

Plasma proteomic profiling reveals distinct molecular signatures associated with coronary collateral circulation in patients with chronic total occlusion

Yi-Ning yang

yangyn5126@163.com

People's Hospital of Xinjiang Uygur Autonomous Region <https://orcid.org/0000-0002-8332-8508>

Jun-Yi Luo

First Affiliated Hospital of Xinjiang Medical University

Peng Ran

Fudan University

Fen Liu

First Affiliated Hospital of Xinjiang Medical University

Ya-Jing Qiu

the First Affiliated Hospital of Xinjiang Medical University

Bin-Bin Fang

Prevention and Treatment of High Incidence Diseases in Central Asia

Yunzhi Wang

School of Life Sciences, Institute of Biomedical Sciences, Human Phenome Institute, Fudan University

<https://orcid.org/0000-0002-1086-8896>

Sha Tian

Fudan University <https://orcid.org/0000-0003-4407-8873>

Xiao-Mei Li

First Affiliated Hospital of Xinjiang Medical University

Chen Ding

State Key Laboratory of Genetics and Development of Complex Phenotypes, School of Life Sciences, Human Phenome Institute, Fudan University <https://orcid.org/0000-0001-8673-3464>

Article

Keywords: coronary collateral circulation, chronic coronary total occlusion, proteomic analysis, machine learning

Posted Date: July 2nd, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10107169/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.
[Read Full License](#)

Additional Declarations: There is **NO** Competing Interest.

Plasma proteomic profiling reveals distinct molecular signatures associated with coronary collateral circulation in patients with chronic total occlusion

Jun-Yi Luo^{1,3,†}, Peng Ran^{2,†}, Fen Liu³, Ya-Jing Qiu¹, Bin-Bin Fang³, Yunzhi Wang², Sha Tian², Xiao-Mei Li^{1,*}, Chen Ding^{2,*}, and Yi-Ning Yang^{4,*}

¹ Department of Cardiology, the First Affiliated Hospital of Xinjiang Medical University, Urumqi, Xinjiang, 830054, China

² Center for Cell and Gene Therapy, Fudan University Clinical Research Center for Cell-based Immunotherapy, Shanghai Pudong Hospital, Human Phenome Institute, State Key Laboratory of Genetic Engineering, School of Life Sciences, Fudan University, Shanghai 200433, China.

³ State Key Laboratory of Pathogenesis, Prevention and Treatment of High Incidence Diseases in Central Asia, Clinical Medical Research Institute, the First Affiliated Hospital of Xinjiang Medical University, Urumqi, Xinjiang, 830054, China.

⁴ Department of Cardiology, Xinjiang Key Laboratory of Cardiovascular Homeostasis and Regeneration Research, People's Hospital of Xinjiang Uygur Autonomous Region, Urumqi, Xinjiang, 830000, China

Jun-Yi Luo and Peng Ran contributed equally to this manuscript.

Jun-Yi Luo: luojunyi3130@126.com

Peng Ran: 21112030023@m.fudan.edu.cn

Fen Liu: 604760840@qq.com

Ya-Jing Qiu: 676366143@qq.com

Bin-Bin Fang: xj_binbin5515@sina.com

Yunzhi Wang: yunzhiw@fudan.edu.cn

Sha Tian: shat@fudan.edu.cn

Chen Ding: chend@fudan.edu.cn

Xiao-Mei Li: lixm505@163.com

26 Yi-Ning Yang: yangyn5126@163.com

27 **Corresponding authors**

28 Yi-Ning Yang, People's Hospital of Xinjiang Uygur Autonomous Region, 91 Tianchi Road, Urumqi, 830054,
29 Xinjiang, China, yangyn5126@163.com.

30 Chen Ding, Center for Cell and Gene Therapy, Fudan University Clinical Research Center for Cell-based
31 Immunotherapy, Shanghai Pudong Hospital, Human Phenome Institute, State Key Laboratory of Genetic
32 Engineering, School of Life Sciences, Fudan University, Shanghai 200433, China.

33 Xiao-Mei Li, Department of Cardiology, the First Affiliated Hospital of Xinjiang Medical University, 137
34 Liyushan South Road, Urumqi, 830054, Xinjiang, China, lixm505@163.com

35 **Abstract**

36 **Background:** Coronary collateral circulation (CCC) plays a critical cardioprotective role in
37 mitigating myocardial ischemia through its dynamic compensatory capacity. Clinical evidence has
38 demonstrated that well-CCC development preserves relatively cardiac functional stability even in
39 the presence of complete native coronary artery occlusion. However, the molecular mechanisms
40 underlying CCC formation remain poorly understood, and there is a lack of clinically actionable
41 non-invasive biomarkers for classifying CCC status in patients with chronic total occlusion (CTO).

42 **Methods:** Using a data-independent acquisition quantitative proteomic approach, this study
43 comprehensively analyzed 235 plasma samples from two independent CTO cohorts. Collateral
44 filling from the contralateral vessel was categorized into four grades according to the Rentrop
45 classification criteria (Grade 0 to 3).

46 **Results:** Proteome profiling revealed enrichment of protein degradation and ion transport
47 pathways in patients with good-CCC CTO, whereas immune responses and vascular remodeling
48 pathways predominated in patients with bad-CCC. Using candidate plasma protein biomarkers

(SOD1, UFD1, DDX5, MIA3, TOM1, and LMAN2) in conjunction with clinical data, CCC classifiers were developed that demonstrated high predictive accuracy, achieving an area under the curve of 0.91 in the discovery cohort. The robustness and generalizability of the model were further confirmed in an independent validation cohort, with an area under the curve (AUC) of 0.82. ELISA analysis revealed that the candidate plasma proteins effectively discriminated CCC status in patients with CTO, achieving an AUC of 0.76 (95% CI, 0.73–0.77) and an area under the precision-recall curve (AUPRC) of 0.79 (95% CI, 0.77–0.81). Additionally, a risk prediction model for main adverse cardiovascular and cerebrovascular events (MACCEs) was constructed based on a panel of nine plasma proteins, providing a potential tool for the early identification of patients at high-risk of CTO.

Conclusion: Machine learning models based on plasma proteomic profiles outperformed existing clinical models in predicting CCC status. The developed risk prediction model for MACCEs provides a non-invasive approach for early stratification of patients at high risk of CTO.

Keywords: coronary collateral circulation, chronic coronary total occlusion, proteomic analysis, machine learning

Introduction

Coronary artery disease (CAD)¹, remains a predominant global cause of mortality, primarily due to flow-limiting obstruction of the coronary arteries that induces myocardial ischemia and infarction^{1, 2}. Among patients with CAD, approximately 30–52% present with chronic total occlusion (CTO) of a coronary artery^{3, 4}. CTO is angiographically defined as complete coronary artery occlusion with no antegrade epicardial flow persisting for more than three months, as confirmed by prior coronary angiography or clinical history⁵. This persistent vascular occlusion not only compromises myocardial perfusion but also heightens cardiovascular risk, manifesting clinically as refractory angina, progressive heart failure, and recurrent infarction⁶. Notably, the clinical trajectory of CTO exhibits substantial heterogeneity, partially attributable to the development of coronary collateral circulation (CCC). These naturally occurring bypass vessels mediate alternative perfusion to jeopardized myocardium, with emerging evidence indicating a substantial impact on patients' quality of life and clinical outcomes⁷. Recent longitudinal data demonstrate that patients with well-developed CCC achieve 10-year survival rates exceeding 50% while maintaining preserved cardiac function, even in the presence of multivessel chronic occlusions^{8, 9}. This adaptive response highlights the therapeutic potential of collateral-promoting strategies in optimizing clinical management approaches for patients with complex CTO.

However, the mechanisms underlying CCC formation and its regulatory determinants remain incompletely understood. Studies have shown that CCC development is influenced by multiple factors, including genetic predisposition, metabolic status, vascular endothelial growth factor expression levels, and the local hemodynamic environment^{10, 11}. Moreover, the dynamic nature of CCC underscores its critical role in tailoring individualized treatment strategies for patients with CTO. Early assessment of CCC development allows clinicians to tailor and optimize therapeutic

92 strategies more accurately, thereby improving patient outcomes^{12, 13}. The current gold standard for
93 assessing CCC, the collateral flow index (CFI), relies on invasive techniques that are technically
94 demanding and carry procedural risks¹⁴, especially in patients with non-occlusive lesions where
95 antegrade and collateral flow are difficult to distinguish. These limitations highlight the urgent
96 need to elucidate the underlying mechanisms and to identify reliable, non-invasive biomarkers for
97 assessing CCC functionality, thereby facilitating early clinical intervention for patients with CTO.

98 Plasma proteins represent dynamic indicators of pathophysiological signatures, serving as
99 accessible biomarkers and direct reporters of tissue-specific responses¹⁵. Their origins, whether
100 actively secreted or passively released from cardiomyocytes during injury, enable proteomic
101 profiling to capture disease states with high biological fidelity. Mass spectrometry-based
102 proteomics has yielded substantial progress in cardiovascular research while simultaneously
103 enabling high-throughput analysis of hundreds to thousands of proteins^{16, 17}. For instance,
104 researchers have identified cardiac myosin-binding protein C (cMyBP-C) as a sensitive marker of
105 myocardial injury through plasma proteomics¹⁸⁻²⁰, offering new avenues for the early diagnosis of
106 myocardial infarction. Recent studies have demonstrated the clinical utility of proteomics in the
107 secondary prevention of atherosclerotic cardiovascular disease (ASCVD)²¹. Systematic proteomic
108 analyses of two ASCVD secondary prevention cohorts revealed that risk prediction models based
109 on proteomic features outperformed traditional clinical risk models in predicting recurrent ASCVD
110 events²¹.

111 Additionally, a large-scale cohort study using the UK Biobank conducted comprehensive plasma
112 proteomic profiling in 44,313 participants and integrated these data with 11.1 years of follow-up.
113 This study constructed a network linking circulating proteomic signatures with the incidence risk
114 of CAD, heart failure, atrial fibrillation, and aortic stenosis²². The findings systematically

elucidated potential molecular mechanisms underlying these cardiovascular diseases and provided evidence-based support for proteomics-driven strategies in cardiovascular disease prevention²².

In this study, data-independent acquisition (DIA) proteomic profiling was performed on 235 plasma samples from two independent multicenter cohorts (124 in the discovery cohort and 111 in the validation cohort) to evaluate the clinical utility of plasma-based biomarkers for assessing CCC status and predicting the risk of major adverse cardiovascular and cerebrovascular events (MACCE) risk in patients with CTO. Collectively, the findings provide novel insights into the molecular characterization of CCC and demonstrate the feasibility of proteomics-integrated machine learning approaches for developing non-invasive diagnostic and prognostic tools. These advances may facilitate personalized therapy and contribute to more effective management of CTO.

Materials and Methods

Sample preparation

Patient cohort and clinical data

CTO was defined as complete vascular occlusion with thrombolysis in myocardial infarction grade 0: indicating no flow distal to the occluded vessel, with an occlusion duration of more than 3 months. The inclusion criteria were as follows: 1) age ≥ 18 years and ≤ 80 years, 2) stable or unstable angina pectoris; and 3) coronary angiography confirming the presence of CTO in at least one major artery (left anterior descending, left circumflex, or right coronary artery). The exclusion criteria included prior coronary artery bypass grafting, acute myocardial infarction, active infectious disease, chronic inflammatory disease, rheumatologic disease, or malignancy. A total of 235 patients with CTO were recruited into two independent CTO cohorts for further analysis, including a discovery cohort (n = 124) from the First Affiliated Hospital of Xinjiang Medical

University and a validation cohort (n = 111) at the People's Hospital of Xinjiang Uygur Autonomous Region Hospital (**Figure 1A; Supplementary Figure 1A**). Additionally, an external validation cohort (cohort 2) was recruited from the People's Hospital of Xinjiang Uygur Autonomous Region Hospital, comprising 291 patients with CTO (**Supplementary Figure 1A**).

Clinical information regarding patient history, CCC status, and relevant diagnostic findings, and surgical records was obtained from all patients. All angiographic data were obtained using a digital angiography systems (GE Medical Systems; Siemens, Forchheim, Germany). At least two experienced pathologists independently reviewed all coronary angiography images from patients with CTO to verify the diagnostic findings. In cases of disagreement, a third reviewer adjudicated the discrepancy. Collateral filling from the contralateral vessel was categorized into four grades according to the Rentrop classification criteria^{23, 24}: Grade 0: No collateral filling; Grade 1: Filling of side branches of the artery to be dilated via collateral channels without visualization of the epicardial segment; Grade 2: Partial filling of the epicardial segment via collateral channels; Grade 3: Complete filling of the epicardial segment of the artery being dilated via collateral channels. According to collateral flow, patients were classified into a good CCC group (Rentrop grades 2–3)⁸ or a bad-CCC group (Rentrop grades 0–1).

Ethical approval

The Research Ethics Committees of the First Affiliated Hospital of Xinjiang Medical University approved this study (Ethics Approval Number: K202101-20), and written informed consent was obtained from all patients.

157 *Study endpoints*

158 Participants were followed up every 3–6 months for endpoint events through telephone interviews
159 or clinic visits. The study endpoint was the occurrence of composite MACCEs during follow-up,
160 including all-cause death, cardiac death, non-fatal myocardial infarction, non-fatal stroke,
161 malignant arrhythmia, revascularization, severe heart failure (New York Heart Association
162 Functional Classification III or IV), and target vessel revascularization²⁵.

163 *Sample collection and storage*

164 A total of 235 plasma samples were collected for further analysis. Blood samples were stored at 4
165 °C after collection, and plasma was extracted within one hour to prevent red blood cell hemolysis.
166 All plasma samples were transferred to the research laboratory and stored at -80 °C ²⁶.

167 **Plasma proteome analysis**

168 *Plasma protein extraction and trypsin digestion*

169 In this procedure, 2 µL of plasma and body fluid samples were mixed with 100 µL of 50 mM ABC
170 buffer, followed by protein inactivation at 95°C for 5 min. After cooling to room temperature, the
171 samples were digested with trypsin at an enzyme-to-protein mass ratio of 1:50 for 16 hours in a
172 37°C incubator. Then, 5 µL of aqueous ammonia was added to each tube and vortexed to quench
173 the digestion reaction. Then, the supernatant was dried at 60°C using a vacuum drier (SpeedVac,
174 Eppendorf). After drying, the peptides were dissolved in 100 µL of 0.1% formic acid and
175 centrifuged for 3 min (12,000 g). The supernatant was transferred to a new tube and then
176 desalinated. Before desalting, the activation of the pillars with 2 slides of 3 M C18 disks was
177 performed, and the lipid content was 100 µL of 100% acetonitrile added twice, 100 µL of 50% and

80% acetonitrile added once, and then 100 μ L of 50% acetonitrile added once. The columns were equilibrated with two 100 μ L washes of 0.1% formic acid, and the sample supernatant was loaded twice onto the columns, which were then washed twice with 100 μ L of 0.1% formic acid for decontamination. Finally, 100 μ L of elution buffer (0.1% formic acid in 50% acetonitrile) was added to the column for elution twice, and only the effluent was collected for mass spectrometry analysis. Then, the collected eluate was dried at 60°C in a vacuum drier.

LC-MS/MS Acquisition of Serum Samples

Samples were measured using LC-MS instrumentation consisting of an EASY-nLC 1200 ultrahigh-pressure system (Thermo Fisher Scientific) coupled via a nanoelectrospray ion source (Thermo Fisher Scientific) to a Fusion Lumos Orbitrap (Thermo Fisher Scientific). The peptides were dissolved in 12 μ L of loading buffer (0.1% formic acid in water), and 5 μ L was loaded onto a 100 μ m ID $\times$ 2.5 cm C18 trap column at a maximum pressure of 280 bar with 14 μ L of solvent A (0.1% formic acid in water). Peptides were separated on a 75 μ m I.D. $\times$ 15 cm column (C18, 1.9 μ m, 120 Å, Dr. Maisch GmbH) with linear 15–30% mobile phase B (ACN and 0.1% formic acid) at 600 nL/min for 10 mins. MS analysis was performed in a data-independent manner. The data-independent acquisition (DIA) method consisted of an MS1 scan from 300–1400 m/z at 60k resolution (AGC target 4e5 or 50 ms). Then, 30 DIA segments were acquired at 15k resolution with an AGC target of 5e4 or 22 ms for maximal injection time. The setting of injections for all available parallelizable times was enabled. HCD fragmentation was set to a normalized collision energy of 27%. The spectra were recorded in profile mode. The default charge state for MS2 was set to 3.

Peptide identification and label-free protein quantification

As mentioned earlier, the spectral library enabled the targeted extraction of precursor ions and protein intensities from the DIA runs. This extraction was executed using DIA-NN software (v1.8.1)²⁷ in a manner consistent with previously established methodologies, and the recommended default settings of DIA-NN were utilized. Carbamidomethylation at cysteine residues was the sole fixed modification allowed. To enhance data quality, interfering precursor peaks were removed, and the robust LC (high accuracy) quantification strategy was employed, filtering precursors at a q-value threshold of 0.01. Protein grouping was conducted at the canonical protein sequence level rather than the gene level. After this analysis, the output underwent stringent filtering based on several criteria: a precursor Q-value of less than 1% per experiment, a global protein Q-value of less than 1%, and a run-specific protein Q-value of less than 1%. For protein abundance calculations, precursor ion intensities were aggregated at the unique gene symbol level using the MaxLFQ normalization method²⁸. This method is known for its intensity-based, label-free quantification, integration of delayed normalization, and maximal peptide ratio extraction. Finally, peak area values were calculated as components of their respective proteins.

