $p < 0.05$) miRNAs were screened. Their predicted target genes were further validated through correlation analysis, and biologically supported miRNAs were integrated with baseline features. On this basis, differentially expressed mRNAs (208 DEGs, adjusted $p < 0.05$) were added as an additional layer to test whether transcriptional features provided further predictive or mechanistic value. The delayed inclusion of miRNA and mRNA features was motivated by their limited statistical significance and the lack of a complete CpG–miRNA or CpG–mRNA annotation framework, making them unsuitable for initial baseline modeling.

All modeling strictly followed the train–test separation principle. For each of 20 independent runs, 80% of samples were used for training and 20% for testing. Feature selection, model training, and hyperparameter tuning were performed exclusively within training sets, ensuring no information leakage. Model performance was evaluated using area under the curve (AUC), accuracy, precision, recall, and F1-score. Comparisons were made both horizontally (across different models and feature strategies) and vertically (between baseline and extended feature sets) to assess marginal performance improvements.

3.4.2 Logistic Regression Model Implementation

We implemented LASSO logistic regression to perform both binary classification and embedded feature selection, leveraging L1 regularization to induce sparsity in model coefficients. All input features, including clinical variables and multi-omics data, were preprocessed using median imputation (for missing values), dummy encoding (for binary categorical variables), and z-score standardization.

The logistic regression model estimates the probability of class membership using the sigmoid function:

$$P(y = 1 \mid \mathbf{x}) = \frac{1}{1 + \exp(-\mathbf{w}^\top \mathbf{x} - b)}, \tag{3.1}$$

where $\mathbf{x} \in \mathbb{R}^p$ is the feature vector, $\mathbf{w} \in \mathbb{R}^p$ is the coefficient vector, and b is the intercept term. Model training minimizes the following regularized loss function:

$$\mathcal{L}(\mathbf{w}, b) = -\frac{1}{n} \sum_{i=1}^{n} [y_i \log \hat{y}_i + (1 - y_i) \log(1 - \hat{y}_i)] + \lambda \|\mathbf{w}\|_1, \tag{3.2}$$

where $\hat{y}_i = P(y_i = 1 \mid \mathbf{x}_i)$, and λ is the regularization strength controlling sparsity via the L1 penalty $\|\mathbf{w}\|_1$.

Model training followed a stratified 80/20 split repeated across 20 runs to ensure robust estimation. Within each training set, 3-fold stratified cross-validation was used to select the optimal inverse regularization parameter $C \in \{0.001, 0.01, 0.1, 1.0, 10.0\}$, where $C = 1/\lambda$. The solver `liblinear` was selected to support the L1 penalty, and `class_weight=balanced` was applied to address class imbalance.

Selected features were tracked across repetitions to assess selection stability and biological relevance. The modeling procedure is formalized in Algorithm 3.1.

Algorithm 3.1: LASSO Logistic Regression with Feature Selection

Data: Dataset $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{N}$, candidate grid $\mathcal{C}$, number of repetitions K
Result: Stable feature set and trained logistic regression models

for $k = 1$ **to** K **do**
  Randomly partition $\mathcal{D}$ into $\mathcal{D}_{\text{train}}^{(k)}$ (80%) and $\mathcal{D}_{\text{test}}^{(k)}$ (20%)
  Apply median imputation, dummy encoding, and z-score normalization
  foreach $C \in \mathcal{C}$ **do**
    Train L1-penalized logistic regression using C with 3-fold CV
    Record cross-validation AUC
  end
  Select optimal C^* with highest AUC
  Retrain model on full $\mathcal{D}_{\text{train}}^{(k)}$ using C^*
  Record non-zero coefficient features
end
Compute feature selection frequency across K repetitions
return *Features selected in ≥70% of runs as stable markers*

3.4.3 Random Forest Model Construction

Random forest classifiers were trained for binary classification using an ensemble of decision trees. Each tree was constructed from a bootstrap sample of the training data, and feature splits were chosen to maximize node purity. The final prediction was obtained via majority voting across all trees.

Formally, given T decision trees $\{h_t(\mathbf{x})\}_{t=1}^{T}$, the predicted class $\hat{y}$ is:

$$\hat{y} = \text{majority\_vote}\left(h_1(\mathbf{x}), h_2(\mathbf{x}), \ldots, h_T(\mathbf{x})\right). \tag{3.3}$$

Each tree h_t is trained on a different bootstrap sample, and at each split, a random subset of features is considered to promote model diversity and reduce overfitting.

Hyperparameters were tuned via grid search for performance-efficiency balance. Candidates included: `n_estimators` $\in$ {100, 200, 300}, `max_depth` $\in$ {3, 5, 10, None}, `min_samples_split` $\in$ {2, 5}, `min_samples_leaf` $\in$ {1, 2}, `max_features` $\in$ {`sqrt`, `log2`}.

Each model was evaluated using 3-fold stratified cross-validation on the training set. Class imbalance was addressed using `class_weight=balanced` and `bootstrap=True` to ensure adequate sample representation.

Feature importance was assessed using both Gini impurity and permutation-based scores. The Gini importance reflects the cumulative reduction in impurity contributed by each feature across all trees, while the permutation importance (averaged over 10 repetitions) provides a model-agnostic estimate by evaluating performance drops after random shuffling of each feature.

The modeling workflow is summarized in Algorithm 3.2.

Algorithm 3.2: Random Forest Training with Importance Evaluation

Data: Dataset $\mathcal{D}$, hyperparameter grid $\mathcal{H}$, number of repetitions K
Result: Trained random forest models and consensus important features

for $k = 1$ **to** K **do**
  Stratified split: $\mathcal{D}_{\text{train}}^{(k)}$ (80%), $\mathcal{D}_{\text{test}}^{(k)}$ (20%)
  foreach *configuration* $h \in \mathcal{H}$ **do**
    Train random forest f_h on $\mathcal{D}_{\text{train}}^{(k)}$
    Evaluate performance via 3-fold CV (AUC)
  end
  Select best configuration h^*
  Retrain f_{h^*} on $\mathcal{D}_{\text{train}}^{(k)}$
  Compute Gini importance and permutation importance (10 runs)
end
Aggregate feature importance across K models
return *Top features by average rank*

3.4.4 Artificial Neural Network Architecture

Artificial neural networks (ANNs) were implemented as shallow three-layer feed-forward structures. The input layer dimension matched the number of selected features. The hidden layer consisted of 60% of input units (minimum 8 nodes), followed by an output layer with a single sigmoid-activated neuron for binary classification. Hidden layers used the tanh activation function, with dropout (0.2–0.5), L2 weight regularization ($\lambda \in [0.001, 0.1]$), and batch normalization to reduce overfitting.

The forward pass of a single hidden layer network can be expressed as:

$$\hat{y} = \sigma\left(\mathbf{w}_2^\top \cdot \phi(\mathbf{W}_1\mathbf{x} + \mathbf{b}_1) + b_2\right), \tag{3.4}$$

where $\mathbf{x} \in \mathbb{R}^p$ is the input vector, $\phi(\cdot)$ is the tanh activation, $\mathbf{W}_1$ and $\mathbf{w}_2$ are weights for the hidden and output layers, and $\sigma(\cdot)$ is the sigmoid function defined as:

$$\sigma(z) = \frac{1}{1 + e^{-z}}. \tag{3.5}$$

The model was trained using weighted binary cross-entropy as the loss function:

$$\mathcal{L}_{\text{BCE}} = -\frac{1}{n}\sum_{i=1}^{n}\left[w_1 \cdot y_i \log \hat{y}_i + w_0 \cdot (1 - y_i)\log(1 - \hat{y}_i)\right], \tag{3.6}$$

where w_1 and w_0 are class weights to address imbalance. The total loss included an L2 penalty:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{BCE}} + \lambda\left(\|\mathbf{W}_1\|_2^2 + \|\mathbf{w}_2\|_2^2\right). \tag{3.7}$$

Optimization used the Adam optimizer with learning rates in [0.0005, 0.005], along with gradient clipping (threshold = 1.0). Early stopping was triggered based on validation AUC (patience = 10 epochs), and learning rate decay was applied when validation loss plateaued.

Hyperparameter search was conducted over hidden units, learning rate, L2 coefficient, and dropout rate. Data were split 80/20 in a stratified manner, with 80% further subdivided for training and validation. Final models were retrained on the full 80% and evaluated on the 20% holdout test set. The training protocol is described in Algorithm 3.3.

