Uncovering CD248, MMP28, and SLC16A10 in Sjögren's disease: a machine learning-driven SHAP approach for CD4+ T cell-associated biomarkers discovery

Q. Wang[1], L. He[2], Y. Han[3]

[1]Department of Rheumatology, Bengbu Hospital of Traditional Chinese Medicine, Anhui; [2]Department of Nephrology and Rheumatology, Shanghai Sixth People's Hospital, Shanghai; [3]Department of Medical Oncology, The First Affiliated Hospital of Bengbu Medical College, Anhui, China.

Abstract

Objective

Sjögren's disease (SjD) is a highly heterogeneous autoimmune disease with substantial challenges in early diagnosis and therapeutic intervention. We developed an integrated approach combining machine learning algorithms, SHAP interpretable modelling, molecular docking, and single-cell analysis to facilitate early diagnosis and treatment of SjD.

Methods

Transcriptomic data and 12 machine learning algorithms were employed to identify diagnostic signature genes. SHAP (Shapley Additive exPlanations) analysis further prioritised hub genes, followed by functional annotation using CIBERSORT, GSVA, and GSEA. Validation was performed using clinical cohorts, single-cell RNA sequencing (scRNA-seq), and molecular docking.

Results

The training cohort comprised 382 samples (61 healthy controls, 321 SjD patients) and 10,015 genes. Machine learning and SHAP analysis identified three hub genes (CD248, MMP28, SLC16A10), validated in external datasets with significant differential expression ($p<0.05$) and robust diagnostic performance (AUC >0.7). Immune infiltration analysis revealed positive correlations between CD248/SLC16A10 and naive CD4+ T cells ($p<0.05$), and between SLC16A10/MMP28 and memory resting CD4+ T cells ($p<0.05$). Single-cell profiling localised CD248 predominantly in naive CD4+ T cells, while SLC16A10 and MMP28 were expressed in both naive and memory CD4+ T cells subsets. Molecular docking demonstrated stable targeting of CD248, MMP28, and SLC16A10 by azathioprine, leflunomide, methotrexate, hydroxychloroquine, iguratimod, pilocarpine, and cevimeline.

Conclusion

Our bioinformatic study identifies CD248, MMP28 and SLC16A10 as candidate biomarkers and therapeutic targets for SjD, with their dysregulation specifically enriched in CD4+ T cell subsets, unveiling a previously underappreciated mechanism in SjD pathogenesis. naive and memory CD4+ T cells emerge as key contributors to inflammatory cascades, with azathioprine, leflunomide, methotrexate, hydroxychloroquine, iguratimod, pilocarpine, and cevimeline predicted to bind potently to these targets. This integrative multi-omics framework, combining machine learning, SHAP, and molecular docking, presents a promising approach for autoimmune disease diagnostics and early therapeutic intervention, although future experimental validation is essential to confirm its translational potential.

Key words

Sjögren's disease, machine learning, SHAP, multi-omics, molecular docking, biomarkers

Qiangqiang Wang, MD
Lingling He, MD
Yajuan Han, MD

Please address correspondence to:
Yajuan Han
The First Affiliated Hospital of Bengbu Medical University, no. 287 Changhuai Road, Longzihu District, Bengbu 233000, Anhui Province, China.
E-mail: 1120761537@qq.com

Received on June 6, 2025; accepted in revised form on October 8, 2025.

© Copyright CLINICAL AND EXPERIMENTAL RHEUMATOLOGY 2025.

Funding: this work was supported by the Key Project of Natural Science of Bengbu Medical College (2024byzd459).

Competing interests: none declared.

Introduction

Sjögren's disease (SjD) is a chronic autoimmune disorder predominantly targeting exocrine glands, such as lacrimal and salivary glands, leading to symptoms including dry mouth, dry eyes, and extra-glandular manifestations such as arthritis, arthralgia, and fatigue (1, 2), with an estimated prevalence of 1–3.0% in the general population (3), ranking as the second most prevalent autoimmune rheumatic disease (4). Its insidious onset and elusive aetiology, coupled with poorly understood pathogenesis, pose significant challenges for early diagnosis. Current therapeutic strategies remain palliative, focusing on symptom alleviation and delaying disease progression via replacement therapies, while genetic susceptibility, viral triggers, and immune dysregulation are implicated as key drivers of pathogenesis (3, 5, 6). Although serological detection of Anti-Ro/SSA antibodies and labial salivary gland biopsies exhibit high diagnostic specificity (7), the former demonstrates limited sensitivity in early-stage SjD identification, whereas the latter is invasive, prone to observer bias, and associated with procedural risks (8-10). The absence of a diagnostic gold standard for SjD (11) highlights the critical need for identifying sensitive and specific biomarkers to enable early detection and the development of precision-targeted therapies. Meanwhile, the treatment of SjD remains a major challenge in modern rheumatology. Current strategies are largely limited to symptomatic local measures and systemic protocols based on organ assessment that repurpose medications from other rheumatic diseases. These approaches often provide limited efficacy and fail to alter the natural course of the disease, leading to widespread patient dissatisfaction (12). Therefore, this study aims to explore novel potential therapeutic targets to lay the foundation for developing therapies that can genuinely improve disease prognosis, thereby addressing this significant unmet clinical need.

Machine learning (ML), as a robust analytical framework integrating ensemble models to enhance predictive accuracy and stability, offers transformative potential for identifying pivotal molecular signatures of disease progression. This approach facilitates the development of precise diagnostic tools and personalised therapeutic regimens (13, 14), while also enabling the discovery of combinatorial biomarkers for SjD. Interpretable ML algorithms leveraging SHapley Additive exPlanations (SHAP) further bridge the gap between predictive performance and clinical translatability, providing clinicians with high-performance, transparent, and actionable decision-support systems (15).

In this study, we aimed to identify robust diagnostic biomarkers for SjD through an integrative bioinformatics approach combined with interpretable machine learning modelling. We performed differential gene expression analysis on publicly available transcriptomic datasets from peripheral blood of SjD patients and healthy controls. Subsequently, multiple machine learning algorithms were employed to prioritise feature genes, which were further validated via single-cell RNA sequencing (scRNA-seq) analysis to clarify their cell-type-specific expression and potential functional roles in immune dysregulation.

Our findings provide new insights into the molecular mechanisms underlying SjD by uncovering the specific role of these biomarkers in $CD4^+$ T cell immunobiology, highlighting their promising diagnostic and therapeutic potential. Furthermore, we developed an interpretable diagnostic model based on SHAP values to facilitate clinical translation and support future precision medicine initiatives in SjD and other autoimmune conditions.

Materials and methods

Data acquisition and processing

All datasets were retrieved from the Gene Expression Omnibus (GEO) database. RNA sequencing (RNA-seq) expression profiles were derived from four public datasets: GSE51092 (16) (32 healthy controls, 190 SjD patients), GSE66795 (17) (29 healthy controls, 131 SjD patients), GSE48378 (18) (16 healthy controls, 11 SjD patients), and