Quality control of the MS data

Quality control of the MS platform

For quality control of MS data, HEK293T cell (National Infrastructure Cell Line Resource) lysates were measured every three days as the quality control standard. The standard was digested and analyzed using the same method and conditions as those used for the CTO patient plasma samples. A pairwise Spearman correlation coefficient was calculated for all quality control runs in the statistical analysis environment R (version 4.1.0).

221 *Quality evaluation of plasma proteomic data*

222 Density plots of peptide counts and protein identifications showed a precise bimodal distribution,
223 indicating that some samples had insufficient peptide or protein yields and should be excluded
224 from further analysis. Therefore, the quality of the generated data was assessed by examining the
225 distribution status of the identified proteins for all samples in the R environment (version 4.1.0).
226 In our cohort, 122 plasma samples in the discovery cohort and 110 plasma samples in the validation
227 cohort passed quality control and were used for further proteomic analysis. Those samples that did
228 not pass the quality control tests were excluded from the subsequent analyses. All the quality
229 control samples showed excellent consistency in terms of proteome quantification
230 (**Supplementary Figure 5G-5H**), exhibiting a typical unimodal (Gaussian or normal) distribution
231 (statistical dip test).

232 *Coefficients of Variance*

233 In DIA-based quantitative proteomics, coefficients of variance (CVs) are instrumental in
234 evaluating the reliability and consistency of experimental outcomes. CVs represent the relative
235 variability surrounding the mean of a dataset. A lower CV signifies greater accuracy and stability,
236 reflecting fewer discrepancies in repeated observations. The CVs were calculated using the
237 standard deviation (SD) divided by the mean values of the peptides and proteins in the dataset.
238 Accordingly, CV values for the quantification intensity of peptides and proteins were ascertained,
239 providing a comprehensive assessment of our measurement precision.

240 *Batch Effects Correction for the Plasma Proteomic Data*

241 Batch effects are a major source of technical bias in proteomic studies, primarily arising from non-
242 biological variations related to sample processing, experimental procedures, and instrument

performance. If not properly identified and corrected, these systematic errors can severely confound the interpretation of true biological differences, leading to biased research conclusions. Therefore, establishing robust strategies for detecting and correcting batch effects is essential to ensure the reliability and reproducibility of proteomic data.

This study employed Principal Component Analysis (PCA), a widely used multidimensional data visualization method, to assess potential batch effects. The analytical workflow involved standardizing the protein expression matrix to eliminate scale differences across features, followed by the computation of the primary principal components (PC1 and PC2) and the generation of two-dimensional scatter plots. Experimental batches were color-coded to identify potential clustering patterns visually. The PCA results indicated that samples from the discovery and validation cohorts exhibited a homogeneous distribution in the principal component space without evident batch-related clustering (**Supplementary Figure 5E-5F**), suggesting consistent experimental procedures and the absence of significant batch effects. However, one outlier sample was identified in each cohort, with principal component scores deviating significantly from the main cluster ($|Z\text{-score}| > 2.58$, $p\text{-value} < 0.01$), suggesting possible technical artifacts. To preserve the integrity of downstream analyses, these outliers were excluded from subsequent analyses.

Missing value imputation

To ensure the reliability of downstream analyses, plasma proteins with a missing rate greater than 50% across all samples were excluded from the dataset. This exclusion criterion is consistent with previously established practices in proteomic studies^{29, 30}. For the remaining proteomic data, missing values were imputed using the K-Nearest Neighbors (K-NN) algorithm^{26, 31}. Specifically, the K-NN imputation was implemented using the open-source scikit-learn package in the Python

programming environment (version 3.8.6). This approach estimates missing values by calculating Euclidean distances between samples and leveraging the feature values of the most similar neighbors, thereby preserving the local structure of the dataset.

We conducted a comprehensive assessment of missing data distribution to evaluate potential biases and ensure data quality. The preprocessed dataset (after excluding proteins with >50% missingness) exhibited high data completeness (**Supplementary Figure 6A, 6B**). Specifically, 82% of proteins have 70–100% completeness across samples. 45.5% of proteins showed 100% completeness without any missing value data. Only a minor fraction of proteins (<5%) fall in the 50–60% completeness range, confirming that most retained data are well-supported (**Supplementary Figure 6C**). To further validate the robustness of the KNN-imputed dataset and assess the clinical relevance of the identified protein biomarkers, we conducted independent experimental verification using the ELISA method in CTO patient plasma samples. The ELISA results exhibited strong concordance with the proteomic findings (**Supplementary Figure 10A, 11A**), which demonstrated the reliability of the imputation strategy.

Data normalization

To more accurately assess relative protein abundance and minimize the impact of data variance, $\log_2$ transformation was applied to normalize the quantified protein expression levels. As a widely used technique in proteomic data analysis, $\log_2$ transformation effectively compresses the high dynamic range of protein expression, thereby enhancing distributional stability and facilitating the identification of underlying biological patterns and differences. Following transformation, protein abundances were further adjusted to ensure comparability across samples. Specifically, the expression levels of all proteins in each sample were normalized to their respective median values.

This normalization approach mitigated the influence of outliers and sample-specific variability, thereby improving the robustness and consistency of the dataset.

Analysis of differentially expressed proteins

Differentially expressed protein (DEP) analysis in experimental samples featuring distinct phenotypes was conducted using the Wilcoxon rank-sum test. This analysis included comparisons between the good-CCC and bad-CCC CTO patient groups. Before the differential expression analysis, proteomic data containing more than 50% missing data were excluded, and 1,633 plasma proteins were utilized as input for DEP analysis. Next, the protein expression matrix was utilized to identify differentially expressed proteins within the comparison groups, employing the “Contrasts” functions incorporated in the limma³² (v.3.48.3) R/Bioconductor package. *P*-values were adjusted using the Benjamini–Hochberg⁶ method, and proteins exhibiting an adjusted *p*-value < 0.05 and a fold change > 1.5 were deemed significantly differentially expressed. As a result, DEP analysis revealed 235 proteins whose levels significantly differed between good-CCC patients and bad-CCC patients.

Pathway enrichment analysis

In this study, pathway enrichment analysis was performed using Over-Representation Analysis (ORA) to systematically investigate the biological functions of differentially expressed proteins. Specifically, functional annotation was conducted through the ConsensusPathDB (CPDB) interactive platform (<http://cpdb.molgen.mpg.de/>), which integrates multiple pathway databases, including KEGG³³, Gene Ontology (GO) biological processes³⁴, and Reactome³⁵ databases. Pathways with a false discovery rate (FDR) < 0.01 were considered significantly enriched.

Construction of the classification model for coronary collateral circulation status

Dataset and model training.

In this study, the gradient-boosted tree algorithm XGBoost (Python package “xgboost” version 1.6.1) was employed to construct a classification model aimed at determining the CCC status (good CCC vs. poor CCC) of CTO patients in the discovery dataset. During the model derivation process, the discovery cohort was stratified into a training set (70%) and an independent test set (30%) using stratified random sampling to ensure balanced distribution of CCC status. To prevent data leakage, all model development processes, including feature selection, hyperparameter tuning, and final model training, were conducted solely on the training set. Once the model was established, it was rigorously tested and validated exclusively on the held-out test set to ensure its complete independence from feature selection and model training, thereby ensuring the authenticity and reliability of model evaluation. The locked models were then evaluated internally on the test set and externally on independent, unseen datasets to assess their robustness and generalizability (Supplementary Figure 2).

XGBoost is a high-performance machine learning algorithm based on a recursive tree structure and offers excellent interpretability, in contrast to typical “black-box” model mechanisms. Beyond accurate prediction of CCC status, our objective was to interpret the decision-making logic of the model. By integrating clinical data and plasma proteome profiles, we identified critical biomarkers indicative of CCC status, enabling differentiation between CTO patients who may require different medical management strategies. This approach ensured precise alignment of features with their corresponding labels, categorizing CTO patients into two groups: good CCC (Label 0) and poor CCC (Label 1).

Model evaluation metrics

The effectiveness of the predictive models in this study was evaluated using a range of evaluation metrics, including the Confusion Matrix, Area Under the Curve³⁶, Accuracy, Precision, Recall, F1 Score, Macro Average Score, and Weighted Average Score. Each metric provides unique insights into the model's performance, catering to various aspects of prediction accuracy and reliability.

To comprehensively evaluate the robustness and generalization ability of the model, this study employed multiple evaluation metrics to assess the performance of the predictive model. These metrics include the confusion matrix, Area Under the Receiver Operating Characteristic Curve ³⁶, Area Under the Precision-Recall Curve (AUPRC), accuracy, precision, recall, F1-score, Matthews Correlation Coefficient (MCC)³⁷, and Balanced Accuracy (BA)³⁸. Each metric reflects different aspects of the model's prediction accuracy and classification performance. Ultimately, the model's classification performance was calculated by computing the geometric mean of four key indicators: AUC, AUPRC, accuracy, and F1-score. This method allows for a balanced consideration of various facets of the model's prediction, ensuring a comprehensive and objective evaluation.

1. Confusion Matrix

The Confusion Matrix is a tabular representation of the model's predictions, detailing the counts of True Positives (TP), True Negatives ³⁹, False Positives (FP), and False Negatives (FN) for each class. It provides insights into the model's performance across different categories. The confusion matrix helps evaluate how well our model predicts the CCC (good-CCC or bad-CCC) status of patients. The axes represent predicted labels versus actual labels.

2. ROC Curves and Area Under the Curve

The ROC curve illustrates the trade-off between the True Positive Rate (TPR) and the False Positive Rate (FPR) of a classification algorithm, serving as a crucial tool for evaluating classifier performance. Ideally, the ROC curve should be close to the upper left corner of the plot, indicating excellent classifier performance; conversely, a curve near the diagonal line suggests that the classifier performs similarly to random guessing. The Area Under the Curve ³⁶ represents the integral area under the ROC curve, providing a comprehensive metric for model performance. AUC values range from 0.5 to 1.0, where 0.5 indicates no discriminative ability, and 1.0 denotes perfect classification. AUC is robust to class imbalance, making it suitable for datasets with varying class proportions. The formula for calculating AUC is as follows:

$$AUC = \frac{\sum_{ins_i \in \text{positive}} \text{rank}_{ins_i} - \frac{M \times (M + 1)}{2}}{M \times N} \quad (1)$$

Where:

- $\sum_{ins_i \in \text{positive}} \text{rank}_{ins_i}$ is the sum of the ranks for all positive instances.
- M is the total number of positive samples.
- N is the total number of negative samples.
- $\frac{M \times (M + 1)}{2}$ is the sum of the first M natural numbers, which adjusts the rank sum to a baseline where all positive samples are ranked lower than any negative sample.

368 3. Accuracy

369 Accuracy measures the ratio of correctly predicted samples (true positives and true negatives) to
370 the total number of predictions. However, it is unsuitable for unbalanced datasets.

$$371 \quad \text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \quad (2)$$

372 4. Precision and Recall

- 373 • Precision: The proportion of correctly predicted positive samples (true positives) among
374 all predicted positive samples.

$$375 \quad \text{Precision} = \frac{TP}{TP + FP} \quad 39$$

376

- 377 • Recall (Sensitivity): The proportion of correctly predicted positive samples among all
378 actual positive samples.

$$379 \quad \text{Recall} = \frac{TP}{TP + FN} \quad (4)$$

380 5. F1 Score

381 The F1 score is the weighted harmonic mean of precision and recall. A higher F1 score indicates
382 a good balance between precision and recall. The formula for calculating the F1 score is as follows:

$$383 \quad F1\text{-Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (5)$$

384 A high F1 score indicates a good balance between precision and recall.

6. Matthews Correlation Coefficient (MCC)

The MCC is an important metric for evaluating the performance of binary classification models. It is mathematically defined as the correlation coefficient between the true class c and the predicted label l . The MCC value ranges from $[-1,1]$, where -1 indicates a perfect negative correlation, 1 indicates a perfect positive correlation, and 0 represents random prediction. The formula for calculating MCC is as follows:

$$\text{MCC} = \frac{\text{Cov}(c, l)}{\sigma_c \cdot \sigma_l} = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP + FP) \cdot (TP + FN) \cdot (TN + FP) \cdot (TN + FN)}} \quad (6)$$

Where $\text{Cov}(c, l)$ represents the covariance between the true class c and the predicted label l , and σ_c and σ_l denote the standard deviations of the true class and predicted label, respectively. The MCC takes into account all four possible classification outcomes—true positives, true negatives, false positives, and false negatives—providing an effective measure of classifier performance, particularly for imbalanced datasets.

7. Balanced Accuracy (BA)

Balanced accuracy is the arithmetic mean of sensitivity and specificity. The range of MCC is $[0, 1]$, and the calculation formula is:

$$\text{Balanced Accuracy} \sim (\text{BA}) = \frac{\text{TPR} + \text{TNR}}{2} \quad (7)$$

By employing these metrics, we ensure a comprehensive evaluation of our model's performance, considering various aspects of predictive accuracy and robustness.

Feature importance evaluation

To evaluate feature importance, we employed an ensemble of different machine learning models (**Supplementary Figure 2**), including Random Forest, XGBoost, LightGBM, CatBoost, Support Vector Classifier (SVC), and Logistic Regression. Each model was trained on the entire training set to assess the relative contribution of each feature to the model's predictions. The method of determining feature importance varied across models: tree-based models (XGBoost, Random Forest, LightGBM, CatBoost) measured importance based on a feature's frequency in making split decisions at each tree node, while linear models (SVC, Logistic Regression) relied on feature coefficients. To reduce the error and generate a more robust feature importance ranking, this process was performed with a 5-fold cross-validation procedure to compute the average importance scores of each feature. The final feature importance ranking was determined by aggregating the results from six different feature-ranking algorithms using a weighted voting approach. All models were implemented in Python using the “scikit-learn” module. By quantifying the relative importance of each feature using this ensemble method, this approach could accurately and robustly identify key features that most significantly contribute to the model's ability to discriminate between classes (Good-CCC and Bad-CCC for CTO patients).

All features were extracted from the data frame "df" and assigned to the variable "X_data_all_features". The target variable was denoted by "Y_data," and the remaining column names (excluding the target) were stored in "x_col." A new data frame, "important_feature," was created to store the features and their corresponding importance scores, initially set to zero for all features. The process was iterated over 100 experiments. In each iteration, a unique random seed was used to partition the data for training and testing, mitigating the potential influence of outliers on feature importance, particularly in smaller datasets. The "train_test_split" function was

employed in each experiment to divide the dataset into training and testing subsets. Each classification model, configured with default hyperparameters, was then trained on the training data. Feature importance scores from each experiment were accumulated by summation and subsequently normalized by division by 100. These normalized scores were then stored in the corresponding feature columns of the "important_feature" data frame. Finally, features were ranked in descending order based on their aggregated relative importance scores, with the top features highlighted. Additionally, bar charts were generated for the 12 most important features to visualize their relative importance.

Feature selection

The model training integrated both clinical and proteomic features (Figure 4A). The discovery cohort was stratified into a training set (70%) and an independent test set (30%) using a stratified random partitioning method. To prevent data leakage, all model development processes—including feature selection, hyperparameter optimization, and final model training—were conducted exclusively on the training set. After identifying differentially expressed proteins, an ensemble learning approach was adopted to further refine the feature selection process. This approach combined the feature importance scores of multiple base learners (XGBoost, Random Forest, and Support Vector Machines) and quantified the marginal contribution of each feature to the model's predictions using Shapley values, thus establishing a multidimensional importance evaluation matrix^{40, 41}. This ensemble method has been successfully applied in several studies⁴², primarily enhancing the robustness and reliability of feature selection by leveraging the strengths of various base learning models.

Based on the feature importance ranking, recursive feature elimination⁴³ was further employed for feature optimization. This method used AUC as the objective function, iteratively removing

features with minimal contributions to model performance. Specifically, during each iteration, the least influential feature on AUC was removed until the removal of any remaining features led to an AUC drop exceeding the predefined threshold. After this optimization process, a streamlined feature subset containing the most discriminative features was obtained. This model performance-based feature selection strategy not only enhanced the stability of the feature set but also effectively mitigated the risk of the curse of dimensionality, providing a reliable foundation for subsequent model construction.

The specific steps are as follows: Initially, the iteration counter "num_i" was set, and two variables were used to track the best AUC scores for the training and validation sets, respectively. Features were sorted in descending order of importance, and a stratified K-fold cross-validation (K=5) strategy was applied for feature selection. In each iteration, the next most important feature was added, and the model's training and validation AUC scores were computed. The iteration process was terminated when the validation AUC score ceased to improve, and the current feature subset was returned. This method, with its strict cross-validation, ensured the stability of feature selection while avoiding the risk of overfitting, providing the optimal feature combination for subsequent modeling.

Hyperparameter optimization

Machine learning models with properly tuned hyperparameters significantly improve the model's accuracy, training speed, and ability to generalize to new data to achieve state-of-the-art results. To optimize the hyperparameters of the XGBoost model, we employed Bayesian optimization using the tree-structured Parzen estimator (TPE) algorithm⁴⁴. The specific hyperparameters are defined as a dictionary with the following hyperparameters:

471 • Number of rounds (num_boost_round): This parameter determines the number of boosting
 472 rounds or trees to build. It is important to tune it properly to avoid overfitting. We explored values
 473 between 100 and 1000.

474 • Maximum depth of trees (max_depth): The maximum depth of the tree is defined as follows:
 475 We tested depths from 3 to 18.

