Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Machine Learning-Based Plasma Metabolomic Biomarker Panels for Auxiliary Diagnosis and Prediction of Acute Antipsychotic Response in Drug-Naïve Schizophrenia

Shuo Wang
Institute of Mental Health, Tianjin Anding Hospital, Mental Health Center of Tianjin Medical University
Boxuan Zhou
Institute of Mental Health, Tianjin Anding Hospital, Mental Health Center of Tianjin Medical University
Xinyi Zhang
Institute of Mental Health, Tianjin Anding Hospital, Mental Health Center of Tianjin Medical University
Yueru Guo
Institute of Mental Health, Tianjin Anding Hospital, Mental Health Center of Tianjin Medical University
Tongxin Li
Institute of Mental Health, Tianjin Anding Hospital, Mental Health Center of Tianjin Medical University
Jiajia Gao
State Key Laboratory of Experimental Hematology, Institute of Hematology & Blood Diseases Hospital, Chinese Academy of Medical Sciences & Peking Union Medical College
Yuying Qiu
Institute of Mental Health, Tianjin Anding Hospital, Mental Health Center of Tianjin Medical University
Meijuan Li
Institute of Mental Health, Tianjin Anding Hospital, Mental Health Center of Tianjin Medical University
Yeqing Dong
Institute of Mental Health, Tianjin Anding Hospital, Mental Health Center of Tianjin Medical University
Jie Li
jieli@tmu.edu.cn

Institute of Mental Health, Tianjin Anding Hospital, Mental Health Center of Tianjin Medical University

Research Article

Keywords: Drug-naïve schizophrenia, Targeted metabolomics, Machine learning, Diagnostic model, Treatment response

Posted Date: November 25th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11624833/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations:
No competing interests reported.

Table 1 is available in the Supplementary Files section.

Machine Learning-Based Plasma Metabolomic Biomarker Panels for Auxiliary Diagnosis and Prediction of Acute Antipsychotic Response in Drug-Naïve Schizophrenia

Shuo Wang[1#], Boxuan Zhou[1#], Xinyi Zhang[1#], Yueru Guo[1], Tongxin Li[1], Jiajia Gao[2], Yuying Qiu[1], Meijuan Li[1]*, Yeqing Dong[1]*, Jie Li[1]*

[1] Institute of Mental Health, Tianjin Anding Hospital, Mental Health Center of Tianjin Medical University, Tianjin, 300222, China;

[2] State Key Laboratory of Experimental Hematology, Institute of Hematology & Blood Diseases Hospital, Chinese Academy of Medical Sciences & Peking Union Medical College, Tianjin, 300020, China;

[#]: contributed equally to this work.

*** Correspondence authors:** Prof. Jie Li, E-mail: jieli@tmu.edu.cn; Prof. Yeqing Dong, E-mail: dongyeqing@tmu.edu.cn; Prof. Meijuan Li, E-mail: meijuanli82@163.com

Abstract

Background: Diagnosing drug-naïve schizophrenia (DNS) and predicting acute-phase antipsychotic response lack objective biomarkers. This study integrated plasma targeted metabolomics with machine learning to develop an auxiliary diagnostic model and a predictive model for acute treatment response.

Methods: Plasma targeted metabolomics (Biocrates P500) was performed on 129 DNS patients and 66 HCs, with external validation in an independent cohort (13 DNS, 20 HCs). Differential metabolites were identified by OPLS-DA (VIP > 1, $P < 0.05$), and a diagnostic metabolite panel was derived by integrating three feature selection algorithms (LASSO, RF-RFE, and SVM-RFE). Five machine learning classifiers were compared using repeated nested cross-validation, followed by SHAP interpretation. Additionally, a two-step strategy combining noise-calibrated random forest with univariate logistic regression was applied to 271 metabolites to identify baseline predictors of 8-week treatment response (≥ 50% reduction in PANSS total score) in 120 DNS patients.

Results: Thirty-four metabolites were significantly dysregulated in DNS patients. The intersection of three algorithms yielded a 10-metabolite diagnostic panel, with the random forest (RF) classifier achieving the highest performance in nested cross-validation (AUC = 0.919, 95% CI: 0.880 - 0.958) and retaining good discrimination in the independent external cohort (AUC = 0.863, 95% CI: 0.724 - 1.000). For treatment response, an exploratory analysis identified a distinct panel of 10 baseline metabolites associated with 8-week response, in which higher amino acid and phosphatidylcholine levels and lower sphingolipid levels were associated with a greater likelihood of response; a logistic model based on these metabolites discriminated responders from non-responders with an AUC of 0.767.

Conclusions: Plasma metabolomic biomarker panels show promising performance for the auxiliary diagnosis of DNS and prediction of early antipsychotic treatment response, potentially facilitating objective diagnosis and personalized treatment, pending confirmation in larger, independent cohorts.

Keywords: Drug-naïve schizophrenia; Targeted metabolomics; Machine learning; Diagnostic model; Treatment response

1. Introduction

Schizophrenia is a severe and chronic psychiatric disorder that typically manifests in late adolescence or early adulthood [1], imposing a profound burden on individuals, families, and healthcare systems [2, 3]. The current diagnosis of schizophrenia relies predominantly on subjective clinical interviews and the observation of characteristic symptoms [4], which can be ambiguous, especially in the prodromal or early stages of the illness [5]. This diagnostic challenge is particularly pronounced in drug-naïve schizophrenia (DNS) patients, wherein the absence of prior pharmacological treatment provides a unique window into the core pathophysiology of the disease, unconfounded by medication effects. Furthermore, beyond the initial diagnosis, the prediction of acute-phase antipsychotic response remains highly challenging. Antipsychotic medications exhibit considerable inter-individual variability in efficacy, and approximately 20-30% of patients fail to achieve an adequate response during the critical early weeks of treatment [6, 7]. The inability to accurately predict who will respond to a given treatment necessitates a prolonged trial-and-error approach, delaying optimal therapeutic interventions and potentially exacerbating the illness trajectory [8]. Therefore, there is an urgent clinical need to identify objective, quantifiable biomarkers that can not only facilitate the early auxiliary diagnosis of DNS but also predict acute-phase treatment responses.

Metabolomics offers a powerful lens through which to explore the dynamic biochemical alterations associated with complex diseases [9]. Given the high metabolic demands of the brain and the well-documented metabolic side effects of antipsychotics, there is a growing consensus that metabolic dysregulation plays a pivotal role in the pathophysiology of schizophrenia [10]. Plasma targeted metabolomics, in particular, provides a highly precise and reproducible quantification of specific metabolic pathways [11], making it an ideal candidate for biomarker discovery [10, 12]. However, previous metabolomic studies in schizophrenia have largely been limited by small sample sizes, cross-sectional designs, and a focus on characterizing general alterations rather than constructing clinically actionable predictive models [13, 14]. Moreover, the high dimensionality and multicollinearity inherent in metabolomic data often render

traditional statistical methods inadequate for robust feature selection and prognostic modeling [15]. Machine learning (ML) has emerged as an indispensable tool for handling high-dimensional biological data, capable of capturing complex, non-linear interactions that conventional statistics might miss [16]. By integrating ML with metabolomics, it is possible to distill vast datasets into compact, highly predictive biomarker panels [17]. While a few studies have attempted to apply ML to metabolomic data in psychiatry, they have predominantly focused on diagnostic classification [18, 19]. The prediction of acute-phase treatment response using machine learning–based metabolomic models remains largely unexplored in DNS patients.

To address these gaps, this study aimed to develop both an auxiliary diagnostic model for DNS and a predictive model for acute-phase antipsychotic response by integrating plasma targeted metabolomics with advanced machine learning algorithms. For diagnostic modeling, we employed a rigorous feature selection strategy intersecting three distinct algorithms to identify highly stable core metabolites, followed by systematic evaluation across multiple ML classifiers. For treatment-response analysis, we combined random forest permutation importance with univariate logistic regression (LR) to identify baseline metabolites associated with 8-week antipsychotic response. This study provides a crucial foundation for early precise diagnosis and individualized therapeutic decision-making in schizophrenia.

2. Methods

2.1 Study design and Participants

The overall workflow and detailed analysis steps of this study are illustrated in the research roadmap in Figure 1. This study comprised 129 DNS patients recruited from Tianjin Anding Hospital and 66 healthy controls (HCs) recruited from the surrounding community. The inclusion criteria for DNS patients were as follows: (1) independently diagnosed with schizophrenia or schizophreniform disorder by two experienced psychiatrists according to the Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition (DSM-5); (2) aged 18 to 65 years; (3) completed primary school education or above; (4) no prior use of antipsychotic medications; and (5) a total score of ≥ 60 on the Positive and Negative Syndrome Scale (PANSS). Exclusion criteria included: (1)

presence of other psychiatric disorders; (2) diabetes mellitus, hyperlipidemia, or other severe physical diseases; (3) pregnancy or lactation; (4) a history of physical therapies, such as modified electroconvulsive therapy (MECT) and transcranial magnetic stimulation (TMS), within the past month; (5) history of alcohol abuse or other psychoactive substance use; and (6) any factors preventing participants from providing informed consent or participating in this study. All DNS patients were hospitalized and received a uniform diet. HCs were recruited through community advertisements and demographically matched to the DNS group. The inclusion criteria for HCs were: (1) aged 18 to 65 years; (2) completed primary school education or above; (3) no lifetime history of psychiatric disorders based on DSM-5 criteria; and (4) no prior use of antipsychotic medications.

To further assess model generalizability, an independent validation cohort of 13 DNS patients and 20 HCs, not overlapping with the discovery sample, was recruited at Tianjin Anding Hospital and the surrounding communities, respectively, under the same inclusion and exclusion criteria, and processed using identical plasma collection and targeted metabolomic protocols.

This study complied with the Declaration of Helsinki and received approval from the Ethics Committee of Tianjin Anding Hospital (Approval No. 2017-03, No. 2023-63). Written informed consent to participate was obtained from all participants in both the discovery and independent validation cohorts.

2.2 Clinical Assessment

General demographic and clinical information, including age, sex, educational level, body mass index (BMI), smoking status, and duration of untreated psychosis (DUP), was collected by specially trained researchers using a self-designed questionnaire. For the 129 DNS patients, longitudinal clinical assessments were conducted at baseline and at weeks 8 during their clinician-prescribed antipsychotic treatment. Clinical psychopathological symptoms were evaluated using the PANSS [20]. The PANSS reduction rate was objectively calculated to assess the acute-phase treatment response: PANSS reduction rate = (baseline PANSS score - follow-up PANSS score) / (baseline PANSS score - 30) × 100%. The DNS patients achieving a ≥

50% reduction in the PANSS total score at week 8 were classified as treatment responders and the remainder as non-responders; patients without an evaluable week-8 assessment were excluded from the treatment-response analysis [21].

To ensure rating consistency and reliability, evaluators jointly attended a training session on the standardized use of the PANSS before the study. Throughout the follow-up period, the intra-class correlation coefficient (ICC) for PANSS scoring among raters was consistently maintained at > 0.8.

2.3 Plasma Sample Collection

After an overnight fast, peripheral venous blood was drawn from patients and HCs into EDTA tubes. Samples were processed within four hours of collection. Plasma was isolated by centrifugation at 1,000 rpm for 15 min at 4 °C, divided into aliquots, and stored at -80 °C until analysis.

