Research Square

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

Diagnostic Biomarkers of Metabolic Reprogramming-Related Genes for Gastroesophageal Reflux Disease: Machine Learning Modeling, Molecular Mechanisms, Drug Prediction, and Experimental Validation

Jinjing Tan
IThe First Affiliated Hospital of Guangxi University of Chinese Medicine
Wenxiao Wang
Graduate School of Guangxi University of Chinese Medicine
Minjian Xie
Graduate School of Guangxi University of Chinese Medicine
Xiaoyun Lei
Graduate School of Guangxi University of Chinese Medicine
Zezhong Teng
Graduate School of Guangxi University of Chinese Medicine
Wenyu Wu
Graduate School of Guangxi University of Chinese Medicine
Xiaomei Tang
Graduate School of Guangxi University of Chinese Medicine
Sheng Xie
IThe First Affiliated Hospital of Guangxi University of Chinese Medicine
Liqun Li
lilq@gxtcmu.edu.cn

IThe First Affiliated Hospital of Guangxi University of Chinese Medicine

Research Article

Keywords: Gastroesophageal reflux disease, Metabolic reprogramming, Machine learning, Artificial neural network, Nomogram, Immune infiltration, AI-based drug prediction, Molecular docking, Animal validation

Posted Date: September 24th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10982253/v1

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

Additional Declarations: No competing interests reported.

Diagnostic Biomarkers of Metabolic Reprogramming-Related Genes for Gastroesophageal Reflux Disease: Machine Learning Modeling, Molecular Mechanisms, Drug Prediction, and Experimental Validation

Jinjing Tan[†]1, Wenxiao Wang[†]2, Minjian Xie[†]2, Xiaoyun Lei2, Zezhong Teng2, Wenyu Wu2, Xiaomei Tang2, Sheng Xie[*]1, Liqun Li[*]1

[†]Jinjing Tan, Wenxiao Wang and Minjian Xie contributed equally to this work and should be considered co-first authors.

[*]Sheng Xie and Liqun Li contributed equally to this work and should be considered joint corresponding authors.

[1]The First Affiliated Hospital of Guangxi University of Chinese Medicine, Nanning,530023, Guangxi, China

[2]Graduate School of Guangxi University of Chinese Medicine, Nanning, 530200, Guangxi, China

[*]Correspondence:

Sheng Xie：xies@gxtcmu.edu.cn, Liqun Li：lilq@gxtcmu.edu.cn

Abstract

Background: Metabolic reprogramming plays a critical role in the pathogenesis of gastroesophageal reflux disease (GERD). This study aimed to identify metabolic reprogramming-related diagnostic biomarkers, construct a machine learning-based diagnostic model, and explore the underlying mechanisms and potential therapeutic agents.

Methods: GERD transcriptomes from the Gene Expression Omnibus (GEO) were corrected for batch effects, and candidate genes were derived by overlapping differentially expressed genes, weighted gene co-expression network analysis (WGCNA) modules, and metabolic reprogramming gene sets. Models built with 107 machine-learning algorithms were ranked by the area under the receiver operating characteristic(ROC) curve (AUC), and the best-performing model was verified by an artificial neural network (ANN); a nomogram was then constructed from the core genes. Mechanisms were explored through protein–protein interaction (PPI) network analysis, functional enrichment, and CIBERSORT immune infiltration analysis. Potential therapeutic agents were predicted by AI-based drug screening and molecular docking. Finally, a rat model of GERD was established, and in vivo validation was performed using real-time PCR, western blot, and immunohistochemistry.

Results: The glmBoost+GBM ensemble model achieved the best performance using the seven screened candidate genes. All seven genes yielded AUC values > 0.7, and the top five genes ranked by AUC (*CKB*, *SH3GL2, JUP*, *LCN2*, and *CPA2*; AUC values > 0.80) were selected as core genes. ANN validation supported the model's robustness. The nomogram showed good calibration and net clinical benefit. Enrichment analysis

revealed pathways associated with cell adhesion, amino acid metabolism, and IL-17 signalling. Immune infiltration analysis indicated immune imbalance, and core genes were closely correlated with immune cell subsets. Three compounds with high binding affinity (alvimopan, baricitinib, and isoevodiamine) were identified, and the differential expression of all five genes in GERD was confirmed by animal experiments.

Conclusions: *CKB*, *SH3GL2*, *JUP*, *LCN2*, and *CPA2* are core diagnostic biomarkers of metabolic reprogramming in GERD. The diagnostic model and nomogram exhibit excellent discriminative performance.These genes may participate in GERD pathogenesis through metabolic–immune crosstalk, providing novel targets and a theoretical basis for early diagnosis, risk stratification, and targeted therapy.

Keywords: Gastroesophageal reflux disease; Metabolic reprogramming; Machine learning; Artificial neural network; Nomogram; Immune infiltration; AI-based drug prediction; Molecular docking; Animal validation

1.Introduction

Gastroesophageal reflux disease (GERD) denotes retrograde flow of gastric or duodenal contents into the oesophagus, manifesting as acid regurgitation and heartburn [1]. Its global prevalence, estimated at 13.98% [2], has climbed by 77.53% over two decades [3], yet its pathogenesis remains poorly defined. Proton pump inhibitors are still first-line therapy, but over half of treated patients remain symptomatic and some develop adverse effects [4]; chronic or refractory GERD also predisposes to coronary artery disease, oesophageal carcinoma, lung cancer, and interstitial pneumonia [5]. At present, no individualised risk-assessment tools or targeted drug-screening strategies are available in

clinical practice. Identifying key driver genes of GERD at the molecular level, discovering novel diagnostic biomarkers, constructing risk prediction models, and employing molecular docking to screen for targeted therapeutic agents therefore carry substantial clinical and research value.

Metabolic reprogramming describes the adaptive remodelling of core metabolic routes—among them glucose, lipid, and mitochondrial energy metabolism—under endogenous and exogenous stimulation, and it is widely involved in the onset and evolution of many diseases [6]. A growing body of work indicates that this process is critically implicated in the pathogenesis and progression of GERD, especially in Barrett's oesophagus (BE). For example, BE cells adapt to hypoxic conditions through the selective upregulation of glycolysis [7]. Meanwhile, the remodelling of mitochondrial energy and lipid metabolism contributes to BE progression [8], and metabolic changes are intimately tied to numerous cellular processes within the BE microenvironment [9]. As a pivotal constituent of this microenvironment, immune cells exert a central regulatory function. Patients with both reflux oesophagitis (RE) and BE display a disturbed T helper 17 (Th17)/regulatory T (Treg) ratio, which may be involved in the pathogenesis through the secretion of inflammatory cytokines [10]. Th17 cells draw on glycolysis and fatty acid synthesis to drive pro-inflammatory responses, whereas Treg cells depend on oxidative phosphorylation and fatty acid oxidation to sustain immunosuppression [11]. Taken together, these findings imply that metabolic reprogramming may promote the onset and progression of GERD by reconfiguring immune cell function and the inflammatory microenvironment. Even so, the core genes and mechanisms that connect metabolic reprogramming with GERD remain obscure. Thus, developing GERD diagnostic

biomarkers on the basis of metabolic reprogramming and cellular immunity, and clarifying the biological mechanisms involved constitute an important direction for further research.

With the widespread availability of high-throughput sequencing, transcriptomic data provide a comprehensive basis for identifying disease-associated signature genes at the molecular level. Traditional analytical methods such as Weighted Gene Co-expression Network Analysis (WGCNA), however, rely largely on linear assumptions and are insufficient for capturing complex non-linear interactions among genes [12]. Machine learning, which excels at recognising intrinsic patterns in high-dimensional data and selecting informative features from numerous variables, has shown considerable potential in diagnostic classification, prognostic assessment, and biomarker discovery [13]. Among these approaches, artificial neural networks (ANNs) can approximate the intricate mapping between input variables and disease status through multiple non-linear transformations, thereby improving discriminative accuracy [14]; when combined with a nomogram, they enable intuitive quantification of GERD risk and facilitate clinical translation [15]. Furthermore, integrating artificial intelligence (AI)-based virtual screening with molecular docking permits high-throughput identification of small-molecule compounds that target disease-relevant proteins, reducing early-stage drug-development costs and providing candidate molecules for targeted GERD intervention [16]. CIBERSORT, a deconvolution-based approach, further derives the fractional composition of 22 immune-cell populations directly from tissue transcriptomes, thereby charting the immune-infiltration landscape in silico [17].

To leverage these technical strengths, the present study integrated transcriptomic data from GERD oesophageal mucosa with a curated set of metabolic reprogramming-related genes. Differential expression screening, WGCNA module identification, and 107 machine-learning algorithms were applied sequentially to construct a diagnostic classifier. An ANN was then used to cross-validate the optimal feature combination, and a risk-prediction nomogram was developed from the core genes. In parallel, AI-based virtual screening coupled with molecular docking was used to identify potential targeted compounds. CIBERSORT was employed to dissect the association network linking the core biomarkers to the immune microenvironment. Finally, multi-level verification was performed through in vivo animal experiments, enabling systematic characterisation of the expression dynamics and potential regulatory pathways of these key genes during GERD pathogenesis.

2. Materials and Methods

2.1 Study design

The overall study design is presented in Figure 1. Initially, GERD expression profiles were retrieved from the Gene Expression Omnibus (GEO) database (http://www.ncbi.nlm.nih.gov/geo), from which differentially expressed genes (DEGs) were identified. WGCNA was subsequently performed to detect gene modules correlated with clinical characteristics, and genes overlapping with metabolic reprogramming-related genes were retained as candidates. These candidates were then entered into 107 machine-learning algorithms to establish a diagnostic model of GERD-associated metabolic reprogramming genes. Model performance was assessed by receiver operating

characteristic (ROC) curves and the corresponding area under the curve (AUC), while an artificial neural network (ANN) was adopted for independent validation. A nomogram intended to estimate disease risk was likewise built and evaluated. The biological roles of the core genes were further explored through protein–protein interaction (PPI) network analysis, Gene Ontology (GO) enrichment, Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway enrichment, Gene Set Variation Analysis (GSVA), and Gene Set Enrichment Analysis (GSEA). CIBERSORT was then used to characterise immune cell infiltration in oesophageal mucosal tissues and the correlations between core genes and immune cells. In addition, AI-based drug prediction and molecular docking were performed. Ultimately, the core genes were subjected to experimental validation. As all public datasets originated from open-access repositories, no further ethical approval was needed. The animal experiments were approved by the Animal Experiment Welfare and Ethics Committee of Guangxi University of Chinese Medicine (approval no. DW20230320-05).