476 • Gamma (gamma): Specifies the minimum loss reduction needed to make a split. The search
 477 space was between 0.1 and 1 in increments of 0.25.

478 • Regularization (L1 and L2: reg_alpha and reg_lambda, respectively): These add regularization
 479 terms to the objective function. We searched within a range of 0 to 1 in increments of 0.01 for both
 480 parameters.

481 • Column Sampling by Tree (colsample_bytree): Denotes the fraction of features to be randomly
 482 sampled for building each tree. We tested values ranging from 0.5 to 1 in increments of 0.05.

483 • Minimum Child Weight (min_child_weight): This parameter determines the minimum sum of
 484 the instance weights needed for a child. A search was performed between 0 and 10.

485 • Learning rate (eta): This makes the optimization more robust by shrinking the weights in each
 486 step. We used a range from 0.025 to 0.5 in increments of 0.05.

487 • Subsample: This denotes the fraction of observations to be randomly sampled for each tree.
 488 The range for this hyperparameter was between 0.5 and 1 in increments of 0.05.

489 This hyperparameter grid will be used as input to a hyperparameter optimization algorithm to
 490 search for the optimal combination of hyperparameters that maximizes the area under the ROC
 491 curve³⁶. We defined the objective function `objective(space)`, which takes a hyperparameter space

as input and returns a dictionary containing the loss and status values. This function is used in Bayesian optimization to evaluate the performance of the XGBoost model. For the training process, we utilized the ‘gbtree’ booster and the exact tree method. Furthermore, since our problem is a binary classification task, we employed the 'binary:logistic' objective function and used the area under the receiver operating characteristic curve (eval_metric: 'auc') as the evaluation metric.

Hyperparameter optimization was performed using the “fmin()” function from the HyperOpt package, which executes the “objective()” function with different hyperparameter values specified in the “hyperparameter_grid”. The “fmin()” function employs the Tree-structured Parzen Estimator (TPE) algorithm to suggest subsequent hyperparameters for evaluation, based on previous results. The optimization continues until either the maximum number of evaluations specified by “max_evals” is reached or convergence is achieved. The “best_hyperparams” dictionary returned by “fmin()” contains the optimal hyperparameters identified during the optimization, within the search space defined by the “hyperparameter_grid”, aiming to minimize the loss value as defined by the objective function.

Hyperparameter optimization was conducted over 200 iterations, and the best set of hyperparameters was selected based on the highest AUC value achieved during these iterations. After optimization, the model was retrained on the entire training set using these optimal hyperparameters.

Benchmarking the performance of the classification model

To establish a performance benchmark, we compared the results of the interpretable model with those of standard machine learning algorithms, including random forest, decision tree, and logistic regression. The generalization performance and prediction accuracy of the constructed

classification models using candidate chronic total occlusion³ biomarkers were comprehensively evaluated on both the hold-out test set and an external independent validation cohort dataset. The benchmark models were trained using the open-source Python package “scikit-learn” (version 1.3.1) with default parameters to create baseline models without hyperparameter optimization. Five-fold cross-validation was employed to enhance the reliability of the performance assessments and mitigate model overfitting. The area under the curve³⁶ was used as a metric to evaluate the overall performance of all models on both datasets.

Enzyme-linked immunosorbent assay (ELISA)

In the external validation cohort, ELISA was used to validate candidate plasma protein biomarkers. The plasma concentrations of SOD1, UFD1, DDX5, MIA3, TOM1, and LMAN2 were quantified using commercially available ELISA kits (Bioswamp Life Science Lab, Wuhan, China). All assays were performed strictly in accordance with the manufacturer’s protocols.

Development of a least absolute shrinkage and selection operator (LASSO) Cox Regression Model for Predicting MACCE in Patients with CTO

A total of 1,644 plasma proteins from the discovery cohort were included in the LASSO regression analysis using the ‘glmnet’ package in R. Five-fold cross-validation was performed to identify the optimal value of the LASSO penalty parameter (λ). Patients with CTO were then stratified based on the median LASSO score. Multivariate analysis of candidate plasma proteins was conducted using a multivariate Cox proportional hazards model. The area under the curve (AUC)³⁶ was calculated using receiver operating characteristic (ROC) analysis with the ‘pROC’ package in R. Time-dependent ROC curves, AUCs, and C-indices were obtained using the ‘cindex’ function within the ‘pec’ package (R). The generalizability of the model was further assessed in an

independent external validation cohort. Kaplan–Meier survival analysis was employed to compare survival curves using the ‘survival’ and ‘survminer’ packages in R. Log-rank test results with p -values < 0.05 were considered to indicate statistical significance.

Statistical methods

All analyses were performed using R (version 4.1.0) and Python (version 3.8.6). Standard statistical tests, including the Wilcoxon rank-sum test, were used to analyze proteomic data, and Student's t -test was used to assess statistically significant differences between the subgroups of continuous variables. The Mann–Whitney U test was utilized to evaluate significant differences in Grantham distances between subgroups of continuous variables, and the Spearman correlation analysis was used to examine consistency and reproducibility across proteome data. Spearman's rank correlation coefficient was calculated. To account for multiple testing, p -values were adjusted using the Benjamini–Hochberg Pathways with a FDR correction method. DeLong's test was employed to determine the significance of differences in the AUCs between the two models⁴⁵. All the statistical tests were two-sided, and p -values < 0.05 were considered to indicate statistical significance.

Results

Clinical baseline characteristics of patients.

The detailed baseline characteristics of the discovery and validation cohorts are summarized in **supplementary Table 1**. There was no significant difference in age, sex, or ethnicity between the discovery cohort and the validation cohort ($p > 0.05$). Similarly, no significant differences were observed in the lateral branch blood supply, heart rate, smoking status, height, weight, and body mass index between the discovery cohort and the validation cohort ($p > 0.05$), further reinforcing

the consistency of physiological characteristics across cohorts. In terms of CCC status, 69 patients (56.6%) in the discovery cohort showed good CCC during a median follow-up of 2.0 years (IQR: 1.0–4.0 years), whereas 55 patients (50%) in the validation cohort demonstrated good CCC over a median follow-up of 2.0 years (IQR: 0.3–6.0 years).

Regarding vascular collateral circulation, although the validation cohort had a higher proportion of patients with Rentrop grade 0 (22.7% vs. 8.2%) and grade 3 (27.3% vs. 19.7%), the overall distribution pattern remained consistent, suggesting no clinically significant differences between the two cohorts in vascular collateral circulation status. Notably, the validation cohort had a significantly lower proportion of comorbidities than the discovery cohort (8.2% vs. 70.5%; $p < 0.001$), while the proportion of patients with a family history was significantly higher (27.3% vs. 2.5%; $p < 0.001$). Additionally, the treatment rate was significantly lower in the validation cohort (55.5% vs. 77.9%; $p = 0.002$). Nevertheless, these differences are of limited clinical relevance and are unlikely to compromise subsequent analyses, given the overall consistency in key clinical characteristics between the two cohorts.

A clinically based classification model exhibited limited predictive accuracy for CCC status.

To evaluate the classification performance of CCC status in patients with CTO, a baseline model utilizing clinical data was developed. Spearman correlation analysis was performed between distinct clinical features to assess interactions and collinearity among these variables (**Figure 1B**). Hierarchical clustering analysis grouped clinical features associated with CCC status into seven major clusters. Feature importance analysis was performed using the XGBoost model to identify clinical markers associated with CCC status. The top 12 clinical features were subsequently plotted in descending order according to their feature importance scores (**Figure 1C**). Considering the

relative importance scores and collinearity among clinical features, ultimately the top six contributing clinical features (basophil counts, prothrombin time (PT) activity, medical history duration, serum chloride concentrations, high-density lipoprotein- cholesterol (HDL-C) levels, and red blood cell distribution width (RDW) , were selected as inputs for constructing the final clinical baseline model. In these selected clinical features, patients with CTO with bad-CCC status exhibited significantly higher levels of basophils, RDW, and serum chloride concentration, as well as a longer duration of medical history and lower serum PT activity and HDL-C levels (**Figure 1D, 1E, p-value < 0.05**).

Moreover, multiple linear and logistic regression analyses identified basophil count, chloride levels, duration of medical history, and PT activity as independent risk factors for CCC status (**Supplementary Figure 3A-3B**). These results were further confirmed by multiple linear regression analysis, demonstrating significant negative coefficients after adjustment for clinical covariates (**Supplementary Figure 3C**), including gender, age, sex, smoking status, body mass index, blood pressure, glycated hemoglobin, and medication history. Basophil count, chloride levels, duration of medical history, and PT activity remained independent risk factors for CCC status in patients with CTO, highlighting the clinical relevance of these variables for risk stratification in this population. However, RDW⁴⁶ and HDL-C did not remain statistically significant after adjustment for other covariates, suggesting that their associations with CCC status may be confounded by additional factors.

Based on the six selected clinical features, the baseline clinical classification model, exclusively using features extracted from the discovery dataset, was trained. This constructed classification model was subsequently tested in an independent external cohort comprising 111 patients with CTO. Model's performance, as assessed by the AUC, yielded values of 0.85, 0.66, and 0.57 for the

training set, test set of the discovery cohort, and the external independent validation cohort, respectively (**Figure 1F, Supplementary Figure 1D**). Overall, CCC status prediction precision, accuracy, recall, and macro-average F1-scores were all below 0.70 in the test set and independent external validation cohort, indicating low model performance, with metric values ranging from approximately 0.50 to 0.60 (**Figure 1G**). Although the clinical feature-based classification model demonstrated good classification performance in the training set, its generalizability to unseen data in the test set and external validation cohort was limited. This limitation was reflected in the confusion matrices (**Supplementary Figure 1B-1C**), which demonstrated a significant decline in performance in the test set (accuracy: 0.63, precision: 0.67) and the external validation cohort (accuracy: 0.53, precision: 0.55). These findings suggest that the clinical baseline model has limited generalizability and performs suboptimally in predicting CCC status under real-world conditions.

In conclusion, the findings identified key clinical features associated with CCC status, which may provide valuable insights for CCC prediction and CTO management, underscoring the potential for tailored therapeutic strategies based on individual CCC profiles. However, the clinical feature-based model faces challenges in accurately classifying CCC status, exhibiting limited predictive accuracy. Thus, there is an urgent need to integrate additional multiomics-level features to refine the clinical diagnosis and classification of CCC status in patients with CTO.

Proteome profiling revealed molecular characteristics associated with distinct CCC statuses.

In this study, a DIA-MS proteomic analysis of plasma samples was performed on samples collected from all participants (n = 235). A total of 1,633 protein groups were identified in more than half of the samples (**Supplementary Figure 4A-4B**). The protein dynamic range spanned

eight orders of magnitude, enabling the detection of low-abundance protein changes (Supplementary Figure 4C-4F). To ensure the dataset reliability and high quality of the dataset, rigorous quality control (QC) procedures were implemented at each stage of plasma proteomic data generation. The average correlation coefficients of the QC proteomic standards in the discovery and validation cohorts were 0.93 (95% CI: 0.91–0.94) and 0.92 (95% CI: 0.90–0.94), respectively (Supplementary Figure 5A-5B). This result demonstrated high reproducibility across repeated measurements of the same samples. These results demonstrated the stability and consistent performance of the MS platform. The plasma proteome abundance exhibited a typical unimodal distribution without significant skewness (Supplementary Figure 5C-5D). PCA was performed on the discovery and validation cohorts to assess potential batch effects, and the results demonstrated excellent consistency in plasma proteome profile (Supplementary Figure 5E-5F). Following QC analysis of the discovery and validation cohorts, one outlier was identified in each cohort. Consequently, these two samples were excluded from all subsequent analyses (Supplementary Figure 5G-5H). Plasma proteomic data containing more than 50% missing data were excluded before further analyses. Therefore, a total of 1,633 plasma proteins were included in subsequent analyses (Supplementary Table 2).

To investigate molecular differences associated with CCC status in patients with CTO, a differential protein expression analysis using plasma proteome profiles was performed and categorized into good-CCC and bad-CCC subgroups. Differential protein expression analysis revealed 25 upregulated and 43 downregulated proteins in patients with good CCC (**p value < 0.05; fold change |FC| > 1.5; Figure 2A**). Predominantly, good-CCC status was associated with overexpression of plasma proteins such as ENPP2, SOD1, MIA3, and PSMA1, and with underexpression of proteins such as DDX5, EPN1, UFD1, TOM1, and UGP2 (**Figure 2B and**

2C). To identify molecular signatures of the biological processes associated with different CCC statuses, plasma protein expression profiles were compared between patients with good-CCC and those with bad-CCC. Pathway enrichment analysis using the KEGG³³, GO-biological process³⁴, and Reactome³⁵ databases highlighted strong associations between proteasome degradation and glucose metabolism pathways and the good-CCC group. Conversely, upregulated plasma proteins in the bad-CCC group were predominantly involved in extracellular matrix organization, MHC class II antigen presentation, and fatty acid degradation pathways (**Figure 2D**). These findings were further validated in the independent validation CTO cohort (**Supplementary Figure 7C**). To further investigate the immune microenvironment in patients with CTO and different CCC statuses, an enrichment analysis of immune cell infiltration was performed in both the discovery and validation cohorts using CIBERSORTx software. These results revealed that patients with bad-CCC exhibited higher proportions of infiltrating B cells, CD4 T cells, and NK cells than patients with good-CCC. This finding further demonstrated increased immune activation in patients with bad-CCC (**Supplementary Figure 7A-7B, 7D**), consistent with the pathway enrichment results. These results suggest a unique plasma proteomic signature in patients with CTO, characterized by enhanced protein degradation and ion transport in patients with good-CCC and increased vascular remodeling and immune responses in those with bad-CCC.

The proteomics-based classification model demonstrated high accuracy in predicting CCC status in patients with CTO.

To identify candidate plasma protein biomarkers for CCC status classification in patients with CTO, the relative importance scores of 122 differentially expressed proteins were evaluated using the XGBoost algorithm. The top ten plasma proteins, including SOD1, UFD1, DDX5, MIA3, TOM1, LMAN2, RAB7A, RAB10, NIF3L1, and ST13P5, demonstrated the highest feature

importance (**Figure 3A; Supplementary Table 3**). Feature selection was further performed using a forward search-based wrapper approach, leveraging previously established feature importance rankings within the training set. A 5-fold cross-validation method was employed to calculate the AUC, which served as the model's overall performance evaluation metric. The performance of the proteomics model demonstrated no improvement in the AUC with the inclusion of the top seven features (**Figure 3B**). Consequently, the top six plasma proteins (SOD1, UFD1, DDX5, MIA3, TOM1, and LMAN2) were selected as key features for the proteomic classification model (**Figure 3C**). The candidate plasma biomarkers SOD1, MIA3, and LMAN2 were significantly elevated in the good-CCC group, whereas UFD1, DDX5, and TOM1 were significantly increased in the bad-CCC group (**Figure 3C**); these findings were further validated in the external validation cohort (**Supplementary Figure 9E**).

Multivariable logistic and linear regression analyses—adjusted for age, sex, smoking status, body mass index, blood pressure, glycated hemoglobin, and medication history—identified three risk-associated and three protective proteins, corresponding to the six candidate plasma proteins as independent biomarkers of CCC status in patients with CTO. Among the upregulated proteins, TOM1 emerged as the strongest positive predictor of poor CCC status (OR = 2.58, 95% CI: 1.05–4.82, $p < 0.001$), with each unit increase in TOM1 levels corresponding to a 158% higher risk of poor CCC status. DDX5 and UFD1 showed the second and third strongest associations (OR = 1.80, 95% CI: 1.12–3.67, $p = 0.003$; OR = 1.51, 95% CI: 1.12–3.23, $p = 0.009$), respectively (**Supplementary Figure 8A-8B**). These associations remained statistically significant after adjusting for clinical confounders. Conversely, MIA3, SOD1, and LMAN2 demonstrated protective effects against poor CCC status. MIA3 exhibited the strongest negative association (OR = 0.49, 95% CI: 0.16–0.58, $p = 0.006$), followed by SOD1 (OR = 0.53, 95% CI: 0.22–0.56, $p =$

0.012). Multiple linear regression analyses confirmed these protective associations through significant negative coefficients after adjusting for clinical covariates (**Supplementary Figure 8C**). These plasma protein biomarkers demonstrated consistent robustness after adjusting for potential confounders, reinforcing their significant predictive value for CCC status in patients with CTO.

To optimize the hyperparameters of the XGBoost model, Bayesian optimization was employed using the tree-structured Parzen estimator algorithm within the training set, with the AUC used as the evaluation metric. ROC curve analysis of the training and test sets in the discovery cohort demonstrated robust performance of the proteomics-based classification model, with AUC values of 0.98 and 0.87, respectively (**Supplementary Figure 9A–9C**). The corresponding confusion matrices showed an ROC-AUC of 87.8%, specificity of 85.2%, and sensitivity of 74.0% in the test set, and an ROC-AUC of 80.4%, specificity of 74.8%, and sensitivity of 79.3% in the external validation cohort (**Supplementary Figure 9D**). SHAP analysis further highlighted plasma protein levels of SOD1, DDX5, and TOM1 as key biomarkers, showing strong associations with the CCC status in patients with CTO (**Figure 3D**). These results confirm the model's strong discriminatory power for classifying patients based on candidate plasma proteomic biomarkers.