Algorithm 3.3: Regularized ANN Training with Early Stopping

Data: Dataset $\mathcal{D}$, hyperparameter space $\mathcal{H}$, maximum epochs E, batch size B
Result: Trained ANN and selected feature importance profiles

Split $\mathcal{D}$ into $\mathcal{D}_{\text{train}}$ (80%) and $\mathcal{D}_{\text{test}}$ (20%)
Further split $\mathcal{D}_{\text{train}}$ into training and validation sets
foreach *configuration* $h \in \mathcal{H}$ **do**
    Initialize ANN with hidden layer size h_{size}, dropout h_{drop}, and L2 penalty h_{l2}
    Train for up to E epochs using Adam optimizer (learning rate h_{lr})
    Monitor validation AUC; apply early stopping (patience = 10)
    Apply learning rate decay if validation loss plateaus
end
Select best model by validation AUC
Retrain on full training set (80%)
Evaluate on test set (20%)
return *Trained ANN model and attention to omics feature contributions*

3.4.5 Performance Evaluation and Stability Analysis

Model performance was evaluated using AUC as the primary metric, complemented by accuracy, precision, recall, and F1 score. Each model–feature combination was repeated 20 times to calculate mean, variability, and confidence intervals. Feature stability was assessed by selection frequency for data-driven methods and coefficient consistency for mechanism-driven strategies. Paired t-tests and effect size analyses were applied for model comparison, while the contribution of miRNA features was evaluated by their incremental improvement over baseline models. This framework enabled the identification of robust B12 deficiency predictors and demonstrated the added value of multi-omics integration for enhancing predictive accuracy and supporting pregnancy nutrition monitoring. A summary of the evaluation metrics and their definitions is provided in Table 3.2.

Table 3.2: Definitions of Performance Evaluation Metrics

Metric	**Definition**
AUC (Area Under Curve)	Measures the ability of the model to distinguish between classes; higher AUC indicates better separability.
Accuracy	Proportion of correctly classified samples among all samples.
Precision	Proportion of true positives among all predicted positives; reflects the model's positive predictive power.
Recall (Sensitivity)	Proportion of true positives among all actual positives; measures how well the model identifies positive cases.
F1 Score	Harmonic mean of precision and recall; balances false positives and false negatives.

Chapter 4

Results and Findings

4.1 Functional Enrichment and Pathway Interpretation

To investigate the biological mechanisms underlying vitamin B12 deficiency, we performed GO and KEGG enrichment analyses on three gene sets: (1) differentially expressed genes (DEGs), (2) CpG-regulated genes annotated from differentially methylated regions (DMRs), and (3) double-regulated genes targeted by both DMRs and miRNAs.

All analyses were conducted using the `clusterProfiler` package, with statistical significance defined as $p.adjust < 0.05$ (Benjamini–Hochberg correction). Background gene sets were specified per omics layer. Detailed enrichment plots are provided in Appendix A.1, covering CpG-regulated genes (Figure A.1), DEGs (Figure A.2), and double-regulated genes (Figure A.3), highlighting both unique and convergent functional patterns across layers.

To identify common biological themes, we integrated enrichment results and identified 18 consensus pathways (Figure 4.1). These clustered into three broad domains: immune and inflammatory signaling (e.g., "regulation of inflammatory response", "JAK–STAT signaling"), hormonal and metabolic regulation (e.g., "steroid hormone biosynthesis", "cytochrome P450 metabolism"), and pathways related to extracellular matrix and vesicle dynamics (e.g., "glycosaminoglycan degradation", "lysosome"). Together, these reflect key processes potentially disrupted under B12 deficiency.

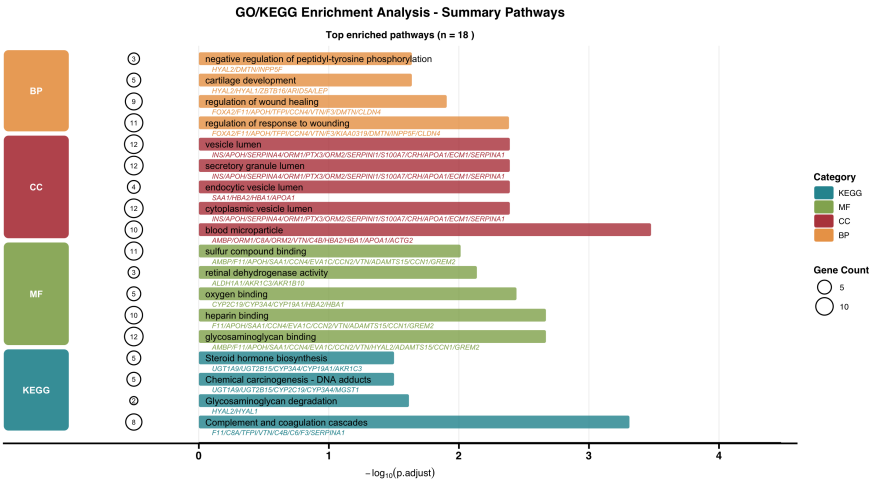

Figure 4.1: Summary of integrative GO/KEGG enrichment analysis. A total of 18 consensus pathways were enriched across DEGs, CpG-regulated, and double-regulated genes, clustering into immune, metabolic, and extracellular matrix-related modules.

In summary, these results suggest that B12 deficiency may affect immune activation, hormonal and metabolic balance, and tissue remodeling, offering mechanistic insight into its broader physiological impacts during pregnancy.

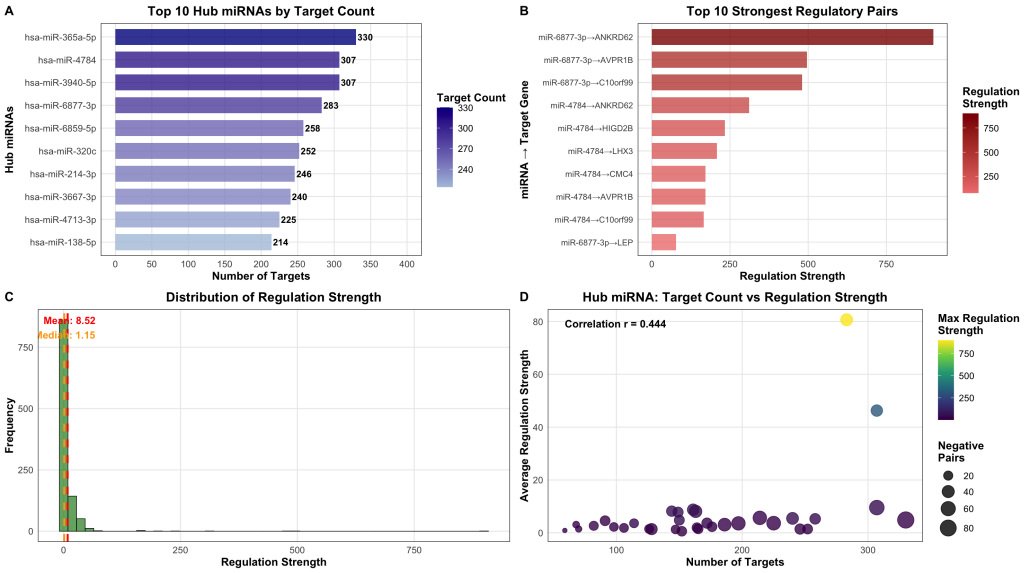

Figure 4.2: Top 10 hub miRNAs ranked by target count (A), top 10 strongest regulatory pairs (B), distribution of regulation strength values (C), and correlation between miRNA hub size and average regulatory strength (D).

4.2 miRNA–mRNA Negative Regulatory Pairs

To understand post-transcriptional regulation under vitamin B12 deficiency, we identified 1,079 significant negative miRNA–mRNA pairs, involving 186 target genes. Among these, 536 pairs featured upregulated miRNAs with downregulated targets, while 543 showed the reverse trend.

Figure 4.2A ranks the top 10 hub miRNAs based on the number of targets. *hsa-miR-365a-5p*, *hsa-miR-3940-5p*, and *hsa-miR-4784* each regulated over 200 genes, suggesting broad regulatory influence. Figure 4.2B and Table 4.1 together highlight the top 10 strongest regulatory pairs, where strength was defined as the product of absolute $\log_2$ fold changes. The pair *hsa-miR-6877-3p* and *ANKRD62* ranked highest, indicating a strong repressive interaction. Figure 4.2C presents the distribution of regulatory strength values, which displayed a right-skewed pattern (mean = 8.52, median = 1.15). Approximately 1.9% of all pairs exhibited strength values greater than one standard deviation above the mean and were defined as strong regulatory axes. In Figure 4.2D, we observed a weak correlation ($r = 0.17$) between miRNA hub size and average regulatory strength, suggesting that regulatory breadth and specificity may operate independently.

Table 4.1: Top 10 strongest miRNA–mRNA regulatory pairs based on regulation strength (product of absolute $\log_2$ fold changes).

Rank	miRNA	Target Gene	Strength	miRNA $\log_2$FC	mRNA $\log_2$FC
1	hsa-miR-6877-3p	ANKRD62	900.00	+30.00	–30.00
2	hsa-miR-6877-3p	AVPR1B	495.73	+30.00	–16.52
3	hsa-miR-6877-3p	C10orf99	480.72	+30.00	–16.02
4	hsa-miR-4784	ANKRD62	310.76	+10.36	–30.00
5	hsa-miR-4784	HIGD2B	233.33	+10.36	–22.52
6	hsa-miR-4784	LHX3	208.49	+10.36	–20.13
7	hsa-miR-4784	AVPR1B	171.13	+10.36	–16.52
8	hsa-miR-4784	CMC4	171.13	+10.36	–16.52
9	hsa-miR-4784	C10orf99	166.00	+10.36	–16.02
10	hsa-miR-6877-3p	LEP	77.10	+30.00	–2.57

In summary, several key miRNAs, particularly *hsa-miR-6877-3p* and *hsa-miR-4784*, emerged as po-

tential regulators of B12-related gene repression. These findings highlight strong candidates for further functional validation.

