

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30

Machine learning for the prediction of spontaneous preterm birth using early second and third trimester maternal blood gene expression: A Cautionary Tale

Kylie K Hornaday*^{1¶}, Ty Werbicki^{1¶}, Suzanne C Tough^{2,3}, Stephen L Wood⁴, David W Anderson^{5#}, Donna M Slater^{1,4}

¹Department of Physiology and Pharmacology, Cumming School of Medicine, University of Calgary, Calgary, Alberta, Canada

²Department of Community of Health Sciences, Cumming School of Medicine, University of Calgary, Calgary, Alberta, Canada

³Department of Pediatrics, Cumming School of Medicine, University of Calgary, Calgary, Alberta, Canada

⁴Department of Obstetrics and Gynaecology, Cumming School of Medicine, University of Calgary, Calgary, Alberta, Canada

⁵Department of Biochemistry and Molecular Biology, Cumming School of Medicine, University of Calgary, Alberta, Canada

[#]Current address: Department of Science, Langara College, Vancouver, British Columbia, Canada

*Corresponding author

E-mail: kylie.hornaday@ucalgary.ca (KKH)

[¶]These authors contributed equally to this work

31 **Abstract**

32 Preterm birth (PTB) remains a significant global health challenge and a leading cause of
33 neonatal mortality and morbidity. Despite advancements in neonatal care, the prediction of PTB
34 remains elusive, in part due to complex etiologies and heterogeneous patient populations. This
35 study aimed to validate and extend information on gene expression biomarkers previously
36 described for predicting spontaneous PTB (sPTB) using maternal whole blood from the All Our
37 Families pregnancy cohort study based in Calgary, Canada. The results of this study are two-
38 fold: first, using additional replicates of maternal blood samples from the All Our Families
39 cohort, we were unable to repeat the findings of a 2016 study which identified top maternal gene
40 expression predictors for sPTB. Second, we conducted a secondary analysis of the original gene
41 expression dataset from the 2016 study, including external validation using a pregnancy cohort
42 based in Detroit, USA. While initial results of our machine learning model suggested promising
43 performance (area under the receiver operating curve, AUC 0.90 in the training set), performance
44 was significantly degraded on the test set (AUC 0.54), and further degraded in external
45 validation (AUC 0.51), suggesting poor generalizability, likely due to overfitting exacerbated by
46 a low feature-to-noise ratio. Prediction was not improved when using machine learning
47 approaches over traditional statistical learning. These findings underscore the challenges in
48 translating biomarker discovery into clinically useful predictive models for sPTB. This study
49 highlights the critical need for rigorous methodological safeguards and external validation in
50 biomarker research. It also emphasizes the impact of data noise and overfitting on model
51 performance, particularly in high-dimensional omics datasets. Future research should prioritize

52 robust validation strategies and explore mechanistic insights to improve our understanding and
53 prediction of PTB.

54

55

56 **Introduction**

57 Preterm birth, defined as delivery of a live infant prior to 37 weeks of gestation, occurs in
58 13.4 million births worldwide, and is a significant contributor to mortality and morbidity in
59 neonates and children under five [1, 2]. While approximately one third of preterm births occur
60 following known maternal or fetal indications, the remaining two thirds occur following
61 spontaneous onset of labour and/or premature rupture of the fetal membranes (sPTB) without
62 known indication, making prediction and subsequent clinical management of those at risk
63 challenging [2]. Child mortality related to PTB complications has declined since 2000, in part
64 due to advancements in treatments for neonatal complications of prematurity such as respiratory
65 distress syndrome (RDS). However, an estimated 900,000 PTB-associated deaths of children
66 under five still occurred in 2019 worldwide [3].

67 As one of the great obstetrical syndromes, the ability to predict sPTB may be key to
68 improving outcomes. Indeed, considerable efforts have aimed to identify predictive biomarkers
69 of sPTB, however none so far have emerged to have clinical utility, possibly due to
70 heterogeneity within both patient populations and preterm birth phenotypes, as well as risk of
71 bias within study design [4-7]. Methodological safeguarding and appropriate validation of
72 models is important to determine the feasibility, repeatability, robustness, and generalizability of
73 prediction [8-11]. Best practices for prediction modelling are well defined in the literature [12,

74 13], and primary research articles reporting external validation of prediction models have been
75 increasingly published over the last five years [14-17]. However, studies externally validating
76 prediction models in the reproductive field are limited [10, 18, 19]. Additionally, strategies to
77 qualify and understand heterogeneity and regular updating and assessment of predictive models
78 is important to ensure best practice and clinical utility for prediction [20]. Thus, we sought herein
79 to determine the predictive relationship between gene expression biomarkers and spontaneous
80 preterm birth, and to repeat and validate previous findings on prediction of sPTB.

81 Gene expression biomarkers have been identified in maternal whole blood for the
82 prediction of sPTB, which presents a promising avenue for minimally invasive prediction as
83 peripheral blood can reflect global and uterine physiological and immunological changes during
84 pregnancy [21]. One example includes eight genes, *LOC100128908*, *MIR3691*, *LOC101927441*,
85 *CST13P*, *ACAP2*, *ZNF324*, *SH3PXD2B*, *TBX21* that were identified as significantly predictive of
86 sPTB (65% sensitivity and 88% specificity after adjusting for history of abortion and anaemia) in
87 a stepwise logistic regression model [22]. These gene expression biomarkers were identified
88 using an Affymetrix chip microarray analysis of maternal whole blood from the All Our Families
89 pregnancy cohort based in Calgary, Canada [22]. The All Our Families pregnancy cohort
90 presents a rare opportunity for testing experimental repeatability, as maternal blood samples
91 were collected and stored in four separate PAXgene RNA tubes, two were used for the original
92 study (22), and one for validating RNA quality and integrity [23], leaving a remaining fourth
93 sample for experimental validation. The study herein sought to use the additional PAXgene tube
94 to repeat and validate this predictive model to test feasibility for clinical use.

