Received: 21 November 2024 | Revised: 24 January 2025 | Accepted: 2 February 2025

DOI: 10.1111/dom.16264

ORIGINAL ARTICLE

WILEY

# Machine learning approach on plasma proteomics identifies signatures associated with obesity in the KORA FF4 cohort

Jiefei Niu MM[1,2,3] | Jonathan Adam MSc[1,2,4] | Thomas Skurk MD[5,6] | Jochen Seissler MD[4,7] | Qiuling Dong MSc[1,2,3] | Esienanwan Efiong PhD[1,2,8,9] | Christian Gieger PhD[1,2,4] | Annette Peters PhD[2,4,10] | Sapna Sharma PhD[1,2,4] | Harald Grallert PhD[1,2,4]

[1]Research Unit of Molecular Epidemiology, Helmholtz Zentrum München, Neuherberg, Germany

[2]Institute of Epidemiology, Helmholtz Zentrum München, Neuherberg, Germany

[3]Faculty of Medicine, Ludwig-Maximilians-University München, Munich, Germany

[4]German Center for Diabetes Research (DZD), Neuherberg, Germany

[5]School of Medicine, Technical University of Munich, Munich, Germany

[6]ZIEL Institute for Food & Health, Core Facility Human Studies, Technical University of Munich, Freising, Germany

[7]Medizinische Klinik und Poliklinik IV, Klinikum der Ludwig-Maximilians-Universität, and Clinical Cooperation Group Diabetes, Ludwig-Maximilians-Universität München, and Helmholtz Zentrum München, Munich, Germany

[8]Faculty of Pharmaceutical, Biomedical and Veterinary Sciences, Department of Pharmaceutical Sciences, Campus Drie Eiken, Universiteitsplein 1, Antwerp, Belgium

[9]Department of Biochemistry, Faculty of Science, Federal University of Lafia, Lafia, Nigeria

[10]Chair of Epidemiology, Institute for Medical Information Processing, Biometry, and Epidemiology (IBE), Faculty of Medicine, Ludwig-Maximilians-University München, Munich, Germany

**Correspondence**
Jiefei Niu, Sapna Sharma, and Harald Grallert, Research Unit of Molecular Epidemiology, Helmholtz Zentrum München, Neuherberg 85764, Germany.
Email: jiefei.niu@helmholtz-munich.de, sapna.sharma@helmholtz-munich.de, and harald.grallert@helmholtz-munich.de

**Funding information**
Helmholtz Zentrum München – German Research Center for Environmental Health; University Hospital of Augsburg; Bavarian State Ministry of Health, Care and Prevention; German Federal Ministry of Education and Research (BMBF)

## Abstract

**Aims:** This study investigated the role of plasma proteins in obesity to identify predictive biomarkers and explore underlying biological mechanisms.

**Methods:** In the Cooperative Health Research in the Region of Augsburg (KORA) FF4 study, 809 proteins were measured in 2045 individuals (564 obese and 1481 non-obese). Multivariate logistic regression adjusted for confounders (basic and full models) was used to identify obesity-associated proteins. Priority-Lasso was applied for feature selection, followed by machine learning models (support vector machine [SVM], random forest [RF], k-nearest neighbour [KNN] and adaptive boosting [Adaboost]) for prediction. Correlation and enrichment analyses were performed to elucidate relationships between protein biomarkers, obesity risk factors and perturbed pathways. Mendelian randomisation (MR) assessed causal links between proteins and obesity.

**Results:** A total of 16 proteins were identified as significantly associated with obesity through multivariable logistic regression in the basic model and subsequent Priority-Lasso analysis. Enrichment analyses highlighted immune response, lipid metabolism and inflammation regulation were linked to obesity. Machine learning models

demonstrated robust predictive performance with area under the curves (AUC) of 0.820 (SVM), 0.805 (RF), 0.791 (KNN) and 0.819 (Adaboost). All 16 proteins correlated with obesity-related risk factors such as blood pressure and lipid levels. MR analysis identified AFM, CRP and CFH as causal and potentially modifiable proteins.

**Conclusions:** The protein signatures identified in our study showed promising predictive potential for obesity. These findings warrant further investigation to evaluate their clinical applicability, offering insights into obesity prevention and treatment strategies.

**KEYWORDS**
cohort study, machine learning, obesity, proteomics

## 1 | INTRODUCTION

The rate of obesity is rising globally.[1] Over the past four decades, the worldwide rate of obesity has surged from below 1% in 1975 to 6%–8% among girls and boys, from 3% to 11% among men and from 6% to 15% among women by 2016.[2] Additionally, there is growing concern that obesity rates have tripled among adults since the 1970s and continue to rise.[3] The World Health Organization (WHO) forecasts that by 2025, one in five adults worldwide will be obese. As obesity rates increase, its negative impact on health has become increasingly apparent, primarily due to higher mortality from noncommunicable diseases like atherosclerotic cardiovascular diseases (CVD), type 2 diabetes (T2D) and several cancers.[4] Moreover, apart from its consequences, the obesity epidemic poses significant health burdens on society.[5] Despite the increasing prevalence and related mortality of obesity, as well as significant advances in the use of molecular phenotyping technologies, for example, proteomics, there remains a need for deeper investigations to give more insights into disease pathophysiological mechanisms and biological sub-phenotypes for personalised medicine.

Given that proteins function as effectors of gene expression and their circulating levels are often modulated by genetic variation, making them a more accurate depiction of biological pathway activity than genetic or transcriptomic data, the proteome serves as a promising intermediary phenotype for uncovering innovative mechanisms underlying the progression of obesity.[6] Additionally, circulating proteins hold significant importance in drug development and are targets of pharmacological interventions.[7] Furthermore, proteomics provides valuable insights into post-translational protein modifications, protein–protein interactions (PPI) and signalling in obesity. Thus, developing high-throughput proteomic methodologies based on liquid chromatography-mass spectrometry (LC–MS) sets up a robust platform for biomarker discovery.[8] Non-targeted MS, specifically data-dependent acquisition (DDA) techniques, has been extensively used to identify prominent peptide ions, which can facilitate the discovery of key biomarkers in obesity.[9]

Obesity is a complex disease shaped by an interplay of genetic, environmental and lifestyle factors. Although many non-targeted proteomic studies have explored obesity, most population-based studies have been restricted by small sample sizes and the absence of reproducibility in both analysis and outcomes.[10–14] In contrast, our current study highlights the role of non-targeted proteomics in a large German-based population study, Cooperative Health Research in the Region of Augsburg (KORA) FF4 study.

In this KORA FF4 study, we used LC–MS/MS-based DDA proteomics to measure 809 proteins in 2045 individuals, with the goal of identifying biomarkers and elucidating underlying biological mechanisms associated with obesity.