2.4 Targeted Metabolomics Analysis

Targeted metabolomic measurements used the Biocrates MxP Quant 500 kit (Biocrates, Innsbruck, Austria). We loaded 10 μL of each sample onto a 96-well extraction plate and derivatized it with 5% phenyl isothiocyanate. Analytes were then extracted with organic solvent and diluted according to the assay workflow. Sample order was randomized, and laboratory personnel were blinded to group assignment. Measurements were obtained on a 6500 QTRAP instrument (AB Sciex, Singapore) using flow injection analysis (FIA) and liquid chromatography-tandem mass spectrometry (LC-MS/MS). The assay monitored 271 metabolites. Initial concentration calculations and peak review were completed in MetIDQ software.

The extracted raw data were exported for further preprocessing. To monitor instrument stability and eliminate batch effects, a quality control (QC) sample with a known concentration was inserted after every 10 test samples in the analytical sequence. Normalization factors were generated by calculating the peak area ratio of the adjacent QC sample to the overall QC samples, which were then used to correct the test samples. Therefore, the data quality could be assessed based on the precision of each metabolite in the QC samples. A 25% filtering standard was applied [22].

We first used principal component analysis (PCA) to inspect the overall distribution of the 129 DNS and 66 HC samples. Group separation was then evaluated with orthogonal partial least squares-discriminant analysis (OPLS-DA) in SIMCA 14.0 (Umetrics, Umeå, Sweden). Model stability was examined using 200 permutations. Metabolites meeting both variable importance in projection (VIP) score > 1 and $P < 0.05$ were carried forward as differential features. Clustering and visualization were performed in MetaboAnalyst 6.0 platform (http://www.metaboanalyst.ca) [23].

2.5 Construction and Interpretation of the Diagnostic Prediction Model

To assist in the early diagnosis of DNS, machine learning algorithms were employed for feature dimensionality reduction and model construction. For the identified differential metabolites, three distinct machine learning algorithms were utilized in parallel: (1) least absolute shrinkage and selection operator (LASSO) regression [24]; (2) random forest recursive feature elimination (RF-RFE) [25]; and (3) support vector machine recursive feature elimination (SVM-RFE) [26]. By intersecting the feature subsets selected by these three algorithms, the core diagnostic metabolites were ultimately identified.

To comprehensively evaluate model performance and validate feature stability, 5000 bootstrap resampling iterations were performed. Subsequently, participants were randomly divided into a training set and a test set at a 7:3 ratio. Five machine learning algorithms, including LR, k-nearest neighbors (KNN), random forest (RF), support vector machine (SVM), and gradient boosting machine (GBM), were evaluated using repeated nested cross-validation [27]. The outer loop comprised five-fold cross-validation repeated three times to estimate model performance, whereas five-fold cross-validation was implemented within each outer training fold for hyperparameter tuning. Predicted probabilities from the held-out outer folds were aggregated to generate receiver operating characteristic (ROC) curves and calculate the area under the curve (AUC), sensitivity, specificity, and F1 score. The best-performing model was further evaluated using confusion matrices, precision–recall curves (PRC), DCA, and calibration curves. Finally, SHapley Additive exPlanations (SHAP) analysis was performed for the optimal RF model to quantify the direction and magnitude of each

metabolite's contribution to model predictions [28]. All diagnostic model development, evaluation, and interpretation procedures were primarily implemented in R (version 4.5.1).

To assess generalizability, the final 10-metabolite RF classifier, trained on the discovery cohort, was applied to an independent validation cohort (13 DNS and 20 HCs). Discriminatory performance was evaluated by the AUC, sensitivity, specificity, and accuracy.

2.6 Construction of a Model for Predicting Treatment Response

To identify baseline metabolites associated with acute-phase treatment response, we analyzed 120 DNS patients with an evaluable week-8 assessment. Based on a $\geq 50\%$ reduction in the PANSS total score at week 8, 58 patients were classified as responders and 62 as non-responders.

A two-step feature selection strategy was applied. First, a RF classifier was trained on the 271 metabolites together with 30 randomly generated Gaussian noise variables [29]. Permutation-based variable importance (VIMP) was averaged across 10 random seeds, and metabolites whose averaged VIMP exceeded the maximum VIMP observed among the noise variables were retained, constituting a noise-calibrated importance filter [30]. Second, univariate LR was performed for each metabolite, and metabolites nominally associated with response at $P < 0.10$ were retained [31, 32]. The intersection of the two sets constituted the final set of candidate metabolites. A multivariable LR model incorporating these candidate metabolites was then constructed to discriminate responders from non-responders. Model performance was summarized by the ROC curve and the corresponding AUC. These analyses were performed in R (version 4.5.1), with RF implemented using the randomForestSRC package.

2.7 Statistical Analysis

Statistical analyses used IBM SPSS 27.0 and R 4.5.1. Distributional assumptions for continuous variables were assessed with the Shapiro–Wilk test. Normally distributed variables were compared with independent-samples t tests, whereas non-normally distributed variables were compared with two-sided Wilcoxon rank-sum tests. Categorical variables were evaluated with chi-square tests. All tests were two-sided,

and $P < 0.05$ defined nominal statistical significance.

3. Results

3.1 Demographic and Clinical Characteristics of Participants in the Discovery Cohort

The discovery cohort comprised 129 DNS patients and 66 HCs (n = 195). As shown in Table 1, the two groups were well-matched in demographic characteristics. Specifically, the mean ages of the DNS and HCs groups were 38.17 ± 10.86 years and 36.48 ± 9.56 years, respectively ($P = 0.287$). The sex distribution (male/female) was 42/87 in the DNS group and 29/37 in the HCs group ($\chi^2 = 2.443$, $P = 0.118$). The average years of education were 12.75 ± 3.29 years and 12.48 ± 3.64 years, respectively ($P = 0.605$).

3.2 Targeted Metabolomic Profiling Reveals Dysregulated Metabolites in DNS Patients

To characterize the plasma metabolic profiles of DNS patients, targeted metabolomic analysis was conducted comparing the DNS and HCs groups. PCA analysis illustrates the overall distribution of the samples (Figure 2A). OPLS-DA revealed a clear separation between the two groups, indicating significant metabolic perturbations in DNS patients (Figure 2B). The validity of the OPLS-DA model was confirmed by a 200-iteration permutation test, which ruled out model overfitting and ensured the statistical reliability of the discrimination (Figure 2C).

Based on strict statistical thresholds of VIP > 1 and $P < 0.05$, a total of 34 significantly dysregulated metabolites were identified between the DNS and HCs groups. Among these, 13 metabolites were up-regulated and 21 were down-regulated (detailed in Supplementary Table 1). Furthermore, hierarchical clustering analysis of these differential metabolites demonstrated distinct abundance patterns, further corroborating the clear metabolic separation between the DNS and HCs groups (Figure 2D).

3.3 Identification of a Machine Learning-Based Metabolomic Biomarker Panel for Auxiliary Diagnosis of DNS

To develop an auxiliary diagnostic tool for DNS, we further mined the

dysregulated metabolic profiles. Three feature selection algorithms were applied to the 34 differential metabolites to identify the most discriminative variables between DNS and HCs. The LASSO regression algorithm optimized the penalty parameter through 10-fold cross-validation, selecting 15 core features (Figure 3A (a) and (b)). Based on the dynamic evaluation of model classification accuracy, RF-RFE (Figure 3B) and SVM-RFE (Figure 3C) selected 16 and 20 features, respectively. The specific metabolites filtered by each algorithm are detailed in Supplementary Table 2. By taking the intersection of the three feature subsets, 10 core metabolites, comprising ADMA, C0, C2, CE (16:1), cortisol, cystine, PC aa C34:4, PC aa C40:2, PC ae C34:3, and PC ae C38:2, were ultimately identified to distinguish between the DNS and HCs (Figure 3D).

To validate the stability of these features, 5000 bootstrap resampling iterations were performed. The results showed that the stability scores of all 10 metabolites were consistently above 0.7, demonstrating the extremely high robustness of the feature selection results (Figure 3E). Furthermore, the expression comparison plot of the 10 core diagnostic metabolites visually demonstrated distinctly different expression trends between the DNS and HCs groups (Figure 3F).

3.4 Diagnostic Evaluation and SHAP Interpretation of the RF Model

To evaluate the diagnostic performance of the 10-metabolite candidate panel, five machine learning classifiers, including LR, KNN, RF, SVM, and GBM, were compared using repeated nested cross-validation. The outer loop comprised five-fold cross-validation repeated three times for performance estimation, whereas five-fold cross-validation was performed within each outer training fold for hyperparameter tuning. RF provided the largest cross-validated AUC: 0.919 (95% CI, 0.880-0.958). (Figure 4A). The complete performance metrics for all five classifiers are presented in Supplementary Table 3. In stratified training–test split analysis, the RF model achieved an accuracy of 84.6%, a sensitivity of 92.1%, and a specificity of 70.2% in the cross-validated training set. In the held-out test set, the corresponding values were 84.2%, 94.7%, and 63.2%, respectively (Figure 4B). In the test set, the precision-recall curve (PRC) showed an AUC of 0. 969 (Figure 4C), and decision curve analysis (DCA)

indicated a positive net benefit across a broad range of threshold probabilities (Figure 4D). Furthermore, the calibration curve revealed a high consistency between the predicted disease probabilities and the actual observed probabilities. The Hosmer–Lemeshow test did not indicate substantial evidence of poor calibration (P = 0.567; Figure 4E).

The SHAP algorithm was applied to interpret the RF model, visualized via beeswarm and dependence plots. As shown in Figure 4F, features were ranked by mean absolute SHAP value in descending order: CE (16:1), C0, C2, PC ae C34:3, cortisol, ADMA, PC ae C38:2, PC aa C34:4, PC aa C40:2, and cystine. Higher levels of CE (16:1), cortisol, ADMA, and cystine generally shifted the model predictions toward the DNS class. In contrast, lower levels of C0, C2, PC ae C34:3, PC ae C38:2, PC aa C34:4, and PC aa C40:2 generally contributed positively to DNS classification.

SHAP dependence plots further illustrated the relationships between individual metabolite concentrations and their contributions to RF predictions (Figure 4G). ADMA, CE (16:1), cortisol, and cystine exhibited generally positive, although partly non-linear, trends, whereas C0, C2, PC aa C34:4, PC aa C40:2, PC ae C34:3, and PC ae C38:2 showed predominantly negative associations with SHAP values. C0, C2, and PC aa C40:2 displayed relatively dispersed or non-linear patterns, while CE (16:1) exhibited a step-like relationship with the model output. Auto-selected interaction features further suggested potential dependence patterns between metabolites, including cortisol–C2, ADMA–PC aa C40:2, CE (16:1)–cystine, and PC ae C34:3–PC ae C38:2, highlighting possible co-variation among metabolite features within the RF model.

3.5 Independent External Validation of the 10-Metabolite Auxiliary Diagnostic Panel

To assess the generalizability of the 10-metabolite auxiliary diagnostic panel, the RF model developed in the discovery cohort was applied, without any retraining or re-tuning, to an independent external validation cohort comprising 13 DNS patients and 20 HCs. The two groups were comparable in demographic and clinical characteristics, with no statistically significant between-group differences in sex, age, educational level,

or BMI (all $P > 0.05$; Supplementary Table 4).

Applied to this independent cohort, the fixed RF model discriminated DNS patients from HCs with an AUC of 0.863 (95% CI: 0.724 - 1.000), a sensitivity of 0.769, a specificity of 0.900, and an overall accuracy of 0.848 (Figure 4H; Supplementary Table 5). Consistent with the discovery findings, the between-group expression patterns of the 10 core metabolites in the external cohort were largely concordant in direction with those observed in the discovery cohort (Figure 4I). Together, these results provide preliminary external validation of the diagnostic panel and support its potential generalizability, although confirmation in larger independent cohorts is warranted.

