


RESEARCH

Open Access



# Precision integrated identification of predictive first-trimester metabolomics signatures for early detection of gestational diabetes mellitus

Sapna Sharma<sup>1†</sup>, Yalamanchili Venkata Subrahmanyam<sup>2,3†</sup>, Payal Gupta<sup>4</sup>, Sangeetha Vadivel<sup>6</sup>, Mohan Deepa<sup>5</sup>, Ansh Tandon<sup>4</sup>, Sreekumar Sreedevi<sup>17</sup>, Uma Ram<sup>7</sup>, Priyanka Narad<sup>8</sup>, Dharmeshkumar Parmar<sup>2,3</sup>, Ranjit Mohan Anjana<sup>6</sup>, Anu Raghunathan<sup>2,3</sup>, Muthuswamy Balasubramanyam<sup>9</sup>, Viswanathan Mohan<sup>6</sup>, Abhishek Sengupta<sup>4</sup>, Jerzy Adamski<sup>10,11,12</sup> , Ponnusamy Saravanan<sup>13,14,15</sup>, Venkateswarlu Panchagnula<sup>2,3</sup>, Dandamudi Usharani<sup>16\*</sup> and Kuppan Gokulakrishnan<sup>17\*</sup>

## Abstract

**Background and aim** Gestational diabetes mellitus (GDM), a common pregnancy-related metabolic disorder, often goes undiagnosed until the second trimester, limiting early intervention opportunities. Given the higher prevalence of GDM in India, there is a critical need to investigate metabolomic biomarkers among Asian Indians, who exhibit greater insulin resistance and are predisposed to developing type 2 diabetes at an earlier age. This study aimed to identify early pregnancy metabolomic signatures predictive of GDM.

**Methods** Among 2115 pregnant women from the STRatification of Risk of Diabetes in Early pregnancy (STRiDE) study, we performed untargeted metabolomic profiling using UPLC-MS/MS at early pregnancy (< 16 weeks) plasma samples from 100 women—comprising 50 with GDM and 50 normal (without GDM) based on oral glucose tolerance test (OGTT) at 24–28 weeks. Statistical and machine learning approaches, including logistic regression and random forest (RF), were applied to identify GDM-associated metabolites and construct predictive models. Pathway enrichment analysis was conducted using KEGG database annotations.

**Results** A total of 49 metabolites were significantly associated with GDM, primarily involving lipid classes such as phosphatidylcholines, sphingomyelins, and triacylglycerols. RF analysis identified a panel of eight metabolites that achieved best predictive performance (AUC 0.880; 95% CI: 0.809–0.951) for GDM. When combined with conventional

<sup>†</sup>Sapna Sharma and Yalamanchili Venkata Subrahmanyam: Joint first authors.

\*Correspondence:  
Dandamudi Usharani  
ushadandamudi.rani@gmail.com  
Kuppan Gokulakrishnan  
gokul@nimhans.ac.in; gokulnimhans@gmail.com

Full list of author information is available at the end of the article



© The Author(s) 2025. **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

clinical risk factors, the integrated model showed comparable prediction of GDM with AUC 0.88; 95% CI: 0.810–0.952). Enrichment analysis highlighted dysregulated pathways including glycerophospholipid and sphingolipid metabolism, autophagy, and insulin resistance.

**Conclusion** This study demonstrates the utility of early-pregnancy metabolomic profiling for predicting GDM in Indian women. The eight-metabolite panel offers a promising tool for early risk stratification of GDM, warranting validation in diverse populations.

**Keywords** Gestational diabetes mellitus, Metabolomics, Prediction, First trimester, Mass spectrometry, Indian women

## Research summary

What is currently known about this topic?

- Gestational diabetes mellitus (GDM) remains undiagnosed until the late second trimester, when glucose tolerance tests are routinely administered.
- This late diagnosis limits opportunities for early intervention, emphasizing the need for novel, non-invasive biomarkers to predict GDM earlier in pregnancy.
- Timely identification and management of GDM risk factors are essential to reducing adverse maternal and fetal outcomes.

What is the key research question?

- Are there any metabolomic signatures in early pregnancy (< 16 weeks) that can serve as a predictive marker for identifying GDM.

What is new?

- This study investigates the predictive performance of metabolomic biomarkers in an Indian cohort.
- We identified a panel of eight metabolites during early pregnancy that could predict GDM with high accuracy, underscoring the potential clinical utility of metabolomic signatures in early risk stratification.

How might this study influence clinical practice?

- Integrating the metabolomic biomarker assessment and conventional clinical risk factors could enhance personalized GDM prevention and management at an early stage.

## Introduction

Gestational diabetes mellitus (GDM), defined as glucose intolerance with onset or first recognition during pregnancy, remains one of the most prevalent pregnancy-related metabolic disorders worldwide [1]. The global prevalence of GDM ranges between 5% and 25%, depending on diagnostic criteria and population demographics, and it continues to rise in parallel with increasing rates of

obesity and sedentary lifestyles [2–4]. GDM poses considerable risks to both mother and fetus [5], including hypertensive disorders, cesarean delivery, macrosomia [6], neonatal hypoglycemia, and a higher predisposition for type 2 diabetes mellitus (T2DM) [7], and cardiovascular diseases [8–11] later in life for both the mother and offspring. Unfortunately, GDM often remains undiagnosed until the late second trimester, when glucose tolerance tests are routinely administered [12]. This delay limits opportunities for early intervention, underscoring the need for novel, non-invasive biomarkers capable of predicting GDM at earlier stages of pregnancy [13].

Recent studies have highlighted the utility of Mass Spectrometry (MS)-based metabolomics in uncovering potential predictive metabolite biomarkers for GDM [14–16]. A study by Zhao et al. used Liquid Chromatography–Mass Spectrometry (LC-MS) to profile maternal serum in the first trimester and identified a panel of metabolites, including branched-chain amino acids and acylcarnitines, that were significantly associated with subsequent GDM development [17]. Similarly, Scholtens et al. from the Hyperglycemia and Adverse Pregnancy Outcome (HAPO) study explored metabolomic signatures and demonstrated distinct metabolic profiles linked to insulin resistance and glucose dysregulation as early as 10–14 weeks of gestation [18]. Another study by Mokkala et al. reported distinctive metabolic signatures in early pregnancy associated with later GDM onset [19]. Additionally, emerging studies have pointed to disruptions in bile acid metabolism [20–22], aromatic amino acids [23], and microbiome-derived metabolites [24] as key components in the metabolic landscape of GDM. Interestingly, our earlier study has shown that lower telomere length and mitochondrial DNA copy number in early pregnancy were significantly linked to a higher risk of developing GDM later in pregnancy [25]. These findings support the notion that metabolic perturbations precede clinical GDM diagnosis and can be leveraged for early risk prediction.

Lipidomics studies have also shown promise in the context of GDM [26, 27]. Among the most promising metabolite classes, small molecular lipids such as sphingolipids, glycerophospholipids, and fatty acids were shown to play a crucial role in metabolic regulation and

disease pathogenesis [28–32]. These molecules hold significant potential as early biomarkers for predicting the onset of GDM, opening new avenues for timely diagnosis and improved maternal and fetal outcomes. A study by Liu et al. identified alterations in phospholipids and sphingolipids in early pregnancy serum samples from women who later developed GDM, indicating potential roles for altered lipid metabolism in GDM pathogenesis [29]. Despite these advancements, there remains a lack of consensus on standardized metabolomic biomarkers for GDM prediction, partly due to heterogeneity in study design, population differences, and analytical platforms. Studies predominantly from European and East Asian cohorts have reported early lipid metabolic alterations associated with GDM, while the data from South Asian populations remain limited despite their high GDM burden. Our study advances this field by aiming to identify a multimetabolite panel in an Indian cohort that not only reflects population-specific metabolic features but also aligns with global metabolomic pathway disturbances linked to GDM. Such cross-ethnic consistency underscores the potential translational value of metabolite panels that capture shared metabolic mechanisms while accounting for population variability.

We therefore, aim to identify early pregnancy (< 16 weeks) metabolomic signatures for GDM in a longitudinal Indian pregnancy cohort from the STRatification of Risk of Diabetes in Early Pregnancy (STRiDE) study [33] using comprehensive mass spectrometry profiling. By leveraging the precision of modern MS techniques, applying robust bioinformatic analysis and integrating conventional maternal risk factors, we strive to advance early prediction and risk stratification, which is expected to result in personalized prenatal care and better maternal-fetal outcomes.

## Method

### Study participants

The study population comprised 2115 pregnant Indian women enrolled in the STRiDE study, a prospective, longitudinal cohort that recruited participants at their first antenatal visit, prior to 16 weeks of gestation. The study was conducted between 2016 and 2019 across seven clinical sites located in three cities in South India. Detailed descriptions of the STRiDE cohort have been published previously [33]. For the present nested cohort study within the STRiDE cohort, 100 pregnant women with fasting glucose levels < 5.1 mmol/L (92 mg/dl) in early pregnancy were randomly selected with similar age and pre-pregnancy body mass index (BMI) to minimize the potential confounding effects of these variables. At 24–28 weeks of gestation (OGTT visit), all participants underwent a 75 g oral glucose tolerance test (OGTT) and were subsequently classified as having either normal glucose

tolerance (NGT,  $n = 50$ ) or gestational diabetes mellitus (GDM,  $n = 50$ ) according to the International Association of Diabetes and Pregnancy Study Groups (IADPSG) criteria [34]. Metabolomic profiling was performed for all 100 participants at early pregnancy, and the overall study design is illustrated in Supplementary Fig. 1. We implemented a nested case–control sampling strategy within the STRiDE cohort to facilitate the matched analysis and control for key confounders. Fifty women who developed GDM at 24–28 weeks of gestation were identified, and for each case, one normoglycemic control was selected from the same cohort, matched on age and pre-pregnancy BMI (group matched) to yield a 1:1 matched set. This procedure led to equal number of cases and controls by design.

A written informed consent was obtained from all participants. The study protocol was approved by the Institutional Ethics Committees of the National Institute of Mental Health and Neuro Sciences (NIMHANS), Bengaluru, and the Madras Diabetes Research Foundation (MDRF), Chennai, India.

### Polar sample preparation

The samples were extracted based on an extraction method reported by Evans et al. [35]. This method was uniformly used for all the samples and quality controls (QCs). The QC samples were prepared by combining equal volume from each individual sample within each category. Briefly, the plasma samples were thawed on ice prior to further processing, and the labelled internal standards ( $^{13}\text{C}_6$  - Glucose and  $\text{D}_5$  - Glutamic acid) were added to 50  $\mu\text{L}$  of the plasma samples and vortexed for 5 minutes. Subsequently, the sample deproteinization was performed by adding 450  $\mu\text{L}$  of ice-cold methanol. The mixture was incubated for 1 h at  $-80^\circ\text{C}$ . The samples were centrifuged at 13,500 rpm for 15 minutes at  $8^\circ\text{C}$ . The supernatant was collected and then transferred into a new sample tube.