## 2 | MATERIALS AND METHODS

### 2.1 | Study population design and sample collections

KORA study is a population-based cohort study. Details about the study population are presented in Methods section in Data S1. In FF4, 2132 individuals had phenotype and protein measurement data, and 87 were excluded from the analysis. This exclusion included 9 underweight participants (body mass index [BMI] <15 kg/m$^2$), 73 with missing covariate information and five samples were collected without at least 8 h of fasting. The final dataset consisted of 2045 participants, with 1481 classified as non-obese (BMI <30 kg/m$^2$) and 564 as obese (BMI ≥30 kg/m$^2$).

### 2.2 | Proteomics measurements

A detailed description of the proteomics measurements can be found in the Methods in Data S1.

### 2.3 | Covariates

The covariates detailed description can be found in Methods section in Data S1.

## 2.4 | Statistical analysis and bioinformatic analysis

### 2.4.1 | Baseline characteristics

Characteristics of the study population were reported as mean ± standard deviation (SD) or median (25th and 75th percentiles) for continuous variables according to normality, respectively, and as numbers (percentages) for categorical variables.

### 2.4.2 | Multivariable logistic regression analysis

The performed details of multivariable logistic regression analysis were described in the Methods section in Data S1.

### 2.4.3 | Feature selection and machine learning algorithms

We employed Priority-Lasso to address multicollinearity among the variables.[15] The conducted details of Priority-Lasso were described in the Methods section in Data S1. The models for obesity were constructed by four machine learning algorithms, including random forest (RF) ('*randomForest*' R package [version 4.7.1.1]),[16] support vector machine (SVM) ('*e1071*' R package [version 1.7.14]),[17] k-nearest neighbour (KNN) ('*kknn*' R package [version 1.3.1])[18] and adaptive boosting (Adaboost) ('*Adabag*' R package [version 5.0]).[19] We performed tenfold cross-validation by '*caret*' (version 6.0.94)[20] on the whole dataset to select the tuned parameters of different models. Then, the parameters were applied to the entire dataset to provide the final metrics of the suitability of the models for classifying individuals with obese and non-obese groups. Using the receiver operating characteristic (ROC) curves to measure models' predictive performance. Utilising the '*pROC*' R package (version 4.3.3) visualise and analyse ROC curves.[21] The area under the curve (AUC) and 95% bootstrap confidence intervals (CI) were also estimated. To model explainability, SHapley Additive exPlanations (SHAP) was utilised in the RF model, using SHAP values were estimated for the 16 obesity-associated protein features.[22]

### 2.4.4 | Association between protein biomarkers and obesity complication risk factors

To understand the correlation between 16 obesity biomarkers and 13 obesity complication risk factors, we performed Spearman's rank correlation and visualised using the '*corrplot*' R package (version 0.92).[23,24] The correlations between BMI and related indicators were calculated separately. Besides, the Mouse Genome Informatics (MGI) database was employed to investigate the expression of 16 obesity-associated proteins using molecular techniques.[25]

### 2.4.5 | Enrichment analyses and protein–protein interaction

Gene ontology (GO) analysis, Kyoto Encyclopedia of Genes and Genomes (KEGG) enrichment analysis of differential expression genes (DEGs) in the key proteins mentioned above were performed by R package '*clusterProfiler*' (Version 4.0.5).[26] The DEGs were classified into functional groups based on molecular functions, biological processes and cellular compartments. All *p* values of the enrichment analyses were corrected for multiple hypothesis tests using Bonferroni correction. PPI network was constructed via the STRING database,[27] setting a confidence score threshold of 0.4 for interactions. Other parameters remained at their default settings. The PPI results were analysed and visualised via Cytoscape (version 3.10.1).[28]

### 2.4.6 | Two-sample Mendelian randomisation

A two-sample Mendelian randomisation (MR) approach was used to explore possible causal interactions among proteins linked to obesity and obesity.[29,30] Details on genome-wide association studies (GWAS) selection are in the Methods section in Data S1. Instrumental variables (IVs) were chosen based on genome-wide significance ($p < 5 \times 10^{-8}$), and subsequent clumping of chosen IVs was performed with $r^2 < 0.001$. Causal estimates were calculated utilising Wald or inverse variance weighted tests, for more than one IV. When more than two IVs were included, Cochran's heterogeneity test was utilised to assess heterogeneity. For three or more IVs that were included, the MR Egger method was employed to evaluate if they had horizontal pleiotropy.[31] IVs were selected for 12 out of 16 obesity-associated proteins. To determine the significance threshold for MR association, we applied Bonferroni correction considering the number of independent sets of IVs tested. For the identified significant links, we also applied opposite direction MR, with obesity IVs as the exposure and proteins IVs as the outcome, to explore possible reverse causality. All analyses were conducted by R package '*TwoSampleMR*' (version 0.5.7).[32]

All analyses were conducted using Python (version 3.8.5), R statistics (version 4.3.3) and RStudio (version 2023.09.1+ 494).

# 3 | RESULTS

## 3.1 | Characteristics of the KORA FF4 Participants

Our overall ($n = 2045$) participants were stratified into non-obese (BMI <30 kg/m$^2$) and obese (BMI ≥30 kg/m$^2$) groups based on their BMI. As indicated in Table 1, a comparison between stratified groups showed smoking status, physical activity, age, weight, height, waist, waist-hip-ratio (WHR), triglycerides (TG), systolic blood pressure (SBP), diastolic blood pressure (DBP), high-density lipoprotein (HDL), triglycerides (TG), C-reactive protein (CRP), body fat percentage (BFP) and T2D status were significantly different. For alcohol consumption, sex, low-

**TABLE 1** The characteristics of the Cooperative Health Research in the Region of Augsburg FF4 participants were analysed based on their body mass index (BMI).