3.6 Identification of a Metabolomic Biomarker Panel for Predicting Treatment Response in DNS Patients

Of the 129 DNS patients, 9 were excluded owing to incomplete follow-up or a missing week-8 assessment, which precluded determination of treatment response, and the remaining 120 patients were included in the analysis. Based on a ≥ 50% reduction in the PANSS total score at week 8, 58 patients were classified as responders and 62 as non-responders (Figure 5A). At baseline, responders were significantly younger (36.69 ± 10.82 vs 40.77 ± 10.50 years, $P = 0.038$) and had a shorter DUP (36.33 ± 44.64 vs 62.48 ± 67.95 months, $P = 0.014$) than non-responders, whereas sex, educational level, BMI, and baseline PANSS scores did not differ significantly between the two groups (Supplementary Table 6).

As an exploratory analysis, a two-step feature selection strategy was applied to the 271 plasma metabolites to identify baseline metabolites associated with treatment response. First, a RF incorporating 30 randomly generated Gaussian noise variables ranked the metabolites by VIMP, and 39 metabolites exceeded the maximum VIMP observed among the noise variables (threshold = 0.000406; Figure 5B). Second, univariate LR identified 21 metabolites nominally associated with response at $P < 0.10$ (Figure 5C). The intersection of the two approaches yielded a final panel of 10 candidate metabolites (Figure 5D; Supplementary Table 7): Cer(d18:2/18:0), beta-Ala, PC ae C42:1, Sarcosine, TG(16:1_32:1), PC ae C42:2, GLCAS, PC ae C38:0, PC aa C40:6, and HexCer(d18:1/16:0). A multivariable LR model incorporating these 10

metabolites discriminated responders from non-responders, with an AUC of 0.767 (95% CI: 0.682 - 0.852) and, at the Youden-optimal cutoff, a sensitivity of 0.966, a specificity of 0.516, and an overall accuracy of 0.733 (Figure 5E; Supplementary Table 8).

In univariate analyses, five metabolites were significantly associated with higher odds of treatment response: beta-Ala (OR = 1.53, 95% CI: 1.05 - 2.23), TG(16:1_32:1) (OR = 1.52, 1.03 - 2.25), Sarcosine (OR = 1.52, 1.04 - 2.21), PC ae C42:2 (OR = 1.50, 1.03 - 2.19), and PC aa C40:6 (OR = 1.46, 1.00 - 2.12). Conversely, higher baseline Cer(d18:2/18:0) was significantly associated with lower odds of response (OR = 0.66, 0.45 - 0.96). The remaining four candidate metabolites, namely PC ae C38:0, PC ae C42:1, GLCAS, and HexCer(d18:1/16:0), showed directionally consistent but non-significant (trend-level) associations (Supplementary Table 9).

4. Discussion

In this study, we performed targeted plasma metabolomic profiling in DNS patients and HCs, and combined multi-algorithm machine learning approaches to construct and validate two distinct metabolite panels: one for the auxiliary diagnosis of DNS and the other for predicting 8-week antipsychotic treatment response. We found that DNS patients exhibited significant metabolic perturbations, with 34 differentially dysregulated metabolites identified. From these, the intersection of three feature selection algorithms yielded a 10-metabolite diagnostic panel, which achieved excellent diagnostic performance under a RF classifier in nested cross-validation and, importantly, retained good discriminative performance in an independent validation cohort. SHAP-based interpretation further revealed the directionality and non-linear interactions of each metabolite's contribution to the model. In addition, a two-step feature selection strategy combining noise-calibrated RF with univariate LR yielded a 10-metabolite panel that predicted 8-week treatment response, and higher baseline amino acid and phospholipid levels and lower sphingolipid levels were associated with a more favorable response. Together, these findings contribute to more objective early diagnosis and provide preliminary metabolic leads for individualized treatment planning in schizophrenia.

A total of 34 metabolites were differentially abundant between DNS patients and HCs according to the predefined screening criteria. The intersection of three feature-selection algorithms retained a 10-metabolite candidate diagnostic panel comprising ADMA, C0, C2, CE (16:1), cortisol, cystine, PC aa C34:4, PC aa C40:2, PC ae C34:3, and PC ae C38:2. The concordant reductions in the diacyl phosphatidylcholines PC aa C34:4 and PC aa C40:2 and the ether-linked phosphatidylcholines PC ae C34:3 and PC ae C38:2 suggest broad perturbations in phospholipid remodeling. As principal structural components of neuronal membranes [33], phosphatidylcholines play essential roles in synaptic vesicle formation and neurotransmitter release [34, 35]; their deficiency may therefore compromise synaptic plasticity and neural signal transduction, representing a molecular basis for the neural circuit dysfunction characteristic of schizophrenia [36]. Ether-linked phospholipids contribute to membrane organization and dynamics, whereas some subclasses, particularly plasmalogens, may also participate in cellular defense against oxidative stress [37, 38]. Together, these changes reflect multi-layered membrane phospholipid dysregulation characteristic of the metabolic phenotype of schizophrenia. Complementarily, the up-regulation of CE (16:1) is consistent with hyperactivation of phospholipase A2 (PLA2) [39, 40], which liberates excess arachidonic acid to fuel pro-inflammatory eicosanoid synthesis via the COX-2 pathway [41], thereby modulating dopaminergic and glutamatergic neurotransmission [42].

Among the metabolites with relatively high SHAP contributions in the diagnostic model, several converge on mitochondrial and oxidative pathways. The down-regulation of the acylcarnitines C2 and C0, key intermediates of mitochondrial fatty-acid β-oxidation, indicates impaired mitochondrial energy metabolism [43]; as neurons rely heavily on oxidative phosphorylation, this deficit may compromise synaptic transmission and accelerate reactive oxygen species (ROS) accumulation [44]. Consistently, the up-regulation of cystine, a biomarker of oxidative stress [45], suggests a sustained oxidative burden that impairs antioxidant defenses and exacerbates neuroinflammation [46, 47]. In parallel, elevated ADMA, an endogenous inhibitor of nitric oxide synthase [48], suppresses nitric oxide production and thereby impairs

cerebrovascular vasodilation and the local blood supply required for neuronal activity [49]; ADMA further engages in a self-reinforcing relationship with oxidative stress, as ROS can oxidize key enzymes of the arginine metabolic pathway to promote additional ADMA accumulation, establishing a vicious cycle between neurovascular dysfunction and oxidative damage [50]. Likewise, elevated cortisol signals chronic hyperactivation of the hypothalamic-pituitary-adrenal (HPA) axis, with sustained glucocorticoid excess inhibiting hippocampal neurogenesis [51] and promoting pro-inflammatory cytokine release that, in turn, impairs mitochondrial oxidative phosphorylation [52]. Notably, the cortisol–C2 interaction identified by SHAP analysis indicates that HPA-axis dysregulation and mitochondrial energy impairment act reciprocally to sustain the metabolic dysregulation in DNS patients.

Beyond its biological plausibility, the diagnostic model developed here offers several methodological strengths. The 10-metabolite panel achieved an AUC of 0.919 under repeated nested cross-validation and preserved an AUC of 0.863 in an independent external cohort, comparable to or exceeding previously reported metabolomics-based diagnostic models for schizophrenia [53]. First, rather than relying on a single algorithm, we derived the core panel from the intersection of three complementary feature-selection methods, mitigating the overfitting and instability inherent to any single method [54]; feature robustness was further supported by 5,000 bootstrap resampling iterations, with all 10 metabolites exceeding a stability score of 0.7. Second, model performance was estimated by repeated nested cross-validation across five classifiers, thereby avoiding the optimism bias of a single train–test split [55]; RF performed best, likely reflecting its capacity to model high-dimensional, non-linear relationships among metabolites [56]. Third, and most importantly, we validated the locked model in an independent external cohort without any retraining or re-tuning. Although modest in size (13 DNS and 20 HCs), such external validation is essential for assessing generalizability [57], and the preserved discrimination provides preliminary evidence supporting future clinical translation.

Beyond early diagnosis, the reliable prediction of acute-phase treatment response remains a critical unmet need in schizophrenia. In this exploratory analysis, a two-step

strategy applied to 271 plasma metabolites yielded a 10-metabolite panel whose multivariable LR model discriminated responders from non-responders with an AUC of 0.767. Although lower than that of the diagnostic model, this performance is noteworthy given that treatment response is shaped by multiple factors beyond baseline biology, including drug class, dosage, and adherence. Directionally, higher baseline amino acid, triacylglycerol, and phosphatidylcholine levels were associated with a greater likelihood of response, whereas higher ceramide levels were associated with a lower likelihood. Several of the individually significant metabolites bear direct mechanistic relevance to antipsychotic action. Sarcosine, an endogenous glycine transporter-1 (GlyT1) inhibitor that potentiates NMDA-receptor function and has been evaluated as an adjunctive treatment for schizophrenia [58, 59], showed one of the strongest positive associations, suggesting that greater baseline glutamatergic tone may favor antipsychotic-induced remission [60]. β-alanine, a precursor of the antioxidant dipeptide carnosine [61], together with the phosphatidylcholines PC ae C42:2 and PC aa C40:6, implicates neuronal membrane integrity and redox homeostasis [36], as these lipids are principal constituents of neuronal membranes and, for the ether-linked (plasmalogen-type) species, contribute to antioxidant defense [37, 62]. Conversely, the inverse association of the ceramide Cer(d18:2/18:0) may reflect altered sphingolipid metabolism, which has been implicated in inflammatory signaling, apoptosis, and metabolic dysfunction [63, 64]. Collectively, these metabolites implicate glutamatergic, membrane-lipid, and neuroinflammatory pathways in treatment response.

This study still has some limitations. First, recruitment occurred at one center, and both the discovery and validation samples were modest. Second, the associations between baseline metabolites and treatment response were exploratory, with modest effects that did not survive multiple-testing correction. The treatment response was also assessed only during the acute phase (8 weeks), precluding evaluation of long-term relapse, functional outcomes, and treatment resistance. Third, the predefined analyte coverage of the targeted metabolomics platform may have overlooked relevant metabolites, necessitating complementary untargeted metabolomic profiling. Fourth,

the mechanistic interpretations remain inferential, and future studies integrating neuroimaging, genetics, and cerebrospinal fluid metabolomics are warranted.

5. Conclusion

In summary, by integrating multi-algorithm feature selection with machine learning, we developed two plasma metabolite panels demonstrating promising performance for the auxiliary diagnosis of DNS and for predicting 8-week antipsychotic treatment response, respectively, with SHAP-based interpretation providing biologically plausible explanations. These findings provide new metabolomic evidence for the objective diagnosis and individualized treatment of schizophrenia and establish a foundation for prospective validation and mechanistic studies. Larger independent cohorts and longitudinal designs are needed to determine the clinical translational value of these metabolite-based biomarkers.

Declarations

Abbreviations

ADMA (asymmetric dimethylarginine), AUC (area under the curve), BMI (body mass index), CE (cholesteryl ester), CI (confidence interval), COX-2 (cyclooxygenase-2), DCA (decision curve analysis), DNS (drug-naïve schizophrenia), DSM-5 (Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition), DUP (duration of untreated psychosis), EDTA (ethylenediaminetetraacetic acid), FIA (flow injection analysis), GBM (gradient boosting machine), HC (healthy control), HPA (hypothalamic-pituitary-adrenal), ICC (intraclass correlation coefficient), KNN (k-nearest neighbors), LASSO (least absolute shrinkage and selection operator), LC-MS/MS (liquid chromatography–tandem mass spectrometry), LR (logistic regression), MECT (modified electroconvulsive therapy), ML (machine learning), NMDA (N-methyl-D-aspartate), OPLS-DA (orthogonal partial least squares-discriminant analysis), OR (odds ratio), PANSS (Positive and Negative Syndrome Scale), PC (phosphatidylcholine), PCA (principal component analysis), PLA2 (phospholipase A2), PRC (precision–recall curve), QC (quality control), RF (random forest), RF-RFE (random forest recursive feature elimination), ROC (receiver operating characteristic), ROS (reactive oxygen species), SHAP (SHapley Additive exPlanations), SVM (support

vector machine), SVM-RFE (support vector machine recursive feature elimination), TG (triacylglycerol), TMS (transcranial magnetic stimulation), VIMP (variable importance), VIP (variable importance in projection).

