nature communications

Article

https://doi.org/10.1038/s41467-026-69075-y

Proteomics-based machine learning model for predicting secondary infection in HBV-related liver failure

Received: 25 July 2025

Accepted: 26 January 2026

Published online: 03 March 2026

Check for updates

Feixiang Xiong[1,7], Jianming Zheng[2,7], Jiajia Chen[3,7], Linhuan Wu[4], Yuyong Jiang[1], Lianhe Lu[1], Tong Zhou[1], Yang Zhou[1], Tong Wu[1], Yamin Sun[1,5,6] ✉, Ronghua Jin[1,5,6] ✉ & Yixin Hou[1] ✉

Patients with Hepatitis B Virus-related liver failure are highly vulnerable to secondary infections (SI), yet early predictive tools remain limited. In this work, we aim to develop and validate a plasma proteomics–based model for early SI risk assessment. In a prospective multicenter study, 114 patients are enrolled in the discovery cohort, 60 each in two validation cohorts. Untargeted proteomics is used to identify SI-related proteins, followed by Minimum Redundancy Maximum Relevance based feature selection and logistic regression modeling. Targeted proteomics and ELISA are applied for external validation. Inflammatory and coagulation pathway dysregulation is strongly associated with SI. A final model including Lysozyme (LYZ), Calmodulin 1 (CALM1), Serpin Family D Member 1 (SERPIND1), Dermatopontin (DPT), total bilirubin, and AST show excellent discrimination (area under the receiver operating characteristic curve (AUROC) 0.980 in discovery; 0.873 in validation), outperforming C-reactive protein (CRP), white blood cell (WBC), and Neutrophil percentage (NE%). It also predicts 28-day mortality better than Chronic Liver Failure–Consortium Acute-on-Chronic Liver Failure score (CLIF-C ACLF) and Model for End-Stage Liver Disease (MELD). ELISA measurements in validation cohort 2 yield consistent trends, and an ELISA-based model achieve an AUROC of 0.883. This proteomics-derived model reliably identifies patients at high SI risk and supports early clinical intervention.

Hepatitis B virus (HBV) infection remains a major global public health burden. As the disease progresses to liver cirrhosis, patients frequently develop serious complications, including ascites, bacterial infections (BI), hepatic encephalopathy, and variceal bleeding. Studies have shown that the incidence of BI in cirrhotic patients ranges from 32 to 40%[1,2]. Compared to those without BI, patients who develop BI show an approximately 30% increase in mortality, a significantly lower rate of clinical remission, and a markedly higher incidence of shock and new-onset organ failure. The initial episode of BI often marks a critical turning point in disease progression[3]. Notably, approximately one-third of BI are secondary infections (SI) that emerge after hospital admission, without clinical signs of infection at the time of

[1]National Center for Infectious Diseases, Beijing Ditan Hospital, Capital Medical University, Beijing, China. [2]Department of Infectious Diseases, Huashan Hospital, Fudan University, Shanghai, China. [3]State Key Laboratory for Diagnosis and Treatment of Infectious Diseases, National Clinical Research Center for Infectious Diseases, The First Affiliated Hospital, Zhejiang University School of Medicine, Hangzhou, Zhejiang, China. [4]Institute of Microbiology, Chinese Academy of Sciences, Beijing, China. [5]National Key Laboratory of Intelligent Tracking and Forecasting for Infectious Diseases, Beijing Ditan Hospital, Capital Medical University, Beijing, China. [6]Beijing Key Laboratory of Viral Infectious Disease, Beijing Institute of Infectious Diseases, Beijing, China. [7]These authors contributed equally: Feixiang Xiong, Jianming Zheng, Jiajia Chen. ✉e-mail: nksunyamin@aliyun.com; ronghuajin@ccmu.edu.cn; xuexin162@163.com

admission[1,4]. Therefore, early identification of patients at high risk for SI, even in the absence of initial signs, holds considerable clinical significance.

Specific risk factors for SI in hospitalized cirrhotic patients remain poorly defined and require further investigation. Several studies have indicated that empirical administration of antibiotics or proton pump inhibitors (PPIs) at admission may increase the risk of infection, particularly when the daily defined dose (DDD) of PPIs exceeds one[5]. A retrospective analysis of 2864 cirrhotic patients identified baseline bacterial infection, Model for End-Stage Liver Disease (MELD) score >20, systemic inflammatory response syndrome (SIRS), and use of lactulose or rifaximin as independent risk factors for SI. However, the prediction model based on these variables demonstrated limited accuracy, with sensitivity and specificity of 0.67 and 0.63, respectively[6]. A separate study employing machine learning reported an area under the receiver operating characteristic (ROC) curve (AUROC) of 0.784, along with sensitivity and specificity of 0.712 and 0.702, respectively, which remained below the threshold for optimal predictive performance[7]. Most existing models depend on non-specific variables, reducing their accuracy in identifying high-risk individuals. Thus, there is an urgent need to explore more specific and predictive biomarkers to enhance early identification and risk stratification of patients susceptible to SI.

With advances in proteomics, identifying potential biomarkers and elucidating disease mechanisms at the molecular level has become increasingly feasible[8]. Proteomics-based prediction models offer superior specificity for disease forecasting and enable mechanistic insights into disease progression[9]. Previous studies have applied proteomic strategies to develop noninvasive diagnostic models for evaluating liver fibrosis, inflammation, and steatosis in patients with alcoholic liver disease, offering new perspectives for clinical screening. Additional research has identified protein biomarkers associated with infection susceptibility in cirrhotic patients, thereby advancing the understanding of underlying molecular mechanisms[10]. Given these unique advantages, this study aims to identify and validate proteomic biomarkers predictive of SI in patients with HBV-related cirrhosis through a prospective, multicenter design. In this work, a prediction model integrating both proteomic and clinical features will be constructed to support early detection and risk stratification of high-risk individuals.

Table 1 | Baseline characteristics between discovery cohort and validation cohort

Variables	Discovery cohort (n = 114)	Validation cohort1 (n = 60)	Validation cohort1 (n = 60)
Age	55.71 ± 10.02	49.97 ± 12.18	51.24 ± 13.14
Gender			
Female	35 (30.70)	15 (25.00)	18 (20.00)
Male	79 (69.30)	45 (75.00)	42 (80.00)
Hypertension	17 (14.91)	11 (18.33)	12 (20.00)
T2DM	19 (16.67)	13 (21.67)	15(25.00)
Coronary Heart Disease	1 (0.88)	0 (0.00)	0 (0.00)
Ascites	78 (68.42)	38 (63.33)	40 (66.67)
HE	15 (13.16)	8 (13.33)	7(11.67)
GOV	10 (8.77)	11 (18.33)	9 (15.00)
ALT	152.57 ± 326.90	190.74 ± 452.54	160.74 ± 314.25
AST	127.12 ± 211.59	189.80 ± 464.16	175.24 ± 325.24
Tbil	141.06 ± 158.90	132.69 ± 158.62	135.72 ± 148.14
Alb	37.47 ± 27.31	35.80 ± 6.79	34.12 ± 8.65
GGT	66.89 ± 103.22	69.88 ± 75.76	72.34 ± 80.45
ALP	123.66 ± 65.60	123.81 ± 66.10	125.47 ± 67.21
TC	2.95 ± 1.06	3.11 ± 1.04	3.14 ± 1.07
TG	0.96 ± 0.78	1.10 ± 0.80	1.14 ± 0.82
Glu	6.32 ± 2.77	6.21 ± 2.30	6.14 ± 2.47
BUN	6.19 ± 2.76	5.56 ± 3.27	6.10 ± 3.45
Cr	68.59 ± 18.45	71.50 ± 36.77	69.12 ± 34.24
K	3.90 ± 0.46	3.76 ± 0.40	3.66 ± 0.45
Na	139.73 ± 4.38	139.27 ± 4.10	139.48 ± 4.19
NH3	30.77 ± 15.67	32.81 ± 18.21	29.14 ± 16.42
INR	1.73 ± 1.68	1.96 ± 2.79	1.85 ± 2.46
PT	18.75 ± 7.40	16.43 ± 4.68	17.47 ± 5.72
WBC	5.12 ± 3.64	4.74 ± 3.25	5.57 ± 3.75
NE%	65.98 ± 12.71	65.17 ± 12.06	65.16 ± 12.98
HB	110.54 ± 30.77	117.14 ± 30.37	115.24 ± 35.27
PLT	91.27 ± 57.71	91.45 ± 43.24	90.24 ± 51.65
HBV-DNA (log_{10})	2.27 ± 1.64	2.38 ± 1.84	2.40 ± 1.41
HBsAg	82 (71.93)	50 (83.33)	52 (86.67)
HBeAg	22 (19.30)	20 (33.33)	18 (30.00)
CRP	9.94 ± 15.37	9.06 ± 12.44	9.14 ± 11.27
MELD	34.04 ± 20.17	33.16 ± 9.46	33.16 ± 9.46
CLIF-C ACLF	39.57 ± 12.90	42.89 ± 13.80	40.75 ± 12.72
SI	38 (33.33)	20 (33.33)	30 (50.00)
28-day mortality	10 (8.77)	6 (10.00)	7 (11.86)

Data are presented as mean ± SD or *n* (%).

