# Multiomics Analysis of Human Pregnancy: Integrating Immunome, Transcriptome, Microbiome, Proteome, and Metabolome Data

OURAHOU MOHAMED

**Abstract**

Pregnancy is a complex physiological process regulated by multiple interconnected biological systems. Throughout a full-term pregnancy, a dynamic interplay of immunological, metabolic, proteomic, genomic, and microbiomic adaptations occurs, collectively orchestrating a healthy gestation. Understanding the precise timing and coordination of these adaptations is critical for investigating deviations that contribute to pregnancy-related complications, such as preterm birth and preeclampsia.

In this project, we conducted a comprehensive multiomics analysis involving 68 samples obtained from 17 pregnant women who successfully delivered at full term. These samples were simultaneously assessed across various biological domains, including the immunome, transcriptome, microbiome, proteome, and metabolome. To evaluate the predictive capacity of each dataset for gestational age, we employed six distinct predictive models (Elastic Net, XGBoostRegressor, GradientBosstingRegressor, Support Vector Regressor, HuberRegressor, and AdaBoostRegressor). Furthermore, we employed a stacked generalization approach to integrate these diverse datasets into a unified model. Notably, the amalgamation of these datasets not only significantly improved the accuracy of gestational age prediction but also unveiled intriguing novel interactions between different biological modalities.

Our findings pave the way for future investigations, including the expansion of our cohort to encompass populations with a higher risk of preterm birth. Additionally, our study sets the stage for in vivo experimentation aimed at modulating the immune response based on the identified mechanistic insights. Ultimately, this research contributes to a deeper understanding of the intricate biological processes underlying pregnancy and offers potential avenues for the prevention and management of pregnancy-related complications.

# Contents



# 1 Introduction

Pregnancy is a complex biological process marked by profound physiological changes in both the mother and the developing fetus. These intricate changes are orchestrated by a network of molecular and cellular systems and are critical determinants of the outcome of

pregnancy. However, pregnancy is not without its challenges, and certain complications can have serious implications for both maternal and fetal health.

Within the domain of maternal-fetal medicine, the precise estimation of fetal age and sex is a challenge of paramount importance.

The primary aim of this project it to examine various approaches for incorporating transcriptomic, immunological, microbiomic, metabolomic, and proteomic datasets into diverse statistical models aimed at predicting gestational age in full-term pregnancies. Additionally, we aimed to pinpoint the most precise method for achieving this prediction.

# 2 Methodes

## 2.1 Dataset

### 2.1.1 Exploration

The dataset under examination in this study encompasses a comprehensive representation of biological processes during pregnancy and postpartum in 17 women. It comprises a rich array of data types, each carefully selected to capture pivotal moments in the maternal-fetal journey. The dataset includes peripheral blood samples for CyTOF analysis, plasma samples for proteomic, cell-free transcriptomics, and metabolomics analyses, serum samples for luminex analyses, and a series of culture swabs for microbiome analysis. These data points were collected at distinct stages of pregnancy, spanning the first trimester (7–14 weeks), second trimester (15–20 weeks), and third trimester (24–32 weeks), as well as a postpartum sample taken at 6 weeks after delivery (Figure 1).

By focusing exclusively on these trimesters and excluding the postpartum period, our dataset provides a unique opportunity to investigate the continuous biological adaptations occurring from early fetal development to the culmination of gestation. This strategic data curation ensures that our analyses are centered around the crucial phases of pregnancy, offering valuable insights into the intricate processes that shape the maternal-fetal relationship.

### 2.1.2 Preprocessing

After conducting an assessment of feature independence, it became evident that there were existing dependencies between the features within the datasets. Recognizing the importance of addressing these dependencies to ensure the robustness of our analysis, we undertook a systematic approach. We embarked on testing various correlation thresholds with each dataset, coupled with employing different modeling techniques. This comprehensive exploration allowed us to identify and select the most suitable correlation threshold values for our specific dataset and research objectives.
After this step, we standardized the data because we observed that the values of individual features were skewed. This standardization, which involved scaling the features to have a mean of 0 and a standard deviation of 1, ensured that our analysis would not be unduly influenced by variations in feature scales.

