Article

Discovery Medicine 2026; 38(214): 3634–3656
https://doi.org/10.24976/Discov.Med.202638214.277

Integrated Transcriptomic and Machine-Learning Analyses Identify Angiogenesis-Associated Molecular Signatures and Immune-Related Characteristics in the Eutopic Endometrium of Patients With Endometriosis

Xiaomin Yang[1,2,†], Ou Jin[3,†], Xiaojun Yang[1,*]

[1]Department of Obstetrics and Gynecology, The First Affiliated Hospital of Soochow University, 215006 Suzhou, Jiangsu, China
[2]Department of Gynecology, Jiaxing Maternity and Child Health Care Hospital, 314000 Jiaxing, Zhejiang, China
[3]Department of Obstetrics and Gynecology, Jiaxing Hospital of Traditional Chinese Medicine, 314000 Jiaxing, Zhejiang, China
*Correspondence: yang_xiaomin@hotmail.com (Xiaojun Yang)
[†]These authors contributed equally.
Submitted: 10 August 2026 Revised: 3 September 2026 Accepted: 23 September 2026 Published: 25 November 2026

Background: Endometriosis (EMs) is a chronic inflammatory gynecological disorder in which angiogenesis, extracellular matrix remodeling, and immune dysregulation contribute to disease biology. However, molecular alterations in the eutopic endometrium of patients with EMs, particularly those related to angiogenesis and immune-associated transcriptional patterns, remain incompletely characterized. We therefore sought to identify angiogenesis-associated molecular signatures and immune-related characteristics in eutopic endometrium using an integrated computational framework.

Methods: We integrated transcriptomic analysis, functional enrichment, machine-learning algorithms (least absolute shrinkage and selection operator (LASSO) and random forest), and weighted gene co-expression network analysis (WGCNA) to prioritize angiogenesis-associated candidate genes in the eutopic endometrium of patients with EMs. GSE51981 served as the discovery cohort, whereas GSE23339 was used for exploratory cross-dataset, cross-tissue evaluation. Cell-type identification by estimating relative subsets of RNA transcripts (CIBERSORT) was applied to characterize immune-cell-associated transcriptional patterns in GSE51981. Exploratory compound prioritization based on L1000CDS2 transcriptomic matching was followed by an independent exploratory structural analysis using molecular docking and molecular dynamics simulations.

Results: A total of 238 angiogenesis-related differentially expressed genes were identified in the discovery cohort and were mainly enriched in vascular development, extracellular matrix organization, inflammatory signaling, and immune regulation. Integrated machine-learning and WGCNA analyses prioritized *COL6A1*, *ERN2*, and *IL34* as candidate genes. In the exploratory cross-dataset cohort, *ERN2* and *IL34* showed between-group expression differences, whereas *COL6A1* did not. After GSE23339 had also been used to compare the three classifier families, the selected random forest model yielded an area under the curve (AUC) of 0.806 in that cohort; this estimate therefore represents exploratory cross-dataset performance rather than independent validation. CIBERSORT identified exploratory differences in inferred immune-cell fractions, with *ERN2* showing the strongest nominal associations. Exploratory transcriptomic matching prioritized several compounds, and the subsequent *COL6A1* structural analysis was interpreted as independent and hypothesis-generating rather than evidence of direct targeting or therapeutic activity.

Conclusions: Integrated computational analyses identified angiogenesis-associated molecular features in the eutopic endometrium of patients with endometriosis. *ERN2* and *IL34* showed stronger cross-dataset support, whereas *COL6A1* remained a discovery-stage candidate. These findings should be interpreted in the context of eutopic endometrial transcriptional characteristics and should not be directly extrapolated to angiogenesis in ectopic lesions or the local lesion immune microenvironment. Further validation in patient-derived tissues and functional experimental models is required.

Keywords: EMs; angiogenesis; biomarker; WGCNA; immune infiltration; compound prioritization

Copyright: © 2026 The Author(s). Published by Discovery Medicine. This is an open access article under the CC BY 4.0 license.
Note: *Discovery Medicine* stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Introduction

Endometriosis (EMs) is a chronic estrogen-dependent gynecological disorder characterized by the presence of endometrial-like tissue outside the uterine cavity. It affects approximately 10% of women of reproductive age and is a major contributor to infertility and chronic pelvic pain [1]. Retrograde menstruation, together with implantation, inflammation, and aberrant tissue remodeling, is widely considered a central mechanism in disease initiation. Angiogenesis is a essential for the survival, growth, and persistence of endometriotic lesions, and vascular endothelial growth factor (VEGF)-mediated neovascularization has been consistently implicated in disease pathogenesis [2]. Beyond VEGF, recent transcriptomic and machine-learning studies have identified additional angiogenesis-associated molecular signatures in endometriosis, supporting their investigation as potential disease-associated biomarkers.

In parallel with angiogenic remodeling, immune dysregulation is a major feature of endometriosis, with macrophages, T-cell subsets, and cytokine networks contributing to the peritoneal and lesional immune microenvironment [3,4,5]. Altered macrophage polarization and dysregulated adaptive immune responses have also been implicated in lesion establishment and progression [4,6].

Advances in high-throughput transcriptomics and computational biology have enabled systematic characterization of disease-associated molecular signatures using publicly available datasets [7]. Integrative approaches combining weighted gene co-expression network analysis (WGCNA) with machine-learning algorithms have proven effective for prioritizing diagnostic and mechanistically relevant candidate genes in endometriosis and related conditions [8].

In this study, we developed an integrated computational framework to investigate angiogenesis-associated molecular features in the eutopic endometrium of patients with EMs by combining differential expression analysis, functional enrichment, two machine-learning algorithms, WGCNA, immune deconvolution, and exploratory cross-dataset evaluation. Because the discovery dataset, GSE51981, comprises eutopic endometrial tissue, all transcriptomic and cell-type identification by estimating relative subsets of RNA transcripts (CIBERSORT)-derived findings from this cohort were interpreted as features of eutopic endometrium rather than as direct characteristics of ectopic lesions.

Materials and Methods

The overall analytical workflow of the study is summarized in Fig. 1.

Dataset Acquisition and Preprocessing

Two human endometriosis-related microarray datasets, GSE51981 and GSE23339, were obtained from the National Center for Biotechnology Information Gene Expression Omnibus (NCBI GEO; https://www.ncbi.nlm.nih.gov/geo/) using the R package GEOquery (version 2.70.0; Bethesda, MD, USA) was used to download datasets from the GEO database. Detailed information on the platforms, tissue sources, and sample composition is provided in Table 1 (Ref. [9,10]). GSE51981, generated on the GPL570 platform, included 77 eutopic early-secretory endometrial samples from patients with endometriosis and 71 normal early-secretory endometrial control samples. GSE23339, generated on the GPL6102 platform, included 10 ovarian endometrioma samples and 9 normal endometrial controls. Because the tissue sources differed substantially between the two datasets, GSE23339 was treated as a cross-dataset, cross-tissue evaluation cohort rather than a tissue-matched replication cohort.

Table 1. GEO microarray chip information.

	GSE51981	GSE23339
Platform	GPL570	GPL6102
Species	Homo sapiens	Homo sapiens
Tissue	Early Secretory Endometrial tissue	Endometrioma
Samples in EMs group	77	10
Samples in Control group	71	9
Reference	PMID: 25243856 [9]	PMID: 21436257 [10]

GEO, Gene Expression Omnibus; EMs, endometriosis.

The two datasets were preprocessed independently. Probe identifiers were mapped to official gene symbols using the corresponding GPL570 and GPL6102 platform annotation files. Probes lacking valid gene symbols or mapped ambiguously to multiple gene symbols were excluded before construction of the gene-level expression matrices. When multiple probes mapped to the same gene symbol, their expression values were averaged using the avereps function in the limma R package (version 3.62.1; Parkville, Victoria, Australia), yielding a single expression profile for each gene.

Probe records containing missing expression values were removed. The distribution of expression values was examined to determine whether the downloaded data had already been log2 transformed; log2 transformation was applied only when the data remained on the original intensity scale. Between-sample normalization was subsequently performed separately for each dataset using the normalizeBetweenArrays function in limma (v3.62.1). Genes with

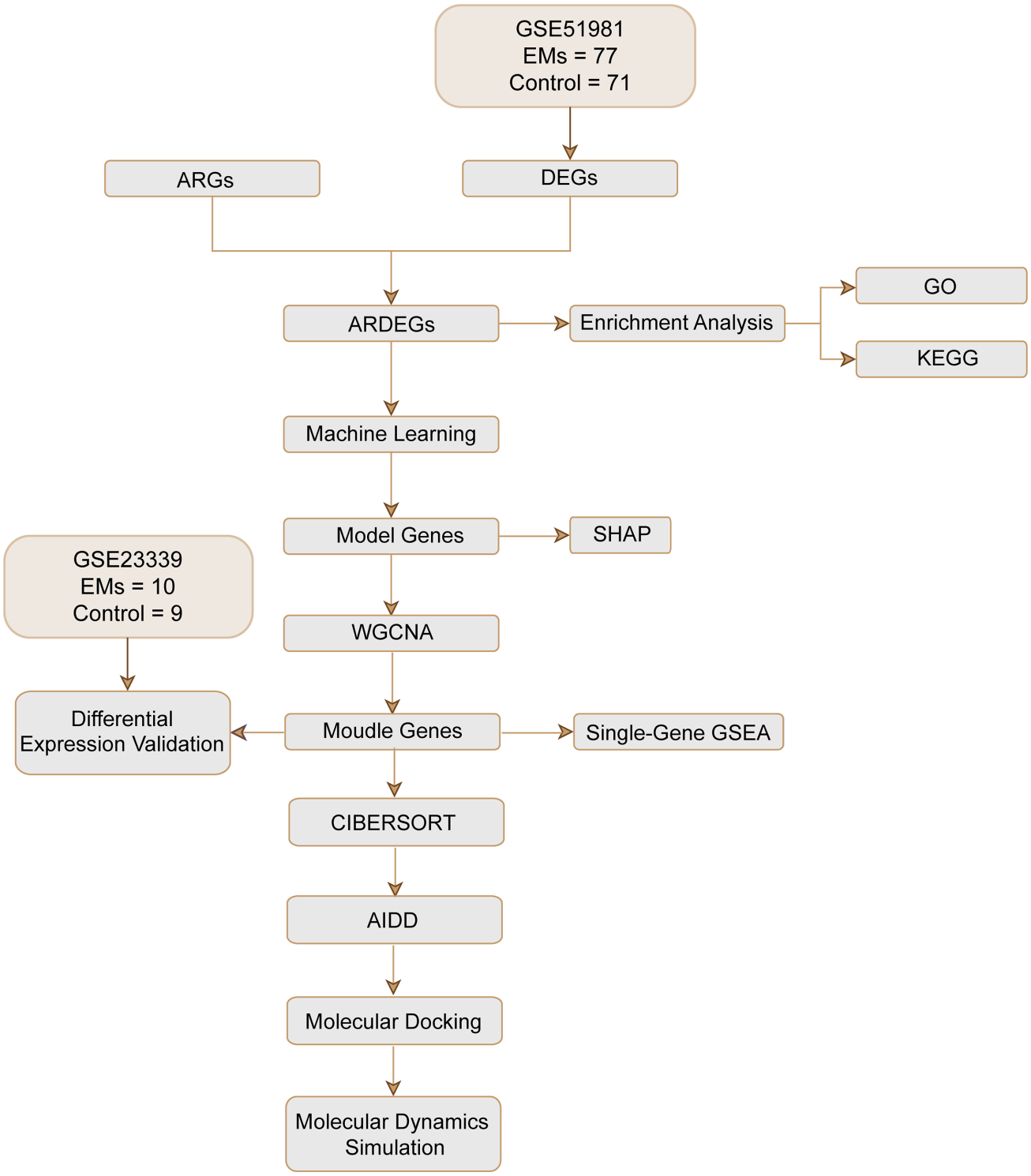

Fig. 1. Flow chart for the comprehensive analysis of angiogenesis-related differentially expressed genes (ARDEGs). This figure was created using Microsoft PowerPoint 2019 (Microsoft Corporation, Redmond, WA, USA) and GraphPad Prism 9.0 (Dotmatics, Boston, MA, USA).

zero expression variance across all samples within a dataset were removed. No additional arbitrary low-expression intensity cutoff was applied.

Because GSE51981 and GSE23339 were analyzed independently rather than merged into a single expression matrix, no cross-dataset batch-effect correction was performed. Each dataset underwent log2 assessment and within-dataset normalizeBetweenArrays processing. During classifier fitting, caret preProcess = c("center", "scale") was specified; centering and scaling parameters were es-

timated from GSE51981 and stored in each fitted model, then automatically applied to GSE23339 during probability prediction. This model-level transformation standardized the selected predictors using training-derived parameters but did not remove tissue- or platform-specific batch effects. Accordingly, cross-dataset results were treated as exploratory.

Acquisition of Angiogenesis-Related Genes

Angiogenesis-related genes (ARGs) were collected from GeneCards and previously published gene sets. The GeneCards database (https://www.genecards.org/) was searched using the keyword "angiogenesis", and entries annotated as protein-coding genes with a relevance score >2 were retained, yielding 2036 candidate ARGs.