Ethics approval and consent to participate

The study was conducted in accordance with the Declaration of Helsinki and approved by the Ethics Committee of Tianjin Anding Hospital (approval numbers 2017-03 and 2023-63). Written informed consent to participate was obtained from all participants in both the discovery and independent validation cohorts.

Consent for publication

Not applicable.

Availability of data and materials

The authors declare that all relevant data of this study are available within the article or from the corresponding author on reasonable request.

Competing Interests

The authors declare that they have no competing interests.

Funding

This work was supported by National Natural Science Foundation of China (82401770), Tianjin Key Medical Discipline Construction Project (Grant No. TJYXZDXK-3-015B), Tianjin Health Research Project (TJWJ2022XK039), and the Special Research Fund for Central Universities, Peking Union Medical College (3332025061).

Authors' contributions

J.L., Y.D., and M.L. contributed to conceptualization, methodology, resources, and manuscript review and editing. S.W., B.Z., and X.Z. performed formal analysis, prepared visualizations, and wrote the original draft. Y.G., T.L., and J.G. contributed to data curation, formal analysis, and software. Y.Q. contributed to data curation, software, and methodology. All authors read and approved the final manuscript.

Acknowledgements

The authors thank all participants for completing assessments and interviews.

Reference:

1. Leucht, S., et al., *Schizophrenia.* Nat Rev Dis Primers, 2025. **11**(1): p. 83.
2. Solmi, M., et al., *Incidence, prevalence, and global burden of schizophrenia - data, with critical appraisal, from the Global Burden of Disease (GBD) 2019.* Molecular Psychiatry, 2023. **28**(12): p. 5319-5327.
3. Krasa, H.B., et al., *National and State Societal Costs of Schizophrenia in the US in 2024.* JAMA Psychiatry, 2026. **83**(4): p. 341-352.
4. Wehr, S., et al., *Clinical Assessment Interview for Negative Symptoms (CAINS): A Systematic Review of Measurement Properties.* Schizophr Bull, 2024. **50**(4): p. 747-756.
5. Bosnjak Kuharic, D., et al., *Interventions for prodromal stage of psychosis.* Cochrane Database Syst Rev, 2019. **2019**(11).
6. Winkelbeiner, S., et al., *Evaluation of Differences in Individual Treatment Response in Schizophrenia Spectrum Disorders: A Meta-analysis.* JAMA Psychiatry, 2019. **76**(10): p. 1063-1073.
7. McCutcheon, R.A., et al., *The efficacy and heterogeneity of antipsychotic response in schizophrenia: A meta-analysis.* Mol Psychiatry, 2021. **26**(4): p. 1310-1320.
8. Williams, L.M., et al., *Precision psychiatry and Research Domain Criteria: Implications for clinical trials and future practice.* CNS Spectr, 2024. **29**(1): p. 26-39.
9. Hajjar, G., et al., *Scaling-up metabolomics: Current state and perspectives.* TrAC Trends in Analytical Chemistry, 2023. **167**: p. 117225.
10. Xu, C.X., et al., *Dysregulation of Serum Exosomal Lipid Metabolism in Schizophrenia: A Biomarker Perspective.* Mol Neurobiol, 2025. **62**(3): p. 3556-3567.
11. McMillen, T.S., et al., *A large-scale, targeted metabolomics method for the analysis and quantification of metabolites in human plasma via liquid chromatography-mass spectrometry.* Anal Chim Acta, 2023. **1279**: p. 341791.
12. Tian, H., et al., *Precise Metabolomics Reveals a Diversity of Aging-Associated Metabolic Features.* Small Methods, 2022. **6**(7): p. e2200130.
13. Dielemans, D.A.M., et al., *Alterations in lipid, redox, and energy metabolism in recent onset psychosis: a metabolomics study in non-smoking individuals and matched controls.* Redox Biol, 2026. **89**: p. 103999.
14. Liu, Y., et al., *Alteration of lipids and amino acids in plasma distinguish schizophrenia patients from controls: A targeted metabolomics study.* Psychiatry Clin Neurosci, 2021. **75**(4): p. 138-144.
15. Rattray, N.J.W., et al., *Beyond genomics: understanding exposotypes through metabolomics.* Hum Genomics, 2018. **12**(1): p. 4.
16. Asif, S., et al., *Advancements and Prospects of Machine Learning in Medical Diagnostics: Unveiling the Future of Diagnostic Precision.* Archives of Computational Methods in Engineering, 2025. **32**(2): p. 853-883.
17. Reel, P.S., et al., *Using machine learning approaches for multi-omics data analysis: A review.* Biotechnol Adv, 2021. **49**: p. 107739.

18. Zhang, S., et al., *Serum lipid metabolic characteristics and potential biomarkers in first-episode schizophrenia.* BMC Psychiatry, 2025. **25**(1): p. 1124.
19. Liu, M., et al., *Targeted metabolomics study on peripheral blood neurotransmitters in early-onset schizophrenia.* J Transl Med, 2025. **23**(1): p. 1238.
20. Kay, S.R., A. Fiszbein, and L.A. Opler, *The Positive and Negative Syndrome Scale (PANSS) for Schizophrenia.* Schizophrenia Bulletin, 1987. **13**(2): p. 261-276.
21. Sun, Y., et al., *Contribution of copy number variants on antipsychotic treatment response in Han Chinese patients with schizophrenia.* eBioMedicine, 2024. **105**: p. 105195.
22. Bijlsma, S., et al., *Large-scale human metabolomics studies: a strategy for data (pre-) processing and validation.* Anal Chem, 2006. **78**(2): p. 567-74.
23. Pang, Z., et al., *MetaboAnalyst 6.0: towards a unified platform for metabolomics data processing, analysis and interpretation.* Nucleic Acids Research, 2024. **52**(W1): p. W398-W406.
24. Wang, Y., et al., *Construction and evaluation of a mortality prediction model for patients with acute kidney injury undergoing continuous renal replacement therapy based on machine learning algorithms.* Ann Med, 2024. **56**(1): p. 2388709.
25. Ge, C., et al., *Machine learning and metabolomics identify biomarkers associated with the disease extent of ulcerative colitis.* J Crohns Colitis, 2025. **19**(2).
26. Yang, M., et al., *Establishing a prediction model of severe acute mountain sickness using machine learning of support vector machine recursive feature elimination.* Sci Rep, 2023. **13**(1): p. 4633.
27. Breiman, L., *Random Forests.* Machine Learning, 2001. **45**(1): p. 5-32.
28. Song, Y., et al., *Prediction models for postoperative delirium in elderly patients with machine-learning algorithms and SHapley Additive exPlanations.* Transl Psychiatry, 2024. **14**(1): p. 57.
29. Kursa, M.B. and W.R. Rudnicki, *Feature Selection with the Boruta Package.* Journal of Statistical Software, 2010. **36**(11): p. 1 - 13.
30. Ishwaran, H. and M. Lu, *Standard errors and confidence intervals for variable importance in random forest regression, classification, and survival.* Stat Med, 2019. **38**(4): p. 558-582.
31. Jin, W., et al., *A clinical prediction model for schizophrenia based on machine learning algorithms.* Front Med (Lausanne), 2025. **12**: p. 1726905.
32. Bursac, Z., et al., *Purposeful selection of variables in logistic regression.* Source Code Biol Med, 2008. **3**: p. 17.
33. Vanherle, S., et al., *Lipid metabolism, remodelling and intercellular transfer in the CNS.* Nat Rev Neurosci, 2025. **26**(4): p. 214-231.
34. Park, S., Y. Jin, and A.D. Chisholm, *Phospholipid biogenesis maintains neuronal integrity during aging and axon regeneration.* Genetics, 2025.

231(1).
35. Dong, Y., et al., *Plasma biomarkers in drug-naïve schizophrenia with severe cognitive impairment via targeted metabolomics.* Psychoneuroendocrinology, 2026. **184**: p. 107711.
36. Li, M., et al., *Impaired Membrane Lipid Homeostasis in Schizophrenia.* Schizophr Bull, 2022. **48**(5): p. 1125-1135.
37. Dean, J.M. and I.J. Lodhi, *Structural and functional roles of ether lipids.* Protein Cell, 2018. **9**(2): p. 196-206.
38. Broniec, A., et al., *Interactions of plasmalogens and their diacyl analogs with singlet oxygen in selected model systems.* Free Radic Biol Med, 2011. **50**(7): p. 892-8.
39. Gao, Z., et al., *Copy number deletion of PLA2G4A affects the susceptibility and clinical phenotypes of schizophrenia.* Schizophrenia (Heidelb), 2024. **10**(1): p. 55.
40. Murakami, M., H. Sato, and Y. Taketomi, *Updating Phospholipase A(2) Biology.* Biomolecules, 2020. **10**(10).
41. Yang, X., et al., *Dysregulation of phospholipase and cyclooxygenase expression is involved in Schizophrenia.* EBioMedicine, 2021. **64**: p. 103239.
42. Nayak, U., et al., *Impact of neuroinflammation on brain glutamate and dopamine signalling in schizophrenia: an update.* Metab Brain Dis, 2025. **40**(2): p. 119.
43. Cao, B., et al., *Characterizing acyl-carnitine biosignatures for schizophrenia: a longitudinal pre- and post-treatment study.* Transl Psychiatry, 2019. **9**(1): p. 19.
44. Mednova, I.A., et al., *Amino Acid and Acylcarnitine Levels in Chronic Patients with Schizophrenia: A Preliminary Study.* Metabolites, 2021. **11**(1).
45. Hung, C.C., C.H. Lin, and H.Y. Lane, *Cystine/Glutamate Antiporter in Schizophrenia: From Molecular Mechanism to Novel Biomarker and Treatment.* Int J Mol Sci, 2021. **22**(18).
46. Fraguas, D., et al., *Oxidative Stress and Inflammation in First-Episode Psychosis: A Systematic Review and Meta-analysis.* Schizophr Bull, 2019. **45**(4): p. 742-751.
47. Yang, X., et al., *Increased serum homocysteine in first episode and drug-naïve individuals with schizophrenia: sex differences and correlations with clinical symptoms.* BMC Psychiatry, 2022. **22**(1): p. 759.
48. Poeggeler, B., et al., *Nitric Oxide as a Determinant of Human Longevity and Health Span.* Int J Mol Sci, 2023. **24**(19).
49. Zhao, Y., et al., *DDAH-1 maintains endoplasmic reticulum-mitochondria contacts and protects dopaminergic neurons in Parkinson's disease.* Cell Death Dis, 2024. **15**(6): p. 399.
50. Zinellu, A., et al., *A systematic review and meta-analysis of nitric oxide-associated arginine metabolites in schizophrenia.* Transl Psychiatry, 2024. **14**(1): p. 439.
51. Pagliarini, M., et al., *Circulating Neuronal Exosome Cargo as Biomarkers of*