95 Further, though the original study, which used a logistic regression based model,
96 presented promising predictive performance, we hypothesized that machine learning approaches

97 could improve predictive performance. Machine learning and other complex data analysis
98 methods are particularly well suited for mining high dimensional datasets, such as transcriptomic
99 datasets, as they do not generally require the data to adhere to any *a priori* assumptions [24].
100 Machine learning allows for the identification of non-obvious, interactive, complex, and/or non-
101 linear patterns which can go undetected when using traditional statistical linear models [24, 25].
102 These patterns can be leveraged both toward outcome prediction (something highly valuable for
103 complex medical conditions such as preterm birth) and characterizing underlying disease
104 mechanisms [26, 27]. This is particularly enticing, as the underlying causes of sPTB remain
105 poorly understood. The authors have identified an external pregnancy cohort based in Detroit,
106 USA for external validation of both regression- and machine learning-based prediction of sPTB
107 to determine the generalizability of prediction. Prediction algorithms that match too closely to
108 the training data, in other words, suffer from overfitting, are not generalizable to other
109 populations, which is one of the major limitations of machine learning and other methods for
110 prediction. This problem is exacerbated by small or non-representative training sets, where
111 patterns identified may not be meaningfully associated with the outcome, or “noise” and thus the
112 prediction does not translate effectively beyond the original training observations. This stresses
113 the importance of external validation in order to identify robust, generalizable models to
114 meaningfully push forward the prediction of preterm birth.

115 The overarching aim is to explore the repeatability, generalization, and robustness of a
116 prediction model for sPTB using maternal blood gene expression biomarkers. Specific aims are
117 as follows:

- 118 1. To test the predictive utility of *LOC100128908*, *MIR3691*, *LOC101927441*, *CST13P*,
119 *ACAP2*, *ZNF324*, *SH3PXD2B*, *TBX21* expression in maternal blood as biomarkers of
120 spontaneous preterm birth.
- 121 2. To externally validate a prediction model for spontaneous preterm birth using maternal
122 blood gene expression data with machine and statistical learning.

123

124

125 **Methods**

126 **Biological samples and validation of top biomarkers**

127 To test the reproducibility of top biomarkers identified in the literature, historical
128 biological samples were collected from the All Our Families cohort [28-30]. In brief, participants
129 were recruited between May 1st, 2008 and December 31st, 2011 at <25 weeks gestation and
130 provided consent for blood sample collection, and complete questionnaires including information
131 related to demographics, emotional and physical health. Participants provided informed written
132 consent at the time of recruitment from healthcare offices, community and through Calgary
133 Laboratory services and were provided copies of their consent forms for their records. This study
134 was approved by the Conjoint Health Research Ethics Board at the University of Calgary
135 #REB15-0248 Predicting Preterm Birth Study. Biological samples were collected at two points
136 in pregnancy, timepoint 1 (T1) at 17-23 weeks gestation and timepoint 2 (T2) 28-32 weeks
137 gestation. Maternal whole blood was collected directly into four separate PAXgene blood RNA
138 tubes which were then stored at -80°C prior to RNA isolation (PAXgene Blood RNA Kit,
139 Qiagen). Samples were de-identified prior to data collection. Two tubes were previously used for
140 the original prediction modelling and biomarker identification by Heng et al., [22], a third tube

141 was used to assess RNA integrity in long term storage [23], and a fourth was collected from
142 storage from August 16th to November 3rd, 2018 for use in the current study. For the current
143 study, n=47 participants who subsequently had an sPTB (<37 weeks) were included (n=44 T1,
144 n=42 T2 samples), in addition to n=45 participants who had a healthy term (38-42 weeks)
145 delivery (n=40 T1, n=44 T2). A total of n=13 samples were missing from storage or insufficient
146 sample remaining (n=3 T1 sPTB, n=3 T2 sPTB, n=5 T1 term, n=2 T2 term), and n=2 T2 samples
147 in the sPTB group were not included as delivery occurred prior to the second sample collection.
148 Maternal blood samples were collected, and RNA was isolated according to manufacturer's
149 instructions (RNAeasy minikit, Qiagen). The following genes were measured using a probe-
150 based assay (Quantigene, Invitrogen, ThermoFisher Scientific), which uses identical probes to an
151 Affymetrix microarray chip: *LOC100128908 (LMLN2)*, *LOC101927441*, *CST13P*, *ACAP2*,
152 *ZNF324*, *SH3PXD2B*, *TBX21*. Due to limitations with measuring microRNA (miRNA) using the
153 assay, *MIR3691* was not measured.

154

155

156 **Novel prediction model**

157 **Population and expression dataset**

158 To test whether machine learning could improve predictive performance, a secondary
159 analysis of the maternal blood microarray data, as previously published [22], was conducted.
160 Gene expression data was downloaded as raw Affymetrix Chip output files from the National
161 Center for Biotechnology Information Gene Expression Omnibus (accession number:
162 GSE59491) (All Our Families, AOF- Calgary cohort) [22]. The dataset used herein contains
163 high-throughput expression data from n=165 subjects (n=51 sPTB, n=114 matched term delivery

164 controls) nested within the Calgary AOF cohort. Matched gene expression data from two
165 timepoints, 17-23 weeks (T1) and 28-33 weeks (T2) was available for each participant. Two
166 observations from the sPTB group were removed from the dataset as these deliveries occurred
167 prior to the T2 collection and therefore have only one expression dataset. An external dataset was
168 additionally identified from a pregnancy cohort based in Detroit, USA [31], which collected
169 maternal blood samples at comparable timepoints for gene expression analysis, (accession
170 number: GSE149440), and was used for external validation of the model. Participants that had at
171 least two matched blood samples collected within the same two timeframes (T1 and T2) were
172 selected from within the Detroit cohort dataset for analysis, for a total of n=98 subjects (n=34
173 sPTB and n=64 matched term delivery controls) included for external validation.