| Characteristic | Overall | Non-obese (BMI <30 kg/m$^2$) | Obese (BMI ≥30 kg/m$^2$) | *p*-Value |
|---|---|---|---|---|
| Sample size | 2045 | 1481 | 564 | |
| Age (years) | 60.1 ± 12.3 | 59.3 ± 12.3 | 62 ± 12 | <0.001 |
| Sex woman (%) | 1042 (51) | 766 (51.7) | 276 (48.9) | 0.282 |
| Weight (kg) | 79.5 ± 16.5 | 73.2 ± 11.8 | 96.1 ± 15.5 | <0.001 |
| Height (cm) | 168.9 ± 9.6 | 169.4 ± 9.6 | 167.7 ± 9.6 | <0.001 |
| Alcohol consumption (g/day) | 14.9 ± 20.2 | 15.2 ± 19 | 14.1 ± 23.3 | 0.288 |
| WC (cm) | 96.8 ± 14.2 | 91.1 ± 10.6 | 111.9 ± 10.8 | <0.001 |
| WHR | 0.9 ± 0.1 | 0.9 ± 0.1 | 1 ± 0.1 | <0.001 |
| BFP (%) | 32.8 ± 7.2 | 30.8 ± 6.5 | 38 ± 6.3 | <0.001 |
| FPG (mmol/L) | 5.7 ± 1.2 | 5.5 ± 0.9 | 6.2 ± 1.6 | <0.001 |
| SBP (mmHg) | 118.9 ± 17.5 | 117.4 ± 17 | 122.9 ± 18.2 | <0.001 |
| DBP (mmHg) | 73 ± 9.6 | 72.4 ± 9.3 | 74.6 ± 10.3 | <0.001 |
| Smoking (%) | | | | |
| Smoker | 866 (42.3) | 589 (39.8) | 277 (49.1) | <0.001 |
| Ex-smoker | 864 (42.2) | 637 (43) | 227 (40.2) | |
| Never-smoker | 315 (15.4) | 255 (17.2) | 60 (10.6) | |
| Physical activities inactive (%) | 878 (42.9) | 560 (37.8) | 318 (56.4) | <0.001 |
| HDL cholesterol (mmol/L) | 1.7 ± 0.5 | 1.8 ± 0.5 | 1.5 ± 0.4 | <0.001 |
| LDL cholesterol (mmol/L) | 3.5 ± 0.9 | 3.5 ± 0.9 | 3.5 ± 0.9 | 0.12 |
| TG (mmol/L) | 1.2 (0.9, 1.6) | 1.1 (0.8, 1.5) | 1.4 (1.1, 2.1) | <0.001 |
| TCHO (mmol/L) | 5.6 ± 1 | 5.6 ± 1 | 5.5 ± 1 | 0.124 |
| CRP (mg/L) | 2.4 ± 4.5 | 1.8 ± 3.7 | 4.1 ± 5.9 | <0.001 |
| T2D status (%) | 272 (13.3) | 123 (8.3) | 149 (26.4) | <0.001 |

*Note*: Quantitative variables are expressed as mean ± SD or median (25th and 75th percentiles); categorical variables are expressed as *n* (%). Statistical analysis was performed to evaluate the significant difference between obese and non-obese participants. Test statistics for categorical variables were calculated via the $\chi^2$ test and Student's *t* test for continuous variables.

Abbreviations: BFP, body fat per cent; CRP, C-reactive protein; DBP, diastolic blood pressure; FPG, fasting plasma glucose; HDL, high-density lipoprotein; LDL, low-density lipoprotein; SBP, systolic blood pressure; T2D, type 2 diabetes; TCHO, total cholesterol; TG, triglyceride; WC, waist circumstance; WHR, waist-hip ratio.

density lipoprotein (LDL) and total cholesterol (TCHO), no significant differences were detected among obese and non-obese groups.

## 3.2 | Proteins associated with obesity

A logistic regression model was employed for obesity associated with proteins. Model assumptions for logistics have been performed and reported (Table S2). The basic model added age and sex, resulting in 25 proteins significantly associated with obesity (Table S3; Figure 1A). Subsequently, we tested how covariates such as lipids, FPG with T2D status and lifestyle influenced the number of significant proteins linked with obesity. More covariates were included, and less significant proteins were reported. Especially the number of significant proteins was influenced by lipids, which showed a dramatic drop when adjusted lipids covariates in the model (Figure S1).

To further explore specific obesity-associated proteins, we performed a logistic regression analysis with the abovementioned obesity-related variables as covariates. In the full model, 11 proteins were observed to have significant associations after conservative Bonferroni correction (Figure 1B, Tables S4, and S5).

APOD, CRP and LGALS3BP showed significant associations with obesity in the basic model (Table S3). Subsequently, in the full model, CRP, APOD and SAA2-SAA4 emerged as the most specific significant proteins (Table S5). We further investigated the significance of these proteins by Student's *t* tests in both obese and non-obese individuals. Figure 1C illustrates that all proteins of the basic and full models mentioned above remained statistically significant in non-obese and obese groups.

## 3.3 | Identification and validation of candidate protein biomarkers

After conducting priority-Lasso feature selection, the 16 selected proteins were regarded as obesity candidate biomarkers (Table S6).