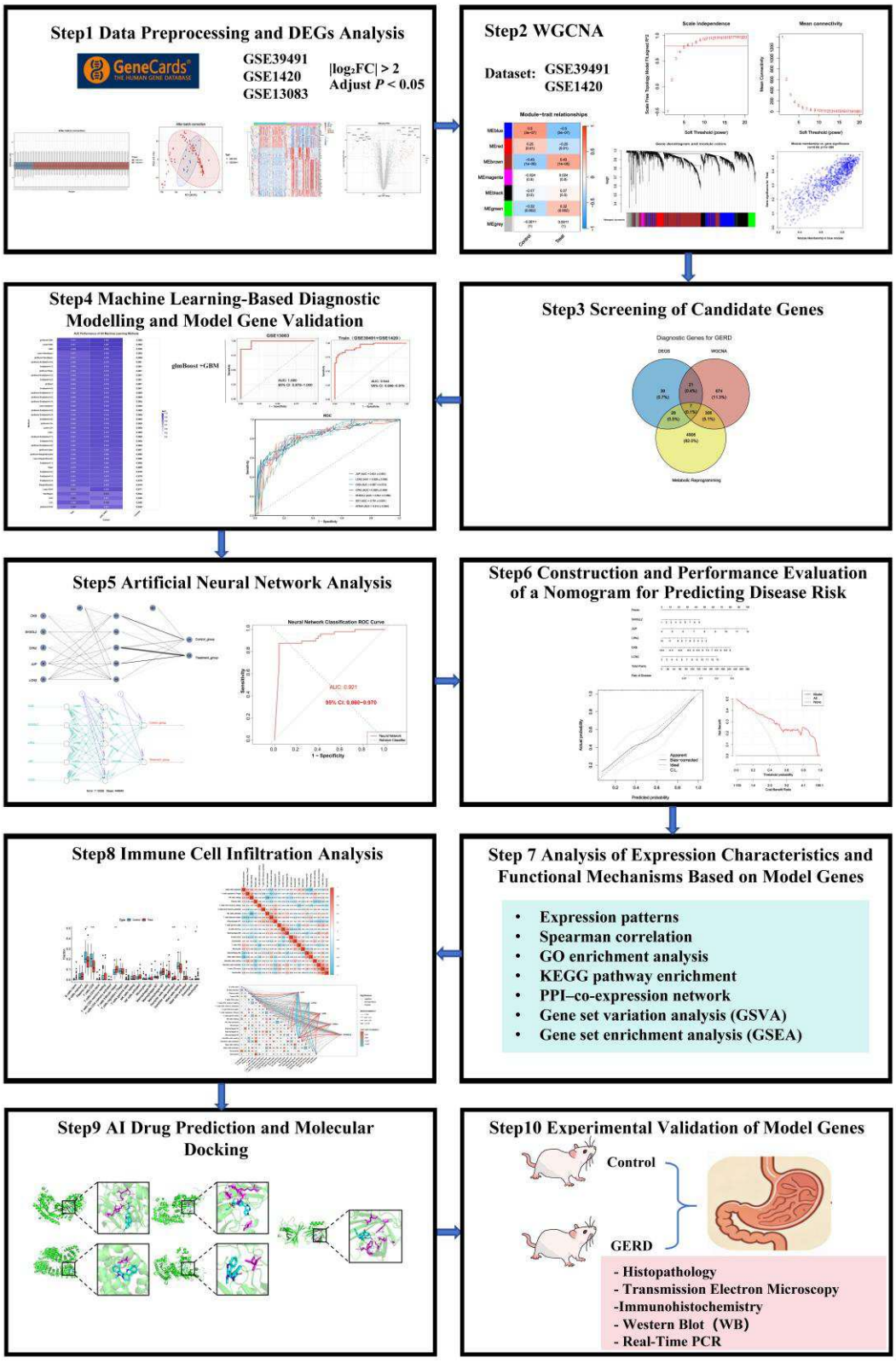


Fig. 1 Flow chart of the investigation.

2.2 Data collection and preprocessing

GERD-related gene expression profiles were systematically downloaded from the GEO database. Eligibility criteria were as follows: (i) human oesophageal mucosal specimens; (ii) microarray-based gene expression data; (iii) enrolment of both GERD patients and healthy controls; and (iv) at least three samples per group. Three datasets were finally included: GSE1420 (8 GERD, 8 controls), GSE39491 (43 GERD, 43 controls), and GSE13083 (7 GERD, 7 controls). GSE1420 and GSE39491 were merged into the training cohort, whereas GSE13083 acted as the validation cohort. Meanwhile, human genes were searched in the GeneCards database (https://www.genecards.org) with "metabolic reprogramming" as the keyword. The GeneCards relevance score integrates genomic, proteomic, and functional annotation evidence to quantify gene–query association strength. To balance coverage against specificity, the top 20% of hits by relevance score were chosen to establish the metabolic-reprogramming-related gene set, keeping genes with well-established metabolic roles and discarding entries with only marginal relevance.

2.3 Differential expression analysis

Following batch-effect correction, DEGs were screened in the merged training cohort (GSE1420 + GSE39491) by comparing disease and control samples with the limma package, applying cut-offs of $|\log_2(\text{fold change})| > 2$ and an FDR-adjusted $P < 0.05$. The pheatmap and ggplot2 packages were used to generate heatmaps and volcano plots, respectively.

2.4 Weighted gene co-expression network analysis

Once the merged GSE1420 and GSE39491 datasets underwent batch correction, genes with a standard deviation of >0.5 were kept as highly variable features. The average-linkage approach was adopted for sample clustering, and outlying samples were excluded using a cluster cut height of 20,000. The soft-thresholding power was determined automatically in accordance with scale-free topology criteria, after which an adjacency matrix was built and converted into a topological overlap matrix (TOM). Co-expression modules were detected through the dynamic tree-cut algorithm, setting the minimum module size to 60 and deepSplit to 2; highly similar modules were then combined under a 0.25 threshold. Pearson correlation analysis was applied to examine the associations of module eigengenes with the GERD phenotype, while gene significance (GS)–module membership (MM) correlations were computed to confirm the biological relevance of the modules. The entire procedure was implemented in R 4.3.3 with the limma and WGCNA packages.

2.5 Candidate gene screening

Candidate genes were obtained by intersecting the DEGs, the key WGCNA module genes, and the metabolic-reprogramming-related gene set. Overlaps were visualised using the ggplot2 and ggvenn packages. The resulting candidate genes were used for subsequent modelling and functional exploration.

2.6 Construction of a gene-based diagnostic model using machine learning

An ensemble of 107 machine-learning algorithms (Table S2) was used to construct a diagnostic model for GERD based on the candidate genes. GSE1420 and GSE39491 were combined as the training cohort, and GSE13083 served as the validation cohort. All

107 algorithms were systematically evaluated, and the models were ranked using the prespecified criterion of mean AUC across training and validation cohorts. The diagnostic performance of the model was evaluated using ROC curves and AUC.

2.7 Evaluation of candidate genes

ROC curves for the candidate genes were generated using the R packages ggplot2, scales, and pROC, and AUC values were calculated. An AUC $\geq$ 0.7 was regarded as clinically meaningful for diagnosing GERD, whereas an AUC $\geq$ 0.8 was considered to indicate a high-value diagnostic gene. To balance model parsimony with experimental feasibility, the top five candidate genes ranked by AUC were selected as core genes for subsequent analyses.

2.8 Validation of the model using an artificial neural network

An ANN was established to verify the diagnostic efficacy of the optimal model. The merged GSE1420 and GSE39491 datasets acted as the training cohort, whereas GSE13083 was used for validation, with the core-gene expression matrix supplied as the input layer. A single-hidden-layer feed-forward network comprising five neurons was implemented in R through the neuralnet package; the output layer used a sigmoid activation function, the iteration limit was set to 1×10^6, and the convergence threshold was 0.01. Classification accuracy was calculated from the confusion matrix. The pROC package was employed to plot ROC curves and obtain AUC values with 95% confidence intervals, and the resulting network structure was depicted with the NeuralNetTools package. R was used throughout the study.

2.9 Nomogram construction and performance evaluation

All computations were carried out in R 4.3.3. A multivariable logistic regression model was fitted with the rms package and used to build a nomogram. Model calibration was examined through bootstrap-based calibration curves, and the rmda package was applied to conduct decision curve analysis (DCA) to quantify net clinical benefit. All tests were two-sided, and $P < 0.05$ denoted statistical significance.

2.10 Expression characteristics and functional analysis of the model genes

Grouped box plots were generated using the ggpubr and ggsci packages to visualise gene expression patterns. Intergroup differences were compared using the Wilcoxon rank-sum test, with $P < 0.05$ considered significant. Co-expression relationships among genes were assessed by Spearman rank-correlation coefficients, and correlation heatmaps were generated using the pheatmap package.

To explore potential functional associations among the core genes, a PPI network was generated using the GeneMANIA online tool (https://genemania.org).

GO functional enrichment and KEGG pathway enrichment analyses were performed on the core gene set using the clusterProfiler and org.Hs.eg.db packages (R version 4.3.3) to elucidate the biological functions and pathways of the core genes.

GSEA and GSVA were applied to investigate the biological pathways in which the model genes participate. GSEA was conducted with clusterProfiler using the c2.cp.kegg.Hs.symbols.gmt gene set and 1,000 permutations ($P < 0.05$), and results were visualised with the gseaplot2 function. GSVA employed the single-sample GSEA (ssGSEA) algorithm to calculate pathway activity; pathways with differential activity

were identified using Student's t-test ($P < 0.05$), and bar plots were generated using ggpubr.

2.11 Immune infiltration landscape analysis

The relative fractions of 22 immune cell types were estimated with the CIBERSORT algorithm, and between-group differences in these fractions were displayed as box plots. In GERD samples, a Spearman correlation heatmap was built to explore interactions among immune cells. Spearman correlation was likewise applied to evaluate the relationships between key metabolic genes and the extent of immune cell infiltration. Correlations reaching statistical significance ($P < 0.05$) were plotted with ggplot2, whereas the linkET package served to construct an integrated correlation network that depicted both interactions among immune cells and gene–immune-cell correlations.

2.12 AI-based drug prediction and molecular docking

To identify candidate compounds that might act on proteins encoded by the core genes, an AI-based drug-screening pipeline (DrugRefLector) was built to carry out virtual prediction using publicly accessible molecular datasets. Molecular docking was subsequently conducted to determine how strongly the predicted compounds bound to these core-gene-encoded proteins. The crystal structures of the target proteins were downloaded from the RCSB Protein Data Bank (PDB), and the ligand structures were acquired from PubChem. AutoDock 4.2.6 was used to prepare the protein and ligand structures, after which docking calculations were run in AutoDock Vina 1.2.7. Conformations exhibiting binding energies lower than −8.0 kcal/mol were examined with

PyMOL; the best binding poses were chosen to characterise protein–ligand interaction modes and thereby pinpoint high-affinity candidate compounds.

2.13 Animal experiments

2.13.1 Animal model construction and validation

A rat GERD model was developed to verify the diagnostic utility of the five model genes with the highest AUC rankings. Twelve sex-balanced specific-pathogen-free (SPF) Sprague-Dawley rats weighing 160–200 g were supplied by Hunan SJA Laboratory Animal Co., Ltd (Hunan, China; licence No. SCXK (Xiang) 2019-0004). The animals underwent 7 days of acclimatisation in an SPF facility with regulated conditions (20–22°C, 50–60% humidity, and a 12-h light/dark cycle) and unrestricted access to food and water, after which they were randomised into a normal control group (n = 6) and a GERD model group (n = 6). Via daily gavage, the model group alternately received 0.1% iodoacetamide (dissolved in 2% sucrose solution) and a pepsin–deoxycholic acid mixture (each at 10 mg/mL in normal saline, pH 5.0), 0.2 mL per administration, for 60 successive days, using a protocol modified from a published study [18]. Controls were gavaged with an equal volume of normal saline.On completion of the 60-day protocol, all rats were humanely euthanised by a two-step sodium pentobarbitone regimen prior to tissue harvesting. Animals were first deeply anaesthetised by intraperitoneal injection at 50 mg/kg, with unconsciousness confirmed by loss of the corneal and pedal withdrawal reflexes, followed by a lethal overdose of the same agent (150 mg/kg i.p.). Death was verified by cessation of spontaneous respiration and cardiac activity.

2.13.2 Haematoxylin and eosin staining and immunohistochemistry

For haematoxylin and eosin staining, oesophageal tissue samples underwent fixation, rinsing, dehydration, and paraffin embedding. A rotary microtome (RM2135, Leica, Germany) was used to cut 3-μm sections, which were slide-mounted, baked, deparaffinised in xylene, and rehydrated through a graded ethanol series. Sections were stained with haematoxylin (BA-4041, Baso, China), differentiated in differentiating solution (C0163M, Beyotime, China), blued and rinsed in water, and then counterstained with eosin (BA-4024, Baso, China). After repeated graded-ethanol dehydration and xylene clearing, the sections were mounted with neutral balsam (ZLI-9555, ZSGB-BIO, China), and images were captured under a light microscope (BX43, Olympus, Japan).