Results

Baseline characteristics

A total of 114 patients were enrolled in the discovery cohort, 60 patients were included in the validation cohort 1, and another 60 patients were included in the validation cohort 2. In all cohorts, patients who developed SI were matched to non-infected controls at a 1:2 or 1:1 ratio. The two groups were comparable in terms of sex and age distribution. In the discovery cohort, the proportion of male patients was 70%, with a median age of 55 years. In the validation cohort 1, 75% were male, with a median age of 50 years. In the validation cohort 2, 80% were male, with a median age of 51 years. Among liver-related complications, ascites was frequently observed, with a prevalence of 68% in the discovery cohort, 63% in the validation cohort 1, and 67% in the validation cohort 2. Most enrolled patients exhibited signs of severe liver dysfunction, as reflected by significantly elevated levels of alanine aminotransferase (ALT), aspartate aminotransferase (AST), total bilirubin (Tbil), and international normalized ratio (INR), indicating hepatocellular injury, impaired bilirubin metabolism, and coagulopathy. Inflammatory markers related to infection, including white blood cell count (WBC), neutrophil percentage (NE%), and C-reactive protein (CRP), were relatively low at admission, suggesting that no patients presented with active infection at the time of enrollment. Both the MELD score and the CLIF-C Acute-on-Chronic Liver Failure (CLIF-C ACLF) score were elevated, indicating a poor overall prognosis in the study population. In-hospital mortality reached 8% in the discovery cohort, 10% in the validation cohort 1, and 11% in the validation cohort 2 (Table 1). Regarding the distribution of SI sites, 16 patients in the discovery cohort developed peritoneal infections, 15 had pulmonary infections, 2 had cholecystitis, 4 developed urinary infections, and 3 had bloodstream infections (Fig. S2a). In the validation cohort 1, 9 patients developed peritoneal infections, 5 had pulmonary infections, 4 had cholecystitis, 1 developed urinary infection, and 1 had bloodstream infections (Fig. S2b). In the validation cohort 2, 9 patients developed peritoneal infections, 5 had pulmonary infections, 4 had cholecystitis, 1 developed urinary infection, and 1 had bloodstream infections (Fig. S2c). In the discovery cohort, the median

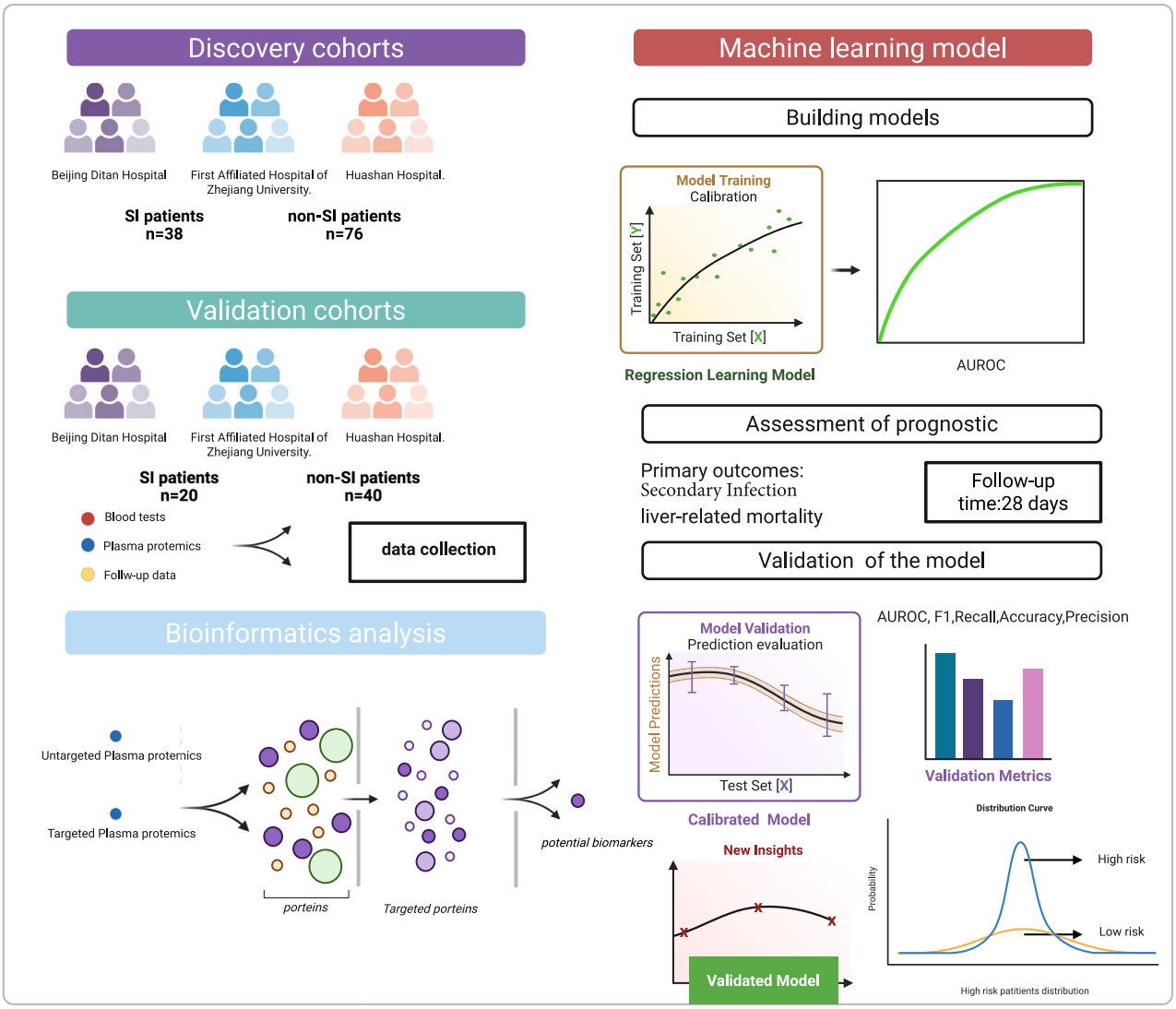

Fig. 1 | Workflow of the study. Created in BioRender. Xiong, J. (2026) https://BioRender.com/8r7n9gc.

time from baseline to the occurrence of SI was 6.5 (5–9) days; In validation cohort 1, the median time was 7 (5–10.25) days; and in validation cohort 2, it was 6.5 (5–9.25) days (Fig. S2d). Overall, peritoneal and pulmonary infections were the most common types of SI, consistent with the known infection patterns in patients with HBV-related liver failure[5].

Subsequently, we compared the baseline biochemical characteristics between patients with and without SI across the three cohorts. The results showed consistent patterns among all cohorts: patients who developed SI exhibited worse liver function, higher levels of jaundice and inflammatory markers, and more impaired coagulation function compared with non-SI patients. Moreover, the 28-day mortality rate was significantly higher in SI patients, indicating that patients who developed SI were in poorer baseline condition (Supplementary Data 1).

Specific proteins were then selected using the minimum redundancy maximum relevance (mRMR) algorithm combined with machine learning methods to construct a predictive model. In the validation cohort, plasma samples were prospectively collected and analyzed using targeted proteomics to quantify the expression levels of the candidate proteins, thereby evaluating the model's predictive performance. The overall study design and analytical workflow are illustrated in Fig. 1.

Plasma untargeted proteomic features in patients with SI

First, we performed quality control on the proteomic data. As shown in Fig. S3a, b, the peptide length distribution exhibited the expected tryptic pattern, and proteins displayed a physiologically consistent molecular weight–pI distribution. After applying data-completeness filtering, 4901 proteins were retained for downstream analysis (Fig. S3c). Protein coverage categories were well balanced, and most proteins were supported by multiple unique peptides (Fig. S3d). Principal component analysis based on all quantified proteins showed no evident site-driven separation among samples from the three study centers. Samples from Beijing Ditan Hospital (site 1), Huashan Hospital (site 2), and the First Affiliated Hospital of Zhejiang University School of Medicine (site 3) were largely intermingled across PC1, PC2, and PC3, indicating minimal batch effects attributable to collection location. Moreover, GR and NGR samples did not cluster by site, further supporting that inter-center differences did not introduce systematic variability into the proteomic profiles. (Fig. S3e). In the discovery cohort, untargeted plasma proteomic analysis identified 974 differentially expressed proteins (DEPs), including 159 upregulated and 815 downregulated proteins in patients with SI. (Fig. 2a and b and Supplementary Data 2). Functional categorization revealed that 27% of the DEPs were associated with inflammation, 19% with metabolic processes, and 9% with coagulation (Fig. 2c). Further pathway enrichment

https://doi.org/10.1038/s41467-026-69075-y

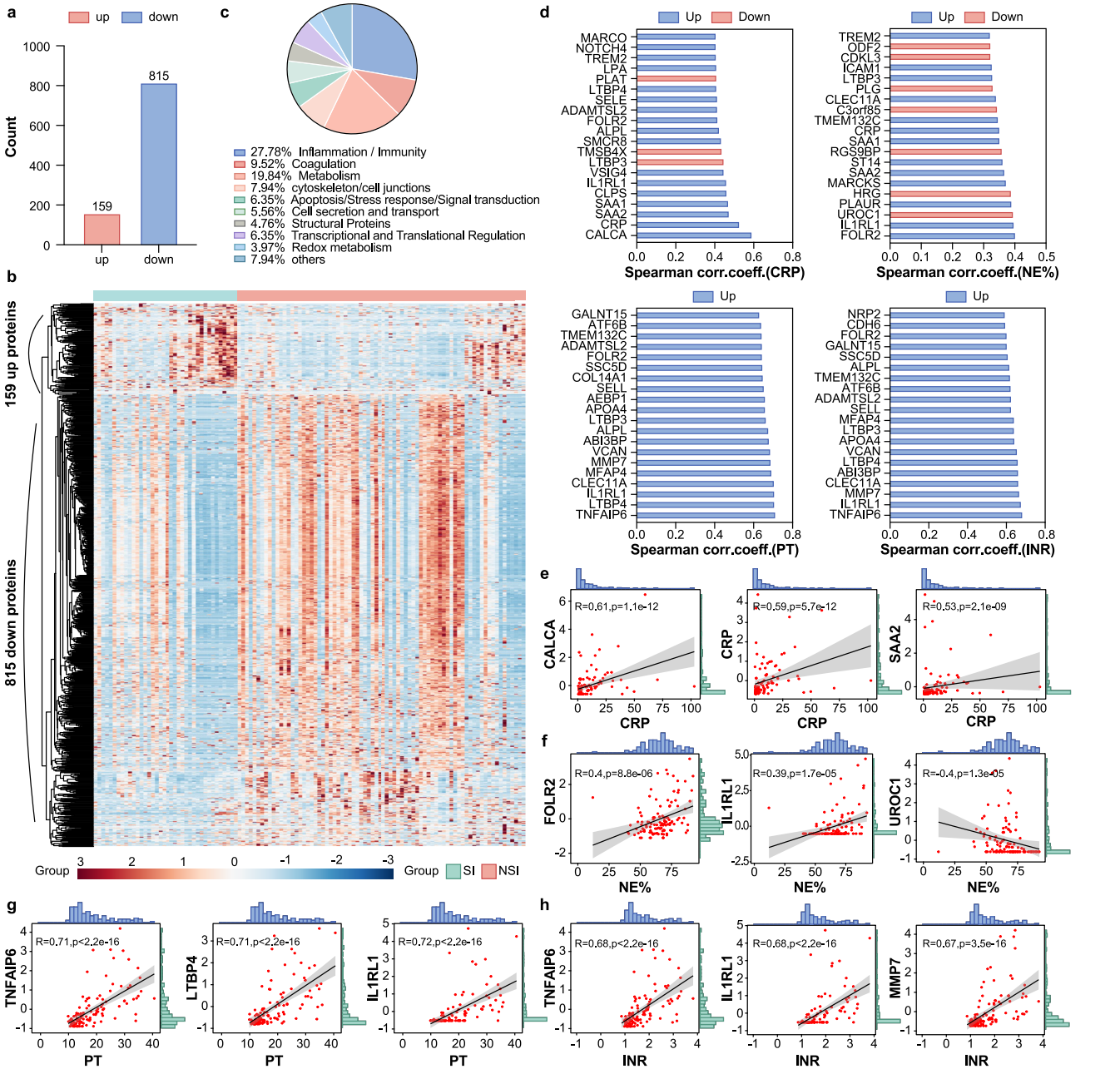

Fig. 2 | Untargeted proteomics in the training cohort. a Number of upregulated and downregulated proteins in patients with and without SI; **b** Heatmap of differentially expressed proteins; **c** Functional classification of proteins; **d** Top 20 proteins ranked by correlation with clinical inflammatory and coagulation markers; **e** Correlation scatter plots for the top 3 proteins associated with CRP; The solid line indicates the linear regression fit, and the shaded area represents the 95% confidence interval of the regression line. Correlation was assessed using Spearman's correlation coefficient, with the corresponding r and P values indicated in the figure. **f** Correlation scatter plots for the top 3 proteins associated with NE%; The solid line indicates the linear regression fit, and the shaded area represents the 95% confidence interval of the regression line. Correlation was assessed using Spearman's correlation coefficient, with the corresponding r and P values indicated in the figure. **g** Correlation scatter plots for the top 3 proteins associated with PT; The solid line indicates the linear regression fit, and the shaded area represents the 95% confidence interval of the regression line. Correlation was assessed using Spearman's correlation coefficient, with the corresponding r and P values indicated in the figure. **h** Correlation scatter plots for the top 3 proteins associated with INR; The solid line indicates the linear regression fit, and the shaded area represents the 95% confidence interval of the regression line. Correlation was assessed using Spearman's correlation coefficient, with the corresponding r and P values indicated in the figure.