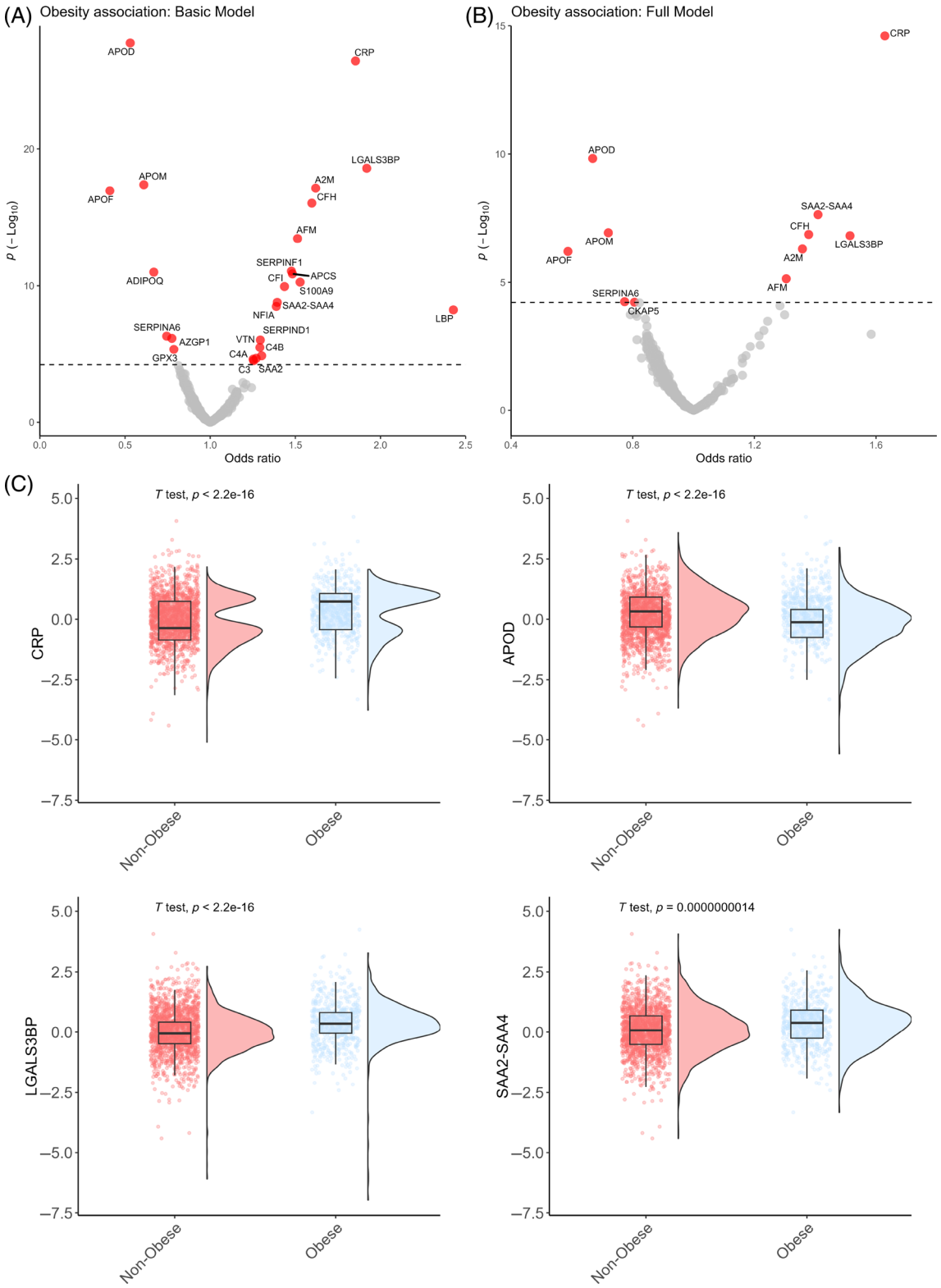


**FIGURE 1** Each dot represents a protein, and they are displayed based on the OR (*x*-axis) and the negative logarithm (base 10) of the *p* value (*y*-axis); Bonferroni correction *p* value cut-off is 0.05/809 = 0.0000618 was considered. The covariates for the basic model are age, sex and (obesity); the covariates for the full model are age, sex, (obesity), physical activities, systolic blood pressure, naturally log-transformed triglycerides, high-density lipoprotein-C, smoking status, fasting plasma glucose, type 2 diabetes status (A, B). (A) Volcano plot shows the association of proteins with obesity in the basic model; (B) Volcano plot shows the association of proteins with obesity in the full model; (C) Raincloud plots show the top three significant proteins in both the basic and full model, Student's *t* tests were utilised to compared in obese and non-obese groups.

**TABLE 2** Diagnostic performance of four machine learning algorithms.

| Machine learning algorithms | Models | Sensitivity | Specificity | Non-error rate | AUC (95% CI) |
|---|---|---|---|---|---|
| RF | Proteins | 0.804 | 0.708 | 0.837 | 0.820 (0.783–0.857) |
| | Risk factors | 0.764 | 0.564 | 0.868 | 0.744 (0.701–0.788) |
| SVM | Proteins | 0.796 | 0.721 | 0.854 | 0.805 (0.765–0.844) |
| | Risk factors | 0.749 | 0.604 | 0.919 | 0.750 (0.706–0.793) |
| KNN | Proteins | 0.817 | 0.623 | 0.780 | 0.791 (0.752–0.831) |
| | Risk factors | 0.762 | 0.459 | 0.812 | 0.683 (0.636–0.731) |
| Adaboost | Proteins | 0.830 | 0.708 | 0.797 | 0.819 (0.782–0.855) |
| | Risk factors | 0.785 | 0.602 | 0.834 | 0.747 (0.704–0.791) |

Abbreviations: Adaboost, adaptive boosting; AUC, area under the curve; KNN, k-nearest neighbour; RF, random forest; SVM, support vector machine.

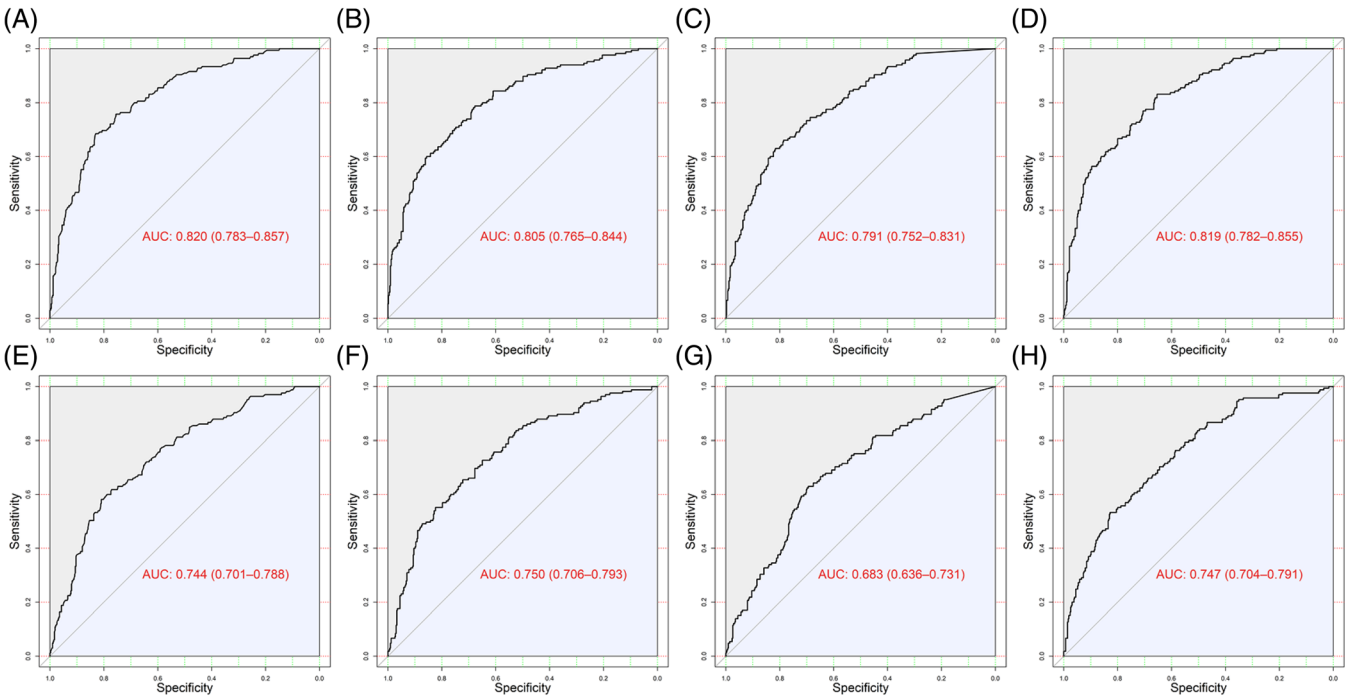


**FIGURE 2** Area under the receiver operating characteristic curves of four machine learning algorithms. (A) Random forest model based on protein biomarkers; (B) random forest model based on clinical covariates; (C) support vector machine model based on protein biomarkers; (D) support vector machine model based on clinical covariates; (E) k-nearest neighbour model based on protein biomarkers; (F) k-nearest neighbour model based on clinical covariates; (G) adaptive boosting model based on protein biomarkers; (H) adaptive boosting model based on clinical covariates. AUC, area under the curve.