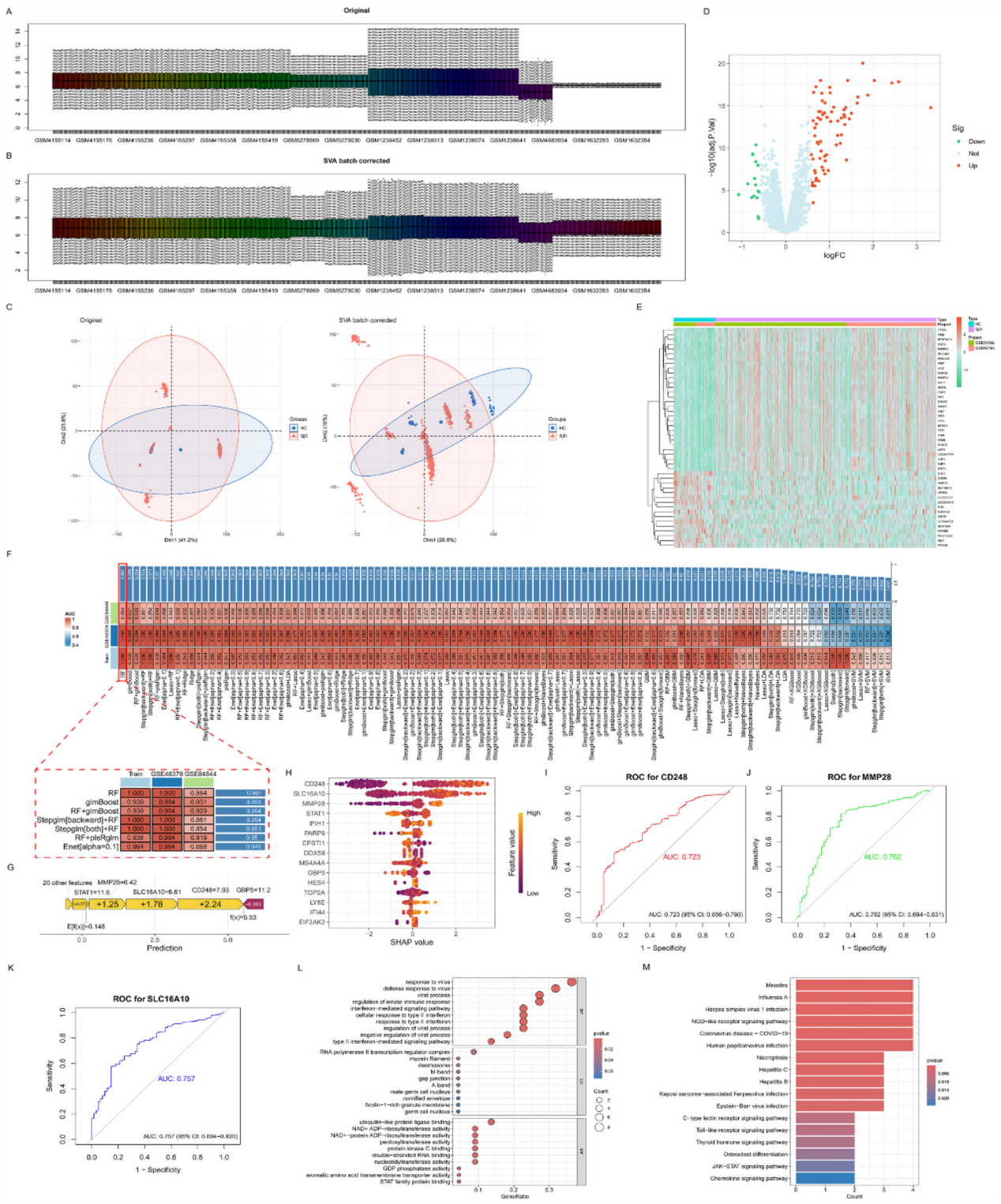

Fig. 1. Data integration, machine learning modelling of SjD hub genes, SHAP analysis, and functional enrichment.
A-B: Sample distribution before and after data integration, highlighting batch-effect correction.
C: Principal component analysis (PCA) plots comparing data variability before and after batch-effect removal.
D: Volcano plot of differential gene expression. Downregulated genes are shown in green; upregulated genes are indicated in red.
E: Heatmap showing expression patterns of the top 30 upregulated and 19 downregulated genes.
F: Heatmap of AUC values from 113 machine learning algorithms for SjD diagnosis in the training cohort. **G**: SHAP force plot visualising individual feature contributions to model predictions.
H: SHAP bee swarm plot summarising global feature importance.
I-K: Diagnostic performance (AUC) of CD248, MMP28, and SLC16A10 in the training set.
L: Gene Ontology (GO) enrichment analysis of hub genes.
M: KEGG pathway enrichment analysis.

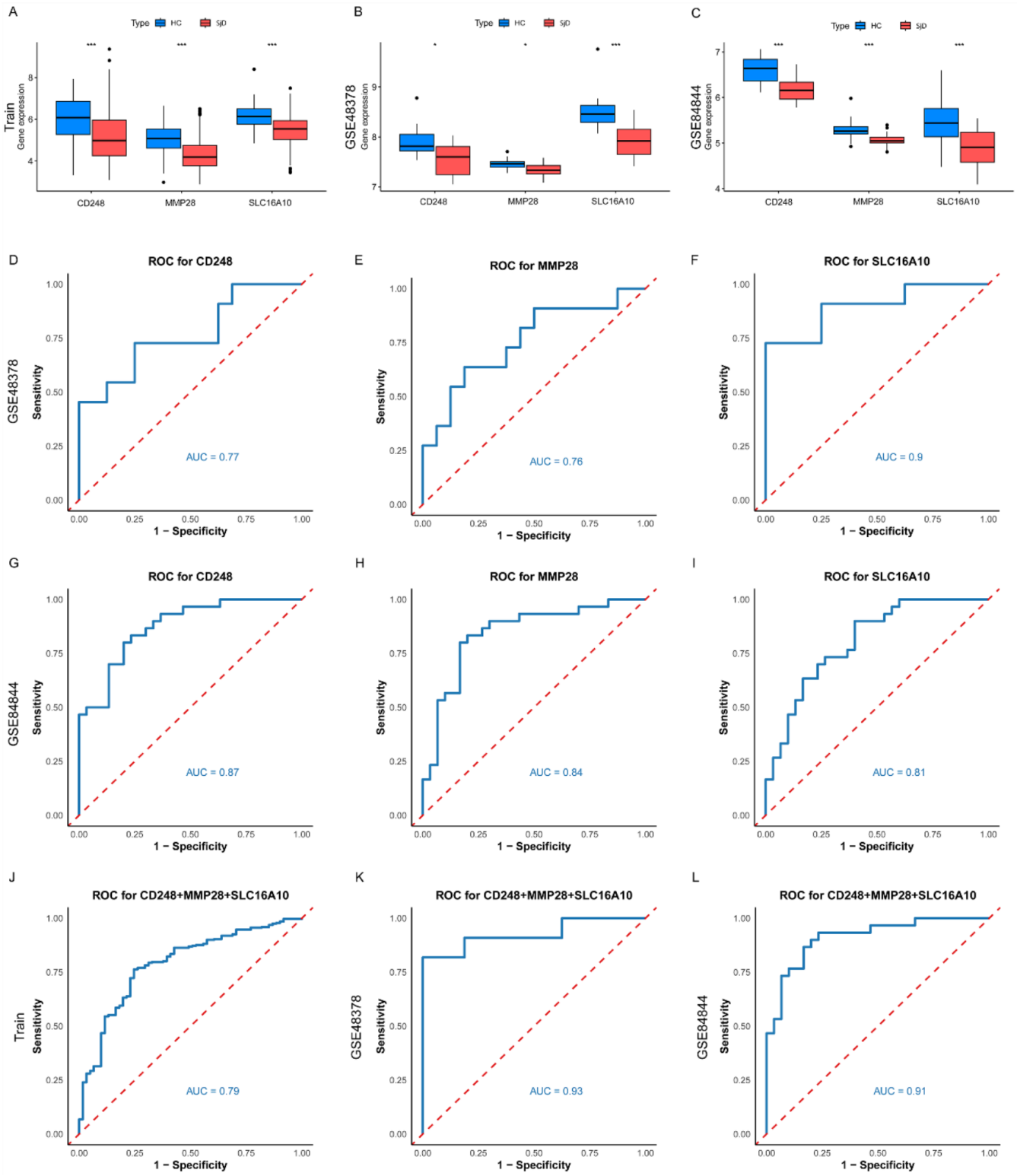

Fig. 2. Gene expression histograms and validation cohort diagnostic performance.
A-**C**: Histograms illustrating differential expression levels of the hub genes between SjD and HC groups ($^{*}p<0.05$, $^{**}p<0.01$, $^{**}p<0.001$). (D-F) AUC of the hub genes validation in the GSE48378 cohort.
G-**I**: AUC of the hub gene validation in the GSE84844 cohort.
J-**L**: AUC of the hub gene validation in the training, GSE48378, and GSE84844 cohort.

GSE84844 (19) (30 healthy controls, 30 SjD patients) (Supplementary Table S1). Single-cell RNA sequencing (scRNA-seq) data were obtained from GSE157278 (20) (5 healthy controls and 5 SjD patients). Probe IDs were converted to gene symbols using platform-specific annotation files. Differentially expressed genes (DEGs)

were identified in the training set (GSE51092, GSE66795), followed by feature gene screening using 12 machine learning (ML) algorithms (113 combinatorial models) and SHapley Additive exPlanations (SHAP)-based interpretable models, ultimately identifying three diagnostic genes. Training datasets were integrated using the ComBat function from the 'sva' package to correct batch effects, while GSE48378 and GSE84844 served as validation sets. As all data were publicly available, anonymised, and de-identified, ethical approval was waived.

Identification of differentially expressed genes (DEGs)

DEGs between SjD and Healthy controls (HC) groups were identified using the limma package (21) with thresholds set at $p<0.05$ and |log2 fold change (FC)| >0.585.

Machine learning algorithms

Twelve ML algorithms were employed for binary classification: Elastic Net (Enet), Ridge Regression, Stepwise Generalised Linear Model (Stepglm), Least Absolute Shrinkage and Selection Operator (LASSO), Support Vector Machine (SVM), Linear Discriminant Analysis (LDA), Generalised Linear Model Boosting (glmBoost), Partial Least Squares Regression Generalised Linear Model (plsRglm), Random Forest (RF), Gradient Boosting Machine (GBM), Extreme Gradient Boosting (XGBoost), and Naive Bayes. A total of 113 combinatorial models were generated to mitigate overfitting.

Model performance was evaluated using the area under the receiver operating characteristic curve (AUC), with optimal models selected to prioritise hub genes.

SHAP-based identification of hub genes

SHAP analysis provided both local and global interpretability, quantifying feature contributions to model predictions via Shapley values (22). This approach enhanced transparency in identifying genes critical to SjD progression (23). SHAP-filtered genes were designated as diagnostic biomarkers.

Validation of hub gene diagnostic potential

The diagnostic efficacy of hub genes was validated in two independent datasets GSE48378, GSE84844, with AUC values calculated to assess discrimination between SjD and HC. Finally, a multivariate logistic regression model was constructed using the hub genes to develop a diagnostic model, and the results were visualised.