174

175 **Differential expression analysis**

176 The Calgary dataset was randomly split into 80:20 training and test sets, and the differential
177 expression analysis was conducted on the training set, with 2 times 5-fold cross validation.

178 Differential expression analysis was performed initially as described previously [22]. In brief,
179 differential expression was explored using the following comparisons:

- 180 • sPTB group compared to term group at T1
- 181 • sPTB group compared to term group at T2
- 182 • T1 compared to T2 in sPTB group
- 183 • T1 compared to T2 in term group
- 184 • dT (T2-T1) in sPTB compared to term

185 Genes with a family-wise error rate less than 0.05 were considered differentially expressed and
186 kept for downstream modelling.

187

188 **Feature selection analysis**

189 Three features from each differentially expressed gene were fed into the downstream
190 modelling pipeline: the log₂ of its intensity value at T1 (T1), the log₂ of its intensity value at T2
191 (T2), and the difference between these two measurements (T2-T1, or dT). Input data for feature
192 selection included all genetic features and their associated target labels (sPTB or term group) for
193 the training set. Feature selection was conducted using a supervised stepwise feature selection
194 approach using two times five-fold cross validation. In brief, this stepwise additive approach to
195 feature selection iteratively includes each feature and retains only those features that significantly
196 improve the training model. The feature selection algorithm assigns a gene score for each feature
197 as a measure of relative importance, and subsequently discards non-explanatory or noisy
198 features.

199

200 **Model training and testing**

201 Each training set was fitted using two learning algorithms: logistic regression (LR) and
202 multilayer perceptron artificial neural network (MLP). Hyperparameters for MLP model training
203 were selected using the *Hyperopt* package, and models were evaluated using two repeats of five-
204 fold stratified cross validation. The resultant two models were assessed for predictive
205 performance by fitting them on the internal (Calgary) test set, or the external (Detroit) dataset.
206 Full details on the computational methods are described elsewhere [32].

207

208 **Results**

209 Validation of biomarkers of preterm birth

210 Demographic characteristics of the population used for biomarker validation are
211 described in Table 1. Participants with an sPTB did not significantly differ from the term group
212 in age, smoking status, alcohol use during pregnancy, history of abortion, history of PTB,
213 gravidity or parity. Of the biomarkers measured, only five of seven were detectable in the study
214 population (Table 2, S2 Table). Samples were tested at four concentrations (1.875, 3.75, 6.25,
215 25ng/uL RNA standards). Biomarkers *CST13P* and *LMLN2* were below the limit of detection
216 (<LOD) in over 50% of the population (68% and 56% respectively) at all concentrations and
217 thus were excluded from further analysis. One sample (term T2) was <LOD across all
218 biomarkers, which was likely a technical issue with sample processing and thus excluded. Levels
219 of *SH3PXD2B* were <LOD in 22% of the population and those <LOD were assigned as one half
220 of the basement level (3 MFI, mean fluorescence index units). The remaining four biomarkers
221 were present above the limit of detection in all samples.

222 **Table 1. Patient demographic and clinical characteristics**

	sPTB (n=47)	Term (n=45)
Maternal age (years)	31.5[30.2-32.8]	31.8[30.6-33.0]
Maternal ethnicity		
White	37, 79[64-89]%	35, 78[63-89]%
Non-white	10, 21[11-36]%	10, 22[11-37]%
Smoking during pregnancy		
Yes	5, 11[4-23]%	10, 22[11-37]%
No	36, 77[62-88]%	34, 76[60-87]%
Unknown	6, 13[5-26]%	1, 2[0.06-12]%
Alcohol during pregnancy		
Yes	17, 36[23-52]%	24, 53[38-68]%
No	22, 47[32-62]%	18, 40[26-56]%
Unknown	8, 17[8-31]%	3, 7[1-18]%
History of abortion		

Yes	4, 9[2-20]%	8, 18[8-32]%
No	43, 91[80-98]%	37, 82[68-92]%
History of PTB		
Yes	8, 17[8-31]%	2, 4[0.5-15]%
No	39, 83[69-92]%	43, 96[85-99]%
Gravidity	1.9[1.6-2.2]	1.9[1.7-2.2]
Parity	0.6[0.3-0.8]	0.5[0.4-0.7]
Gestational age at delivery (weeks)	33.9[33.2-34.6]	39.2[38.9-39.4]

223 Values are represented as mean[95% confidence interval] for continuous variables or n, %[95%
224 confidence interval] for categorical variables.