### Non-polar sample preparation

The samples were extracted based on an extraction method reported by Cajka et al. [36]. This method was uniformly used for all the samples and QCs. The QC samples were prepared by combining equal volume from each individual sample within each category. The sample deproteinization was performed by adding 450  $\mu\text{L}$  of ice-cold methanol containing Sphingomyelin (d18:0/17:0), Sphingosine (d17:1) and Ceramide (d18:0/17:0) standards. 1 mL of MTBE was added. Phase separation was induced by adding 250  $\mu\text{L}$  of LC-MS grade water, followed by centrifugation at 14,000 rpm for 5 min at  $4^\circ\text{C}$ . The upper organic phase was collected and dried under vacuum at  $10^\circ\text{C}$ . Finally, the dried samples were reconstituted with ACN and IPA (2:1 v/v). 50  $\mu\text{L}$  of each

sample was added to the autosampler vial prior to sample analysis.

#### UPLC-MS/MS analysis

Polar and non-polar analysis was performed on Ultra Performance Liquid Chromatography (UPLC) (Water® ACQUITY I-Class), equipped with a binary solvent manager with online degasser, flow-through needle (FTN) sample manager, and column oven compartment. In polar metabolite analysis, Waters Acquity BEH HILIC column (50 mm\*2.1 mm, 1.7 µm) was used for chromatography separation and column temperature was maintained at 45 °C. 1 µL of sample was injected for analysis. The polar analytes were separated via multistep gradient elution and flow rate. The mobile phase A was composed of 0.1% formic acid in water (v/v) with 10 mM ammonium acetate, and mobile phase B was 0.1% formic acid in acetonitrile: water (95:5) with 10 mM ammonium acetate. The gradient conditions were; at 0 min-0.15 mL, 10% (A); at 0.25 min-0.15 mL, 10% (A); at 0.5 min-0.15 mL, 20% (B); at 4.0 min-0.15 mL, 25% (A); at 4.5 min-0.35 mL, 40% (A); at 7.0 min-0.35 mL, 90% (A); at 7.10 min-0.35 mL, 10% (A); at 8.0 min-0.35 mL, 10% (A); at 8.01 min-0.15 mL, 10% (A); at 10.0 min-0.15 mL, 10% (A).

For non-polar and polar lipid analysis, a Waters Acquity BEH Amide column (50 mm\*2.1 mm, 1.7 µm) was used for chromatography separation, and the column temperature was maintained at 45 °C. 2 µL of the sample was injected for analysis. The lipids were separated via multi-step gradient elution. The mobile phase A was composed of 10 mM ammonium acetate in acetonitrile: water (95:5) and mobile phase B was 10 mM ammonium acetate in acetonitrile: water (95:50). Flow rate was set at 0.6 mL/min. The gradient conditions were at 0 min, 99.9% (A); at 1.0 min, 80% (A); at 2.0 min, 20% (B); at 2.50 min, 99.9% (A); at 4. min, 99.9% (A). For polar, non-polar and polar lipid analysis, the purge solvent was a mixture of 50% acetonitrile in water, while the needle wash was 80% acetonitrile in water.

The polar, nonpolar and polar lipid metabolite analysis was performed on Xevo TQ-XS tandem Mass Spectrometer (Waters Corporation) with an ESI ionization source in both positive and negative mode MRM. The source parameters for each analyte were optimized using standards and pooled samples. The MS source capillary was maintained for positive and negative ion mode at 3.0 kV, 2.5 kV respectively. The desolvation, cone, and collision gas were maintained at 1000 L/h, 150 L/h, and 0.13 mL/min, respectively, and the nebulizer gas was operated at 7.0 bar. UPLC-MS/MS analysis of the raw data was processed with MassLynx v4.2 and TargetLynx (Waters Corporation) for peak detection, alignment, and quantification. Internal standards were employed to maintain uniformity across runs, and pooled QC samples were

incorporated throughout the study to assess technical variability and instrument performance.

#### Statistical analysis

The metabolomic data comprised two datasets: Polar and Non-Polar Dataset. The Non-Polar dataset was further subdivided into positive ion mode and negative ion mode subsets. Initial quality assessment was conducted using relative standard deviation (RSD) values calculated from QC samples. The following filtering criteria were applied: (a) Metabolite abundance with the RSD greater than 20% was replaced with missing values ("Null"). (b) Metabolites with an RSD of 20% or less were retained without modification. These adjustments were made to the averaged dataset using the RSD values as a reference. The same preprocessing procedure was consistently applied across all three datasets: polar, positive non-polar lipids, and negative polar lipids.

#### Data cleaning and imputation

The metabolites inclusion workflow is presented in Supplementary Fig. 2. Out of 651 metabolites, 278 passed QC assessment. The RSD-filtered datasets were imported into Python for further processing. For handling missing data, we excluded metabolites with >40% missing values, and for the remaining metabolites with ≤40% missing values, imputation was performed by replacing missing entries with half the minimum observed value for that metabolite. Boxplots were generated prior to and following data transformations to evaluate distributions and to identify outliers (Supplementary Figs. 3, 4 and 5). To stabilize variance and normalize distributions, all metabolite values were log-transformed plus one to handle infinity. To address potential batch effects, the pyComBat package was employed. Post-adjustment boxplots were generated to assess batch normalization both across batches and individual metabolites (Supplementary Fig. 3). Subsequently, data was standardized using the StandardScaler function from the scikit-learn package, which scales features to zero mean and unit variance. Principal Component Analysis (PCA) was performed on the standardized datasets to reduce dimensionality and identify key contributors to variance. Metabolites were ranked according to their loading scores on the principal components. Following preprocessing, the three datasets were merged using unique Batch IDs and Sample IDs to maintain alignment and prevent redundancy, metabolite full annotation is given in the Supplementary Table 1. For the characteristics of clinical variables: Continuous variables were summarized as mean ± standard deviation (SD), and categorical variables were presented as frequencies and percentages (n [%]). The Kolmogorov–Smirnov test was used to assess data normality, confirming the use of parametric tests due to the normal distribution of the



variables. Comparisons of continuous variables between groups were performed using independent t-tests, while categorical variables were compared using the Chi-square test or Fisher's exact test, as appropriate.

### Association and prediction analysis

Logistic regression analysis was conducted to examine the associations between individual metabolites and GDM. Odds ratios (ORs) were estimated for each metabolite by modeling GDM status as the dependent variable and metabolite levels as independent variables. A significance threshold of FDR adjusted  $p$  value  $< 0.05$  was applied following correction for multiple testing. In addition, a Random Forest (RF) classification algorithm was employed to evaluate the predictive performance of metabolites for GDM. For the machine learning analysis, we used a stratified random partitioning approach. In each of ten iterations, the dataset was split into 70% training ( $n = 70$ ) and 30% test ( $n = 30$ ) subsets, stratified by case/control status. Within each training subset, we performed stratified 10-fold cross-validation for model tuning. Random forest hyperparameters were tuned using a grid search over  $mtry$  values and  $ntree$  values. Hyperparameter optimisation was performed exclusively within the training folds to prevent information leakage. A fixed random seed was applied to ensure reproducibility across iterations. Evaluation metrics included the area under the receiver operating characteristic curve (AUROC), F1 score, and overall accuracy, providing a comprehensive assessment of both sensitivity and specificity of the classification model. Feature importance scores were derived to identify metabolites contributing most to the prediction of GDM. To identify robust biomarker panels, a systematic approach involving ten iterations of random data partitioning into training and testing sets were employed, with reproducibility ensured through fixed random seeds. In each iteration, panels comprising 1 to 10 metabolites were constructed and evaluated for classification performance based on the AUC.

Model stability was assessed using bootstrap validation with 1000 resampling runs.

For each panel, performance metrics including AUC, sensitivity, specificity, and accuracy were aggregated, and their SDs were calculated. The most stable model within each panel size was defined as the one with the lowest SD. From these, the final biomarker panel was selected based on the highest AUC among the top models across all iterations. To assess whether the predictive performance of our machine learning model exceeded chance, we performed permutation testing. Outcome labels were randomly shuffled 1,000 times, and the model was refitted for each permutation using the same cross-validation and preprocessing procedures. The distribution of performance metrics (e.g., accuracy, AUROC) from the

permuted datasets were compared with the performance of the original model. A model performance significantly higher than the permuted distribution indicates that the observed patterns are unlikely to have arisen by chance, supporting the robustness of the identified signals.

All analyses were executed using R (version 4.2.2), employing the "train" function from the caret package to implement the RF. The predictive performance of the logistic regression model was assessed by constructing AUROC and corresponding 95% confidence intervals (CI) reported as primary measures of model discrimination. All statistical analyses were performed using R software (version 4.3.2), and a two-tailed  $p$  value  $< 0.05$  was considered statistically significant.

### Clinical model developments

Conventional risk predictors based on clinical accessibility null We considered the clinically accessible maternal risk factors and blood-based metabolomics signatures associated with the development of GDM, utilizing data from the original STRiDE cohort of 2,115 pregnant women [33]. These factors were included in the nested GDM case-control subset as potential conventional risk predictors and comprised maternal age, waist-to-height ratio, family history of diabetes, and venous HbA1c. All selected predictors are routinely measured in clinical practice at relatively low cost.

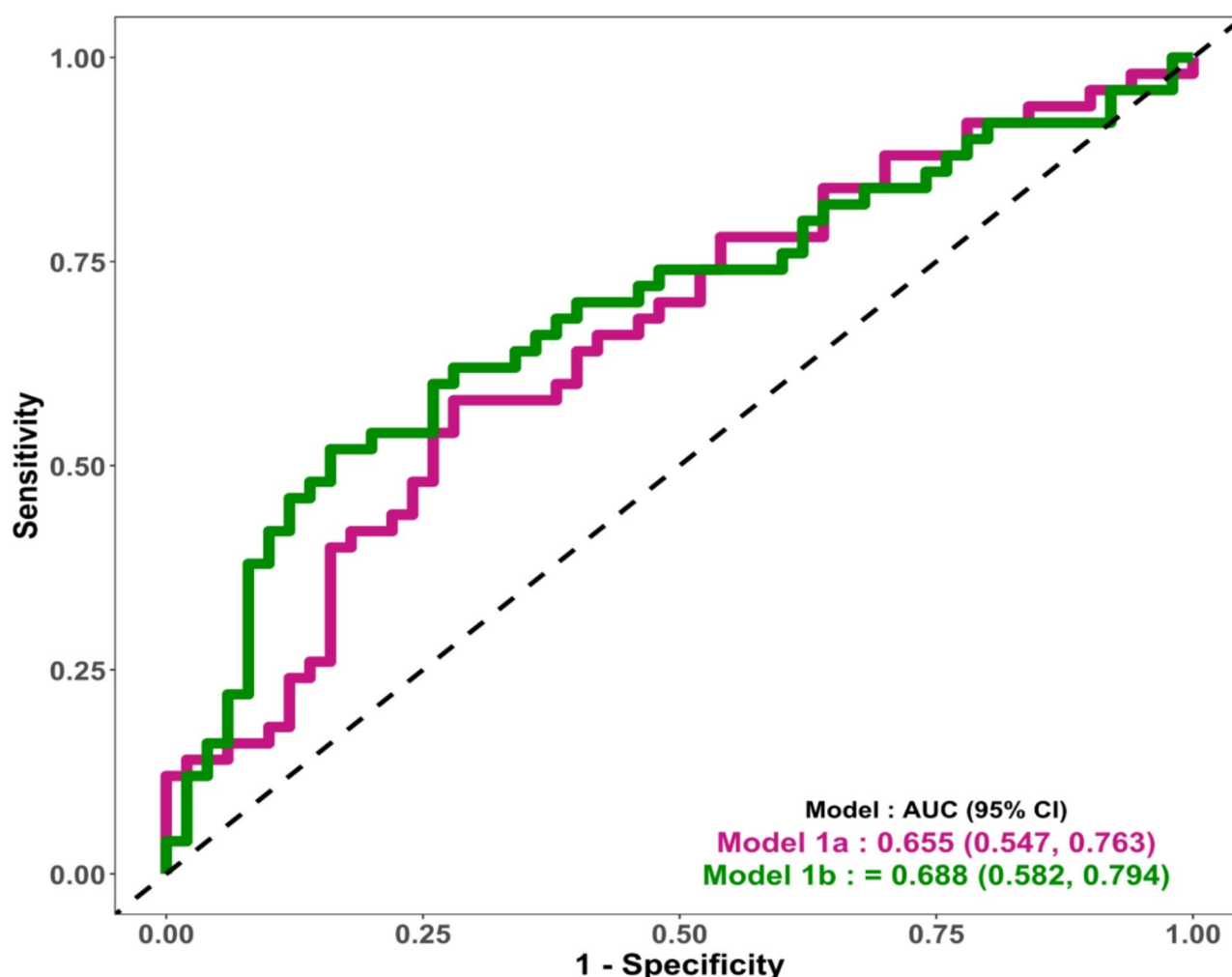
Initially, logistic regression models were constructed using maternal risk factors assessed during early pregnancy, with GDM status at the OGTT visit as the outcome. In Model 1a, maternal age, pre-pregnancy BMI, family history of diabetes, and venous HbA1c were included as covariates. Model 1b tested substituting waist-to-height ratio for pre-pregnancy BMI to evaluate the potential influence of central adiposity while adjusting for the same covariates from Model 1a (Fig. 1).