analysis (KEGG and Wiki Pathway) showed significant enrichment in platelet activation, complement and coagulation cascades, and various inflammation-related signaling pathways (Fig. S4a and b, Supplementary Data 3-4). Gene set enrichment analysis (GSEA) further confirmed significant enrichment of coagulation- and platelet activation-related gene sets in the SI group (Fig. S4c), suggesting a potential role for coagulation system activation in the development of SI in patients with cirrhosis. These findings indicated that inflammatory and coagulation pathways may contribute to SI pathogenesis in patients with HBV-related liver failure. Previous studies have demonstrated a bidirectional interaction between inflammation and coagulation, wherein inflammatory stimuli activate the coagulation cascade, and coagulation proteases can trigger inflammatory signaling via cell surface receptors[11].

To further investigate the interplay between inflammation and coagulation pathways, PPI analysis was performed on the DEPs. Functional enrichment focused on proteins involved in complement activation, coagulation, and platelet function. Several immunoregulatory proteins, including PLAT, CFHR3, VSIG4, and CLU, were enriched within complement and coagulation networks (Fig. S4d). In addition, PTGS1, an inflammation-associated protein, was enriched in the platelet activation pathway, suggesting that this pathway may link inflammatory and coagulative responses and support the hypothesis of cross-regulation (Fig. S4d). To determine the relationships between commonly used clinical markers of inflammation and coagulation and the identified proteins, Spearman correlation analyses were conducted to evaluate associations between protein levels and CRP, NE%, INR, and prothrombin time (PT). Among the top 20 proteins correlated with CRP, PLAT (a key protein in fibrinolysis) showed a significant negative correlation (Fig. 2d, Supplementary Data 5). Previous studies have reported decreased PLAT expression during BI and systemic inflammatory responses, which potentially exacerbates the inflammatory state[12]. In contrast, scatter plot analysis revealed strong positive correlations between CRP and several inflammatory proteins, including CALCA, CRP, and SAA2 ($P < 0.001$) (Fig. 2e).

Among proteins correlated with NE%, HRG and PLG, both closely related to coagulation, exhibited negative correlations, whereas PLAUR, a fibrinolysis-related protein, showed a positive correlation (Fig. 2d, Supplementary Data 6). Previous studies have indicated that PLG and PLAUR contribute to the resolution of inflammation and may exert protective effects in infectious diseases characterized by inflammatory responses[13]. In contrast, HRG has been associated with macrophage activation and inflammation in chronic liver disease[14]. Further analysis showed a significant positive correlation between NE% and IL1RL1, a pro-inflammatory receptor ($P < 0.001$) (Fig. 2f).

Correlation analyses of PT and INR identified a range of inflammatory modulators among the top 20 associated proteins, including TNFAIP6, IL1RL1, CLEC11A, and SELL, all of which demonstrated significant positive correlations (Fig. 2d, Supplementary Data 7–8). Downregulation of TNFAIP6 has been shown to promote fibrinolytic pathway activation[15], while IL1RL1 has been incorporated into biomarker models for predicting venous thromboembolism. The top three proteins most strongly correlated with PT and INR also exhibited strong positive associations with inflammation-related proteins in scatter plot analyses ($P < 0.001$) (Fig. 2e–h).

Collectively, these findings revealed strong associations between inflammation, coagulation, and their regulatory proteins, suggesting that the pathogenesis of SI may be driven by concurrent dysregulation of both pathways. These results underscore the potential utility of jointly targeting inflammatory and coagulation-related proteins to improve the identification of high-risk patients. To further screen for clinically applicable biomarkers, targeted proteomic analyses were conducted, focusing on proteins related to inflammation, coagulation, and liver function, to construct a more specific and robust predictive model.

Targeted plasma proteomic features in patients with SI

A total of 60 patients were prospectively enrolled from three clinical centers, and plasma samples were collected within 48 h of admission for targeted proteomic analysis to validate DEPs identified during the untargeted discovery phase. In the validation cohort, 31 DEPs were successfully detected, including 5 upregulated and 26 downregulated proteins in patients with SI. Protein abundance changes were visualized using heatmaps (Fig. 3a and b), with detailed information provided in Supplementary Data 9. Among these 31 proteins, 24% were related to inflammation or immune responses, while 27% were involved in coagulation pathways, together accounting for over half of the identified proteins. These findings further emphasized the pivotal role of inflammation–coagulation imbalance in the development of SI in patients with HBV-related cirrhosis (Fig. 3c).

To clarify the relationship between protein expression and clinical inflammatory/coagulation markers, we performed Spearman correlation analyses between protein levels and CRP, NE%, INR, and PT. The top 20 proteins correlated with each marker were identified (Supplementary Data 10–13). Among the proteins most strongly correlated with CRP and NE%, several were coagulation-related, including F2, ITGA2B, SERPINC1, TF, and PLG (Fig. 3d). F2 serves as the precursor of thrombin and has been implicated in the progression of HBV-related liver disease[16]. ITGA2B has been identified as a key component of monocyte–platelet aggregates in single-cell RNA sequencing studies of CHB[17]. SERPINC1 has been reported to attenuate hepatic inflammation by inhibiting leukocyte adhesion and migration[18]. Both TF and PLG are key components of the coagulation and fibrinolytic systems, and their decreased expression in infected patients suggested a close link between coagulation dysfunction and SI. Correlation scatter plots confirmed significant positive associations between CRP, NE%, and these inflammation/coagulation-related proteins ($P < 0.001$) (Fig. 3e and f).

Among the top proteins correlated with PT and INR, several acute-phase inflammatory proteins (such as SAA1, S100A9, CRP, and LYZ) showed strong positive associations (Fig. 2d). SAA1 and CRP are classic acute-phase reactants, and their strong correlation with coagulation parameters reflects the synergistic role of systemic inflammation and coagulation dysfunction in disease progression during infection. LYZ, a lysozyme with broad-spectrum antibacterial activity, has been shown in previous studies to promote hemostasis and wound healing when combined with Cu2 + -loaded nanofibers, suggesting its potential role in coagulation regulation[19]. S100A9 has also been reported in multiple proteomic studies to covary with coagulation-associated proteins[20]. Scatter plots for the top three most strongly correlated proteins further confirmed these associations, all showing significant positive correlations ($P < 0.001$) (Fig. 3e–h).

Identification of predictive factors for SI

To more effectively identify high-risk patients with HBV-related cirrhosis who may develop SI, we constructed a machine learning prediction model based on proteomic data. Partial least squares discriminant analysis (PLS-DA) was first applied to reduce dimensionality and classify protein expression profiles. The analysis revealed a clear separation between the SI and non-SI groups at the proteomic level, indicating strong discriminatory power of protein features (Fig. 4a).

Among the five machine learning models, the LR algorithm demonstrated the best overall performance and was selected as the core classifier for model construction (Fig. 4b and c). The mRMR algorithm was applied to select the top 20 candidate proteins most relevant to SI (Supplementary Data 14). These proteins were sequentially incorporated into the LR model, and five-fold cross-validation was repeated 10 times to calculate average AUROC values across different combinations. The model achieved the highest AUROC when 10 proteins were included (Fig. 4d, Supplementary Data 15). As a result, a final panel of 10 proteins was selected as model input features: CALM1, LYZ, TMSB4X, GC, CRP, SERPIND1, DPT, A2M, PF4V1, and SNCA. Among them, CRP and LYZ were upregulated in the SI group, while the remaining eight proteins were downregulated (Fig. 4e). All 10 proteins showed significant differences in expression between the SI and non-SI groups, indicating their strong discriminatory ability (Fig. 4f).

To enhance the clinical applicability of the model, conventional laboratory parameters were integrated with proteomic biomarkers. Following normalization of protein expression data, univariate logistic regression analyses were conducted, incorporating both proteomic and clinical variables. Significant associations with secondary infection were identified for the following factors: ascites (OR: 2.73, 95% CI:

https://doi.org/10.1038/s41467-026-69075-y

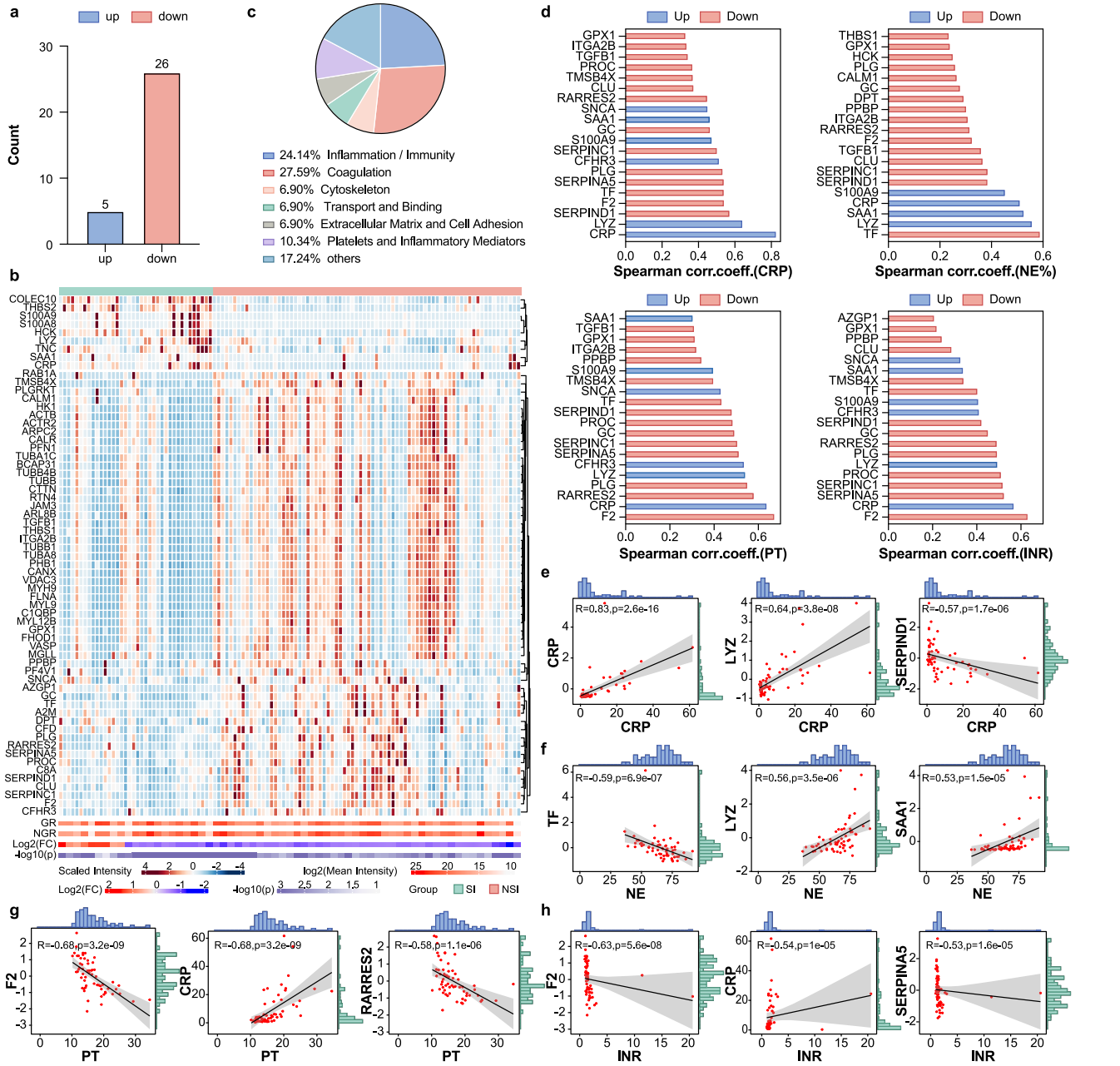

Fig. 3 | Targeted proteomics in the validation cohort. a Number of upregulated and downregulated proteins in patients with and without secondary infection; **b** Heatmap of differentially expressed proteins; **c** Functional classification of proteins; **d** Top 20 proteins ranked by correlation with clinical inflammatory and coagulation markers; **e** Correlation scatter plots for the top 3 proteins associated with CRP; The solid line indicates the linear regression fit, and the shaded area represents the 95% confidence interval of the regression line. Correlation was assessed using Spearman's correlation coefficient, with the corresponding r and P values indicated in the figure. **f** Correlation scatter plots for the top 3 proteins associated with NE%; The solid line indicates the linear regression fit, and the shaded area represents the 95% confidence interval of the regression line. Correlation was assessed using Spearman's correlation coefficient, with the corresponding r and P values indicated in the figure. **g** Correlation scatter plots for the top 3 proteins associated with PT; The solid line indicates the linear regression fit, and the shaded area represents the 95% confidence interval of the regression line. Correlation was assessed using Spearman's correlation coefficient, with the corresponding r and P values indicated in the figure. **h** Correlation scatter plots for the top 3 proteins associated with INR; The solid line indicates the linear regression fit, and the shaded area represents the 95% confidence interval of the regression line. Correlation was assessed using Spearman's correlation coefficient, with the corresponding r and P values indicated in the figure.

1.07–7.01, $P = 0.036$), ALT (OR: 1.01, 95% CI: 1.01–1.01, $P = 0.024$), AST (OR: 1.01, 95% CI: 1.01–1.01, $P = 0.022$), TBIL (OR: 1.01, 95% CI: 1.01–1.01, $P < 0.001$), INR (OR: 2.74, 95% CI: 1.50–4.99, $P = 0.001$), WBC (OR: 1.19, 95% CI: 1.06–1.34, $P = 0.003$), NE% (OR: 1.04, 95% CI: 1.01–1.08, $P = 0.031$), and the 10 candidate proteins (Table 2). Subsequently, variables with $P < 0.05$ were included in a multivariate logistic regression model. This analysis identified six independent predictors: four protein biomarkers—CALM1 (OR: 0.01, 95% CI: 0.00–0.14, $P < 0.001$), LYZ (OR: 17.26, 95% CI: 2.13–140.02, $P = 0.008$), SERPIND1 (OR: 0.06, 95% CI: 0.00–0.14, $P = 0.001$), DPT (OR: 0.36, 95% CI: 0.02–0.61, $P = 0.012$)—and two clinical indicators: AST (OR: 1.01, 95% CI: 1.01–1.01, $P = 0.030$) and Tbil (OR: 1.01, 95% CI: 1.01–1.01, $P = 0.006$) (Table 2).

Construction of ML models for predicting SI

Based on the results of prior analyses, two predictive models for SI were constructed. Model 1 combined two clinical indicators (AST and

https://doi.org/10.1038/s41467-026-69075-y

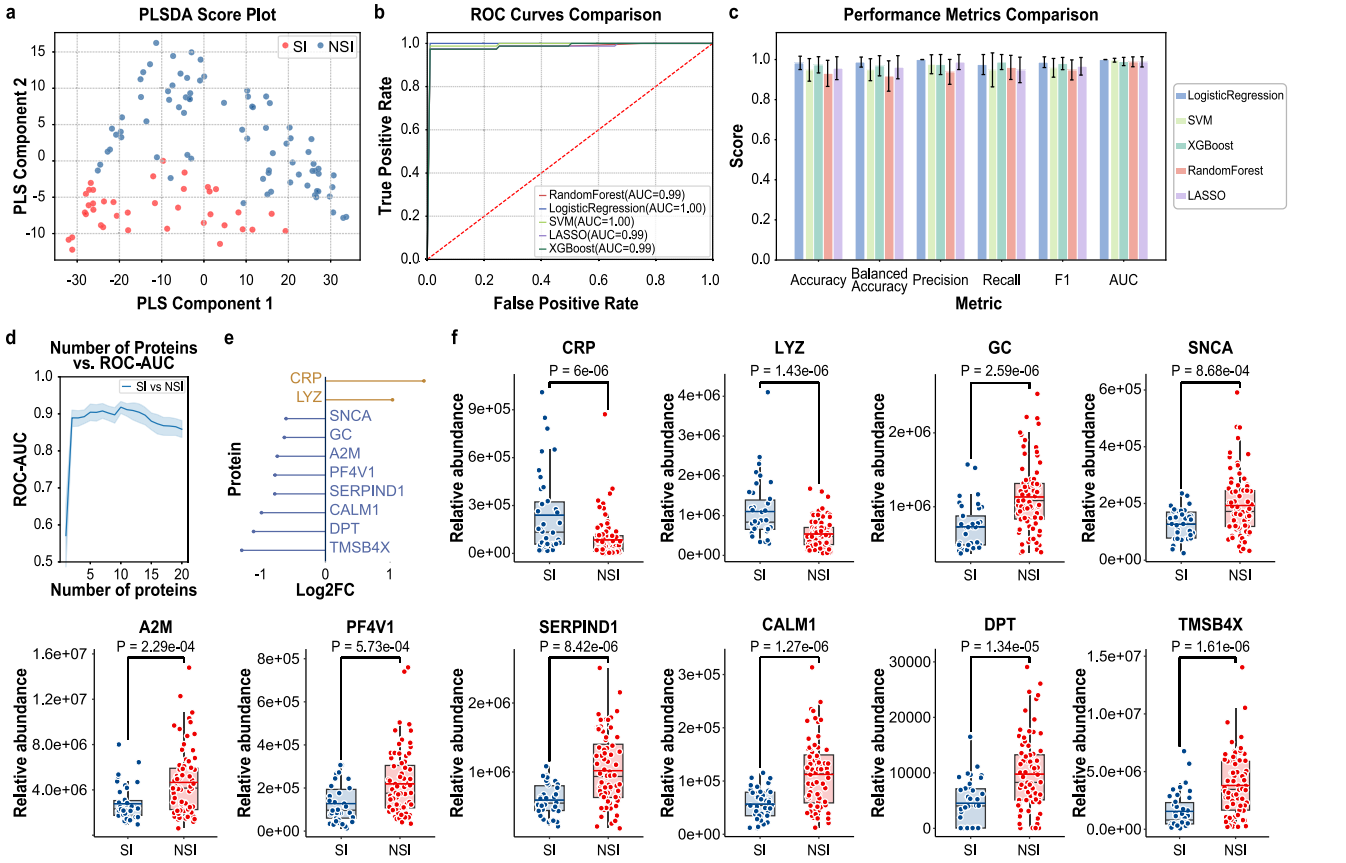

Fig. 4 | Model performance in discovery cohort. a PLS-DA scores based on protein levels in patients with and without SI; **b** Comparison of AUROC among five machine learning models: Random Forest, Logistic Regression, SVM, LASSO, and XGBoost; **c** Comparison of model performance metrics including accuracy, balanced accuracy, precision, recall, F1-score, and AUROC, showing the distribution across the 10 folds ($n = 10$ folds). Statistical analyses were presented as mean ± SD; **d** AUROC comparison using different numbers of proteins for prediction; Statistical analyses were presented as mean ± SD. **e** Lollipop plot showing the expression changes of the top 10 proteins achieving the highest AUROC; **f** Expression levels of the top 10 proteins in the model in SI ($n = 38$) and non-SI groups ($n = 76$). For all statistical analyses, n denotes the number of biological replicates, defined as independent human subjects. Statistical analyses were performed by two-tailed unpaired Student's t test, and data were presented as mean ± SD. The center line denotes the median; box limits indicate the upper and lower quartiles; whiskers extend to 1.5× the interquartile range; points represent individual samples. Source data are provided as a Source Data file.

Tbil) with four protein biomarkers (CALM1, LYZ, SERPIND1, and DPT), whereas Model 2 incorporated the previously identified panel of 10 candidate proteins. To evaluate predictive performance, both models were compared with conventional inflammatory markers (CRP, WBC, NE%) by plotting ROC curves (Fig. 5a). The area under the ROC curve (AUROC) for Model 1 reached 0.980, significantly outperforming CRP (AUROC: 0.645, $P < 0.001$), WBC (AUROC: 0.674, $P < 0.001$), and NE% (AUROC: 0.642, $P < 0.001$). Although the performance of Model 1 was statistically comparable to that of Model 2 (AUROC: 0.973, $P = 0.323$), the reduced number of protein inputs in Model 1 suggested greater feasibility for clinical implementation (Supplementary Data 16). Further comparison of performance metrics, including F1 score, balanced accuracy, positive predictive value (PPV), negative predictive value (NPV), and recall, revealed that Model 1 outperformed all other comparator models across these measures (Fig. 5b and c; Fig. S5a; Supplementary Data).