A complementary literature search was completed on May 29, 2026, using the terms "angiogenesis" and "GEO". Studies were considered eligible if they reported a clearly defined angiogenesis-related gene set that could be extracted at the gene-symbol level. Studies were excluded if they did not provide an extractable gene list, reported only pathway-level findings, or lacked relevant transcriptomic or gene-set information. Seven eligible studies were retained. Their corresponding gene sets contained 24, 510, 3, 9, 20, 15, and 6 entries, respectively, yielding 587 gene entries before de-duplication.

Gene symbols from the seven literature-derived lists were standardized and merged, and duplicate symbols were removed using the distinct() function, yielding 547 unique literature-derived ARGs. These genes were subsequently combined with the 2036 GeneCards-derived ARGs using the union() function. After cross-source de-duplication based on official gene symbol, 2212 unique ARGs were retained for intersection with the differentially expressed genes.

The seven included studies are listed in **Supplementary Table 1** [11,12,13,14,15,16,17], and-gene-level provenance is provided in **Supplementary Table 2**

Screening of Differentially Expressed Genes (DEGs) and ARDEGs

Based on the predefined sample stratification in GSE51981, limma was used to compare transcriptomic profiles of eutopic endometrium from patients with EMs with those of normal endometrial controls. Genes with |logFC| >0.5 and a Benjamini-Hochberg (BH)-adjusted $p < 0.05$ were considered DEGs. Genes with logFC >0.5 were classified as upregulated, whereas those with logFC <–0.5 were classified as downregulated. Genome-wide expression changes were visualized using a volcano plot. DEGs were then intersected with the curated angiogenesis-related gene set, and the overlapping genes were defined as angiogenesis-related DEGs (ARDEGs). The pheatmap package (version 1.0.12; Tartu, Estonia) was used to visualize the 20 most variable ARDEGs.

GO and KEGG Functional Enrichment Analysis

R package clusterProfiler (version 4.10.0; Guangzhou, China) was used to perform Gene Ontology (GO, including Biological Process (BP), Cellular Component (CC), Molecular Function (MF)) and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway enrichment analyses on 238 ARDEGs. Screening criteria included: adjusted p-value < 0.05, false discovery rate (FDR) (q-value) <0.25, with BH correction. Enrichment results were ranked by adjusted p-value, and the top five entries from BP, CC, MF and KEGG were extracted for visualization. Bubble chart and gene-term interaction network diagrams were plotted to display enrichment characteristics, with node size representing enriched gene count and node color representing logFC value of corresponding genes.

All 238 ARDEGs were used as the initial input with gene symbols converted to ENTREZ identifiers before GO enrichment analysis. Owing to differences in annotation coverage across GO ontologies, 237 genes were included in the BP and CC enrichment calculations, whereas 238 genes were included in the MF analysis. For KEGG enrichment, 175 input genes had valid pathway annotations in the human KEGG TERM2GENE mapping and were included in the corresponding enrichment calculation. Thus, differences in GeneRatio denominators reflect annotation availability rather than pre-enrichment filtering of the original ARDEG set.

Dual Machine Learning Screening of Candidate Model Genes

Two machine learning algorithms were used to screen feature genes from 238 ARDEGs:

Least absolute shrinkage and selection operator (LASSO) Logistic Regression: The glmnet R package (version 5.1; Stanford, CA, USA) was used with set.seed(1012) and family = "binomial". The minimum lambda was selected as the optimal penalty parameter to retain model genes. The results were visualized using a LASSO variable trajectory plot and a diagnostic cross-validation curve.

Random Forest (RF): The randomForest package (version 4.7-1.2; NJ, USA (Merck)) was used with set.seed (1012) and ntree = 1000. The MeanDecreaseGini index was extracted to evaluate gene feature importance, and the top 30 genes with the highest Gini reduction were retained.

A Venn diagram was used to identify the intersecting genes between the LASSO and RF results, which were defined as preliminary machine learning model genes for subsequent integration with WGCNA [18]. These model genes were subsequently used for both classification model development with Shapley Additive Explanations (SHAP) analysis (see the Kernel SHAP Model Interpretability Kernel SHAP and Shapviz subsection) and intersection with WGCNA module genes (see the Weighted Gene Co-Expression Network Analysis WGCNA subsec-

tion). The SHAP Python package (version 0.42.1; GitHub, CA, USA) was employed to interpret machine-learning model outputs. The WGCNA R package (version 1.72-1; Bioconductor, Boston, MA, USA) was used for weighted gene co-expression network analysis.

Weighted Gene Co-Expression Network Analysis WGCNA

R package was used to construct a co-expression network based on the GSE51981 transcriptome data:

Genes with the top 60% variance across all samples were pre-filtered to reduce computational noise;

Pairwise gene correlations were calculated, and a soft-thresholding power of $\beta = 21$ was selected because it achieved a scale-free topology fit index ≥ 0.8. The corresponding mean-connectivity profile was also inspected to avoid selecting the threshold based on scale-free fit alone.

(1) A hierarchical clustering tree was constructed with a minimum module gene number of 25 and a module merging threshold of 0.2, assigning all genes to distinct color-coded modules;

(2) Module-trait correlations between each module and EMs/control phenotype were calculated, and modules with $|r| > 0.3$ and $p < 0.05$ were selected;

(3) A Venn intersection was performed between genes within the selected modules and the preliminary machine learning model genes. The overlapping genes were defined as the final hub module genes.

Construction and Evaluation of Exploratory Multi-Algorithm Classification Models and SHAP Interpretability Analysis

Model Training and Exploratory Cross-Dataset Evaluation

Using the 14 machine-learning-selected genes as predictors, the caret R package (version 6.0-88; CRAN) was used to fit random forest (method = "rf"), recursive partitioning and regression tree (method = "rpart"), and support vector machine with a radial kernel (method = "svmRadial") classifiers in GSE51981. Reproducibility was controlled with set.seed(2024). Resampling used repeated fivefold cross-validation with five repeats. trainControl was configured with classProbs = TRUE, summaryFunction = twoClassSummary, savePredictions = "final", search = "random", and the models were optimized for receiver operating characteristic (ROC). For each algorithm, caret generated a random search with tuneLength = 5. Predictors were centered and scaled using preProcess = c("center", "scale"); the preprocessing parameters were estimated in the training cohort and applied automatically to new data. Control and Case were the class labels, Case was the event class, and predicted Case probabilities were obtained using predict(..., type = "prob")$Case. The code archive did not include the saved model object or model$results output; therefore, the exact sampled tuning values and final mtry (RF), cp (RPART), and sigma/C (SVM radial) cannot be recovered from the script alone. Importantly, GSE23339 was used both to compare the three classifier families and to describe the selected RF model, so it was involved in model selection and was not an independent or external test set.

For exploratory cross-dataset evaluation, discrimination in GSE23339 was summarized using the area under the ROC curve with a bootstrap 95% confidence interval based on 1000 replicates. The optimal threshold was determined using the Youden index, and the corresponding sensitivity and specificity were calculated. Training-derived centering and scaling were applied through the fitted caret preprocessing object. Nevertheless, because GSE23339 was involved in classifier-family selection, contained only 19 samples, and differed in tissue source and microarray platform, these indices describe exploratory cross-dataset discrimination rather than independent validation, generalizable performance, or clinical diagnostic accuracy.

Kernel SHAP Model Interpretability Kernel SHAP and Shapviz

Kernel SHAP and shapviz were applied to the RF model because RF showed the highest area under the curve (AUC) among the three classifiers in GSE23339 and was therefore selected for interpretation. Because GSE23339 contributed to this selection, the SHAP results describe the selected exploratory model and do not establish independently validated feature importance. SHAP analysis was used to characterize feature contributions within the selected model:

(1) A SHAP summary bar plot was constructed to rank model genes based on their mean absolute SHAP values, representing feature importance;

(2) A SHAP swarm plot was generated to illustrate the direction and magnitude of the effects of gene expression on disease prediction probability;

(3) SHAP waterfall plots were generated for representative samples near the decision boundary (predicted probability ≈ 0.5) to quantify the individual contribution of each model gene to the classification results, thereby improving the interpretability of the machine learning model.

Differential Expression Assessment and Exploratory Cross-Dataset ROC Analysis

Training Cohort GSE51981

Group comparison boxplots of the three hub genes were generated to verify expression differences between the EMs and control groups, with statistical significance defined as $p < 0.001$. The pROC package (version 1.18.5; Geneva, Switzerland) was used to plot single-gene ROC curves and calculate AUC values to evaluate the apparent discriminative ability of each candidate gene within the discovery cohort. AUC classification standard: 0.5–0.7 low diagnostic accuracy, 0.7–0.9 moderate accuracy, >0.9 high accuracy.

Exploratory Cross-Dataset Cohort GSE23339

In GSE23339, expression comparisons and single-gene ROC analyses were repeated for *COL6A1*, *ERN2*, and *IL34* as an exploratory assessment of cross-dataset reproducibility. Because this cohort contains ovarian endometrioma rather than eutopic endometrial tissue and differs in microarray platform, these analyses were not interpreted as independent validation, tissue-matched replication, or evidence of generalizable diagnostic performance.

Single-Gene GSEA Enrichment Analysis

For each hub gene, Pearson correlation coefficients between the target gene and all other transcripts in GSE51981 were calculated, and genes were ranked in descending order by correlation coefficient. The clusterProfiler package was sued to perform single-gene GSEA based on the Molecular Signatures Database (MSigDB; https://www.gsea-msigdb.org/gsea/msigdb) c2 reference gene set v2026.1.Hs. (Parameters: gene set size range 10–500, screening threshold adjusted $p < 0.05$, FDR <0.05). The top five positively and negatively enriched gene sets, ranked by NES, were extracted and visualized using Gene Set Enrichment Analysis (GSEA) enrichment curves to identify biological pathways potentially associated with each module gene.

Immune-Cell Deconvolution Using CIBERSORT

The CIBERSORT R script (version 1.03; Stanford, CA, USA) was used to estimate the relative proportions of 22 immune-cell subsets in the GSE51981 expression matrix. The LM22 leukocyte signature matrix, comprising 547 signature genes, was used as the reference. The analysis was performed with 1000 permutations and quantile normalization enabled (QN = TRUE). CIBERSORT generated an empirical deconvolution p value, correlation coefficient, and root-mean-square error (RMSE) for each sample.

A CIBERSORT p value < 0.05 was used only as the sample-level reliability criterion. All 148 GSE51981 samples met this criterion; therefore, no samples were excluded. For group comparisons of immune-cell fractions and gene-cell correlation analyses, reported p values were nominal and were not adjusted for the number of cell subsets or correlations tested. These analyses were therefore considered exploratory, and $p < 0.05$ was described as a nominal association rather than confirmatory statistical significance. Immune-cell subsets with a total inferred abundance of zero across all samples were removed before visualization and analysis.

CIBERSORT estimates derived from bulk transcriptomic data represent inferred immune-cell-associated transcriptional features and should not be interpreted as direct measurements of tissue immune-cell abundance, infiltration, or polarization.

Exploratory Compound Prioritization Based on L1000CDS2 Transcriptomic Matching

All DEGs between eutopic endometrium from patients with EMs and normal controls in GSE51981 were submitted to the L1000CDS2 platform as an EMs-associated eutopic-endometrial transcriptional signature. The platform was used as an exploratory prioritization tool to identify small-molecule perturbation profiles showing transcriptomic similarity to the input signature. The top 50 compounds ranked by matching score were retained for subsequent analysis. This similarity-based screening was not interpreted as evidence that the compounds would reverse disease biology or provide therapeutic benefit.

A compound-candidate gene transcriptomic matching matrix was constructed according to the directional patterns reported by L1000CDS2. The matrix represented concordant transcriptional directions, such as up/up and dn/dn matches, and did not indicate direct compound-gene targeting or physical interaction. Duplicate perturbation records generated under different cell lines, doses, or exposure durations were collapsed by retaining the maximum transcriptomic matching score for each compound-gene pair. Compounds without a transcriptomic match with any candidate gene were excluded.

Compounds were prioritized according to the number of candidate genes represented in the transcriptomic matching matrix and then by the L1000CDS2 score. The top three compounds-ingenol 3,20-dibenzoate (CID5225086), phorbol12-myristate 13-acetate (PMA; CID 27924), and BRD-K25737009 (CID24868309)-were carried forward to an independent exploratory structural analysis. COL6A1 was chosen for this structural analysis because it was a discovery-stage candidate linked to extracellular-matrix organization and angiogenesis-related transcriptional changes and has established extracellular-matrix biological relevance in the literature. This choice was not based on evidence that L1000CDS2 compounds directly target COL6A1. The most favorable modeled complex was subsequently examined by molecular dynamics simulation.

Molecular Docking

Molecular docking was performed using the CB-Dock2 web server as an independent exploratory analysis of predicted structural compatibility between the prioritized compounds and the predicted *COL6A1* structure. It was not used to infer that the compounds directly target *COL6A1*. Three-dimensional compound structures were obtained from PubChem (https://pubchem.ncbi.nlm.nih.gov/), and the predicted *COL6A1* structure was obtained from the AlphaFold Protein Structure Database (https://alphafold.ebi.ac.uk/) [19].