Functional annotation and pathway enrichment

Hub genes were functionally annotated using clusterProfiler (24) for Gene Ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) analyses. Significantly enriched terms were defined by $p<0.05$ and adjusted $p<0.05$.

Immune infiltration analysis

Immune cell composition in tissues was quantified via CIBERSORT (25), with samples retained only if $p<0.05$. Correlations between hub genes and immune cell infiltration were assessed using Pearson/Spearman coefficients.

Gene set enrichment analysis (GSEA) and gene set variation analysis (GSVA)

GSEA evaluated pathway enrichment using KEGG gene sets from MSigDB (Human v2023.2.Hs) (26). GSVA (27) (v. 2.1.6) was performed to establish molecular signatures of SjD based on SHAP-derived hub genes.

Naive Bayes single-cell RNA sequencing analysis (scRNA-seq)

scRNA-seq data were processed using Seurat (v. 5), with quality control thresholds: 300 < nFeature_RNA <3000 and 3000 <nCount_RNA <10000. Expression matrices were normalised via LogNormalize. Batch effects were corrected using Harmony (28), while DecontX (29) removed ambient RNA and DoubletFinder (30) excluded doublets. Cell types were annotated via SingleR, CellMarker 2.0, and ScType (31-33), $CD4^+$ T cell subpopulations were analysed for ligand-receptor interactions using CellChat (34).

Molecular docking

3D conformations of leflunomide, azathioprine, and cyclophosphamide (ligands) were obtained from PubChem, while hub gene protein structures (receptors) were retrieved from UniProt (35) and ChemSpider (36). Molecular docking was performed using CB-Dock2 (37), visualised via Discovery Studio. Binding affinities ≤ -5 kcal/mol were considered significant (38).

Statistical analysis

All analyses were conducted in R (v. 4.4). Group differences were assessed via Student's t-test or Wilcoxon test, with correlations analysed using Pearson/Spearman methods. Two-tailed $p<0.05$ was deemed statistically significant ($^*p<0.05$, $^{**}p<0.01$, $^{***}p<0.001$).

Results

Data integration and differential analysis

The GSE51092 and GSE66795 datasets were combined to establish a training cohort. Following batch-effect correction (Fig. 1A-C), 85 differentially expressed genes (DEGs) were identified (Suppl. Table S2), including 69 upregulated genes and 16 downregulated genes in SjD compared to HC. Expression patterns of these DEGs were visualised in a heatmap and volcano plot (Fig. 1D-E).

Machine learning and SHAP analysis

The RF algorithm demonstrated superior predictive performance among the 113 tested models (Fig. 1F). Using the 26 genes selected by RF, the diagnostic model achieved AUC values of 1.000, 1.000 and 0.884 in the training set and validation cohorts (GSE48378, GSE84844), respectively (Suppl. Fig. S1, Table S3). The consistent performance across all cohorts indicated a lower risk of overfitting with the RF-based model. SHAP analysis further prioritised the top three hub genes-CD248, MMP28, and SLC16A10-based on their global feature importance (Fig. 1G-H). These genes exhibited robust diagnostic potential in the training set, with individual AUC values of 0.723, 0.762, and 0.757 (Fig. 1I-K). Validation in independent cohorts confirmed their reliability (AUC>0.7), solidifying their utility as biomarkers for SjD.

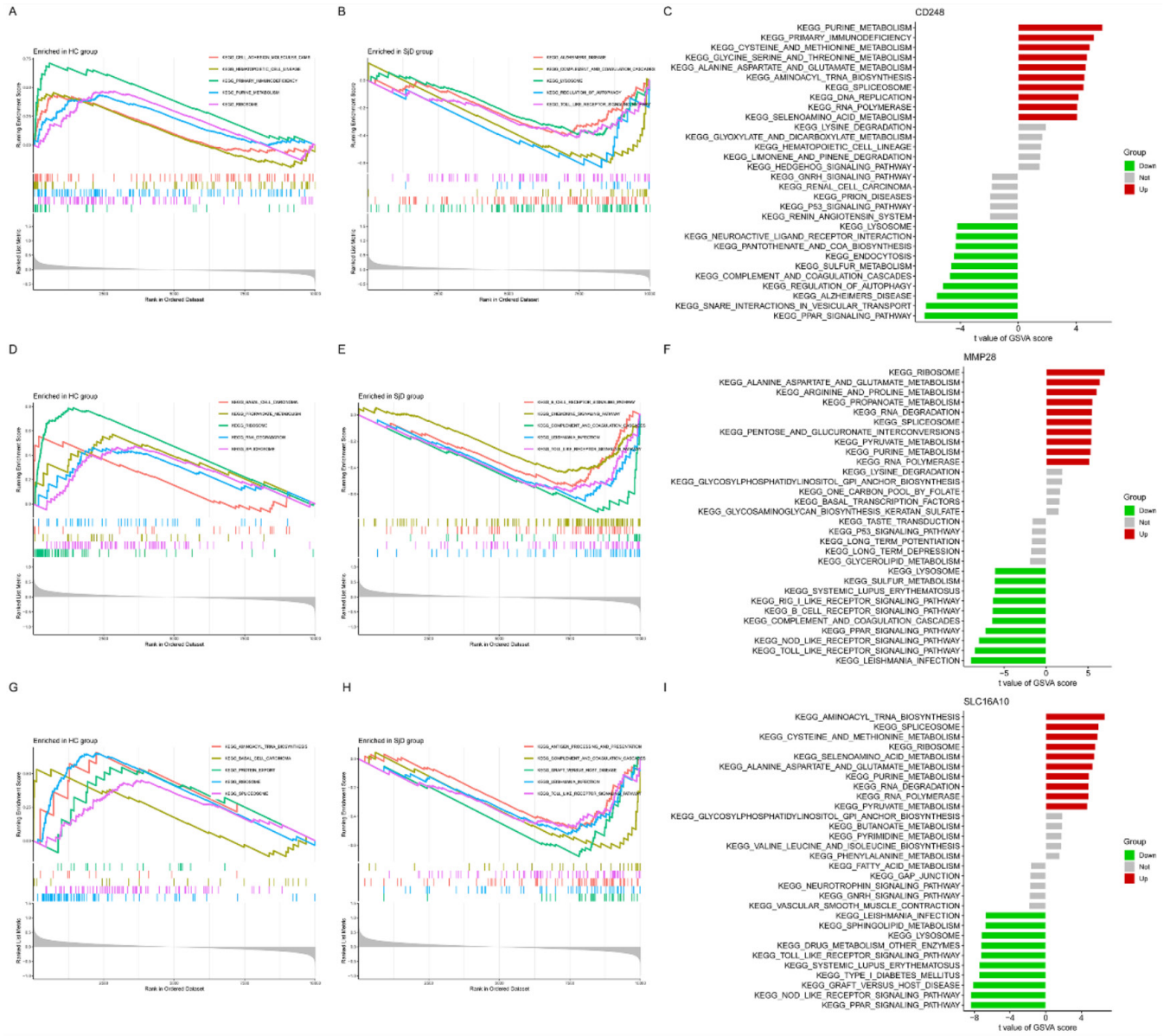

Fig. 3. GSEA and GSVA analysis of high and low expression of CD248, MMP28 and SLC16A10.
A-**B**: GSEA of CD248. **C**: GSVA of CD248. **D**-**E**: GSEA of MMP28. **F**: GSVA of MMP28. **H**-**I**: GSEA of SLC16A10. **J**: GSVA of SLC16A10.

GO and KEGG enrichment analysis

GO enrichment analysis revealed that DEGs were predominantly enriched in viral response pathways, including response to virus, defense response to virus, and viral processes (biological process). Cellular component terms highlighted RNA polymerase II transcription regulator complex, myosin filament, and desmosome, while molecular functions centred on ubiquitin-like protein ligase binding and NAD+-dependent ADP-ribosyltransferase activity (Fig. 1L, Suppl. Table S4). KEGG pathway analysis further identified hub genes enrichment in critical immune and inflammatory signalling pathways, notably the NOD-like receptor signalling pathway, Toll-like receptor signalling pathway, and JAK-STAT signalling pathway (Fig. 1M, Suppl. Table S5), underscoring their roles in SjD pathogenesis.