Neuroplasticity in Cushing's Syndrome. Mol Neurobiol, 2025. **62**(10): p. 12498-12510.
52. Auger, J.P., et al., *Metabolic rewiring promotes anti-inflammatory effects of glucocorticoids.* Nature, 2024. **629**(8010): p. 184-192.
53. Du, Y., et al., *Metabolomic Identification of Exosome-Derived Biomarkers for Schizophrenia: A Large Multicenter Study.* Schizophr Bull, 2021. **47**(3): p. 615-623.
54. Budhraja, S., et al., *Filter and Wrapper Stacking Ensemble (FWSE): a robust approach for reliable biomarker discovery in high-dimensional omics data.* Brief Bioinform, 2023. **24**(6).
55. Lewis, M.J., et al., *nestedcv: an R package for fast implementation of nested cross-validation with embedded feature selection designed for transcriptomics and high-dimensional data.* Bioinform Adv, 2023. **3**(1): p. vbad048.
56. Bifarin, O.O., *Interpretable machine learning with tree-based shapley additive explanations: Application to metabolomics datasets for binary classification.* PLoS One, 2023. **18**(5): p. e0284315.
57. Riley, R.D., et al., *Evaluation of clinical prediction models (part 2): how to undertake an external validation study.* Bmj, 2024. **384**: p. e074820.
58. Lane, H.Y., et al., *Glycine transporter I inhibitor, N-methylglycine (sarcosine), added to clozapine for the treatment of schizophrenia.* Biol Psychiatry, 2006. **60**(6): p. 645-9.
59. Pinna, A. and A. Pałasz, *Modulation of glycine transporters as a novel therapeutic strategy in neuropsychiatry.* Psychopharmacology (Berl), 2026. **243**(3): p. 495-512.
60. Liang, C.W., H.Y. Cheng, and M.M. Tseng, *Augmentation with glutamatergic modulators for schizophrenia: A network meta-analysis.* Prog Neuropsychopharmacol Biol Psychiatry, 2025. **142**: p. 111495.
61. Cesak, O., et al., *Carnosine and Beta-Alanine Supplementation in Human Medicine: Narrative Review and Critical Assessment.* Nutrients, 2023. **15**(7).
62. Caruso, G., F. Caraci, and R.B. Jolivet, *Pivotal role of carnosine in the modulation of brain cells activity: Multimodal mechanism of action and therapeutic potential in neurodegenerative disorders.* Prog Neurobiol, 2019. **175**: p. 35-53.
63. Turpin-Nolan, S.M. and J.C. Brüning, *The role of ceramides in metabolic disorders: when size and localization matters.* Nat Rev Endocrinol, 2020. **16**(4): p. 224-233.
64. Kuo, A. and T. Hla, *Regulation of cellular and systemic sphingolipid homeostasis.* Nat Rev Mol Cell Biol, 2024. **25**(10): p. 802-821.

Figure 1. Flowchart of the study.

Figure 2. Plasma targeted metabolomics analysis of schizophrenia patients and healthy controls.

(A) Unsupervised principal component analysis (PCA) score plot of the baseline plasma targeted metabolomics data comparing drug-naïve schizophrenia (DNS) patients and healthy controls (HCs).

(B) Orthogonal partial least squares-discriminant analysis (OPLS-DA) score plot distinguishing DNS patients from the HCs group.

(C) A 200-iteration permutation test was performed to evaluate the robustness of the OPLS-DA model and verify the absence of overfitting.

(D) Heat map of differential metabolites by clustering analysis according to their variation trends. Red indicates high levels of metabolite expression, while blue indicates low levels of expression. The tree diagram on the left shows the hierarchical clustering of differential metabolites.

Figure 3. Screening and stability validation of core diagnostic biomarkers based on integrated machine learning algorithms.

(A) Feature screening based on the least absolute shrinkage and selection operator (LASSO) regression algorithm. The left panel (a) shows the binomial deviance curve versus the penalty parameter log(λ) based on 10-fold cross-validation; the right panel (b) displays the LASSO coefficient profile plot of metabolite features shrinking as the penalty increases.

(B) Feature screening process using the random forest recursive feature elimination (RF-RFE) algorithm.

(C) Feature screening process using the support vector machine recursive feature elimination (SVM-RFE) algorithm.

(D) Venn diagram displaying the subsets of metabolite features selected by three independent machine learning algorithms (LASSO, RF-RFE, and SVM-RFE) and their intersection to identify 10 core differential metabolites

(E) Feature stability evaluation using 5000 bootstrap resampling iterations, showing the selection frequency of each candidate metabolite. The dashed line indicates the high stability threshold (0.7). Red dots represent the 10 core metabolites; grey dots represent other metabolites.

(F) Comparison of normalized intensity levels of the 10 core diagnostic metabolites between the DNS and HCs groups. Statistical significance: *$p \leq 0.05$, **$p \leq 0.01$, ***$p \leq 0.001$.

Figure 4. Construction and multidimensional performance evaluation of the machine learning-based auxiliary diagnostic model using the core metabolite panel.

(A) Receiver operating characteristic (ROC) curves of five machine learning classifiers evaluated by nested cross-validation: random forest (RF, AUC = 0.919), gradient boosting machine (GBM, AUC = 0.915), k-nearest neighbors (KNN, AUC = 0.896), support vector machine (SVM, AUC = 0.893), and logistic regression (LR, AUC = 0.880). The RF model achieved the highest discriminative performance.

(B) Confusion matrices of the best-performing RF model on the cross-validated training set (left; Accuracy = 84.6%, Sensitivity = 92.1%, Specificity = 70.2%) and the independent test set (right; Accuracy = 84.2%, Sensitivity = 94.7%, Specificity = 63.2%). Color intensity indicates column-wise classification proportions.

(C) Precision-recall curve (PRC) of the RF model, with an area under the precision-recall curve (AUPRC) of 0.969.

(D) Decision curve analysis (DCA) of the RF model, plotting net benefit across a range of threshold probabilities; the red curve represents the RF model, compared with the "treat all" and "treat none" reference strategies.

(E) Calibration curve of the RF model assessing the agreement between predicted and observed probabilities; the Hosmer-Lemeshow test indicated good calibration ($p = 0.567$). The dashed diagonal line represents perfect calibration.

(F) SHAP beeswarm plot of the RF model illustrating the contribution of each metabolite to individual predictions. Each dot represents a single sample; the color gradient (blue to red) indicates the feature value from low to high. Features are ranked by mean absolute SHAP value. A SHAP value greater than 0 pushes the prediction toward schizophrenia, whereas a value less than 0 indicates the opposite.

(G) SHAP dependence plots for all 10 core metabolites based on the RF model. The x-axis represents the feature value, the y-axis the corresponding SHAP value, and the color the value of the automatically selected interacting feature.

(H) ROC curve of the locked 10-metabolite RF panel applied to the independent external validation cohort (13 DNS and 20 HCs), achieving an AUC of 0.863 (95% CI: 0.724 - 1.000).

(I) Comparison of the normalized intensity of the 10 core metabolites between DNS and HCs groups in the external validation cohort. Statistical significance: ns, $p \geq 0.05$; *$p < 0.05$; **$p < 0.01$; ***$p < 0.001$.

Figure 5. Construction and performance evaluation of an acute-phase prognostic model for drug-naïve schizophrenia based on multidimensional features.

(A) Flowchart of the treatment-response cohort. Of 129 DNS patients, 9 were excluded because of incomplete follow-up or a missing week-8 assessment, leaving 120 patients, who were classified as treatment responders (n = 58) or non-responders (n = 62) according to a ≥ 50% reduction in the PANSS total score at week 8.

(B) Baseline metabolites ranked by permutation variable importance (VIMP) from a random forest incorporating 30 randomly generated noise variables. The top 35 of 271 metabolites are shown; the red dashed line indicates the noise threshold (VIMP = 0.000406), and 39 metabolites exceeded this threshold (red bars).

(C) Volcano plot of the univariate logistic regression screening of the 271 metabolites; the x-axis shows the odds ratio per 1 SD and the y-axis the −log10(P). Metabolites associated with response at $P < 0.10$ are highlighted in red (n = 21).

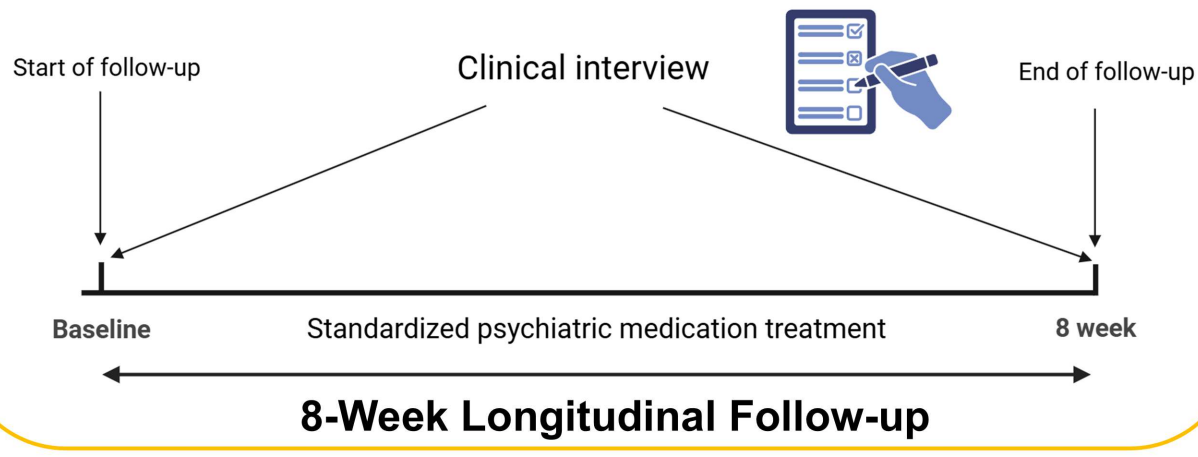
(D) Venn diagram of the intersection between metabolites exceeding the RF-VIMP noise threshold (n = 39) and those with univariate $P < 0.10$ (n = 21), yielding 10 final candidate metabolites.

(E) ROC curve of the logistic regression model based on the 10 candidate metabolites for discriminating responders from non-responders (AUC = 0.767).

Cohort Enrollment & Design

195 Participants

Drug-Naïve Schizophrenia (DNS) Patients (n=129)



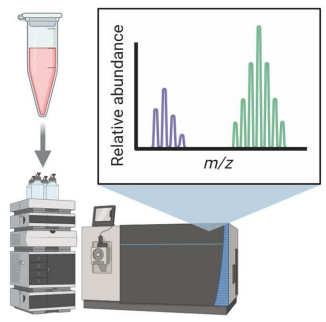
8-Week Longitudinal Follow-up

Healthy Controls (HCs) (n=66)

- **Clinical Interview**
- **Demographic Data**

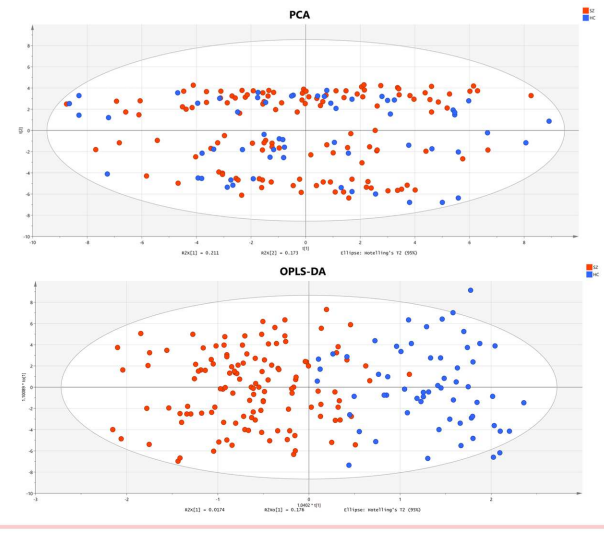
- **Plasma Sample**

Data Acquisition & Analysis



Targeted Metabolomics Analysis (Biocrates MxP Quant 500 kit)