CB-Dock2 automatically detected potential protein cavities and generated the corresponding docking search spaces; therefore, no single binding site was manually predefined. For BRD-K25737009, the docking-box center was

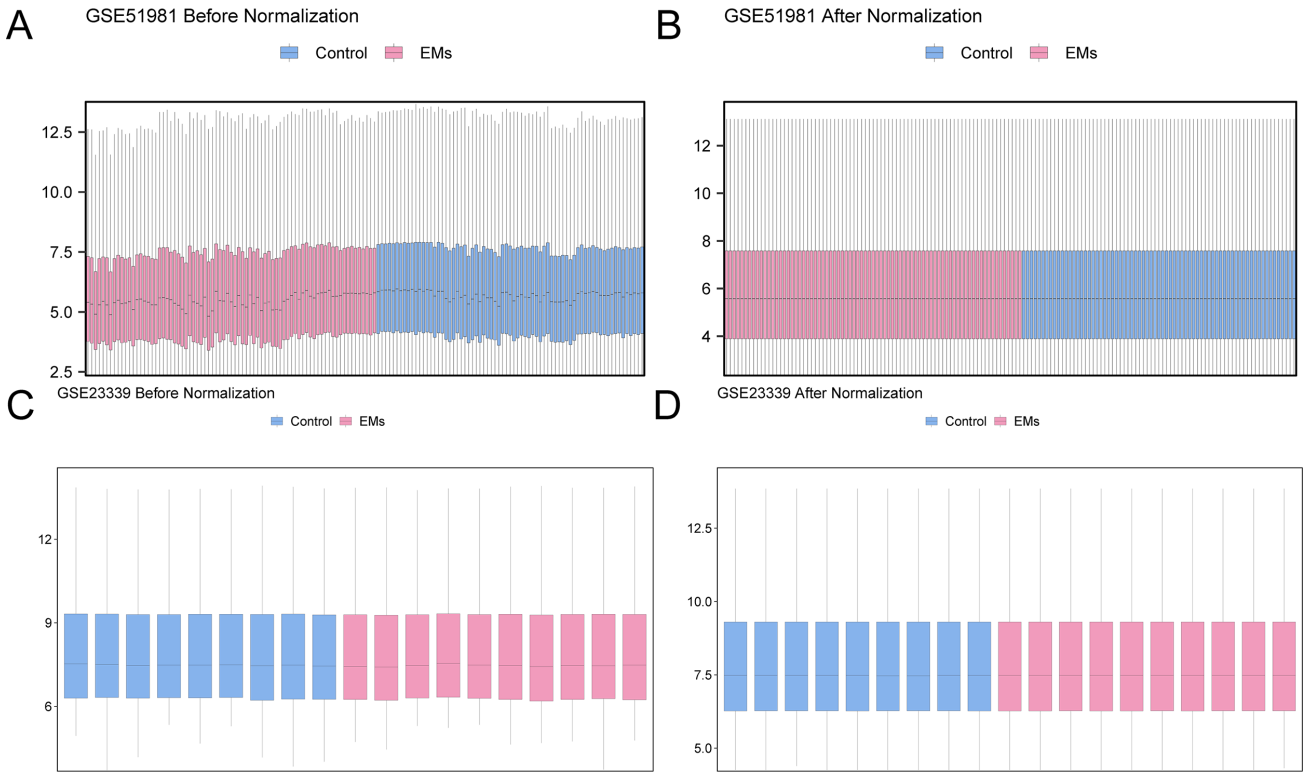

Fig. 2. Data cleaning of GSE51981&GSE23339. (A) Boxplot of GSE51981 distribution before standardized processing. (B) Boxplot of GSE51981 distribution after standardized treatment. (C) Boxplot of GSE23339 distribution before standardized treatment. (D) Boxplot of GSE23339 distribution after standardized treatment.

(37, 23, –29), with a search-space size of 25 × 25 × 25 Å. For ingenol 3,20-dibenzoate, the center was (50, 31, 7), with a search-space size of 23 × 23 × 23 Å. For phorbol 12-myristate 13-acetate, the center was (47, 25, 6), with a search-space size of 26 × 26 × 26 Å. The compound-specific docking centers and search-space dimensions are provided in **Supplementary Table 3** [20,21].

Before docking, water molecules and nonessential heteroatoms were removed from the receptor structure, followed by hydrogen addition and charge assignment. Ligand preparation and subsequent blind docking were performed using the automated CB-Dock2 workflow. Five candidate poses were returned for each compound, and docking was conducted using AutoDock Vina version 1.1.2. The exhaustiveness parameter was controlled internally by the CB-Dock2 server and was not reported in the available task output; therefore, no unverified value was assigned. Docking scores were interpreted as computational estimates of relative structural compatibility rather than experimental evidence of direct binding.

50 ns Molecular Dynamics (MD) Simulation

The compound-protein complex with the optimal docking score (*COL6A1* + BRD-K25737009) was selected for molecular dynamics simulation using GROMACS software (version 2021.4; GROMACS development team, KTH Royal Institute of Technology and Science for Life Laboratory, Stockholm, Sweden) was used for molecular dynamics simulation [22]:

(1) Simulation conditions: constant temperature of 310 K, constant pressure of 101 kPa, periodic boundary conditions, and the TIP3P water model;

(2) The total simulation duration was 50 ns, with trajectory file recorded every 20 ps;

(3) DuIvyTools (version 0.5.1) was used to process the trajectory data and calculate four core indicators:

Root mean square deviation (RMSD) was used to evaluate overall system equilibrium and structural stability; root mean square fluctuation (RMSF) was used to assess residue-level conformational flexibility; radius of gyration (Rg) was used to characterize the compactness of the protein tertiary structure; and solvent-accessible surface area (SASA) was used to monitor changes in solvent exposure during ligand binding.

Results

Dataset Preprocessing and Quality Assessment

Expression profiles in GSE51981 and GSE23339 were normalized separately using limma. Within each dataset, expression distributions became more consistent after normalization (Fig. 2A–D). For classification, caret

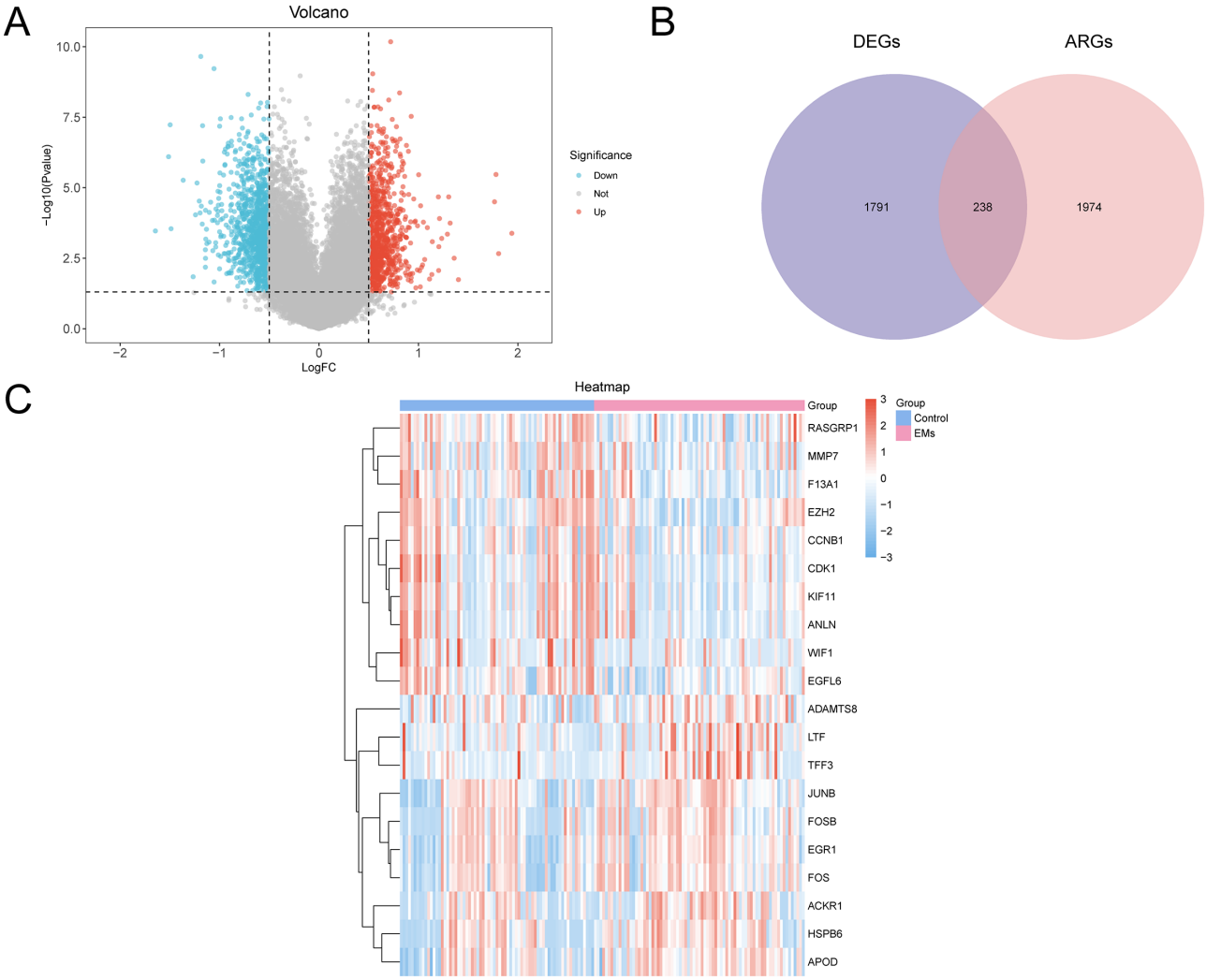

Fig. 3. Differential gene expression analysis. (A) Volcano plot of differentially expressed genes between eutopic endometrium from patients with EMs and normal endometrium in GSE51981. (B) Venn diagram showing the intersection of differentially expressed genes (DEGs) and angiogenesis-related genes (ARGs). (C) Heatmap of representative angiogenesis-related differentially expressed genes (ARDEGs) in GSE51981. EMs, endometriosis. Red represents relatively high expression and blue represents relatively low expression.

additionally centered and scaled the selected predictors using parameters estimated from GSE51981 and applied these stored transformations to GSE23339 during prediction. This training-derived feature standardization did not constitute cross-platform batch-effect correction and could not eliminate tissue- or platform-related heterogeneity.

Identification of Angiogenesis-Related DEGs in Eutopic Endometrium

To investigate angiogenesis-related transcriptional alterations in eutopic endometrium, differential expression analysis was performed between eutopic endometrial samples from patients with endometriosis and normal endometrial controls in GSE51981. A total of 2029 DEGs were identified, including 913 upregulated and 1116 downregulated genes (Fig. 3A).

By intersecting the identified DEGs with angiogenesis-related genes, 238 ARDEGs were obtained (Fig. 3B). The heatmap showed distinct expression patterns of representative ARDEGs between eutopic endometrium from patients with endometriosis and normal endometrium (Fig. 3C), indicating angiogenesis-associated transcriptional differences at the eutopic endometrial level.

Functional Characterization of Angiogenesis-Associated Genes

To explore the biological significance of ARDEGs, GO and KEGG enrichment analyses were performed. The enriched biological processes were primarily associated with the regulation of angiogenesis, vasculature development, and epithelial cell proliferation (Fig. 4A).

Regarding cellular components, ARDEGs were significantly enriched in the collagen-containing extracellular