To rigorously assess the performance, generalizability, and clinical utility of the proteomics-based classification model, both the internal hold-out test set and the external validation cohort were applied. In the independent multicenter validation cohort comprising 111 patients with CTO, the model achieved robust predictive performance, with an ROC-AUC of 0.81 (95% CI: 0.76–0.87) and a PR-AUC of 0.83 (**Figure 3E–3F**). Furthermore, the model yielded balanced accuracies of 0.83 in the test set and 0.76 in the validation cohort, with F1-scores of 0.82 and 0.71, respectively

(**Figure 3G**). These findings confirm the model's robust and consistent predictive performance across both internal and external datasets.

To further evaluate clinical utility, the six-protein biomarker panel was compared with a baseline clinical model using DeLong's test⁴⁵. The proteomics model significantly outperformed the clinical model, achieving an AUC of 0.87 (95% CI: 0.78–0.98) versus 0.66 (95% CI: 0.65–0.71) in the test set ($p = 4.3 \times 10^{-4}$), and an AUC of 0.80 (95% CI: 0.76–0.87) versus 0.57 (95% CI: 0.55–0.59) in the external validation cohort ($p = 2.1 \times 10^{-5}$) (**Supplementary Figure 9F–9G**). Collectively, these results demonstrate that integrating plasma proteomic features with an advanced XGBoost algorithm yields a robust, generalizable, and clinically superior model for predicting CCC status in patients with CTO.

Machine learning integrates clinical and proteomic features for enhanced CCC prediction

In this study, identifying a comprehensive set of biomarkers associated with CCC status in patients with CTO was the focus, encompassing both clinical and proteomic data. Although the proteomic feature-based model individually achieved excellent performance, this study sought to further enhance prediction accuracy by integrating biomarkers from multiple platforms using a machine learning framework (**Figure 4A**). The multi-platform CCC status prediction model was trained using combined features from clinical and proteomic profiles, including basophil count, medical history duration, serum chloride levels, serum HDL-C levels, RDW, and candidate plasma protein biomarkers, namely SOD1, UFD1, DDX5, MIA3, TOM1, and LMAN2.

Therefore, the trained combined model exhibited excellent classification performance, with an AUC of 0.91 (95% CI: 0.87–0.93) in the test set of the discovery dataset. In the external independent cohort, the combined model achieved an AUC of 0.82 (95% CI: 0.79–0.84) (**Figure**

4B), demonstrating good generalizability to previously unseen CTO datasets. The performance of the combined classification model, which integrated both clinical and proteomic features, demonstrated significant improvement compared with single-platform models based solely on clinical or proteomic data ($p < 0.05$, DeLong's test; **Figure 4C**). Furthermore, confusion matrix analysis revealed high sensitivity (90.9%) and specificity (92.3%) for identifying patients with good-CCC status in the discovery dataset (**Figure 4D**). This performance generalized well to the external validation cohort, with 81.4% sensitivity and 84.2% specificity (**Figure 4E**). These findings highlighted the combined model's enhanced performance in accurately classifying CCC status in patients with CTO based on multi-platform features.

To develop a clinically operable and interpretable model, a simplified single-tree XGBoost model was constructed using integrated features. The simplified model yielded a classification decision tree structure (**Figure 4F**). For model benchmarking, the performance of the single-tree and multitree XGBoost classification models were compared with that of other machine learning algorithms, including random forest, decision tree, and logistic regression. Overall, multitree models outperformed single-tree models. The single-tree XGBoost model exhibited excellent performance and generalizability, with AUC values of 0.86 and 0.82 in the test set and external validation cohort, respectively. Notably, the performance of the single-tree XGBoost model combining clinical and proteomic features significantly surpassed that of models based solely on clinical or proteomic features (**Figure 4G**). Despite using a single-tree method, the model exhibited high classification accuracy, highlighting its strong ability to classify the CCC status in patients with CTO. Ultimately, the model combining with multi-platform features demonstrated excellent classification performance, highlighting its promising potential for the clinical diagnosis and classification of CTO.

ELISA validation for clinical applicability and generalizability in an independent large-scale validation cohort

To verify the robustness and reproducibility of the proteomic findings, additional validations were conducted using ELISA in both the discovery and independent external cohorts. SOD1, MIA3, and LMAN2 were significantly elevated in the good-CCC group, whereas UFD1, DDX5, and TOM1 were higher in the bad-CCC group. The expression patterns of all six biomarkers were consistent across both analytical methods, with statistically significant differences observed for each biomarker (**Supplementary Figure 10A-10B**). Building on these results, an ELISA-based classification model was developed to predict CCC status. The model achieved strong predictive performance in the external validation cohort, with an AUC of 0.78 (95% CI: 0.77–0.93), an AUPRC of 0.79 (95% CI: 0.77–0.81), an accuracy of 0.72, and an F1-score of 0.74 (**Supplementary Figure 10C–10E; Supplementary Figure 10F–10H**).

To further evaluate generalizability in a broader clinical setting, an independent large-scale cohort of 291 patients with CTO was recruited from a separate medical center (55% male, mean age 65 $\pm$ 10 years; 54% good-CCC, 46% bad-CCC) (**Supplementary Table 1**). Consistent with prior results, ELISA validation revealed that SOD1, MIA3, and LMAN2 were significantly elevated in the good-CCC group, whereas UFD1, DDX5, and TOM1 were higher in the bad-CCC group (**Supplementary Figure 11A**). The ELISA-based classification model maintained high performance in this cohort, achieving an AUC of 0.75 (95% CI: 0.73–0.77) and an AUPRC of 0.78 (95% CI: 0.77–0.81) (**Supplementary Figure 11B–11D**). Collectively, these results provide compelling evidence that the six plasma protein biomarkers are robust, reproducible, and clinically generalizable predictors of CCC status across diverse populations of patient with CTO.

Proteomic analysis revealed molecular features associated with MACCE and established a predictive risk model for patients with CTO

Accurate prediction of MACCEs is essential for optimizing treatment strategies in patients with CTO. To elucidate plasma proteomic characteristics associated with MACCEs, a comprehensive correlation analysis was conducted between protein expression levels and MACCE occurrence. A total of 234 proteins were positively correlated with MACCEs, suggesting a potential role in disease progression, whereas 163 proteins were negatively correlated, indicating potential protective effects (**Figure 5A, Supplementary Figure 12A, and Supplementary Table 4**). Pathway enrichment analysis revealed that positively correlated proteins were primarily involved in immune activation-related pathways, including complement and coagulation cascades, the innate immune system, and extracellular matrix organization, which may accelerate disease progression through inflammation and vascular remodeling. Conversely, negatively correlated proteins were enriched in metabolic pathways such as the pentose phosphate pathway, the Warburg effect, and fat-soluble vitamin metabolism (**Figure 5B**), suggesting that impaired energy metabolism may contribute to cellular dysfunction in patients with CTO.

To determine whether clinical intervention-related factors influenced MACCE outcomes, the effects of vessel reperfusion status was assessed. Comparative analyses of MACCE incidence and survival between these subgroups revealed no significant differences (**Supplementary Figure 12C–12F**), suggesting that clinical intervention alone may not substantially affect long-term MACCE risk in patients with CTO.

Building on these findings, a multistep Cox regression analysis was conducted to construct a plasma proteome-based MACCE risk prediction model. Univariate Cox regression identified 138 candidate proteins that were significantly associated with MACCEs. Subsequent LASSO-Cox

regression minimized overfitting and selected 31 core proteins, from which nine key proteins were retained through multivariable analysis (**Figure 5C, Supplementary Figure 12B**). Elevated levels of CPM, F10, NDUFAF3, and MAVS were associated with increased MACCE risk, whereas higher levels of ITGB4, MASP1, IGHV1-46, AP3B1, and ABI3BP were associated with reduced risk. Based on these proteins, a nine-protein LASSO risk score model was developed that effectively stratified patients into high- and low-risk groups (**Figure 5D**). Patients with high-risk scores had significantly poorer outcomes, with a hazard ratio (HR) of 5.324 (95% CI: 3.648–6.412, $p < 0.0001$) for overall MACCE-free survival (**Figure 5E**). Multivariate analysis confirmed the LASSO score as an independent prognostic predictor of OS.

To validate the performance of the established model, the predictive ability of the model was rigorously assessed using time-dependent ROC analysis in an independent external validation cohort of 110 patients with CTO. The results demonstrated that the identified biomarkers exhibited strong prognostic value at 12-, 24-, and 48-month follow-up intervals. The model achieved AUC values of 0.79, 0.91, and 0.92 for 1-year, 2-year, and 4-year MACCE risk prediction, respectively (**Figure 5F**). The recalibrated models showed improved calibration, with observed MACCE risk closely aligning with predicted values at each time point (**Supplementary Figure 12G**). Notably, the recalibrated models outperformed the initial models, indicating enhanced predictive accuracy. These findings underscore the robustness and long-term reliability of the MACCE risk prediction model, demonstrating consistent and excellent prediction performance across follow-up periods. To assess clinical applicability, individual risk scores were calculated for all patients in the external independent validation cohort. These 110 patients were stratified into high- and low-risk groups based on their median LASSO scores. Kaplan–Meier survival curve analysis and log-rank tests

demonstrated significantly greater survival in the low-risk group than in the high-risk group ($P < 0.0001$, **Figure 5E**).

Taken together, a nine plasma protein-based model was developed that accurately predicts MACCE risk and effectively stratifies patients with CTO into prognostic subgroups. This proteomics-driven model demonstrates strong predictive performance, independence from clinical intervention factors, and excellent reproducibility, highlighting its potential for early risk identification and personalized management of patients with CTO.

Discussion

The global burden of CAD underscores the necessity for innovative diagnostic and therapeutic strategies. While CCC acts as a natural bypass in the setting of coronary obstruction, the underlying molecular mechanisms and relevant clinical markers remain elusive. Traditionally, research on CCC has been predominantly observational, offering limited insight into molecular mechanisms or clinically actionable markers. In this study, a label-free, quantitative proteomic approach using DIA was employed to elucidate a distinct plasma proteomic signature associated with CCC status in patients with CTO. A strong association was found between protein degradation and ion transport pathways and good-CCC status, whereas immune responses and vascular remodeling pathways predominated in patients with bad-CCC. Furthermore, classification models were established using a panel of non-invasive plasma protein biomarkers combined with clinical data, achieving high predictive accuracy in identifying CCC status in patients with CTO. The models were validated in an independent external validation cohort and demonstrated excellent classification performance and generalizability.

Consistent with previous studies, several clinical biochemical indicators were associated with cardiovascular disease. Specifically, RDW correlated with carotid atherosclerosis⁴⁷, serum chloride served as a prognostic marker in heart failure,⁴⁸ and basophil count was associated with higher mortality in stable CAD⁴⁹. Patients with CTO and poor collateral circulation (bad-CCC) generally exhibited longer medical histories, suggesting a chronic pattern of disease progression. Therefore, early and accurate classification of patients with CTO may improve treatment efficiency and clinical management. In contrast, patients with CTO and good-CCC showed significantly higher PT activity and HDL-C levels. Importantly, previous research has consistently shown that low HDL-C is a recognized risk factor for CAD and adverse outcomes^{50, 51}. These findings highlight the potential of key clinical biochemical markers to serve as a valuable clinical basis for risk stratification and individualized management in patients with CTO.

Plasma and serum are rich sources of proteins that are commonly used for clinical proteomic profiling and biomarker discovery⁵². The biological plausibility of proteomic models is high, because proteins orchestrate biological processes that, under adverse conditions, lead to disease and its clinical manifestations⁵³. Proteins link the effects of genes with the environment, age, comorbidities, lifestyle factors, and pharmacologic therapies, making them proximate to disease pathogenesis and clinical outcomes. Unlike traditional risk factors, many of which are nonmodifiable (e.g., age, sex, history of diabetes, and hypertension), proteins are well-suited for risk modeling because their expression can change with physiological conditions⁵⁴. Proteomic profiling revealed 25 upregulated and 43 downregulated proteins in patients with CTO with good CCC. Further analysis demonstrated a strong association between proteasome degradation and glucose metabolism pathways in patients with good CCC. In contrast, upregulated plasma proteins in the bad-CCC cohort were predominantly involved in extracellular matrix organization and MHC

class II antigen presentation pathways. Previous studies have reported that proteasome degradation can promote angiogenesis⁵⁵. In recent years, the role of endothelial cell metabolism in angiogenesis has received considerable attention. Accumulating evidence suggests that endothelial cells rely on aerobic glycolysis, rather than oxidative phosphorylation, to generate ATP during angiogenesis⁵⁶. Understanding these roles and underlying mechanisms will facilitate the development of new therapeutic strategies to modulate CCC formation during coronary occlusion.

The prevailing gold standard for evaluating CCC in patients with CTO involves an invasive method for determining the CFI¹⁴. Coronary arteriography is also extensively utilized for direct CCC assessment; however, its invasive nature may pose potential risks for specific patient groups⁵⁷. Therefore, identifying non-invasive CCC biomarkers is imperative for accurately classifying CCC status and facilitating risk stratification in patients with CTO. Precise determination of CCC status enables physicians to tailor treatment strategies, which may yield substantial clinical benefits and improve quality of life for individuals with CTO⁵⁸. Recently, mass spectrometry-based proteomic analysis has emerged as a robust method for identifying non-invasive circulating biomarkers in body fluids or plasma⁵⁹ and has become instrumental in the diagnosis and classification of clinical diseases^{26, 60}. Given the limitations of invasive CCC assessment methods, such as coronary arteriography and CFI measurement¹⁴, there remains an unmet need for reliable non-invasive biomarkers. Leveraging the high sensitivity and biological relevance of plasma proteomics, a clinical baseline model was developed, along with a six-protein proteomic model, and an integrated multi-platform model. Notably, the proteomic model substantially outperformed the clinical model, and the integrated classifier achieved the highest predictive performance (AUC = 0.91 in the test set; AUC = 0.82 in an external validation cohort).

894 These results highlight the translational potential of plasma-based proteomic classifiers for non-
895 invasive CCC assessment and individualized clinical management of CTO.

896 While patients with poor collateral circulation exhibit a higher incidence of MACCEs, the timing
897 of these events varies widely within both good and bad-CCC groups. This variability highlights
898 the need for individualized risk prediction, as a purely categorical approach based solely on CCC
899 status is insufficient for precise clinical decision-making. To address this limitation, this study
900 extended its framework to develop a prognostic model specifically designed to predict MACCEs
901 in patients with CTO. Accurately predicting the timing of MACCEs is crucial for enaling timely
902 interventions and optimizing therapeutic strategies, ultimately improving patient outcomes. While
903 traditional risk factors and CCC status provide valuable insights into event risk, the ability to
904 forecast the exact timing of these events remains a significant challenge. Recent biomarker
905 discovery efforts have embraced a systems-wide approach that integrates machine learning and
906 MS-based proteomics, to link circulating protein levels with organ dysfunction^{61,17,18}. Its
907 application to large patient cohorts, combined with its robustness, positions this approach as an
908 effective clinical biomarker discovery platform⁶². This study demonstrated the potential of a
909 plasma protein-based signature to accurately predict MACCEs in patients with CTO. This non-
910 invasive approach, which achieved a high AUC of 0.98 and 0.96 for one- and three-year forecasts,
911 respectively, offers a powerful tool and holds promise for personalized risk stratification and
912 improved clinical management.

913 Despite these promising results, several limitations warrant careful consideration. First, although
914 a multicenter clinical cohort design was employed, the overall sample size remained relatively
915 small, which may limit the generalizability of these findings. While the current sample size is
916 statistically adequate to support data-driven biological discovery in CTO research, further

validation in larger and more diverse patient populations is essential before these biomarkers can be translated into clinical practice. Such efforts would contribute to the development of more robust and generalizable clinical prediction models. At the mechanistic level, although the present study identified molecular pathways associated with CCC, the clinical relevance of these findings requires confirmation through prospective interventional studies. In particular, in-depth mechanistic investigations are needed to elucidate the specific molecular roles of these pathways in CCC formation. Moreover, the cross-sectional design of this study limited the current understanding of the dynamic progression of CCC. As samples were collected at a single time point, the study could not fully capture the gradual biological processes underlying collateral vessel development. Future research should prioritize longitudinal cohort studies with continuous proteomic monitoring to enhance the reliability of biomarker discovery and clarify the temporal relationship between molecular changes and CCC progression, ultimately informing more accurate patient stratification and personalized treatment strategies in CTO.

Conclusion

This study represents a seminal effort to employ integrated proteomics and machine learning techniques to unravel the complexity of CCC in patients with CTO and bridge a critical gap in understanding CCC in the context of CTO. Key protein biomarkers and plasma proteome signatures associated with CCC were elucidated, thereby providing a molecular explanation for the observed complex traits and paving the way for novel therapeutic strategies. Importantly, machine learning models based on plasma proteomic profiles outperformed existing clinical models in predicting CCC status. The developed risk prediction model for MACCEs provides a non-invasive approach for early stratification of patients at high risk of CTO. Future work should

939 focus on functional validation of the identified protein markers and the development of targeted
940 therapies aimed at modulating CCC.

941 **Data availability**

942 The newly generated mass spectrometry-based proteomics data, comprising 124 patients in the
943 discovery CTO cohorts, 111 patients in independent multicenter validation cohorts, and quality
944 control datasets for both cohorts, have been deposited in the ProteomeXchange Consortium via
945 the iProx repository⁶³. The dataset identifier is IPX0008889001 (URL:
946 <https://www.iprox.cn/page/DSV021.html?url=1716634412467s4Y3>; PASSWORD: 60BD) and
947 IPX0008889002 (URL: <https://www.iprox.cn/page/DSV021.html?url=1716634524682HPsf>;
948 PASSWORD: 76WL).

949 **Code availability**

950 To ensure reproducibility, the source code utilized for all data analyses and for generating both
951 main and supplementary figures is available in a public GitHub repository via the following link:
952 https://github.com/Ran485/plasma_proteomics_analysis_CTO.