For immunohistochemistry (IHC), paraffin-embedded oesophageal sections were deparaffinised and rehydrated, followed by high-pressure antigen retrieval in 3% citric acid buffer. Endogenous peroxidase was inactivated with 3% methanol–hydrogen peroxide for 20 min at room temperature. After phosphate-buffered saline (PBS) washes, sections were blocked with 5% bovine serum albumin (BSA; AR0004, Boster, China) for 30 min at room temperature and then incubated overnight at 4°C with the following primary antibodies: anti-CPA2 (822139, ZEN-BIO, China; 1:90), anti-CKB (23981, ZEN-BIO, China; 1:100), anti-JUP (AF0126, Affinity Biosciences, China; 1:110), anti-LCN2 (DF6816, Affinity Biosciences, China; 1:90), and anti-SH3GL2 (12435-1-ap, Proteintech, China; 1:130). Following three PBS rinses, a secondary antibody (PV-6000, ZSGB-BIO, China) was applied for 20 min at 37°C. Immunoreactivity was developed with 3,3′-diaminobenzidine (DAB) substrate (ZLI-9018, ZSGB-BIO, China), and sections were counterstained with haematoxylin before being differentiated, blued,

dehydrated, cleared, and mounted. Positive staining was brownish-yellow, while nuclei were stained blue.

2.13.3 Real-time PCR

Total RNA was isolated from rat oesophageal tissues with TRIzol reagent (15596-026, Invitrogen, USA), and its concentration and purity were measured on a microvolume spectrophotometer (DS-11, Denovix, USA). A 1-μg aliquot of RNA was reverse-transcribed to cDNA with SuperRT III Reverse Transcription Premix (BL1020B, Biosharp, China), which simultaneously eliminates genomic DNA. Quantitative real-time PCR was run on a CFX Connect system (Bio-Rad, USA) with an RT-PCR kit (BL697A, Biosharp, China). With GAPDH as the internal reference, relative expression was quantified by the $2^{-}\Delta\Delta Ct$ method, and all primer sequences are provided in Table 1.

Table 1 Primers sequences

Gene name	Forward primer (5'→3')	Reverse primer (5'→3')	Amplicon length (bp)
CPA2	AGACTGCACGGCACCAGTTATA	ATTTGATGCCAAGGTCGTAAGC	100
CKB	TTCAGAAGCGAGGCACAGG	CACTCCGTCCACCACCATC	122
SH3GL2	AAGTGAGTGAGAAGGTGGGAGGAG	GCTGGATTGGGTTGGAGGTATT	142
JUP	AAAACAAGACGCTGGTGACGC	CCGGGATTCCATAGTTGAGGC	180
LCN2	GCCGACACTGACTACGAC	GAGACTTGGCAAAGCTGAC	145

2.13.4 Western blot

Approximately 100 mg of oesophageal tissue was homogenised and lysed on ice in RIPA lysis buffer (R0020, Solarbio, China) supplemented with PMSF (BL507A, Biosharp, China) and phosphatase inhibitors (BL636A, Biosharp, China). Following centrifugation (4°C, 12,000 rpm, 10 min), the supernatant was harvested and protein concentrations were measured with a spectrophotometer (DS-11, Denovix, USA). Protein separation was achieved by SDS-PAGE, comprising a 10% separating gel and a 5% stacking gel, on an electrophoresis apparatus (DYCZ-24DN, Beijing Liuyi, China), after which proteins were wet-transferred onto PVDF membranes (IPVH00010, Millipore, USA). After 1 h of blocking at room temperature, membranes were probed overnight at 4°C with the following primary antibodies: anti-SH3GL2 (12435-1-ap, Proteintech; 1:1100), anti-CPA2 (822139, ZEN-BIO; 1:1000), anti-CKB (R23981, ZEN-BIO; 1:1000), anti-JUP (AF0125, Affinity Biosciences; 1:1000), anti-LCN2 (DF6816, Affinity Biosciences; 1:1000), together with the loading controls GAPDH (T0004, Affinity Biosciences; 1:10000) and β-tubulin (AF7011, Affinity Biosciences; 1:8000). The membranes were washed three times for 10 min each in TBST and then incubated for 1 h at room temperature with HRP-conjugated goat anti-rabbit or anti-mouse secondary antibodies (ZB2301/ZB2305, ZSGB-BIO; 1:3000), followed by additional TBST washes. Signals were detected with ECL substrate (BMU102-CN, Abbkine, China) and exposed on X-ray film (SUPER RX-N, FUJIFILM, Japan). Films were scanned (V19, EPSON, Japan), band density was quantified in ImageJ, and relative protein expression was normalised to the corresponding loading controls.

2.13.5 Statistical analysis

Each experiment was conducted in triplicate. Continuous variables are presented as the mean ± standard deviation (SD), and all animal data were processed in SPSS 22.0. For control-versus-model comparisons, the independent-samples t-test was used when normality and variance homogeneity held; failing these assumptions, the Mann–Whitney *U* test was adopted instead. A *P* value below 0.05 was regarded as statistically significant.

3. Results

3.1 Differential expression analysis, WGCNA, and candidate gene screening

Following normalisation and batch-effect removal, box plots indicated broadly comparable expression levels among all training-cohort samples (Fig. S1A), while PCA clustering demonstrated that batch effects had been effectively eliminated (Fig. S1B). Under the cut-offs of $|\log_2 FC| > 2$ and adjusted $P < 0.05$, 97 significant DEGs were detected in the training cohort, 75 upregulated and 22 downregulated (Supplementary Table 1). The top 50 DEGs are shown in the heatmap (Fig. 2A), whereas the volcano plot marks the 10 most upregulated and 10 most downregulated genes (Fig. 2B).
During WGCNA, the soft-thresholding power was chosen as $\beta = 6$, the value at which the network met the scale-free topology criterion (Fig. 2C). Dynamic tree-cut clustering of genes produced several co-expression modules, among which the blue module was the largest (Fig. 2D). Module–trait analysis (Fig. 2E) showed a significant association between the blue module and the target trait ($r = 0.5$, $P = 3 \times 10^{-7}$), and MM and GS within this module exhibited a strong positive correlation ($r = 0.84$, $P < 1 \times 10^{-200}$; Fig. 2F). Taking the overlap among DEGs, blue-module genes, and the metabolic-

reprogramming-related gene set, seven candidate genes emerged: *SH3GL2*, *JUP*, *CPA2*, *CKB*, *LCN2*, *ATP4A*, and *SST* (Fig. 2G).

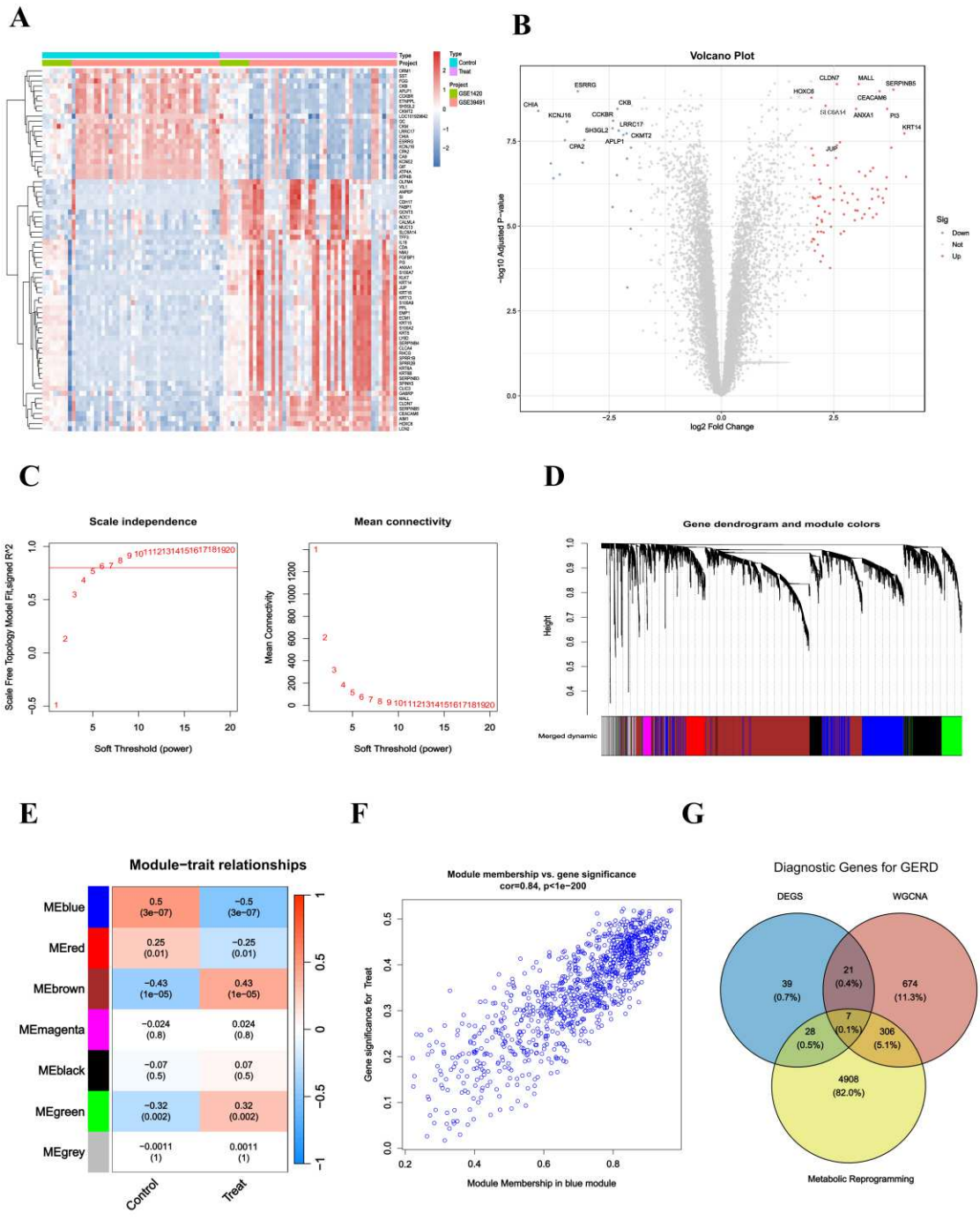


Fig. 2 Differential expression analysis, WGCNA, and candidate gene screening. A, Heatmap of DEGs; B, Volcano plot of DEGs; C, Soft-thresholding power selection; D, Gene clustering dendrogram and module identification; E, Module–trait correlations; F, MM versus GS scatter plot for the blue module; G, Venn diagram of overlapping genes.

3.2 Machine learning model construction and ANN validation

Using the seven candidate genes as input features, 107 machine-learning algorithms were evaluated. After pre-screening, 39 diagnostic models met the inclusion criteria (Fig. 3A, Tables S3 and S4). Based on mean AUC, the top five algorithms were glmBoost+GBM, Lasso+GBM, GBM, Lasso+NaiveBayes, and glmBoost+NaiveBayes. Among these, glmBoost+GBM performed best, with an AUC of 0.944 (95% CI: 0.898–0.978) in the training cohort and 1.000 (95% CI: 0.878–1.000) in the independent validation cohort (Fig. 3B).