225

226

227 **Table 2. Biomarker levels in maternal blood**

	<i>ACAP2</i>	<i>LOC101927441</i>	<i>ZNF324</i>	<i>SH3PXD2B</i>	<i>TBX21</i>	
TERM	T1 (n=40)	11649[10656-12641]	546[479-613]	572[513-631]	27[18-36]	1128[1002-1254]
	T2 (n=43)	10711[9727-11695]	465[397-533]	484[432-536]	24[17-31]	989[852-1126]
SPTL	T1 (n=44)	9868[8744-10629]	422[364-480]	444[394-494]	21[17-25]	895[757-1033]
	T2 (n=42)	11827[10049-13605]	471[388-554]	515[443-587]	39[26-52]	1042[864-1220]

228 Values are represented as mean [95% confidence interval] of mean fluorescence index MFI.

229

230

231 Four of the five measured biomarkers, *ACAP2* (p=0.0068), *LOC101927441* (p=0.0082),

232 *ZNF324* (p=0.0019), and *TBX21* (p=0.0182) exhibited significantly lower levels in the sPTL

233 group compared to the term group at T1, and not at T2. When assessing biomarkers as a

234 measurement of T2/T1 ratios, *ACAP2* (p=0.0074), *LOC101927441* (p=0.0273), *ZNF324*

235 (p=0.0170), *TBX21* (p=0.0119) ratios were significantly higher in the sPTL group than the term

236 group, suggesting a greater trajectory of increased expression through gestation in those with

237 sPTL (Fig 1). Though we do observe some differences in biomarker levels between sPTL and

238 term samples, only five of the eight originally identified biomarkers [22] could be measured
239 using the same population and methodology, and only four exhibited significant differences
240 between term and preterm groups.

241

242

243 **Figure 1. Biomarkers of preterm birth.** Values are reported as mean fluorescence index (MFI)
244 at either timepoint or a ratio of MFI values at T2 over T1. Analysed by one-way ANOVA
245 followed by Dunnett correction for multiple comparisons. *p-value<0.05, **p-value<0.01.
246 *ACAP2*: ArfGAP with coiled-coil, ankyrin repeat and PH domains 2, *ZNF324*: zinc finger
247 protein 324. *SH3PCD2B*: SH3 and PX domains 2B, *TBX21*: T-box transcription factor 21.

248

249 **Novel prediction model: Feature selection**

250 In the interest of identifying other potential biomarkers that can robustly predict sPTB,
251 we conducted feature selection analysis on the publicly available AOF microarray dataset
252 (GSE59491) [22]. The top gene features selected from the complete microarray dataset, along
253 with their assigned gene scores for each iteration of cross validation (two times five-fold cross
254 validation for a total of ten iterations), are represented in Table 3. Notably, the top predictive
255 genes did not show consistency in assigned gene scores across iterations. The topmost
256 explanatory feature as selected by the feature selection algorithm, *FPR3* dT, was assigned a score
257 of 1 (#1 most predictive) and 2 (#2 most predictive) but was assigned a score of zero
258 (uninformative) the remaining eight iterations, indicating that top features are not robust to noise
259 within the dataset.

260 **Table 3. Top overall explanatory features and assigned feature selection scores by iteration**

<i>Feature</i>	<i>Iteration:</i>	<i>1</i>	<i>2</i>	<i>3</i>	<i>4</i>	<i>5</i>	<i>6</i>	<i>7</i>	<i>8</i>	<i>9</i>	<i>10</i>
<i>FPR3</i> dT		0	0	0	0	0	1	0	2	0	0
Formyl peptide receptor 3											
<i>TRBJ2-6</i> T1		0	0	0	1	0	0	0	0	0	2
T cell receptor beta joining 2-6											
<i>PLD4</i> T2		0	0	0	12	0	0	4	0	0	0
phospholipase D family member 4											
<i>ZFP14</i> T2		0	0	0	0	1	0	0	0	0	0
zinc finger protein											

261 Rows represent top four overall explanatory genetic features selected. Columns represent
262 individual analysis (2 times 5-times cross validation for a total of 10 iterations). Values represent
263 predictive score assigned by the feature selection algorithm where 1 is the highest possible value,
264 indicating the most predictive genetic feature and a score of 0 indicates that the feature was not
265 considered predictive for that given iteration.
266

267 **Model performance for prediction of sPTB**

268 The MLP model showed promising performance in the training set (area under the
269 receiver operating curve, AUC 0.9), and an improvement over traditional LR (AUC 0.85),
270 though performance was notably degraded when applied to the internal test set (AUC 0.54), and
271 further degraded when validated externally (AUC 0.51 MLP, AUC 0.53 LR), which indicates a
272 high degree of overfitting (Table 4).

273 **Table 4. Model Performance**

<i>Model</i>	<i>Training Set</i>	<i>Test Set</i>	<i>Training AUC</i>	<i>Test AUC</i>
MLP	Calgary	Calgary	0.90	0.54
LR	Calgary	Calgary	0.85	0.54

MLP	Calgary	Detroit	0.74	0.51
LR	Calgary	Detroit	0.77	0.53

274 Model performance as reported by area under the receiver operating curve (AUC) for two
275 prediction algorithms, multilayer perceptron (MLP) and logistic regression (LR) validated
276 internally and externally.
277

278

279 **Assessing overfitting**

280 Both LR and MLP models showed significant degradation of performance in both
281 internal and external test sets as compared to training performance, indicating a high degree of
282 overfitting during training, particularly in the MLP model. To test the degree of overfitting,
283 models were retrained using permuted data. In brief, target labels (sPTL or term) were scrambled
284 to remove any potential true pattern within the data before proceeding with model training as
285 before. Using scrambled data, high performance was still observed in the training set, with the
286 highest performance by the MLP algorithm (AUC 0.80). Model performance was degraded when
287 applied to either the internal or external test sets (Table 5).

288 **Table 5. Model performance using permuted data**

<i>Model</i>	<i>Training Set</i>	<i>Test Set</i>	<i>Training AUC</i>	<i>Test AUC</i>
MLP	Calgary	Calgary	0.80	0.49
LR	Calgary	Calgary	0.75	0.50
MLP	Calgary	Detroit	0.72	0.52
LR	Calgary	Detroit	0.63	0.52

289 Model performance as reported by area under the receiver operating curve (AUC) for two
290 prediction algorithms, multilayer perceptron (MLP) and logistic regression (LR) tested internally
291 and validated externally.