953 **Funding**

954 The work in this paper was supported by grants from Tianshan Talent Cultivation Plan for Young
955 Top Talents Project of Xinjiang Uygur Autonomous Region (2023TSYCCX0049) and National
956 Natural Science Foundation of China (82560086, 82160054, 82260073). This study was funded
957 by Technology Innovation Leading Talent Project-High level Leading Talents (2022195955) and
958 Outstanding Talent Program of the First Affiliated Hospital of Xinjiang Medical University
959 (ZYRC202417).

Author contributions

Jun-Yi Luo and Ya-Jing Qiu had full access to all of the data in the study and take responsibility for the integrity of the data. Study concept and design: Yi-Ning Yang, Xiao-Mei Li, and Chen Ding. Collation of data: Fen Liu and Bin-bin Fang. Sample preparation and statistical analysis: Peng Ran and Sha Tian. Drafting of the manuscript: Jun-Yi Luo, Peng Ran, and Yunzhi Wang. Funding: Yi-Ning Yang, Xiao-Mei Li and Jun-Yi Luo.

Conflict of interest

The authors declare no competing financial interests.

Abbreviations

CCC,coronary collateral circulation; CTO, chronic total occlusion; AUC, area under the curve, AUPRC, area under the precision-recall curve; MACCEs, main adverse cardiovascular and cerebrovascular events; CAD, coronary artery disease; CFI, collateral flow index; cMyBP-C, cardiac myosin-binding protein C; ASCVD, atherosclerotic cardiovascular disease; DIA, data-independent acquisition; RDW, red blood cell distribution width; QC, quality control; ENPP2, ectonucleotide pyrophosphatase/phosphodiesterase 2; SOD1, superoxide dismutase 1; MIA3, melanoma inhibitory activity family member 3; PSMA1, proteasome subunit alpha type-1; DDX5, DEAD-box helicase 5; EPN1, epsin 1; UFD1, Ubiquitin recognition factor in ER-associated degradation 1; TOM1, Target of Myb protein 1; UGP2, UDP-glucose pyrophosphorylase 2; LMAN2, lectin, mannose binding 2, RAB7A, Ras-associated binding protein 7A; RAB10, member RAS oncogene family; NIF3L1, NGG1 interacting factor 3 like 1; ST13P5, Hsp70 interacting protein pseudogene 5; SHAP, SHapley Additive exPlanations; HDL-C, High-density

lipoprotein cholesterol; ELISA, enzyme-Linked Immunosorbent Assay; ITGB4, integrin subunit beta 4 ; MASP1, mannan-binding lectin associated serine protease 1; IGHV1-46, immunoglobulin Heavy Variable 1-46; AP3B1, adaptor-Related Protein Complex 3 Subunit Beta 1; ABI3BP, ABI family member 3 binding protein; LASSO, Least Absolute Shrinkage and Selection Operator.

References

1. Jones GN, Rooney C, Griffin N, et al. pRAD50: a novel and clinically applicable pharmacodynamic biomarker of both ATM and ATR inhibition identified using mass spectrometry and immunohistochemistry. *Br J Cancer*. 2018;119(10):1233-1243.
2. Tomaniak M, Katagiri Y, Modolo R, et al. Vulnerable plaques and patients: state-of-the-art. *Eur Heart J*. 2020;41(31):2997-3004.
3. Borner K, Contractor N, Falk-Krzesinski HJ, et al. A multi-level systems perspective for the science of team science. *Sci Transl Med*. 2010;2(49):49cm24.
4. Cheney A, Kearney KE, Lombardi W. Sex-Based Differences in Chronic Total Occlusion Management. *Curr Atheroscler Rep*. 2018;20(12):60.
5. Liu T, Wu Z, Liu J, et al. Metabolic syndrome and its components reduce coronary collateralization in chronic total occlusion: An observational study. *Cardiovasc Diabetol*. 2021;20(1):104.
6. Brilakis ES, Mashayekhi K, Tsuchikane E, et al. Guiding Principles for Chronic Total Occlusion Percutaneous Coronary Intervention. *Circulation*. 2019;140(5):420-433.
7. Haubner BJ, Schuetz T, Penninger JM. A reproducible protocol for neonatal ischemic injury and cardiac regeneration in neonatal mice. *Basic Res Cardiol*. 2016;111(6):64.
8. Berry C, Balachandran KP, L'Allier PL, et al. Importance of collateral circulation in coronary heart disease. *Eur Heart J*. 2007;28(3):278-291.

- 1004 9. Hansen JF. Coronary collateral circulation: clinical significance and influence on survival in
1005 patients with coronary artery occlusion. *Am Heart J.* 1989;117(2):290-295.
- 1006 10. He W, Chen P, Chen Q, et al. Cytokine storm: behind the scenes of the collateral circulation
1007 after acute myocardial infarction. *Inflamm Res.* 2022;71(10-11):1143-1158.
- 1008 11. Pries AR, Badimon L, Bugiardini R, et al. Coronary vascular regulation, remodelling, and
1009 collateralization: mechanisms and clinical implications on behalf of the working group on
1010 coronary pathophysiology and microcirculation. *Eur Heart J.* 2015;36(45):3134-3146.
- 1011 12. Zimarino M, D'Andre Matteo M, Waksman R, et al. The dynamics of the coronary collateral
1012 circulation. *Nat Rev Cardiol.* 2014;11(4):191-197.
- 1013 13. Braun MM, Stevens WA, Barstow CH. Stable Coronary Artery Disease: Treatment. *Am*
1014 *Fam Physician.* 2018;97(6):376-384.
- 1015 14. Meier P, Schirmer SH, Lansky AJ, et al. The collateral circulation of the heart. *BMC Med.*
1016 2013;11:143.
- 1017 15. Marshall KD, Edwards MA, Krenz M, et al. Proteomic mapping of proteins released during
1018 necrosis and apoptosis from cultured neonatal cardiac myocytes. *Am J Physiol Cell Physiol.*
1019 2014;306(7):C639-647.
- 1020 16. Rost HL, Rosenberger G, Navarro P, et al. OpenSWATH enables automated, targeted
1021 analysis of data-independent acquisition MS data. *Nat Biotechnol.* 2014;32(3):219-223.
- 1022 17. Schubert OT, Rost HL, Collins BC, et al. Quantitative proteomics: challenges and
1023 opportunities in basic and applied research. *Nat Protoc.* 2017;12(7):1289-1294.
- 1024 18. Jacquet S, Yin X, Sicard P, et al. Identification of cardiac myosin-binding protein C as a
1025 candidate biomarker of myocardial infarction by proteomics analysis. *Mol Cell Proteomics.*
1026 2009;8(12):2687-2699.

- 1027 19. Kaier TE, Twerenbold R, Puelacher C, et al. Direct Comparison of Cardiac Myosin-Binding
1028 Protein C With Cardiac Troponins for the Early Diagnosis of Acute Myocardial Infarction.
1029 Circulation. 2017;136(16):1495-1508.
- 1030 20. Marjot J, Kaier TE, Martin ED, et al. Quantifying the Release of Biomarkers of Myocardial
1031 Necrosis from Cardiac Myocytes and Intact Myocardium. Clin Chem. 2017;63(5):990-996.
- 1032 21. Nurmohamed NS, Belo Pereira JP, Hoogeveen RM, et al. Targeted proteomics improves
1033 cardiovascular risk prediction in secondary prevention. Eur Heart J. 2022;43(16):1569-1577.
- 1034 22. Schuermans A, Pournamdari AB, Lee J, et al. Integrative proteomic analyses across common
1035 cardiac diseases yield mechanistic insights and enhanced prediction. Nat Cardiovasc Res.
1036 2024;3(12):1516-1530.
- 1037 23. Cohen M, Rentrop KP. Limitation of myocardial ischemia by collateral circulation during
1038 sudden controlled coronary artery occlusion in human subjects: a prospective study.
1039 Circulation. 1986;74(3):469-476.
- 1040 24. Rentrop KP, Cohen M, Blanke H, et al. Changes in collateral channel filling immediately
1041 after controlled coronary artery occlusion by an angioplasty balloon in human subjects. J
1042 Am Coll Cardiol. 1985;5(3):587-592.
- 1043 25. Gong X, Zhou L, Ding X, et al. A sex-stratified long-term clinical outcome analysis in
1044 coronary chronic total occlusion patients. Biol Sex Differ. 2021;12(1):9.
- 1045 26. Nakayasu ES, Gritsenko M, Piehowski PD, et al. Tutorial: best practices and considerations
1046 for mass-spectrometry-based protein biomarker discovery and validation. Nat Protoc.
1047 2021;16(8):3737-3760.

1048 27. Demichev V, Messner CB, Vernardis SI, et al. DIA-NN: neural networks and interference
1049 correction enable deep proteome coverage in high throughput. *Nat Methods*. 2020;17(1):41-
1050 44.

1051 28. Cox J, Hein MY, Luber CA, et al. Accurate proteome-wide label-free quantification by
1052 delayed normalization and maximal peptide ratio extraction, termed MaxLFQ. *Mol Cell*
1053 *Proteomics*. 2014;13(9):2513-2526.

1054 29. Li Y, Porta-Pardo E, Tokheim C, et al. Pan-cancer proteogenomics connects oncogenic
1055 drivers to functional states. *Cell*. 2023;186(18):3921-3944 e3925.

1056 30. Petralia F, Ma W, Yaron TM, et al. Pan-cancer proteogenomics characterization of tumor
1057 immunity. *Cell*. 2024;187(5):1255-1277 e1227.

1058 31. Harris L, Fondrie WE, Oh S, et al. Evaluating Proteomics Imputation Methods with
1059 Improved Criteria. *J Proteome Res*. 2023;22(11):3427-3438.

1060 32. Ritchie ME, Phipson B, Wu D, et al. limma powers differential expression analyses for
1061 RNA-sequencing and microarray studies. *Nucleic Acids Res*. 2015;43(7):e47.

1062 33. Kanehisa M, Goto S. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res*.
1063 2000;28(1):27-30.

1064 34. Ashburner M, Ball CA, Blake JA, et al. Gene ontology: tool for the unification of biology.
1065 The Gene Ontology Consortium. *Nat Genet*. 2000;25(1):25-29.

1066 35. Fabregat A, Jupe S, Matthews L, et al. The Reactome Pathway Knowledgebase. *Nucleic*
1067 *Acids Res*. 2018;46(D1):D649-D655.

1068 36. Lawrence MS, Stojanov P, Polak P, et al. Mutational heterogeneity in cancer and the search
1069 for new cancer-associated genes. *Nature*. 2013;499(7457):214-218.

37. Chicco D, Totsch N, Jurman G. The Matthews correlation coefficient (MCC) is more reliable than balanced accuracy, bookmaker informedness, and markedness in two-class confusion matrix evaluation. *BioData Min.* 2021;14(1):13.
38. Brodersen KH, Ong CS, Stephan KE, et al. The Balanced Accuracy and Its Posterior Distribution. 2010 20th International Conference on Pattern Recognition 2010:3121-3124.
39. Go AS, Mozaffarian D, Roger VL, et al. Heart disease and stroke statistics--2014 update: a report from the American Heart Association. *Circulation.* 2014;129(3):e28-e292.
40. Chen H, Covert IC, Lundberg SM, et al. Algorithms to estimate Shapley value feature attributions. *Nature Machine Intelligence.* 2023;5(6):590-601.
41. Chen H, Lundberg SM, Lee SI. Explaining a series of models by propagating Shapley values. *Nat Commun.* 2022;13(1):4512.
42. Bolón-Canedo V, Alonso-Betanzos A. Ensembles for feature selection: A review and future trends. *Information Fusion.* 2019;52:1-12.
43. Collins RL, Glessner JT, Porcu E, et al. A cross-disorder dosage sensitivity map of the human genome. *Cell.* 2022;185(16):3041-3055 e3025.
44. Yang L, Shami A. On hyperparameter optimization of machine learning algorithms: Theory and practice. *Neurocomputing.* 2020;415:295-316.
45. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics.* 1988;44(3):837-845.
46. Sammut SJ, Crispin-Ortuzar M, Chin SF, et al. Multi-omic machine learning predictor of breast cancer therapy response. *Nature.* 2022;601(7894):623-629.

1092 47. Wen Y. High red blood cell distribution width is closely associated with risk of carotid artery
1093 atherosclerosis in patients with hypertension. *Exp Clin Cardiol.* 2010;15(3):37-40.

1094 48. Zandijk AJL, van Norel MR, Julius FEC, et al. Chloride in Heart Failure: The Neglected
1095 Electrolyte. *JACC Heart Fail.* 2021;9(12):904-915.

1096 49. Pizzolo F, Castagna A, Olivieri O, et al. Basophil Blood Cell Count Is Associated With
1097 Enhanced Factor II Plasma Coagulant Activity and Increased Risk of Mortality in Patients
1098 With Stable Coronary Artery Disease: Not Only Neutrophils as Prognostic Marker in
1099 Ischemic Heart Disease. *J Am Heart Assoc.* 2021;10(5):e018243.

1100 50. Natarajan P, Ray KK, Cannon CP. High-density lipoprotein and coronary heart disease:
1101 current and future therapies. *J Am Coll Cardiol.* 2010;55(13):1283-1299.

1102 51. Wilson PW, Abbott RD, Castelli WP. High density lipoprotein cholesterol and mortality.
1103 The Framingham Heart Study. *Arteriosclerosis.* 1988;8(6):737-741.

1104 52. Soni RK. High-Throughput Plasma Proteomic Profiling. *Methods Mol Biol.* 2022;2546:411-
1105 420.

1106 53. Williams SA, Kivimaki M, Langenberg C, et al. Plasma protein patterns as comprehensive
1107 indicators of health. *Nat Med.* 2019;25(12):1851-1857.

1108 54. Williams SA, Ostroff R, Hinterberg MA, et al. A proteomic surrogate for cardiovascular
1109 outcomes that is sensitive to multiple mechanisms of change in risk. *Sci Transl Med.*
1110 2022;14(639):eabj9625.

1111 55. Lai CR, Wang HH, Chang HH, et al. Enhancement of Farnesoid X Receptor Inhibits
1112 Migration, Adhesion and Angiogenesis through Proteasome Degradation and VEGF
1113 Reduction in Bladder Cancers. *Int J Mol Sci.* 2022;23(9).

- 1114 56. Du W, Ren L, Hamblin MH, et al. Endothelial Cell Glucose Metabolism and Angiogenesis.
1115 Biomedicines. 2021;9(2).
- 1116 57. Ghobrial M, Haley HA, Gosling R, et al. The new role of diagnostic angiography in coronary
1117 physiological assessment. Heart. 2021;107(10):783-789.
- 1118 58. van Praag HM. CSF 5-HIAA and suicide in non-depressed schizophrenics. Lancet.
1119 1983;2(8356):977-978.
- 1120 59. Khoo A, Liu LY, Nyalwidhe JO, et al. Proteomic discovery of non-invasive biomarkers of
1121 localized prostate cancer using mass spectrometry. Nat Rev Urol. 2021;18(12):707-724.
- 1122 60. Deng M, Ran P, Chen L, et al. Proteogenomic characterization of cholangiocarcinoma.
1123 Hepatology. 2022.
- 1124 61. Niu L, Thiele M, Geyer PE, et al. Noninvasive proteomic biomarkers for alcohol-related
1125 liver disease. Nat Med. 2022;28(6):1277-1287.
- 1126 62. Messner CB, Demichev V, Bloomfield N, et al. Ultra-fast proteomics with Scanning
1127 SWATH. Nat Biotechnol. 2021;39(7):846-854.
- 1128 63. Ma J, Chen T, Wu S, et al. iProX: an integrated proteome resource. Nucleic Acids Res.
1129 2019;47(D1):D1211-D1217.