Evaluation of individual genes in the glmBoost+GBM model (Fig. 3C) showed that *CKB* had the highest AUC (0.857 ± 0.073), followed by *SH3GL2* (0.841 ± 0.080), *JUP* (0.831 ± 0.081), *LCN2* (0.828 ± 0.085), *CPA2* (0.826 ± 0.084), *ATP4A* (0.813 ± 0.090), and *SST* (0.791 ± 0.091). To balance model simplicity with experimental feasibility, the top five genes by AUC were selected as core genes for subsequent analysis.

An ANN model was then constructed using the five core genes (Figs. S2A–S2B). In the training cohort, the model error converged to 7.12026 after 446,945 iterations (Fig. S2C), with an AUC of 0.921 (95% CI: 0.860–0.970) (Fig. 3D). In the validation cohort, the model error decreased to 0.004028 after 6,009 iterations (Fig. S2D), yielding an AUC of 1.000; given the small validation cohort (14 samples), the 95% CI was not reliably

estimated. These results indicate that the five-gene panel accurately distinguishes patients with GERD from healthy controls, demonstrating excellent discriminative performance and clinical translation potential.

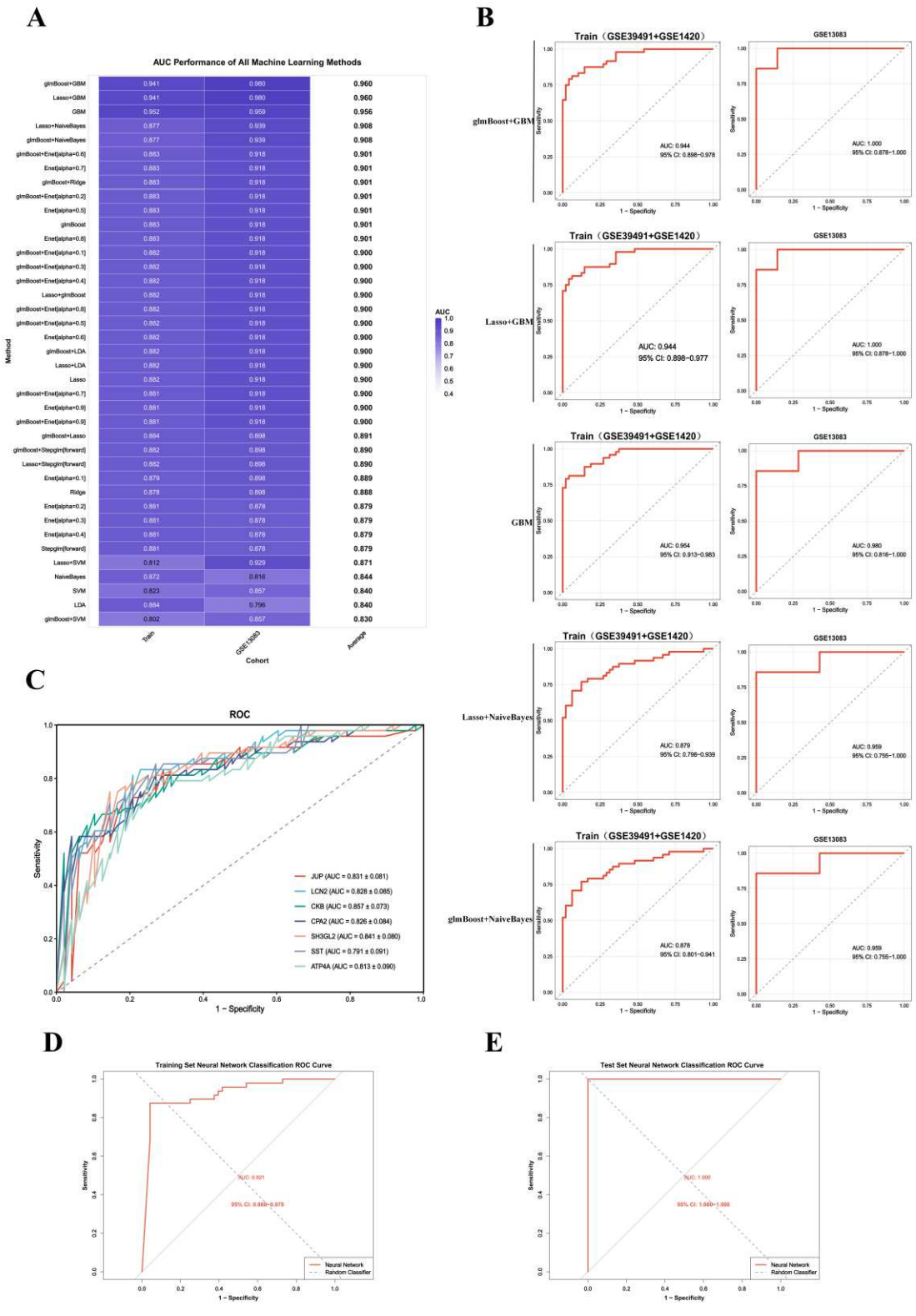

A
AUC Performance of All Machine Learning Methods
Method
Cohort
Train
GSE13083
Average
AUC
B
Train（GSE39491+GSE1420）
GSE13083
glmBoost+GBM
Lasso+GBM
GBM
Lasso+NaiveBayes
glmBoost+NaiveBayes
Sensitivity
1 − Specificity
C
ROC
JUP (AUC = 0.831 ± 0.081)
LCN2 (AUC = 0.828 ± 0.085)
CKB (AUC = 0.857 ± 0.073)
CPA2 (AUC = 0.826 ± 0.084)
SH3GL2 (AUC = 0.841 ± 0.080)
SST (AUC = 0.791 ± 0.091)
ATP4A (AUC = 0.813 ± 0.090)
D
Training Set Neural Network Classification ROC Curve
E
Test Set Neural Network Classification ROC Curve
Neural Network
Random Classifier

Fig. 3 Machine learning model construction and ANN validation. A, Performance comparison of 39 algorithms; B, ROC curves of the top five models; C, Single-gene ROC curves for the seven candidates; D, ANN ROC curve in the training cohort; E, ANN ROC curve in the validation cohort.

3.3 Nomogram for disease-risk prediction

Using the five core genes—*SH3GL2*, *JUP*, *CPA2*, *CKB*, and *LCN2*—a nomogram was established (Fig. 4A). As shown by the calibration curve (Fig. 4B), predicted and observed probabilities were in close concordance. Decision curve analysis (Fig. 4C) indicated that the model yielded higher net benefit than either the treat-all or the treat-none strategy over the clinically meaningful threshold interval, thereby confirming its predictive performance and prospective clinical value.

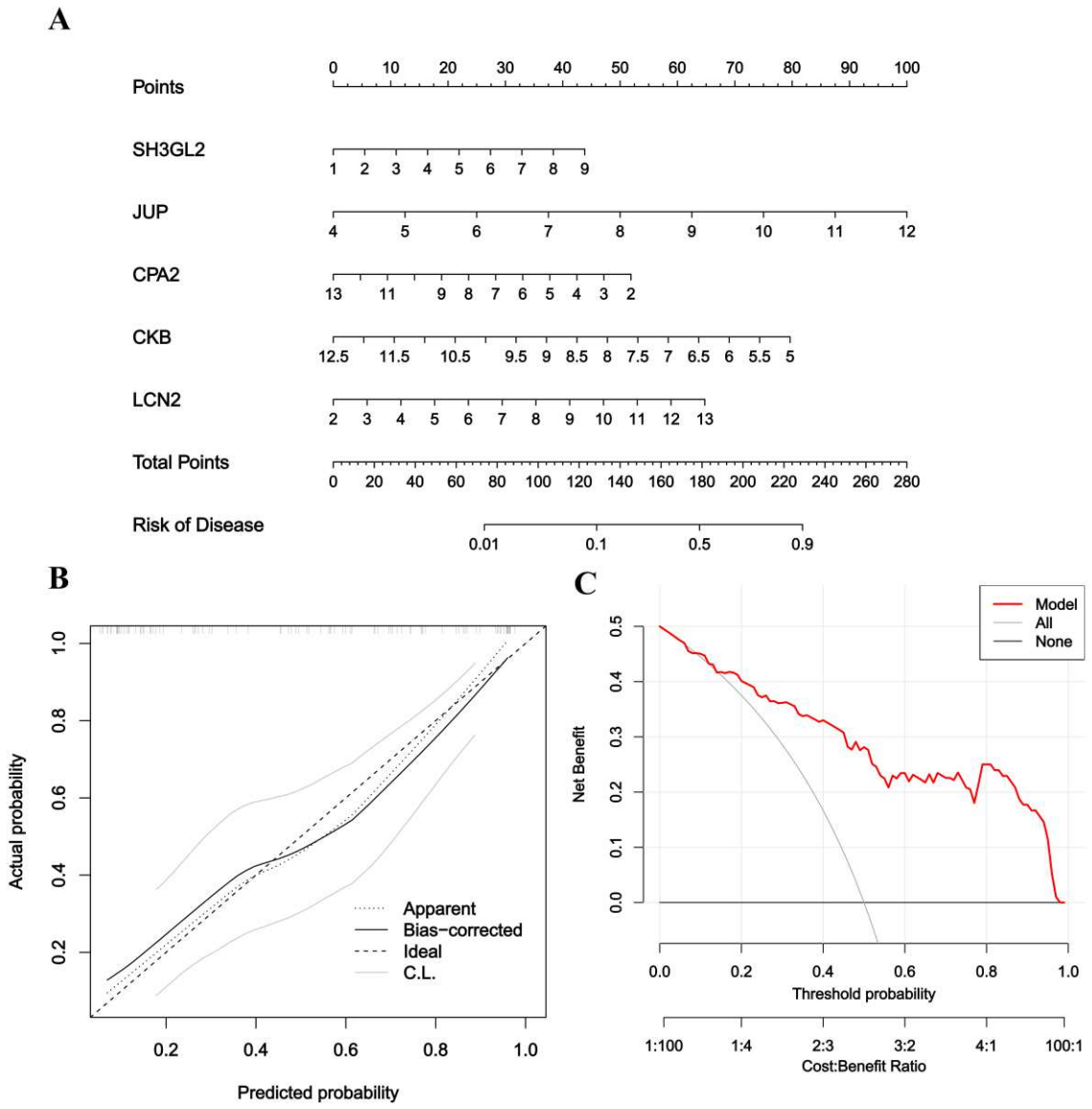


Fig. 4 Nomogram construction and evaluation. A, Risk-prediction nomogram; B, Calibration curve; C, Decision curve analysis.

3.4 Expression characteristics and functional annotation of core genes

Box plots showed that *JUP* and *LCN2* were significantly upregulated in the GERD group, whereas *CKB*, *CPA2*, and *SH3GL2* were significantly downregulated (Fig. 5A). Correlation analysis revealed the strongest co-expression relationship between *CKB* and

SH3GL2 (r = 0.87; Fig. 5B). The PPI network further illustrated the interactions among the corresponding proteins (Fig. 5C).

GO enrichment analysis identified 202 enriched terms (Fig. 5D, Table S5): 143 biological process terms, examples of which include cell adhesion, iron-ion transport, metabolic regulation, and vesicle-mediated signalling; 37 cellular component terms, mainly localised to secretory granules, vesicles, cell junctions, and membrane-associated structures; and 22 molecular function terms, largely related to ion binding, cell-adhesion mediator activity, carboxypeptidase activity, and protein–protein binding.

KEGG analysis yielded 10 enriched pathways (Fig. 5E, Table S6), including arginine and proline metabolism, digestive-system-related pathways, oncogenic pathways, the IL-17 signalling pathway, cytoskeletal regulation, and endocytosis.