### Metabolomic signature predictors

This includes novel metabolomic signatures identified through mass spectrometry profiling. These biomarkers are currently not available for routine clinical use and require specialized laboratory facilities for measurement. We further integrated conventional maternal risk factors with novel metabolomic signatures to develop a clinically applicable and comprehensive model for GDM risk prediction.

### Metabolic pathway enrichment analysis

Metabolic pathway enrichment analysis was performed using KEGG compound identifiers (e.g., C00031, C00186). Metabolite-to-pathway associations were retrieved via the KEGG REST API, and a custom enrichment approach was implemented in R specifically for the human analysis. For each KEGG pathway, the overlap



**Fig. 1** GDM prediction using conventional risk factors. Model 1a: Composite risk score (venous HbA1c, age, BMI, and family history of diabetes) for GDM prediction. 1b: Conventional risk predictors (age, waist-to-height ratio, family history of diabetes and venous HbA1c at early pregnancy) for GDM prediction

between input metabolites and known pathway-associated metabolites were evaluated using Fisher's exact test. *p* values were adjusted for multiple testing using the Benjamini-Hochberg method to control the false discovery rate (FDR). Pathways with an FDR-adjusted *p* value < 0.05 were considered significantly enriched. Pathway annotations were retrieved using the *keggList* function to associate KEGG map IDs with pathway names.

## Results

### Characteristics of the study participants

Baseline and follow-up characteristics of the study population are shown in Table 1. The nested case-control samples included 100 women (50 GDM cases and 50 NGT controls), drawn from the larger STRiDE cohort (*n* = 2115). In early pregnancy (< 16 weeks), there were no significant differences in mean age or gestational age at inclusion between GDM and NGT groups. The average

age was approximately 26–27 years across all groups, with recruitment occurring at a mean gestational age of 10 weeks. Anthropometric measures showed a trend towards higher pre-pregnancy weight, BMI and waist circumference in the GDM group compared to the NGT group. Blood pressure measurements were comparable between groups at both visits. Fasting plasma glucose and HbA1c levels during early pregnancy were similar in both groups (FPG: 83 mg/dL; HbA1c: 5.02%), indicating that glycemic profiles were largely indistinguishable prior to the development of GDM. A higher proportion of GDM participants reported a family history of T2DM (54%) compared to NGT participants (40%). Additionally, a greater number of GDM cases belonged to the lower socioeconomic class (20% vs. 8%), suggesting a potential role for familial and socio-demographic factors in GDM risk. At the OGTT visit (24–28 weeks), a significant difference emerged in glucose regulation. Fasting,

**Table 1** Baseline characteristics at different visits of pregnancy

| Variables                               | All STRiDE participants (n=2115) | All (Nested case -control) participants (n=100) | NGT (n=50) | GDM (n=50) |
|---|----------------------------------|---|------------|------------|
| <i>Early pregnancy (&lt; 16 weeks)</i>  |                                  |   |            |            |
| Age (years)                             | 27±4.1                           | 26±2.5  | 26±2.4     | 27±2.6     |
| Gestational Age (weeks) at inclusion    | 10.5±3                           | 10.2±3  | 10±3.1     | 10.5±3.1   |
| Pre-pregnancy weight (kg)               | 59.2±12.6                        | 55.8±8.1  | 55.3±7.6   | 56.4±8.6   |
| Body mass index (kg/m <sup>2</sup> )    | 24.2±4.6                         | 23.3±2.5  | 22.8±2.1   | 23.8±2.9   |
| Waist Circumference (cm)                | 86±10                            | 83±8  | 82±7.5     | 85±8.2     |
| Systolic Blood Pressure (mm Hg)         | 102±10                           | 102±10  | 101±10.4   | 104±10     |
| Diastolic Blood Pressure (mm Hg)        | 69±9                             | 67±8  | 66±8       | 68±7.5     |
| Fasting plasma glucose (mg/dl)          | 83±5.2                           | 83±4.5  | 83±5.2     | 83±3.9     |
| HbA <sub>1c</sub> (Venous. %)           | 5.08±0.3                         | 5.02±0.3  | 5.02±0.3   | 5.02±0.3   |
| Family history of type 2 diabetes n [%] | 860 (40.7)                       | 47 (47)   | 20 (40)    | 27 (54)    |
| Socio Economic Status (SES) n (%)       |                                  |   |            |            |
| Lower Class                             | 499 (23.6)                       | 14 (14)   | 4 (8)      | 10 (20)    |
| Middle Class                            | 1177 (55.7)                      | 60 (60)   | 32 (64)    | 28 (56)    |
| Upper Class                             | 439 (20.8)                       | 26 (26)   | 14 (28)    | 12 (24)    |
| Nulliparous n (%)                       | 1291 (61)                        | 84 (84)   | 50 (100)   | 34 (68)*   |
| <i>OGTT Visit (24–28 weeks)</i>         |                                  |   |            |            |
| Systolic Blood Pressure (mm Hg)         | 102±10.5                         | 101±9.5   | 100±9.4    | 102±9.7    |
| Diastolic Blood Pressure (mm Hg)        | 68.8±9.4                         | 66±7.25   | 65.6±7.9   | 66.5±6.6   |
| OGTT Fasting plasma glucose (mg/dl)     | 81.5±8.1                         | 82.7±6.1  | 80.5±5.7   | 85±6.5**   |
| OGTT 1 h Venous plasma glucose (mg/dl)  | 137±30.8                         | 154±26  | 129±23.3   | 179±28.7** |
| OGTT 2 h Venous plasma glucose (mg/dl)  | 118±26.1                         | 130±22.7  | 112±16.8   | 148±28.6** |
| HbA <sub>1c</sub> (Venous. %)           | 5.08±0.3                         | 4.84±0.3  | 4.76±0.3   | 4.92±0.3*  |

Data presented as mean ± SD. \* $p < 0.01$ , \*\* $p < 0.001$  compared to NGT. OGTT—Oral Glucose Tolerance Test

1-hour, and 2-hour plasma glucose levels were higher in the GDM group compared to the NGT group ( $p < 0.01$  for all). Venous HbA<sub>1c</sub> was also modest but significantly higher in GDM participants ( $4.92 \pm 0.3\%$ ) compared to controls ( $4.76 \pm 0.3\%$ ,  $p < 0.05$ ). These findings highlight that while early pregnancy glycemic markers were similar across groups, clear metabolic differences emerged by mid-pregnancy, reinforcing the importance of dynamic glucose testing (i.e., OGTT) for timely identification of GDM.

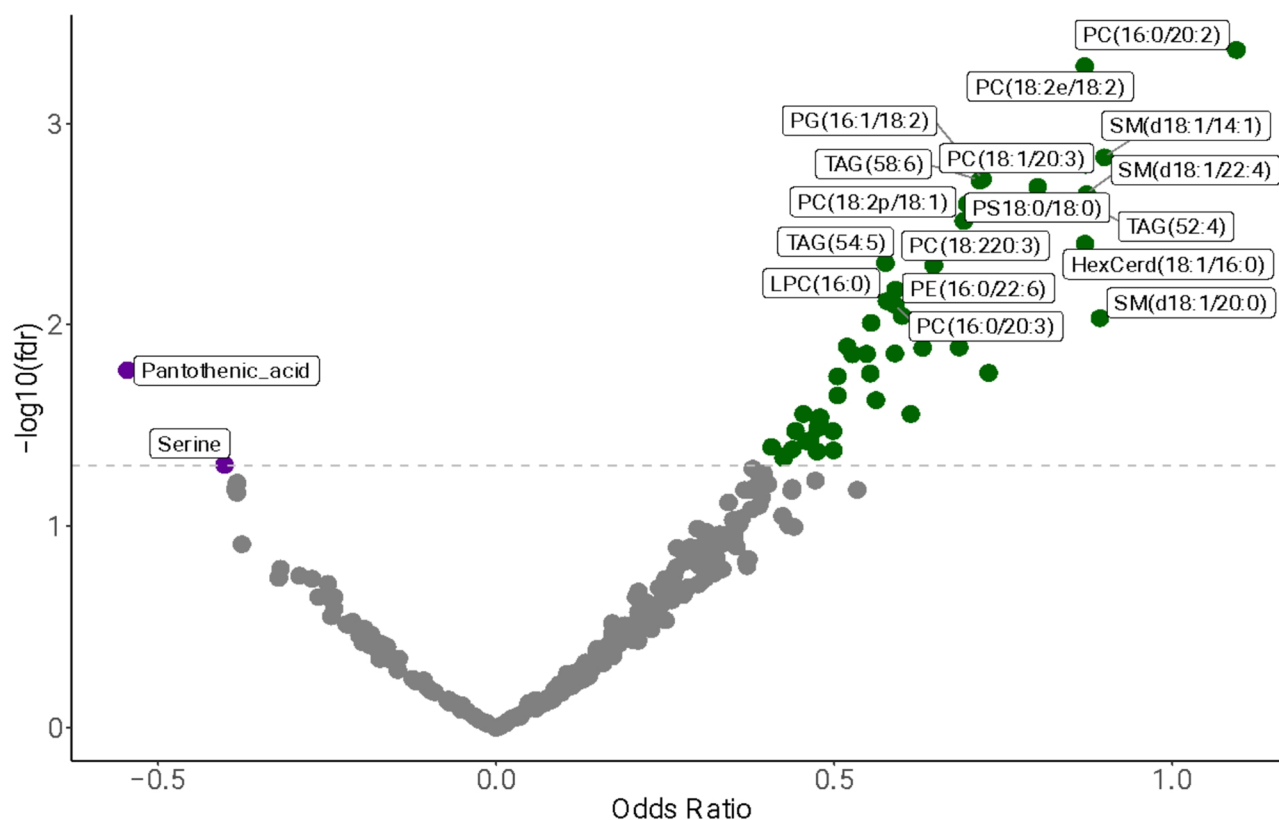
#### Metabolomic signatures in association with GDM