Diagnostic validation of hub genes

The diagnostic validity of the identified biomarkers was assessed in independent validation cohorts (GSE48378 and GSE84844) using AUC analysis. Genes with AUC>0.7 were considered robust for SjD diagnosis, demonstrating excellent specificity and sensitivity. Expression analysis revealed significant downregulation of CD248, MMP28, and SLC16A10 in SjD compared to HC (Fig. 2A-C). In the GSE48378 cohort, these genes achieved AUC values of 0.77, 0.76, and 0.90, respectively (Fig. 2D-F). Similarly, in the GSE84844 cohort, AUC values were 0.87, 0.84, and 0.81 (Fig. 2G-I). The multivariate regression analysis revealed that the combined score of CD248, MMP28, and SLC16A10 achieved AUC values of 0.79, 0.93, and 0.91 in the training set, GSE48378, and GSE84844, respectively, although future experimental validation is essential to confirm its translational potential. Collectively, CD248, MMP28, and SLC16A10 demonstrated robust diagnostic accuracy for both established and early-stage SjD, sug-

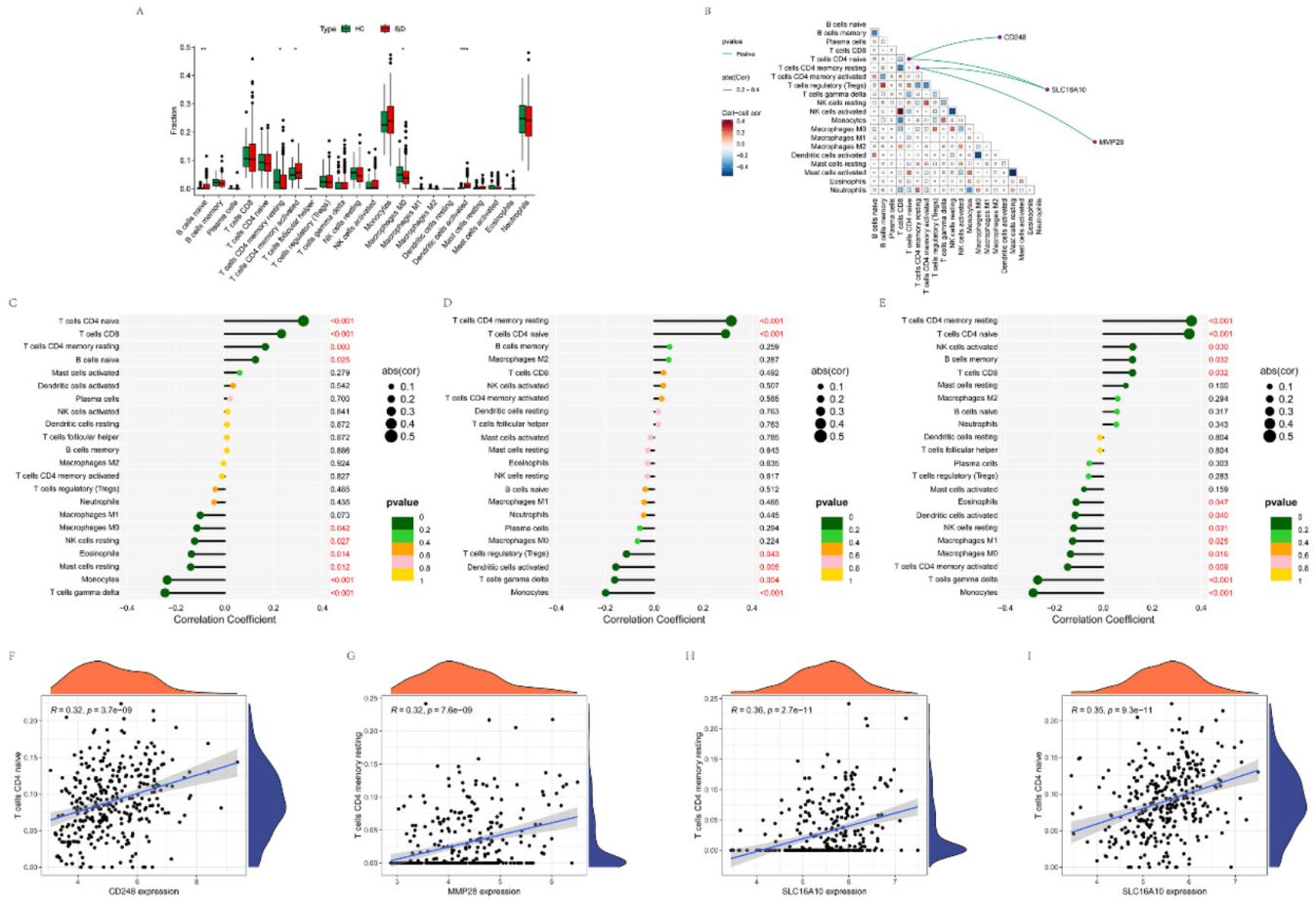

Fig. 4. Immune cell infiltration in SjD and HC.
A: Immune cell infiltration between SjD and HC. ($^{*}p<0.05$, $^{**}p<0.01$, $^{***}p<0.001$).
B: Heatmap of correlations among immune cells. red: positive correlation; blue: negative correlation.
C-**E**: Lollipop plots of correlations between hub genes and immune infiltrating cells (larger circles indicate stronger correlations; darker green indicates *p*-values closer to zero).
F-**I**: Scatter plots of correlations between hub genes and immune infiltrating cells.

gesting their potential clinical utility as biomarkers.

Hub gene-associated signalling mechanisms

We conducted an in-depth analysis of the specific signalling mechanisms associated with the three hub genes, CD248, MMP28, and SLC16A10, to explore their potential roles in disease progression and their influence on relevant signalling pathways. Through GSEA analysis, we found that high expression of CD248 was primarily enriched in signalling pathways such as cell adhesion molecules, hematopoietic cell lineage, and primary immunodeficiency, whereas low expression was mainly enriched in complement and coagulation cascade, lysosome, and Toll-like receptor (TLR) signalling pathways (Fig. 3A-B). For MMP28, high expression was primarily enriched in basal cell carcinoma, propanoate metabolism, and ribosome pathways, while low expression was mainly enriched in B cell receptor, chemokine, and complement and coagulation cascade signalling pathways (Fig. 3D-E). High expression of SLC16A10 was primarily enriched in aminoacyl-tRNA biosynthesis, basal cell carcinoma, and protein export pathways, whereas low expression was mainly enriched in antigen processing and presentation, complement and coagulation cascade, and graft-*versus*-host disease pathways (Fig. 3G-H). Additionally, GSVA analysis further revealed the enrichment characteristics of these hub genes (Fig. 3C, F, I). Specifically, low expression of CD248 was primarily enriched in PPAR signalling pathway and SNARE interactions in vesicular transport, while high expression was mainly enriched in purine metabolism, primary immunodeficiency, and cysteine and methionine metabolism pathways. Low expression of MMP28 was primarily enriched in NOD-like receptor signalling pathway and Toll-like receptor signalling pathway, while high expression was mainly enriched in ribosome, alanine, aspartate and glutamate metabolism, and arginine and proline metabolism pathways. Low expression of SLC16A10 was primarily enriched in PPAR signalling pathway and Toll-like receptor signalling pathway, while high expression was mainly enriched in aminoacyl-tRNA biosynthesis, spliceosome, and cysteine and methionine metabolism pathways.

Correlation of hub genes with immune cell infiltration in SjD

Immune infiltration analysis revealed

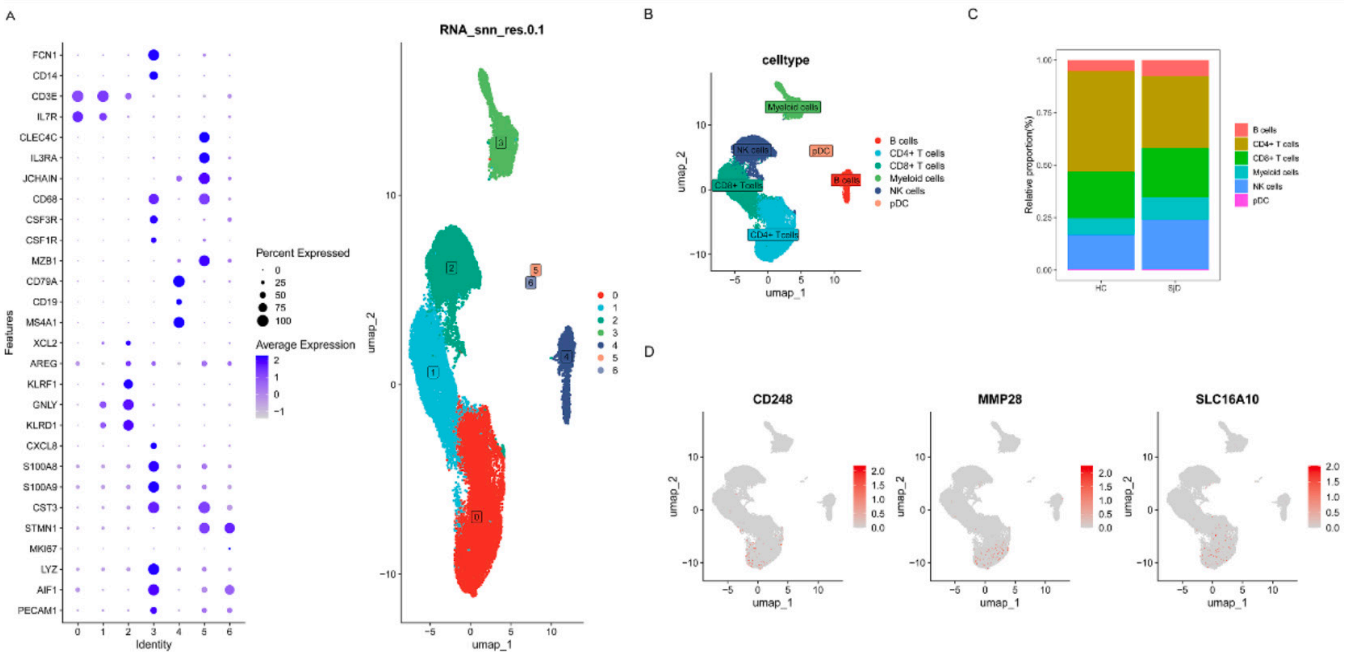

Fig. 5. Quality assurance for single cell isolation and sequencing.
A: Dot plot and UMAP showing marker genes and distribution of each cluster. **B**: UMAP plot of cell type annotation with different cell types colour-coded. **C**: Cell distribution in Normal and SjD groups. **D**: Expression of hub genes in single cells.