1130

1131

Figure 1

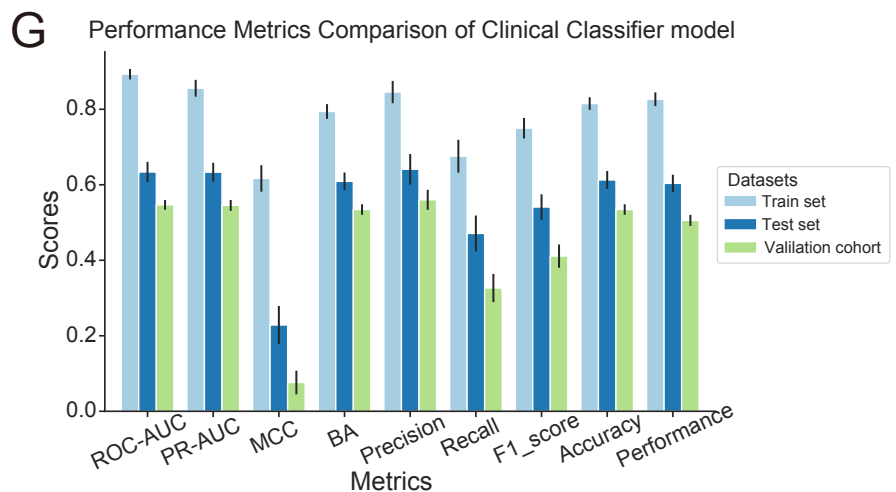
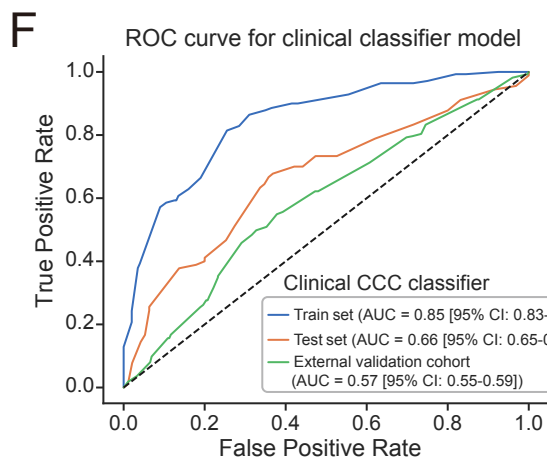
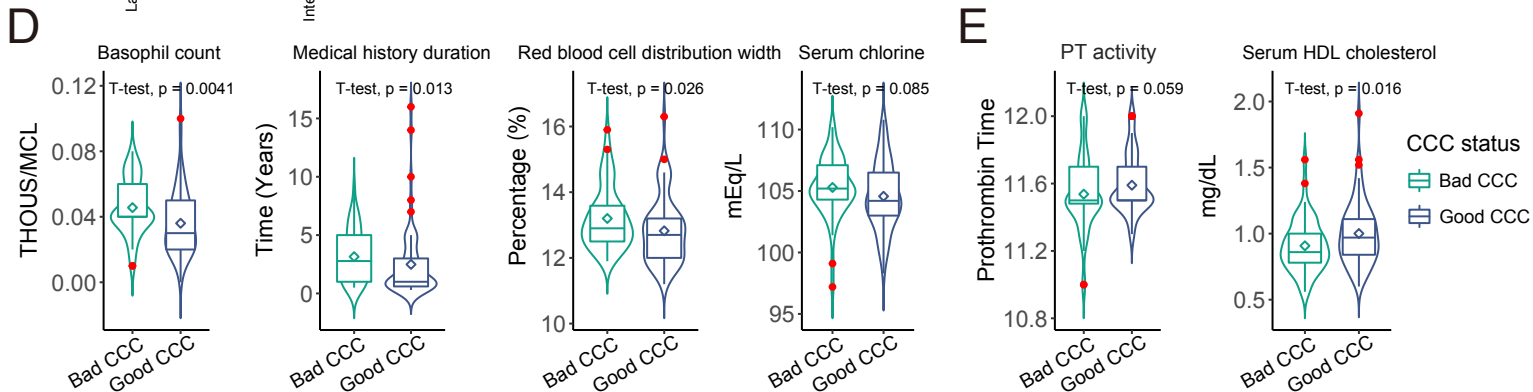
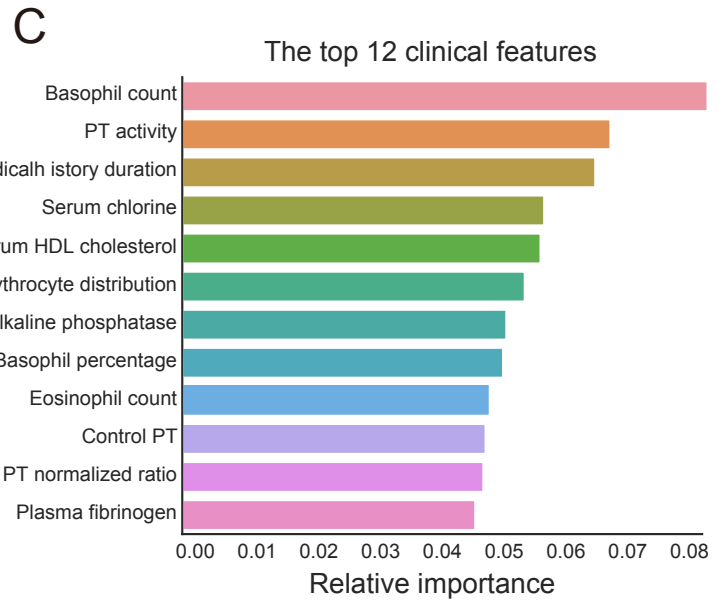
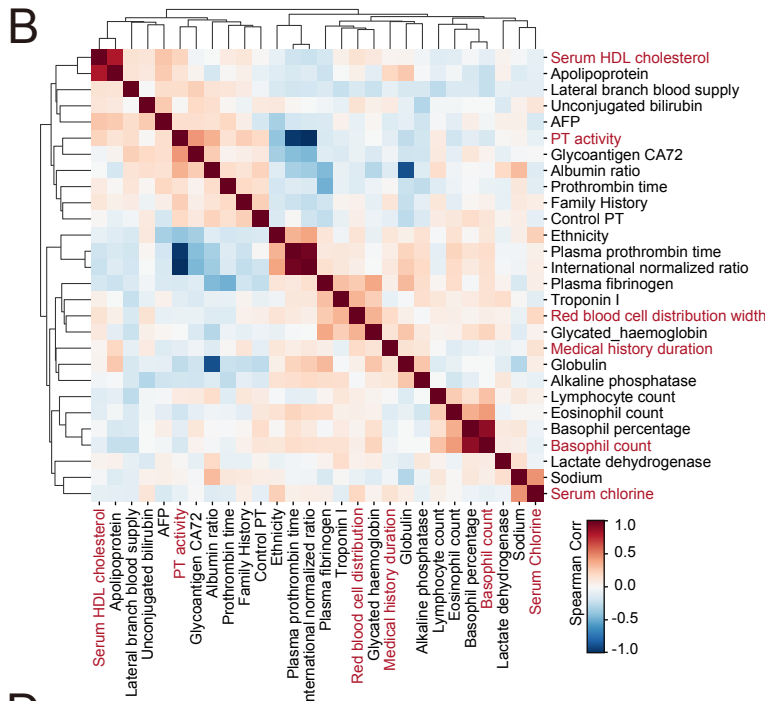
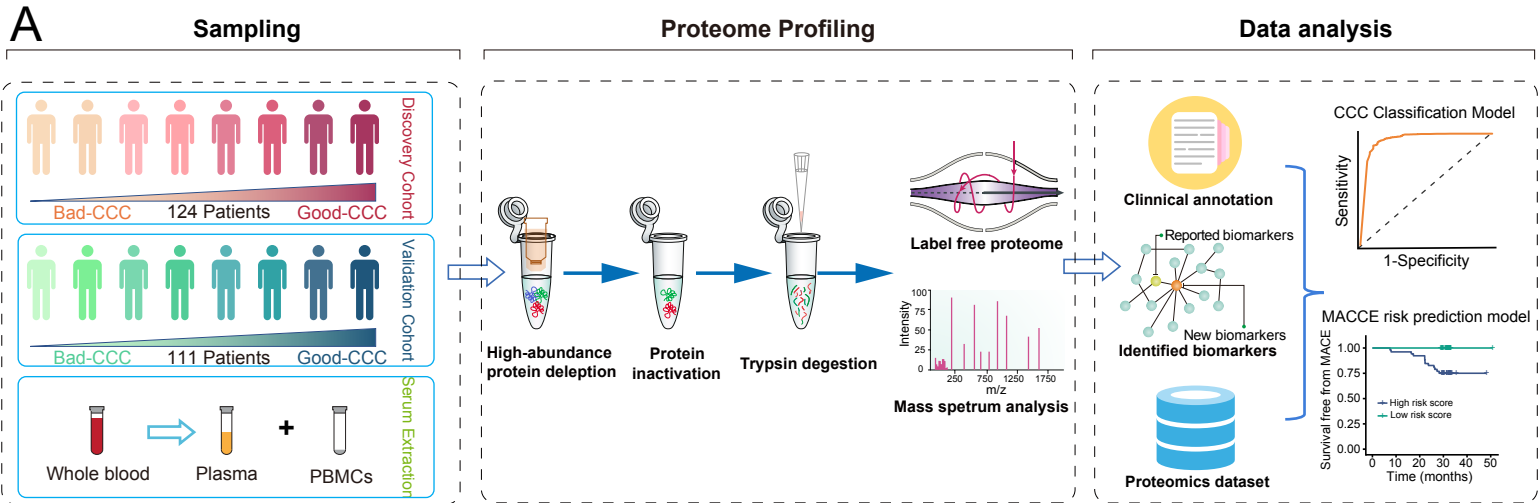


Figure 2

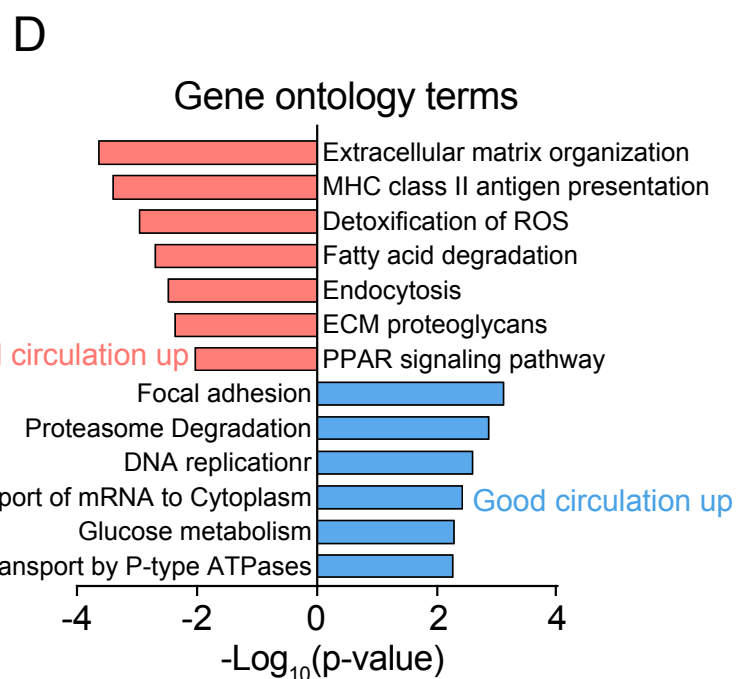
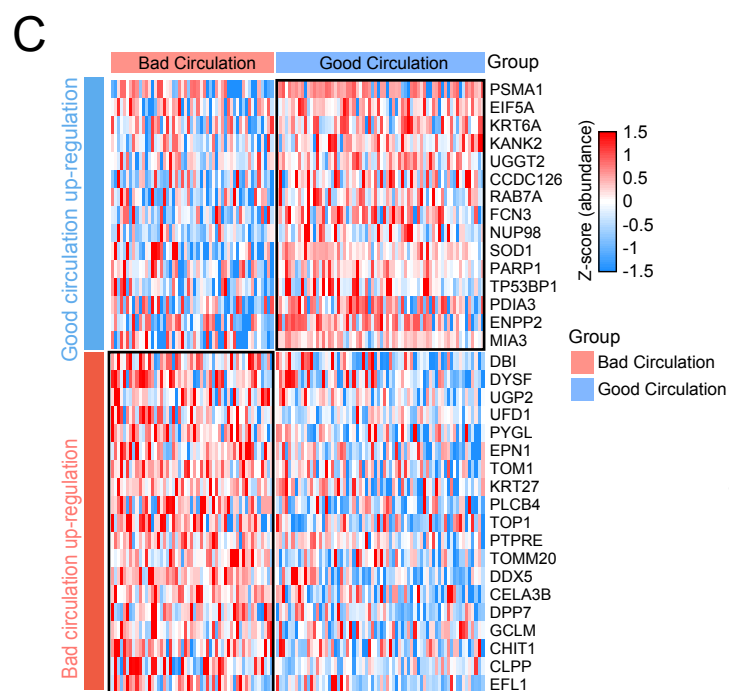
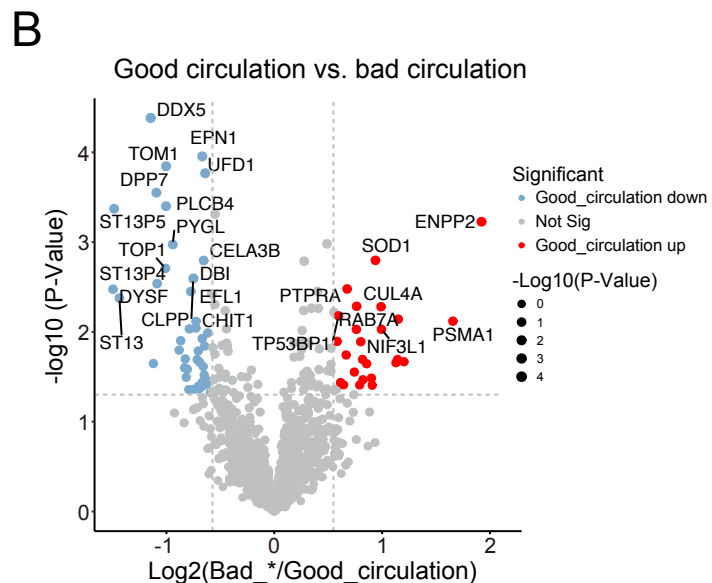
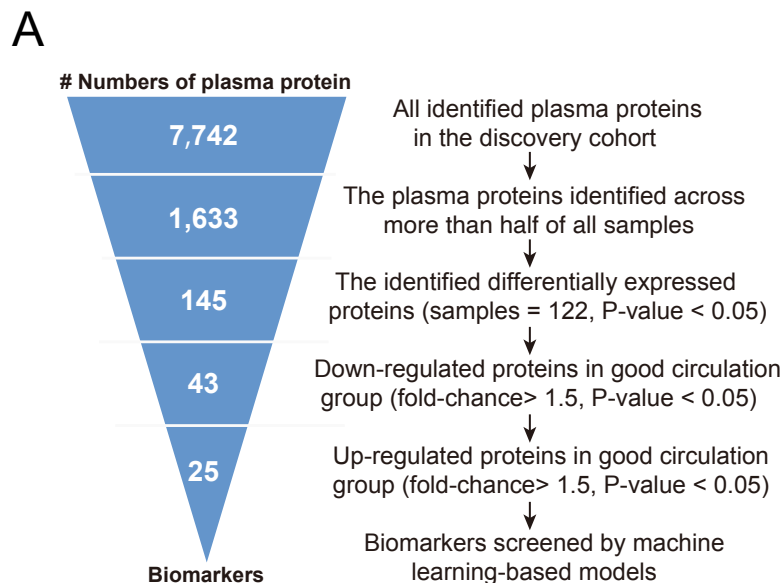
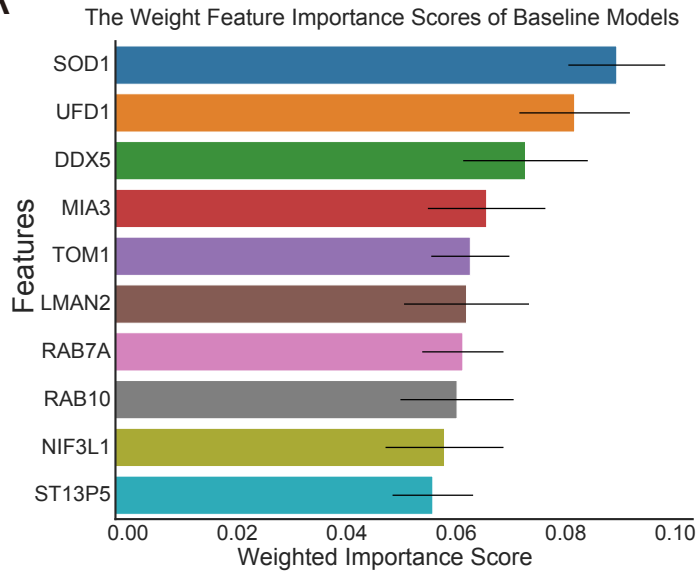
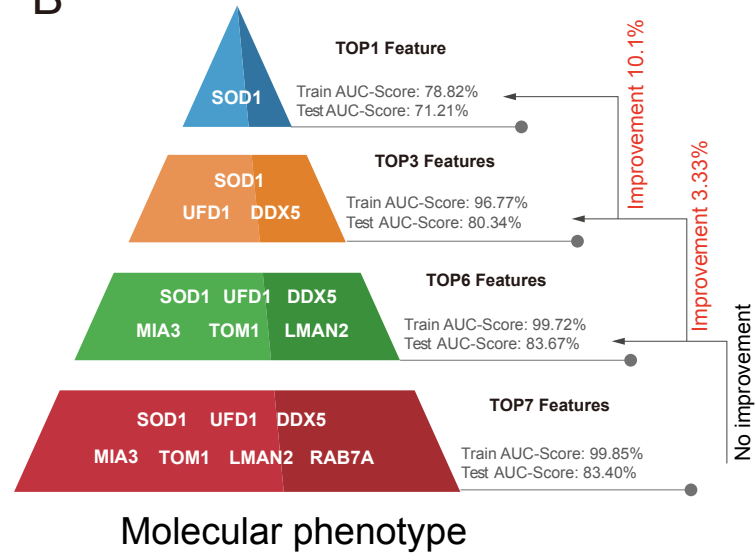