A total of 49 metabolites were found to be significantly associated with GDM based on logistic regression analysis (Fig. 2). Notably, lipid classes such as phosphatidylcholines (PC), sphingomyelins (SM), and triacylglycerols (TAG) were predominantly positively associated with GDM. Among the positively associated metabolites, PC(16:0/20:2) (OR=1.09; 95% CI: [0.54, 1.76]; FDR-adjusted  $p = 4.28 \times 10^{-4}$ ), PC(18:2/18:2) (OR=0.87; 95% CI: [0.41, 1.33]; FDR-adjusted  $p = 5.1 \times 10^{-4}$ ), and SM(d18:1/14:1) (OR=0.89; 95% CI: [0.40, 1.52]; FDR-adjusted  $p$  value=0.001) demonstrated the strongest associations. Pantothenic acid (Estimate = -0.54; 95% CI: [-1.02, -0.12]; FDR-adjusted  $p = 0.016$ ) and Serine (OR = -0.40; 95% CI: [-0.82, -0.017]; FDR-adjusted  $p = 0.049$ ) were negatively associated with GDM. Full statistical

details for the significant 49 metabolites are provided in Table 2, while comprehensive results for all metabolites are available in Supplementary Table 2.

#### Metabolomic association with other clinical variables

To investigate temporal changes in metabolite signatures across pregnancy, we generated a circular heatmap (Circos plot) illustrating differentially abundant metabolites across seven study timepoints: follow-up HbA<sub>1c</sub> (T1), oral glucose tolerance test phases (T2–T4), baseline HbA<sub>1c</sub> (T5), baseline glucose (T6) and pre-pregnancy BMI (T7). Metabolites, including PCs, phosphatidylethanolamines (PEs), SMs, TAGs, and other small molecules such as serine and pantothenic acid, were mapped based on their relative abundance and statistical significance at each timepoint. The plot revealed dynamic and time-specific shifts in lipid profiles. Several PCs (e.g., PC(16:0/22:6), PC(18:2e/18:3) and PEs (e.g., PE(18:0p/22:6)) were significantly upregulated during OGTT phases (T2–T4) and baseline (T5–T6), suggesting enhanced lipid remodeling or signaling activity during glucose challenge. Conversely, certain SMs (e.g., SM(d18:1/22:1) and amino acids like serine were down-regulated at early time points, particularly in the pre-pregnancy group (T7), indicating potential preclinical metabolic alterations (Fig. 3).



**Fig. 2** Volcano plot showing the association of metabolites with GDM. Each point represents a metabolite, plotted according to the odds ratio (OR) on the x-axis and the  $-\log_{10}$  of the FDR-adjusted  $p$  value on the y-axis from logistic regression analysis. Metabolites with OR > 1 and significant  $p$  values (above the horizontal threshold line) are considered positively associated with GDM and are colored in green; those with OR < 1 and significant  $p$  values are negatively associated and shown in blue. Non-significant metabolites are shown in gray. The horizontal dashed line denotes the significance threshold (FDR-adjusted  $p$  values < 0.05)

### Predictive analysis of metabolite signatures for GDM using machine learning

For the development of a conventional risk prediction model for GDM, we initially built a model using the maternal risk factors and blood-based biomarkers, linked to the development of GDM (AUC: 0.65) from the original cohort of 2,115 pregnant women (Fig. 1: Model 1a). The nested GDM case-control subset included the same factors and other potential conventional risk predictors. The predictors considered were age, waist-to-height ratio, family history of diabetes, and venous HbA1c, all routinely assessed at relatively affordable clinical costs. This prediction model had an AUC of 0.688 for GDM (Fig. 1: Model 1b).

For the development of the metabolomic signature panel and to evaluate its performance for early GDM prediction, we applied the RF machine learning method and bootstrap validation with 1000 repetitions, respectively, on a metabolomics dataset of 49 metabolites that were significantly associated. The RF method identified a panel of eight metabolites (PC(18:1/20:3), PC(16:0e/18:2), TAG(50:2), Serine, PE(18:0/20:4), HexCer(d18:1/16:0), SM(d18:1/14:1), and PS(18:0/18:0) with the best

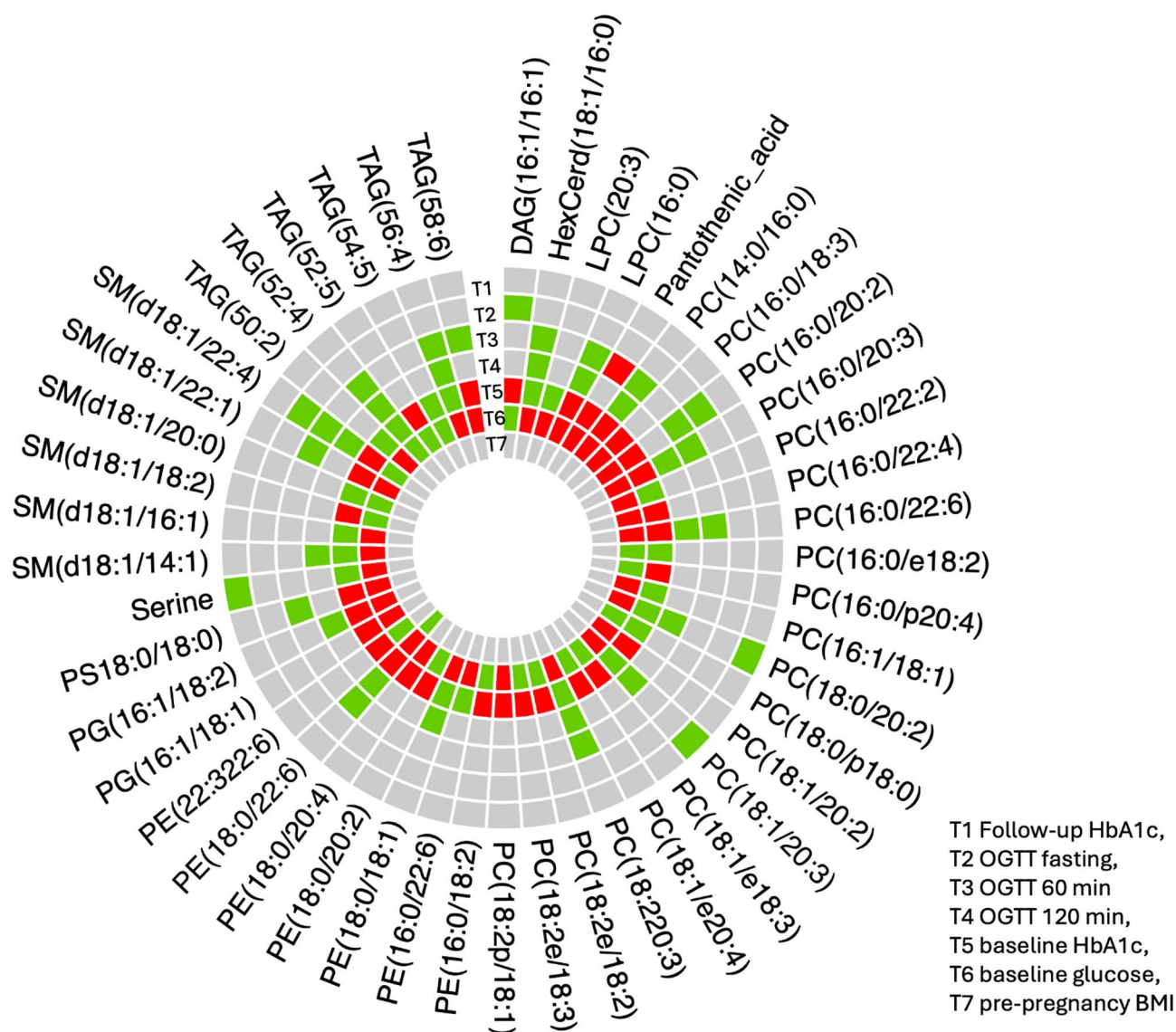
prediction performance for the GDM. During bootstrap validation, the panel maintained robust performance, yielding a mean AUC of 0.83, a mean sensitivity of 0.73 and a specificity of 0.75. Table 3 summarises the performance metrics for the selected 8 metabolite panel. Predictions for all other metabolites panels are given in the Supplementary Table 3. The mean levels of eight metabolites are presented as a violin plot in the Supplementary Fig. 6. In the iterative train/test splits, the eight-metabolite panel achieved a mean test-set AUC of 0.98. However, when assessed by 1,000-repeat bootstrap resampling, the mean AUC was 0.83 (95% CI: 0.77–0.89; SD: 0.06). The bootstrap estimate provides a more conservative measure of performance and reflects the variability expected in independent samples. The discrepancy between the single split test AUC and the bootstrap AUC likely reflects sampling variability due to the modest sample size. Further, permutation testing demonstrated that the model's predictive performance was significantly higher than expected by chance ( $p < 0.001$ ) (Supplementary Fig. 7, indicating that the identified metabolomic/lipidomic patterns are robust despite the limited sample size.