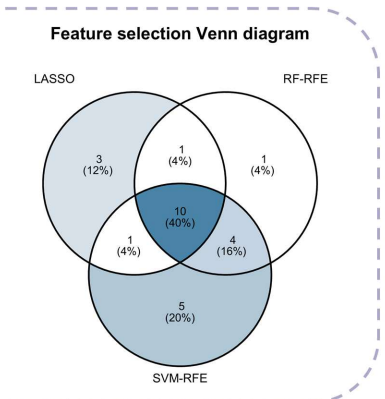
271 Metabolites Monitored



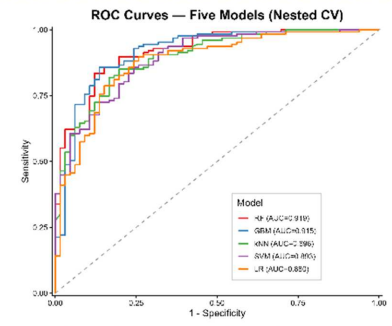
Model construction based on Machine Learning

Diagnostic model

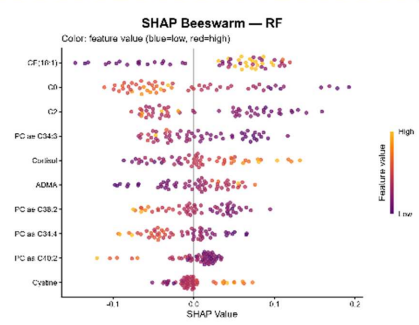
10 core diagnostic metabolites screened



Optimal RF Diagnostic Model

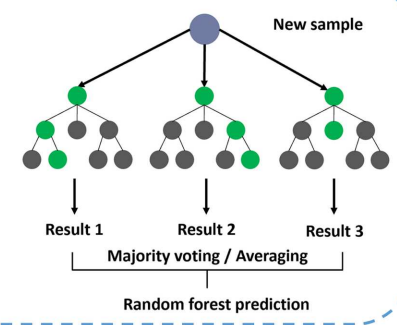


SHAP analysis of RF model

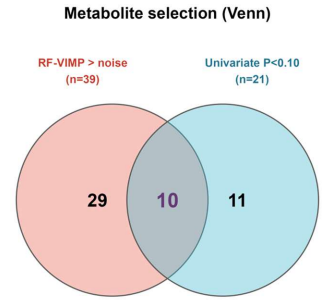


Prediction model

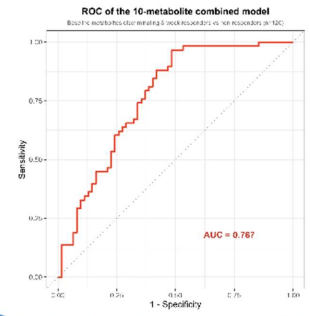
Construction based on RF model



10 core predictive metabolites for treatment



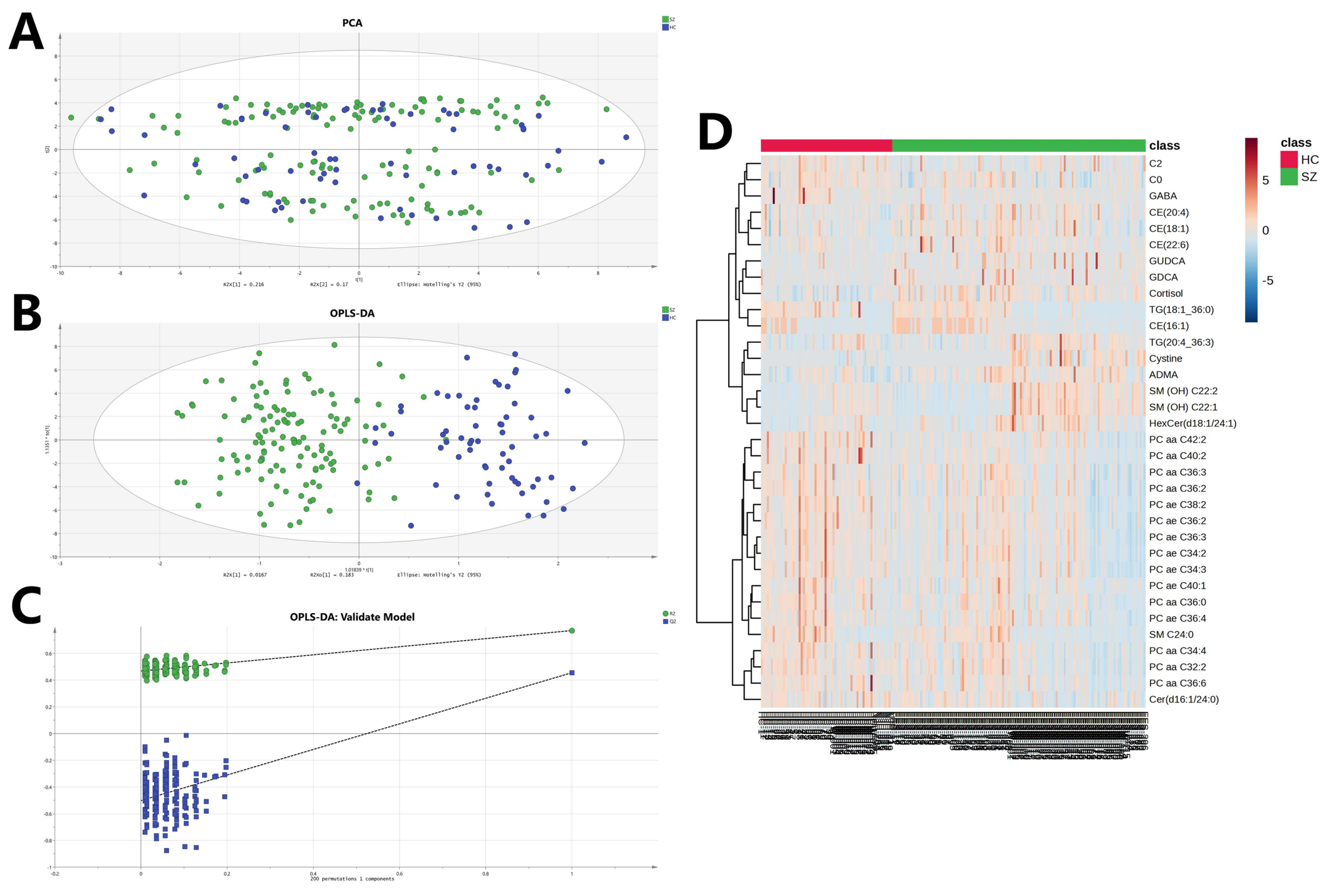
Validation of the validity of the model



Supplementary Table Univariate associations of the 10 final metabolites with 8-week treatment response.

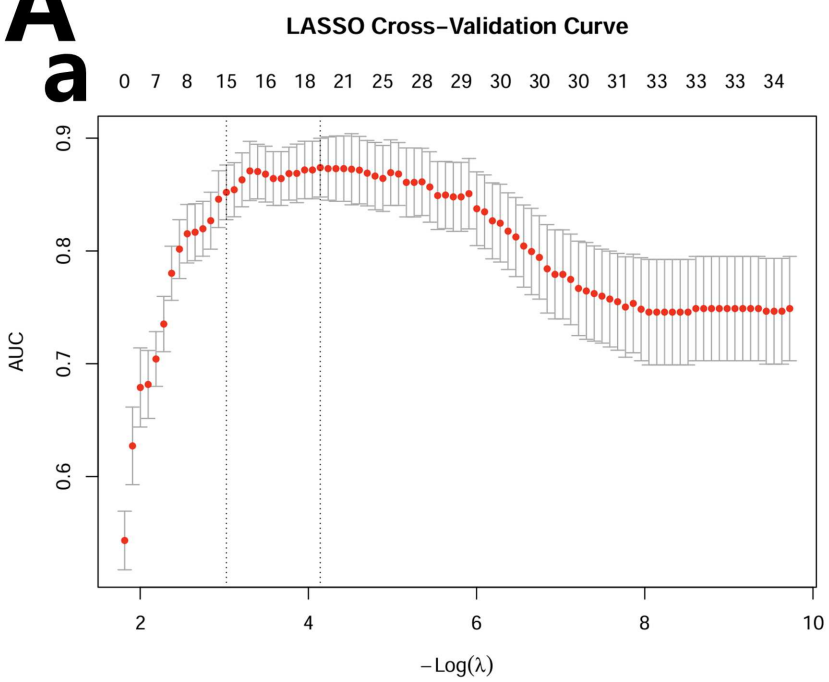
Metabolites	OR (95% CI)	Direction
beta-Ala	1.53 (1.05 - 2.23)	Higher response
TG (16:1_32:1)	1.52 (1.03 - 2.25)	Higher response
Sarcosine	1.52 (1.04 - 2.21)	Higher response
PC ae C42:2	1.50 (1.03 - 2.19)	Higher response
PC aa C40:6	1.46 (1.00 - 2.12)	Higher response
PC ae C38:0	1.43 (0.98 - 2.08)	Higher response
PC ae C42:1	1.40 (0.97 - 2.04)	Higher response
GLCAS	0.72 (0.49 - 1.04)	Lower response
HexCer (d18:1/16:0)	0.69 (0.44 - 1.07)	Lower response
Cer(d18:2/18:0)	0.66 (0.45 - 0.96)	Lower response

Note: OR: odds ratios; CI: confidence interval.

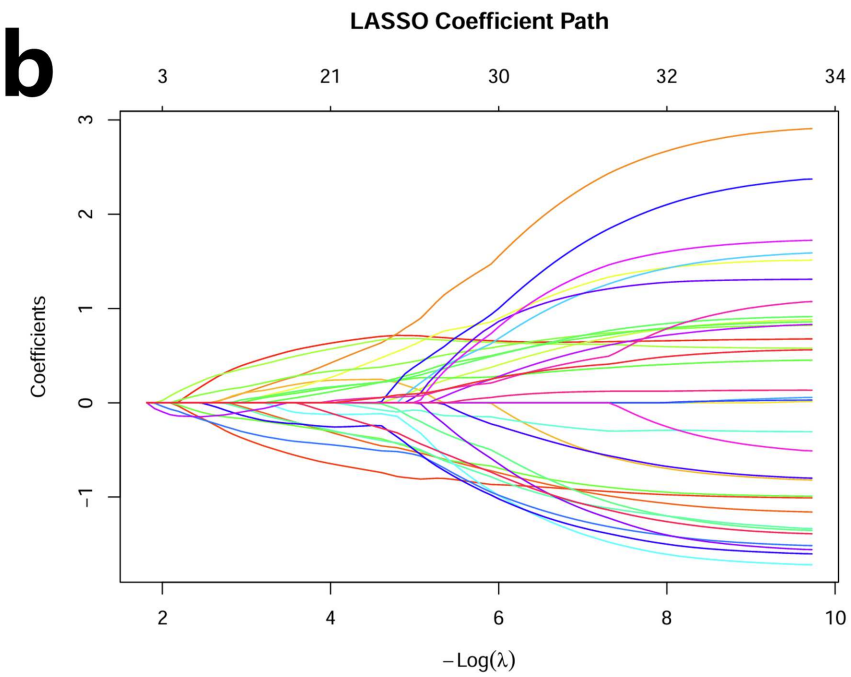

A
PCA
SZ
HC
t[2]
t[1]
R2X[1] = 0.216
R2X[2] = 0.17
Ellipse: Hotelling's T2 (95%)
B
OPLS-DA
SZ
HC
1.1351 * to[1]
1.01839 * t[1]
R2X[1] = 0.0167
R2Xo[1] = 0.183
Ellipse: Hotelling's T2 (95%)
C
OPLS-DA: Validate Model
R2
Q2
200 permutations 1 components
D
class
C2
C0
GABA
CE(20:4)
CE(18:1)
CE(22:6)
GUDCA
GDCA
Cortisol
TG(18:1_36:0)
CE(16:1)
TG(20:4_36:3)
Cystine
ADMA
SM (OH) C22:2
SM (OH) C22:1
HexCer(d18:1/24:1)
PC aa C42:2
PC aa C40:2
PC aa C36:3
PC aa C36:2
PC ae C38:2
PC ae C36:2
PC ae C36:3
PC ae C34:2
PC ae C34:3
PC ae C40:1
PC aa C36:0
PC ae C36:4
SM C24:0
PC aa C34:4
PC aa C32:2
PC aa C36:6
Cer(d16:1/24:0)
class
HC
SZ
5
0
-5