To assess prognostic potential, all patients were followed for 28 days, and model performance in predicting short-term mortality was evaluated (Fig. 5d). The AUROC of Model 1 for 28-day mortality prediction reached 0.883, significantly higher than that of the CLIF-C ACLF score[21] (AUROC: 0.545, $P < 0.001$), a prognostic tool for HBV-related acute-on-chronic liver failure (ACLF). Although no statistically significant differences were found when compared with the MELD score (AUROC: 0.824, $P = 0.959$) or Model 2 (AUROC: 0.937, $P = 0.252$), Model 1 maintained excellent predictive performance (Supplementary Data 16). Further comparison of F1 score, PPV, NPV, and recall indicated that Model 1 outperformed both MELD and CLIF-C AD scores and performed comparably to Model 2 (Fig. 5d–f; Fig. S5b; Supplementary Data 16).

Model 1 correctly identified 32 of 38 patients with SI, yielding a prediction accuracy of 84% and correctly identified 9 of 10 patients who died within 28 days, achieving a mortality prediction accuracy of 90% (Fig. 5g and h). Analysis of feature importance within Model 1 revealed that the four protein biomarkers contributed most substantially to predictive performance, highlighting their central role in identifying high-risk individuals (Fig. 5i). The DCA confirmed that both Model 1 and Model 2 provided greater net clinical benefit compared with conventional clinical indicators, with Model 1 showing a slight advantage over Model 2 (Fig. S5c). Moreover, the calibration curve demonstrated strong concordance between predicted probabilities and observed outcomes (Fig. S5d). These results highlight the substantial clinical value of integrating proteomic features with routine laboratory parameters for the early prediction of SI in HBV-related cirrhosis.

Validation of the ML model for predicting SI

To assess the generalizability of Model 1 across independent cohorts, we conducted external validation cohort 1 using a prospective, multicenter cohort comprising 60 patients with HBV-related cirrhosis. Model performance in predicting SI was first evaluated by comparing

Table 2 | Univariate and multivariate logistic regression analyses for predicting SI

Variables	*P*	OR (95%CI)	*P*	OR (95%CI)
Sex	0.316	0.65 (0.29–1.50)		
Ascites	0.036	2.73 (1.07–7.01)		
GOV	0.073	3.37 (0.89–12.79)		
HE	0.085	2.63 (0.87–7.91)		
Hypertension	0.148	0.38 (0.10–1.41)		
Coronary Heart Disease	0.992	0.00 (0.00 ~ Inf)		
Diabetes	0.479	0.67 (0.22–2.03)		
Age	0.432	1.02 (0.98–1.06)		
ALT	0.024	1.01 (1.01–1.01)		
AST	0.022	1.01 (1.01–1.01)	0.030	1.01 (1.01–1.01)
Tbil	<0.001	1.01 (1.01–1.01)	0.006	1.01 (1.01–1.01)
Alb	0.503	1.00 (0.99–1.02)		
GGT	0.280	1.00 (1.00–1.01)		
ALP	0.271	1.00 (1.00–1.01)		
TC	0.710	0.93 (0.63–1.37)		
TG	0.753	1.08 (0.66–1.78)		
Glu	0.455	0.94 (0.81–1.10)		
BUN	0.974	1.00 (0.87–1.15)		
Cr	0.701	1.00 (0.97–1.02)		
K	0.888	1.06 (0.46–2.48)		
Na	0.067	0.92 (0.84–1.01)		
NH3	0.761	1.00 (0.98–1.03)		
INR	0.001	2.74 (1.50–4.99)		
WBC	0.003	1.19 (1.06–1.34)		
NE	0.031	1.04 (1.01–1.08)		
HB	0.313	0.99 (0.98–1.01)		
PLT	0.146	0.99 (0.99–1.00)		
CRP	0.138	1.02 (0.99–1.05)		
CALM1	<0.001	0.22 (0.11–0.44)	<0.001	0.01 (0.00–0.14)
LYZ	<0.001	3.74 (2.00–6.97)	0.008	17.26 (2.13–140.02)
TMSB4X	<0.001	0.26 (0.13–0.49)		
GC	<0.001	0.27 (0.14–0.51)		
CRP	<0.001	2.45 (1.44–4.16)		
SERPIND1	<0.001	0.26 (0.13–0.50)	0.001	0.06 (0.00–0.14)
DPT	<0.001	0.30 (0.16–0.55)	0.012	0.36 (0.02–0.61)
A2M	<0.001	0.34 (0.18–0.64)		
PF4V1	0.002	0.39 (0.21–0.70)		
SNCA	0.001	0.37 (0.20–0.68)		
PT	<0.001	1.12 (1.05–1.18)		

P values were obtained using the Wald test (two-sided). No adjustment was made for multiple comparisons.

Model 1 with other models. ROC curves (Fig. 6a). Model 1 achieved an AUROC of 0.873, which was significantly higher than that of traditional inflammatory markers, including WBC (AUROC: 0.702, $P = 0.027$), NE% (AUROC: 0.793, $P = 0.035$), and CRP (AUROC: 0.694, $P = 0.026$). Moreover, Model 1 also outperformed Model 2 (AUROC: 0.815, $P = 0.042$), further validating its accuracy and robustness in predicting SI (Supplementary Data 17). Evaluation of other performance metrics, including F1 score, balanced accuracy, PPV, NPV, and recall, also demonstrated that Model 1 outperformed all comparator models (Fig. 6b and c; Fig. S6a; Supplementary Data 17).

In addition, the ability of Model 1 to predict 28-day mortality was assessed in the validation cohort. ROC curve analysis (Fig. 6d) showed that the AUROC of Model 1 reached 0.957, significantly higher than that of the CLIF-C AD score (AUROC: 0.651, $P = 0.018$) and MELD score (AUROC: 0.864, $P = 0.036$). While no statistically significant difference was observed between Model 1 and Model 2 (AUROC: 0.874, $P = 0.154$), Model 1 exhibited superior accuracy and sensitivity (Supplementary Data 17). Further comparisons of F1 score, balanced accuracy, PPV, NPV, and recall also indicated that Model 1 outperformed both the CLIF-C AD and MELD scores and showed comparable performance to Model 2 (Fig. 6d–f; Fig. S6b; Supplementary Data 17). The confusion matrices of Model 1 for predicting SI and 28-day mortality in the validation cohort 1 further demonstrated the model's discriminatory performance (Fig. 6g and h). In this cohort, targeted proteomic analysis confirmed that all four proteins included in Model 1 (LYZ, CALM1, SERPIND1, and DPT) were significantly differentially expressed between patients with and without SI (Fig. 6i). Specifically, LYZ levels were elevated, while CALM1, SERPIND1, and DPT levels were reduced in the SI group, consistent with findings from the discovery cohort. Overall, these findings validated the predictive performance and

https://doi.org/10.1038/s41467-026-69075-y

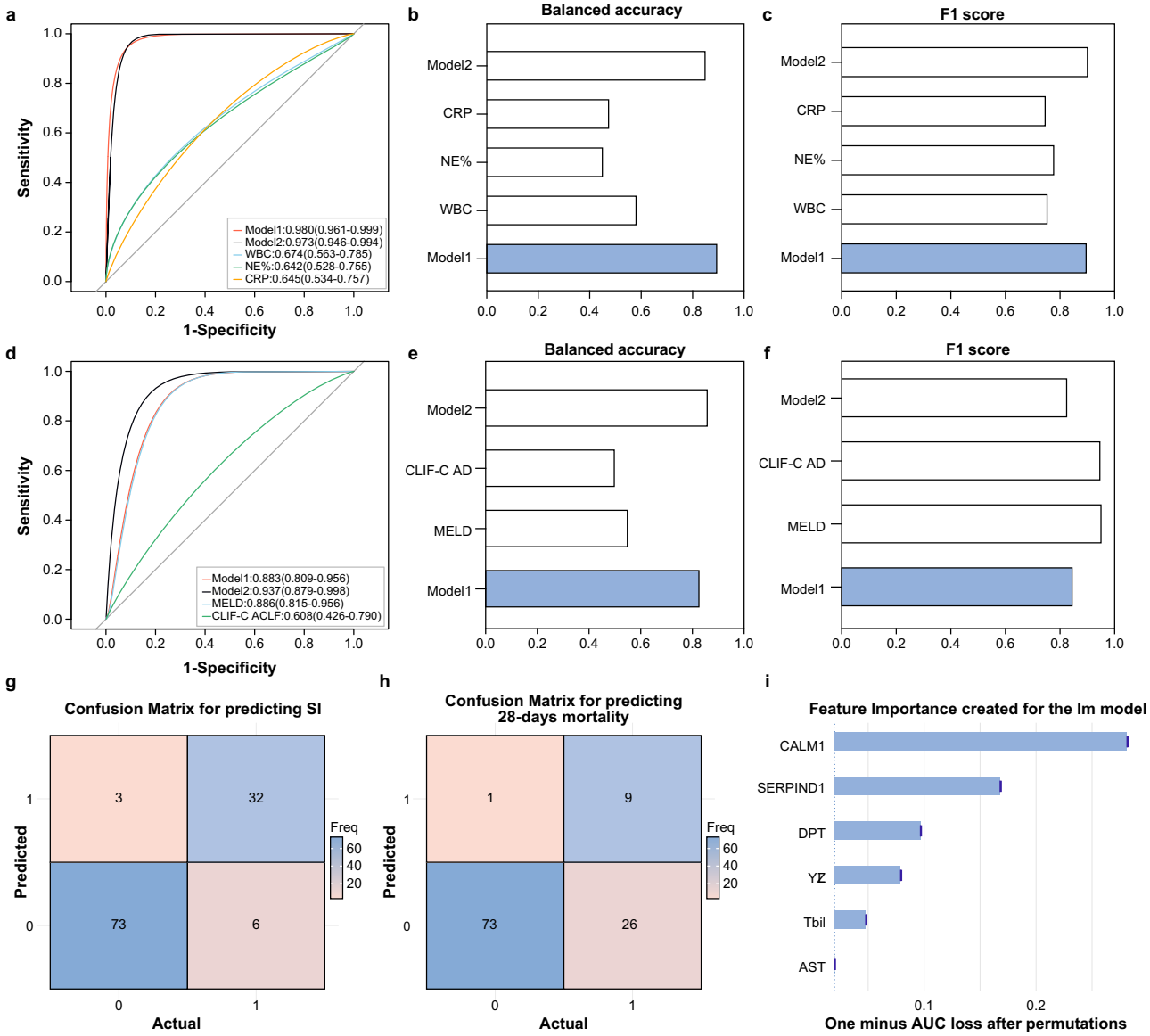

Fig. 5 | Model performance in the discovery cohort. a Comparison of AUROC for different proteomics-based models in predicting SI; **b**, **c** Comparison of balanced accuracy and F1-score across models for predicting SI; **d** Comparison of AUROC for different proteomics-based models in predicting 28-day mortality; **e**, **f** Comparison of balanced accuracy and F1-score across models for predicting 28-day mortality; **g** Confusion matrix for SI prediction; **h** Confusion matrix for 28-day mortality prediction; **i** Variable importance weights in the predictive model.

clinical utility of Model 1 in real-world patients with advanced HBV-related liver failure, supporting its potential application for early identification of SI and prediction of short-term mortality risk in clinical practice.