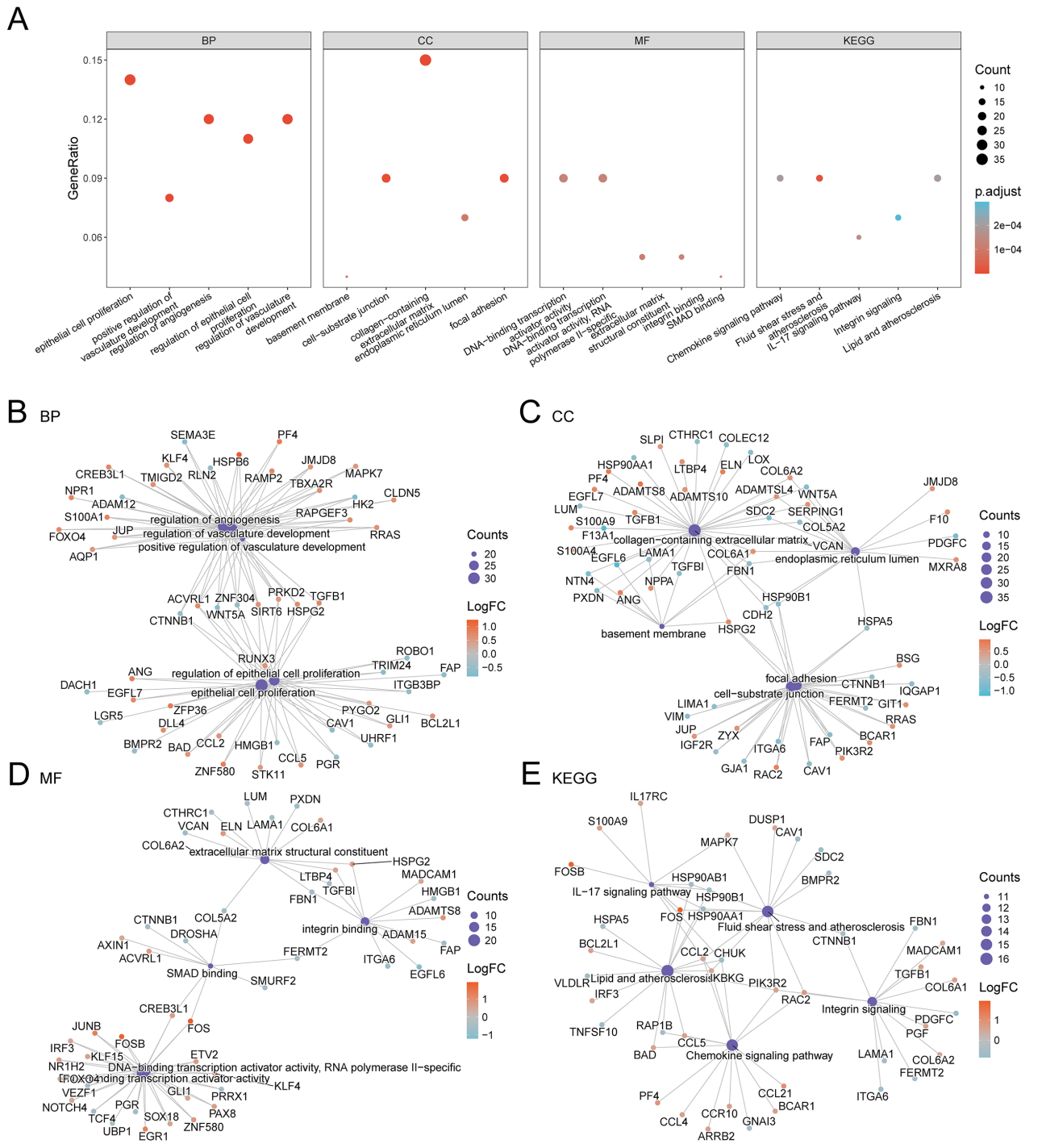

Fig. 4. GO and KEGG enrichment analysis of ARDEGs. (A) Bubble plot showing GO and KEGG enrichment results of ARDEGs, including biological process (BP), cellular component (CC), molecular function (MF), and KEGG pathways. Bubble size represents the number of enriched genes, and bubble color represents the *p* value, with redder colors indicating smaller *p* values. (B–E) Gene-term interaction networks showing the relationships between ARDEGs and enriched terms in BP (B), CC (C), MF (D), and KEGG pathways (E). Purple nodes represent enriched terms, and connected gene nodes represent ARDEGs involved in the corresponding terms. Red gene nodes indicate upregulated genes, whereas blue gene nodes indicate downregulated genes. ARDEGs, angiogenesis-related differentially expressed genes; GO, Gene Ontology; KEGG, Kyoto Encyclopedia of Genes and Genomes.

matrix, focal adhesion, cell-substrate junction, and endoplasmic reticulum lumen, suggesting potential involvement in extracellular matrix remodeling and cellular stress responses. Molecular function analysis further revealed enrichment in extracellular matrix structural components and integrin binding.

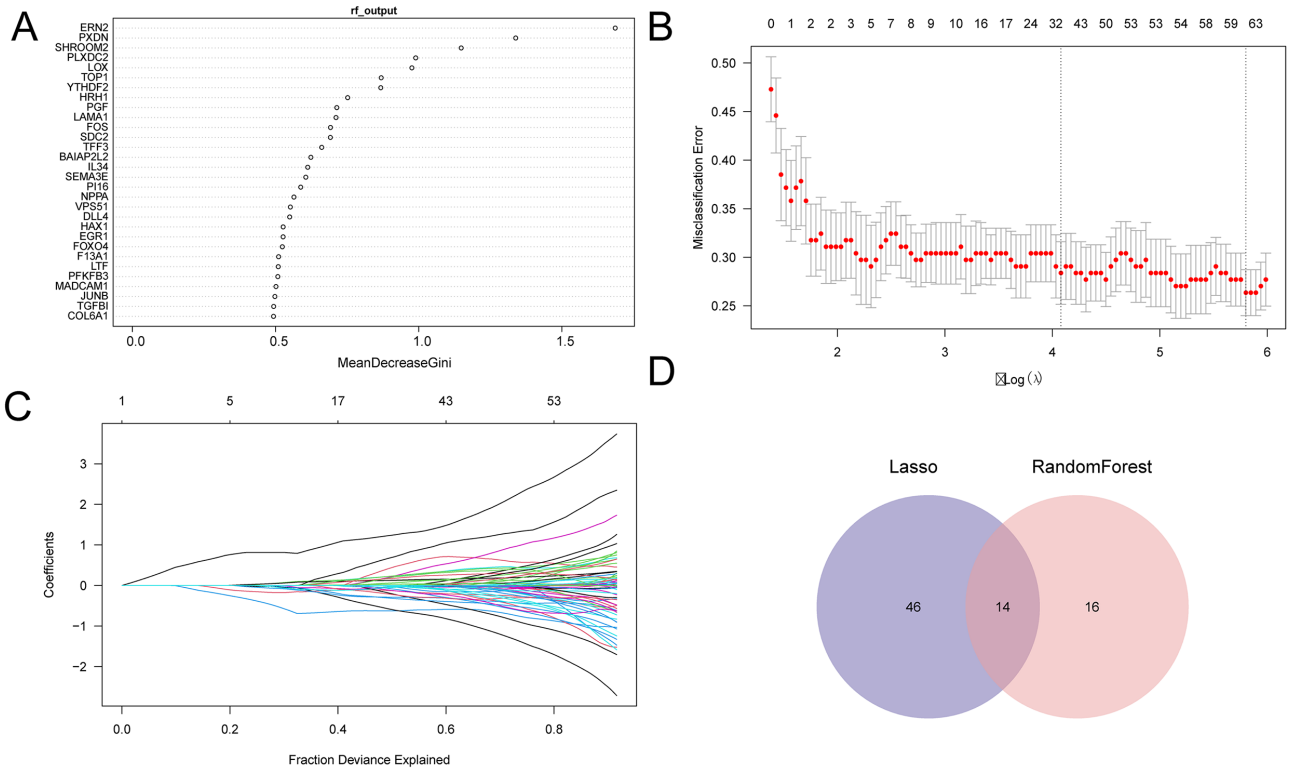

Fig. 5. Identification of angiogenesis-associated candidate genes using machine-learning approaches. (A) Feature importance ranking of angiogenesis-related differentially expressed genes based on the mean decrease Gini index in the random forest (RF) algorithm. (B,C) LASSO regression analysis showing model performance and variable selection process. (B) Cross-validation curve identifying the optimal lambda value. (C) LASSO coefficient trajectory plot showing changes in gene coefficients with increasing penalty strength. (D) Venn diagram showing the intersection between genes selected by RF and LASSO regression, yielding 14 machine-learning-selected candidate genes for subsequent WGCNA analysis. LASSO, least absolute shrinkage and selection operator; WGCNA, weighted gene co-expression network analysis.

KEGG analysis demonstrated significant enrichment in pathways related to inflammatory regulation and vascular-associated signaling, including IL-17 signaling, chemokine signaling, and fluid shear stress-related pathways (Fig. 4A–E). Detailed enrichment information on the top-ranked GO terms and KEGG pathways is summarized in Table 2.

Identification of Candidate Genes Using Dual Machine-Learning Approaches

To identify robust candidate genes associated with angiogenesis-related alterations, two complementary machine-learning algorithms were applied to the 238 ARDEGs. Random forest analysis ranked genes according to feature importance based on mean decrease Gini values, and the top 30 genes were selected (Fig. 5A). In parallel, LASSO regression identified informative genes under the optimal penalty conditions, generating a set of model-associated genes (Fig. 5B,C).

The intersection of the RF-derived and LASSO-selected gene sets yielded 14 overlapping genes (Fig. 5D). These genes were defined as machine-learning-selected candidate genes and were subsequently used for both classification model construction and integration with WGCNA modules.

Integration of WGCNA and Machine-Learning Analysis Identifies Angiogenesis-Associated Hub Genes

To further prioritize biologically relevant genes, WGCNA was performed using the GSE51981 expression dataset. Gene modules were constructed based on expression similarity, and module–trait correlation analysis was conducted to identify modules associated with the endometriosis phenotype (Fig. 6A–D).

Two modules, MElightgreen and MEdarkmagenta, showed significant associations with endometriosis status. By intersecting genes from these phenotype-associated modules with the 14 machine-learning-selected candidate genes, three overlapping genes were identified: COL6A1, ERN2, and IL34 (Fig. 6E,F).

Because these genes were supported by both predictive modeling and co-expression network analysis, *COL6A1*, *ERN2*, and *IL34* were selected as candidate

Table 2. Top five GO and KEGG enrichment results of ARDEGs.

ONTOLOGY	ID	Description	GeneRatio	BgRatio	*p* value	*p*.adjust	q value
BP	GO:1901342	regulation of vasculature development	29/237	369/18,986	2.81×10^{-15}	1.14×10^{-11}	7.22×10^{-12}
BP	GO:0050673	epithelial cell proliferation	32/237	482/18,986	1.10×10^{-14}	1.53×10^{-11}	9.74×10^{-12}
BP	GO:0045765	regulation of angiogenesis	28/237	360/18,986	1.14×10^{-14}	1.53×10^{-11}	9.74×10^{-12}
BP	GO:1904018	positive regulation of vasculature development	20/237	192/18,986	3.87×10^{-13}	3.91×10^{-10}	2.48×10^{-10}
BP	GO:0050678	regulation of epithelial cell proliferation	27/237	406/18,986	1.48×10^{-12}	9.59×10^{-10}	6.09×10^{-10}
CC	GO:0062023	collagen-containing extracellular matrix	35/237	430/19,960	2.01×10^{-19}	6.80×10^{-17}	4.73×10^{-17}
CC	GO:0005925	focal adhesion	22/237	424/19,960	7.95×10^{-9}	1.35×10^{-6}	9.37×10^{-7}
CC	GO:0030055	cell-substrate junction	22/237	434/19,960	1.21×10^{-8}	1.37×10^{-6}	9.52×10^{-7}
CC	GO:0005788	endoplasmic reticulum lumen	16/237	314/19,960	1.19×10^{-6}	1.01×10^{-4}	7.00×10^{-5}
CC	GO:0005604	basement membrane	9/237	92/19,960	1.49×10^{-6}	1.01×10^{-4}	7.02×10^{-5}
MF	GO:0005201	extracellular matrix structural constituent	13/238	166/18,737	2.09×10^{-7}	1.15×10^{-4}	9.04×10^{-5}
MF	GO:0046332	SMAD binding	9/238	79/18,737	7.03×10^{-7}	1.28×10^{-4}	1.01×10^{-4}
MF	GO:0001228	DNA-binding transcription activator activity, RNA polymerase II-specific	21/238	480/18,737	9.25×10^{-7}	1.28×10^{-4}	1.01×10^{-4}
MF	GO:0005178	integrin binding	12/238	161/18,737	1.05×10^{-6}	1.28×10^{-4}	1.01×10^{-4}
MF	GO:0001216	DNA-binding transcription activator activity	21/238	487/18,737	1.17×10^{-6}	1.28×10^{-4}	1.01×10^{-4}
KEGG	hsa05418	Fluid shear stress and atherosclerosis	15/175	142/9364	5.65×10^{-8}	1.46×10^{-5}	9.45×10^{-6}
KEGG	hsa04657	IL-17 signaling pathway	11/175	94/9364	1.28×10^{-6}	1.65×10^{-4}	1.07×10^{-4}
KEGG	hsa05417	Lipid and atherosclerosis	16/175	216/9364	2.63×10^{-6}	1.95×10^{-4}	1.27×10^{-4}
KEGG	hsa04062	Chemokine signaling pathway	15/175	193/9364	3.03×10^{-6}	1.95×10^{-4}	1.27×10^{-4}
KEGG	hsa04518	Integrin signaling	13/175	154/9364	5.75×10^{-6}	2.97×10^{-4}	1.92×10^{-4}

GO, Gene Ontology; BP, Biological Process; MF, Molecular Function; KEGG, Kyoto Encyclopedia of Genes and Genomes; GeneRatio denominators indicate the number of input ARDEGs with valid annotations in the corresponding database or ontology: 237 for GO-BP and GO-CC, 238 for GO-MF, and 175 for KEGG. All 238 ARDEGs were submitted as the initial enrichment input.

angiogenesis-associated hub genes for subsequent validation and biological interpretation.

Construction and Interpretation of Machine-Learning Classification Models

To evaluate whether the machine-learning-selected genes could discriminate endometriosis from control samples, three classification models were constructed using the 14 candidate genes, including RF, recursive partitioning and regression tree (RPART), and support vector machine (SVM).

Among the evaluated models, RF yielded the highest AUC in GSE23339 (0.806; Fig. 7A) and was selected for further interpretation. Because GSE23339 was used for this classifier-family comparison, the same cohort was not an independent test set; the reported RF AUC is an exploratory, selection-influenced cross-dataset estimate.