A

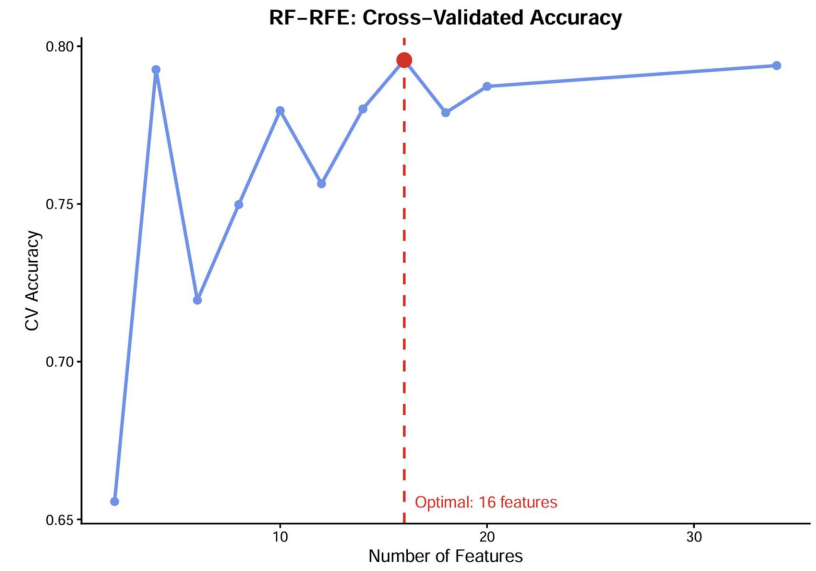
a


LASSO Cross-Validation Curve
0 7 8 15 16 18 21 25 28 29 30 30 30 31 33 33 33 34
AUC
0.6
0.7
0.8
0.9
2
4
6
8
10
−Log(λ)

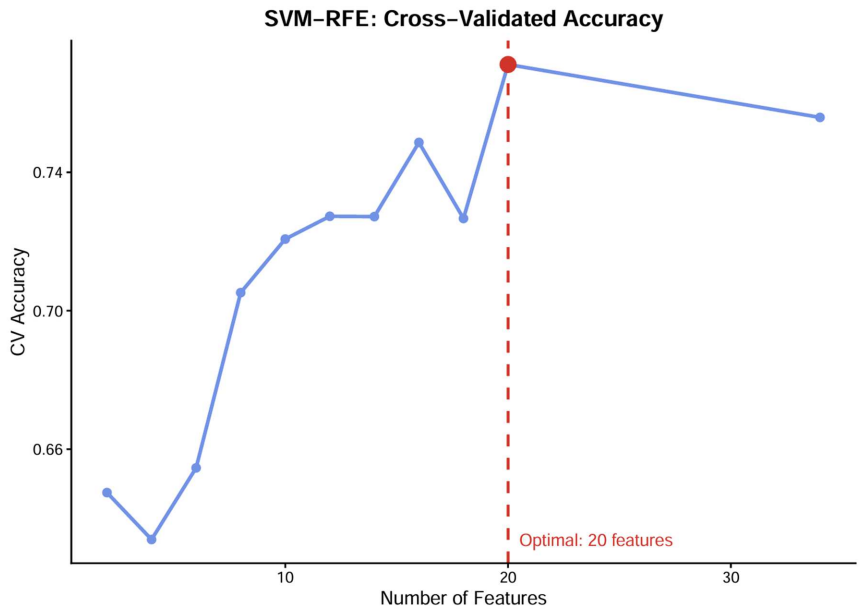
b


LASSO Coefficient Path
3 21 30 32 34
Coefficients
−1
0
1
2
3
2
4
6
8
10
−Log(λ)

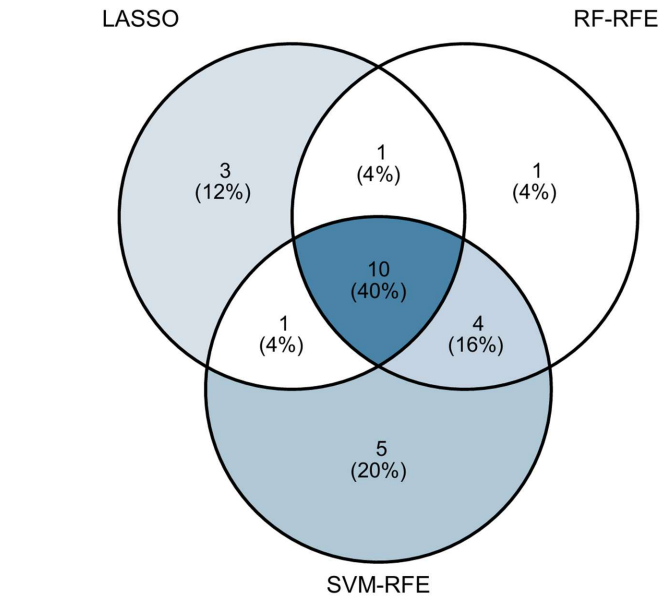
B


RF-RFE: Cross-Validated Accuracy
CV Accuracy
0.65
0.70
0.75
0.80
Optimal: 16 features
10
20
30
Number of Features

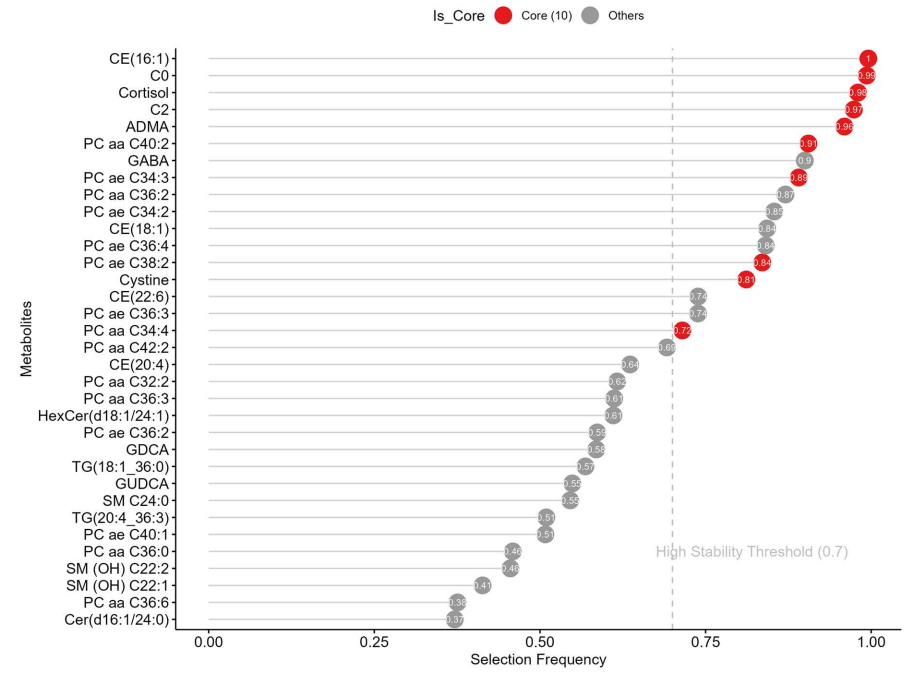
C


SVM-RFE: Cross-Validated Accuracy
CV Accuracy
0.66
0.70
0.74
Optimal: 20 features
10
20
30
Number of Features