292

293

294 **Discussion**

295 We were unable to repeat the findings of Heng et al., [22] to predict spontaneous preterm
296 birth using maternal blood gene expression. Most alarmingly, two of the eight topmost predictive
297 genes were not detectable in blood samples from the same patients, suggesting issues with
298 repeatability in probe-based RNA array methods, despite validation of RNA integrity over long-
299 term storage [23]. Indeed, array reliability may be particularly problematic in lowly expressed
300 genes and certain genes may be more subject to poor probe specificity [33, 34], and we were
301 unable to conduct assessment of gene-specific expression levels over long-term storage.
302 Additionally, we were unable to produce a more generalizable model through secondary analysis
303 of the microarray data using machine learning, and our results suggest a high degree of
304 overfitting following external validation. This highlights the importance of repeat and validation
305 studies in order to meaningfully progress the field of preterm birth prediction.

306 One of the primary limitations to prediction using maternal blood gene expression
307 encountered was overfitting and noise within the dataset, which significantly skewed
308 performance estimates. A noisy dataset likely also contributed to the inability to repeat and/or
309 validate previous findings. Possible consequences of data noise are further exacerbated when
310 using advanced methods such as machine learning, and, as evidenced in the study herein, high
311 complexity/machine learning approaches often do not demonstrate improved predictive

312 performance over traditional statistical methods [35]. Feature selection approaches were unable
313 to effectively reduce the noise within this dataset to obtain clinically useful patterns as markers
314 for spontaneous preterm birth. Many prediction studies, at least in the field of reproduction, are
315 preceded by observational experiments to explore biomarker patterns as possible predictors, such
316 as differential gene expression analysis to identify genes associated with an outcome. Often,
317 these observational experiments are conducted on the whole dataset, not the training set, and as
318 such, prediction models trained on this data are biased by patterns that exist in the test set. This
319 phenomenon, in which information from outside the training dataset is used to create the
320 prediction model, is known as data leakage. Consequently, the training dataset contains
321 information about the outcome that would not be otherwise available when using the model for
322 prediction, artificially overinflating the predictive performance when the model is applied to the
323 test set. Unintentional data leakage likely contributes to the lack of reproducibility in such
324 prediction studies. High levels of noise inherent in gene expression data exacerbated the
325 consequences of data leakage during differential expression analysis and feature selection,
326 highlighting the importance of cautious interpretation and robust methodological safeguards.

327 To illustrate the extent to which data leakage may impact performance results, we also
328 performed model training in the training set in which differential expression analysis was
329 conducted *before* the splitting the training and test sets (S1 Table). While we observe high
330 predictive performance in the training set (AUC 0.72 with LR, 0.79 with MLP), performance
331 was not significantly degraded in the test set (AUC 0.65 with LR, 0.85 with MLP). Note that
332 with the presence of data leakage, the machine learning MLP model had substantially higher
333 performance (AUC 0.85) as compared to analysis without data leakage (AUC 0.54 for
334 comparable dataset). This stresses the importance of methodological safeguarding and careful

335 study design, to avoid possible sources of bias and data leakage, particularly in omics or similar
336 datasets that are prone to a high degree of noise. External validation is also a highly powerful
337 tool for testing the generalizability of models that may have been subject to data leakage.

338 Our findings also underscore the broader implications for omics studies for discovery
339 analysis, where high feature-to-observation ratios are common, which exacerbates the challenge
340 for mitigating bias and ensuring the reliability of predictive models. For example, differential
341 expression analysis without appropriate training and testing sets for validation introduces
342 inherent bias and limits the generalizability of patterns identified. Testing on internal test sets
343 alone is insufficient for measuring generalizability, especially in the instance of data leakage.
344 Yet, assessments of overfitting and external validation are not standard practice in preterm birth
345 prediction, and the authors stress their importance for meaningful future work in this field. As
346 such, our study serves as a cautionary tale for researchers, emphasizing the need for
347 transparency, rigorous methodological standards, as well as not only repeating results but
348 validation in external cohorts in order to advance the field of spontaneous preterm birth
349 prediction responsibly.

350 While maternal blood presents an enticing opportunity for minimally invasive prediction,
351 peripheral blood is subject to a high noise signal from various physiological processes occurring
352 within the body possibly unrelated to uterine function during pregnancy. This stresses the need
353 for improved feature identification. Biological compartments including cervicovaginal fluid,
354 amniotic fluid, and the vaginal microbiome may better reflect the physiology of pregnancy [36,
355 37] though sample availability of reproductive and gestational tissues for research purposes are
356 limited. An emerging strategy involves the use of cell-free nucleic acid biomarkers, which can be
357 utilized to identify biomarkers with uterine origin in maternal blood for improved prediction of

358 adverse pregnancy outcomes [38, 39]. Considerable research has been conducted to review the
359 most robust predictors for sPTB, including but not limited to inflammatory biomarkers, maternal
360 characteristics and genetic contributions [40-42], yet the most frequently used risk factors in
361 current literature show variable predictive performance and poor robustness [43]. A recent meta-
362 analysis identified the most robust predictors of PTB, including low gestational weight gain,
363 interpregnancy interval following miscarriage <6months, and sleep-disordered breathing [42],
364 and it is likely that combined biomarker approaches are necessary for prediction [44, 45].
365 Additionally, current literature often does not distinguish those predictors for general PTB from
366 those for sPTB, despite likely distinct aetiologies. It is also worth noting that the pervasive use of
367 convenience sampling in reproductive studies (e.g. secondary analysis of biosamples used for
368 routine antenatal screening) are not necessarily performed proximal to the outcome of interest
369 (PTB). For many subjects, the delay from testing to outcome may make identifying true
370 associations difficult.