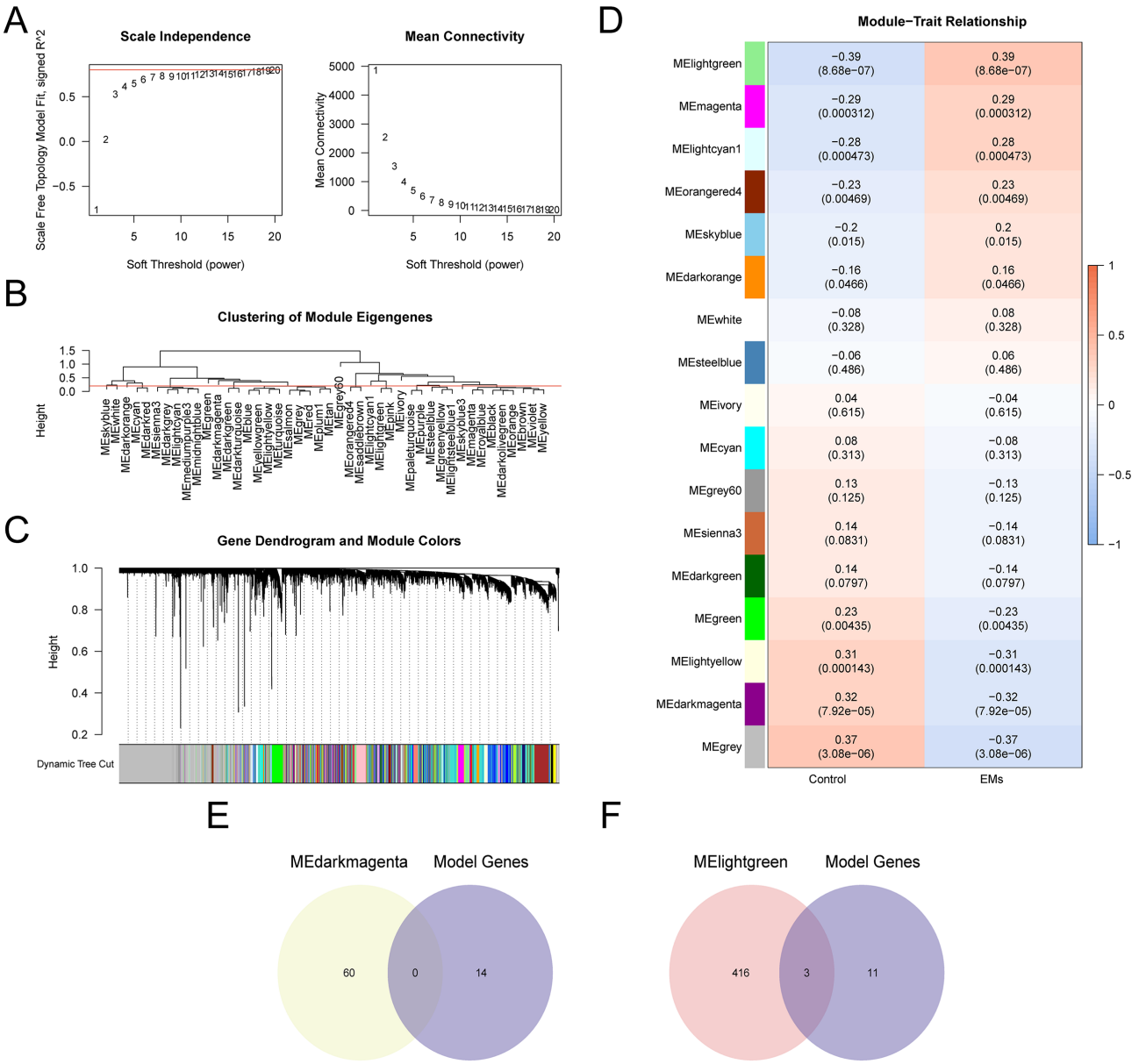

Fig. 6. WGCNA analysis. (A) Analysis of scale-free topology to determine the optimal soft-thresholding power for WGCNA network construction. (B) Gene clustering and module detection based on expression similarity. (C) Hierarchical clustering dendrogram showing gene module assignment. (D) Module–trait correlation heatmap identifying WGCNA modules associated with the endometriosis phenotype. (E,F) Venn diagrams showing the intersection between the 14 machine-learning-selected genes and genes from the MElightgreen and MEdarkmagenta modules, identifying three hub genes: *COL6A1*, *ERN2*, and *IL34*. WGCNA, weighted gene co-expression network analysis.

SHAP analysis was subsequently applied to characterize the contribution of individual features within the RF model. The SHAP summary plot ranked genes according to their contribution to model prediction, while the SHAP swarm plot illustrated the direction and magnitude of gene effects across individual samples (Fig. 7B,C). The waterfall plot further demonstrated how individual features collectively influenced the prediction outcome in representative samples (Fig. 7D).

These results provided an interpretable framework for understanding the contribution of machine-learning-selected genes to endometriosis classification.

In the exploratory cross-dataset, cross-tissue cohort GSE23339, the selected RF model yielded an AUC of 0.806 (95% CI: 0.572–1.000). The Youden-index threshold was 0.494, with sensitivity of 90.0% and specificity of 77.8%. Although training-derived centering and scaling were applied through the fitted caret model, the wide confidence interval, small sample size, tissue and platform differences,

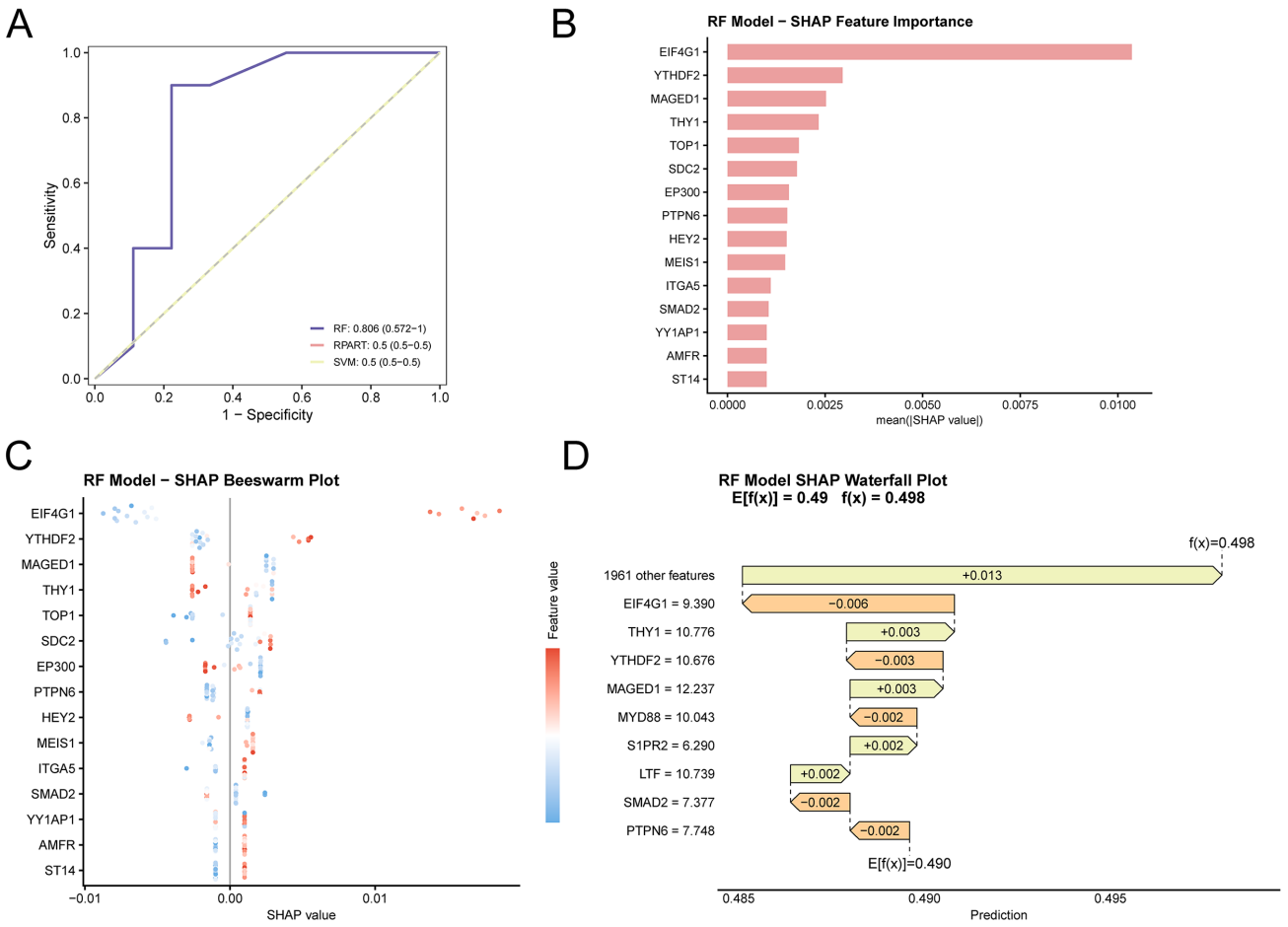

Fig. 7. Performance evaluation and interpretability analysis of the random forest classification model. (A) ROC curves comparing the discriminative performance of RF, RPART, and SVM models based on the 14 machine-learning-selected genes. (B) SHAP summary bar plot ranking feature importance in the RF model according to mean absolute SHAP values. (C) SHAP swarm plot showing the direction and magnitude of individual gene contributions to model prediction. Red and blue colors represent relatively high and low gene expression levels, respectively. (D) SHAP waterfall plot illustrating feature-specific contributions to the prediction of representative samples. Red features increase in the predicted probability of endometriosis, whereas blue features decrease in the predicted probability. ROC, receiver operating characteristic; AUC, area under the curve; SHAP, Shapley Additive Explanations; RF, random forest; RPART, recursive partitioning and regression tree; SVM, support vector machine. In GSE23339, the RF model achieved an AUC of 0.806 (95% CI: 0.572–1.000), with a sensitivity of 90.0% and specificity of 77.8% at the Youden-index-derived threshold of 0.494.

and use of this cohort during classifier-family selection create substantial uncertainty. The result should not be interpreted as independent validation, established model generalizability, or clinical diagnostic performance.

Expression and Exploratory Discriminative Analysis of Angiogenesis-Associated Candidate Genes

Expression Validation of *COL6A1*, *ERN2*, and *IL34*

Group comparison boxplots demonstrated that *COL6A1*, *ERN2*, and *IL34* were significantly different between eutopic endometrium from patients with EMs and normal endometrial controls in the discovery cohort (all $p < 0.001$, Fig. 8A). Single-gene ROC curves showed moderate discrimination for each discovery-stage candidate ($0.7 < \text{AUC} < 0.9$, Fig. 8B–D). In the exploratory cross-dataset cohort GSE23339 (10 ovarian endometrioma samples and 9 controls), *ERN2* and *IL34* showed between-group expression differences, whereas *COL6A1* did not (Fig. 9A). *ERN2* yielded an AUC of 0.956 (95% CI: 0.861–1.000), and *IL34* yielded an AUC of 0.889 (95% CI: 0.692–1.000) (Fig. 9B,C). These findings provide exploratory cross-dataset support for *ERN2* and *IL34*, but not for *COL6A1*. They should not be interpreted as independent validation, tissue-matched confirmation, or evidence of clinical diagnostic utility.

Single-Gene GSEA Pathway Analysis

Single-gene GSEA identified shared and distinct transcriptional programs associated with *COL6A1, ERN2*, and *IL34* (Fig. 10A–C). These gene-set associations were in-

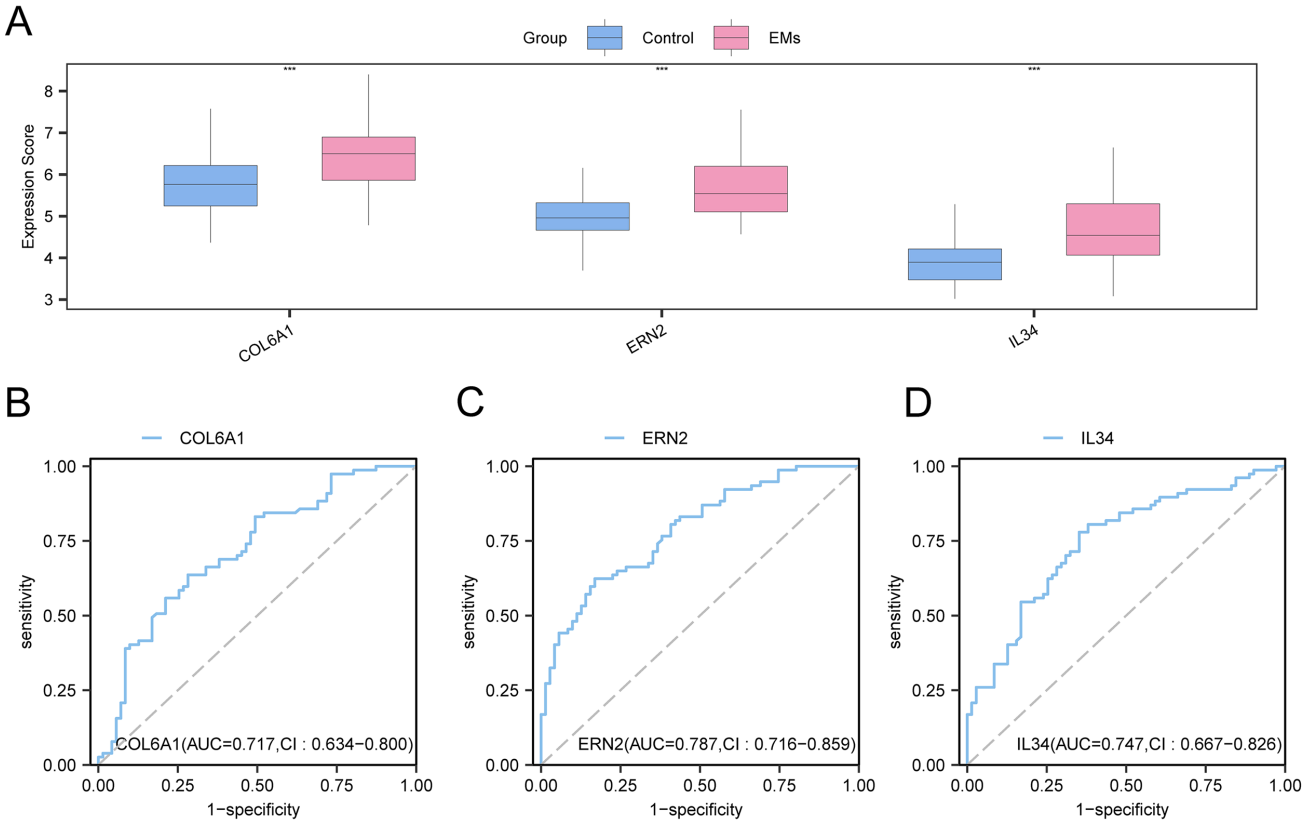

Fig. 8. Expression and discriminative performance of the candidate genes in the discovery cohort. (A) Expression levels of *COL6A1*, *ERN2*, and *IL34* in eutopic endometrial samples from patients with endometriosis and normal endometrial controls in GSE51981. (B–D) ROC curve analyses of *COL6A1* (B), *ERN2* (C), and *IL34* (D), respectively, for distinguishing endometriosis from control samples in the discovery cohort. ***$p < 0.001$. ROC, receiver operating characteristic; AUC, area under the curve.