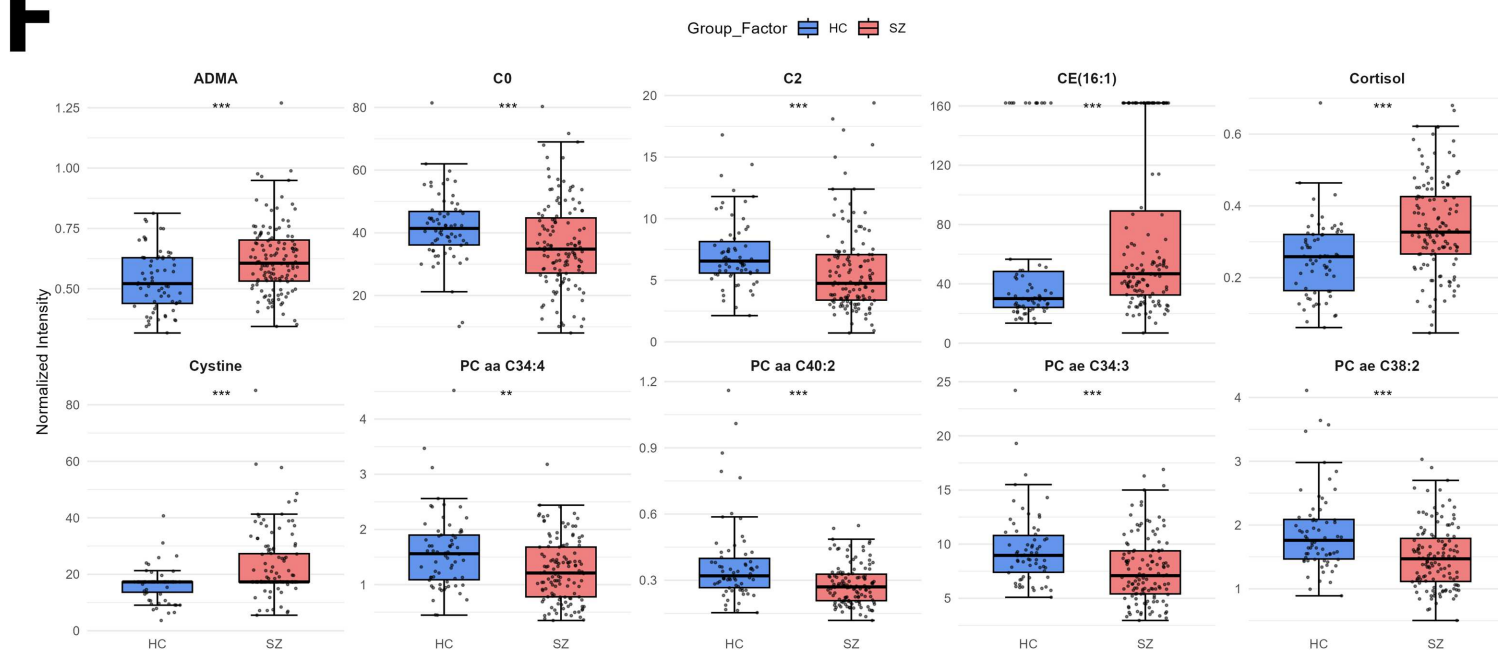
D

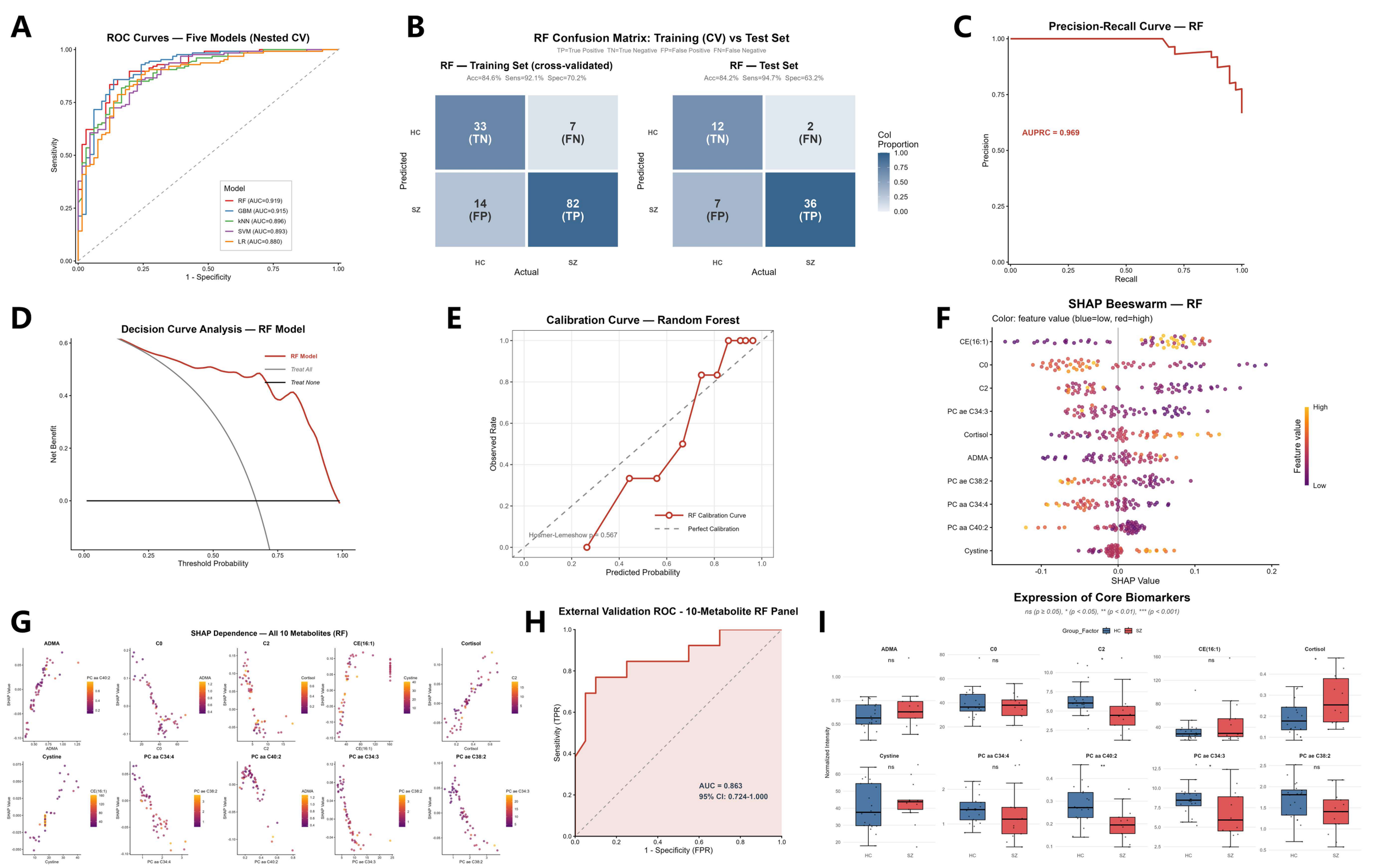

Feature selection Venn diagram
LASSO
RF-RFE
SVM-RFE
3
(12%)
1
(4%)
1
(4%)
10
(40%)
1
(4%)
4
(16%)
5
(20%)

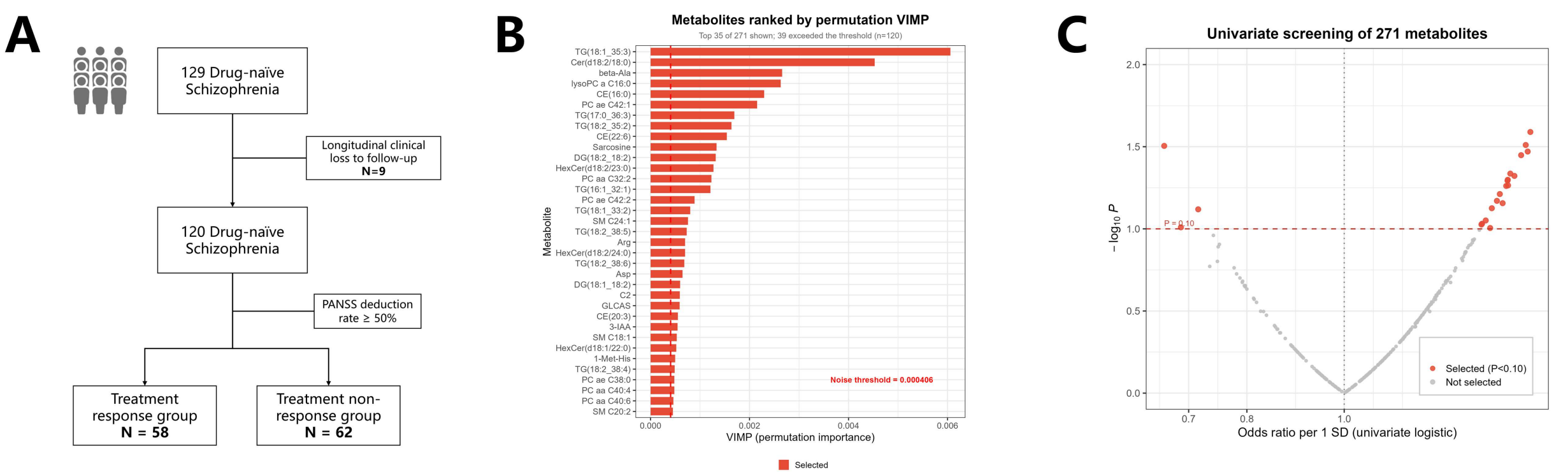
E


Feature Selection Stability Analysis (5000 Bootstraps)
Is_Core
Core (10)
Others
Metabolites
CE(16:1)
C0
Cortisol
C2
ADMA
PC aa C40:2
GABA
PC ae C34:3
PC aa C36:2
PC ae C34:2
CE(18:1)
PC ae C36:4
PC ae C38:2
Cystine
CE(22:6)
PC ae C36:3
PC aa C34:4
PC aa C42:2
CE(20:4)
PC ae C32:2
PC aa C36:3
HexCer(d18:1/24:1)
PC ae C36:2
GDCA
TG(18:1_36:0)
GUDCA
SM C24:0
TG(20:4_36:3)
PC ae C40:1
PC aa C36:0
SM (OH) C22:2
SM (OH) C22:1
PC aa C36:6
Cer(d16:1/24:0)
High Stability Threshold (0.7)
0.00
0.25
0.50
0.75
1.00
Selection Frequency

F


Expression of Core Biomarkers
ns (p ≥ 0.05), * (p < 0.05), ** (p < 0.01), *** (p < 0.001)
Group_Factor
HC
SZ
ADMA
C0
C2
CE(16:1)
Cortisol
Cystine
PC aa C34:4
PC aa C40:2
PC ae C34:3
PC ae C38:2
Normalized Intensity
HC
SZ


A
ROC Curves — Five Models (Nested CV)
Sensitivity
1 - Specificity
Model
RF (AUC=0.919)
GBM (AUC=0.915)
kNN (AUC=0.896)
SVM (AUC=0.893)
LR (AUC=0.880)
B
RF Confusion Matrix: Training (CV) vs Test Set
TP=True Positive TN=True Negative FP=False Positive FN=False Negative
RF — Training Set (cross-validated)
Acc=84.6% Sens=92.1% Spec=70.2%
RF — Test Set
Acc=84.2% Sens=94.7% Spec=63.2%
33 (TN)
7 (FN)
14 (FP)
82 (TP)
12 (TN)
2 (FN)
7 (FP)
36 (TP)
Predicted
Actual
HC
SZ
Col Proportion
C
Precision-Recall Curve — RF
AUPRC = 0.969
Precision
Recall
D
Decision Curve Analysis — RF Model
RF Model
Treat All
Treat None
Net Benefit
Threshold Probability
E
Calibration Curve — Random Forest
RF Calibration Curve
Perfect Calibration
Hosmer-Lemeshow p = 0.567
Observed Rate
Predicted Probability
F
SHAP Beeswarm — RF
Color: feature value (blue=low, red=high)
CE(16:1)
C0
C2
PC ae C34:3
Cortisol
ADMA
PC ae C38:2
PC aa C34:4
PC aa C40:2
Cystine
SHAP Value
Feature value
High
Low
G
SHAP Dependence — All 10 Metabolites (RF)
ADMA
C0
C2
CE(16:1)
Cortisol
Cystine
PC aa C34:4
PC aa C40:2
PC ae C34:3
PC ae C38:2
SHAP Value
H
External Validation ROC - 10-Metabolite RF Panel
AUC = 0.863
95% CI: 0.724-1.000
Sensitivity (TPR)
1 - Specificity (FPR)
I
Expression of Core Biomarkers
ns (p ≥ 0.05), * (p < 0.05), ** (p < 0.01), *** (p < 0.001)
Group_Factor HC SZ
Normalized Intensity
ADMA
C0
C2
CE(16:1)
Cortisol
Cystine
PC aa C34:4
PC aa C40:2
PC ae C34:3
PC ae C38:2
ns
HC
SZ


A
129 Drug-naïve Schizophrenia
Longitudinal clinical loss to follow-up N=9
120 Drug-naïve Schizophrenia
PANSS deduction rate ≥ 50%
Treatment response group N = 58
Treatment non-response group N = 62
B
Metabolites ranked by permutation VIMP
Top 35 of 271 shown; 39 exceeded the threshold (n=120)
TG(18:1_35:3)
Cer(d18:2/18:0)
beta-Ala
lysoPC a C16:0
CE(16:0)
PC ae C42:1
TG(17:0_36:3)
TG(18:2_35:2)
CE(22:6)
Sarcosine
DG(18:2_18:2)
HexCer(d18:2/23:0)
PC aa C32:2
TG(16:1_32:1)
PC ae C42:2
TG(18:1_33:2)
SM C24:1
TG(18:2_38:5)
Arg
HexCer(d18:2/24:0)
TG(18:2_38:6)
Asp
DG(18:1_18:2)
C2
GLCAS
CE(20:3)
3-IAA
SM C18:1
HexCer(d18:1/22:0)
1-Met-His
TG(18:2_38:4)
PC ae C38:0
PC aa C40:4
PC aa C40:6
SM C20:2
Metabolite
Noise threshold = 0.000406
0.000
0.002
0.004
0.006
VIMP (permutation importance)
Selected
C
Univariate screening of 271 metabolites
P = 0.10
Selected (P<0.10)
Not selected
2.0
1.5
1.0
0.5
0.0
0.7
0.8
1.0
Odds ratio per 1 SD (univariate logistic)
D
Metabolite selection (Venn)
RF-VIMP > noise (n=39)
Univariate P<0.10 (n=21)
29
10
11
E
ROC of the 10-metabolite combined model
Baseline metabolites discriminating 8-week responders vs non-responders (n=120)
AUC = 0.767
Sensitivity
1 - Specificity
0.00
0.25
0.50
0.75
1.00

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Table.docx
- SupplementaryTables1104.docx