Figure 3

A

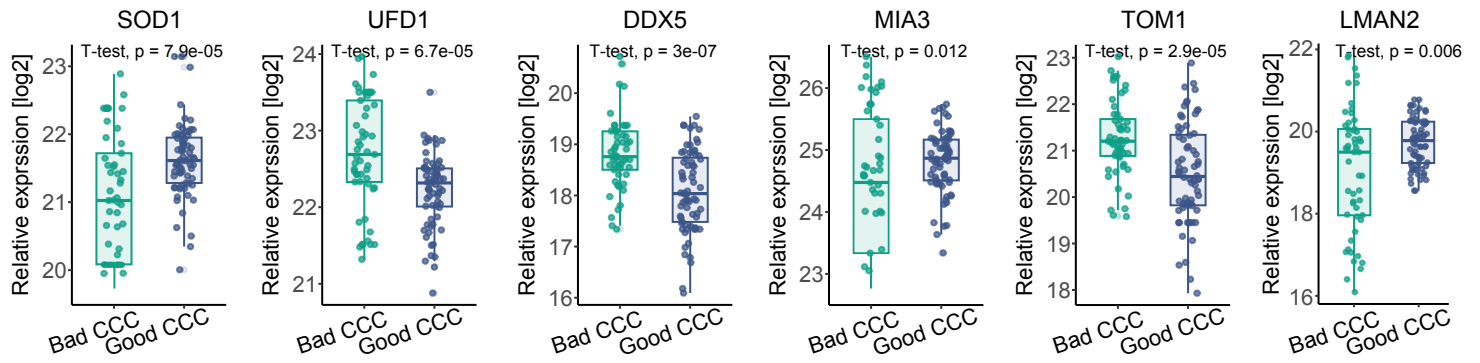


B

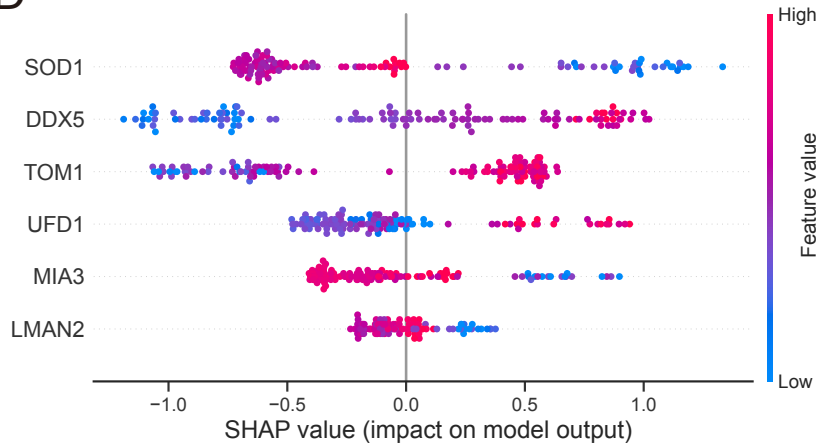


C

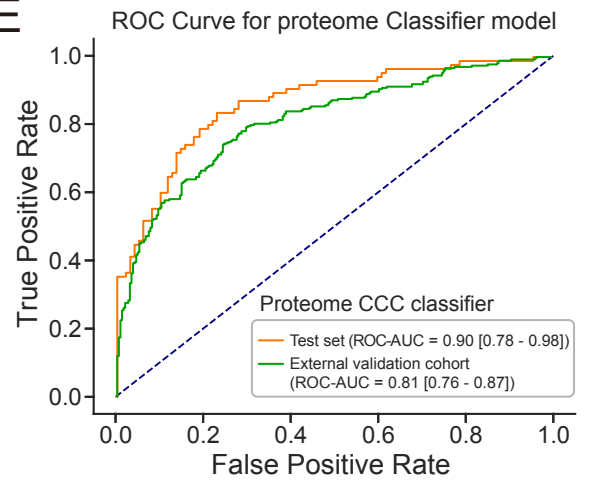
The candidate protein plasma biomarkers in discovery cohort



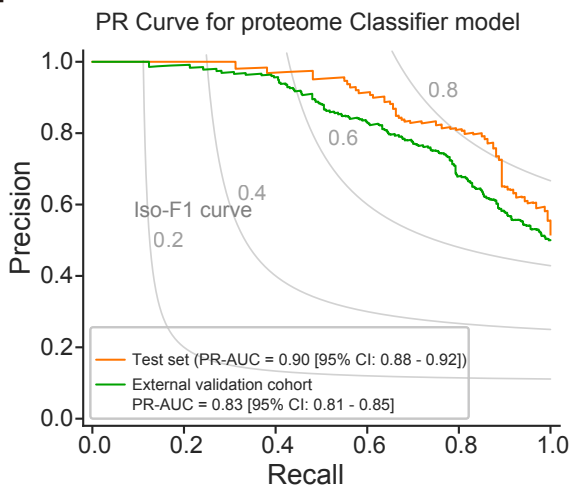
D



E



F



G

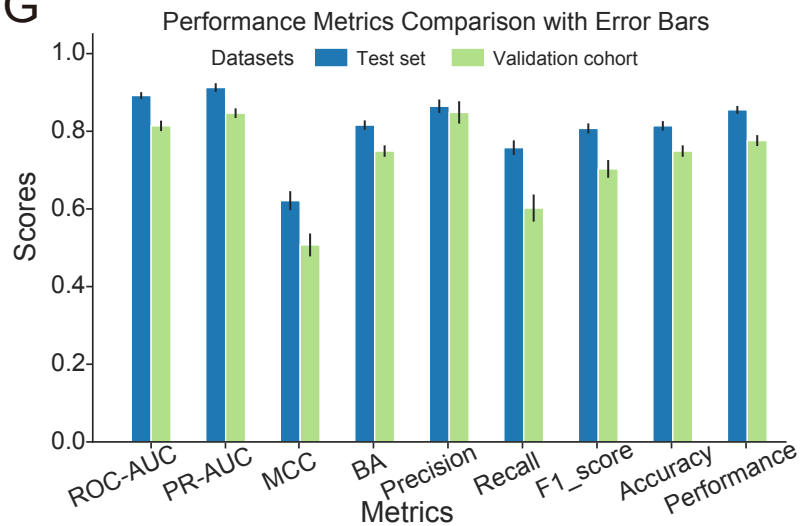
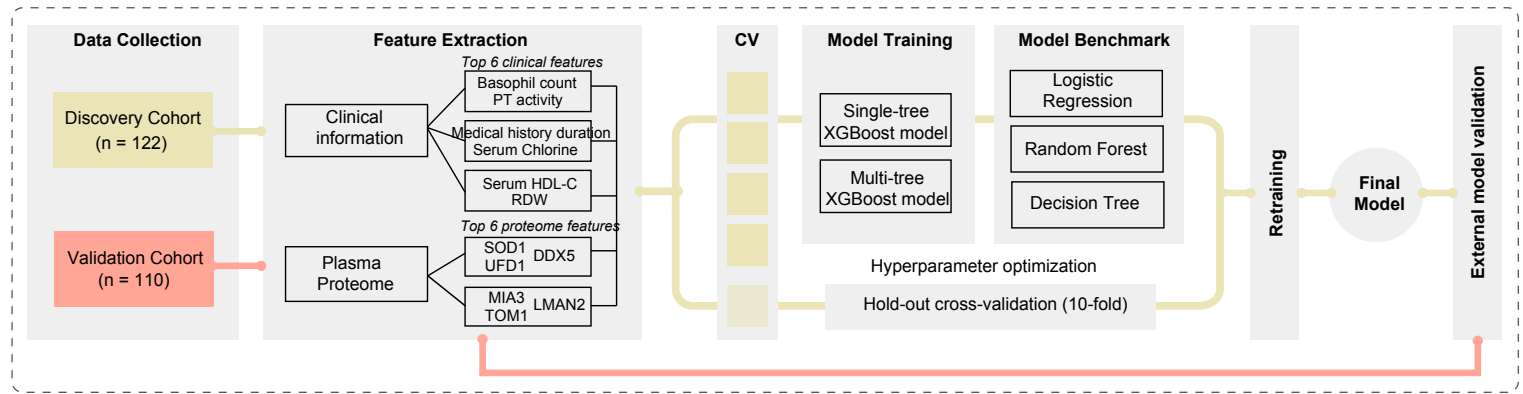
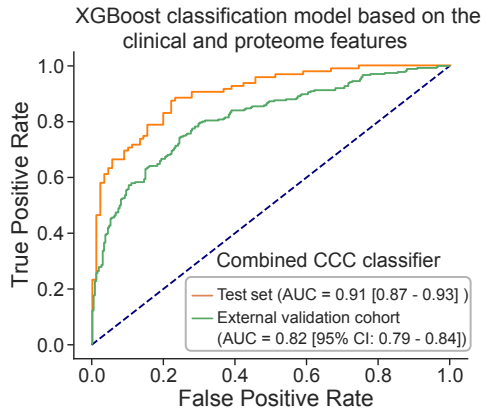


Figure 4

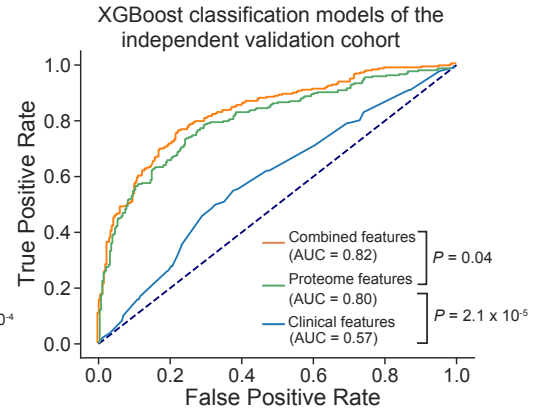
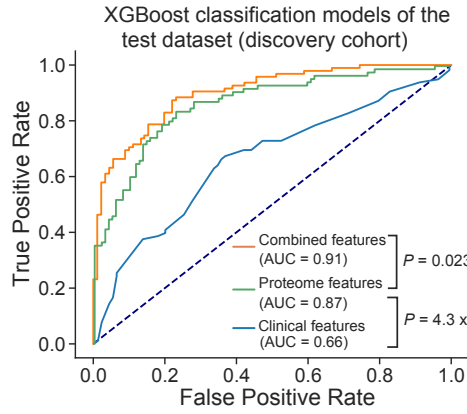
A



B



C



D

Confusion matrix of the test dataset

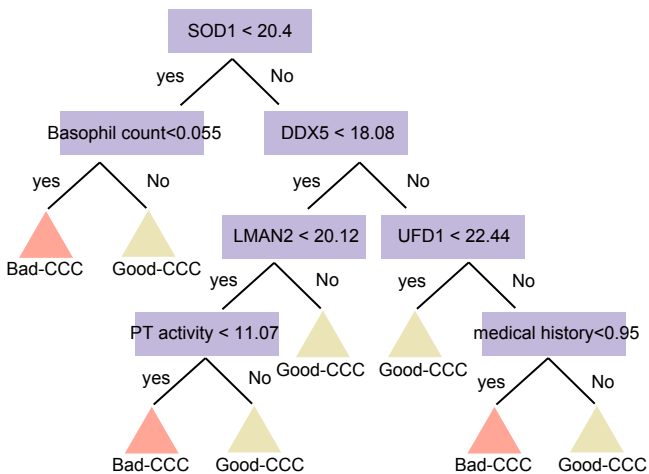
Discovery cohort	Good circulation	Bad circulation	
Good circulation	20	1	
Bad circulation	2	12	Totals
Totals	22	13	35
Correct	20	12	32
Sensitivity (%)	90.9		
Specificity (%)		92.3	91.4

E

Confusion matrix of the external validation cohort

Validation cohort	Good circulation	Bad circulation	
Good circulation	57	6	
Bad circulation	13	32	Totals
Totals	70	38	108
Correct	57	32	89
Sensitivity (%)	81.4		
Specificity (%)		84.2	82.4

F The top6 clinical and top6 plasma proteins



G

The AUC Scores of model benchmark with 100 iterations

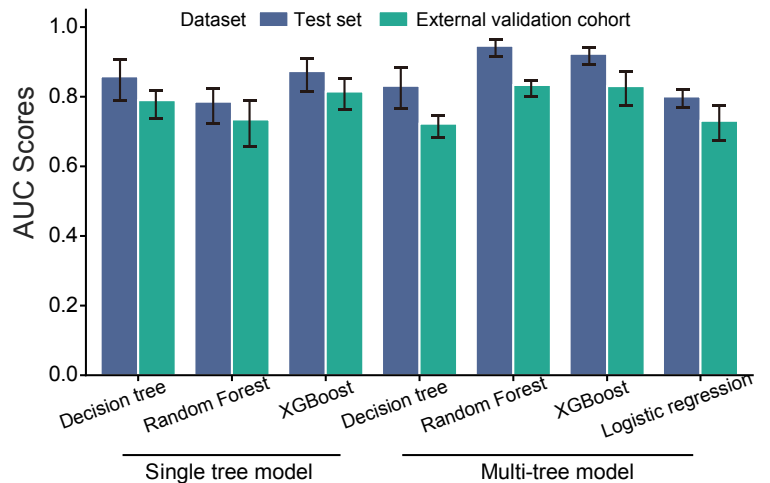


Figure 5

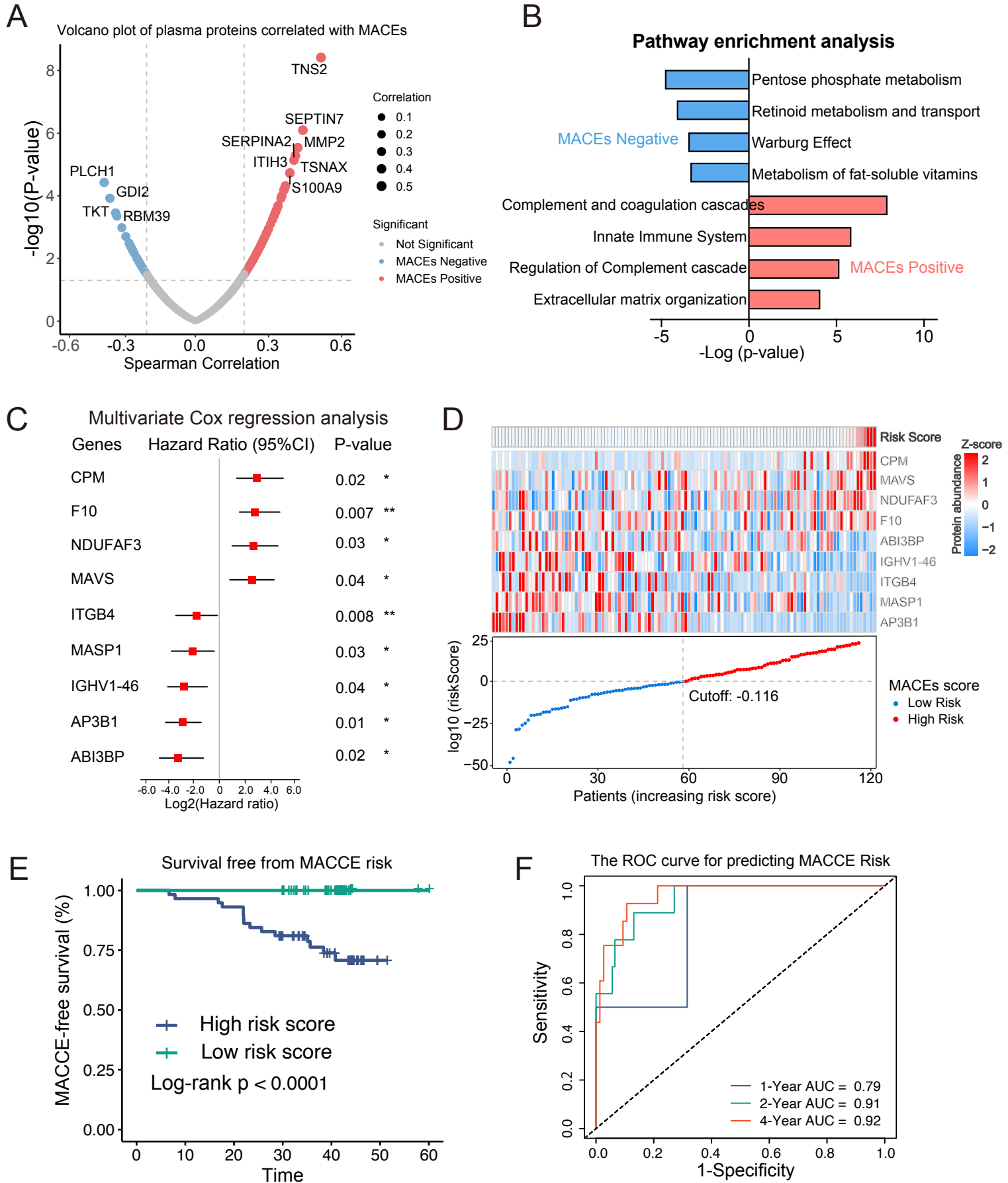


Figure Legends

Figure 1. Overview of the research workflow and study design.

Figure 1A: Representation of the research workflow, beginning with the collection of two independent prospective cohorts of patients with CTO (discovery cohort: n = 124 patients recruited from the First Affiliated Hospital of Xinjiang Medical University; validation cohort: n = 111 patients recruited from the People's Hospital of Xinjiang Uygur Autonomous Region Hospital). Plasma samples were isolated from peripheral blood collected from patients with CTO. The isolated proteins were then enzymatically digested with trypsin. The resulting peptides were desalted using a pre-prepared C18 column and subsequently aliquoted for mass spectrometry analysis. Finally, we analyzed biological mechanisms based on plasma proteomics and constructed machine learning-based classifiers and predictive models.

Figure 1B: The heatmap displays the Spearman correlations of the clinical features in the discovery cohort.

Figure 1C: Bar plot illustrating the top 12 clinical features based on the XGBoost model; the X-axis depicts the relative importance of the model features.

Figure 1D-E: Violin plot illustrating the distributions of clinical features in the good-CCC and bad-CCC groups.

Figure F: The receiver operating characteristic (ROC) curves for the training and test datasets, derived from the discovery cohort (121 patients with CTO), represent a 70/30 split for the training and testing datasets, respectively. The model was further evaluated in the external independent validation cohort.