To validate the clinical applicability of these four biomarkers, we utilized an additional validation cohort 2 consisting of 30 SI patients and 30 non-SI patients. Plasma levels of the four biomarkers were measured using ELISA. Consistent with the findings from the discovery and validation cohorts, the level of LYZ was significantly higher in SI patients than in non-SI patients ($P < 0.05$; Fig. S7a), while CALM1, SERPIND1, and DPT levels were significantly lower in SI patients compared with non-SI patients ($P < 0.05$; Fig. S7b–d). Furthermore, an LR model constructed using these four biomarkers demonstrated robust discrimination between SI and non-SI patients, with an AUROC of 0.883 (0.798–0.968) based on ELISA-measured levels (Fig. S7e). These results indicate that the four biomarkers can effectively distinguish SI from non-SI patients, and that ELISA provides a convenient approach for clinical application.

In conclusion, targeted proteomic validation highlighted the critical interplay between inflammation and the coagulation system. In patients with HBV-related liver failure, coagulation dysfunction may amplify inflammatory responses by promoting the recruitment of inflammatory mediators, while persistent inflammation may aggravate coagulation abnormalities, forming a vicious cycle that drives the onset and progression of SI. Additionally, the use of ELISA in validation cohort 2 further confirmed the potential of these four biomarkers for predicting the occurrence of SI, demonstrating their feasibility for clinical application.

Discussion

In the present study, protein biomarkers associated with SI in patients with HBV-related failure were identified through the integration of untargeted and targeted plasma proteomic analyses. A high-performance machine learning-based predictive model was subsequently developed by combining these biomarkers with conventional clinical indicators. Compared with traditional inflammatory markers

https://doi.org/10.1038/s41467-026-69075-y

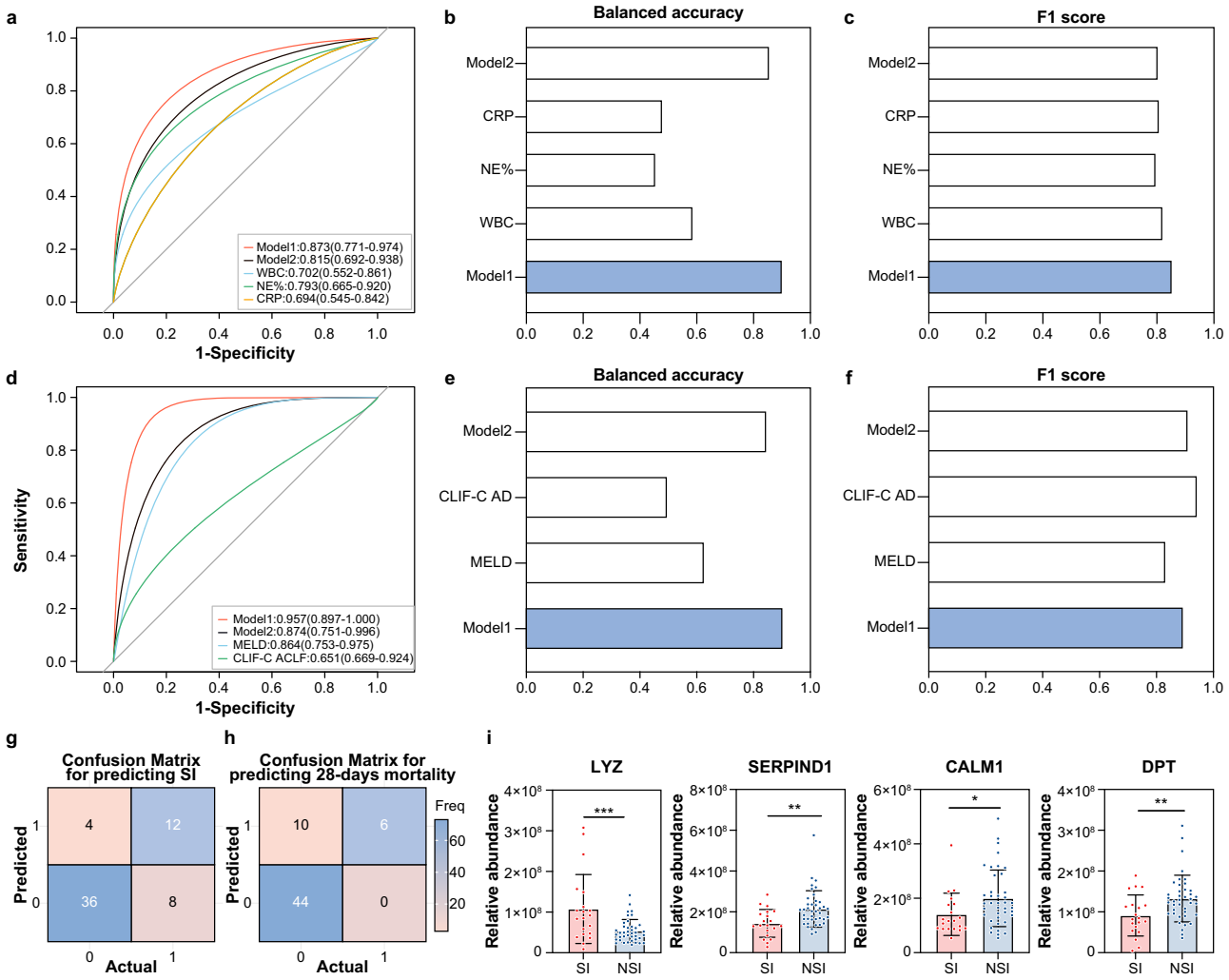

Fig. 6 | Model performance in the validation cohort. a Comparison of AUROC for different proteomics-based models in predicting SI; **b**, **c** Comparison of balanced accuracy and F1-score across models for SI prediction; **d** Comparison of AUROC for different proteomics-based models in predicting 28-day mortality; **e**, **f** Comparison of balanced accuracy and F1-score across models for 28-day mortality prediction; **g** Confusion matrix for SI prediction; **h** Confusion matrix for 28-day mortality prediction; **i** Comparison of the expression levels of four selected proteins between patients with ($n = 20$) and without secondary infection ($n = 40$) in the validation cohort 1. For all statistical analyses, n denotes the number of biological replicates, defined as independent human subjects. Statistical analyses were performed by a two-tailed unpaired Student's t test, and data were presented as mean ± SD. Statistical significance was observed for LYZ ($P = 0.0005$), SERPIND1 ($P = 0.0030$), CALM1 ($P = 0.0295$), and DPT ($P = 0.0080$). The center line denotes the median; box limits indicate the upper and lower quartiles; whiskers extend to 1.5× the interquartile range; points represent individual samples. Source data are provided as a Source Data file.

such as CRP, WBC, and NE%, the model demonstrated superior predictive capacity across multiple metrics, including AUROC, F1 score, and balanced accuracy. These findings were robustly validated in an independent, prospective multicenter cohort. DCA and calibration curves further confirmed its potential clinical utility. Notably, the model enabled early risk prediction based on plasma samples collected within 48 h of hospital admission, preceding changes in routine laboratory markers. The application of ELISA in the validation cohort further underscores the potential of these biomarkers as promising indicators for identifying patients at risk of SI. This early stratification capacity provides an opportunity for timely clinical intervention, which may reduce the likelihood of clinical deterioration and improve patient outcomes.

The findings also underscored the pivotal role of inflammation–coagulation crosstalk in the pathogenesis of SI. Previous studies have reported associations between CRP levels and the risk of bleeding and venous thromboembolism in cirrhotic patients, and SI is frequently accompanied by elevated INR, coagulopathy, and gastrointestinal bleeding[5–7,22]. In the current analysis, most coagulation-related proteins were downregulated and negatively correlated with inflammatory markers, reinforcing the mechanistic link between coagulation dysfunction and systemic inflammation. Moreover, the inclusion of Tbil in the final predictive model suggested that progressive jaundice may contribute to infection susceptibility, which aligns with prior studies identifying Tbil levels exceeding 188 μmol/L as a risk factor for infection in decompensated cirrhosis[23]. These findings highlight the contribution of hepatic dysfunction to immune vulnerability in this population.

The protein markers included in the predictive model were selected using machine learning algorithms based on their contribution to predictive performance, thereby minimizing subjective bias. Each of the four proteins has previously been associated with inflammation, coagulation, or tissue repair. CALM1 has been implicated in thrombotic complications and systemic inflammation in patients with COVID-19[24] and may also contribute to lung inflammation and secondary Pseudomonas aeruginosa infections[25]. LYZ levels in saliva have been associated with systemic inflammation in cirrhosis and proposed as a prognostic indicator in patients with arthritis[26]. SERPIND1 has been

identified as a predictive marker for adverse cardiovascular events following acute myocardial infarction[27], and DPT plays a role in vascular repair and wound healing[28]. Notably, among the four proteins, LYZ was the only one upregulated in the SI group, suggesting its potential utility as a sensitive predictive biomarker. Several studies have explored its clinical relevance in inflammation and infection[29–31], although further investigation is warranted to determine its clinical applicability.

This study offers several strengths. First, patient recruitment was conducted prospectively and continuously across three major liver disease centers in China, ensuring representative sampling. All plasma samples were obtained before the onset of SI, which provided strong temporal validity. Second, none of the enrolled patients had received antibiotic treatment within six months prior to sampling, minimizing potential confounding and selection bias. Third, the identified proteins have been consistently reported in previous studies related to inflammation, coagulation, and tissue repair, supporting their biological plausibility. The study was the first to systematically define the proteomic landscape of SI in HBV-related cirrhosis and to construct a clinically relevant early-warning model, thereby addressing a significant knowledge gap in the field.