Prediction models constructed according to 16 protein biomarkers were using four machine-learning algorithms: RF, SVM, KNN and AdaBoost. The model parameters were optimised through tenfold cross-validation on the KORA FF4 data. Lastly, trees number in the RF model was 500; for each decision tree, the number of features considered in splitting nodes is 5. In SVM, the Radial kernel was utilised and the sigma number was $1/n$ ($n$ = protein numbers). The KNN model was configured with nine neighbours. The iterations in AdaBoost were 150, and the decision maximum depth tree was 3. Subsequently, the diagnostic models incorporating clinical covariates, including age, sex, physical activity, SBP, naturally log-transformed TG, HDL-C, smoking status, FPG and T2D status were also constructed to compare the predictive performance with those of the protein biomarker-based model. The detailed results of these eight models are reported below (Table 2 and Figure 2), AUCs ranged from 0.683 to 0.820. Especially, the AUCs of proteins models ranged from 0.791 to 0.820, higher than each risk factors model ranging from 0.683 to 0.750. Besides, for model explainability, we used Tree SHAP algorithms to generate an importance ranking that explains the output of the RF model based on SHAP values estimated for the 16 obesity-associated protein features (Figure S2). Proteins such as LGALS3BP, CRP and APOD exhibit the highest SHAP values, indicating a strong contribution to obesity prediction. These proteins play key roles in pathways associated with obesity, such as inflammation and lipid metabolism.

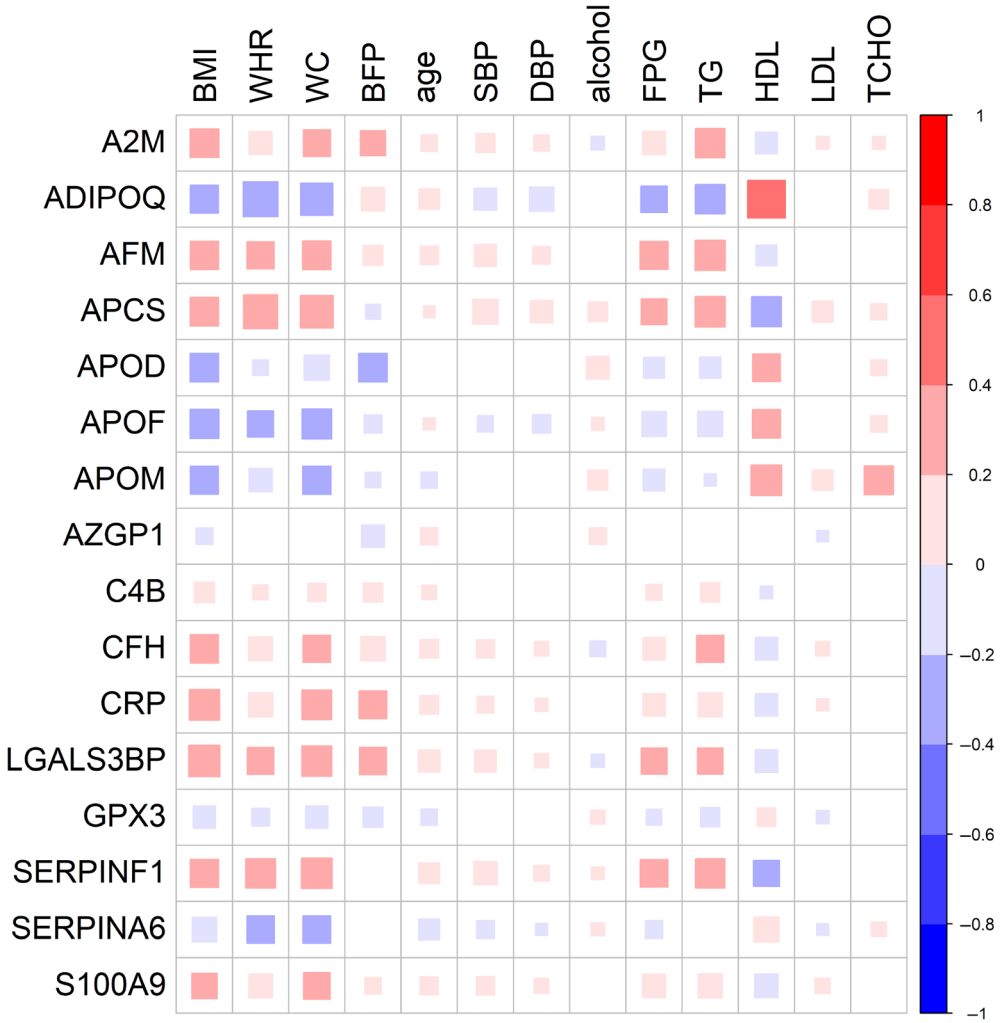


**FIGURE 3** The correlation matrix between 16 potential proteins and 13 obesity risk factors is depicted. Statistically significant correlations between two proteins are indicated, while insignificant *r* is left blank in the boxes. Positive correlations are denoted by red, while negative correlations are represented by blue. BFP, body fat per cent; BMI, body mass index; DBP, diastolic blood pressure; FPG, fasting plasma glucose; SBP, systolic blood pressure; TCHO, total cholesterol; TG, triglycerides; WC, waist circumstance; WHR, waist-hip ratio. $p < 0.05$ is considered statistically significant. The detailed *r* and *p* values are shown in Tables S7 and S8.

### 3.4 | Association between proteins and obesity complication risk factors

Spearman's correlation coefficient (*r*) was computed to examine the possible associations between 16 protein biomarkers and 13 obesity risk factors (Table S8). The resulting *r* matrix is visualised in Figure 3. All 16 proteins were significantly correlated with BMI, and 15 were correlated with WHR, waist circumstance (WC), age, FPG and HDL. Moreover, 14 proteins were associated with TG, 11 proteins were associated with SBP, DBP, alcohol consumption and nine proteins were associated with LDL, seven proteins were associated with TCHO (Table S7). Additionally, we can see the direction of 16 proteins associated with BMI is the same in FPG and TG but not in HDL. For example, ADIPOQ, APOD, APOF and APOM were observed to be negatively associated with HDL. The significant *r* ranged from −0.385 to 0.451 (positive coefficients are in red and negatives are in blue) (Figure 3). Notably, APCS and A2M demonstrated the strongest correlation with all 13 obesity risk factors. The *r* of HDL and ADIPOQ exhibited the highest magnitude ($r = 0.451$, *p* value < 0.001). There were seven proteins correlated with TCHO (*p* values < 0.05), with the *r* ranging from −0.033 to 0.270, indicating relatively weak associations. APOF and CFH were linked to 10 obesity risk factors (Figure 3 and Table S7). In the KORA FF4 cohort, BMI showed strong correlation with WC ($r = 0.865$) and moderate correlations with WHR ($r = 0.542$) and BFP ($r = 0.501$), all highly significant ($p < 0.001$), underscoring its association with abdominal fat distribution and overall adiposity (Tables S7 and S8). All 16 proteins were expressed in relevant tissues such as plasma, fat, liver and immune system (more details in Results section in Data S1).

### 3.5 | Enrichment analyses and protein–protein interaction

Sixteen GO terms were statistically significant when using 16 protein biomarkers (Table S9, Figure S3). The enriched GO terms were characterised by processes relating to the response of complement regulation, lipid lipoprotein particles and antioxidant cellular detoxification (Figure S3A). The top three significant GO terms were GO:0072562 (blood microparticle), GO:0001848 (complement binding) and GO:0062023 (collagen-containing extracellular matrix) (Figure S3B). In the KEGG pathway analysis, complement and coagulation cascades

and staphylococcus aureus infection two pathways were statistically significant in two groups (Table S10). In the PPI network depicted in Figure S4, all 16 proteins consisted of 16 nodes and 40 edges, with APCS being associated with nine other proteins, the highest degree observed.