4.3 CpG–mRNA Negative Regulatory Pairs

We next explored transcriptional regulation via DNA methylation. Linear regression was applied to assess CpG–mRNA relationships in 49 pregnancy samples. Only significant negative regulatory pairs ($\beta < 0$) were retained for analysis.

Out of 16,168 annotated CpG–mRNA pairs, 14,471 passed quality filters and were modeled. From these, 109 pairs showed significant negative effects (FDR < 0.05), involving 109 target genes. This reflects a discovery rate of 0.75%. Though sparse, these regulatory effects were highly specific.

The estimated regression coefficients showed a strong negative bias, with an average β of –0.606. Most genes clustered in the extreme negative zone ($\beta < --0.5$, $-\log_{10}(\text{FDR}) > 2$). Genes such as *IQCG*, *ZNF154*, *TSTD1*, *GTSF1*, and *NUDT10* were among the most repressed, as illustrated in the volcano plot (Figure 4.3). We also evaluated effect size and model fit. As summarized in Table 4.2, 87.2% of significant pairs showed large effects ($|\beta| > 0.5$), with no small-effect interactions observed. The mean R^2 across all models was 0.421, indicating good explanatory power. Notably, higher $|\beta|$ values were positively correlated with R^2, suggesting that stronger effects were also more stable.

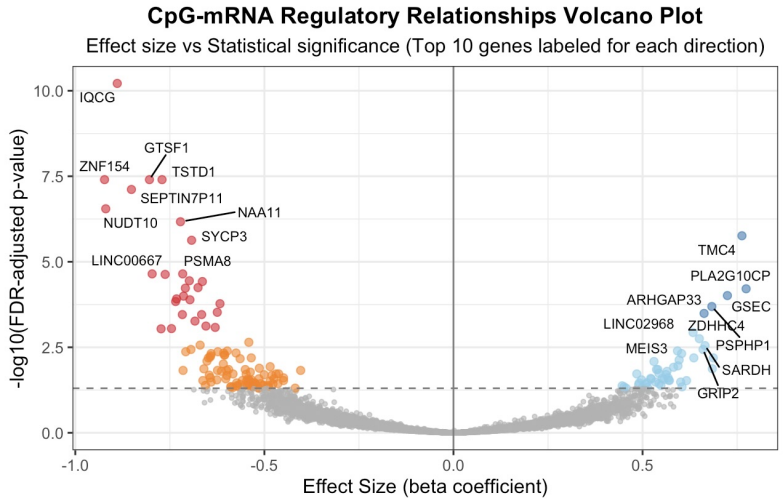

Figure 4.3: Volcano plot of CpG–mRNA regulatory relationships. Most significant pairs exhibited strong negative effects ($\beta < -0.5$ and FDR < 0.05), with top genes including *IQCG*, *ZNF154*, and *GTSF1*.

Table 4.2: Top 10 most significant CpG–mRNA negative regulatory pairs ranked by FDR. All exhibited large effects and strong model fit.

Rank	Gene	β	FDR	R^2	Effect Size
1	IQCG	–0.889	6.11×10^{-11}	0.785	Large
2	ZNF154	–0.923	3.97×10^{-8}	0.727	Large
3	TSTD1	–0.771	3.97×10^{-8}	0.702	Large
4	GTSF1	–0.804	3.97×10^{-8}	0.735	Large
5	SEPTIN7P11	–0.852	7.70×10^{-8}	0.704	Large
6	NUDT10	–0.920	2.82×10^{-7}	0.717	Large
7	NAA11	–0.722	6.75×10^{-7}	0.684	Large
8	SYCP3	–0.692	2.35×10^{-6}	0.751	Large
9	PSMA8	–0.716	2.26×10^{-5}	0.591	Large
10	LINC00667	–0.797	2.26×10^{-5}	0.550	Large

4.4 Prediction Performance of B12 Deficiency Models

4.4.1 Overall Model Performance and Feature Importance Comparison

This study developed three machine learning models to predict vitamin B12 deficiency during pregnancy. These included logistic regression (LR), random forest (RF), and artificial neural network (ANN). Each model was evaluated under two strategies: a data-driven strategy and a mechanism-driven strategy. All evaluations were based on 20 independent stratified train–test splits.

Table 4.3: Overall performance comparison of machine learning models (LR = Logistic Regression, RF = Random Forest, ANN = Neural Network; D = Data-driven, M = Mechanism-driven).

Model	Strat.	AUC	Acc.	Prec.	Rec.	F1	Rank
LR	D	0.620±0.191	59.5±16.3	61.1±21.0	58.0±19.3	59.0±19.5	6
LR	M	0.766±0.121	70.0±13.0	73.6±14.2	72.0±16.0	72.0±13.8	4
RF	D	0.698±0.212	64.5±16.0	69.2±20.6	63.0±18.2	63.9±15.0	5
RF	M	**0.774±0.133**	**74.5±12.8**	**77.6±16.2**	72.0±16.0	**73.7±13.0**	**1**
ANN	D	0.686±0.226	63.0±18.2	62.9±28.1	62.0±28.2	59.8±25.3	6
ANN	M	0.784±0.121	72.5±13.7	73.2±16.0	**74.0±18.0**	72.5±14.1	2

As shown in Table 4.3, the mechanism-driven strategy consistently outperformed the data-driven strategy across all models, achieving higher AUC, accuracy, precision, and recall. These improvements suggest that integrating biological knowledge into feature selection enhances both performance and robustness. Mechanism-driven models rely on regulatory interactions and omics coherence, rather than purely statistical associations. This helps reduce noise and improves interpretability.

Among all configurations, the mechanism-driven random forest model performed best. It achieved an average AUC of 0.774 and the highest overall accuracy. The mechanism-driven ANN achieved the highest recall (74.0%) and the highest peak AUC (0.784). Random forest likely benefited from its ensemble structure, which combines multiple decision trees. This design reduces variance and improves generalization, especially in small or heterogeneous datasets. Random forest also captures non-linear interactions and feature dependencies, which are common in multi-omics data.

To better understand model behavior, we examined which features were ranked highest across models. As shown in Figure 4.4, several features were consistently assigned high importance across multiple model types and strategies. Table 4.4 summarizes the top consensus features across all configurations. BMI and age consistently ranked among the top clinical predictors, appearing in all six models. Parity and B12 supplementation were also frequently selected, confirming their relevance to maternal B12 status.

Several methylation features also showed high importance and stability. For example, CCNG1 and KCMF1 were selected in four or more configurations. Among mechanism-derived features, IQCG, NUDT10, and GTSF1 were consistently ranked high in mechanism-driven models. These genes are linked to cell cycle control, germline development, and nucleotide metabolism—functions known to be affected by vitamin B12 levels.

In conclusion, the mechanism-driven random forest model provided the best combination of predictive accuracy, feature stability, and biological relevance. These results support its use as a practical tool for identifying pregnant women at risk of vitamin B12 deficiency.

4.4.2 Logistic Regression Model Performance

Logistic regression showed limited performance under the data-driven strategy. Across 20 train-test splits, the model achieved an average AUC of 0.620 ± 0.191 and accuracy of 59.5% ± 16.3%. However, performance varied widely. AUC ranged from 0.320 to 0.920, and only 25% of splits exceeded an AUC of 0.8 (Table 4.5).

Feature stability was also low. All clinical covariates were consistently selected, but only one methylation feature, *LOC102725048*, showed high stability. Other CpG features, such as *KCMF1*, *SLF2*, *RHBDF2*, and *CCNG1*, appeared inconsistently, with selection frequencies below 50%.

In contrast, the mechanism-driven logistic regression model performed much better. The average AUC increased to 0.766 ± 0.121, and accuracy rose to 70.0% ± 13.0%. Precision and recall improved to 73.6% and 72.0%, respectively. The coefficient of variation dropped by nearly half, indicating greater