## 2.2 Features Selection

Optimizing the feature set was a key focus for us, acknowledging the importance of enhancing the efficiency and effectiveness of our machine-learning models. This crucial step involved evaluating our datasets to identify redundant or irrelevant information.

In this process, we employed both Pearson and Spearman correlation coefficients to track both linear and monotone relationships

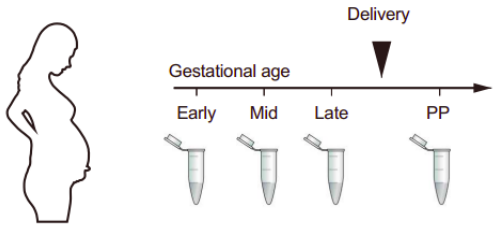



Figure 1: Overview of the study design.

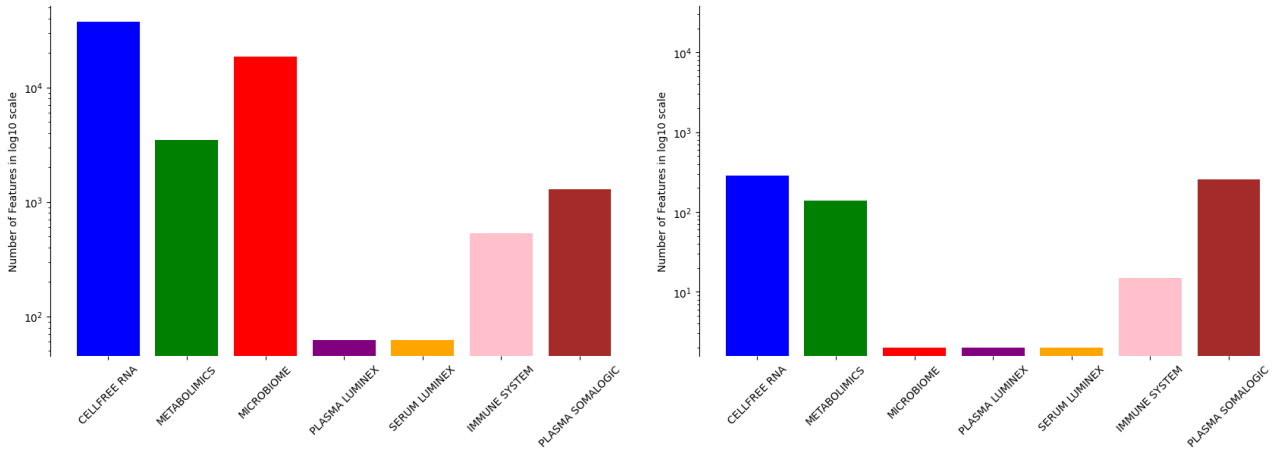



Figure 2: Number of Features in Different Omics dataset before and after Feature reduction

between our features and the target variable gestation age. Pearson correlation helped us determine the strength and significance of linear relationships, while Spearman correlation enabled us to capture any monotone relationships that might exist.

This systematic approach allowed us to identify and quantify the relationships between the features and the target variable, guiding us in the removal of redundant variables and simplifying our model. By conducting this feature selection using both correlation methods, we aimed to improve model interpretability, mitigate the risk of overfitting, and ultimately contribute to more robust and efficient machine learning outcomes.

## 2.3 Stack Generalization

In computer science, "stacked generalization" means combining weaker predictors to make better predictions. It's like asking multiple people for their opinions and then using all those opinions to make a final decision. [1] [2] [3]

We took this concept and applied it to the analysis of multiple types of biological data from a group of patients. For each dataset, we rigorously evaluated six different models (Elastic Net[4], XGBoostRegressor[5], GradientBosstingRegressor[6], Support Vector Regressor[7], HuberRegressor[8], and AdaBoostRegressor[9]). We then selected the best-performing model for predicting gestational ages. This winning model was subsequently employed in stack generalization, allowing us to harness the collective insights from diverse sources and produce a more comprehensive prediction.

## 2.4 Cross Validation

The ML algorithms rely on the assumption of statistical independence among all observations. However, in our analysis, there is a departure from this assumption as the samples collected from different trimesters of the same

subject exhibit interdependence. To address this issue, we devised a leave-one-subject-out cross-validation approach (Figure 3).