Several limitations should also be considered. First, no internationally recognized tool currently exists for assessing the risk of SI in HBV-related cirrhosis. Conventional inflammatory markers such as WBC, NE%, and CRP were used for comparison. However, these markers often lack sensitivity in the early stages of infection, which limits their utility as reference standards. Besides, the CRP values obtained from routine clinical testing did not show the highest correlation with the MS-derived CRP measurements. One important reason is that the clinical tests were not always performed at the same time as the plasma sampling used for MS, and in some patients, the clinical CRP result was simply not available for that time point. As a result, the normal biological fluctuations of CRP during acute inflammation likely introduced additional variability, which explains the modest concordance observed between the two methods. Second, given the multicenter design, minor inconsistencies in blood sample processing may have introduced variability in protein measurements. Nevertheless, targeted proteomic and ELISA test validation demonstrated consistent expression patterns of key biomarkers across centers, supporting the robustness of the findings. Third, LYZ has not yet undergone systematic evaluation in clinical practice. Future studies should aim to develop accessible and cost-effective assays for LYZ quantification to support its potential clinical application.

In summary, a multifactorial machine learning model was developed to predict SI in patients with HBV-related failure, based on integrated proteomic profiling and robust algorithmic selection. Validation in a prospective, multicenter cohort using ELSIA tests demonstrated high predictive accuracy and strong clinical potential. The identified protein biomarkers are functionally linked to inflammation, coagulation, and immune regulation, and may serve as early-warning indicators to facilitate timely infection management and improve clinical outcomes in patients with advanced HBV-related liver disease.

Method

Patients

This study was approved by the Ethics Committees of Beijing Ditan Hospital, Capital Medical University (DTECKY2023-045-01), Huashan Hospital, Fudan University (KY2020-679), and the First Affiliated Hospital, Zhejiang University (IIT0230519B-R1). A prospective, multicenter cohort study was conducted. Patients in the discovery cohort were enrolled from Beijing Ditan Hospital (Capital Medical University), the First Affiliated Hospital of Zhejiang University School of Medicine, and Huashan Hospital (Fudan University) between June 2024 and February 2025. The validation cohort 1 was recruited from February to April 2025. The validation cohort 2 was recruited from May to June 2025. In the discovery cohort, validation cohort, and validation cohort 2, plasma samples were collected within 48 h of admission and analyzed using untargeted proteomics to identify candidate biomarkers associated with SI. Inclusion criteria were as follows: (1) age between 18 and 75 years; (2) diagnosis of liver failure based on the European Association for the Study of the Liver (EASL) guidelines[32]; (3) absence of clinical or laboratory evidence of bacterial infection at admission; and (4) provision of written informed consent. Exclusion criteria included: (1) age <18 or >75 years; (2) coexistence of other liver diseases (e.g., hepatitis C, alcoholic, drug-induced, autoimmune, or hereditary/metabolic liver diseases); (3) presence of hepatocellular carcinoma (HCC) or other malignancies; (4) co-infection with human immunodeficiency virus (HIV); and (5) pregnancy or lactation. The written informed consent was obtained from all participants prior to enrollment in the study. This study had been registered in the Chinese National Medical Research Registration Information System (MR-11-24-043987).

Diagnosis of BI was based on established criteria[33]. (1) Spontaneous bacterial peritonitis: ascitic fluid polymorphonuclear leukocyte count >250/μL, regardless of culture results; (2) Pneumonia: new pulmonary infiltrates on imaging plus either (a) at least one respiratory symptom (e.g., cough, sputum production, dyspnea, chest pain), or (b) at least one physical sign (e.g., rales, crackles) or systemic infection indicator (e.g., body temperature >38 °C or <36 °C, chills, WBC > 10,000/mm^3 or <4000/mm^3), after excluding antibiotic interference; (3) Enteric infection: diarrhea or dysentery with stool cultures positive for Salmonella, Shigella, Campylobacter, or pathogenic Escherichia coli; (4) Skin infection: fever with signs of cellulitis; (5) Urinary tract infection: symptomatic presentation with pyuria (urine WBC > 15/high-power field) and a positive Gram stain or urine culture; (6) Bloodstream infection: positive blood culture in the absence of an identifiable infection source elsewhere.

Proteomics

Plasma sample preparation. Plasma samples from patients at the three participating centers were collected following standardized SOPs to ensure the accuracy of proteomic analyses. Blood samples were collected within 48 h of patient admission using EDTA anticoagulant tubes. Within 4 h of collection, samples were centrifuged at 750 × g for 10 min in the laboratory, and the resulting plasma was stored at −80 °C.

For each sample, 5 μL of plasma was denatured in 50 μL of buffer containing 8 M urea and 100 mM ammonium bicarbonate. Proteins were reduced with 10 mM DTT at 37 °C for 30 min and subsequently alkylated with 50 mM IAA in the dark at room temperature. The solution was then diluted with 50 mM ammonium bicarbonate, followed by digestion with trypsin (Promega) at a protein-to-enzyme ratio of 50:1 overnight at 37 °C. Peptides were desalted using C18 StageTips and prepared for LC-MS analysis. Peptide concentrations were measured at 280 nm using a Nanodrop One spectrophotometer.

LC-MS/MS analysis for DIA

LC–MS/MS analysis was performed using an Orbitrap Astral mass spectrometer coupled with a Vanquish Neo UHPLC system (Thermo Fisher Scientific). Peptides from each sample were loaded onto a 50 cm Low-Load μPAC™ Neo HPLC column (Thermo Scientific) at a flow rate of 2.2 μL/min. The reversed-phase HPLC mobile phases consisted of 0.1% formic acid in water (buffer A) and 0.1% formic acid in 80% acetonitrile (buffer B). Peptides were separated over 8 min using a linear gradient of buffer B at 1.25 μL/min. The gradient was programmed as follows: 0–0.1 min, 4 to 6% B; 0.1–1.1 min, 6% to 12% B; 1.1–4.3 min, 12% to 25% B; 4.3–6.1 min, 25% to 45% B; 6.1–6.5 min, 45% to 99% B; 6.5–8 min, held at 99% B. The eluted peptides were analyzed in

data-independent acquisition (DIA) mode. The survey scan was acquired from 380 to 980 m/z at a resolution of 240,000, with an AGC target of 500% and an injection time of 5 ms. DIA MS/MS scans were acquired from 150 to 2000 m/z using a 2 m/z isolation window, an AGC target of 500%, and an injection time of 3 ms. Normalized collision energy was set to 25, and the total cycle time was 0.6 s. Full MS scans were recorded in profile mode, while DIA MS/MS spectra were recorded in centroid mode.

Plasma samples (50 µL) were processed using the DBE kit following the manufacturer's instructions (Shanghai Bioprofile, China). The magnetic bead–based protein enrichment procedure followed the methodology described by Omid et al[34]. First, 20 µL of magnetic beads were washed three times with 200 µL of Washing-Buffer in tubes. Each wash involved mixing at 317 × g for 30 s, followed by magnetic separation using a magnetic rack for 1 min and removal of the supernatant. This washing step was repeated for a total of three washes. Then, 50 µL of plasma and 50 µL of Incubation-Buffer were both added to the beads and mixed at 317 × g for 10 min at room temperature. Tubes were placed on the magnetic rack and the supernatant was discarded. Samples were washed with 150 µL of Washing-Buffer for a total of four washes. For lysis, 40 µL of Elution Buffer was added to the beads and mixed at 317 × g for 20 min at room temperature. For protein digestion, 1 µL 400 mM TCEP and 2 µL 800 mM CAA were added into the eluted protein buffer, vortexed, and 95 °C for 3 min (Reduction and Alkylation). Then 160 µL of 50 mM NH_4HCO_3 was added, and then proteins were digested with 0.5 µg Lys-C/Trypsin (1:2) at 37 °C overnight. The digestion reaction was stopped by adding 60 µL of STOP solution and by mixing the tubes at 317 x g for 1 min at RT. The whole mixture was transferred to the C18 stage tip to clean the peptides. Samples were washed sequentially with 100 µL of methanol and 100 µL of 0.1% TFA, and peptides were eluted with 100 µL of 50% Acetonitrile (0.1% TFA) twice in the same collection tube. Samples were dried down in a SpeedVac (Eppendorf), and peptides were redissolved in 10 µL of buffer with 2% ACN, 0.1%FA. Peptide concentration was determined using a Nanodrop and then analyzed by LC-MS/MS.

The DIA MS data were analyzed using DIA-NN 1.8.1. MS data were searched against the UniProtKB reviewed (Swiss-Prot) database (Homo sapiens (species), 20422 total entries, downloaded 12/2024. The trypsin was selected as the digestion enzyme. The maximum of 1 missed cleavage sites and a mass tolerance of 10 ppm for precursor ions and 10 ppm for fragment ions were defined for database search. Carbamidomethylation of cysteines was defined as a fixed modification, while acetylation of protein N-terminal, oxidation of Methionine was set as variable modifications for database searching. The maximum number of variable modifications was 1. The peptide length range was set from 7 to 30. The charge of the peptide was from 1 to 4. The fragment ion m/z range was from 150 to 2000. The database search results were filtered and exported with <1% false discovery rate (FDR) at the peptide-spectrum-matched level and the protein level, respectively. Quantification was based on MS1 signal intensity. A schematic workflow is presented in Fig. S1. The figures were generated using BioRender.

Targeted proteomics (PRM)

PRM analyses were performed on a Thermo Scientific Q-Exactive HF-X mass spectrometer using MaxQuant 2.0.1.0. Proteins were digested with trypsin allowing up to two missed cleavages. Precursor mass tolerances were set to 20 ppm for the first search and 4.5 ppm for the main search; MS/MS tolerance was 20 ppm. Carbamidomethylation of cysteine was set as a fixed modification, and oxidation (M) and N-terminal acetylation were variable modifications. Database searching used the *Homo sapiens* SwissProt database (9606, 20,418 entries, downloaded 2025-04-14) with a target-reverse pattern. PSM, protein, and site-level FDR were all controlled at 1%.

Quality control

Peptide selection for protein analysis followed the criteria outlined below. Peptides were required to be 6–24 amino acids in length, contain no missed K/R cleavages (missed cleavages = 0), and be protein-specific (unique peptides). Peptides containing modification-prone residues such as methionine or cysteine were excluded. In addition, peptide scores were required to meet the following thresholds: Mascot ion score >20 and MaxQuant score >40. After filtering peptides based on these criteria, proteins were further enriched according to the number of remaining unique peptides. For proteins, those with >50% missing values across samples were removed, whereas proteins with <50% missing values were retained and subjected to imputation. Imputation was performed only when a protein was detected in at least two treatment groups. If this condition was met, missing values for that protein were imputed by replacing them with the minimum observed intensity within each group[35]. Using this workflow, a total of 4,901 proteins were ultimately retained for downstream analyses.