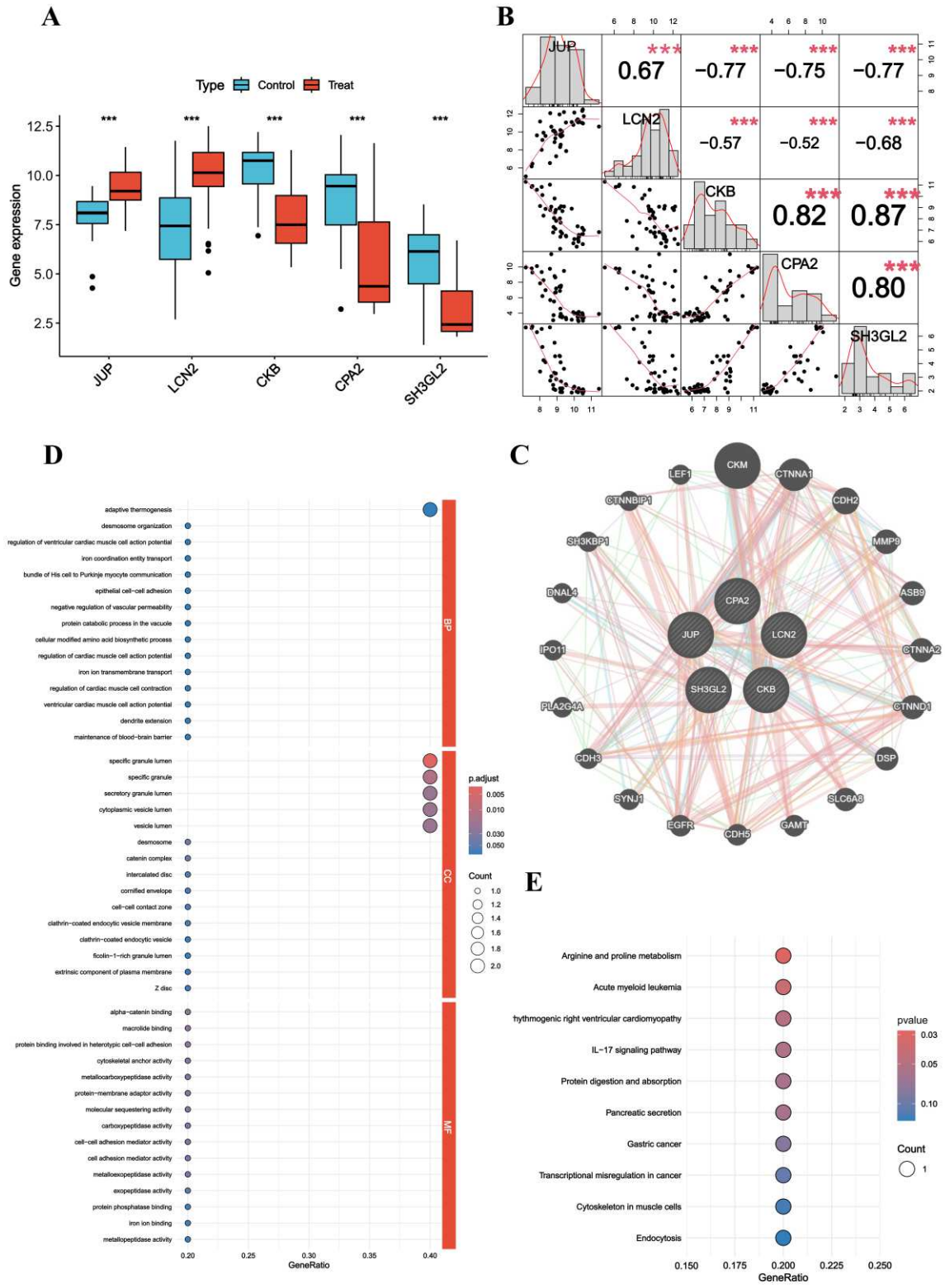

A
Type Control Treat
Gene expression
JUP
LCN2
CKB
CPA2
SH3GL2
B
0.67
−0.77
−0.75
−0.77
−0.57
−0.52
−0.68
0.82
0.87
0.80
C
CKM
LEF1
CTNNA1
CTNNBIP1
CDH2
SH3KBP1
MMP9
DNAL4
ASB9
CPA2
JUP
LCN2
IPO11
CTNNA2
SH3GL2
CKB
PLA2G4A
CTNND1
CDH3
DSP
SYNJ1
SLC6A8
EGFR
CDH5
GAMT
D
adaptive thermogenesis
desmosome organization
regulation of ventricular cardiac muscle cell action potential
iron coordination entity transport
bundle of His cell to Purkinje myocyte communication
epithelial cell−cell adhesion
negative regulation of vascular permeability
protein catabolic process in the vacuole
cellular modified amino acid biosynthetic process
regulation of cardiac muscle cell action potential
iron ion transmembrane transport
regulation of cardiac muscle cell contraction
ventricular cardiac muscle cell action potential
dendrite extension
maintenance of blood−brain barrier
specific granule lumen
specific granule
secretory granule lumen
cytoplasmic vesicle lumen
vesicle lumen
desmosome
catenin complex
intercalated disc
cornified envelope
cell−cell contact zone
clathrin−coated endocytic vesicle membrane
clathrin−coated endocytic vesicle
ficolin−1−rich granule lumen
extrinsic component of plasma membrane
Z disc
alpha−catenin binding
macrolide binding
protein binding involved in heterotypic cell−cell adhesion
cytoskeletal anchor activity
metallocarboxypeptidase activity
protein−membrane adaptor activity
molecular sequestering activity
carboxypeptidase activity
cell−cell adhesion mediator activity
cell adhesion mediator activity
metalloexopeptidase activity
exopeptidase activity
protein phosphatase binding
iron ion binding
metallopeptidase activity
BP
CC
MF
p.adjust
0.005
0.010
0.030
0.050
Count
1.0
1.2
1.4
1.6
1.8
2.0
GeneRatio
E
Arginine and proline metabolism
Acute myeloid leukemia
hythmogenic right ventricular cardiomyopathy
IL−17 signaling pathway
Protein digestion and absorption
Pancreatic secretion
Gastric cancer
Transcriptional misregulation in cancer
Cytoskeleton in muscle cells
Endocytosis
pvalue
0.03
0.05
0.10
Count
1
GeneRatio

Fig. 5 Expression characteristics and functional annotation. A, Box plots of core gene expression; B, Correlation matrix among core genes; C, PPI network (GeneMANIA); D, GO enrichment; E, KEGG enrichment.

3.5 GSEA, GSVA, and immune infiltration analyses

GSEA (Fig. 6A, Table S7) showed that *CKB*, *CPA2*, and *SH3GL2* were negatively enriched in the cell cycle and p53 signalling pathways but positively enriched in oxidative phosphorylation and ECM–receptor interaction pathways. *JUP* and *LCN2* showed the opposite pattern, with positive enrichment in the cell cycle pathway and negative enrichment in the p53 and oxidative phosphorylation pathways.

GSVA confirmed these findings (Fig. 6B, Table S8). In the *CKB*-high group, cell-cycle pathway activity was reduced whereas oxidative phosphorylation activity was elevated. In the *CPA2*- and *SH3GL2*-high groups, cell-cycle-related activities (including oocyte meiosis and p53 signalling) were decreased and oxidative phosphorylation increased. In the *LCN2*-high group, cell-cycle and arachidonic acid metabolism activities were elevated; in the *JUP*-high group, dorsoventral axis formation and ErbB signalling were upregulated.

CIBERSORT estimation of the 22 immune cell types revealed that, relative to controls, the GERD group displayed markedly decreased fractions of plasma cells and $CD8^+$ T cells and markedly increased fractions of T follicular helper cells, activated dendritic cells, activated mast cells, and neutrophils (all $P < 0.05$; Fig. 6C). Among inter-cellular relationships, the most pronounced negative association was observed between activated and resting NK cells ($r = -0.69$; Fig. 6D). The correlation analysis of core metabolic

genes and immune cell infiltration (Fig. 6E) showed the following patterns. *JUP* correlated negatively with plasma cells, $CD8^+$ T cells, and M1 macrophages, and positively with resting dendritic cells, activated dendritic cells, naïve $CD4^+$ T cells, and neutrophils. *LCN2* showed positive associations with resting dendritic cells and neutrophils. For *CKB*, negative associations were found with naïve $CD4^+$ T cells, M0 macrophages, resting dendritic cells, activated dendritic cells, and neutrophils, whereas a positive association was observed with $CD8^+$ T cells. *CPA2* was inversely associated with both resting and activated dendritic cells but positively associated with plasma cells. As for *SH3GL2*, it was inversely related to resting dendritic cells, activated dendritic cells, and naïve $CD4^+$ T cells, and directly related to plasma cells and $CD8^+$ T cells.

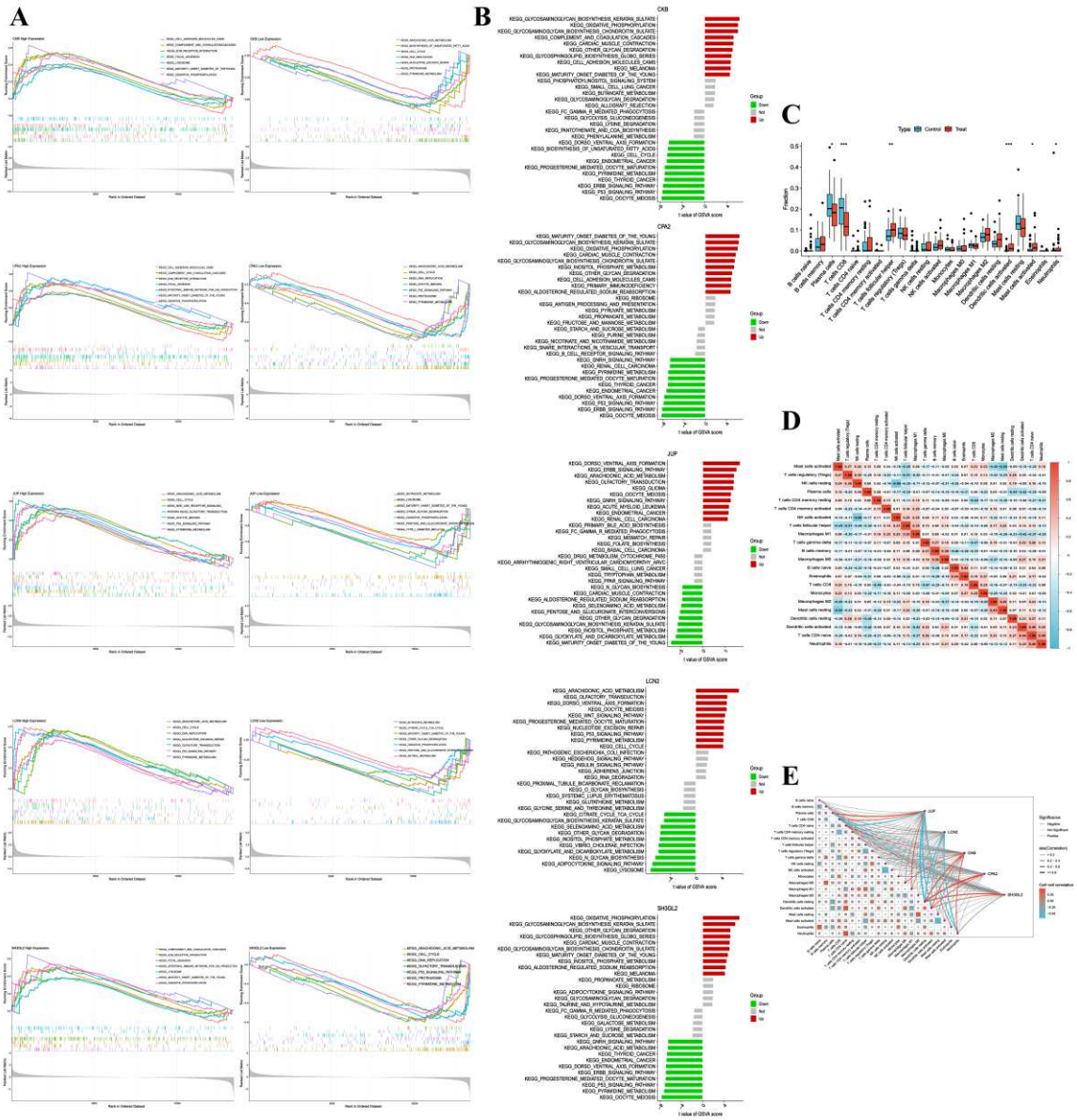


Fig. 6 GSEA, GSVA, and immune infiltration. A, GSEA results; B, GSVA results; C, Immune cell proportions; D, Immune cell correlation matrix; E, Gene–immune cell correlation network. *P < 0.05, **P < 0.01, ***P < 0.001.