### 3.6 | Mendelian randomisation causally implicates proteins

For the 16 top candidate proteins strongly linked to obesity, a two-sample MR approach was performed to evaluate the potential causal relations between these proteins and obesity. IVs for MR were derived from protein quantitative trait loci identified in the Icelandic studies and the INTERVAL study.[33,34] In the causal direction that proteins lead to obesity, after checking in the instrument dataset, only 12 proteins had available single nucleotide polymorphisms (SNPs) for further analysis. The Bonferroni correction significance $p$ value was adjusted to $0.05/12 = 0.004$. However, after multiple corrections, no significant associations were observed for any of the proteins (Table S11). For the opposite direction that obesity leads to proteins, six proteins could find SNPs in exposure data, and the Bonferroni correction significance $p$ value was adjusted to $0.05/6 = 0.0083$. Our results indicated that change in obesity caused changes in proteins such as AFM ($\beta = 3.06$, $p$ value < 0.001), CRP ($\beta = 9.96$, $p$ value < 0.001) and CFH ($\beta = 0.06$, $p$ value = 0.001) (Table S11). Sensitivity analysis was further conducted to assess the robustness of the results against heterogeneity or horizontal pleiotropy. For the direction of BMI to AFM and CRP, sensitivity analysis wasn't performed because each protein only obtained one SNP. For CFH, the $Q$ statistic from the heterogeneity test ($p$_Het = 0.78) indicated no heterogeneity. Additionally, the MR–Egger intercept test ($p$_Pleio = 0.53) also suggested no directional pleiotropy for this protein.

## 4 | DISCUSSION

To investigate altered protein profiles associated with obesity, we utilised LC–MS/MS-based DDA proteomics, identifying a total of 16 proteins significantly associated with obesity. Notably, all 16 proteins have been previously reported implicated in obesity.[10,35–37] Enrichment analysis of these biomarkers revealed disruptions in lipid metabolism, immune response and inflammation regulation in obesity. We applied four machine learning algorithms, including SVM, RF, KNN and Adaboost, to develop obese predictive models, with AUCs ranging from 0.791 to 0.820, surpassing classical obesity risk factors. All 16 proteins correlated with obesity-related risk factors such as blood pressure (BP) and lipid levels. MR analysis provided suggestive evidence that obesity can lead to changes in AFM, CRP and CFH.

To our knowledge, our study is the first to comprehensively identify protein biosignatures of obesity using LC–MS/MS-based proteome profiling in a large, well-established German cohort. Our findings underscore the potential of the proteome as a valuable resource for identifying biomarkers of obesity, which could be utilised for both prevention and treatment. These protein biosignatures provide critical insights into the metabolic pathways disrupted in obesity and its progression toward associated complications, such as T2D and CVD. Consequently, obesity-associated proteins and their metabolic patterns have the potential to serve as predictive models for risk stratification in populations and to inform targeted interventions for obesity and its comorbidities.

Our results also revealed significant differences between obese and non-obese participants across various demographic and clinical factors such as smoking status, physical activity and others (Table 1). These differences highlight the multifactorial nature of obesity and its associated factors. Furthermore, these 16 proteins for obesity also presented significant correlations with obesity indexes and complications associated factors (Figure 3), including BMI, WHR, WC, BFP, age, SBP, DBP, alcohol consumption, FPG, TG, HDL, LDL and TCHO (Table 1). Given that obesity is closely linked to multiple related risk factors, we reasonably hypothesise that these biomarkers may also serve as possible biomarkers for fatty liver, coronary heart diseases, hyperglycaemia and dyslipidaemia.

The results of enrichment analyses can be summarised into three main processes: immune response, lipid metabolism and inflammation regulation. Innate and adaptive immunity dysregulation, resulting in chronic, low-level, tissue-specific and systemic inflammation, plays a key role in contributing to the development of multiple obesity-associated disorders and metabolic diseases.[38] Dyslipidaemia affects approximately 60%–70% of patients with obesity,[39] and is also verified as a characteristic of obesity and CVD. In the obese state, hypertrophic adipocytes secrete increased levels of proinflammatory adipokines and free fatty acids, leading to dyslipidaemia, inflammation and ectopic fat accumulation.[40] Since obesity is an inflammatory condition,[41] adipose expansion and chronic obesity trigger an early inflammatory response, leading to a lasting change in the immune system toward a proinflammatory state.[42] All the functions and pathways associated with obesity interact with each other and play significant roles in the onset, progression and prognosis of obesity.

MR analyses supported that BMI could lead to AFM, CRP and CFH alterations in relative protein levels. In the opposite direction, no protein can lead to BMI changes. Afamin (AFM) is the fourth member of the human albumin gene family, encompassing metabolic syndrome and associated diseases such as obesity, T2D, hypertension and dyslipidaemia.[43] Transgenic mice with AFM overexpression exhibited gained body weight, lipid and glucose levels, and meta-analysis of the population-based Bruneck ($n = 826$), Salzburg Atherosclerosis Prevention Program in Subjects at High Individual Risk ($n = 1499$) and KORA F4 studies ($n = 3060$) indicated high serum AFM levels were positively related with metabolic syndrome components, including obesity and BMI.[44] Obesity is marked by persistent low-grade inflammation. CRP, a key indicator of systemic inflammation, has consistently emerged as the primary factor linked to overweight and obesity in human epidemiological studies.[45] Statistical analyses suggest that elevated CRP levels are a result of obesity rather than a cause. Our

study results further corroborate this perspective. Human complement factor H (CFH) protects cells from unintended complement system damage and is linked to metabolic disruptions in obesity. Higher CFH levels are associated with increased BMI, waist circumference, triglycerides and inflammation, indicating its potential role in insulin resistance and metabolic issues in obesity.[46]

Our study offers several significant advantages. Firstly, we leveraged a well-characterised, population-based cohort that allows for adjustments across various demographic and clinical parameters, enhancing the robustness of our findings. Rigorous quality control measures were applied to proteome profiles collected across three time points, effectively minimising measurement noise and increasing the reliability of the data. We used multivariable logistic regression in the basic model, followed by Priority-Lasso feature selection, machine learning algorithms and correlation analyses to validate and confirm the candidate protein biomarkers associated with obesity. These complementary approaches increase our results' confidence and potential clinical relevance. Although the study has several strengths, there are also limitations. We used tenfold cross-validation to optimise predictive accuracy, but external replication studies are still needed to further confirm our findings. Additionally, due to the study's observational nature, we are unable to delve into the complex mechanisms behind these results using animal models or cellular experiments. Furthermore, as the KORA cohort primarily consists of a German population, future studies should aim to validate these findings in more diverse populations and consider additional obesity-related indicators, such as WHR and BFP, to enhance the generalisability of our results. The identified protein biomarkers present promising opportunities to revolutionise obesity research and inform targeted prevention and treatment strategies, although challenges such as standardisation and broader validation remain to be addressed for clinical implementation.