Figure G: Bar plot illustrating the performance metrics comparison of the XGBoost classifier model based on clinical features.

Figure 2. Proteome profiling revealed molecular characteristics associated with diverse CCC status.

Figure 2A: Schematic showing the triage of plasma proteins to identify differentially expressed proteins and candidate biomarkers for CTO patients.

Figure 2B: Volcano plot showing the identified significantly differentially expressed proteins between good-CCC and bad-CCC groups (Benjamini–Hochberg adjusted $p < 0.05$). Red and blue indicate plasma proteins that were upregulated and downregulated, respectively, in the good-CCC subgroup.

Figure 2C: Heatmap illustrating the significantly differentially expressed plasma proteins.

Figure 2D: Pathway enrichment analysis revealing pathways significantly enriched in good-CCC (blue) and bad-CCC (red) patients with CTO. A p value < 0.05 was considered significant.

Figure 3. Development of a classification model utilizing plasma proteomics data to predict CCC status in patients with CTO.

Figure 3A: The bar chart depicts the top 12 plasma proteins identified by multiple baseline models (XGBoost, random forest, LightGBM, CatBoost, SVC, and logistic regression) as the most important features. The feature importance is represented by the average absolute value of multiple baseline models for each feature, ranked from highest to lowest. The error bars indicate the standard errors of the importance estimates.

Figure 3B: The feature pyramid graph illustrating the AUC score improvement of feature selection based on plasma protein profiling. A multitree XGBoost model utilizing the top-X features from 121 CTO samples was trained with specific parameters: a maximum depth of 4, a learning rate of 0.2, and a regularization parameter (α) value of 1.

Figure 3C: The boxplot shows that the candidate protein plasma biomarkers exhibited significant differences between the good-CCC (blue) and bad-CCC (green) subgroups of CTO patients. Student's t-test was used to assess statistical significance, considering p -values < 0.05 to indicate statistical significance.

Figure 3D: The distribution of SHAP (Shapley Additive exPlanations) values illustrates the impact of individual features on the model's output. This scatter plot visualizes the impact of individual features on the model's output, with the SHAP values plotted on the x-axis. Each dot represents a sample, where its position on the x-axis indicates the SHAP value and, hence, the magnitude and direction of the feature's impact. Features are listed on the y-axis and are color-coded; pink

indicates a higher feature value, and blue indicates a lower feature value. The distribution of SHAP values for each feature illustrates the variability of the feature's influence across the dataset.

Figure 3E: The receiver operating characteristic (ROC) curves for the proteome XGBoost classifier model on the training and test datasets and the external independent validation cohort.

Figure 3F: Precision-Recall curves for the XGBoost model in the test set (orange) and an external validation cohort (green). The model demonstrates strong classification performance, with an Area Under the Curve (AUC) of 0.9 (95% CI: 0.88–0.92) in the test set and an AUC of 0.83 (95% CI: 0.81–0.85) in the external validation cohort. These results highlight the model's robustness and potential clinical applicability for CCC classification.

Figure 3G: Bar plot illustrating the performance metrics comparison of the XGBoost classifier model based on the candidate plasma biomarkers.

Figure 4. Prediction of CTO patients' CCC status with an ensemble machine learning model integrating clinical and proteomics data.

Figure 4A: Schematic of the machine learning framework. CV, cross-validation.

Figure 4B: The receiver operating characteristic (ROC) curves for the test set and the external independent validation cohort, respectively, with the combined clinical and plasma proteome features.

Figure 4C: The ROC curves demonstrate the superior generalization performance of the constructed ensemble classification model that combines clinical and plasma proteome features, compared to models based solely on either clinical or plasma proteome features.

Figure 4D-4E: Confusion matrix of the constructed ensemble classification model in the test set and the external independent validation cohort.

Figure 4F: A decision rule using the combined clinical and plasma protein features and their absolute threshold values.

Figure 4G: The model's classification performance was benchmarked by comparing the AUC of XGBoost with those of logistic regression, random forest, and decision tree models. The bar plot illustrates these comparisons, showcasing single-tree models in the left and multitree models in the right. The models were trained using combined clinical and plasma proteome features, with

each model being iterated 100 times using different random number seeds to assess the robustness of these models.

Figure 5. The development of a MACCE risk prediction model for CTO patients utilizing plasma proteomic data.

Figure 5A: The volcano plot demonstrates a significant correlation between plasma proteins and MACCEs. A p value < 0.05 and an absolute value of Pearson's correlation coefficient > 0.3 were considered significant.

Figure 5B: Pathway enrichment analysis of positively (red) and negatively (blue) proteins correlated with MACCEs in CTO patients. A p value < 0.05 was considered significant.

Figure 5C: Forest plots of the multivariate Cox regression analyses in the discovery cohort.

Figure 5D: Establishing an optimal proteomics-level MACCE risk model using LASSO–Cox regression. The heatmap displays the protein abundance of these 9 plasma proteins in each patient with CTO. The bottom scatter plot presents the LASSO score across CTO patients, with patients ordered according to their LASSO score. The CTO patients were then stratified into two risk groups based on the median LASSO score.

Figure 5E: Kaplan–Meier survival curves comparing survival free from MACCEs between the independent validation cohort's LASSO score-low and score-high groups. P values were calculated using the log-rank test.

Figure 5F: The time-dependent ROC curves of the predicted plasma protein signature in the independent validation cohort.

Supplementary Figure 1. Performance of the XGBoost classification model using clinical features.

Supplementary Figure 1A: Flow diagram of cohort recruitment and analysis. Patients with chronic total occlusion (CTO) were consecutively enrolled from three independent cohorts: the Discovery Cohort (First Affiliated Hospital of Xinjiang Medical University, 2017–2019; $n = 338$), Validation Cohort 1 (People's Hospital of Xinjiang Uygur Autonomous Region, 2019–2023; $n = 268$), and Validation Cohort 2 (People's Hospital of Xinjiang Uygur Autonomous Region, 2020–2023; $n = 468$). Patients were excluded for prior coronary artery bypass grafting, acute myocardial infarction,

complications of active infectious disease, rheumatologic disease or malignancy, and unavailability of Rentrop grades. After exclusions, 216, 158, and 177 eligible CTO patients were identified in the Discovery Cohort, Validation Cohort 1, and Validation Cohort 2, respectively. Additional exclusions were made for absence of clinical information or failure to meet quality-control (QC) criteria. The final analyses included 122 patients in the Discovery Cohort, 110 patients in Validation Cohort 1, and 291 patients in Validation Cohort 2. Associations with coronary collateral circulation (CCC) were investigated across clinical features, plasma proteome profiling, and ELISA validation of candidate proteins, with plasma proteomics performed only in the Discovery and Validation Cohort 1.

Supplementary Figure 1B-1C: The confusion matrix of the constructed XGBoost classification model for the training, test, and the external independent validation cohort sets.

Supplementary Figure 1D: Precision-Recall curves for the XGBoost model in the test set (orange) and an external validation cohort (green).

Supplementary Figure 2. Schematic workflow for developing the CCC machine learning model and evaluating it in the external independent validation cohort.

Supplementary Figure 2A: The discovery cohort was divided into a training set (70%) and a test set (30%) with the stratified random split approach during model derivation. The feature selection was conducted on the training set to identify the most significant features contributing to the model's performance in distinguishing between different classes. Subsequently, hyperparameter optimization was performed on the training set, and the optimal parameters were selected. In the machine learning architecture, hyperparameters were tuned using the Bayesian optimization with the tree-structured Parzen estimator (TPE) algorithm, employing five-fold cross-validation. Finally, the final models were then tested on external, independent, unseen datasets to evaluate their generalizability performance. The external independent validation CTO cohorts (n = 110 samples).

Supplementary Figure 3. Multivariate analyses for the clinical risk factors with CCC status in patients with CTO.

Supplementary Figure 3A: Logistic regression analysis of clinical factors associated with CCC status, presenting odds ratios (ORs) with 95% confidence intervals (CIs). Basophil count, chlorine

levels, medical history duration, and PT activity were identified as independent predictors of CCC status. Statistically significant p-values ($p < 0.05$) are indicated.

Supplementary Figure 3B: Multiple linear regression analysis of the same clinical factors. Coefficients are shown with 95% CIs, with significant p values ($p < 0.05$) indicated.

Supplementary Figure 3C: Multiple linear regression analysis of the same clinical factors, adjusted for potential confounders, including age, sex, smoking status, body mass index (BMI), blood pressure, glycated hemoglobin, and medication history. Coefficients are shown with 95% CIs, with significant p values ($p < 0.05$) indicated.

Supplementary Figure 4. Quantitative and qualitative analysis of plasma proteomics-based proteins in the CTO cohort.

Supplementary Figure 4A-4B: The bar plot illustrates the quantified protein numbers in the discovery and validation cohorts.

Supplementary Figure 4C-4D: The scatter plot illustrates the protein dynamic range in the discovery and validation cohorts.

Supplementary Figure 4E-4F: The boxplot depicts the distribution of quantified protein LFQ intensity in the discovery and validation cohorts.

Supplementary Figure 5. Quality control of the proteomics data of the CTO discovery and validation cohorts.

Supplementary Figure 5A-5B: Pearson correlation analysis of the quality control data of the CTO discovery and validation cohorts.

Supplementary Figure 5C-5D: The density plot illustrates the distribution of label-free quantification (LFQ) abundance in the quality control data for both the CTO discovery and validation cohorts.

Supplementary Figure 5E-5F: Principal component analysis (PCA) was employed to detect batch effects in the CTO proteomic data from both the discovery and validation cohorts.

Supplementary Figure 5G-5H: The density plot depicts the distribution of LFQ abundance across the proteomic profiling data of the CTO patients in both the discovery and validation cohorts.

Supplementary Figure 6. Data Completeness Analysis for Proteomic Dataset.

Supplementary Figure 6A: Histogram showing the distribution of data completeness across proteins. The x-axis represents the data completeness, defined as the proportion of non-missing values, while the y-axis indicates the proportion of proteins with corresponding data completeness. Most proteins have data completeness close to 1.0, with a significant proportion showing completeness between 0.9 and 1.0.

Supplementary Figure 6B: Density plot illustrating the data completeness distribution. The plot shows peaks around completeness values of 0.7 and 1.0, indicating clusters of proteins with relatively high completeness.

Supplementary Figure 6C: Table RL1 summarizes the data completeness by bin centers. The table includes the bin center for data completeness, the number of proteins in each bin, and the proportion of total proteins. Notably, 45.5% of proteins exhibit 100% data completeness, 36.4% of proteins exhibit 70%-90% data completeness, while only smaller proportions of proteins have completeness between 0.5 and 0.6.

Supplementary Figure 7. Proteome profiling revealed molecular characteristics associated with diverse CCC statuses.

Supplementary Figure 7A: The stacked plot shows the proportions of infiltrating immune cell types identified with CIBERSORT in the discovery cohort.

Supplementary Figure 7B: Bar plot depicting the cell composition differences between the good-CCC (green) and bad-CCC (red) subgroups of CTO patients in the discovery CTO cohort.

Supplementary Figure 7C: Pathway enrichment analysis revealing significantly enriched pathways related to good-CCC (red) and bad-CCC (blue) in CTO patients in the independent validation cohort. A p value < 0.05 was considered significant.

Supplementary Figure 7D: Bar plot depicting the differences in cell composition between the good-CCC (red) and bad-CCC (green) subgroups of CTO patients in the independent validation cohort.

Supplementary Figure 8. Multivariate analyses on the candidate plasma protein biomarkers to adjust for confounding factors.

Supplementary Figure 8A: Logistic regression analysis of candidate plasma protein biomarkers associated with CCC status, presenting odds ratios (ORs) with 95% confidence intervals (CIs). TOM1, DDX5, UFD1, LMAN2, SOD1, and MIA3 were identified as independent predictors of CCC status. Statistically significant p values ($p < 0.05$) are indicated.

Supplementary Figure 8B: Multiple linear regression analysis of the candidate plasma biomarkers. Coefficients are shown with 95% CIs, with significant p values ($p < 0.05$) indicated.

Supplementary Figure 8C: Multiple linear regression analysis of the candidate plasma protein biomarkers, adjusted for potential confounders, including age, sex, smoking status, body mass index (BMI), blood pressure, glycated hemoglobin, and medication history. Coefficients are shown with 95% CIs, with significant p values ($p < 0.05$) indicated.

Supplementary Figure 9. Hyperparameter optimization in the classification model and evaluating its performance.

Supplementary Figure 9A: XGBoost training and test accuracy. This line graph depicts the training and test accuracy of an XGBoost model over 250 iterations. The blue and orange lines represent the training and test accuracy, respectively, indicating the model's performance stability across iterations.

Supplementary Figure 9B: Training loss and test accuracy over iterations. This dual-axis line graph presents the training log loss (red line) and test accuracy (blue line) over 250 iterations. The left y-axis represents the training log loss, while the right y-axis shows the test accuracy. This visualization depicts the model's convergence, aiming to minimize loss while maximizing accuracy.

Supplementary Figure 9C: Training and validation loss curves for the XGBoost model across boosting iterations. The blue line represents the loss on the training set, while the orange line shows the loss on the validation set. Both curves demonstrate rapid convergence within the first 50 iterations, with the training loss continuing to decrease and the validation loss stabilizing after approximately 100 iterations, indicating an appropriate early stopping point to prevent overfitting.

Supplementary Figure 9D: The confusion matrix demonstrating the generalization performance of the developed proteome classification models on the test set and the external independent validation cohort.

Supplementary Figure 9E: Boxplot depicting the distribution of candidate protein plasma biomarkers between the good-CCC (cyan) and bad-CCC (navy) subgroups of CTO patients in the validation cohort.

Supplementary Figure 9F: Receiver-operating characteristic curves for the clinical and proteome models applied on the test set and the external independent validation cohort.

Supplementary Figure 9G: Comparison between AUCs of the clinical and proteome models. The measure of the center is the parameter estimate, and error bars represent 95% DeLong confidence intervals. The p-value was calculated using the DeLong's test.

Supplementary Figure 10. ELISA validation of candidate plasma biomarkers in the ELISA-based discovery and validation cohorts.

Supplementary Figure 10A-10B: Box plots depicting the ELISA validation of candidate plasma biomarkers in the discovery and validation cohort of CTO patients, stratified by good CCC group (light blue) and poor CCC group (blue). Statistical significance was assessed using Student's t-test, with p-value < 0.05 considered statistically significant.

Supplementary Figure 10C, 10F: ROC curves for the XGBoost classifier model based on proteomic data and ELISA discovery and validation cohorts, respectively.

Supplementary Figure 10D, 10G: Precision-Recall (PR) curves for the XGBoost classifier model using proteomic data and ELISA on the discovery and validation cohort.

Supplementary Figure 9E, 10H: Comparison of performance metrics with error bars for different datasets (Proteome vs. ELISA). The metrics include ROC-AUC, PR-AUC, accuracy, precision, recall, F1-score, MCC (Matthews correlation coefficient), sensitivity, and specificity. The blue bars represent the proteome-based model, the green bars represent the ELISA-based model, and the light blue bars represent the validation set. Error bars indicate standard deviation across the metrics.

Supplementary Figure 11. ELISA Validation of candidate plasma biomarkers in the external validation cohort 2.

Supplementary Figure 11A: Box plots depicting the ELISA validation of candidate plasma biomarkers in another external validation cohort of CTO patients, stratified by good CCC group (light blue) and poor CCC group (blue).

Supplementary Figure 11B: ROC curves of the XGBoost classifier model based on proteomic data and ELISA results from validation cohorts 1 and 2. The orange curve denotes classifier performance in the discovery cohort using proteomic data, while the green and blue curves indicate performance based on ELISA measurements in validation cohorts 1 and 2, respectively.

Supplementary Figure 11C: Precision–Recall (PR) curves of the XGBoost classifier model based on proteomic and ELISA data from validation cohorts 1 and 2. The orange curve shows the PR performance (AUPRC) of the proteomic model, whereas the green and blue curves indicate the ELISA-based performance in validation cohorts 1 and 2, respectively. Mean AUPRC values for both models are shown with 95% confidence intervals.

Supplementary Figure 11D: Comparison of performance metrics with error bars across different datasets (proteomic vs. ELISA). Metrics include ROC–AUC, PR–AUC, accuracy, precision, recall, F1 score, Matthews correlation coefficient (MCC), sensitivity, and specificity. Blue bars indicate the proteome-based model, whereas green and light blue bars represent ELISA-based validation cohorts 1 and 2, respectively. Error bars denote the standard deviation of each metric.

Supplementary Figure 12. Development of a risk prediction model for MACCEs in CTO patients based on plasma proteomic data.

Supplementary Figure 12A: Schematic showing the triage of plasma proteins to identify survival-related plasma proteins and candidate biomarkers for predicting CTO patients' MACCEs.

Figure 12B: LASSO coefficient profiles of the 19 survival-related plasma proteins.

Figure 12C: The stacked plot illustrates the distribution of MACCEs in the vessel reperfusion status and revascularization method subgroups.

Figure 12D-12E: The Kaplan–Meier survival curves comparing survival free from MACCEs in the vessel reperfusion status and revascularization method subgroups.

Figure 12F: Calibration plots for the MACCE risk prediction model at 1-year, 2-year, and 4-year time points. Each plot compares the observed versus predicted MACCE risk, with calibration curves plotted for different time points. For each time point, the black line represents perfect calibration, while the blue dots represent the observed risk at specific predicted values. The results indicate excellent calibration with observed MACCE risk closely aligning with predicted risk for each time period, demonstrating the model's robust performance in risk prediction.