In this approach, a model is initially trained using all available samples except those from the three trimesters of a specific subject. Subsequently, the model is tested on the samples from the subject it had not been exposed to during training. This process is iteratively repeated for all subjects until we generate predictions for all samples. Our final results are then based on these blinded predictions, ensuring our analysis remains entirely independent of any intra-subject correlations.

# 3 Results

## 3.1 Features Selection

We employed both Spearman and Pearson correlation methods to identify relevant features within our dataset. The selection criteria were based on the significance level of the features, with a p-value threshold set at 0.01, a common practice often seen in medical literature.

For each of the seven datasets, we conducted feature selection separately using both Spearman and Pearson correlations, applying a significance threshold of 0.01 for the p-value, except for Dataset SERUM LUMINEX, where we adjusted the p-value threshold to 0.05 to ensure the inclusion of meaningful variables (no features met the 0.01 p-value threshold) (Figure 2).

## 3.2 Per-dataset analysis

Our study focused on the development of 6 predictive models (Elastic Net, XGBoostRegressor, GradientBosstingRegressor, Support Vector Regressor, HuberRegressor, and AdaBoostRegressor) to estimate the gestational age of subjects during pregnancy at various stages. To optimize the performance of the models, we employed a random search approach to optimize the models' hyperparameters instead of grid search since it is more exhaustive to find the best values in the searching space. To ensure that gestational age predictions were made on completely independent samples we employed a leave-one-subject-out cross-validation approach.

The effectiveness of our predictive models was visually assessed by comparing their predictions on test samples to the clinical estimates of gestational age, and the statistical significance of these predictions was quantified using p-values. Specifically, we evaluated the correlations between our model's predictions and the actual gestational age at the time of sampling. The results of the testing procedure can be found in (Figure 4).

Remarkably, our analysis unveiled that no single model consistently outperforms all others. Instead, for each dataset, there exists a model that excels relative to the rest. Consequently, we opt for the top-performing model specific to each dataset for utilization in the stacked generalization process.

## 3.3 Stacked generalization

Within the figures, it is evident that this approach proves beneficial in refining results, particularly in addressing high p-values across individual model tests in comparison to tests conducted on a single dataset, while also yielding low MSE values. AdaboostRegressor demonstrated superior performance relative to other models, with HubberRegressor trailing closely behind.

The METABOLIMICS dataset displayed the highest predictive capability (as depicted in Figure 2B and Figure 5), followed by the MICROBIOME and CELLFREE RNA datasets, respectively. Additionally, the figure illustrates that datasets containing fewer features exhibit greater predictive power than