**Table 2** Summary of logistic regression analysis evaluating the association between individual metabolites and GDM

| Metabolite         | Estimate | CI 2.5% | CI 97.5% | SE   | p-value  | FDR-adjusted p-values |
|--------------------|----------|---------|----------|------|----------|-----------------------|
| PC(16:0/20:2)      | 1.09     | 0.54    | 1.76     | 0.31 | 4.28E-04 | 4.28E-04              |
| PC(18:2e/18:2)     | 0.87     | 0.42    | 1.40     | 0.25 | 5.15E-04 | 5.15E-04              |
| SM(d18:1/14:1)     | 0.90     | 0.40    | 1.53     | 0.28 | 1.46E-03 | 1.46E-03              |
| PC(18:1/20:3)      | 0.87     | 0.37    | 1.46     | 0.28 | 1.59E-03 | 1.59E-03              |
| PG(16:1/18:2)      | 0.72     | 0.29    | 1.20     | 0.23 | 1.88E-03 | 1.88E-03              |
| TAG(58:6)          | 0.71     | 0.30    | 1.21     | 0.23 | 1.90E-03 | 1.90E-03              |
| PC(16:0/22:6)      | 0.80     | 0.33    | 1.36     | 0.26 | 2.05E-03 | 2.05E-03              |
| SM(d18:1/22:4)     | 0.87     | 0.38    | 1.50     | 0.29 | 2.23E-03 | 2.23E-03              |
| PC(18:2p/18:1)     | 0.70     | 0.28    | 1.19     | 0.23 | 2.50E-03 | 2.50E-03              |
| TAG(52:4)          | 0.89     | 0.38    | 1.55     | 0.29 | 2.65E-03 | 2.65E-03              |
| PS(18:0/18:0)      | 0.69     | 0.25    | 1.17     | 0.23 | 3.03E-03 | 3.03E-03              |
| HexCerd(18:1/16:0) | 0.87     | 0.35    | 1.55     | 0.30 | 3.92E-03 | 3.92E-03              |
| TAG(54:5)          | 0.58     | 0.19    | 1.00     | 0.20 | 4.90E-03 | 4.90E-03              |
| PC(18:22:0:3)      | 0.65     | 0.22    | 1.13     | 0.23 | 5.02E-03 | 5.02E-03              |
| PE(16:0/22:6)      | 0.59     | 0.17    | 1.03     | 0.22 | 6.63E-03 | 6.63E-03              |
| LPC(16:0)          | 0.58     | 0.17    | 1.02     | 0.22 | 7.56E-03 | 7.56E-03              |
| PC(16:0/20:3)      | 0.59     | 0.18    | 1.06     | 0.22 | 7.91E-03 | 7.91E-03              |
| PC(18:2e/18:3)     | 0.60     | 0.19    | 1.09     | 0.23 | 8.96E-03 | 8.96E-03              |
| SM(d18:1/20:0)     | 0.89     | 0.34    | 1.71     | 0.34 | 9.19E-03 | 9.19E-03              |
| PE(22:322:6)       | 0.55     | 0.14    | 0.99     | 0.21 | 9.70E-03 | 9.70E-03              |
| PC(18:1/e18:3)     | 0.52     | 0.14    | 0.96     | 0.21 | 1.27E-02 | 1.27E-02              |
| PC(18:0/p18:0)     | 0.68     | 0.21    | 1.31     | 0.28 | 1.29E-02 | 1.29E-02              |
| TAG(56:4)          | 0.63     | 0.17    | 1.18     | 0.25 | 1.29E-02 | 1.29E-02              |
| PC(16:0/22:4)      | 0.59     | 0.15    | 1.10     | 0.24 | 1.38E-02 | 1.38E-02              |
| PE(18:0/20:2)      | 0.55     | 0.13    | 1.01     | 0.22 | 1.38E-02 | 1.38E-02              |
| SM(d18:1/16:1)     | 0.53     | 0.14    | 0.98     | 0.21 | 1.39E-02 | 1.39E-02              |
| Pantothenic_acid   | -0.55    | -1.02   | -0.12    | 0.23 | 1.67E-02 | 1.67E-02              |
| SM(d18:1/22:1)     | 0.73     | 0.24    | 1.45     | 0.31 | 1.72E-02 | 1.72E-02              |
| PC(16:0/p20:4)     | 0.55     | 0.14    | 1.05     | 0.23 | 1.73E-02 | 1.73E-02              |
| SM(d18:1/18:2)     | 0.50     | 0.11    | 0.95     | 0.21 | 1.79E-02 | 1.79E-02              |
| PC(14:0/16:0)      | 0.50     | 0.08    | 0.95     | 0.22 | 2.22E-02 | 2.22E-02              |
| PC(18:0/20:2)      | 0.56     | 0.12    | 1.09     | 0.25 | 2.35E-02 | 2.35E-02              |
| PE(18:0/22:6)      | 0.45     | 0.06    | 0.87     | 0.21 | 2.75E-02 | 2.75E-02              |
| DAG(16:1/16:1)     | 0.61     | 0.13    | 1.24     | 0.28 | 2.75E-02 | 2.75E-02              |
| PE(18:0/18:1)      | 0.48     | 0.06    | 0.92     | 0.22 | 2.85E-02 | 2.85E-02              |
| LPC(20:3)          | 0.48     | 0.08    | 0.95     | 0.22 | 3.03E-02 | 3.03E-02              |
| TAG(52:5)          | 0.47     | 0.05    | 0.93     | 0.22 | 3.24E-02 | 3.24E-02              |
| PC(18:1/e20:4)     | 0.44     | 0.06    | 0.88     | 0.21 | 3.34E-02 | 3.34E-02              |
| PE(18:0/20:4)      | 0.50     | 0.06    | 0.99     | 0.23 | 3.34E-02 | 3.34E-02              |
| PC(18:1/20:2)      | 0.46     | 0.04    | 0.92     | 0.22 | 3.72E-02 | 3.72E-02              |
| TAG(50:2)          | 0.46     | 0.04    | 0.91     | 0.22 | 3.76E-02 | 3.76E-02              |
| PC(16:0/e18:2)     | 0.41     | 0.05    | 0.85     | 0.20 | 4.01E-02 | 4.01E-02              |
| PG(16:1/18:1)      | 0.44     | 0.04    | 0.89     | 0.21 | 4.12E-02 | 4.12E-02              |
| PE(16:0/18:2)      | 0.50     | 0.04    | 1.01     | 0.24 | 4.17E-02 | 4.17E-02              |
| PC(16:1/18:1)      | 0.47     | 0.05    | 0.98     | 0.23 | 4.23E-02 | 4.23E-02              |
| PC(16:0/22:2)      | 0.43     | 0.02    | 0.85     | 0.21 | 4.44E-02 | 4.44E-02              |
| PC(16:0/18:3)      | 0.42     | 0.02    | 0.86     | 0.21 | 4.55E-02 | 4.55E-02              |
| Serine             | -0.40    | -0.83   | -0.02    | 0.20 | 4.92E-02 | 4.92E-02              |

The table includes the estimated regression coefficient (Estimate), standard error (SE), 95% confidence interval bounds (CI 2.5% and CI 97.5%), nominal *p* values, and false discovery rate FDR adjusted *p* values for multiple testing correction.



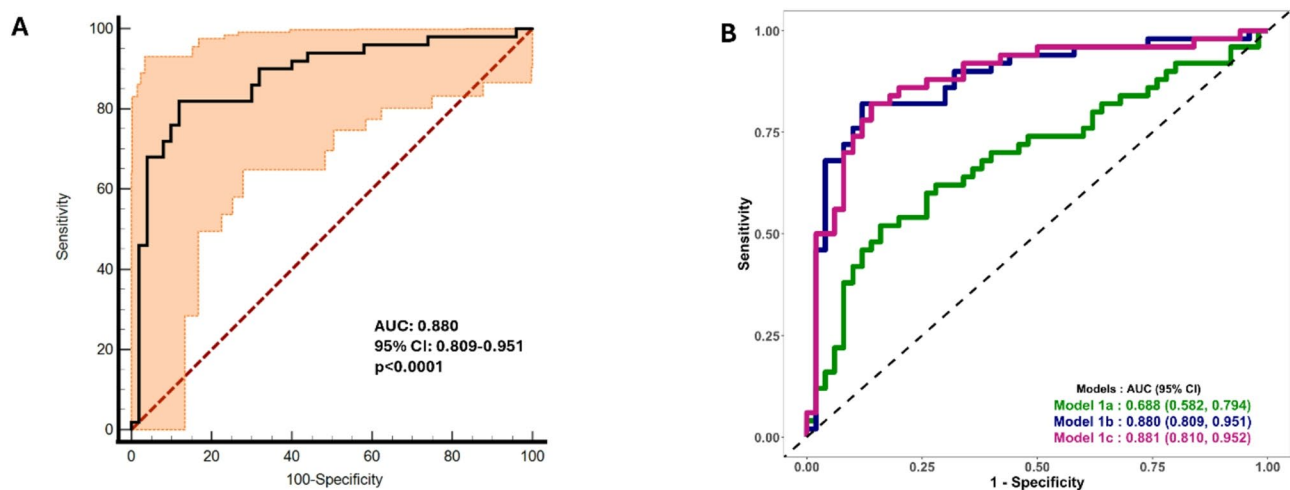
**Fig. 3** Circos heatmap plot illustrating associations between baseline metabolites and clinical variables related to glycemic outcomes. The outermost circle displays metabolite names. Inner rings represent clinical variables including follow-up HbA1c, OGTT fasting, 60 min and 120 min, baseline HbA1c, baseline glucose, and pre-pregnancy BMI. Colored bars within each ring indicate the direction of association between each metabolite and the corresponding variable (green: upregulated; red: downregulated; grey: no change)

**Table 3** Performance of the 8-Metabolite panel selected by random forest for GDM prediction

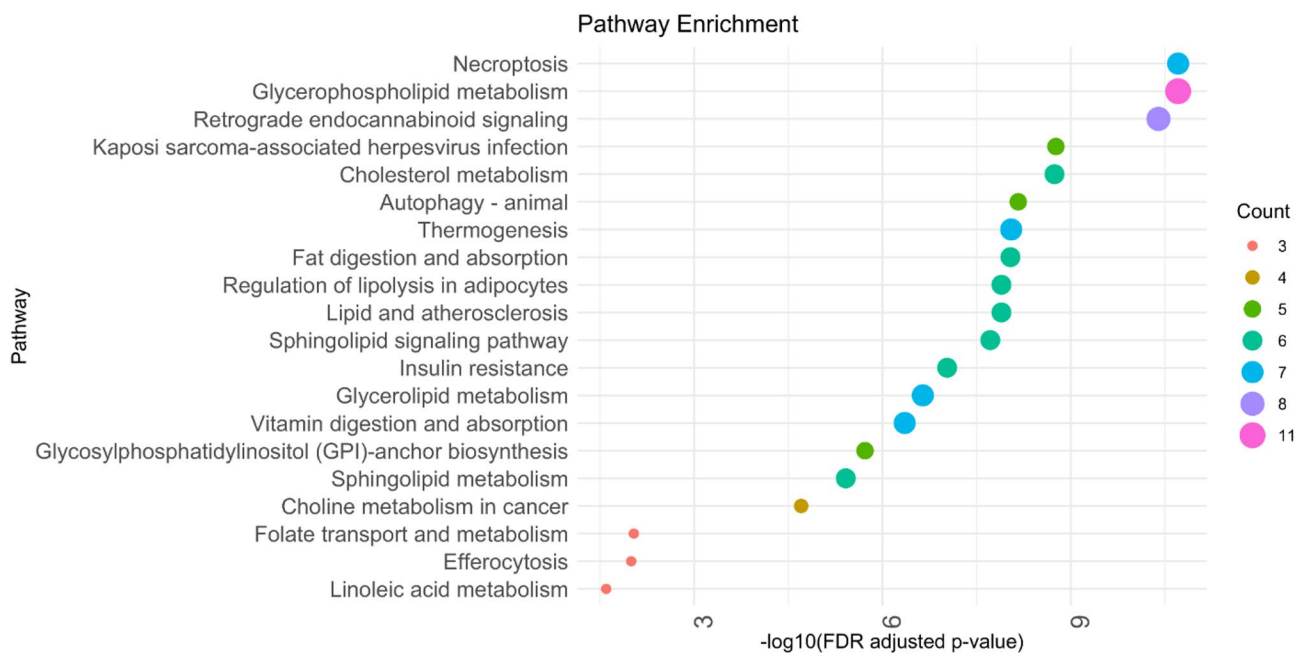
| Method                           | Dataset          | AUC  | Sensitivity | Specificity | F1 Score | Accuracy |
|----------------------------------|------------------|------|-------------|-------------|----------|----------|
| Random Forest<br>(8 metabolites) | Test Set         | 0.98 | 1           | 0.56        | 0.82     | 0.78     |
|                                  | Bootstrap (Mean) | 0.83 | 0.73        | 0.75        | —        | 0.74     |

After obtaining the best panel of 8 metabolites through building step-by-ML approaches and bootstrap validation, further logistic regression models were constructed to deepen the understanding of these metabolites and their association with other maternal risk factors linked to GDM prediction. ROC analysis demonstrated a robust AUC of 0.880 (95% CI: 0.809– 0.951) with a sensitivity of 82% and a specificity of 88% (Fig. 4a).