Bioinformatics analysis

Bioinformatic analysis was carried out with Microsoft Excel and R statistical computing software. Sequence annotation was performed based on information retrieved from the Kyoto Encyclopedia of Genes and Genomes (KEGG) database. KEGG pathway enrichment analysis was conducted using Fisher's exact test, followed by false discovery rate (FDR) correction to account for multiple testing. Pathways with a Fisher's exact test $p < 0.01$ were considered nominally significant. Protein–protein interaction (PPI) networks were constructed using the STRING database and visualized with Cytoscape software.

Machine learning models

Next, we systematically evaluated the performance of five mainstream machine learning algorithms: Logistic Regression (LR), Support Vector Machine (SVM), Extreme Gradient Boosting (XGBoost), Random Forest (RF), and Least Absolute Shrinkage and Selection Operator (LASSO) regression. Evaluation metrics included AUROC, accuracy, balanced accuracy, precision, recall, and F1 score. By comparing these metrics, we will select the most suitable machine learning algorithm for constructing the predictive model.

ELISA analysis

The ELISA kits used for the validation experiments were purchased from Abcam, including LYZ (ab267798), DPT (ab272206), and SERPIND1 (ab277402). The CALM1 kit was purchased from Cloud Clone (catalog no. SEG095Hu). ELISA readings were performed using a Bio-Tek 800 TS microplate reader. Samples, reagents, and standards were prepared according to the manufacturer's instructions. Specific antibodies against the target proteins were coated onto the microplate wells. Standards and samples were then added to the wells, allowing the immobilized antibodies to bind the target proteins. After washing, a horseradish peroxidase (HRP) conjugate mixture was added, and the optical density (OD) was measured at 450 nm using the microplate reader.

Statistical analysis

All statistical analyses were performed using R software (version 4.4.0). Model performance was assessed using receiver operating characteristic (ROC) curves and area under the ROC curve (AUROC). Comparisons between models were conducted using the DeLong test. Clinical utility and diagnostic accuracy were further evaluated using decision curve analysis (DCA) and calibration plots. A two-sided $P < 0.05$ was considered statistically significant.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

The Proteomics data that support the findings of this study have been deposited in IPROX with the PXD067174 accession codes. Source data are provided with this paper.

Code availability

The code used for the analyses in this study is publicly available on GitHub at: https://github.com/xiang-x98/machine-learning.

References

1. Fernández, J., Piano, S., Bartoletti, M. & Wey, E. Q. Management of bacterial and fungal infections in cirrhosis: the MDRO challenge. *J. Hepatol.* **75**, S101–S117 (2021).
2. Piano, S. et al. Epidemiology and effects of bacterial infections in patients with cirrhosis worldwide. *Gastroenterology* **156**, 1368–1380.e10 (2019).
3. Nahon, P. et al. Bacterial infection in compensated viral cirrhosis impairs 5-year survival (ANRS CO12 CirVir prospective cohort). *Gut* **66**, 330–341 (2017).
4. Piano, S., Tonon, M. & Angeli, P. Changes in the epidemiology and management of bacterial infections in cirrhosis. *Clin. Mol. Hepatol.* **27**, 437–445 (2021).
5. Sargenti, K., Prytz, H., Strand, A., Nilsson, E. & Kalaitzakis, E. Healthcare-associated and nosocomial bacterial infections in cirrhosis: predictors and impact on outcome. *Liver Int.* **35**, 391–400 (2015).
6. Bajaj, J. S. et al. Nosocomial infections are frequent and negatively impact outcomes in hospitalized patients with cirrhosis. *Am. J. Gastroenterol.* **114**, 1091–1100 (2019).
7. Li, S. et al. Development and validation of prediction models for nosocomial infection and prognosis in hospitalized patients with cirrhosis. *Antimicrob. Resist. Infect. Control* **13**, 85 (2024).
8. Altelaar, A. F., Munoz, J. & Heck, A. J. Next-generation proteomics: towards an integrative view of proteome dynamics. *Nat. Rev. Genet* **14**, 35–48 (2013).
9. Niu, L. et al. Noninvasive proteomic biomarkers for alcohol-related liver disease. *Nat. Med.* **28**, 1277–1287 (2022).
10. Garg, P. et al. Proteomics-guided biomarker discovery, validation, and pathway perturbation in infection-related acute decompensation of cirrhosis. *Clin. Gastroenterol. Hepatol.* **24**, 407–420.e39 (2026).
11. Rangaswamy, C. et al. The contact system in liver injury. *Semin. Immunopathol.* **43**, 507–517 (2021).
12. Sillen, M. & Declerck, P. J. A narrative review on plasminogen activator inhibitor-1 and its (patho)physiological role: to target or not to target? *Int. J. Mol. Sci.* **22**, 5 (2021).
13. Perucci, L. O., Vago, J. P., Miles, L. A. & Sousa, L. P. Crosstalk between the plasminogen/plasmin system and inflammation resolution. *J. Thromb. Haemost.* **21**, 2666–2678 (2023).
14. Bartneck, M. et al. Histidine-rich glycoprotein promotes macrophage activation and inflammation in chronic liver disease. *Hepatology* **63**, 1310–1324 (2016).
15. Memon, A. A. et al. Identification of novel diagnostic biomarkers for deep venous thrombosis. *Br. J. Haematol.* **181**, 378–385 (2018).
16. Zhan, J. et al. Noninvasive diagnosis of significant liver inflammation in patients with chronic hepatitis B in the indeterminate phase. *Virulence* **14**, 2268497 (2023).
17. Zhuo, Y. et al. Single-cell profiling reveals monocyte heterogeneity and association with liver fibrosis in patients with chronic HBV. *Hepatol. Commun.* **9**, e0672 (2025).
18. Maksan, S. M., Ulger, Z., Gebhard, M. M. & Schmidt, J. Impact of antithrombin III on hepatic and intestinal microcirculation in experimental liver cirrhosis and bowel inflammation: an in vivo analysis. *World J. Gastroenterol.* **11**, 4997–5001 (2005).
19. He, W. et al. Fabrication of Cu^{2+}-loaded phase-transited lysozyme nanofilm on bacterial cellulose: antibacterial, anti-inflammatory, and pro-angiogenesis for bacteria-infected wound healing. *Carbohydr. Polym.* **309**, 120681 (2023).
20. Nunes, M. et al. Data-independent LC–MS/MS analysis of ME/CFS plasma reveals a dysregulated coagulation system, endothelial dysfunction, and downregulation of complement machinery. *Cardiovasc. Diabetol.* **23**, 254 (2024).
21. Jalan, R. et al. Development and validation of a prognostic score to predict mortality in patients with acute-on-chronic liver failure. *J. Hepatol.* **61**, 1038–1047 (2014).
22. Zanetto, A. et al. Severity of systemic inflammation is the main predictor of ACLF and bleeding in individuals with acutely decompensated cirrhosis. *J. Hepatol.* **78**, 301–311 (2023).
23. Fuller, H. et al. Hyperbilirubinemia at hospitalization predicts nosocomial infection in decompensated cirrhosis: data from ATTIRE trial. *Hepatol. Commun.* **9**, e0648 (2025).
24. Eyileten, C. et al. Thrombosis-related circulating miR-16-5p is associated with disease severity in patients hospitalised for COVID-19. *RNA Biol.* **19**, 963–979 (2022).
25. Bajaj, J. S. et al. Salivary microbiota reflects changes in gut microbiota in cirrhosis with hepatic encephalopathy. *Hepatology* **62**, 1260–1271 (2015).
26. Yılmaz Tuğan, B. et al. Is tear proteome profile a predictor of developing uveitis in ANA-positive patients with oligoarticular juvenile idiopathic arthritis? *Graefes Arch. Clin. Exp. Ophthalmol.* **262**, 211–221 (2024).
27. Xu, S. et al. Proteomics analysis of five potential plasma-derived exosomal biomarkers for acute myocardial infarction. *Curr. Med. Chem.* **32**, 4816–4835 (2025).
28. Krishnaswamy, V. R., Balaguru, U. M., Chatterjee, S. & Korrapati, P. S. Dermatopontin augments angiogenesis and modulates the expression of transforming growth factor beta 1 and integrin alpha 3 beta 1 in endothelial cells. *Eur. J. Cell Biol.* **96**, 266–275 (2017).
29. Chen, J. et al. Converting lysozyme to hydrogel: a multifunctional wound dressing that is more than antibacterial. *Colloids Surf. B Biointerfaces* **219**, 112854 (2022).
30. Chen, T., Wang, Y., Xie, J., Qu, X. & Liu, C. Lysozyme amyloid fibril-integrated PEG injectable hydrogel adhesive with improved anti-swelling and antibacterial capabilities. *Biomacromolecules* **23**, 1376–1391 (2022).
31. Li, S. et al. Development of crosslinker-free polysaccharide–lysozyme microspheres for treatment enteric infection. *Polymers* **15**, 1077 (2023).
32. EASL Clinical Practice Guidelines on acute-on-chronic liver failure. *J. Hepatol.* **79**, 461–491 (2023).
33. Bajaj, J. S. et al. Second infections independently increase mortality in hospitalized patients with cirrhosis: the North American consortium for the study of end-stage liver disease (NACSELD) experience. *Hepatology* **56**, 2328–2335 (2012).
34. Blume, J. E. et al. Rapid, deep and precise profiling of the plasma proteome with multi-nanoparticle protein corona. *Nat. Commun.* **11**, 3662 (2020).
35. Liu, M. & Dongre, A. Proper imputation of missing values in proteomics datasets for differential expression analysis. *Brief. Bioinform.* **22**, bbaa112 (2021).

Acknowledgements

This work was supported by the National Key Research and Development Program of China (2023YFC2308800; R.J.), the Beijing Municipal Administration of Hospitals Incubating Program (PZ2024034; Y.H.), the High-Level Public Health Technical Personnel Construction Project (Discipline Backbone-03-07; Y.H.), and the Laboratory for Clinical Medicine, Capital Medical University (Y.H.). We thank all patients who participated in the study. We thank the technical team at Shanghai Bioprofile Technology (Shanghai, China) for the collaboration in the untargeted and targeted proteomics analysis.

Author contributions

Y.H., R.J., and Y.S. designed and supervised the research. Y.H. and R.J. are the principal investigators of the study. L.W., Y.J., L.L., T.Z., Y.Z., and T.W. collected the clinical data and samples. F.X., J.Z., and J.C. performed statistical analyses and Y.S. assisted in the bioinformatics analyses. F.X. drafted the manuscript. Y.H. revised the manuscript. All authors provided constructive criticism throughout the project and approved the final version of the paper.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41467-026-69075-y.

Correspondence and requests for materials should be addressed to Yamin Sun, Ronghua Jin or Yixin Hou.

Peer review information *Nature Communications* thanks the anonymous reviewers for their contribution to the peer review of this work. A peer review file is available.