371 Looking ahead, our recommendations for future research include safeguarding against
372 sources of data leakage, implementing cross-validation techniques as a measure of robustness,
373 and prioritizing repeatability and reproducibility of findings. This likely includes incentivizing
374 repeated studies in published literature and improving data management, storage, and sharing
375 infrastructure [10, 11]. Additionally, as unsupervised feature selection techniques were not
376 shown to be beneficial in improving prediction of spontaneous preterm birth, future research in
377 identifying biomarkers for the mechanism preterm labour are important. The best models
378 combine an understanding of the features (such as genes, proteins, or patient characteristics) that
379 are most important for determining the outcome and robust methodologies. In the case of
380 spontaneous preterm birth, this should involve a return to the bench to better elucidate those

381 pathways and biomarkers and their possible contribution to preterm birth outcomes. In better
382 understanding the mechanisms of labour and preterm birth, we stand to better approach its
383 prediction, and consequently, improving maternal and neonatal health outcomes for those
384 impacted by preterm birth.

385

386

387 **References**

- 388 1. Ohuma EO, Moller AB, Bradley E, Chakwera S, Hussain-Alkhateeb L, Lewin A, et al. National,
389 regional, and global estimates of preterm birth in 2020, with trends from 2010: a systematic analysis.
390 Lancet. 2023;402(10409):1261-71.
- 391 2. Purisch SE, Gyamfi-Bannerman C. Epidemiology of preterm birth. Seminars in Perinatology.
392 2017;41(7):387-91.
- 393 3. Perin J, Mulick A, Yeung D, Villavicencio F, Lopez G, Strong KL, et al. Global, regional, and
394 national causes of under-5 mortality in 2000-19: an updated systematic analysis with implications for the
395 Sustainable Development Goals. Lancet Child Adolesc Health. 2022;6(2):106-15.
- 396 4. Hornaday KK, Wood EM, Slater DM. Is there a maternal blood biomarker that can predict
397 spontaneous preterm birth prior to labour onset? A systematic review. PLOS ONE. 2022;17(4):e0265853.
- 398 5. Marić I, Stevenson DK, Aghaepour N, Gaudillière B, Wong RJ, Angst MS. Predicting Preterm
399 Birth Using Proteomics. Clin Perinatol. 2024;51(2):391-409.
- 400 6. Ramachandran A, Clotney KD, Gordon A, Hyett JA. Prediction and prevention of preterm birth:
401 Quality assessment and systematic review of clinical practice guidelines using the AGREE II framework.
402 Int J Gynaecol Obstet. 2024.

- 403 7. Yang Q, Fan X, Cao X, Hao W, Lu J, Wei J, et al. Reporting and risk of bias of prediction models
404 based on machine learning methods in preterm birth: A systematic review. *Acta Obstet Gynecol Scand.*
405 2023;102(1):7-14.
- 406 8. Staffa SJ, Zurakowski D. Statistical Development and Validation of Clinical Prediction Models.
407 *Anesthesiology.* 2021;135(3):396-405.
- 408 9. Steyerberg EW, Harrell FE, Jr. Prediction models need appropriate internal, internal-external, and
409 external validation. *J Clin Epidemiol.* 2016;69:245-7.
- 410 10. Sharifi-Heris Z, Laitala J, Airola A, Rahmani AM, Bender M. Machine Learning Approach for
411 Preterm Birth Prediction Using Health Records: Systematic Review. *JMIR Med Inform.*
412 2022;10(4):e33875.
- 413 11. Mennickent D, Rodríguez A, Opazo MC, Riedel CA, Castro E, Eriz-Salinas A, et al. Machine
414 learning applied in maternal and fetal health: a narrative review focused on pregnancy diseases and
415 complications. *Front Endocrinol (Lausanne).* 2023;14:1130139.
- 416 12. Leisman DE, Harhay MO, Lederer DJ, Abramson M, Adjei AA, Bakker J, et al. Development
417 and Reporting of Prediction Models: Guidance for Authors From Editors of Respiratory, Sleep, and
418 Critical Care Journals. *Crit Care Med.* 2020;48(5):623-33.
- 419 13. Collins GS, Dhiman P, Ma J, Schlüssel MM, Archer L, Van Calster B, et al. Evaluation of
420 clinical prediction models (part 1): from development to external validation. *Bmj.* 2024;384:e074819.
- 421 14. Lenain R, Dantan E, Giral M, Foucher Y, Asar Ö, Naesens M, et al. External Validation of the
422 DynPG for Kidney Transplant Recipients. *Transplantation.* 2021;105(2):396-403.
- 423 15. Russell FM, Herbert A, Kennedy S, Nti B, Powell M, Davis J, et al. External validation of the
424 ultrasound competency assessment tool. *AEM Educ Train.* 2023;7(3):e10887.
- 425 16. Yun JS, Han K, Choi SY, Cha SA, Ahn YB, Ko SH. External validation and clinical application
426 of the predictive model for severe hypoglycemia. *Front Endocrinol (Lausanne).* 2022;13:1006470.