To develop a clinically efficient comprehensive GDM prediction model, we integrated clinically accessible maternal risk factors (conventional risk predictors) with novel 8 metabolomics signatures. Interestingly, the addition of the conventional risk predictors has slightly increased the AUC of the predicted GDM risk to 0.881, suggesting that the full GDM prediction model at early pregnancy improved the discrimination even after adjusting for the potential maternal risk factors (Fig. 4b:



**Fig. 4** ROC analysis of the 8 metabolomic markers and conventional risk factors at early pregnancy. **a** ROC showing the novel 8 metabolomic signatures as metabolomic biomarker predictors of GDM mean in black curve and CI in orange area. **b** ROC curves and AUC of different prediction models. Model 1a: Conventional risk predictors (age, waist-to-height ratio, family history of diabetes and HbA1c), Model 1b: Novel metabolomic biomarker panel of 8 markers at early pregnancy. Model 1c: Full GDM prediction model - integrated clinically accessible maternal risk factors (conventional risk predictors) with novel metabolomic panel of 8 signatures (1a & 1b combined)



**Fig. 5** Dot plot showing the top 10 enriched KEGG pathways based on significantly altered metabolites. The y-axis represents enriched pathways, and the x-axis shows  $-\log_{10}$  (FDR adjusted  $p$  value) from a hypergeometric test. Dot sizes correspond to the number (as a count) of metabolites associated with each pathway

Model 1c). Parity, a known risk factor for GDM, was summarized in Table 1. As a sensitivity analysis, we have adjusted our model for parity, and no significant differences were observed in the associations of the eight metabolites with GDM status.

#### Pathway enrichment analysis

Pathway enrichment analysis identified several significantly enriched metabolic and signaling pathways

(FDR-adjusted  $p$  values  $< 0.05$ ) associated with the studied condition (Fig. 5). Notably, lipid-related pathways were highly represented, including Glycerophospholipid metabolism (FDR-adjusted  $p$  values =  $1.96 \times 10^{-11}$ ), Glycerolipid metabolism (FDR-adjusted  $p$  values =  $2.28 \times 10^{-7}$ ), Sphingolipid metabolism (FDR-adjusted  $p$  values =  $3.85 \times 10^{-6}$ ), and Cholesterol metabolism (FDR-adjusted  $p$  values =  $1.82 \times 10^{-9}$ ).

In addition, signaling and regulatory pathways such as Sphingolipid signaling (FDR-adjusted  $p$  values =  $1.92 \times 10^{-8}$ ), Autophagy (FDR-adjusted  $p$  values =  $6.88 \times 10^{-9}$ ), Necroptosis (FDR-adjusted  $p$  values =  $1.96 \times 10^{-11}$ ), and Retrograde endocannabinoid signaling (FDR-adjusted  $p$  values =  $4.02 \times 10^{-11}$ ) were significantly enriched. Other relevant findings included pathways involved in insulin resistance, fat/vitamin digestion and absorption, regulation of lipolysis, and folate metabolism. These results underscore the involvement of lipid metabolism, cell death pathways, and metabolic regulation in the underlying GDM pathophysiology.

## Discussion

Recent studies have consistently highlighted disruptions in lipid metabolism [28, 29], amino acid turnover [37], and energy homeostasis [38] in the pathogenesis of GDM. In this context, using high-throughput UPLC-MS/MS within the nested STRiDE cohort, we identified 49 metabolites, measured during early pregnancy, significantly associated with later GDM development. The results emphasise the contribution of lipid metabolism, cell death mechanisms, and lipid metabolic regulation to the underlying pathophysiology of GDM. Importantly, machine learning models revealed that a panel of eight metabolites could predict GDM with high accuracy, underscoring the potential clinical utility of metabolomics in early risk stratification. Traditional clinical risk factors alone provided the predictive power (AUC = 0.688), their integration with the metabolomic panel robustly improved prediction (AUC = 0.881), reinforcing the added value of molecular biomarkers in refining risk assessment. Findings from this metabolomic study in an Indian pregnancy cohort highlight the potential of early pregnancy novel metabolomic biomarkers, and this is the first comprehensive study to report such associations in early GDM prediction.

Studies by Zhao et al. and Scholtens et al., have demonstrated that metabolic derangements precede hyperglycemia, supporting the rationale for early screening using molecular signatures [17, 18]. Other studies, such as those by Guasch-Ferré et al., have integrated metabolomics with clinical risk factors to enhance predictive models, showing that a combination of metabolite panels and maternal characteristics can significantly improve early diagnostic accuracy compared to traditional screening alone [39]. Our findings corroborate and extend previous studies that identified early-pregnancy lipid perturbations associated with GDM. Like prior reports from Zhao et al., Scholtens et al. and Liu et al., we observed alterations in glycerophospholipids and sphingolipids. The novelty of the present study lies in its focus on an Indian population, among whom GDM prevalence is high, and in the combined profiling of polar and

non-polar metabolites within a single analytical framework. The eight-metabolite panel we identified is therefore best viewed as a population-specific signature for an Indian population [40], that aligns with broader evidence linking early lipid metabolism to GDM risk. The eight metabolites identified by the random forest algorithm were PC(18:1/20:3), PC(16:0e/18:2), TAG(50:2), Serine, PE(18:0/20:4), HexCer(d18:1/16:0), SM(d18:1/14:1), and PS(18:0/18:0), represent diverse biochemical classes, including glycerophospholipids, sphingolipids, triacylglycerols, and amino acid derivatives. These classes of metabolites are central to GDM pathogenesis, such as insulin resistance [41] lipid imbalance [42], and mitochondrial dysfunction [43, 44].

PCs, both diacyl and ether-linked, were significantly elevated in women who later developed GDM. These findings align with previous studies that reported higher PCs in GDM and T2DM cohorts, such as those by Liu et al. and Zhao et al. [17, 29]. PC(18:1/20:3) may reflect enhanced hepatic lipid synthesis or impaired lipoprotein metabolism [45], both of which are common in insulin-resistant states. The ether-linked PC(16:0e/18:2) also signals oxidative stress and disrupted peroxisomal function, consistent with inflammatory activation observed in early GDM [46–48].

Triacylglycerols, such as TAG(50:2), are classical indicators of insulin resistance and dyslipidemia. Elevated TAG levels may indicate hepatic steatosis or impaired lipid clearance mechanisms, which are frequently observed in obese and GDM pregnancies [19, 39]. These results further validate the role of altered lipid storage and mobilization as early drivers of metabolic dysfunction in pregnancy.

The L-serine metabolism has shown to be altered in type 1, type 2, and gestational diabetes mellitus. Hølem et al. [49], study suggests that L-serine supplementation can improve glucose homeostasis and mitochondrial function, indicating its potential therapeutic role in diabetes management. Ethanolamine is a precursor for PE, a mitochondrial membrane lipid essential for bioenergetics and insulin sensitivity [50]. Elevated PE(18:0/20:4), containing arachidonic acid, suggests a link to pro-inflammatory eicosanoid production and mitochondrial stress [51–53]. Barranco-Altirriba et al. [54] and Chang et al. [55] have highlighted the association of altered PE species with metabolic diseases, supporting our findings.

Hexosylceramides, a subset of sphingolipids, were increasingly recognized for their role in insulin resistance and  $\beta$ -cell dysfunction [56]. The strong positive association of HexCer(d18:1/16:0) with GDM from our study mirrors the findings by Wittenbecher et al., who demonstrated their predictive role for T2DM [57]. SMs modulate membrane microdomain structure, and their dysregulation may disrupt insulin receptor localization,

thereby contributing to insulin resistance [58]. Beyond GDM, SMs are known to be associated with T2DM and mediating the effect of BMI on T2DM [59, 60]. We observed that SM(d18:1/14:1), were elevated in the GDM group, consistent with previous metabolomic studies in pregnant and non-pregnant populations [61]. Phosphatidylserine, though less studied in the context of GDM, plays a key role in cell signaling, apoptosis, and membrane structure [62]. Lappas et al. further supported these findings by showing that cholesteryl ester species, alkenyl phosphatidylethanolamines, and phosphatidylserine species were most strongly associated with the risk of developing T2DM after a GDM pregnancy [28]. Its elevation in GDM pregnancies may reflect altered placental membrane integrity or increased apoptotic signaling both potential contributors to impaired glucose homeostasis. Zhao et al., identified similar trends in related glycerophospholipid classes [17]. In this context, the pathway enrichment analysis in the present study revealed key pathways, including glycerophospholipid and sphingolipid metabolism, cholesterol metabolism, insulin resistance, necroptosis, and autophagy. These findings highlight the central role of lipid-related and signaling pathways in the GDM studies. Our findings serve as proof of concept and may inform future mechanistic studies exploring the potential role of these metabolomic signatures in the development of GDM across different ethnic groups. Additionally, the identified metabolomic panel may prove valuable for screening high-risk individuals. However, further validation in larger, independent cohorts and different ethnicities is needed to confirm these results.

The strengths of this study include that metabolomics results were drawn from a well-characterized, a nested case-control design embedded within a large, diverse cohort, enhancing internal validity. A high-resolution mass spectrometry capturing both polar and non-polar metabolite classes, enabling a broad metabolic landscape and rigorous statistical and machine learning validation strategies, including bootstrapping and cross-validation, increased robustness and minimised overfitting. Finally, our study focused on a high-risk population, Indian women, who exhibit higher GDM susceptibility and are underrepresented in metabolomic studies, enhancing the relevance and novelty of the findings.

This study has important limitations. First, the sample size of 100 women (50 GDM, 50 controls) is relatively modest. Although the nested case-control design strengthens internal validity, it also heightens the risk of overfitting in the machine learning analyses. Nevertheless, the statistical power estimation study by Blaise et al. [63] supports the adequacy of our sample size and validates the benchmark employed for the machine learning approach. Second, although we employed both stratified

cross-validation and extensive bootstrap resampling to estimate performance, these are internal validation techniques. We did not validate the eight-metabolite panel in an independent external cohort, and therefore its clinical utility and generalisability remain speculative. Prospective validation in larger, geographically and ethnically diverse cohorts, ideally using targeted assays for the candidate metabolites, will be required before the panel can be considered for clinical translation. Third, our matching accounted for age and BMI but did not explicitly match on parity, although parity status was summarised and found not to differ materially between groups. Finally, our findings are derived from a single population and should be interpreted as exploratory. Despite stated limitations, the strength includes that this study was designed as an initial, hypothesis-generating nested case-control analysis within the large STRiDE cohort. The nested design enabled strict matching on key confounders and high-quality, early-pregnancy biosamples, which were profiled on the same UPLC-MS/MS platform, thereby reducing technical heterogeneity. These features enhance internal validity and render the identified metabolite classes biologically plausible (lipids, sphingolipids, phosphatidylcholines), consistent with prior literature. We performed extensive internal validation and complemented it with permutation testing, which confirmed that the model's predictive performance was significantly greater than expected by chance. These results indicate that the identified metabolomic signatures are robust despite the modest sample size. We acknowledge that external validation is currently not feasible due to the lack of pregnancy focused cohorts with comparable metabolomics data in Indian populations, underscoring the novelty and unique contribution of our study. By generating high-resolution metabolomic profiles in this cohort, we provide a valuable metabolomics dataset that can serve as a foundation for future larger studies and for the upscaling of metabolomics research in this population. Moreover, our findings reinforce and extend earlier metabolomics studies. Prior studies by Scholtens et al. [18] and Zhao et al. [17] identified similar metabolic disruptions as an individual metabolite. Interestingly, our study is the first to identify an integrated 8-metabolite panel representing meaningful biology signals even with a modest sample size, in an Indian cohort, a population characterized by a high prevalence of GDM and distinct metabolic phenotypes.