terpreted as correlation-based pathway signals rather than evidence of direct regulation of the corresponding biological processes by the candidate genes:

(1) *COL6A1* and *ERN2* were positively enriched in the HAMAI_APOPTOSIS_VIA_TRAIL_DN gene set, suggesting a potential association with apoptosis-related pathways;

(2) All three candidate genes showed negative enrichment in MILI_PSEUDOPODIA_HAPTOTAXIS_UP, WILCOX_RESPONSE_TO_PROGESTERONE_UP, and WP_GENES_RELATED_TO_PRIMARY_CILIUM_DEVELOPMENT. These findings suggest inverse associations between higher candidate-gene expression and transcriptional programs related to cell motility/adhesion, progesterone responsiveness, and ciliary biology. Their biological relevance to endometriosis remains exploratory and requires targeted functional validation;

(3) *IL34* was negatively enriched in PLASARI_TGFB1_TARGETS_10HR_DN, suggesting a potential association with TGF-β-related transcriptional signatures.

Immune Infiltration Landscape and Hub Gene-Immune Cell Correlation

CIBERSORT deconvolution of GSE51981 estimated the relative fractions of 22 immune-cell subsets in eutopic endometrial samples (Fig. 11A). At an unadjusted nominal $p < 0.05$, exploratory differences were observed for naive B cells, CD8+ T cells, gamma-delta T cells, monocytes, and M2 macrophages (Fig. 11B). Gamma-delta T cells and monocytes showed the strongest nominal inverse correlation (r = –0.707, unadjusted $p < 0.05$; Fig. 11C). Because multiple cell subsets and correlations were examined without multiplicity correction, these findings are vulnerable to false-positive results. They reflect exploratory immune-related transcriptional characteristics of eutopic endometrium and should not be interpreted as confirmatory differences or as direct measurements of the immune microenvironment in ectopic lesions.

Among the three candidate genes, *ERN2* showed the strongest nominal associations with inferred immune-cell fractions, including a positive correlation with monocytes and a negative correlation with gamma-delta T cells. *COL6A1* and *IL34* showed comparatively weaker nominal associations (Fig. 11D). These unadjusted multiple comparisons are exploratory. Bulk-transcriptomic deconvolu-

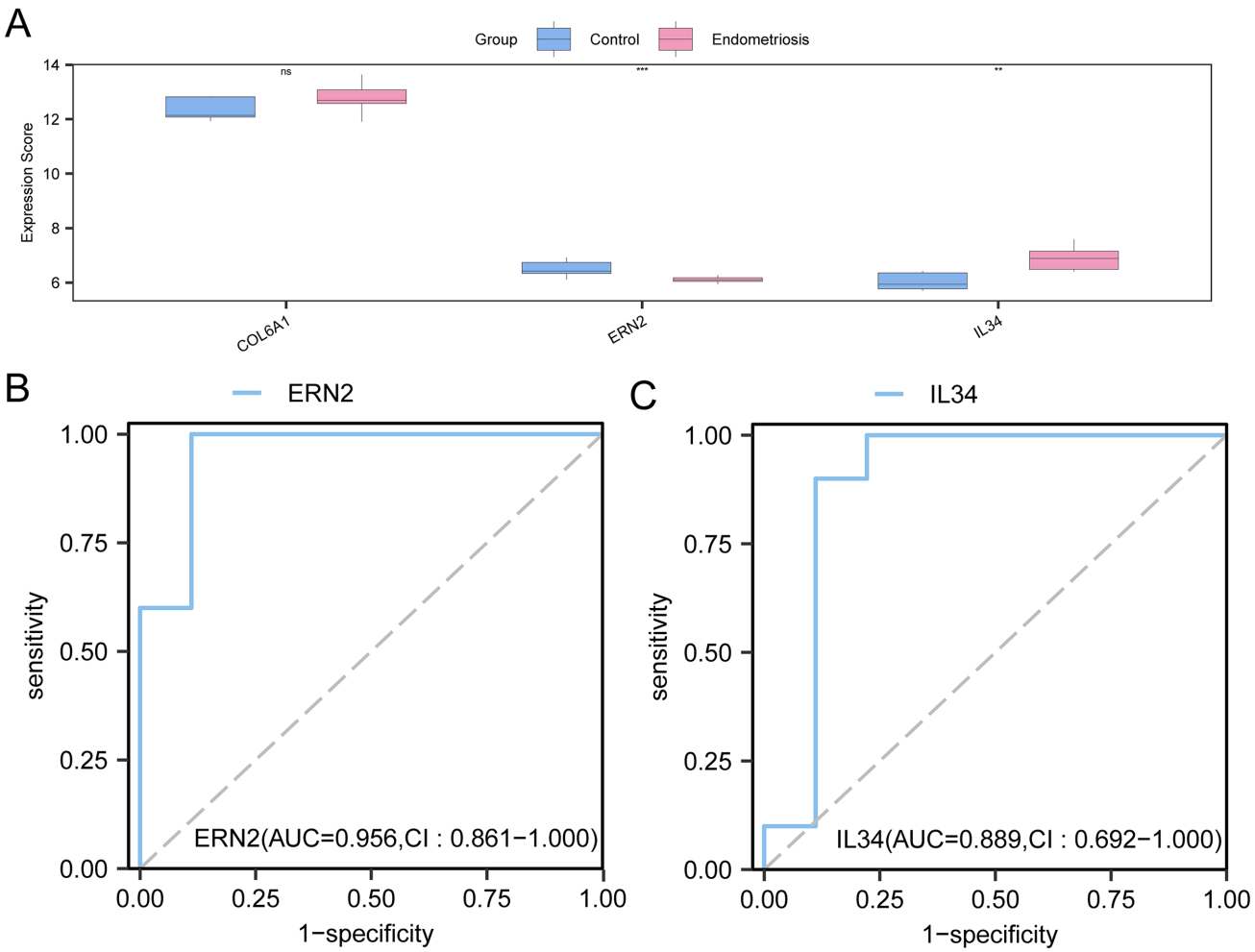

Fig. 9. Exploratory cross-dataset assessment of candidate genes in GSE23339. (A) Expression levels of *COL6A1*, *ERN2*, and *IL34* in ovarian endometrioma and control samples from GSE23339. *ERN2* and *IL34* showed between-group differences, whereas *COL6A1* did not. (B,C) ROC curve analyses of *ERN2* (B) and *IL34* (C). ROC, receiver operating characteristic; AUC, area under the curve. Symbols shown in panel (A): ns, $p \geq 0.05$; **, $p < 0.01$; ***, $p < 0.001$.

tion cannot establish monocyte-to-macrophage transition, macrophage polarization, or causality.

Exploratory Compound Prioritization Based on Transcriptomic Similarity

L1000CDS2 identified the top 50 small-molecule perturbation profiles showing transcriptomic similarity to the GSE51981 eutopic-endometrial EMs signature. After integration and de-duplication, three compounds with the highest candidate-gene matching coverage and L1000CDS2 matching scores were prioritized: BRD-K25737009, ingenol 3,20-dibenzoate, and phorbol 12-myristate 13-acetate (PMA). The compound-gene heatmap showed concordant transcriptomic matching patterns between the prioritized candidate compounds and module genes (Fig. 12A–D). The three prioritized compounds were carried forward to independent exploratory molecular docking analysis with *COL6A1*. These findings indicate transcriptomic similarity only and should not be interpreted as therapeutic predictions or evidence of drug efficacy.

Independent Exploratory Structural Analysis of Prioritized Compounds With COL6A1

As an independent exploratory structural analysis, *COL6A1* was docked with the three prioritized compounds. This analysis was not evidence that any compound directly targets *COL6A1*:

(1) BRD-K25737009 exhibited the most favorable Vina score (–7.2 kcal/mol) among the three tested compounds, with predicted contacts involving residues including LEU52, LYS53, PRO54, ASP60, ARG68, TRP218, GLN220, and ARG222 (Fig. 13B). Because no experimentally validated *COL6A1* ligand or positive-control inhibitor was included, this result should be interpreted as relative computational prioritization rather than evidence of binding specificity.

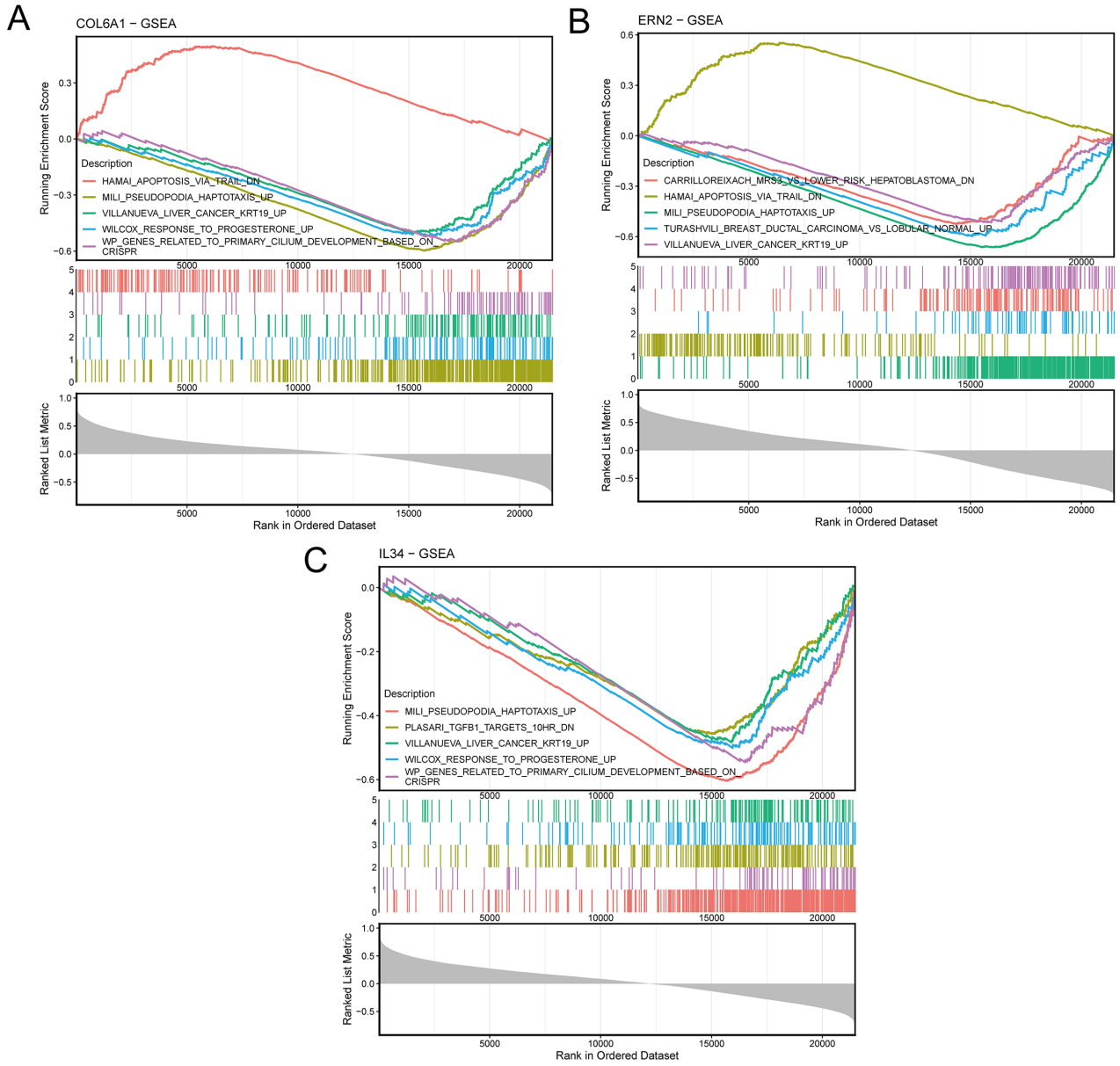

Fig. 10. Single-Gene GSEA. (A–C) Single-Gene gene set enrichment analysis (single-gene GSEA) results of Module Genes *COL6A1* (A), *ERN2* (B), and *IL34* (C) in GSE51981. GSEA, Gene Set Enrichment Analysis.