- 427 17. Sliker RC, van der Heijden A, Siddiqui MK, Langendoen-Gort M, Nijpels G, Herings R, et al.
428 Performance of prediction models for nephropathy in people with type 2 diabetes: systematic review and
429 external validation study. *Bmj*. 2021;374:n2134.
- 430 18. Chaemsaitong P, Sahota DS, Poon LC. First trimester preeclampsia screening and prediction.
431 *Am J Obstet Gynecol*. 2022;226(2s):S1071-S97.e2.
- 432 19. Neary C, Naheed S, McLernon DJ, Black M. Predicting risk of postpartum haemorrhage: a
433 systematic review. *Bjog*. 2021;128(1):46-53.
- 434 20. Van Calster B, Steyerberg EW, Wynants L, van Smeden M. There is no such thing as a validated
435 prediction model. *BMC Medicine*. 2023;21(1):70.
- 436 21. Feyaerts D, Marić I, Arck PC, Prins JR, Gomez-Lopez N, Gaudillière B, et al. Predicting
437 Spontaneous Preterm Birth Using the Immunome. *Clin Perinatol*. 2024;51(2):441-59.
- 438 22. Heng YJ, Pennell CE, McDonald SW, Vinturache AE, Xu J, Lee MWF, et al. Maternal whole
439 blood gene expression at 18 and 28 weeks of gestation associated with spontaneous preterm birth in
440 asymptomatic women. *PLoS ONE*. 2016;11(6).
- 441 23. Stephenson NL, Hornaday KK, Doktorchik CTA, Lyon AW, Tough SC, Slater DM. Quality
442 assessment of RNA in long-term storage: The All Our Families biorepository. *PLoS One*.
443 2020;15(12):e0242404.
- 444 24. Alpaydin E. Introduction to machine learning. Ieee Xplore d, Mit Press p, ebrary I, editors:
445 Cambridge, Massachusetts : MIT Press; Third edition.; 2014.
- 446 25. Theodoridis S. Machine learning : a Bayesian and optimization perspective: Amsterdam,
447 Netherlands : Academic Press; First edition.; 2015.
- 448 26. Dhar V. Data science and prediction. *Commun ACM*. 2013;56(12):64–73.
- 449 27. Arain Z, Iliodromiti S, Slabaugh G, David AL, Chowdhury TT. Machine learning and disease
450 prediction in obstetrics. *Curr Res Physiol*. 2023;6:100099.

- 451 28. McDonald CR, Darling AM, Conroy AL, Tran V, Cabrera A, Liles WC, et al. Inflammatory and
452 angiogenic factors at mid-pregnancy are associated with spontaneous preterm birth in a cohort of
453 Tanzanian women. PLoS ONE. 2015;10(8).
- 454 29. Gracie SK, Lyon AW, Kehler HL, Pennell CE, Dolan SM, McNeil DA, et al. All Our Babies
455 Cohort Study: recruitment of a cohort to predict women at risk of preterm birth through the examination
456 of gene expression profiles and the environment. BioMed Central;[http://www.biomedcentral.com/1471-](http://www.biomedcentral.com/1471-2393/10/87)
457 [2393/10/87](http://www.biomedcentral.com/1471-2393/10/87);University of Calgary;Medicine; 2011.
- 458 30. Tough SC, McDonald SW, Collisson BA, Graham SA, Kehler H, Kingston D, et al. Cohort
459 Profile: The All Our Babies pregnancy cohort (AOB). International Journal of Epidemiology.
460 2017;46(5):1389-90k.
- 461 31. Tarca AL, Pataki BÁ, Romero R, Sirota M, Guan Y, Kutum R, et al. Crowdsourcing assessment
462 of maternal blood multi-omics for predicting gestational age and preterm birth. Cell Reports Medicine.
463 2021;2(6):100323-.
- 464 32. Werbicki T. Slater-Lab-SPTB2022. Available from: [https://github.com/tywerbicki/Slater-Lab-](https://github.com/tywerbicki/Slater-Lab-SPTB)
465 [SPTB](https://github.com/tywerbicki/Slater-Lab-SPTB).
- 466 33. Draghici S, Khatri P, Eklund AC, Szallasi Z. Reliability and reproducibility issues in DNA
467 microarray measurements. Trends Genet. 2006;22(2):101-9.
- 468 34. Kothapalli R, Yoder SJ, Mane S, Loughran TP, Jr. Microarray results: how accurate are they?
469 BMC Bioinformatics. 2002;3:22.
- 470 35. Christodoulou E, Ma J, Collins GS, Steyerberg EW, Verbakel JY, Van Calster B. A systematic
471 review shows no performance benefit of machine learning over logistic regression for clinical prediction
472 models. J Clin Epidemiol. 2019;110:12-22.
- 473 36. Chakoory O, Barra V, Rochette E, Blanchon L, Sapin V, Merlin E, et al. DeepMPTB: a vaginal
474 microbiome-based deep neural network as artificial intelligence strategy for efficient preterm birth
475 prediction. Biomarker Research. 2024;12.