## Conclusion

In summary, our study provides novel insights into early metabolic alterations associated with GDM in Indian women. An eight-metabolite panel identified through affordable MS-based profiling offers excellent predictive performance and reflects core biological processes



implicated in GDM pathogenesis. These findings support the use of metabolomics as a powerful tool for early GDM screening and underscore the need for external validation and translational efforts to integrate these markers into clinical practice.

### Supplementary Information

The online version contains supplementary material available at <https://doi.org/10.1186/s12933-025-02978-0>.

Supplementary Material 1: Table S1: Complete List of Metabolites with Annotations (Post-Preprocessing and Merging). Table S2: Details of all metabolites associated with GDM. Table S3: Prediction Performance Metrics of All Metabolite Panels for GDM. Figure S1: Study design. Figure S2: Metabolomics data QC work flow. Figure S3: Non-Polar Negative Data - Standardized. Boxplot of all the non-polar negative metabolites on x-axis and adjusted for the batch effect and scaled metabolite frequencies on y-axis. Figure S4: Non-Polar Positive Data - Standardized. Boxplot of all the non-polar positive metabolites on x-axis and adjusted for the batch effect and scaled metabolite frequencies on y-axis. Figure S5: Polar Positive Data - Standardized. Boxplot of all the polar metabolites on x-axis and adjusted for the batch effect and scaled metabolite frequencies on y-axis. Figure S6: A violin plot representing metabolites distribution across normal and GDM. Figure S7: Permutation test distribution of AUC.

### Acknowledgements

Dr. Gokulakrishnan Kuppan is a current recipient of a DBT-Wellcome Trust India Alliance Intermediate Clinical & Public Health Fellowship (Grant Number IA/CPHI/18/1/503964) and acknowledges the funding received from the Anusandhan National Research Foundation (ANRF) for this study. STRiDE study was funded by the MRC-DBT Newton fund (MRC Newton Fund MR/N006232/1). The authors would like to thank all the study participants and the staff of the STRiDE study. Dr. Usharani Dandamudi acknowledges the funding received from the Anusandhan National Research Foundation (ANRF) SERB—POWER GRANT (File No: SPG/2021/002205).

### Author contributions

KG has conceptualized, written, and taken the lead in completing manuscripts through all stages of preparation and submission. YS, SV, DP, VP and DU conducted mass spectrometry measurements. SS, PG, AT, PN, AR, AS and KG did statistical analysis for the study and conducted biostatistical analysis. KG, SS, PS, JA, SS, VP, MB, RA, UR, and VM have reviewed the project from the conceptualization stage and have contributed to each version of the manuscript. MD, RA, and UR are members of the research team who participated in the conduct of the study. All authors have contributed to the article critically for intellectual content and have provided final approval of the version to be published.

### Funding

Open Access funding enabled and organized by Projekt DEAL.

### Data availability

No datasets were generated or analysed during the current study.

### Declarations

### Competing interests

The authors declare no competing interests.

### Author details

<sup>1</sup>Institute of Epidemiology, Research Unit of Molecular Epidemiology, Helmholtz Zentrum München, German Research Center for Environmental Health, 85764 Neuherberg, Germany

<sup>2</sup>CEPD Division, CSIR-National Chemical Laboratory, Dr. Homi Bhabha Road, Pune 411008, India

<sup>3</sup>Academy of Scientific and Innovative Research (AcSIR), Ghaziabad 201002, India

<sup>4</sup>Systems Biology and Data Analytics Research Lab, Amity Institute of Biotechnology, Amity University, Uttar Pradesh, Noida 201313, India

<sup>5</sup>Department of Epidemiology, Madras Diabetes Research Foundation, Chennai, India

<sup>6</sup>Department of Diabetology, Madras Diabetes Research Foundation (ICMR-Collaborating Centre of Excellence) & Dr. Mohan's Diabetes Specialities Centre (IDF Centre of Excellence in Diabetes Care), Chennai, India

<sup>7</sup>Seethapathi Clinic & Hospital, Chennai, India

<sup>8</sup>Indian Council of Medical Research, Ansari Nagar, New Delhi, India

<sup>9</sup>Department of Cell and Molecular Biology, Madras Diabetes Research Foundation, Chennai, India

<sup>10</sup>Institute of Experimental Genetics, Helmholtz Zentrum München, German Research Center for Environmental Health, Ingolstädter Landstraße 1, 85764 Neuherberg, Germany

<sup>11</sup>Department of Biochemistry, Yong Loo Lin School of Medicine, National University of Singapore, 8 Medical Drive, Singapore 117597, Singapore

<sup>12</sup>Institute of Biochemistry, Faculty of Medicine, University of Ljubljana, Vrazov Trg 2, 1000 Ljubljana, Slovenia

<sup>13</sup>Populations, Evidence and Technologies, Division of Health Sciences, Warwick Medical School, University of Warwick, Coventry, UK

<sup>14</sup>Department of Diabetes, Endocrinology and Metabolism, George Eliot Hospital, Nuneaton, UK

<sup>15</sup>Centre for Global Health, University of Warwick, Coventry, UK

<sup>16</sup>Department of Food Safety and Analytical Quality Control Laboratory, CSIR-Central Food Technological Research Institute (CFTRI), Mysore, Karnataka 570020, India

<sup>17</sup>Department of Neurochemistry, National Institute of Mental Health and Neuro Sciences (NIMHANS), Hosur Road, Bengaluru, Karnataka 560029, India

Received: 2 September 2025 / Accepted: 10 October 2025

Published online: 14 November 2025

### References

1. Buchanan TA, Xiang AH, Page KA. Gestational diabetes mellitus: risks and management during and after pregnancy. *Nat Rev Endocrinol*. 2012;8(11):639–49.
2. Bhavadharini B, Mahalakshmi MM, Anjana RM, Maheswari K, Uma R, Deepa M, et al. Prevalence of gestational diabetes mellitus in urban and rural Tamil Nadu using IADPSG and WHO 1999 criteria (WINGS 6). *Clin Diabetes Endocrinol*. 2016;2:8.
3. Choudhury AA, Devi Rajeswari V. Gestational diabetes mellitus - A metabolic and reproductive disorder. *Biomed Pharmacother Biomedicine Pharmacother*. 2021;143:112183.
4. Seshiah V, Balaji V, Balaji MS, Paneerselvam A, Arthi T, Thamizharasi M, et al. Gestational diabetes mellitus manifests in all trimesters of pregnancy. *Diabetes Res Clin Pract*. 2007;77(3):482–4.
5. Saravanan P, Diabetes in Pregnancy Working Group, Maternal Medicine Clinical Study Group. Royal college of obstetricians and Gynaecologists, UK. Gestational diabetes: opportunities for improving maternal and child health. *Lancet Diabetes Endocrinol*. 2020 Sept;8(9):793–800.
6. Ye W, Luo C, Huang J, Li C, Liu Z, Liu F. Gestational diabetes mellitus and adverse pregnancy outcomes: systematic review and meta-analysis. *BMJ*. 2022;377:e067946.
7. Diaz-Santana MV, O'Brien KM, Park YMM, Sandler DP, Weinberg CR. Persistence of risk for type 2 diabetes after gestational diabetes mellitus. *Diabetes Care*. 2022;45(4):864–70.
8. Kramer CK, Campbell S, Retnakaran R. Gestational diabetes and the risk of cardiovascular disease in women: a systematic review and meta-analysis. *Diabetologia*. 2019;62(6):905–14.
9. Yu Y, Soohoo M, Sørensen HT, Li J, Arah OA. Gestational diabetes mellitus and the risks of overall and type-specific cardiovascular diseases: a population- and sibling-matched cohort study. *Diabetes Care*. 2022;45(1):151–9.
10. Chen A, Tan B, Du R, Chong YS, Zhang C, Koh AS, et al. Gestational diabetes mellitus and development of intergenerational overall and subtypes of cardiovascular diseases: a systematic review and meta-analysis. *Cardiovasc Diabetol*. 2024;23(1):320.
11. Lorenzo-Almorós A, Hang T, Peiró C, Soriano-Guillén L, Egidio J, Tuñón J, et al. Predictive and diagnostic biomarkers for gestational diabetes and its