(2) Ingenol 3,20-dibenzoate showed a Vina score of –7.0 kcal/mol, with predicted docking contacts involving multiple *COL6A1* residues (Fig. 13C);

(3) PMA showed a Vina score of –5.3 kcal/mol, indicating a less favorable predicted docking interaction with fewer contacts (Fig. 14). The Vina docking-score heatmap ranked BRD-K25737009 as the top-scoring ligand among the three compounds (Fig. 13A). Among the three compounds, BRD-K25737009 yielded the most favorable docking score (−7.2 kcal/mol) and was therefore selected for subsequent exploratory molecular dynamics simulation.

Exploratory 50 ns Molecular Dynamics Simulation of the Modeled COL6A1-BRD-K25737009 Complex

The *COL6A1*-BRD-K25737009 complex was subjected to a 50 ns MD simulation. Four trajectory metrics indicated a relatively stable modeled complex conformation over the simulated interval (Fig. 15):

(1) RMSD curve: Fluctuation amplitude remained limited during the later stage of the simulation, the system reached dynamic equilibrium without sustained structural drift;

(2) RMSF curve: Terminal regions and residues 174–185 showed relatively high fluctuations, whereas most other residues showed lower RMSF values, suggesting that the overall complex structure remained relatively stable;

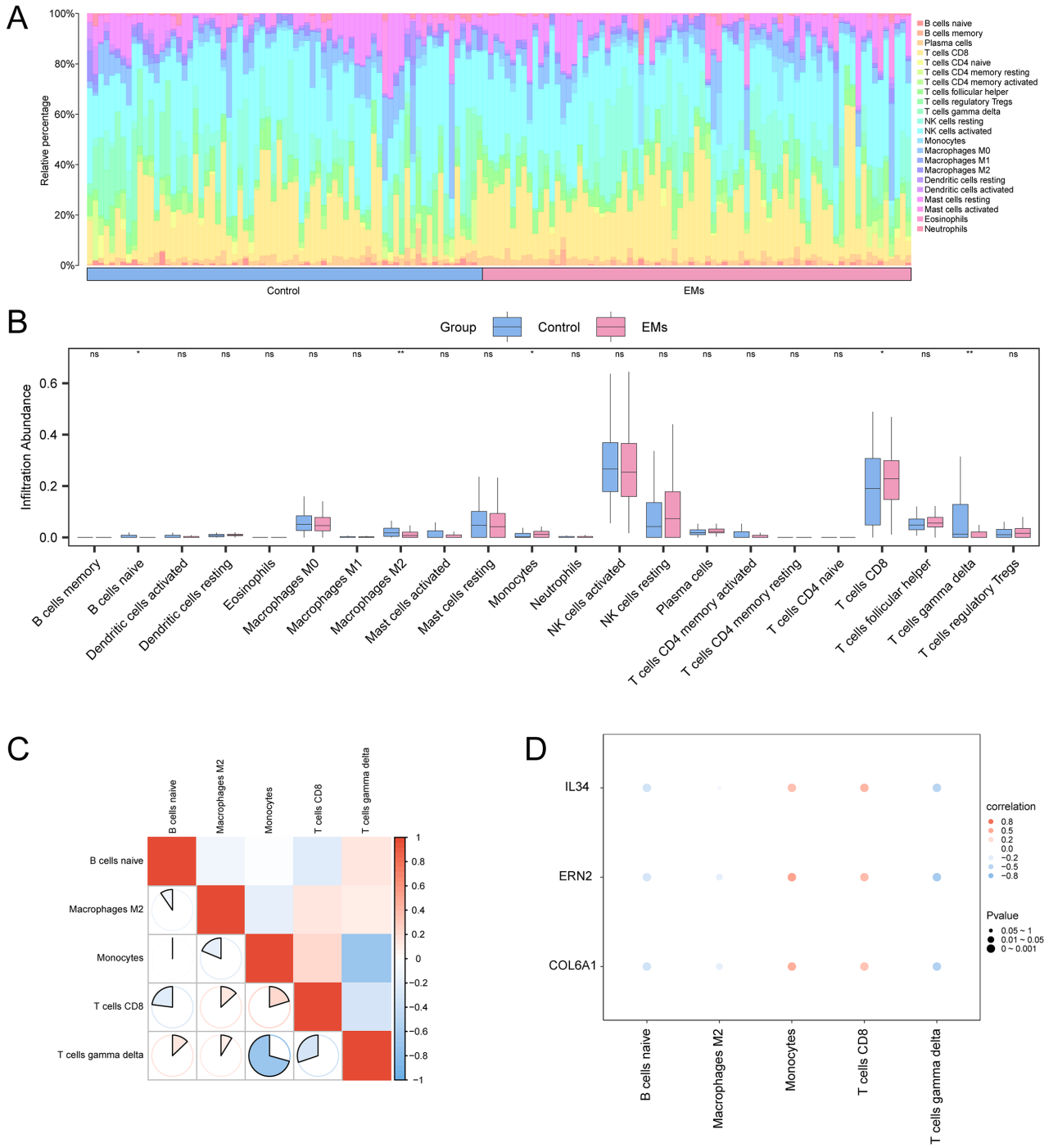

Fig. 11. Immune infiltration profiling of GSE51981 via CIBERSORT. (A) Bar chart of inferred immune-cell proportions per sample. (B) Exploratory group comparisons of inferred immune-cell fractions. (C) Correlation matrix of immune-cell subsets. (D) Bubble plot of nominal correlations between candidate-gene expression and inferred cell fractions. *p* values in panel (B) are unadjusted for multiple testing: ns, $p \geq 0.05$; *, $p < 0.05$; **, $p < 0.01$. Correlation intensity by |r|: <0.3 (negligible), 0.3–0.5 (weak), 0.5–0.8 (moderate), ≥0.8 (strong). Red and blue denote positive and negative correlations, respectively.

(3) Radius of gyration (Rg): Minor overall fluctuation, protein tertiary structure remained compact without large-scale unfolding;

(4) SASA curve: Fluctuated steadily between 132.2–141.4 nm^2, maintaining stable solvent contact surface without drastic conformational rearrangement.

Collectively, the 50-ns MD trajectory suggested that the modeled BRD-K25737009–*COL6A1* complex maintained a relatively stable conformation during the simulated interval. However, this simulation samples only a limited conformational timescale and does not establish long-term binding stability or biological activity.

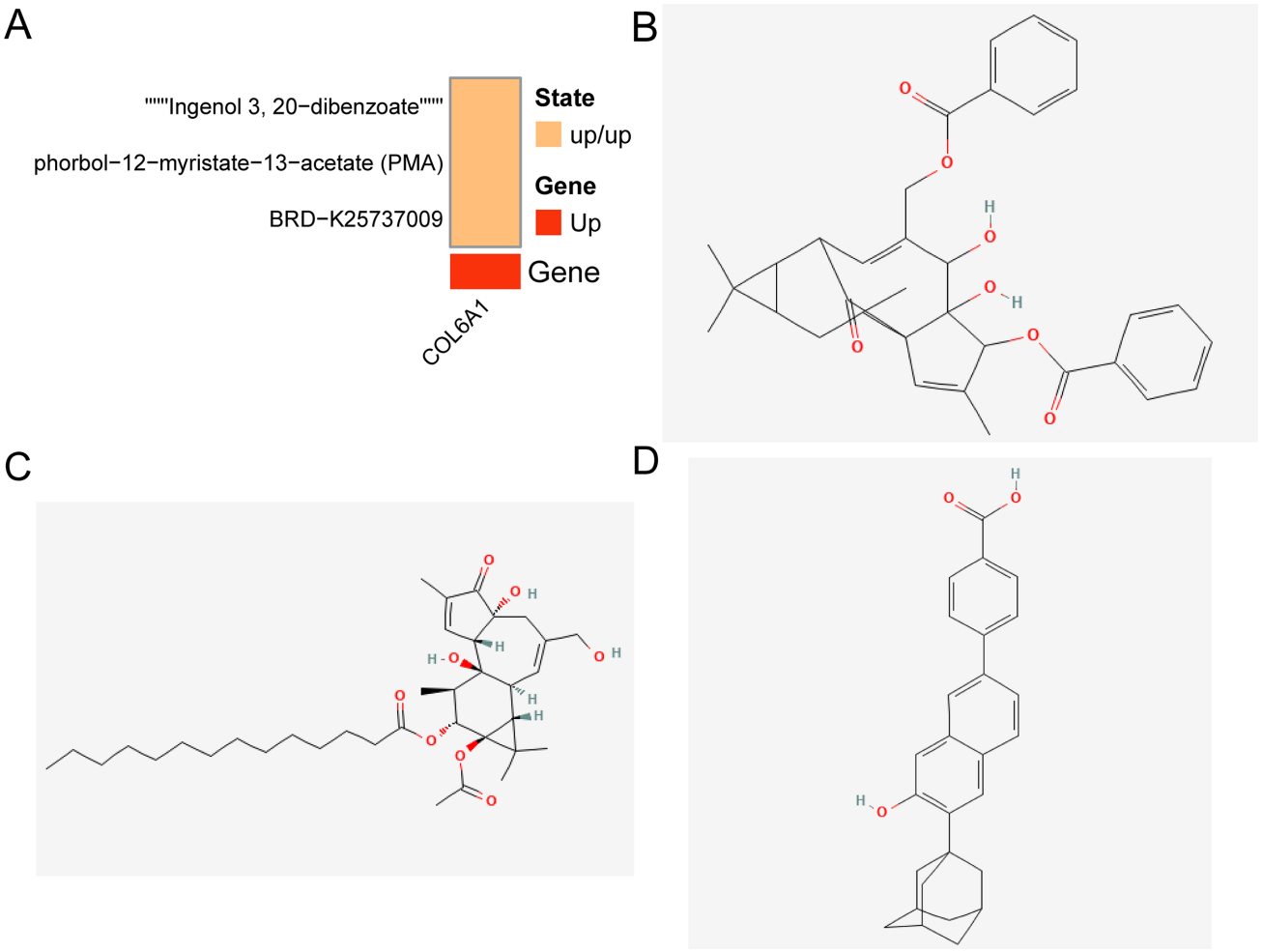

Fig. 12. Exploratory compound prioritization based on L1000CDS2 transcriptomic matching. (A) Heatmap showing predicted transcriptomic matching between L1000CDS2-prioritized small-molecule perturbations and candidate module genes. Red indicates concordant upregulation (up/up), green indicates concordant downregulation (dn/dn), and gray indicates no predicted match. The lower color bar shows the differential-expression direction of module genes in the GSE51981 eutopic-endometrial comparison. (B–D) Two-dimensional chemical structures of ingenol 3,20-dibenzoate (B), phorbol 12-myristate 13-acetate (PMA) (C), and BRD-K25737009 (D). The three prioritized compounds were carried forward to exploratory molecular docking analysis. BRD-K25737009 was subsequently selected for molecular dynamics simulation based on its docking score. These transcriptomic matching results are exploratory and do not imply therapeutic efficacy.

Discussion

A central limitation in interpreting this study is the tissue source of the discovery cohort. GSE51981 represents eutopic endometrium from patients with endometriosis rather than ectopic endometriotic lesions. Therefore, the angiogenesis-associated transcriptional patterns, enrichment results, and CIBERSORT-derived immune-cell estimates obtained from GSE51981 should be interpreted as molecular features of eutopic endometrium. They do not provide direct evidence of angiogenesis, vascular remodeling, or the local immune microenvironment within ectopic lesions. Within this defined tissue context, our integrated analyses identified 238 ARDEGs enriched in vascular-development, extracellular-matrix, inflammatory, and immune-related pathways, supporting an association among these transcriptional programs in eutopic endometrium.