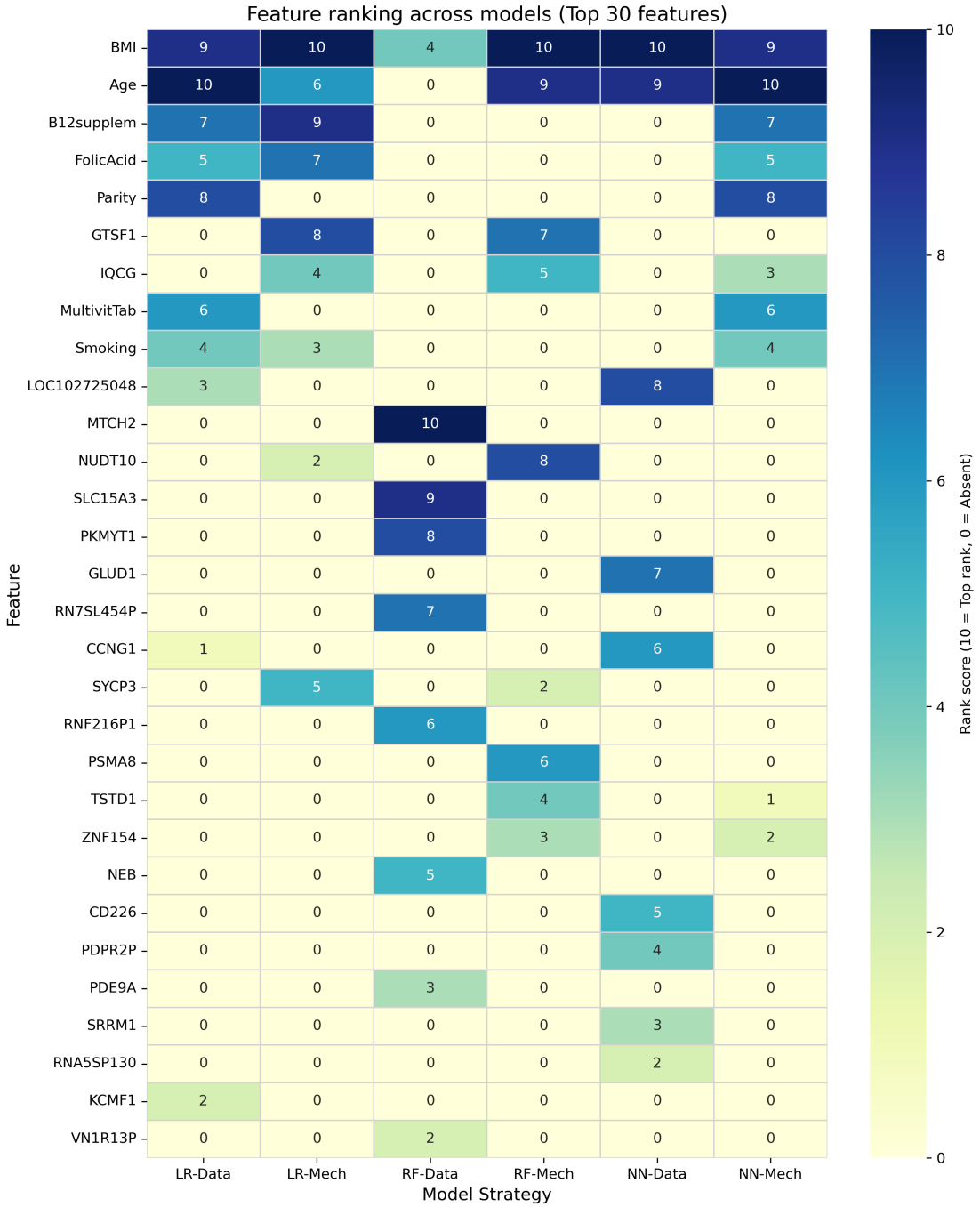

Figure 4.4: Heatmap of feature ranking across models (Top 30 features). Darker colors indicate higher feature importance (0 = absent, 10 = top rank).

Table 4.4: Consensus features across models and strategies.

Rank	LR-Data	LR-Mech	RF-Data	RF-Mech	ANN-Data	ANN-Mech
1	Age	BMI	MTCH2	BMI	BMI	Age
2	BMI	B12supplem	SLC15A3	Age	Age	BMI
3	Parity	GTSF1	PKMYT1	NUDT10	LOC102725048	Parity
4	B12supplem	FolicAcid	RN7SL454P	GTSF1	GLUD1	B12supplem
5	MultivitTab	Age	RNF216P1	PSMA8	CCNG1	MultivitTab
6	FolicAcid	SYCP3	NEB	IQCG	CD226	FolicAcid
7	Smoking	IQCG	BMI	TSTD1	PDPR2P	Smoking
8	LOC102725048	Smoking	PDE9A	ZNF154	SRRM1	IQCG
9	KCMF1	NUDT10	VN1R13P	SYCP3	RNA5SP130	ZNF154
10	CCNG1	SEPTIN7P11	RHBDF2	LINC00667	ZCCHC10P1	TSTD1

robustness. More splits achieved high performance: 35% exceeded AUC 0.8, and 20% exceeded 0.9. Because feature sets were predefined, selection was fully stable across all runs.

Table 4.5: Performance comparison of logistic regression models

Metric	Data-driven	Mechanism-driven	Difference
Mean AUC	0.620±0.191	0.766±0.121	+23.5%
Accuracy	59.5±16.3%	70.0±13.0%	+17.6%
Precision	61.1±21.0%	73.6±14.2%	+20.5%
Coefficient of variation	30.8%	15.8%	-48.7%
Splits with AUC > 0.8	5/20	7/20	+40%

Coefficient analysis showed partial inconsistency with known regulatory directions. Only three features—*IQCG*, *SEPTIN7P11*, and *SYCP3*—had regression signs consistent with their original CpG–mRNA effects. In contrast, *ZNF154*, *GTSF1*, and *NUDT10* displayed opposite effects (Table 4.6). This suggests that while biological priors improved model accuracy, statistical coefficients may still deviate from mechanistic expectations.

Among clinical variables, BMI was the most important predictor (coefficient = -1.046). B_{12} supplementation was protective (coefficient = -0.759), while maternal age showed a positive association with deficiency risk (coefficient = 0.248). These trends align with known nutritional patterns.

In summary, the mechanism-driven logistic regression model significantly improved accuracy and robustness. However, mismatches between model coefficients and biological expectations highlight the limits of statistical learning in capturing mechanistic signals.

4.4.3 Random forest model performance

The data-driven random forest model performed moderately well but showed high variability. Its average AUC was 0.698 ± 0.212, and accuracy reached 64.5% ± 16.0%. AUC ranged widely, from 0.160 to 1.000. Notably, 20% of splits performed below the random baseline (AUC ¡ 0.5), and the coefficient of variation was the highest among all models (30.4%).

BMI consistently ranked as the most important feature, with a mean Gini importance of 0.088 ± 0.033. Among methylation features, *CCNG1* was selected in 95% of splits, followed by *ZNF747-DT* (60%) and *KCMF1* (55%). Other features showed low stability.

In contrast, the mechanism-driven random forest outperformed its data-driven counterpart in every evaluation metric. As shown in Table 4.7, the average AUC increased to 0.774 ± 0.133, and accuracy rose to 74.5% ± 12.8%. Precision improved to 77.6% ± 16.2%, and performance became more stable, with the coefficient of variation reduced to 17.2%. Half of the splits achieved AUC above 0.8, and 20% exceeded 0.9.

Feature importance in the mechanism-driven model was also more consistent. BMI remained the

Table 4.6: Consistency of coefficients in mechanism-driven logistic regression

Gene	Original Effect (β)	LR Coefficient	Consistency	Stability
IQCG	-0.889	-0.098±0.027	Consistent	High
SEPTIN7P11	-0.852	-0.049±0.008	Consistent	Medium
SYCP3	-0.692	-0.125±0.011	Consistent	High
ZNF154	-0.923	+0.187±0.029	Inconsistent	Low
GTSF1	-0.804	+0.503±0.021	Inconsistent	Medium
NUDT10	-0.920	+0.316±0.024	Inconsistent	Medium
TSTD1	-0.771	-0.074±0.016	Consistent	Low
NAA11	-0.722	+0.002±0.006	Inconsistent	Low
PSMA8	-0.716	-0.011±0.004	Consistent	Low
LINC00667	-0.797	-0.026±0.008	Consistent	Low

Note: "Consistency" indicates whether the model coefficient matches the original regulatory direction. "Stability" refers to selection frequency across splits.

Table 4.7: Performance comparison and feature importance of random forest models

Metric	Data-driven	Mechanism-driven	Difference
Mean AUC	0.698±0.212	0.774±0.133	+10.9%
Accuracy	64.5±16.0%	74.5±12.8%	+15.5%
Precision	69.2±20.6%	77.6±16.2%	+12.1%
Coefficient of variation	30.4%	17.2%	–43.4%
Splits with AUC > 0.8	8/20	10/20	+25%

top predictor, followed by age and several CpG-regulated genes including *IQCG*, *NUDT10*, and *GTSF1*. These features were frequently selected across splits and aligned well with the results from prior regulatory analysis, supporting their relevance to vitamin B_{12} metabolic mechanisms.

Overall, the mechanism-driven random forest model demonstrated the best combination of predictive accuracy, reduced variance, and biological interpretability, making it the strongest and most reliable model in this study.

4.4.4 Artificial Neural Network Model Performance

The data-driven neural network model showed high variability. It achieved an average AUC of 0.686 ± 0.226 and accuracy of 63.0% ± 18.2%. AUC values ranged from 0.240 to 1.000. While 35% of splits exceeded 0.8, 20% fell below 0.5. This instability limited its clinical utility.

Feature stability was low. Only *KCMF1* and *CCNG1* appeared in 70% of splits. *LAMA3* reached 50%, while other features varied widely.

The mechanism-driven neural network showed better performance. As summarized in Table 4.8, its average AUC improved to 0.784 ± 0.121, and accuracy reached 72.5% ± 13.7%. Recall rose to 74.0% ± 18.0%, the highest among all models. Stability also increased, with a reduced coefficient of variation of 15.4%.