- 476 37. Chang Y, Li W, Shen Y, Li S, Chen X. Association between interleukin-6 and preterm birth: a
477 meta-analysis. *Ann Med*. 2023;55(2):2284384.
- 478 38. Cowan AD, Rasmussen M, Jain M, Tribe RM. Predicting Preterm Birth Using Cell-Free
479 Ribonucleic Acid. *Clin Perinatol*. 2024;51(2):379-89.
- 480 39. Moufarrej MN, Vorperian SK, Wong RJ, Campos AA, Quaintance CC, Sit RV, et al. Early
481 prediction of preeclampsia in pregnancy with cell-free RNA. *Nature*. 2022;602(7898):689-94.
- 482 40. Tang ID, Mallia D, Yan Q, Pe'er I, Raja A, Salieb-Aouissi A, et al. A Scoping Review of Preterm
483 Birth Risk Factors. *Am J Perinatol*. 2024;41(S 01):e2804-e17.
- 484 41. Li J, Ge J, Ran N, Zheng C, Fang Y, Fang D, et al. Finding the priority and cluster of
485 inflammatory biomarkers for infectious preterm birth: a systematic review. *J Inflamm (Lond)*.
486 2023;20(1):25.
- 487 42. Mitrogiannis I, Evangelou E, Efthymiou A, Kanavos T, Birbas E, Makrydimas G, et al. Risk
488 factors for preterm birth: an umbrella review of meta-analyses of observational studies. *BMC Med*.
489 2023;21(1):494.
- 490 43. Ferreira A, Bernardes J, Gonçalves H. Risk Scoring Systems for Preterm Birth and Their
491 Performance: A Systematic Review. *J Clin Med*. 2023;12(13).
- 492 44. Mirzaei A, Hiller BC, Stelzer IA, Thiele K, Tan Y, Becker M. Computational Approaches for
493 Connecting Maternal Stress to Preterm Birth. *Clin Perinatol*. 2024;51(2):345-60.
- 494 45. Creswell L, Rolnik DL, Lindow SW, O'Gorman N. Preterm Birth: Screening and Prediction. *Int J*
495 *Womens Health*. 2023;15:1981-97.

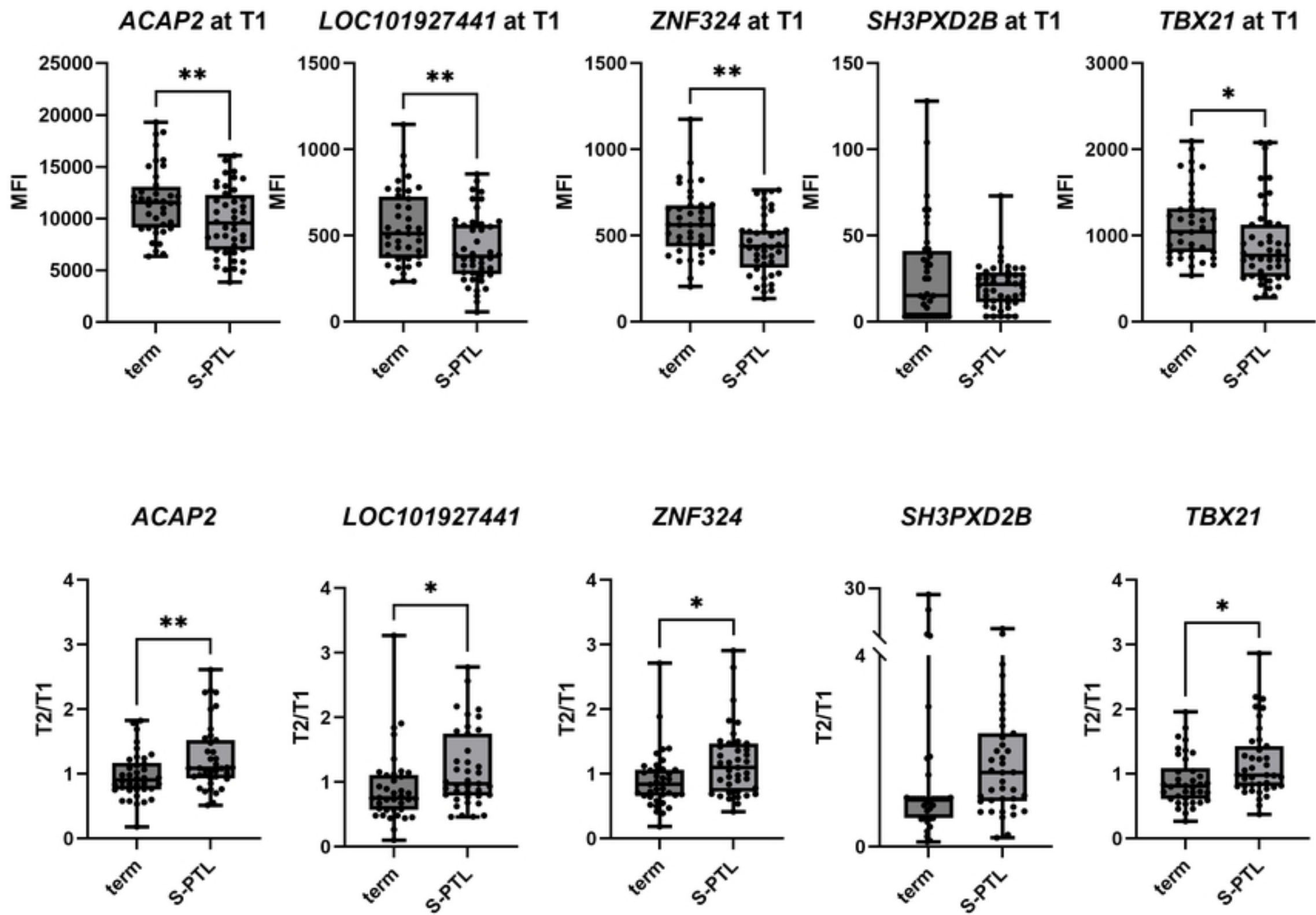
496

497

498 **Supporting Information**

499 **S1 Table. Model performance with data leakage.** LR: logistic regression, MLP: multilayer
500 perceptron, AUC: area under the receiver operating curve. Trained on the Calgary dataset and
501 tested on the Calgary test set.

502 **S2 Table. Raw fluorescence index for predictive genes tested in maternal blood.** Isolated
503 RNA from whole maternal blood was analyzed for gene expression using a QuantiGene Plex
504 custom assay (Qiagen).



Figure