3.6 AI-based drug prediction and molecular docking

AI-assisted screening identified three compounds—alvimopan, baricitinib, and isoevodiamine—which were docked against all five target proteins. Binding energies

ranged from −9.1 to −6.4 kcal/mol; alvimopan showed the strongest affinity for CKB (−9.1 kcal/mol). Inspection of complexes with binding energies < −8 kcal/mol (Figs. 7A–E) revealed stable hydrogen bonds with key amino acid residues, indicating that all three compounds stably bound the targets and warrant further investigation as candidate therapeutic molecules. The binding energies of all combinations are summarised in Table 2.

Table 2. Molecular docking binding energies.

Target protein	**Alvimopan (kcal/mol)**	**Baricitinib (kcal/mol)**	**Isoevodiamine (kcal/mol)**
CKB	−9.1	−8.9	−8.3
SH3GL2	−7.1	−6.7	−7.8
JUP	−7.8	−7.1	−8.6
CPA2	−6.6	−6.4	−6.5
LCN2	−8.2	−7.8	−8.0

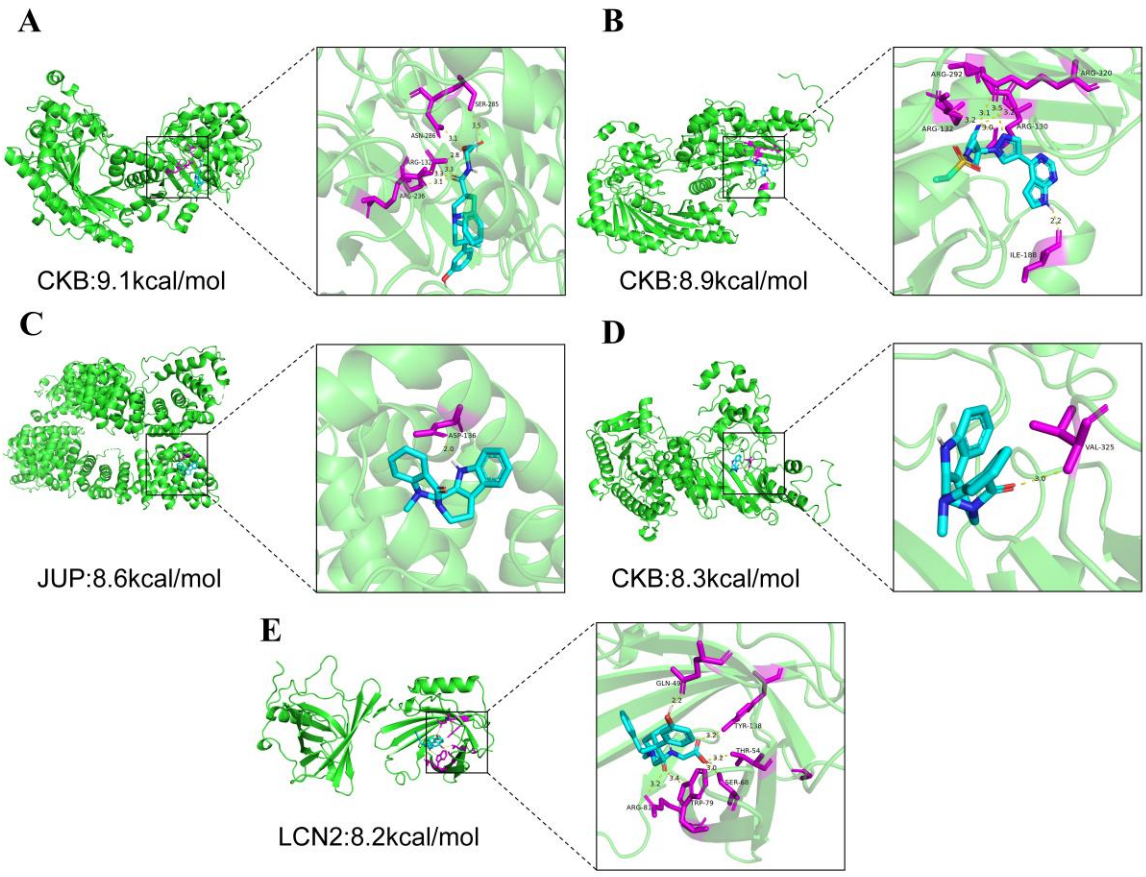


Fig. 7 Molecular docking. A, CKB–alvimopan; B, CKB–baricitinib; C, JUP–isoevodiamine; D, CKB–isoevodiamine; E, LCN2–alvimopan.

3.7 Animal experiment results

H&E staining revealed marked pathological changes in the GERD model group, including disorganised squamous epithelium, basal cell hyperplasia, elongated papillae, widened intercellular spaces, and inflammatory cell infiltration in the epithelium and lamina propria (Fig. 8A). IHC showed that *JUP* and *LCN2* protein levels were elevated in the model group (both $P < 0.0001$), whereas *CKB*, *CPA2*, and *SH3GL2* levels were decreased (all $P < 0.0001$; Figs. 8B, 8C). Real-time PCR confirmed upregulation of *JUP* ($P < 0.01$) and *LCN2* ($P < 0.001$) and downregulation of *CKB* ($P < 0.0001$), *CPA2* ($P <$

0.001), and *SH3GL2* ($P < 0.001$) at the mRNA level (Fig. 8D). Western blot analysis showed concordant changes at the protein level, all statistically significant (all $P < 0.0001$) (Fig. 8E).

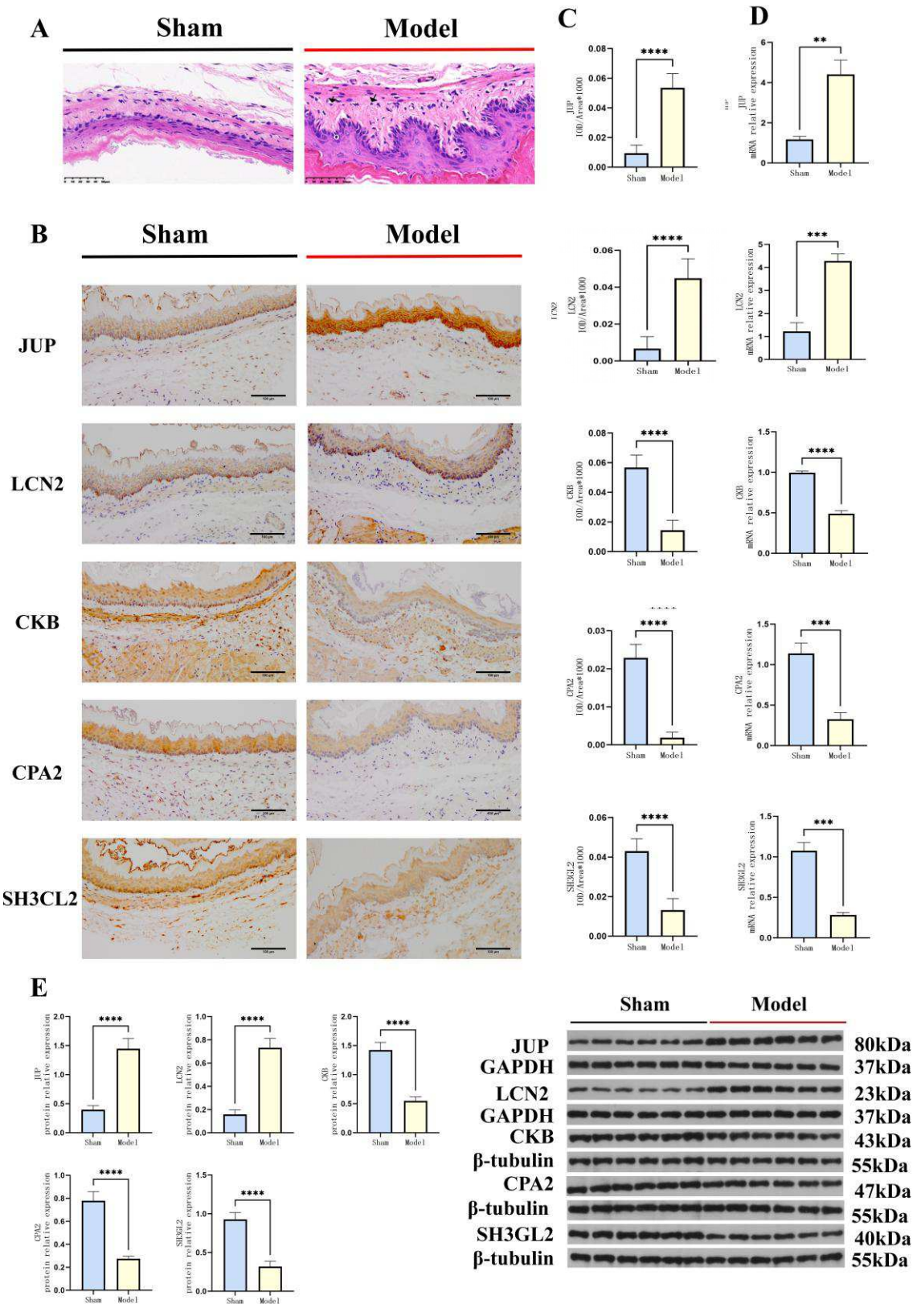

A
Sham
Model
B
Sham
Model
JUP
LCN2
CKB
CPA2
SH3CL2
C
D
E
Sham
Model
JUP
GAPDH
LCN2
GAPDH
CKB
β-tubulin
CPA2
β-tubulin
SH3GL2
β-tubulin
80kDa
37kDa
23kDa
37kDa
43kDa
55kDa
47kDa
55kDa
40kDa
55kDa

Fig. 8 Animal validation. A, H&E staining (×400; scale bar 50 μm); B, IHC (×200; scale bar 100 μm); C, IHC quantification; D, Relative mRNA expression (RT-PCR); E, Western blot. GAPDH and β-tubulin served as loading controls. *P < 0.05, **P < 0.01, ***P < 0.001, ****P < 0.0001 vs. control.

4. Discussion

Previous work has established that metabolic reprogramming contributes to oesophageal mucosal injury and persistent inflammation in GERD [19], but a systematic strategy for identifying diagnostically valuable biomarkers from metabolic-reprogramming gene sets is lacking. We addressed this gap through a multi-stage approach. Starting from oesophageal tissue microarray profiles, we first identified disease-relevant modules by differential expression analysis and WGCNA, and then intersected them with a metabolic-reprogramming gene set to obtain seven candidate genes. Iterative comparison of 107 machine-learning algorithms yielded five hub genes—*SH3GL2*, *JUP*, *CPA2*, *CKB*, and *LCN2*—with reliable diagnostic performance. We then integrated ANN cross-validation, nomogram construction, CIBERSORT immune-infiltration analysis, and multidimensional enrichment analysis, coupled with AI-based virtual screening and molecular docking, to systematically elucidate the diagnostic and therapeutic potential of these genes from multiple dimensions, including model robustness, clinical interpretability, association with the immune microenvironment, in-depth exploration of molecular mechanisms, and targeted drug prediction. Together, these findings position the five-gene signature as a candidate diagnostic and therapeutic target panel for GERD that merits further translational investigation.