## 5 | CONCLUSIONS

In summary, we conducted an LC–MS/MS-based protein analysis, identifying a total of 16 proteins associated with obesity. Enrichment analyses revealed that these proteins are involved in lipid metabolism, immune response and inflammation regulation. Machine learning models based on these biomarkers achieved higher AUCs for obesity prediction than traditional risk factors models. Furthermore, these proteins demonstrated significant correlations with obesity-related factors, such as BP, and lipid levels. MR analysis provided suggestive evidence that obesity causally influences the levels of AFM, CRP and CFH. Although these biomarkers have been previously reported, our study is the first to establish a direct link between them and obesity in a large-scale German population-based cohort. These obesity-associated biomarkers present opportunities for clinical diagnosis and personalised treatment strategies. Future research should focus on exploring the underlying mechanisms of these biomarkers and assessing their clinical applicability across diverse populations, paving the way for more tailored obesity management approaches.

## AUTHOR CONTRIBUTIONS

Conceptualisation: J.N., S.S., H.G., A.P., C.G., T.S., J.S. and Q.D. Methodology: J.N., S.S. and Q.D. Software, validation, formal analysis and writing original draft preparation: J.N. Investigation: J.N. and S.S. Resources and funding acquisition: C.G., A.P. and H.G. Data curation: C.G., A.P. and H.G. Data: Q.C. and J.A. Writing review and editing: J.N., S.S., H.G., C.G., J.A., T.S., J.S., Q.D., E.E. and A.P. Visualisation: J.N. and Q.D. Supervision: S.S. and H.G. Project administration: H.G. All authors have read and agreed to the published version of the manuscript.

## ACKNOWLEDGEMENTS

We thank all participants for their long-term commitment to the KORA study, the staff for data collection and research data management, and the members of the KORA Study Group (https://www.helmholtz-munich.de/en/epi/cohort/kora) who are responsible for the study's design and conduct. We also would like to thank the China Scholarship Council (CSC) for its financial support. Open Access funding enabled and organized by Projekt DEAL.

## FUNDING INFORMATION

The KORA study was initiated and financed by Helmholtz Zentrum München-German Research Center for Environmental Health, which is financed by the German Federal Ministry of Education and Research (BMBF) and by the State of Bavaria. Data collection in the KORA study is done in cooperation with the University Hospital of Augsburg. This study was funded by the Bavarian State Ministry of Health, Care and Prevention through the research project DigiMed Bayern (www.digimed-bayern.de).

## CONFLICT OF INTEREST STATEMENT

The authors declare no conflicts of interest.

## PEER REVIEW

The peer review history for this article is available at https://www.webofscience.com/api/gateway/wos/peer-review/10.1111/dom.16264.

## DATA AVAILABILITY STATEMENT

The KORA and the proteomics datasets are not publicly available but can be accessed upon application through the KORA-PASST use and access hub subject to KORA Board approval (https://helmholtz-muenchen.managed-otrs.com/external/).

## ORCID

*Jiefei Niu* https://orcid.org/0000-0003-4104-1426
*Qiuling Dong* https://orcid.org/0000-0002-3369-4120

## REFERENCES

1. Roberto CA, Swinburn B, Hawkes C, et al. Patchy progress on obesity prevention: emerging examples, entrenched barriers, and new thinking. *Lancet*. 2015;385(9985):2400-2409. doi:10.1016/S0140-6736(14)61744-X