Table 4.8: Performance comparison of neural network models

Metric	Data-driven	Mechanism-driven
Mean AUC	0.686±0.226	0.784±0.121
Accuracy	63.0±18.2%	72.5±13.7%
Recall	62.0±28.2%	74.0±18.0%
Coefficient of variation	32.9%	15.4%
Feature stability	2 high-stability features	Fully stable

Hyperparameter optimization revealed that simpler architectures performed better. Most optimal models used 8 hidden units (about 60% of input size). Learning rates clustered around 0.0005, with L2 regularization typically set at 0.001. Dropout ranged from 0.2 to 0.5. Models generally required 30 or more iterations, and early stopping was triggered in most runs.

Although the mechanism-driven ANN achieved the highest peak AUC, its sensitivity to hyperparameters and training instability limited its clinical applicability. Validation–test AUC correlations were stronger in mechanism-driven models (r = 0.76 vs. 0.68), indicating better generalization. Still, for small-sample prediction tasks, the mechanism-driven random forest offered a better balance between accuracy, stability, and interpretability.

4.5 Model optimization

After constructing and evaluating the baseline predictive models, we further explored the potential of multi-omics feature integration. The primary objective was to determine whether incorporating miRNA and mRNA expression features could enhance the performance of the mechanism-driven random forest model.

The optimization process strictly adhered to the principle of training–testing separation. All feature selection, model training, and hyperparameter tuning were performed exclusively within the training sets, and 20 independent stratified random splits were used to ensure robustness and generalizability.

Differential expression analysis identified 46 miRNAs with nominal significance among the 2,201 detected species, of which only two remained significant after false discovery rate (FDR) correction. To address this limitation, a three-tier confidence strategy was applied: high confidence (FDR < 0.05), moderate confidence (uncorrected $p < 0.01$), and exploratory (uncorrected $p < 0.05$). These candidate miRNAs were then combined with clinical variables and mechanism-driven methylation features to construct the enhanced feature set. Similarly, the top 30 most significant differentially expressed genes (DEGs) were selected based on adjusted p-values and $\log_2$ fold-change to form the final multi-omics input.

Given the expanded feature space, the model complexity was adjusted by increasing the number of trees in the random forest to 300 and setting the maximum tree depth to 7. This configuration was chosen to balance model capacity and the risk of overfitting.

As shown in Table 4.9, neither the addition of miRNA features nor the further inclusion of mRNA features led to improvements in model performance. The base model (clinical + methylation features) achieved an average AUC of 0.798, which remained unchanged after miRNA integration. Adding mRNA features slightly reduced the AUC to 0.772, along with minor decreases in accuracy and recall. Feature importance analysis further revealed that the cumulative contribution of the newly added transcriptomic features was below 5%, which was substantially lower than that of BMI, age, and CpG-regulated genes.

Several factors may explain this limited improvement. First, the small number of differentially expressed miRNAs and the modest sample size constrained their predictive power. Second, mechanism-driven methylation features may have already captured the essential molecular signatures associated with vitamin B_{12} deficiency. Third, miRNA and mRNA regulation is often tissue-specific and temporally dynamic, which may not be fully captured in peripheral blood samples collected at a single time point.

Although the inclusion of miRNA and mRNA features did not enhance predictive accuracy, the enhanced and multi-omics models maintained stability comparable to the baseline model, confirming the robustness of the hierarchical feature integration strategy. Importantly, these results highlight that increasing the number of features does not necessarily improve model performance. The key determinant remains the biological relevance and quality of the input features. From a clinical perspective, the baseline model—composed of clinical and mechanism-driven methylation features—already demonstrates strong predictive performance and interpretability, making it a practical tool for identifying pregnant women at risk of vitamin B_{12} deficiency.

Table 4.9: Performance comparison of random forest models under different feature integration strategies

Model	**AUC**	**Accuracy**	**Precision**	**Recall**
Base model (clinical + methylation)	0.798	0.710	0.745	0.640
Enhanced model (base + miRNA)	0.798	0.710	0.745	0.640
Multi-omics model (enhanced + mRNA)	0.772	0.685	0.745	0.610

Chapter 5

Conclusion

5.1 Main Findings

This study systematically investigated the molecular phenotypes and predictive modeling of vitamin B_{12} deficiency during pregnancy by integrating transcriptomic (mRNA), small RNA (miRNA), and DNA methylation data. Under rigorous covariate adjustment and batch correction, we performed differential feature identification, regulatory network construction, and mechanism-driven predictive modeling. Using peripheral blood samples from 50 pregnant women within the same cohort, this study established a consistent multi-omics framework that enabled comprehensive characterization of B_{12} deficiency at the molecular level, while also exploring prediction strategies under small-sample conditions with the incorporation of mechanistic priors.

In differential feature identification, a total of 208 significant differentially expressed genes (DEGs) were identified at the mRNA level, including 138 upregulated and 70 downregulated transcripts, most of which exhibited strong effect sizes. For miRNAs, only 46 candidates passed initial screening among 2,201 detected molecules, and after multiple testing correction only 2 remained significant. This highlights the sparsity of miRNA signals under limited sample sizes and underscores the need for target prediction and correlation validation to improve reliability. At the DNA methylation level, 493,648 differentially methylated regions (DMRs) were identified genome-wide, spanning promoters, exons, introns, upstream regulatory regions, and CpG islands. Within promoter and upstream 5 kb windows, 100 significant DMRs were further refined, providing multi-layered evidence from global to local levels and establishing the foundation for regulatory analysis and mechanistic interpretation.

Functional enrichment revealed a highly consistent "immunity–metabolism–development" axis. Key pathways included immune regulation processes such as JAK–STAT signaling and inflammatory responses, metabolic pathways like steroid hormone biosynthesis and cytochrome P450 metabolism, and structural processes such as glycosaminoglycan degradation and lysosomal activity. These findings suggest that B_{12} deficiency may influence pregnancy-related biological processes through immune activation, metabolic disruption, and tissue remodeling.

In regulatory mechanism modeling, a total of 1,079 miRNA–mRNA negative regulatory relationships were confirmed, involving 186 target genes. Central hub miRNAs such as *hsa-miR-4784* and *hsa-miR-6877-3p* exhibited strong regulatory strength and broad target ranges. On the DNA methylation side, among 16,168 annotated CpG–mRNA pairs, 109 were identified as significant with negative regulatory direction using linear regression (FDR $<$ 0.05). The overall effects were predominantly negative (average $\beta \approx -0.606$), with 87.2% classified as large effect sizes ($|\beta| > 0.5$). The average R^2 was 0.421, indicating that a single CpG site could explain over 40% of the expression variance of its target gene. Key genes such as *IQCG*, *ZNF154*, *TSTD1*, *GTSF1*, and *NUDT10* exhibited strong negative effects and high model fit, with 9 pairs exceeding $R^2 > 0.6$, suggesting robust and biologically credible epigenetic regulation pathways. These core genes were further enriched in functional modules related to transcription factor activity, RNA processing, protein modification, and cytoskeletal regulation—highly consistent with the known roles of vitamin B_{12} in embryonic development, metabolic homeostasis, and methylation control.

In predictive modeling, mechanism-driven strategies consistently outperformed data-driven strategies. The mechanism-driven random forest achieved the best balance of average performance and stability (AUC = 0.774 ± 0.133), while the mechanism-driven neural network achieved the highest peak performance (AUC = 0.784 ± 0.121). Both models clearly outperformed their data-driven counterparts, demonstrating that incorporating mechanistic priors improves both performance and stability under

small-sample conditions. Cross-model consistency analysis further showed that body mass index (BMI) was the most stable clinical predictor. It ranked highly across all models and displayed a negative association with B_{12} deficiency risk, aligning with nutritional expectations. Age also consistently appeared across models, reflecting its general role in the decline of B_{12} metabolism. For methylation features, data-driven methods repeatedly identified *CCNG1*, *KCMF1*, and *LOC102725048*, showing high stability across partitions. Mechanism-driven methods highlighted genes such as *GTSF1*, *IQCG*, *NUDT10*, and *SYCP3*, which not only had high weights in model training but also aligned strongly with CpG–mRNA regulatory evidence. These features formed a core set combining predictive power with mechanistic interpretability. Notably, some mechanistic features exhibited regression coefficients in logistic regression models that contradicted the expected negative regulatory direction. This suggests that purely statistical learning may deviate from biological mechanisms and that future models may benefit from coefficient constraints or mechanism-weighted approaches.

Importantly, the subsequent optimization analysis revealed that adding miRNA and mRNA features did not improve predictive performance compared to the baseline model of clinical and methylation features. Both enhanced and multi-omics models maintained similar stability but showed no measurable gains in accuracy, precision, recall, or AUC. In fact, the integration of mRNA features slightly decreased AUC (from 0.798 to 0.772). Feature importance analysis indicated that the cumulative contribution of miRNA and mRNA features was negligible, accounting for less than 5% of the total predictive power—far below that of BMI, age, and key CpG–mRNA features. These results emphasize a critical principle in multi-omics modeling: feature quantity does not equate to predictive quality. Instead, biological relevance and signal strength are decisive factors.