that naive B cells, resting memory CD4⁺ T cells, activated memory CD4⁺ T cells, M0 macrophages, and activated dendritic cells were the predominant infiltrating immune cell populations (Fig. 4A). Subsequently, correlations among 22 immune cell types in SjD were evaluated (Fig. 4B). For example, memory B cells exhibited a negative correlation with naive B cells. Activated memory CD4⁺ T cells showed negative correlations with memory B cells and regulatory T cells (Tregs), while M0 macrophages were negatively correlated with activated dendritic cells. We further analysed correlations between three hub genes (CD248, MMP28, and SLC16A10) and two significantly differential immune cell types (naive CD4⁺ T cells and resting memory CD4⁺ T cells) (Fig. 4B).

The results demonstrated that CD248 was positively correlated with naive CD4⁺ T cells (p=3.7e-09) (Fig. 4C and F), MMP28 showed a positive correlation with resting memory CD4⁺ T cells (p=7.6e-09) Fig. 4D and G); and SLC16A10 exhibited positive correlations with both naive CD4⁺ T cells (p=9.3e-11) and resting memory CD4⁺ T cells (p=2.7e-11) (Fig. 4E and H-I).

Single-cell profiling reveals hub genes expression in CD4⁺ T cell subsets and altered cell-cell communication

A total of 51,585 cells were retrieved from GSE157278 dataset, with 24,349 cells from the HC group and 27,236 from the SjD group. After stringent quality control procedures (Supplementary Fig. S2A), harmony integration, and exclusion of cells in the cell cycle (Suppl. Fig. S2B), the cell population was reduced to 41,293 (20,344 in HC and 20,949 in SjD). Following the removal of 2,942 doublets (Supplementary Fig. S2C), 38,351 high-quality cells remained, consisting of 17,402 in the HC group and 20,949 in the SjD group. With a resolution set at 0.1, these cells were partitioned into seven distinct cell clusters (Fig. 5A). Classification based on specific marker genes identified six major cell types, including B cells, CD4⁺ T cells, CD8⁺ T cells, myeloid cells, NK cells, and pDC (Fig. 5B). Notably, the cellular distribution patterns differed markedly between the HC and SjD groups (Fig. 5C). Using UMAP visualisation, we mapped the distribution of these hub genes across cell types, revealing that CD248, MMP28, and SLC16A10 were preferentially expressed in CD4⁺ T cells (Fig. 5D).

Consequently, we conducted a more in-depth analysis of CD4⁺ T cells. After additional quality control, 14,432 CD4⁺ T cells were retained. At a resolution of 1.2, these cells were clustered into 14 subgroups, which were manually annotated as 11 distinct CD4⁺ T cell subsets, including naive CD4⁺ T cells, memory CD4⁺ T cells, central memory CD4⁺ T cells, effector CD4⁺ T cells, follicular helper CD4⁺ T cells, Tem/Th1 CD4⁺ T cells, Treg CD4⁺ T cells, and four uncharacterised subsets (CD4⁺ T cells 1-4) (Fig. 6A-B) . The disparity in cellular composition between the HC and SjD groups was again evident (Fig. 6C). Visualisation of hub genes expression indicated that CD248 was predominantly expressed in naive CD4⁺ T cells and CD4⁺ T cells 1, while MMP28 and SLC16A10 were mainly expressed in memory CD4⁺ T cells and CD4⁺ T cells 1 (Fig. 6D). To elucidate the underlying mechanisms, we performed cell-cell communication analysis to quantify the interaction frequency and intensity between the HC and SjD groups. Compared with the HC group, the SjD group showed a decreased communication frequency but enhanced interaction

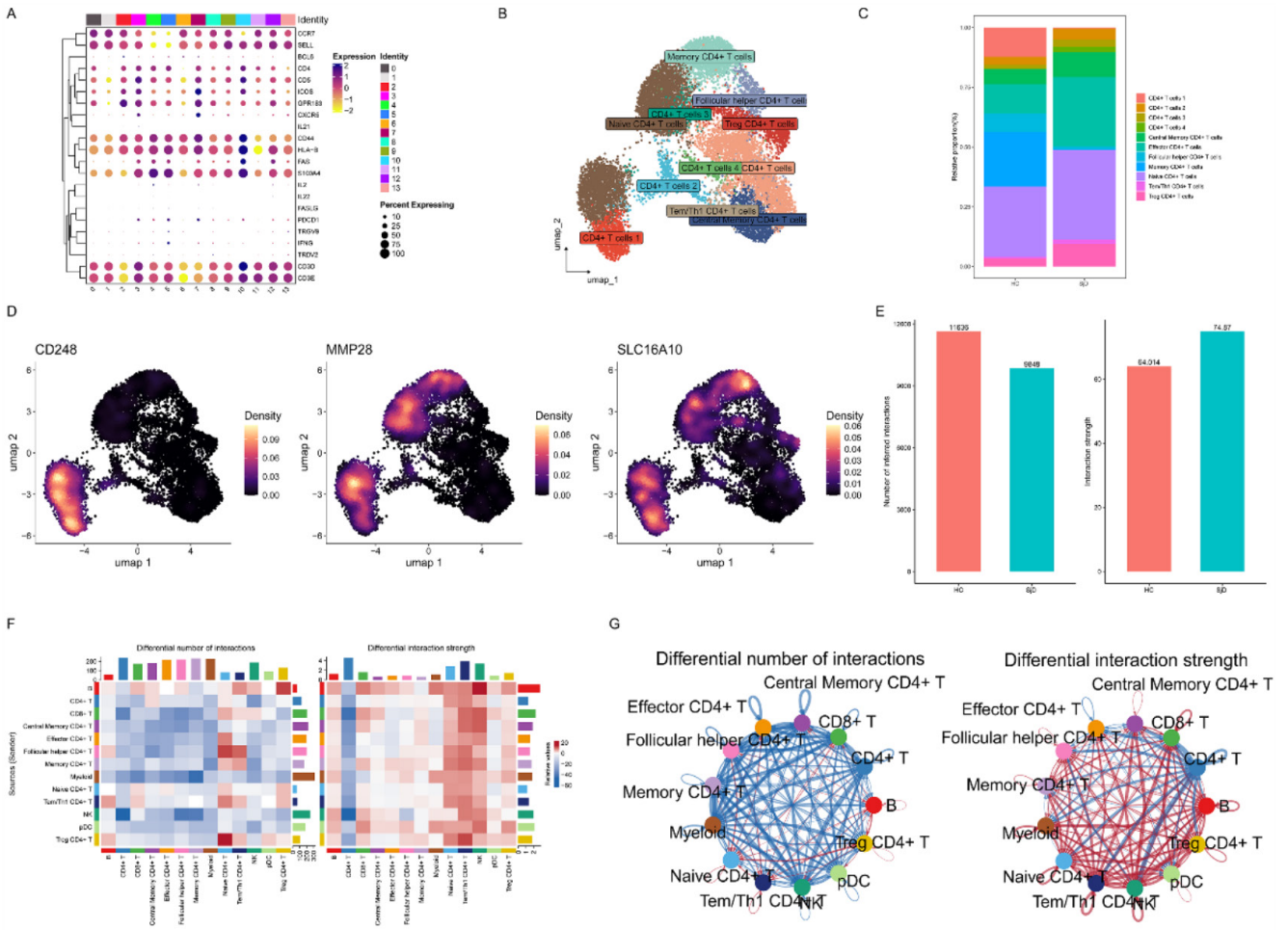

Fig. 6. CD4⁺ T cell subpopulation analysis, hub genes expression, and cell-cell communication.
A-B: Bubble plot and UMAP visualisation illustrating marker gene distribution across CD4⁺ T cell clusters.
C: UMAP projection of annotated cell types, colour-coded by distinct populations.
D: Hub genes (CD248, MMP28, SLC16A10) expression levels across CD4⁺ T cell subsets.
E: Bar plot quantifying interaction frequency and strength in HC *vs*. SjD groups.
F: Heatmap depicting cell-cell communication networks.
G: Cell-cell communication network (circle plot). Each colour represents a distinct cell type, arrows indicate directionality, and line thickness corresponds to interaction number/strength.

strength (Fig. 6E). Overall, cells in the SjD group engaged in more active communication with each other, with all cell types except CD4⁺ T cells-showing varying degrees of increased signalling (Fig. 6F-G).