SH3GL2 encodes endophilin A1, an endocytosis-regulatory protein involved in receptor signalling and cell adhesion [20]. In this study, *SH3GL2* was significantly downregulated in the oesophageal mucosa of patients with GERD, with a diagnostic AUC of 0.841; animal experiments confirmed reductions at both mRNA and protein levels. This is consistent with previous reports that repeated exposure to acid and bile acids lowers *SH3GL2* expression [21]. Downregulation of *SH3GL2* can activate STAT3/MMP2 signalling, promoting cell migration and invasion, and STAT3 is aberrantly activated in the oesophageal tissue of patients with GERD [22,23]. In addition, mitochondrial translocation and phosphorylation of endophilin A1 can induce ROS production and regulate mitochondrial fusion and apoptosis [24], suggesting a direct role in mitochondrial metabolism. Our GSEA and GSVA results showed that *SH3GL2* was positively enriched in oxidative phosphorylation and negatively enriched in cell cycle pathways, suggesting that *SH3GL2* normally promotes oxidative phosphorylation. Its downregulation in GERD may therefore impair mitochondrial oxidative capacity in the oesophageal epithelium, shifting metabolism from oxidative phosphorylation towards glycolysis and contributing to metabolic reprogramming.

CKB encodes the B subunit of creatine kinase, which maintains intracellular ATP homeostasis and energy buffering through reversible phosphotransfer [25]. *CKB* was significantly downregulated in the oesophageal mucosa of patients with GERD (AUC 0.857), and animal experiments confirmed this finding. Endoplasmic reticulum stress has been shown to epigenetically silence *CKB* through the IRE1–XBP1s–DNMT3A axis, driving a shift towards glycolysis and increasing the secretion of the pro-inflammatory chemokine CCL2 [26]. Mechanistically, *CKB* promotes mitochondrial ATP production

and reduces glycolytic ATP generation by inhibiting AKT activity and lowering mitochondrial calcium levels, thereby sustaining oxidative phosphorylation [27]. *CKB* deficiency not only causes ATP depletion, impaired oxidative phosphorylation, and compromised barrier integrity in epithelial cells [28] but also promotes epithelial–mesenchymal transition and migration through AKT activation [29]. Our GSEA results similarly showed positive enrichment of *CKB* in oxidative phosphorylation and negative enrichment in the cell cycle pathway, supporting the interpretation that *CKB* downregulation contributes to metabolic reprogramming and tissue damage in GERD by suppressing oxidative phosphorylation, enhancing glycolysis, and weakening barrier function.

CPA2 belongs to the zinc metalloprotease carboxypeptidase A family, which hydrolyses aromatic amino acids from the C-terminus of peptides and participates in protein digestion and metabolism [30]. *CPA2* was significantly downregulated in GERD, a finding confirmed in our animal model. KEGG enrichment showed that DEGs were significantly enriched in arginine and proline metabolism, suggesting that *CPA2* may participate in metabolic network remodelling in GERD through effects on amino acid metabolism. Direct evidence linking *CPA2* to metabolic reprogramming in GERD is currently scarce; nevertheless, both GSEA and GSVA consistently associated high *CPA2* expression with enhanced oxidative phosphorylation and suppressed cell cycle pathways, suggesting that *CPA2* may act in concert with *CKB* and *SH3GL2* to regulate oxidative metabolism. The underlying mechanisms warrant further investigation.

JUP encodes plakoglobin, a core structural protein of desmosomes and adherens junctions in the armadillo family. Through interactions with desmogleins and the

intermediate filament system, plakoglobin maintains epithelial cell–cell adhesion and tissue integrity [31]. *JUP* expression was elevated in the oesophageal mucosa of patients with GERD (AUC 0.831). Plakoglobin can enhance insulin signalling and glucose uptake [32], raising the possibility that it upregulates glycolytic metabolism in the oesophageal epithelium. GSEA showed that high *JUP* expression was negatively enriched in oxidative phosphorylation, further supporting a role in metabolic regulation. In the intestinal epithelium, loss of plakoglobin relieves suppression of the p38MAPK–NLRP1 axis, activating inflammasomes and potentiating IL-23/Th17 signalling [33]. These observations provide plausible mechanistic links to our findings.

LCN2 is an innate immune protein that is markedly upregulated in inflammatory conditions [34]. *LCN2* expression was elevated in both GERD patients and model rats (AUC 0.828). Previous studies have shown that *LCN2* can mediate inflammation and M1 macrophage polarisation through HIF-1α-induced glycolysis [35], promote aerobic glycolysis via JAK2/STAT3 signalling [36], and suppress pyruvate dehydrogenase activity through the MAPK/ERK pathway, thereby reducing mitochondrial ATP production [37]. In GERD, acidic bile salts can induce DRP1 Ser637 phosphorylation, driving mitochondrial fragmentation and mtROS-dependent NF-κB activation with increased secretion of cytokines such as IL-8 [38,39]. As a downstream effector, *LCN2* may thus couple metabolic reprogramming to inflammatory responses through these signalling axes, exacerbating oesophageal mucosal injury.

The immune infiltration analysis showed increased fractions of activated mast cells, neutrophils, and activated dendritic cells in GERD, accompanied by reduced fractions of $CD8^+$ T cells and plasma cells, which points to disturbed crosstalk among several

immune cell subsets. In the correlation analyses, *JUP* exhibited an inverse correlation with $CD8^+$ T cells, whereas *LCN2* was positively associated with neutrophils. *CKB* and *SH3GL2* were inversely associated with dendritic cells and naïve $CD4^+$ T cells but positively associated with $CD8^+$ T cells. Together, these results lend further support to the idea that such metabolic genes contribute to reshaping the GERD immune microenvironment through functional regulation of distinct immune cell subsets. Metabolic reprogramming is known to regulate the Th17/Treg balance and, in turn, influence inflammatory responses and mucosal injury [10,11]. Even though the Th17/Treg ratio was not assessed directly in this work, the metabolic gene–immune infiltration network identified here offers new molecular insights into this mechanism.

Alvimopan, a peripherally acting μ-opioid receptor antagonist developed for postoperative ileus [40], may ameliorate aberrant oesophagogastric motility in GERD. Baricitinib selectively inhibits JAK1/2, thereby attenuating pro-inflammatory cytokines including IL-6, IL-17A, and TNF-α via the JAK–STAT axis [41,42]. Evodiamine (isoevodiamine), an alkaloid from Evodia rutaecarpa, suppresses NF-κB activation and pro-inflammatory mediator release [43,44], potentially preserving the oesophagogastric mucosal barrier. Together, these agents offer complementary strategies for GERD through motility modulation and inflammation suppression.

This study has several strengths. First, by integrating differential expression analysis, WGCNA, and 107 machine-learning algorithms from the dual perspectives of metabolic reprogramming and immune infiltration, we systematically constructed a GERD diagnostic model centred on five genes and achieved high classification accuracy in the analysed cohorts. Model robustness was evaluated through a multi-dimensional strategy:

ANN cross-validation compensated for the limited ability of conventional linear models to capture gene–gene interactions, while the nomogram translated model outputs into individualised disease probabilities for potential clinical use. Second, AI-based virtual screening coupled with molecular docking identified candidate compounds targeting the core proteins, providing high-confidence candidate molecules for subsequent drug development and effectively reducing early-stage screening costs. On this basis, CIBERSORT further uncovered a regulatory network linking these genes to immune cell infiltration, offering novel insights into how metabolic reprogramming drives the remodelling of the GERD immune microenvironment and laying a theoretical foundation for subsequent targeted drug development and precision diagnostic strategies.

Several limitations should be acknowledged. First, the independent validation cohort (GSE13083) comprised only 14 samples (7 GERD, 7 controls), which limits robust assessment of model generalisability; the AUC of 1.000 observed in this cohort may partly reflect the small sample size. External validation in larger cohorts is required. Second, the analysis relied solely on public GEO data, and the combined training cohort may be subject to selection bias. Third, although the biomarkers were validated in an animal model, large-scale clinical and prospective studies are lacking; the nomogram requires external clinical validation to assess its discrimination and calibration. Fourth, the pathways implicated by bioinformatic analysis provide only preliminary indications of metabolic–immune crosstalk; the precise mechanisms of each gene and the therapeutic value of the predicted compounds require further in vitro and in vivo functional studies. Future work should include patient cohort studies and in vitro experiments to define the

specific roles of these genes in metabolic reprogramming and immune regulation, thereby laying a foundation for the precision diagnosis and treatment of GERD.

5. Conclusions

By integrating transcriptomics and machine learning, this study identified a five-gene diagnostic signature—*SH3GL2*, *JUP*, *CPA2*, *CKB*, and *LCN2*—linking metabolic reprogramming to oesophageal microenvironmental dysregulation in GERD. The diagnostic model and nomogram showed promising discriminative performance in the analysed cohorts, and a metabolic–immune network and candidate therapeutic compounds were identified. These findings provide a basis for molecular subtyping and precision intervention, pending validation in larger clinical cohorts.

Declarations

Ethics approval and consent to participate

All animal experiments were approved by the Animal Experiment Welfare and Ethics Committee of Guangxi University of Chinese Medicine (Approval No. DW20230320-05) and performed under its supervision. This study is reported in accordance with ARRIVE Guidelines. This study did not involve human participants, so consent to participate is not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Funding

This work was supported by the National Natural Science Foundation of China (Grant Nos. 82560908 and 82360910), the Bagui Young Top Talents Training Programme (Guirencaiban〔2025〕31), the Guangxi Natural Science Foundation (Grant No. 2026GXNSFAA00641081), and Wang Wenxiao Graduate Research Project (Grant No. YCSY2026067).

Authors' contributions

Jinjing Tan: Funding acquisition. Wenxiao Wang: Writing – original draft. Minjian Xie: Model construction and testing. Xiaoyun Lei: Model construction. Zezhong Teng: Model construction. Wenyu Wu: Model construction. Xiaomei Tang: Model construction. Sheng Xie: Funding acquisition. Liqun Li: Study conception, funding acquisition, Writing – review & editing.

[†]Jinjing Tan, Wenxiao Wang and Minjian Xie contributed equally to this work and share first authorship.

[*]Sheng Xie and Liqun Li contributed equally to this work and share corresponding authorship.

All authors read and approved the final manuscript.

Availability of data and materials

The transcriptomic datasets analysed during the current study are available in the Gene Expression Omnibus (GEO,http://www.ncbi.nlm.nih.gov/geo) repository under accession

numbers GSE1420, GSE39491 and GSE13083. Metabolic reprogramming-related genes were retrieved from the GeneCards database (https://www.genecards.org/).All additional data supporting the findings of this study are included within the article and its additional files.

Acknowledgments

We thank all the researchers who provided data on gastroesophageal reflux disease and metabolic reprogramming.