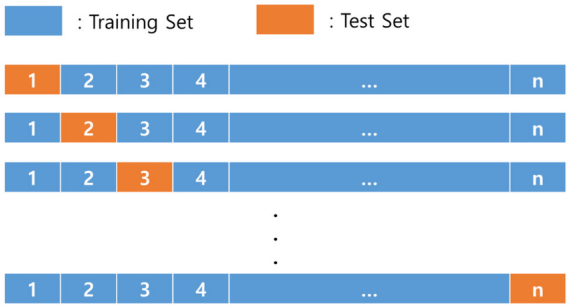
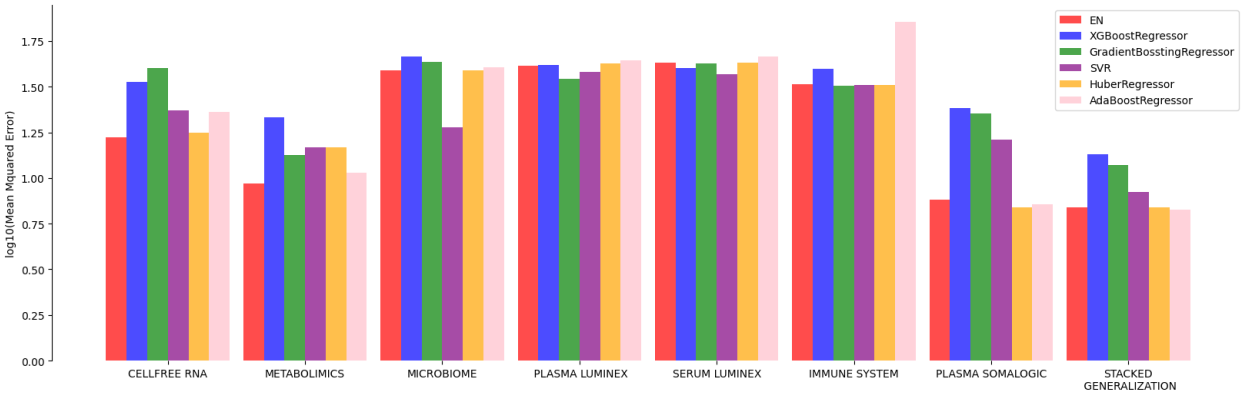
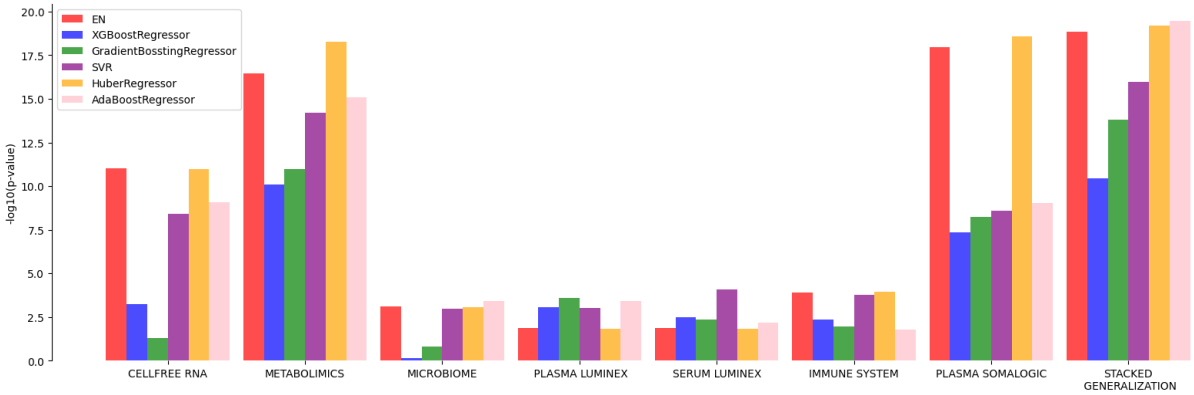



Figure 3: Subject-Level Leave-One-Out Cross-Validation: In this cross-validation approach, we leave out one entire subject (comprising its three samples) for testing in each iteration, while using the remaining subjects for training.




(A) Mean Squared Error Loss




(B) P-values

Figure 4: Empirical evaluation of the sex models on each dataset, and the combination of all of them (Stacked generalization). (A) and (B) represent the Mean Squared Error Loss and P-values between predicted and clinical gestational ages respectively. The hyperparameters of each model were tuned by hundreds of random search experiences.

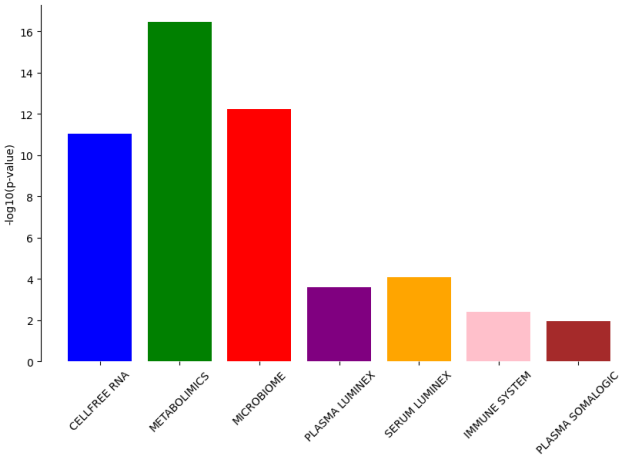



Figure 5: The contribution power of each dataset on predicting gestational age.

those with a higher number of features. This underscores the significance of each feature within each dataset in contributing to the prediction of gestational age.