Overall, this study leveraged cohort-based multi-omics data to systematically characterize vitamin B_{12} deficiency. From differential feature discovery and pathway analysis to regulatory modeling and predictive modeling, it revealed a consistent "immunity–metabolism–development" axis and identified a stable set of features including BMI, age, and core methylation signals. This integrative framework achieved a balance between predictive performance, stability, and interpretability under small-sample conditions, offering a methodological pathway and translational potential for early risk prediction and mechanistic understanding of B_{12} deficiency during pregnancy.

5.2 Innovations and Contributions

This study introduces a series of methodological and application-level innovations in exploring molecular mechanisms and building predictive models for vitamin B12 deficiency during pregnancy.

First, at the level of study design, this work breaks through the traditional limitation of nutritional epidemiology that focuses on single biomarkers. By integrating transcriptomic (mRNA), small RNA (miRNA), and DNA methylation data with clinical variables, a multi-omics regulatory framework was established. Unlike prior studies that relied on statistical associations or single-omics analyses, this study combined negative regulation filtering, functional annotation, and network modeling to build an epigenetic network centered on CpG–mRNA interactions. This framework offers a more interpretable basis for understanding the biological regulation underlying B12 deficiency during pregnancy.

Second, in feature engineering and modeling, a staged integration strategy was proposed. Initially, a baseline model was built using clinical variables and differentially methylated regions. This model was then expanded by introducing differentially expressed miRNAs, creating a hierarchical feature structure. Additionally, CpG–mRNA pairs identified through mechanistic analysis were directly included in the models. These features were selected not just for statistical relevance, but for their biological plausibility. This mechanism-driven approach helped address instability issues common in data-driven models, especially under small-sample conditions, while enhancing the interpretability of the results.

Third, at the modeling and evaluation stage, three representative classifiers—logistic regression, random forest, and artificial neural network—were systematically compared. Each model was tested under both data-driven and mechanism-driven strategies. Among them, the mechanism-driven random forest showed the best average performance and stability, while the mechanism-driven neural network achieved the highest peak performance. This comprehensive comparison framework enhances robustness and offers empirical insights into choosing suitable modeling approaches for similar research contexts. Performance across 20 independent training-test splits also revealed notable variability, highlighting the need for repeated validation when working with small sample sizes.

Finally, a cross-model feature consistency analysis was conducted to identify stable and biologically meaningful predictors. Among clinical variables, BMI and age consistently ranked as top predictors, reaffirming their central role in maternal nutrition and metabolism. For methylation features, CCNG1 and

LOC102725048 frequently emerged in data-driven models, while IQCG, NUDT10, GTSF1, and SYCP3 were highlighted through mechanism-driven selection. This dual-track outcome—combining data-driven discovery with mechanistic validation—offers a reliable pathway for identifying biomarkers that are both predictive and biologically interpretable.

In summary, the key innovations of this study include:

- Proposing a staged feature integration and mechanism-driven modeling strategy to balance performance and interpretability under small-sample conditions;
- Conducting systematic multi-model and multi-strategy comparisons to evaluate robustness and guide model selection;
- Identifying a consistent set of features across models, strengthening both predictive stability and biological relevance.

Together, these contributions go beyond advancing early prediction of vitamin B12 deficiency. They offer a methodological framework for integrating predictive modeling and mechanistic interpretation in small-sample multi-omics studies, with broad implications for nutritional and systems biology research.

5.3 Recommendations for Future Research

This study combined multi-omics data with clinical variables to build predictive models for vitamin B12 deficiency during pregnancy. It also explored possible molecular mechanisms. Despite these advances, some goals remain unmet due to limitations in sample size, omics coverage, and modeling design. These limitations point to clear directions for future research.

The most pressing issue is the limited sample size. Only 50 participants were included, which led to unstable model performance. Some data splits performed no better than random guessing. This highlights how small datasets can reduce the reliability and generalizability of machine learning. To address this, future studies should recruit larger cohorts, ideally across multiple centers. This would allow for external validation and improve population-level relevance. Evaluation strategies should also include leave-batch or leave-center validation. These can help identify and adjust for batch effects, making models more robust to different populations and lab conditions.

Another important gap lies in feature integration. This study originally aimed to build a CpG–miRNA–mRNA mediation model. The goal was to investigate how DNA methylation might regulate gene expression indirectly through miRNAs. However, this framework could not be completed. Two factors were responsible: too few significant miRNAs were detected, and existing databases lacked annotations linking promoter methylation to miRNA regulation. As a result, miRNAs were only included as features in the later stages of modeling. Future work should focus on collecting multi-omics data that includes miRNA methylation information. This would enable more complete mediation analyses and support reconstruction of full regulatory chains. It would also improve both interpretability and predictive power.

The feature optimization analysis provided another insight. Adding miRNA and mRNA features did not significantly improve model performance over the baseline. Several reasons may explain this. First, the number of meaningful miRNA signals was limited. Second, the sample size constrained model learning. Third, the CpG–mRNA features may have already captured most of the important biological variation. These results suggest that simply adding more features is not enough. Future research should focus on increasing sample size and improving feature quality. New omics layers should provide independent and non-redundant information. Longitudinal study designs and tissue-specific data collection could also help. These strategies would capture regulatory changes over time and in relevant biological contexts—insights that were not possible in this study's single-time-point, peripheral blood data.

Modeling strategies also offer room for improvement. Although both data-driven and mechanism-driven feature selection methods were tested, mechanistic knowledge was not formally built into the models. Known regulatory patterns, such as "increased methylation leads to decreased expression," were sometimes reversed in the results. This mismatch suggests a need for stronger biological constraints in future models. Techniques such as coefficient sign constraints, weighted sparsity, or monotonic penalties can help align model outputs with biological expectations. For complex models like neural networks, attention layers or specialized regularization could better capture regulatory relevance while maintaining performance.

Finally, model evaluation and clinical translation deserve more attention. This study focused mainly on classification metrics such as AUC and accuracy. However, it did not fully assess calibration or clinical

utility. Future studies should include metrics like the Brier score, calibration curves, and decision curve analysis. These provide a more complete picture of model reliability and usefulness. On the clinical side, researchers should evaluate the feasibility and cost-effectiveness of proposed biomarkers. Lightweight risk scores that combine a few key molecular and clinical features could support real-world screening. These tools should be simple enough for use in pregnancy care and early intervention programs.

In summary, future research should focus on four key areas:

- Expanding sample size to improve model stability and generalizability;
- Enhancing multi-omics data collection, especially incorporating underrepresented layers such as miRNA methylation;
- Embedding biological knowledge directly into modeling strategies through mechanism-informed constraints and feature selection;
- Broadening model evaluation to include clinical relevance, such as calibration, net benefit, and translational feasibility.

These improvements will help produce models that are more accurate, stable, interpretable, and useful in clinical settings. Ultimately, this will support earlier detection and better management of vitamin B12 deficiency during pregnancy.

Bibliography

[1] Antonysunil Adaikalakoteswari, Manu Vatish, Mohammad Tauqeer Alam, Sascha Ott, Sudhesh Kumar, and Ponnusamy Saravanan. Low vitamin b12 in pregnancy is associated with adipose-derived circulating mirs targeting pparγ and insulin resistance. *The Journal of Clinical Endocrinology & Metabolism*, 102(11):4200–4210, 2017.

[2] Antonysunil Adaikalakoteswari, Catherine Wood, Theresia H Mina, Craig Webster, Ilona Goljan, Yonas Weldeselassie, Rebecca M Reynolds, and Ponnusamy Saravanan. Vitamin b12 deficiency and altered one-carbon metabolites in early pregnancy is associated with maternal obesity and dyslipidaemia. *Scientific Reports*, 10(1):12548, 2020.

[3] Maider Aguerralde-Martin, Mónica Clemente-Císcar, Luis Lopez-Cárcel, Ana Conesa, and Sonia Tarazona. More: Interpretable multi-omic regulatory networks to characterize phenotypes. *bioRxiv*, pages 2024–01, 2024.

[4] Lindsay H Allen. Causes of vitamin b12 and folate deficiency. *Food and Nutrition Bulletin*, 29(2_suppl1):S20–S34, 2008.

[5] Charlotte L Ars, Ilse M Nijs, Hanan E Marroun, Ryan Muetzel, Marcus Schmidt, Jolien Steenweg-de Graaff, Aad van der Lugt, Vincent W Jaddoe, Albert Hofman, Eric A Steegers, Frank C Verhulst, Henning Tiemeier, and Tonya White. Prenatal folate, homocysteine and vitamin b12 levels and child brain volumes, cognitive development and psychological functioning: the generation r study. *British Journal of Nutrition*, 115(7):1245–1253, 2016.

[6] Y Ba, H Yu, F Liu, X Geng, C Zhu, Q Zhu, T Zheng, S Ma, G Wang, Z Li, and Y Zhang. Relationship of folate, vitamin b12 and methylation of insulin-like growth factor-ii in maternal and cord blood. *European Journal of Clinical Nutrition*, 65(4):480–485, 2011.

[7] Joseph Boachie, Antonysunil Adaikalakoteswari, Jinous Samavat, and Ponnusamy Saravanan. Low vitamin b12 and lipid metabolism: evidence from pre-clinical and clinical studies. *Nutrients*, 12(8):2220, 2020.