- associated metabolic and cardiovascular diseases. *Cardiovasc Diabetol.* 2019;18(1):140.
12. Moon JH, Choi SH. Advancing early prediction of gestational diabetes mellitus with circular RNA biomarkers. *Diabetes Metab J.* 2025;49(3):403–4.
  13. Pramodkumar TA, Hannah W, Anjana RM, Ram U, Tiwaskar M, Gokulakrishnan K, et al. Biomarkers of gestational diabetes mellitus: mechanisms, advances, and clinical utility. *J Assoc Physicians India.* 2025;73(2):56–67.
  14. Yang J, Cao Y, Qian F, Grewal J, Sacks DB, Chen Z et al. Early prediction of gestational diabetes mellitus based on systematically selected multi-panel biomarkers and clinical accessibility—a longitudinal study of a multi-racial pregnant cohort. *BMC Med.* 2025;23(1):430.
  15. Razo-Azamar M, Nambo-Venegas R, Meraz-Cruz N, Guevara-Cruz M, Ibarra-González I, Vela-Amieva M et al. An early prediction model for gestational diabetes mellitus based on metabolomic biomarkers. *Diabetol Metab Syndr.* 2023;15(1):116.
  16. Jung Y, Lee SM, Lee J, Kim Y, Lee W, Koo JN, et al. Metabolomic profiling reveals early biomarkers of gestational diabetes mellitus and associated hepatic steatosis. *Cardiovasc Diabetol.* 2025;24(1):125.
  17. Zhao H, Zheng Y, Zhu L, Xiang L, Xu S, Cai Z. Trimester-specific urinary metabolome alterations associated with gestational diabetes mellitus: a study in different pregnancy stages. *Chin Chem Lett.* 2022;33(6):3139–43.
  18. Scholtens DM, Kuang A, Lowe LP, Hamilton J, Lawrence JM, Lebenthal Y, et al. Hyperglycemia and adverse pregnancy outcome follow-up study (HAPO FUS): maternal glycemia and childhood glucose metabolism. *Diabetes Care.* 2019;42(3):381–92.
  19. Morkkala K, Vahlberg T, Pellonperä O, Houttu N, Koivuniemi E, Laitinen K. Distinct metabolic profile in early pregnancy of overweight and obese women developing gestational diabetes. *J Nutr.* 2020;150(1):31–7.
  20. Kong M, Lu Z, Zhong C, Gao Q, Zhou X, Chen R, et al. A higher level of total bile acid in early mid-pregnancy is associated with an increased risk of gestational diabetes mellitus: a prospective cohort study in Wuhan, China. *J Endocrinol Invest.* 2020;43(8):1097–103.
  21. Li J, Huo X, Cao YF, Li SN, Du Z, Shao P, et al. Bile acid metabolites in early pregnancy and risk of gestational diabetes in Chinese women: a nested case-control study. *EBioMedicine.* 2018;35:317–24.
  22. Lu C, Li C, Lei X. Bile acids and gestational diabetes mellitus: exploring the link and implications—a review. *Front Endocrinol.* 2025;16:1574228.
  23. Gao B, Shen Q, Wu Y, Cao M, Zhang Q, Chen L. Branched-chain and aromatic amino acid levels response to an oral glucose load associated with gestational diabetes mellitus. *Sci Rep.* 2022;12(1):12257.
  24. Susarla SM, Fiehn O, Thiele I, Ngo AL, Barupal DK, Chehab RF, et al. Microbiome-derived metabolites in early to mid-pregnancy and risk of gestational diabetes: a metabolome-wide association study. *BMC Med.* 2024;22(1):449.
  25. Thirumoorthy C, Rekha RP, Deepa M, Ram U, Shalu D, Venkatesan U, et al. Association of early pregnancy telomere length and mitochondrial copy number with gestational diabetes mellitus and depressive symptoms. *Psychoneuroendocrinology.* 2025;176:107431.
  26. Yu J, Ren J, Ren Y, Wu Y, Zeng Y, Zhang Q, et al. Using metabolomics and proteomics to identify the potential urine biomarkers for prediction and diagnosis of gestational diabetes. *eBioMedicine.* 2024;101:105008.
  27. Borges Manna L, Syngelaki A, Würtz P, Koivu A, Sairanen M, Pölonen T et al. First-trimester nuclear magnetic resonance-based metabolomic profiling increases the prediction of gestational diabetes mellitus. *Am J Obstet Gynecol.* 2025;233(1):71.e1–71.e14.
  28. Lappas M, Munda PA, Wong G, Huynh K, Jinks D, Georgiou HM, et al. The prediction of type 2 diabetes in women with previous gestational diabetes mellitus using lipidomics. *Diabetologia.* 2015;58(7):1436–42.
  29. Liu J, Li J, Li S, Leng J, Li W, Yang W, et al. Circulating lysophosphatidylcholines in early pregnancy and risk of gestational diabetes in Chinese women. *J Clin Endocrinol Metab.* 2020;105(4):dgaa058.
  30. Rahman ML, Feng YCA, Fiehn O, Albert PS, Tsai MY, Zhu Y, et al. Plasma lipidomics profile in pregnancy and gestational diabetes risk: a prospective study in a multiracial/ethnic cohort. *BMJ Open Diabetes Res Care.* 2021;9(1):e001551.
  31. Wang Y, Huang Y, Wu P, Ye Y, Sun F, Yang X, et al. Plasma lipidomics in early pregnancy and risk of gestational diabetes mellitus: a prospective nested case-control study in Chinese women. *Am J Clin Nutr.* 2021;114(5):1763–73.
  32. Dudzik D, Atanasova V, Barbas C, Bartha JL. First-trimester metabolic profiling of gestational diabetes mellitus: insights into early-onset and late-onset cases compared with healthy controls. *Front Mol Biosci.* 2024;11:1452312.
  33. Saravanan P, Deepa M, Ahmed Z, Ram U, Surapaneni T, Kallur SD, et al. Early pregnancy HbA1c as the first screening test for gestational diabetes: results from three prospective cohorts. *Lancet Diabetes Endocrinol.* 2024;12(8):535–44.
  34. International Association of Diabetes and Pregnancy Study Groups Consensus Panel, Metzger BE, Gabbe SG, Persson B, Buchanan TA, Catalano PA, et al. International association of diabetes and pregnancy study groups recommendations on the diagnosis and classification of hyperglycemia in pregnancy. *Diabetes Care.* 2010;33(3):676–82.
  35. Evans AM, DeHaven CD, Barrett T, Mitchell M, Milgram E. Integrated, nontargeted ultrahigh performance liquid chromatography/electrospray ionization tandem mass spectrometry platform for the identification and relative quantification of the small-molecule complement of biological systems. *Anal Chem.* 2009;81(16):6656–67.
  36. Cajka T, Smilowitz JT, Fiehn O. Validating quantitative untargeted lipidomics across nine liquid chromatography-high-resolution mass spectrometry platforms. *Anal Chem.* 2017;89(22):12360–8.
  37. Zimmer DM, Golichowski AM, Karn CA, Brechtel G, Baron AD, Denne SC. Glucose and amino acid turnover in untreated gestational diabetes. *Diabetes Care.* 1996;19(6):591–6.
  38. Shryack GE, Krause AA, Hernandez Ruano S, Schulz LC, Pennington KA, Rector RS. A murine model of gestational diabetes reveals MASLD risk and alterations in markers of hepatic mitochondrial metabolism. *Front Endocrinol [Internet].* 2025 June 25 [cited 2025 Aug 3];16. <https://www.frontiersin.org/journals/endocrinology/articles/https://doi.org/10.3389/fendo.2025.1498764/full>
  39. Guasch-Ferré M, Hruby A, Toledo E, Clish CB, Martínez-González MA, Salas-Salvadó J, et al. Metabolomics in prediabetes and diabetes: a systematic review and meta-analysis. *Diabetes Care.* 2016;39(5):833–46.
  40. Unnikrishnan R, Anjana RM, Mohan V. Diabetes mellitus and its complications in India. *Nat Rev Endocrinol.* 2016;12(6):357–70.
  41. Duft RG, Castro A, Bonfante ILP, Lopes WA, da Silva LR, Chacon-Mikahil MPT, et al. Serum metabolites associated with increased insulin resistance and low cardiorespiratory fitness in overweight adolescents. *Nutr Metab Cardiovasc Dis NMCD.* 2022;32(1):269–78.
  42. Zhang Z, Zhou Z, Li H. The role of lipid dysregulation in gestational diabetes mellitus: early prediction and postpartum prognosis. *J Diabetes Investig.* 2023;15(1):15–25.
  43. McElwain C, McCarthy CM. Investigating mitochondrial dysfunction in gestational diabetes mellitus and elucidating if BMI is a causative mediator. *Eur J Obstet Gynecol Reprod Biol.* 2020;251:60–5.
  44. Sharma K, Karl B, Mathew AV, Gangoiti JA, Wassel CL, Saito R, et al. Metabolomics reveals signature of mitochondrial dysfunction in diabetic kidney disease. *J Am Soc Nephrol JASN.* 2013;24(11):1901–12.
  45. Cole LK, Vance JE, Vance DE. Phosphatidylcholine biosynthesis and lipoprotein metabolism. *Biochim Biophys Acta.* 2012;1821(5):754–61.
  46. Kleiboecker B, Lodhi J. Peroxisomal regulation of energy homeostasis: effect on obesity and related metabolic disorders. *Mol Metab.* 2022;65:101577.
  47. Shi X, Yao D, Gosnell BA, Chen C. Lipidomic profiling reveals protective function of fatty acid oxidation in cocaine-induced hepatotoxicity. *J Lipid Res.* 2012;53(11):2318–30.
  48. Rong X, Albert CJ, Hong C, Duerr MA, Chamberlain BT, Tarling EJ, et al. LXR $\alpha$ s regulate ER stress and inflammation through dynamic modulation of membrane phospholipid composition. *Cell Metab.* 2013;18(5):685–97.
  49. Holm LJ, Buschard K. L-serine: a neglected amino acid with a potential therapeutic role in diabetes. *APMIS Acta Pathol Microbiol Immunol Scand.* 2019;127(10):655–9.
  50. Decker ST, Funai K. Mitochondrial membrane lipids in the regulation of bioenergetic flux. *Cell Metab* 2024;36(9):1963–78.
  51. Calzada E, Onguka O, Claypool SM. Phosphatidylethanolamine metabolism in health and disease. *Int Rev Cell Mol Biol.* 2016;321:29–88.
  52. Maurício T, Guerra IMS, Pinho M, Melo T, Bonciarelli S, Goracci L, et al. Phosphatidylethanolamine species with n-3 and n-6 fatty acids modulate macrophage lipidome and attenuate responses to LPS stimulation. *Biochim Biophys Acta Mol Cell Biol Lipids.* 2025;1870(4):159614.
  53. Rong J, He T, Zhang J, Bai Z, Shi B. Serum lipidomics reveals phosphatidylethanolamine and phosphatidylcholine disorders in patients with myocardial infarction and post-myocardial infarction-heart failure. *Lipids Health Dis.* 2023;22(1):66.
  54. Barranco-Altirriba M, Alonso N, Weber RJM, Lloyd GR, Hernandez M, Yanes O, et al. Lipidome characterisation and sex-specific differences in type 1 and type 2 diabetes mellitus. *Cardiovasc Diabetol.* 2024;23:109.

55. Chang W, Hatch GM, Wang Y, Yu F, Wang M. The relationship between phospholipids and insulin resistance: from clinical to experimental studies. *J Cell Mol Med*. 2019;23(2):702–10.
56. Yun H, Sun L, Wu Q, Zong G, Qi Q, Li H, et al. Associations among circulating sphingolipids,  $\beta$ -cell function, and risk of developing type 2 diabetes: a population-based cohort study in China. *PLoS Med*. 2020;17(12):e1003451.
57. Wittenbecher C, Guasch-Ferré M, Haslam DE, Dennis C, Li J, Bhupathiraju SN, et al. Changes in metabolomics profiles over ten years and subsequent risk of developing type 2 diabetes: results from the nurses' health study. *EBioMedicine*. 2022;75:103799.
58. Mitsutake S, Zama K, Yokota H, Yoshida T, Tanaka M, Mitsui M, et al. Dynamic modification of sphingomyelin in lipid microdomains controls development of obesity, fatty liver, and type 2 diabetes. *J Biol Chem*. 2011;286(32):28544–55.
59. Dong Q, Sidra S, Gieger C, Wang-Sattler R, Rathmann W, Prehn C, et al. Metabolic signatures elucidate the effect of body mass index on type 2 diabetes. *Metabolites*. 2023;13(2):227.
60. Sharma S, Dong Q, Haid M, Adam J, Bizzotto R, Fernandez-Tajes JJ, et al. Role of human plasma metabolites in prediabetes and type 2 diabetes from the IMI-DIRECT study. *Diabetologia*. 2024;67(12):2804–18.
61. Enthoven LF, Shi Y, Fay E, Kim A, Moreni S, Mao J et al. Effects of pregnancy on plasma sphingolipids using a metabolomic and quantitative analysis approach. *Metabolites*. 2023;13(9):1026.
62. Kay JG, Fairn GD. Distribution, dynamics and functional roles of phosphatidylserine within the cell. *Cell Commun Signal CCS*. 2019;17(1):126.
63. Blaise BJ, Correia G, Tin A, Young JH, Vergnaud AC, Lewis M, et al. Power analysis and sample size determination in metabolic phenotyping. *Anal Chem*. 2016;88(10):5179–88.

## Publisher's note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.