Molecular docking and visualisation

Ligands included drugs commonly used in SjD treatment: azathioprine, cyclophosphamide, leflunomide, methotrexate, hydroxychloroquine, iguratimod, pilocarpine, and cevimeline (39). These were docked against the three hub gene-encoded proteins serving as receptors: CD248, MMP28, and SLC16A10. The binding energy, calculated using the CB-Dock2 web tool and expressed as binding energy, was used to evaluate the interactions. A more negative binding energy indicates stronger binding affinity and higher complex stability. The docking results for all ligand-receptor pairs are shown in Figure 7A. The complexes of SLC16A10-methotrexate, CD248-iguratimod, and SLC16A10-iguratimod demonstrated the highest binding affinity, with binding energies <-8.0 kcal/mol (Fig. 7A, F, S, U). Furthermore, leflunomide, methotrexate, and iguratimod showed strong binding affinity (<-7.0 kcal/mol) with all three receptors: CD248, MMP28, and SLC16A10 (Fig. 7A, C, D, F, J, K, M, R, S, U). With the exception of cyclophosphamide (binding energy >-5.0 kcal/mol with CD248 and MMP28), all other ligand-receptor pairs exhibited binding energies <-5.0 kcal/mol (Fig. 7B-W). Notably, SLC16A10 demonstrated consistently high binding affinity (binding energy <-5.0 kcal/mol) across all drugs, highlighting its role as a central node in the drug-target interaction network (Fig. 7P-W). Visualisation of the binding modes confirmed stable interactions and plausible spatial orientations at the predicted binding sites for all compounds, supporting the reliability of the docking results.

Discussion

In this study, we analysed human blood transcriptomic datasets from the Gene Expression Omnibus (GEO) database to identify differentially expressed genes (DEGs). Feature genes were ex-

tracted from DEGs using 12 distinct ML algorithms, followed by systematic interrogation of their biological functions, molecular mechanisms, and immunological relevance in SjD pathogenesis. An interpretable SHAP-based diagnostic model was developed to quantify the contributions of hub genes to SjD initiation and progression, validated further through single-cell RNA sequencing (scRNA-seq) data, thereby establishing a decision-making framework for precision therapeutics.

Our integrative methodology combined DEG analysis with a multi-algorithm ML pipeline to screen feature genes and prioritise biologically significant hub genes. The interpretable diagnostic model not only enhanced diagnostic accuracy but also elucidated mechanistic insights into hub gene-driven disease dynamics, advancing early detection capabilities. External validation across independent cohorts and scRNA-seq confirmation supported their potential as robust SjD biomarkers. Finally, molecular docking simulations with clinically approved drugs predicted potential therapeutic targets among hub genes, suggesting a potential roadmap for developing future precision-targeted interventions. This study pioneers the convergence of multi-omics data, interpretable ML, and computational drug discovery to address unmet clinical needs in SjD. The framework establishes a paradigm for biomarker identification, diagnostic innovation, and therapeutic development in complex autoimmune disorders.

The aetiology and pathogenesis of SjD remain complex, and traditional diagnostic approaches are limited by invasiveness, procedural risks, and heterogeneity in classification criteria across geographic regions. Epidemiological studies of SjD are scarce, with over 50% of patients remaining undiagnosed due to inconsistent diagnostic frameworks (40). This underscores the urgent need for reliable, non-invasive biomarkers to improve diagnostic accuracy. In this study, we employed an integrative pipeline combining 12 machine learning (ML) algorithms (113 combinatorial models) to identify CD248, MMP28, and SLC16A10 as novel diagnostic biomarkers for SjD, validated through SHapley Additive exPlanations analysis. Expression patterns of these hub genes in independent validation cohorts (GSE48378, GSE84844) aligned with training set trends, confirming the robustness of our screening strategy. While these genes have been documented in other inflammatory contexts (41-43), our study is the first to implicate CD248, MMP28, and SLC16A10 in SjD pathogenesis through a specific association with $CD4^+$ T cell dysfunction, opening new avenues for mechanistic exploration.

Functional enrichment analyses revealed that SjD-associated pathways are closely tied to viral lifecycle regulation, with KEGG highlighting critical roles for the NOD-like receptor signalling pathway, Epstein-Barr virus (EBV) infection, Toll-like receptor (TLR) signalling, JAK-STAT signalling and hepatitis B virus (HBV) interactions. The NLRP3 inflammasome, a key mediator of innate immunity, is upregulated in SjD patients' peripheral blood mononuclear cells (PBMCs) and ocular tissues, exacerbating disease progression (44). EBV, a known trigger of epithelial damage and immune dysregulation, may initiate SjD by activating both innate and adaptive immune responses (45). Similarly, HBV infection correlates with elevated B-cell activating factor (BAFF) levels, which drive autoantibody production and interferon responses-hallmarks of SjD severity (46). While TLR ligands in SjD remain unidentified, aberrant activation by damage-associated molecular patterns (DAMPs) likely contributes to chronic inflammation (47). The JAK-STAT pathway further amplifies disease via STAT1/3/5-mediated B-cell activation, linking genetic susceptibility, viral triggers, and immune dysregulation (48, 49).

Immune infiltration analysis demonstrated strong associations between hub genes and naive $CD4^+$ T cell /memory $CD4^+$ T cell subsets. In SjD, naive $CD4^+$ T cells exhibit telomere shortening, reduced IL-7R expression, and senescence-associated β-galactosidase (SA-β-Gal) accumulation, reflecting thymic insufficiency and impaired lymphopoiesis (50). Our findings align with these observations, showing hub gene enrichment in $CD4^+$ T cell subpopulations. However, the precise mechanisms by which these genes modulate immune homeostasis warrant further investigation.

Our molecular docking analysis, which included a wider range of small-molecule drugs used clinically in SjD, revealed favourable binding affinities for most compounds (except Cyclophosphamide) with the hub targets CD248, MMP28, and SLC16A10. The strong in silico binding of azathioprine, leflunomide, methotrexate, hydroxychloroquine, iguratimod, pilocarpine, and cevimeline suggests a potential structural basis for their established clinical efficacy, which may involve modulation of these targets. These results provide a preliminary computational rationale for repurposing existing drugs to mitigate SjD progression, which merits further investigation in experimental settings.

However, a key finding of our study is that these three hub genes are downregulated at the transcript level in SjD, which raises a valid question regarding the therapeutic rationale of targeting underexpressed proteins. We hypothesise that the binding of these drugs may not follow a conventional inhibitory mechanism. Instead, it could lead to stabilisation of protein structure, allosteric modulation of residual activity, interference with degradation pathways, or indirect feedback upregulation of gene expression. Thus, the molecular docking presented here should not be interpreted as evidence for inhibition, but rather as a computational assessment of binding potential, providing a structural hypothesis for these complex and non-canonical mechanisms of action. This hypothesis requires rigorous functional validation in future studies.

In conclusion, by leveraging a multi-algorithm ML framework, we have not only identified robust biomarkers for SjD but also delineated their specific involvement in $CD4^+$ T cell subsets, offering novel insights into its immunopathology and therapeutic targeting. While this study advances precision diagnostics and drug discovery, clinical translation requires rigorous validation