[8] Dayanne M Castro, Nicholas R de Veaux, Emily R Miraldi, and Richard Bonneau. Multi-study inference of regulatory networks for more accurate models of gene regulation. *PLoS Computational Biology*, 15(1):e1006591, 2019.

[9] Muller Fabbri, Ramiro Garzon, Amelia Cimmino, Zhongfa Liu, Nicola Zanesi, Elisa Callegari, Shujun Liu, Hansjuerg Alder, Stefan Costinean, Cecilia Fernandez-Cymering, Stefano Volinia, Gulnur Guler, Carl D Morrison, Kenneth K Chan, Guido Marcucci, George A Calin, Kay Huebner, and Carlo M Croce. Microrna-29 family reverts aberrant methylation in lung cancer by targeting dna methyltransferases 3a and 3b. *Proceedings of the National Academy of Sciences*, 104(40):15805–15810, 2007.

[10] Michael Fenech. Folate (vitamin b9) and vitamin b12 and their function in the maintenance of nuclear and mitochondrial genome integrity. *Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis*, 733(1-2):21–33, 2012.

[11] Liangfeng Han, P Dane Witmer, Emily Casey, David Valle, and Saraswati Sukumar. Dna methylation regulates microrna expression. *Cancer Biology & Therapy*, 6(8):1284–1288, 2007.

[12] Marta Jeruszka-Bielak, Carly Isman, Theresa H Schroder, Wangyang Li, Tim J Green, and Yvonne Lamers. South asian ethnicity is related to the highest risk of vitamin b12 deficiency in pregnant canadian women. *Nutrients*, 9(4):317, 2017.

[13] Satyajeet P Khare, Ayush Madhok, Indumathi Patta, Krishna K Sukla, Vipul V Wagh, Pooja S Kunte, Deepa Raut, Dattatray Bhat, Kalyanaraman Kumaran, Caroline Fall, Utpal Tatu, Giriraj R Chandak, Chittaranjan S Yajnik, and Sanjeev Galande. Differential expression of genes influencing mitotic processes in cord blood mononuclear cells after a pre-conceptional micronutrient-based randomised controlled trial: Pune rural intervention in young adolescents (priya). *Journal of Developmental Origins of Health and Disease*, 14(5):530–539, 2023.

[14] Aatish Mahajan, Divika Sapehia, Beenish Rahat, and Jyotdeep Kaur. Altered dietary ratio of folic acid and vitamin b12 during pregnancy influences the expression of imprinted h19/igf2 locus in c57bl/6 mice. *British Journal of Nutrition*, 128(8):1600–1609, 2022.

[15] Aatish Mahajan, Divika Sapehia, Shilpa Thakur, Palani Selvam Mohanraj, Rashmi Bagga, and Jyotdeep Kaur. Effect of imbalance in folate and vitamin b12 in maternal/parental diet on global methylation and regulatory mirnas. *Scientific Reports*, 9(1):17602, 2019.

[16] Pooja R Mandaviya, Roby Joehanes, Jennifer Brody, Juan E Castillo-Fernandez, Koen F Dekkers, Anh N Do, Mariaelisa Graff, Ismo K Hänninen, Toshiko Tanaka, Ester AL de Jonge, et al. Association of dietary folate and vitamin b-12 intake with genome-wide dna methylation in blood: a large-scale epigenome-wide association analysis in 5841 individuals. *The American Journal of Clinical Nutrition*, 110(3):437–450, 2019.

[17] Jill A McKay, Alexandra Groom, Catherine Potter, Lisa J Coneyworth, Dianne Ford, John C Mathers, and Caroline L Relton. Genetic and non-genetic influences during pregnancy on infant global and site-specific dna methylation: role for folate gene variants and vitamin b12. *PLoS One*, 7(3):e33290, 2012.

[18] Giulietta S Monasso, Thanh T Hoang, Giulia Mancano, et al. A meta-analysis of epigenome-wide association studies on pregnancy vitamin b12 concentrations and offspring dna methylation. *Epigenetics*, 18(1):2207028, 2023.

[19] Pauline Mosca, Aurélie Robert, Jean-Marc Alberto, Marie Meyer, Urbi Kundu, Sébastien Hergalant, Rémy Umoret, David Coelho, Jean-Louis Guéant, Bruno Leheup, and Natacha Dreumont. Vitamin b12 deficiency dysregulates m6a mrna methylation of genes involved in neurological functions. *Molecular Nutrition & Food Research*, 65(22):2100453, 2021.

[20] Jinous Samavat, Antonysunil Adaikalakoteswari, Joseph Boachie, Philip McTernan, Mark Christian, and Ponnusamy Saravanan. Vitamin b12 deficiency leads to fatty acid metabolism dysregulation and increased pro-inflammatory cytokine production in human adipocytes and maternal adipose tissue. *Journal of the Endocrine Society*, 4(Suppl_1):A611–A612, 2020.

[21] Nithya Sukumar, Snorri B Rafnsson, Ngianga-Bakwin Kandala, Raj Bhopal, Chittaranjan S Yajnik, and Ponnusamy Saravanan. Prevalence of vitamin b-12 insufficiency during pregnancy and its effect on offspring birth weight: a systematic review and meta-analysis. *The American Journal of Clinical Nutrition*, 103(5):1232–1251, 2016.

[22] Marjolein M van Vliet, Sam Schoenmakers, Joost Gribnau, and Régine PM Steegers-Theunissen. The one-carbon metabolism as an underlying pathway for placental dna methylation: a systematic review. *Epigenetics*, 19(1):2256036, 2024.

[23] Rency S Varghese, Megan E Barefoot, Sidharth Jain, Yifan Chen, Yunxi Zhang, Amber Alley, Alexander H Kroemer, Mahlet G Tadesse, Deepak Kumar, Zaki A Sherif, and Habtom W Ressom. Integrative analysis of dna methylation and microrna expression reveals mechanisms of racial heterogeneity in hepatocellular carcinoma. *Frontiers in Genetics*, 12:691460, 2021.

Appendix A

Appendix

This appendix presents supplementary results that provide additional evidence for the findings in the main text. It includes extended analyses of detailed enrichment plots for key gene sets, CpG–mRNA regulatory relationships, and cross-model comparisons of top-ranked features.

A.1 Functional Enrichment of Key Gene Sets

We performed GO and KEGG enrichment analysis for three key gene sets discussed in Section 4.1: (1) CpG-regulated genes, (2) differentially expressed genes (DEGs), and (3) double-regulated genes (subject to both miRNA and CpG regulation).

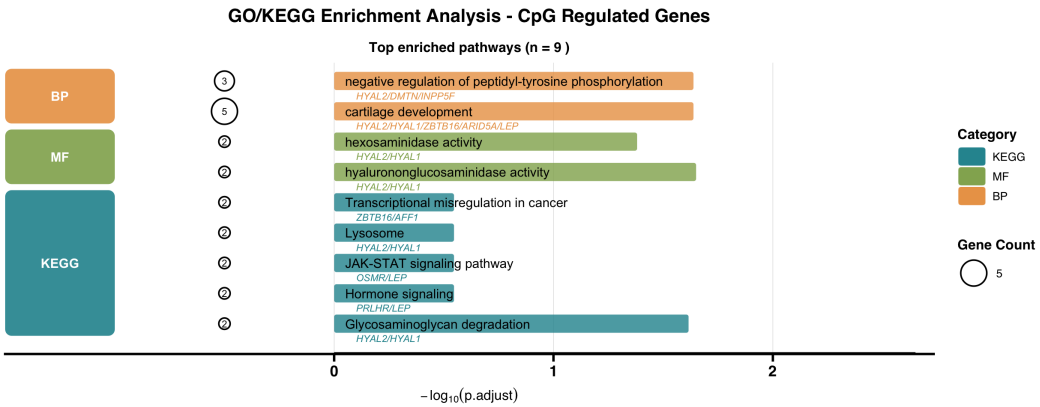

Figure A.1: Enriched pathways among CpG-regulated genes ($n = 38$), highlighting immune and developmental processes.