References

1. Katz PO, Dunbar KB, Schnoll-Sussman FH, Greer KB, Yadlapati R, Spechler SJ. ACG clinical guideline for the diagnosis and management of gastroesophageal reflux disease. Am J Gastroenterol 2022;117:27–56.
2. Nirwan JS, Hasan SS, Babar ZU, Conway BR, Ghori MU. Global prevalence and risk factors of gastro-oesophageal reflux disease (GORD): systematic review with meta-analysis. Sci Rep 2020;10:5814.
3. Zhang D, Liu S, Li Z, Wang R. Global, regional and national burden of gastroesophageal reflux disease, 1990–2019: update from the GBD 2019 study. Ann Med 2022;54:1372–84.
4. Dhar A, Maw F, Dallal HJ, Attwood S. Side effects of drug treatments for gastro-oesophageal reflux disease: current controversies. Frontline Gastroenterol. 2022;13(1):45–9.

5. Tan J, Li L, Huang X, Yang C, Liang X, Zhao Y, et al. Associations between gastro-oesophageal reflux disease and a range of diseases: an umbrella review of systematic reviews and meta-analyses. BMJ Open 2020;10:e038450.
6. Huai J, Chang E, Zhu Y, Ma W, Lv T, Sun J, et al. Diabetic encephalopathy: metabolic reprogramming as a potential driver of accelerated brain aging and cognitive decline. Front Cell Dev Biol 2025;13:1701406.
7. Suchorolski MT, Paulson TG, Sanchez CA, Hockenbery D, Reid BJ. Warburg and Crabtree effects in premalignant Barrett's esophagus cell lines with active mitochondria. PLoS One 2013;8:e56884.
8. Crispino P, Ciarambino T, Giordano M. Variation in fatty acid synthase, Ki67 and p53 esophageal mucosa expressions in Barrett's esophagus patients treated for one year with two esomeprazole regimens. Curr Issues Mol Biol 2023;45:4701–15.
9. Phelan JJ, Feighery R, Eldin OS, Ó Meachair S, Cannon A, Byrne R, MacCarthy F, O'Toole D, Reynolds JV, O'Sullivan J. Examining the connectivity between different cellular processes in the Barrett tissue microenvironment. Cancer Lett. 2016;371(2):334–46.
10. Liu J, Luo Y, Wang J, Xi C, Chen Y, Yang G, et al. Key molecules involved in the Th17/Treg balance are associated with the pathogenesis of reflux esophagitis and Barrett's esophagus. Esophagus 2021;18:388–97.
11. Hu Y, Zhao Q, Dai H, Wu Y, Tang X, Zhang N, et al. Metabolic reprogramming as a therapeutic target for modulating the Th17/Treg balance in autoimmune diseases: a comprehensive review. Front Immunol 2025;16:1687755.

12. Zhang T, Wong G. Gene expression data analysis using Hellinger correlation in weighted gene co-expression networks (WGCNA). Comput Struct Biotechnol J 2022;20:3851–63.
13. Wang J, Zhang Z, Wang Y. Utilizing feature selection techniques for AI-driven tumor subtype classification: enhancing precision in cancer diagnostics. Biomolecules 2025;15:81.
14. Yu Z, Wang K, Wan Z, Xie S, Lv Z. Popular deep learning algorithms for disease prediction: a review. Cluster Comput 2023;26:1231–51.
15. Shang M, Li Z, Xu G, Lian D, Liao Z, Wang D, et al. A predictive nomogram for the occurrence of gastroesophageal reflux disease after sleeve gastrectomy. Diabetes Metab Syndr Obes 2024;17:4135–47.
16. Suay-García B, Falcó A. Topological deep learning for drug-target interaction, virtual screening, and docking scoring: a practical, benchmark-driven review. Brief Bioinform 2026;27:bbag370.
17. Newman AM, Liu CL, Green MR, Gentles AJ, Feng W, Xu Y, et al. Robust enumeration of cell subsets from tissue expression profiles. Nat Methods 2015;12:453–7.
18. Xie S, Han BJ, Wang JX, Yan J, Liu Z, Wang YQ, et al. Mechanism of Back-Shu finger-pressure therapy in GERD rats via NLRP3/caspase-1 pathway-mediated esophageal ICC pyroptosis. Liaoning Zhong Yi Za Zhi. 2025:1–13 (in Chinese). https://link.cnki.net/urlid/21.1128.R.20251121.1639.047. Accessed 4 Sep 2026.

19. Xun J, Gao Q, Yang H, Zhang J, Wang Y, Liu B, et al. Esophageal epithelial cell-derived kynurenine drives Th17 inflammation in reflux esophagitis and is targeted by Xuanfu Daizhe decoction. Chin Med 2026;21:192.
20. Xu S, Rigaux E, Hène D, Renard HF, Thines L. Bending the boundaries: the many facets of endophilin-As from membrane dynamics to disease. Cell Mol Life Sci. 2025;82(1):339.
21. Shang C, Guo Y, Fu S, Fu W, Sun K. SH3GL2 gene participates in MEK–ERK signal pathway partly by regulating EGFR in the laryngeal carcinoma cell line Hep2. Med Sci Monit 2010;16:BR168-73.
22. Dvorak K, Chavarria M, Payne CM, Ramsey L, Crowley-Weber C, Dvorakova B, et al. Activation of the interleukin-6/STAT3 antiapoptotic pathway in esophageal cells by bile acids and low pH: relevance to Barrett's esophagus. Clin Cancer Res 2007;13:5305-13.
23. Zhu Y, Zhang X, Wang L, Ji Z, Xie M, Zhou X, et al. Loss of SH3GL2 promotes migration and invasion of glioblastoma cells through activating STAT3/MMP2 signalling. J Cell Mol Med 2017;21:2685-94.
24. Kannan A, Wells RB, Sivakumar S, Komatsu S, Singh KP, Samten B, et al. Mitochondrial reprogramming regulates breast cancer progression. Clin Cancer Res 2016;22:3348-60.
25. Wallimann T, Wyss M, Brdiczka D, Nicolay K, Eppenberger HM. Intracellular compartmentation, structure and function of creatine kinase isoenzymes in tissues with high and fluctuating energy demands: the 'phosphocreatine circuit' for cellular energy homeostasis. Biochem J 1992;281:21-40.

26. Renzi G, Vlassakev I, Hansen M, Higos R, Lecoutre S, Elmastas M, et al. Epigenetic suppression of creatine kinase B in adipocytes links endoplasmic reticulum stress to obesity-associated inflammation. Mol Metab 2025;92:102082.
27. He L, Lin J, Lu S, Li H, Chen J, Wu X, et al. CKB promotes mitochondrial ATP production by suppressing permeability transition pore. Adv Sci 2024;11:e2403093.
28. Bhagavatula G, Worledge CS, Schaepe C, Murphy EM, Neuhart RM, Lee JS, et al. Phosphocreatine rescues intestinal epithelial metabolic dysfunction related to creatine kinase loss and is protective in murine colitis. Cell Mol Gastroenterol Hepatol 2025;19:101557.
29. Wang Z, Hulsurkar M, Zhuo L, Xu J, Yang H, Naderinezhad S, et al. CKB inhibits epithelial–mesenchymal transition and prostate cancer progression by sequestering and inhibiting AKT activation. Neoplasia 2021;23:1147-65.
30. Reverter D, García-Sáez I, Catasús L, Vendrell J, Coll M, Avilés FX. Characterisation and preliminary X-ray diffraction analysis of human pancreatic procarboxypeptidase A2. FEBS Lett 1997;420:7-10.
31. Aktary Z, Alaee M, Pasdar M. Beyond cell-cell adhesion: plakoglobin and the regulation of tumorigenesis and metastasis. Oncotarget 2017;8:32270-9.
32. Negoita F, Vavakova M, Säll J, Laurencikiene J, Göransson O. JUP/plakoglobin is regulated by salt-inducible kinase 2 and is required for insulin-induced signalling and glucose uptake in adipocytes. Cell Signal 2020;76:109786.

33. Kelm M, Burkard N, Hörner M, Kollmann C, Otto C, Saravi B, et al. Junctional epithelial plakoglobin facilitates intestinal inflammation by p38MAPK-dependent activation of the inflammasome. Mucosal Immunol 2025;18:1257-70.
34. Asaf S, Maqsood F, Jalil J, Sarfraz Z, Sarfraz A, Mustafa S, et al. Lipocalin 2—not only a biomarker: a study of current literature and systematic findings of ongoing clinical trials. Immunol Res 2023;71:287-313.
35. Yang Y, Yan F, Shi P, Yang L, Cui D. HIF-1α pathway orchestration by LCN2: a key player in hypoxia-mediated colitis exacerbation. Inflammation 2024;47:1491-519.
36. Han B, An Z, Gong T, Pu Y, Liu K. LCN2 promotes proliferation and glycolysis by activating the JAK2/STAT3 signalling pathway in hepatocellular carcinoma. Appl Biochem Biotechnol 2024;196:717-28.
37. Tsai C, Lee C, Wu JCC. Astrocyte-secreted lipocalin-2 elicits bioenergetic failure-induced neuronal death in a mouse model of hepatic encephalopathy. Neurochem Int 2024;178:105800.
38. Isomoto H, Saenko VA, Kanazawa Y, Nishi Y, Ohtsuru A, Inoue K, et al. Enhanced expression of interleukin-8 and activation of nuclear factor-κB in endoscopy-negative gastroesophageal reflux disease. Am J Gastroenterol 2004;99:589-97.
39. Yang Y, Zhu Z, Li X, Li Q, Zhu H. Targeting DRP1 Ser637 phosphorylation alleviates acidic bile salt-induced NF-κB activation via mitochondrial protection. J Gastroenterol Hepatol 2025;40:2507-17.

40. Viscusi ER, Goldstein S, Witkowski T, Andonakakis A, Jan R, Gabriel K, et al. Alvimopan, a peripherally acting mu-opioid receptor antagonist, compared with placebo in postoperative ileus after major abdominal surgery: results of a randomized, double-blind, controlled study. Surg Endosc 2006;20:64-70.
41. Fridman JS, Scherle PA, Collins R, Burn TC, Li Y, Li J, et al. Selective inhibition of JAK1 and JAK2 is efficacious in rodent models of arthritis: preclinical characterization of INCB028050. J Immunol 2010;184:5298-307.
42. Deepak HB, Prince SE, Deshpande P. Effect of baricitinib in regulating programmed death 1 and ligand programmed cell death ligand 1 through JAK/STAT pathway in psoriasis. Indian J Pharmacol 2022;54:183-93.
43. Fan X, Zhu J, Sun Y, Luo L, Yan J, Yang X, et al. Evodiamine inhibits zymosan-induced inflammation in vitro and in vivo: inactivation of NF-κB by inhibiting IκBα phosphorylation. Inflammation 2017;40:1012-27.
44. Zhang Y, Zhang Y, Zhao Y, Wu W, Meng W, Zhou Y, et al. Protection against ulcerative colitis and colorectal cancer by evodiamine via anti-inflammatory effects. Mol Med Rep 2022;25:188.

Figure S1: Results of batch effect correction for the GERD dataset. (A) Boxplots of the GERD dataset before and after correction; (B) PCA analysis plots of the GERD dataset before and after correction.

Figure S2: Validation model for artificial neural network analysis. A–B: 5–5–2 topology ANN network architecture; C: Visualization of iterative weights and errors for the training set; D: Visualization of iterative weights and errors for the test set.

Supplementary Table S1. Differentially expressed genes in GERD.

Supplementary Table S2. All algorithms in machine learning.

Supplementary Table S3.Genes of GERD models with different algorithm combinations.

Supplementary Table S4.All algorithms in machine learning.

Supplementary Table S5. Intersection gene GO enrichment analysis.

Supplementary Table S6. Intersection gene KEGG enrichment analysis.

Supplementary Table S7.Results of GSEA analysis of hub genes.

Supplementary Table S8.Results of GSVA analysis of hub genes.

Supplementary Files

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

- FigureS1.tif
- supplementarytables.xlsx
- FigureS2.tif