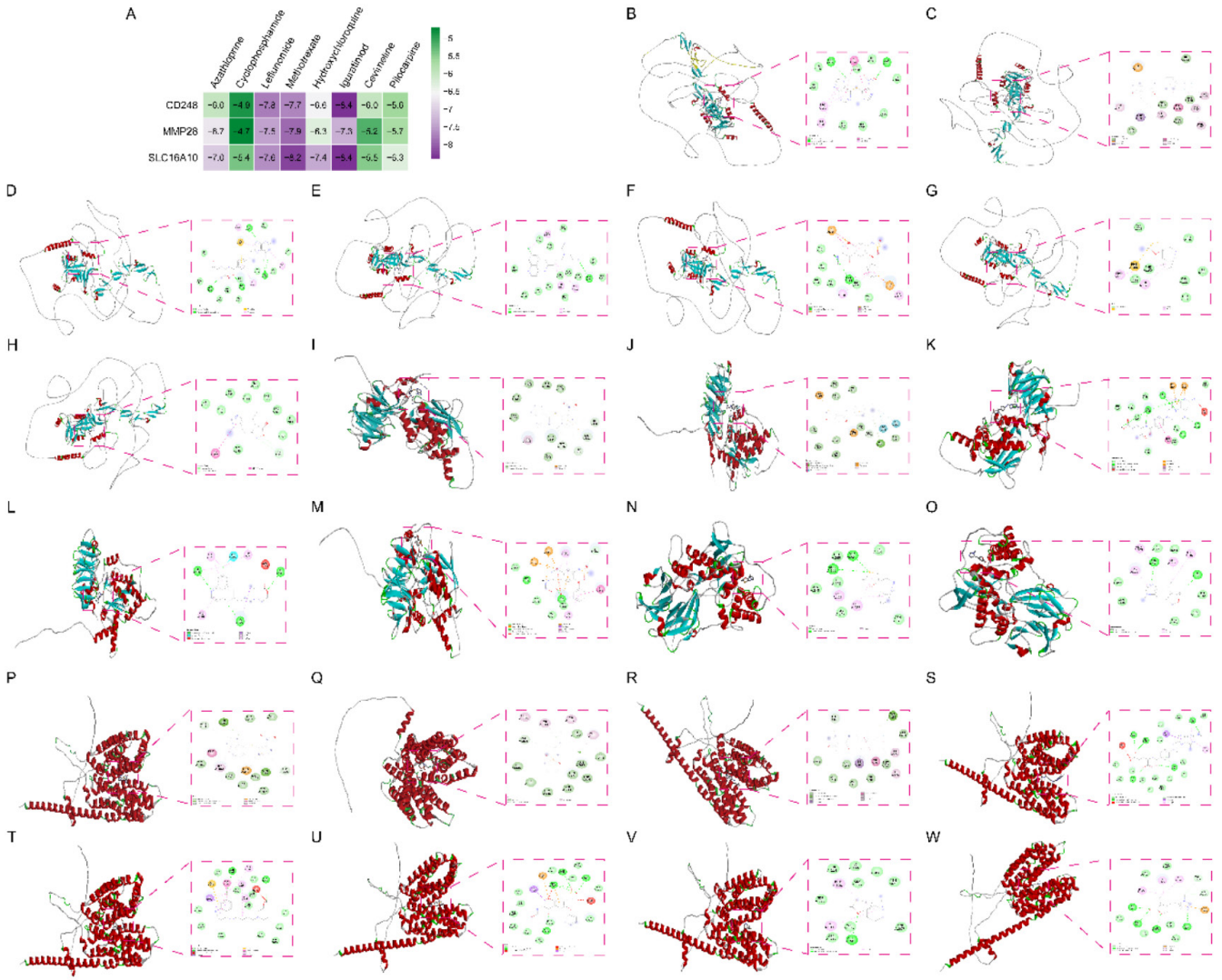

Fig. 7. Molecular docking and visualisation.
A: Binding affinity heatmap of molecular docking interactions. **B**: CD248-azathioprine. **C**: CD248-leflunomide. **D**: CD248-Methotrexate. **E**: CD248-Hydroxychloroquine. **F**: CD248-Iguratimod. **G**: CD248-Cevimeline. **H**: CD248-pilocarpine. **I**: MMP28-azathioprine. **J**: MMP28-leflunomide. **K**: MMP28-methotrexate. **L**: MMP28-hydroxychloroquine. **M**: MMP28-iguratimod. **N**: MMP28-cevimeline. **O**: MMP28-pilocarpine. **P**: SLC16A10-azathioprine. **Q**: SLC16A10-cyclophosphamide. **R**: SLC16A10-leflunomide. **S**: SLC16A10-methotrexate. **T**: SLC16A10-hydroxychloroquine. **U**: SLC16A10-iguratimod. **V**: SLC16A10-cevimeline. **W**: SLC16A10-pilocarpine.

through large-scale trials to confirm biomarker reliability and therapeutic efficacy. This integrative approach exemplifies the power of AI-driven biomarker discovery in addressing unmet needs in autoimmune diseases.

Study limitations

This study has several limitations that should be acknowledged. First, our analysis is based solely on publicly available genomic datasets. While we employed robust batch-effect correction methods and validated findings in independent cohorts, the inherent heterogeneity in sample sources, sequencing platforms, and demographic backgrounds across these datasets may introduce residual confounding effects and limit the generalisability of our results. Second, and most importantly, our conclusions are derived from in silico bioinformatic analyses and computational predictions. The absence of experimental validation, such as qPCR for gene expression, western blot for protein levels, or functional assays in cellular or animal models, means that the diagnostic and therapeutic potential of the identified hub genes (CD248, MMP28, SLC16A10) remains hypothetical at this stage. Finally, the mechanistic insights into immune dysregulation and drug-target interactions, while supported by statistical correlations and docking simulations, require direct biological experimentation to establish causality. Future studies with new patient cohorts and laboratory validation are essential to translate these findings into clinical practice.

Conclusion

In summary, our bioinformatic study proposes CD248, MMP28, and SLC16A10 as candidate diagnostic biomarkers for SjD, suggesting their potential association with the disease. Mechanistically, these genes likely contribute to SjD pathogenesis by modulating immune dysregulation, particu-

larly through functional alterations in memory CD4+ T cells and naive CD4+ T cell subpopulations. This regulatory interplay highlights their pivotal role in disrupting immune homeostasis, a hallmark of SjD progression. Collectively, our findings not only advance the understanding of SjD immunopathology but also hypothesise actionable targets for developing precision diagnostics and immunomodulatory therapies, although these hypotheses require further experimental confirmation.

Acknowledgements

We thank Dr Jianming Zeng (University of Macau), and all the members of his bioinformatics team, biotrainee, for generously sharing their experience and codes.

References

1. TRUTSCHEL D, BOST P, MARIETTE X *et al*.: Variability of primary Sjögren's syndrome is driven by interferon-alpha and interferon-alpha blood levels are associated with the class II HLA-DQ locus. *Arthritis Rheumatol* 2022; 74: 1991-2002. https://doi.org/10.1002/art.42265
2. TERSLEV L, SCHMIDT NS, AMMITZBOLL-DANIELSEN M, FANA V: Diagnostic value of high-frequency ultrasound assessment of the lacrimal glands for primary Sjögren's disease. *RMD Open* 2025; 11. https://doi.org/10.1136/rmdopen-2025-005884
3. SARKAR I, DAVIES R, AAREBROT AK *et al*.: Aberrant signaling of immune cells in Sjögren's syndrome patient subgroups upon interferon stimulation. *Front Immunol* 2022; 13: 854183. https://doi.org/10.3389/fimmu.2022.854183
4. XU J, SHEN Z, DU Y *et al*.: Huoxue Jiedu Recipe represses mitochondrial fission to alleviate submandibular gland inflammation in Sjögren's syndrome. *Microbiol Immunol* 2023; 67: 377-87. https://doi.org/10.1111/1348-0421.13084
5. MACKIEWICZ Z, MAZUL J, NARKEVICIUTE I *et al*.: Sjögren's syndrome: concerted triggering of sicca conditions. *J Immunol Res* 2019; 2019: 2075803. https://doi.org/10.1155/2019/2075803
6. ZHANG N, JI C, BAO X *et al*.: Uncovering potential new biomarkers and immune infiltration characteristics in primary Sjögren's syndrome by integrated bioinformatics analysis. *Medicine* (Baltimore) 2023; 102: e35534. https://doi.org/10.1097/md.0000000000035534
7. SHIBOSKI CH, SHIBOSKI SC, SEROR R *et al*.: 2016 American College of Rheumatology/European League Against Rheumatism Classification Criteria for Primary Sjögren's Syndrome: a consensus and data-driven methodology involving three international patient cohorts. *Arthritis Rheumatol* 2017; 69: 35-45. https://doi.org/10.1002/art.39859
8. PIJPE J, KALK WW, VAN DER WAL JE *et al*.: Parotid gland biopsy compared with labial biopsy in the diagnosis of patients with primary Sjögren's syndrome. *Rheumatology* (Oxford) 2007; 46: 335-41. https://doi.org/10.1093/rheumatology/kel266
9. COSTA S, QUINTIN-ROUE I, LESOURD A *et al*.: Reliability of histopathological salivary gland biopsy assessment in Sjögren's syndrome: a multicentre cohort study. *Rheumatology* (Oxford) 2015; 54: 1056-64. https://doi.org/10.1093/rheumatology/keu453
10. JIN Y, LI J, CHEN J *et al*.: Tissue-specific autoantibodies improve diagnosis of primary Sjögren's syndrome in the early stage and indicate localized salivary injury. *J Immunol Res*. 2019; 2019: 3642937. https://doi.org/10.1155/2019/3642937
11. SEROR R, NOCTURNE G, MARIETTE X: Current and future therapies for primary Sjögren syndrome. *Nat Rev Rheumatol* 2021; 17: 475-86. https://doi.org/10.1038/s41584-021-00634-x
12. TROMBY F, MANFRE V, CHATZIS LG *et al*.: Clinical manifestations, imaging and treatment of Sjögren's disease: one year in review 2024. *Clin Exp Rheumatol* 2024; 42: 2322-35. https://doi.org/10.55563/clinexprheumatol/5xq3fb
13. ZHANG H, ZHANG H, YANG H *et al*.: Machine learning-based integrated identification of predictive combined diagnostic biomarkers for endometriosis. *Front Genet* 2023; 14: 1290036. https://doi.org/10.3389/fgene.2023.1290036
14. LIU J, FENG J, ZHU F *et al*.: Analysis of the relationships between interferon-stimulated genes and anti-SSA/Ro 60 antibodies in primary Sjögren's syndrome patients via multi-omics and machine learning methods. *Int Immunopharmacol* 2025; 144: 113652. https://doi.org/10.1016/j.intimp.2024.113652
15. WANG F, LIANG Y, WANG QW: Interpretable machine learning-driven biomarker identification and validation for Alzheimer's disease. *Sci Rep* 2024; 14: 30770. https://doi.org/10.1038/s41598-024-80401-6
16. LESSARD CJ, LI H, ADRIANTO I *et al*.: Variants at multiple loci implicated in both innate and adaptive immune responses are associated with Sjögren's syndrome. *Nat Genet* 2013; 45: 1284-92. https://doi.org/10.1038/ng.2792
17. JAMES K, AL-ALI S, TARN J *et al*.: A transcriptional signature of fatigue derived from patients with primary Sjögren's syndrome. *PLoS One* 2015; 10: e0143970. https://doi.org/10.1371/journal.pone.0143970
18. SJOSTRAND M, AMBROSI A, BRAUNER S *et al*.: Expression of the immune regulator tripartite-motif 21 is controlled by IFN regulatory factors. *J Immunol* 2013; 191: 3753-63. https://doi.org/10.4049/jimmunol.1202341
19. TASAKI S, SUZUKI K, NISHIKAWA A *et al*.: Multiomic disease signatures converge to cytotoxic CD8 T cells in primary Sjögren's syndrome. *Ann Rheum Dis* 2017; 76: 1458-66. https://doi.org/10.1136/annrheumdis-2016-210788
20. HONG X, MENG S, TANG D *et al*.: Single-cell RNA sequencing reveals the expansion of cytotoxic CD4(+) T lymphocytes and a landscape of immune cells in primary Sjögren's syndrome. *Front Immunol* 2020; 11: 594658. https://doi.org/10.3389/fimmu.2020.594658
21. RITCHIE ME, PHIPSON B, WU D *et al*.: limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res* 2015; 43: e47. https://doi.org/10.1093/nar/gkv007
22. YAGIN B, YAGIN FH, COLAK C *et al*.: Cancer metastasis prediction and genomic biomarker identification through machine learning and eXplainable artificial intelligence in breast cancer research. *Diagnostics* (Basel) 2023; 13. https://doi.org/10.3390/diagnostics13213314
23. GOKCE E, FRERET T, LANGEARD A: Blood biomarker signatures for slow gait speed in older adults: an explainable machine learning approach. *Brain Behav Immun* 2025; 124: 295-304. https://doi.org/10.1016/j.bbi.2024.12.007
24. XU S, HU E, CAI Y *et al*.: Using clusterProfiler to characterize multiomics data. *Nat Protoc* 2024; 19: 3292-320. https://doi.org/10.1038/s41596-024-01020-z
25. NEWMAN AM, LIU CL, GREEN MR *et al*.: Robust enumeration of cell subsets from tissue expression profiles. *Nat Methods* 2015; 12: 453-57. https://doi.org/10.1038/nmeth.3337
26. SUBRAMANIAN A, TAMAYO P, MOOTHA VK *et al*.: Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci USA* 2005; 102: 15545-50. https://doi.org/10.1073/pnas.0506580102
27. ZHAO P, ZHEN H, ZHAO H, HUANG Y, CAO B: Identification of hub genes and potential molecular mechanisms related to radiotherapy sensitivity in rectal cancer based on multiple datasets. *J Transl Med* 2023; 21: 176. https://doi.org/10.1186/s12967-023-04029-2
28. KORSUNSKY I, MILLARD N, FAN J *et al*.: Fast, sensitive and accurate integration of single-cell data with Harmony. *Nat Methods* 2019; 16: 1289-96. https://doi.org/10.1038/s41592-019-0619-0
29. YANG S, CORBETT SE, KOGA Y *et al*.: Decontamination of ambient RNA in single-cell RNA-seq with DecontX. *Genome Biol* 2020; 21: 57. https://doi.org/10.1186/s13059-020-1950-6
30. XI NM, LI JJ: Benchmarking computational doublet-detection methods for single-cell RNA sequencing data. *Cell Syst* 2021; 12: 176-94 e6. https://doi.org/10.1016/j.cels.2020.11.008
31. ARAN D, LOONEY AP, LIU L *et al*.: Reference-based analysis of lung single-cell sequencing reveals a transitional profibrotic macrophage. *Nat Immunol* 2019; 20: 163-72. https://doi.org/10.1038/s41590-018-0276-y
32. HU C, LI T, XU Y *et al*.: CellMarker 2.0: an updated database of manually curated cell markers in human/mouse and web tools based on scRNA-seq data. *Nucleic Acids Res* 2023; 51: D870-D6. https://doi.org/10.1093/nar/gkac947
33. IANEVSKI A, GIRI AK, AITTOKALLIO T: Fully-automated and ultra-fast cell-type identification using specific marker combinations from single-cell transcriptomic data. *Nat Commun* 2022; 13: 1246.