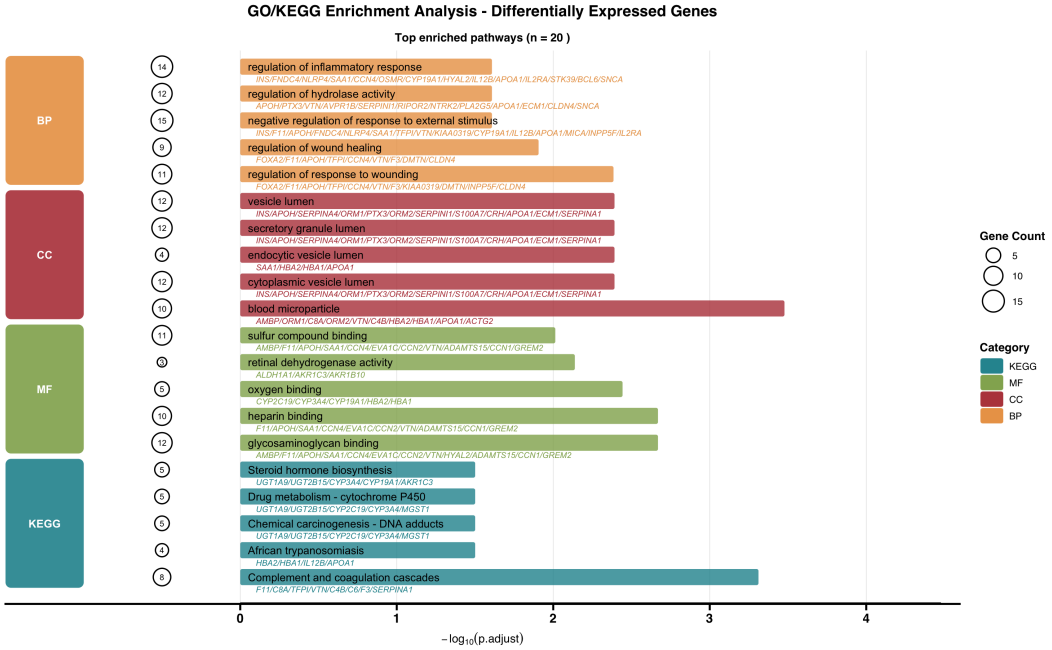

Figure A.2: GO/KEGG enrichment results for DEGs ($n = 208$), including pathways involved in immune response and hormone signaling.

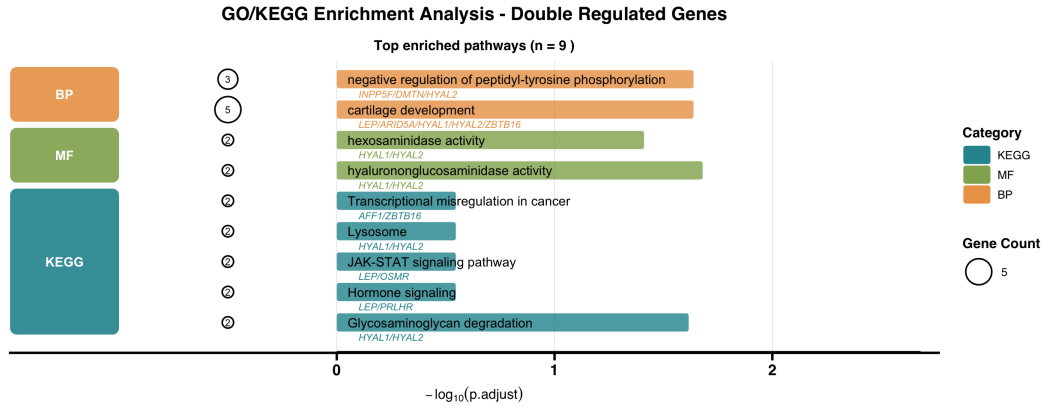

Figure A.3: Enrichment of double-regulated genes ($n = 9$), with significant hits in cartilage development and lysosomal function.

A.2 CpG–mRNA Regulatory Effect Summary

To better understand the strength and consistency of epigenetic regulation, we evaluated the distribution of effect sizes (β coefficients) and model fit (R^2) across all CpG–mRNA pairs. As shown in Figure A.4, the majority of significant regulatory pairs exhibited large effect sizes ($|\beta| > 0.5$), with over 40% of them showing $R^2 > 0.4$, indicating strong explanatory power.

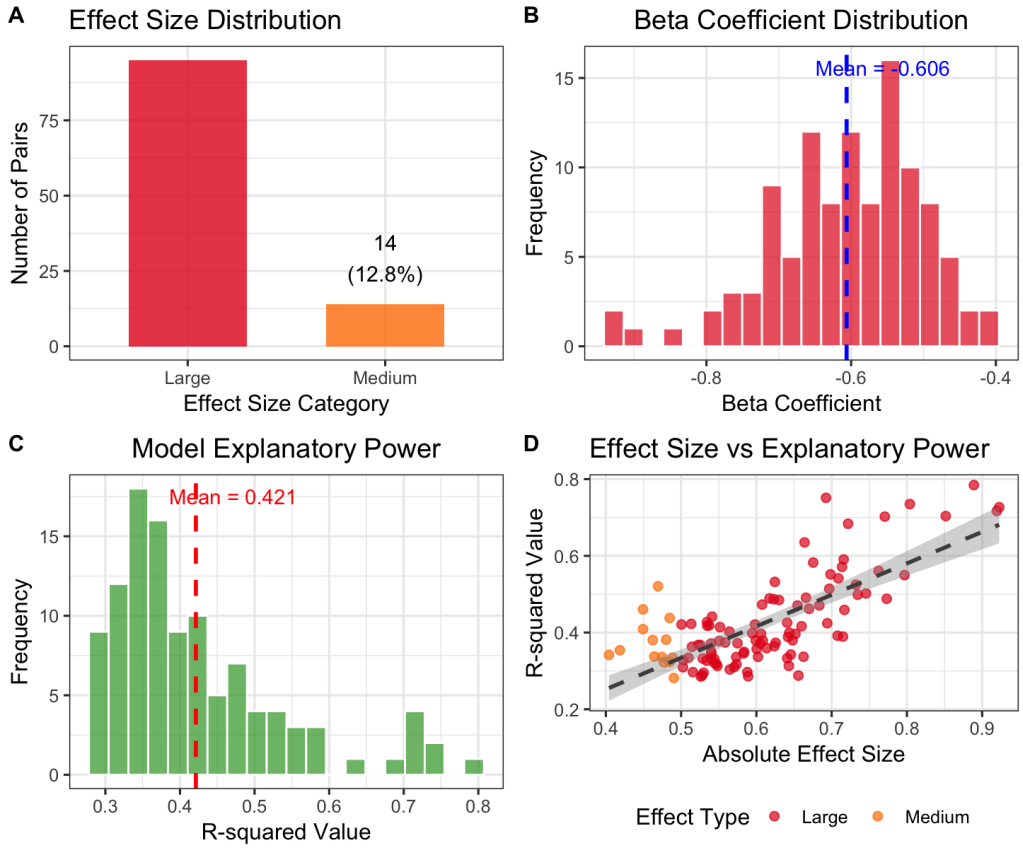

Figure A.4: Distribution of effect sizes and model R^2 in CpG–mRNA regulatory pairs.

A.3 Consensus Feature Interpretation Across Models

To assess the consistency and biological relevance of model-selected features, we compared the top 10 ranked predictors across six modeling configurations (logistic regression, random forest, and neural network; each under data-driven and mechanism-driven strategies).

Table A.1: Top consensus features across all models and strategies

Feature	Models	Mean Rank	Type	Biological Role
(A) Overall Top 10 Features (All 6 Models)				
BMI	6/6	2.2±1.3	Clinical	Nutritional status
Age	5/6	2.6±1.0	Clinical	Physiological aging
GTSF1	3/6	4.7±1.2	Methylation	Epigenetic regulation
IQCG	3/6	5.0±1.0	Methylation	GTPase signaling
NUDT10	3/6	5.7±1.2	Methylation	Nucleotide metabolism
SYCP3	3/6	8.0±1.0	Methylation	Synaptonemal complex
CCNG1	3/6	5.3±1.2	Protein-coding	Cell cycle regulation
LOC102725048	3/6	5.6±0.6	lncRNA	Unknown function
TSTD1	2/6	7.0±0.0	Methylation	Sulfur metabolism
ZNF154	2/6	9.0±0.0	Methylation	Zinc finger protein
(B) Mechanism-Driven Models Only (LR-Mech, RF-Mech, NN-Mech)				
BMI	3/3	1.3±0.5	Clinical	Nutritional status
GTSF1	3/3	4.0±1.0	Methylation	Epigenetic regulation
IQCG	3/3	5.3±1.0	Methylation	GTPase signaling
NUDT10	3/3	6.3±1.5	Methylation	Nucleotide metabolism
SYCP3	3/3	9.0±0.0	Methylation	Synaptonemal complex
TSTD1	2/3	7.5±0.5	Methylation	Sulfur metabolism
ZNF154	2/3	9.5±0.5	Methylation	Zinc finger protein
PSMA8	2/3	5.0±1.0	Methylation	Proteasome subunit
LINC00667	2/3	9.5±0.5	lncRNA	Unknown function
NAA11	1/3	10.0	Methylation	N-acetyltransferase
(C) Data-Driven Models Only (LR-Data, RF-Data, NN-Data)				
BMI	3/3	2.3±0.6	Clinical	Nutritional status
Age	3/3	2.7±0.6	Clinical	Physiological aging
CCNG1	3/3	3.7±1.5	Protein-coding	Cell cycle regulation
LOC102725048	3/3	4.3±1.0	lncRNA	Unknown function
Parity	2/3	3.5±0.5	Clinical	Reproductive history
KCMF1	2/3	5.5±0.5	Methylation	Neural regulation
RHBDF2	1/3	9.5±0.5	Methylation	Protease regulation
SLF2	1/3	10.0	Methylation	Genome stability
SLC15A3	1/3	9.0±0.0	Methylation	Solute transport
ZNF747-DT	1/3	6.0±0.0	lncRNA	Unknown function