# 4 Discussion

## 4.1 Key Findings and Implications

Our analysis sought to address the challenge of predicting gestational age accurately using a multiomic approach. We combined various types of biological data, including transcriptomic, immunological, microbiomic, metabolomic, and proteomic datasets, to build predictive models. Our results indicate that this integrative approach has the potential to enhance the precision of gestational age prediction.

One noteworthy observation is that no single predictive model consistently outperformed all others across different datasets. Instead, the choice of the best-performing model varied depending on the specific dataset being analyzed. This finding underscores the importance of tailoring the choice of machine learning model to the characteristics of the data at hand. It also highlights the complexity of the biological processes underlying gestational age, which may require different modeling strategies for optimal prediction.

Furthermore, our feature selection process, which involved both Pearson and Spearman correlation coefficients, allowed us to identify relevant features while mitigating the risk of overfitting. This step improved the interpretability of our models and contributed to their robustness.

## 4.2 The Value of Stacked Generalization

Stacked generalization proved to be a valuable technique for improving the accuracy of gestational age prediction. By combining the insights from multiple predictive models, we were able to generate more comprehensive and reliable predictions.

The choice of AdaboostRegressor as the top-performing model in the stacked generalization process was intriguing. It demonstrated superior performance relative to other models, with HubberRegressor as a close contender. This suggests that ensemble methods, such as Adaboost, may be particularly effective in capturing the complex relationships between multiomic data and gestational age.

Our analysis also revealed that certain

datasets, such as METABOLIMICS, exhibited higher predictive capability than others. This underscores the importance of considering the quality and informativeness of individual omics datasets when building predictive models.

Overall, our study highlights the potential of multiomic analysis in improving gestational age prediction. It also emphasizes the need for a tailored approach, considering both the choice of predictive model and the characteristics of the data. Stacked generalization emerges as a promising technique for harnessing the collective insights from diverse biological datasets. Future research in this area could further refine these models and explore additional omics data sources to enhance the accuracy of gestational age estimation.

# 5 Conclusions

In conclusion, this study delved into the challenging task of predicting gestational age in full-term pregnancies through a multiomic approach. Our findings reveal the complexity of this endeavor, where no single predictive model consistently outperformed others across diverse datasets. Flexibility in model selection, tailored to the data's unique characteristics, emerged as a crucial consideration.

The rigorous feature selection process enhanced the interpretability and robustness of our models, yet it remains an ongoing challenge in multiomic analysis.

However, our project had limitations. The longitudinal nature of the data, encompassing samples from the same subjects, introduced complexities requiring further exploration. The relatively small number of subjects in our 'proof-of-concept' cohort limited the generalizability of our findings, and recruitment from a single-care center constrained dataset diversity.

In summary, our work highlights the potential of multiomic analysis in gestational age prediction. Addressing these limitations and expanding our dataset could refine our models, potentially benefiting prenatal care and maternal-fetal health outcomes.

# References

[1] Breiman,L. (1996) *Stacked regressions.* Mach. Learn., 24, 49–64

[2] Sharkey,A.J.C. (1996) *On combining artificial neural nets.* Connect. Sci., 8, 299–314

[3] Wolpert,D.H. (1992) *Stacked generalization. Neural Netw.*, 5, 241–259.

[4] Zou, H., and Hastie, T. (2005). *Regularization and variable selection via the elastic net.* Journal of the Royal Statistical Society: Series B (Statistical Methodology), 67(2), 301-320.

[5] Chen, T., and Guestrin, C. (2016). *XGBoost: A Scalable Tree Boosting System.* In Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining (pp. 785-794).

[6] Friedman, J. H. (2001). *Greedy Function Approximation: A Gradient Boosting Machine.* Annals of Statistics, 29(5), 1189-1232.

[7] Vapnik, V. (1997). *Support-Vector Networks.* Machine Learning, 20(3), 273-297.

[8] Huber, P. J. (1964). *Robust Estimation of a Location Parameter.* Annals of Mathematical Statistics, 35(1), 73-101.

[9] Freund, Y., and Schapire, R. E. (1996). *Experiments with a New Boosting Algorithm.* In Proceedings of the Thirteenth International Conference on Machine Learning (ICML) (pp. 148-156).