https://doi.org/10.1038/s41467-022-28803-w
34. JIN S, GUERRERO-JUAREZ CF, ZHANG L *et al*.: Inference and analysis of cell-cell communication using CellChat. *Nat Commun* 2021; 12: 1088. https://doi.org/10.1038/s41467-021-21246-9
35. AHMAD S, DA COSTA GONZALES JL, BOWLER-BARNETT EH *et al*.: The UniProt website API: facilitating programmatic access to protein knowledge. *Nucleic Acids Res* 2025; 53: W547-W53. https://doi.org/10.1093/nar/gkaf394
36. BISWAS R, BAGCHI A: Inhibition of TRAF6-Ubc13 interaction in NFkB inflammatory pathway by analyzing the hotspot amino acid residues and protein-protein interactions using molecular docking simulations. *Comput Biol Chem* 2017; 70: 116-24. https://doi.org/10.1016/j.compbiolchem.2017.08.014
37. LIU Y, YANG X, GAN J *et al*.: CB-Dock2: improved protein-ligand blind docking by integrating cavity detection, docking and homologous template fitting. *Nucleic Acids Res* 2022; 50: W159-W64. https://doi.org/10.1093/nar/gkac394.
38. LUO W, DENG J, HE J *et al*.: Integration of molecular docking, molecular dynamics and network pharmacology to explore the multi-target pharmacology of fenugreek against diabetes. *J Cell Mol Med* 2023; 27: 1959-74. https://doi.org/10.1111/jcmm.17787.
39. MALEKI-FISCHBACH M, KASTSIANOK L, KOSLOW M, CHAN ED: Manifestations and management of Sjögren's disease. *Arthritis Res Ther* 2024; 26: 43. https://doi.org/10.1186/s13075-024-03262-4
40. PSIANOU K, PANAGOULIAS I, PAPANASTASIOU AD *et al*.: Clinical and immunological parameters of Sjögren's syndrome. *Autoimmun Rev* 2018; 17: 1053-64. https://doi.org/10.1016/j.autrev.2018.05.005
41. ZENG H, ZHOU Y, LIU Z, LIU W: MiR-21-5p modulates LPS-induced acute injury in alveolar epithelial cells by targeting SLC16A10. *Sci Rep* 2024; 14: 11160. https://doi.org/10.1038/s41598-024-61777-x
42. MOMOHARA S, OKAMOTO H, KOMIYA K *et al*.: Matrix metalloproteinase 28/epilysin expression in cartilage from patients with rheumatoid arthritis and osteoarthritis: comment on the article by Kevorkian *et al*. *Arthritis Rheum* 2004; 50: 4074-75; author reply 5. https://doi.org/10.1002/art.20799
43. MYLES A, TUTEJA A, AGGARWAL A: Synovial fluid mononuclear cell gene expression profiling suggests dysregulation of innate immune genes in enthesitis-related arthritis patients. *Rheumatology* (Oxford) 2012; 51: 1785-89. https://doi.org/10.1093/rheumatology/kes151
44. LI Z, GUO J, BI L: Role of the NLRP3 inflammasome in autoimmune diseases. *Biomed Pharmacother* 2020; 130: 110542. https://doi.org/10.1016/j.biopha.2020.110542
45. MASLINSKA M: The role of Epstein-Barr virus infection in primary Sjögren's syndrome. *Curr Opin Rheumatol* 2019; 31: 475-83. https://doi.org/10.1097/bor.0000000000000622
46. TUNG CH, LI CY, CHEN YC, CHEN YC: Association between nucleos(t)ide analogue therapy for hepatitis B and Sjögren's syndrome: 15-year analysis of the national database of Taiwan. *J Viral Hepat* 2021; 28: 809-16. https://doi.org/10.1111/jvh.13481
47. KIRIPOLSKY J, KRAMER JM: Current and emerging evidence for toll-like receptor activation in Sjögren's syndrome. *J Immunol Res* 2018; 2018: 1246818. https://doi.org/10.1155/2018/1246818.
48. LI M, LI M, QIAO L *et al*.: Role of JAK-STAT signaling pathway in pathogenesis and treatment of primary Sjögren's syndrome. *Chin Med J* (Engl) 2023; 136: 2297-306. https://doi.org/10.1097/cm9.0000000000002539
49. WU KY, KULBAY M, TANASESCU C *et al*.: An overview of the dry eye disease in Sjögren's syndrome using our current molecular understanding. *Int J Mol Sci* 2023; 24. https://doi.org/10.3390/ijms24021580
50. FESSLER J, FASCHING P, RAICHT A *et al*.: Lymphopenia in primary Sjögren's syndrome is associated with premature aging of naive CD4$^+$ T cells. *Rheumatology* (Oxford) 2021; 60: 588-97. https://doi.org/10.1093/rheumatology/keaa105