2. Jaacks LM, Vandevijvere S, Pan A, et al. The obesity transition: stages of the global epidemic. *Lancet Diabetes Endocrinol*. 2019;7(3):231-240. doi:10.1016/S2213-8587(19)30026-9
3. Collaboration NCDRF. Trends in adult body-mass index in 200 countries from 1975 to 2014: a pooled analysis of 1698 population-based measurement studies with 19.2 million participants. *Lancet*. 2016;387(10026):1377-1396. doi:10.1016/S0140-6736(16)30054-X
4. Prospective Studies C, Whitlock G, Lewington S, et al. Body-mass index and cause-specific mortality in 900 000 adults: collaborative analyses of 57 prospective studies. *Lancet*. 2009;373(9669):1083-1096. doi:10.1016/S0140-6736(09)60318-4
5. Swinburn BA, Sacks G, Hall KD, et al. The global obesity pandemic: shaped by global drivers and local environments. *Lancet*. 2011;378(9793):804-814. doi:10.1016/S0140-6736(11)60813-1
6. Shah AM, Myhre PL, Arthur V, et al. Large scale plasma proteomics identifies novel proteins and protein networks associated with heart failure development. *Nat Commun*. 2024;15(1):528. doi:10.1038/s41467-023-44680-3
7. Swinney DC, Anthony J. How were new medicines discovered? *Nat Rev Drug Discov*. 2011;10(7):507-519. doi:10.1038/nrd3480
8. Kang C, Lee Y, Lee JE. Recent advances in mass spectrometry-based proteomics of gastric cancer. *World J Gastroenterol*. 2016;22(37):8283-8293. doi:10.3748/wjg.v22.i37.8283
9. Doerr A. Mass spectrometry-based targeted proteomics. *Nat Methods*. 2013;10(1):23. doi:10.1038/nmeth.2286
10. Geyer PE, Wewer Albrechtsen NJ, Tyanova S, et al. Proteomics reveals the effects of sustained weight loss on the human plasma proteome. *Mol Syst Biol*. 2016;12(12):901. doi:10.15252/msb.20167357
11. Sahebekhtiari N, Saraswat M, Joenväärä S, et al. Plasma proteomics analysis reveals dysregulation of complement proteins and inflammation in acquired obesity—a study on rare BMI-discordant monozygotic twin pairs. *Proteomics Clin Appl*. 2019;13(4):1800173. doi:10.1002/prca.201800173
12. Garrison CB, Lastwika KJ, Zhang Y, Li CI, Lampe PD. Proteomic analysis, immune dysregulation, and pathway interconnections with obesity. *J Proteome Res*. 2017;16(1):274-287. doi:10.1021/acs.jproteome.6b00611
13. Al-Daghri NM, Manousopoulou A, Alokail MS, et al. Sex-specific correlation of IGFBP-2 and IGFBP-3 with vitamin D status in adults with obesity: a cross-sectional serum proteomics study. *Nutr Diabetes*. 2018;8(1):54. doi:10.1038/s41387-018-0063-8
14. Doumatey AP, Zhou J, Zhou M, Prieto D, Rotimi CN, Adeyemo A. Proinflammatory and lipid biomarkers mediate metabolically healthy obesity: a proteomics study. *Obesity*. 2016;24(6):1257-1265. doi:10.1002/oby.21482
15. Klau S, Jurinovic V, Hornung R, Herold T, Boulesteix AL. Priority-lasso: a simple hierarchical approach to the prediction of clinical outcome using multi-omics data. *BMC Bioinformatics*. 2018;19(1):322. doi:10.1186/s12859-018-2344-6
16. Liaw A, Wiener M. Classification and regression by randomForest. *R News*. 2002;2(3):18-22.
17. Chang CC, Lin CJ. LIBSVM: a library for support vector machines. *ACM Trans Intell Syst Technol*. 2011;2(3):1-27.
18. Hechenbichler K, Schliep K. *Weighted k-Nearest-Neighbor Techniques and Ordinal Classification*. Discussion Paper 399, SFB 386. Ludwig-Maximilians University of Munich; 2004 https://epub.ub.uni-muenchen.de/1769/1/paper_399.pdf
19. Alfaro E, Gamez M, Garcia N. Adabag: an R package for classification with boosting and bagging. *J Stat Softw*. 2013;54:1-35.
20. Kuhn M. Caret: classification and regression training. 2015. Astrophysics Source Code Library. ascl: 1505.003.
21. Robin X, Turck N, Hainard A, et al. pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics*. 2011;12:1-8.
22. Scott M, Su-In L. A unified approach to interpreting model predictions. *Adv Neural Inf Process Syst*. 2017;30:4765-4774.
23. Corrgrams FM, Friendly M. Corrgrams: exploratory displays for correlation matrices. *Am Stat*. 2002;56(4):316-324.
24. Murdoch DJ, Chow E. A graphical display of large correlation matrices. *Am Stat*. 1996;50(2):178-180.
25. Baldarelli RM, Smith CM, Finger JH, et al. The mouse gene expression database (GXD): 2021 update. *Nucleic Acids Res*. 2021;49(D1):D924-d931. doi:10.1093/nar/gkaa914
26. Wu T, Hu E, Xu S, et al. clusterProfiler 4.0: a universal enrichment tool for interpreting omics data. *Innovation (Camb)*. 2021;2(3):100141. doi:10.1016/j.xinn.2021.100141
27. Szklarczyk D, Gable AL, Nastou KC, et al. The STRING database in 2021: customizable protein-protein networks, and functional characterization of user-uploaded gene/measurement sets. *Nucleic Acids Res*. 2020;49(D1):D605-D612. doi:10.1093/nar/gkaa1074
28. Shannon P, Markiel A, Ozier O, et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res*. 2003;13(11):2498-2504. doi:10.1101/gr.1239303
29. Davey Smith G, Ebrahim S. 'Mendelian randomization': can genetic epidemiology contribute to understanding environmental determinants of disease? *Int J Epidemiol*. 2003;32(1):1-22. doi:10.1093/ije/dyg070
30. Hartwig FP, Davies NM, Hemani G, Davey Smith G. Two-sample Mendelian randomization: avoiding the downsides of a powerful, widely applicable but potentially fallible technique. *Int J Epidemiol*. 2016;45(6):1717-1726. doi:10.1093/ije/dyx028
31. Bowden J, Smith GD, Burgess S. Mendelian randomization with invalid instruments: effect estimation and bias detection through egger regression. *Int J Epidemiol*. 2015;44(2):512-525. doi:10.1093/ije/dyv080
32. Hemani G, Zhengn J, Elsworth B, et al. The MR-base platform supports systematic causal inference across the human phenome. *Elife*. 2018;7:e34408. doi:10.7554/eLife.34408
33. Ferkingstad E, Sulem P, Atlason BA, et al. Large-scale integration of the plasma proteome with genetics and disease. *Nat Genet*. 2021;53(12):1712-1721. doi:10.1038/s41588-021-00978-w
34. Sun BB, Maranville JC, Peters JE, et al. Genomic atlas of the human plasma proteome. *Nature*. 2018;558(7708):73-79. doi:10.1038/s41586-018-0175-2
35. Aleksandrova K, Egea Rodrigues C, Floegel A, Ahrens W. Omics biomarkers in obesity: novel etiological insights and targets for precision prevention. *Curr Obes Rep*. 2020;9(3):219-230. doi:10.1007/s13679-020-00393-y
36. Rodriguez-Munoz A, Motahari-Rad H, Martin-Chaves L, et al. Correction: a systematic review of proteomics in obesity: unpacking the molecular puzzle. *Curr Obes Rep*. 2024;13(3):439. doi:10.1007/s13679-024-00575-y
37. Zaghlool SB, Sharma S, Molnar M, et al. Revealing the role of the human blood plasma proteome in obesity using genetic drivers. *Nat Commun*. 2021;12(1):1279. doi:10.1038/s41467-021-21542-4
38. Shaikh SR, Beck MA, Alwarawrah Y, MacIver NJ. Emerging mechanisms of obesity-associated immune dysfunction. *Nat Rev Endocrinol*. 2024;20(3):136-148. doi:10.1038/s41574-023-00932-2
39. Nussbaumerova B, Rosolova H. Obesity and dyslipidemia. *Curr Atheroscler Rep*. 2023;25(12):947-955. doi:10.1007/s11883-023-01167-2
40. Choi SH, Hong ES, Lim S. Clinical implications of adipocytokines and newly emerging metabolic factors with relation to insulin resistance and cardiovascular health. *Front Endocrinol Lausanne*. 2013;4:97. doi:10.3389/fendo.2013.00097
41. Dandona P, Aljada A, Bandyopadhyay A. Inflammation: the link between insulin resistance, obesity and diabetes. *Trends Immunol*. 2004;25(1):4-7. doi:10.1016/j.it.2003.10.013
42. Saltiel AR, Olefsky JM. Inflammatory mechanisms linking obesity and metabolic disease. *J Clin Invest*. 2017;127(1):1-4. doi:10.1172/JCI92035
43. Dieplinger H, Dieplinger B. Afamin–a pleiotropic glycoprotein involved in various disease states. *Clin Chim Acta*. 2015;446:105-110. doi:10.1016/j.cca.2015.04.010

44. Kronenberg F, Kollerits B, Kiechl S, et al. Plasma concentrations of afamin are associated with the prevalence and development of metabolic syndrome. *Circ Cardiovasc Genet*. 2014;7(6):822-829. doi:10.1161/CIRCGENETICS.113.000654
45. Timpson NJ, Nordestgaard BG, Harbord RM, et al. C-reactive protein levels and body mass index: elucidating direction of causation through reciprocal Mendelian randomization. *Int J Obes (Lond)*. 2011;35(2):300-308. doi:10.1038/ijo.2010.137
46. Moreno-Navarrete JM, Martinez-Barricarte R, Catalán V, et al. Complement factor H is expressed in adipose tissue in association with insulin resistance. *Diabetes*. 2010;59(1):200-209. doi:10.2337/db09-0700

## SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

**How to cite this article:** Niu J, Adam J, Skurk T, et al. Machine learning approach on plasma proteomics identifies signatures associated with obesity in the KORA FF4 cohort. *Diabetes Obes Metab*. 2025;27(5):2626-2636. doi:10.1111/dom.16264