COL6A1, *ERN2*, and *IL34* were prioritized through the combined machine-learning and WGCNA workflow, but the level of evidence was not equivalent across the three genes. *ERN2* and *IL34* showed expression differences in the exploratory GSE23339 assessment, whereas *COL6A1* did not and should therefore be regarded as a discovery-stage candidate. Previous literature provides biological context for these genes: *COL6A1* encodes a principal component of collagen VI, an extracellular-matrix protein involved in matrix organization, cell-matrix interactions, and tissue homeostasis [23]; *ERN2* encodes IRE1beta, an endoplasmic-reticulum-resident sensor in-

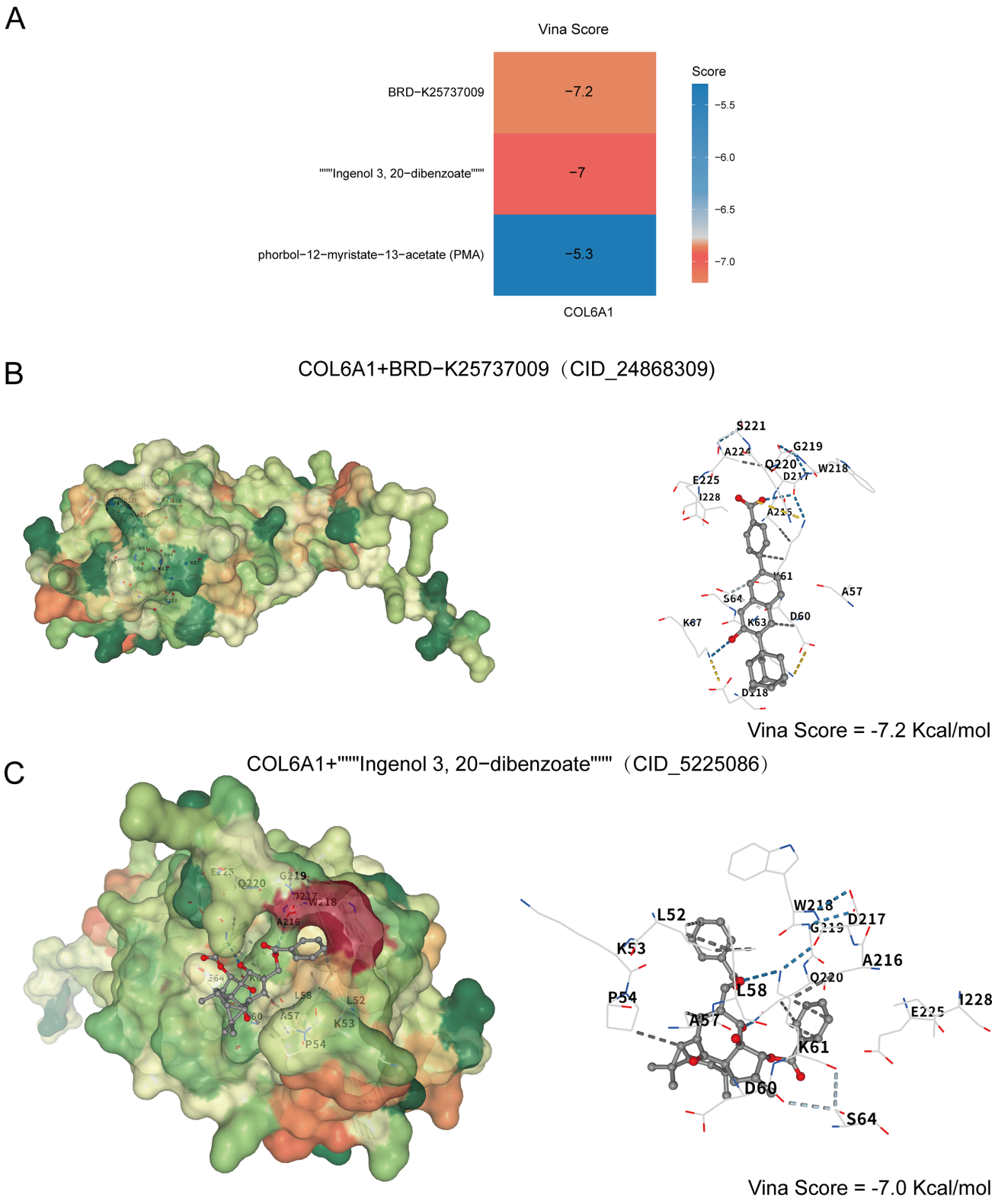

Fig. 13. Molecular docking of COL6A1 with BRD-K25737009 and ingenol 3,20-dibenzoate. (A) Vina docking scores of *COL6A1* with candidate compounds. (B,C) Visualization of the docking results of *COL6A1* with BRD-K25737009 (B) and ingenol 3,20-dibenzoate (C). For each panel, the global docking view and the interaction map are shown from left to right. The protein surface color changes from green to orange to red, indicating increasing hydrophobicity. Dashed blue lines indicate hydrogen bonds, dashed light blue lines indicate weak hydrogen bonds, dashed gray lines indicate hydrophobic interactions, dashed cyan lines indicate halogen bonds, dashed yellow lines indicate ionic bonds, dashed orange lines indicate cation–π interactions, and dashed green lines indicate π–π interactions.

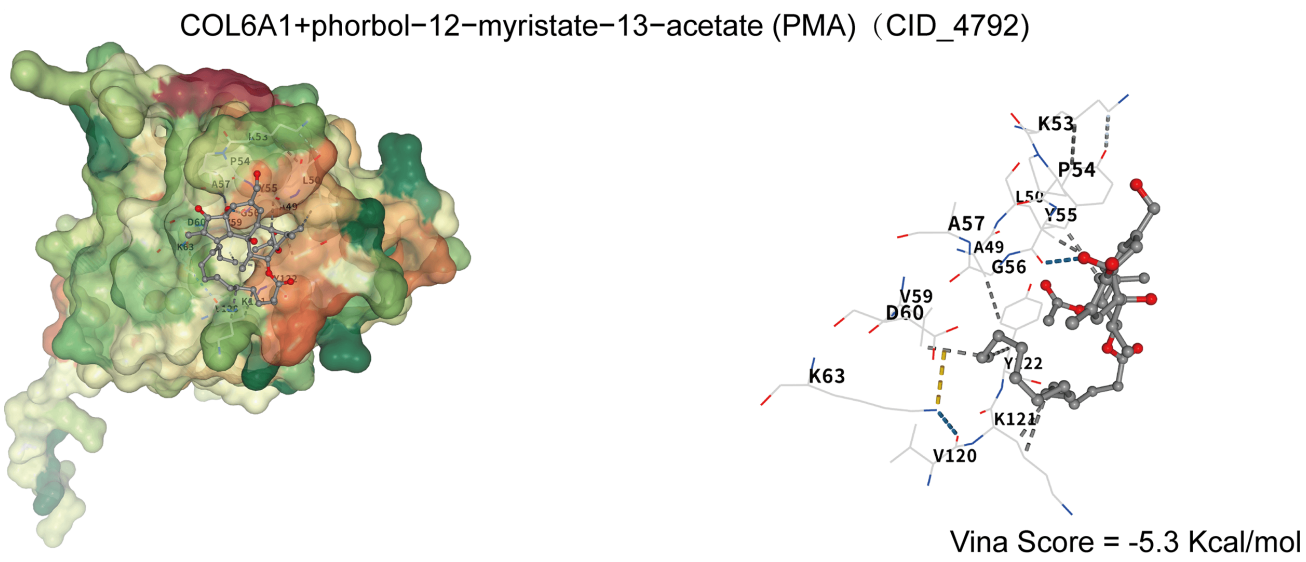

Fig. 14. Molecular docking of *COL6A1* with phorbol-12-myristate-13-acetate (PMA). Visualization of the docking results of *COL6A1* with phorbol 12-myristate 13-acetate (PMA). The global docking view and the interaction map are shown from left to right, respectively. The protein surface color changes from green to orange to red, indicating increasing hydrophobicity. Dashed blue lines indicate hydrogen bonds, dashed light blue lines indicate weak hydrogen bonds, dashed gray lines indicate hydrophobic interactions, dashed cyan lines indicate halogen bonds, dashed yellow lines indicate ionic bonds, dashed orange lines indicate cation–π interactions, and dashed green lines indicate π–π interactions.

volved in epithelial stress responses and regulation of the unfolded-protein response [24,25]; and *IL34* is involved in monocyte/macrophage survival, differentiation, and immune regulation [26,27,28]. The lack of cross-dataset support for *COL6A1* may reflect the small comparison cohort, tissue heterogeneity, platform-related variation, or biological heterogeneity. Nevertheless, *COL6A1* was selected for independent exploratory structural analysis because it linked the discovery-stage signature to extracellular-matrix and angiogenesis-related biology. This biological rationale did not establish a direct relationship between *COL6A1* and the L1000CDS2-prioritized compounds. Supporting literature is summarized in **Supplementary Table 4** [24,29,30,31,32]. Because GSE23339 compares ovarian endometrioma with normal endometrial tissue, the observed expression differences may be influenced by tissue-source differences. Therefore, these findings should be interpreted as exploratory cross-tissue evidence rather than direct replication of the eutopic endometrial findings.

From a classification perspective, the RF model based on the 14 machine-learning-selected genes yielded an AUC of 0.806 in GSE23339. The code confirms repeated five-fold cross-validation with five repeats, ROC optimization, random tuning with tuneLength = 5, class-probability output, and training-derived centering and scaling. However, GSE23339 was also used to compare RF, RPART, and SVM radial and thus contributed to model selection; the estimate is not independent validation and may be optimistic. Additionally, feature selection was performed using the full discovery cohort before classifier cross-validation, which may introduce selection bias and lead to optimistic estimates of internal model performance. Moreover, this cohort was small, cross-tissue, and cross-platform. The result therefore indicates only exploratory cross-dataset discrimination. Comparable integrative WGCNA-machine-learning frameworks have been reported in endometriosis cohorts [33,34,35]. The apparently high single-gene AUC for *ERN2* (0.956) likewise requires caution. Larger prospective, tissue-matched cohorts with prespecified models and locked preprocessing are required before diagnostic utility can be assessed. L1000CDS2 prioritized compounds solely from transcriptomic matching; subsequent docking and molecular dynamics constituted an independent exploratory structural analysis of *COL6A1* rather than evidence of direct compound targeting.

These analyses do not predict treatment response, demonstrate reversal of endometriosis biology, or establish drug efficacy. In particular, BRD-K25737009 is not an established therapy for endometriosis; no positive-control *COL6A1* ligand was included; the *COL6A1* structure was computationally predicted; and the 50-ns trajectory provides only limited conformational sampling. Accordingly, this part of the study should be viewed strictly as hypothesis-generating computational prioritization.

Several limitations should be considered. First, GSE51981 consists of eutopic endometrium; its findings cannot be directly extrapolated to ectopic-lesion angiogenesis or immunity. Second, GSE23339 was small, cross-tissue, and cross-platform; it was also used to compare classifier families before RF performance was reported, so it

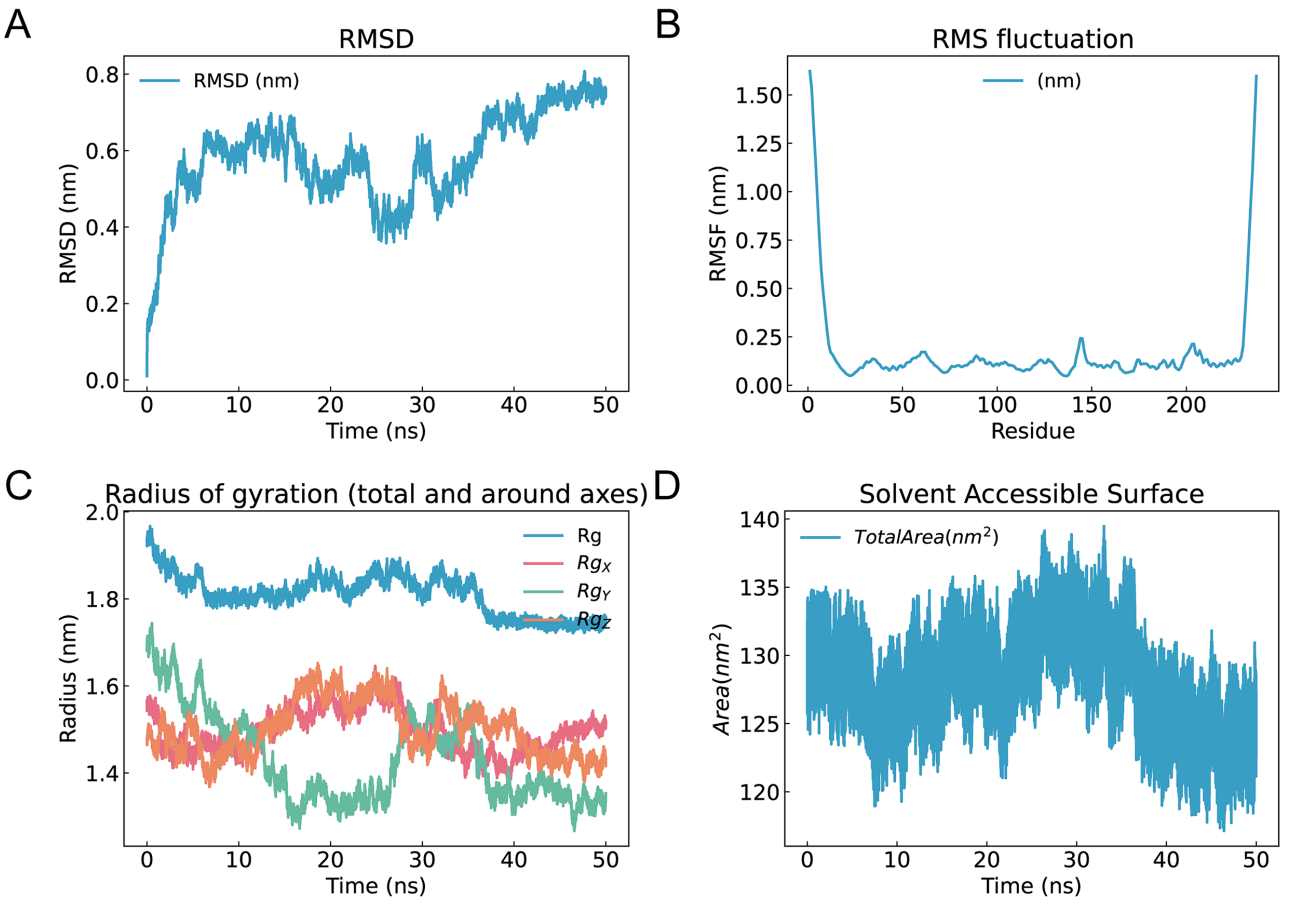

Fig. 15. Stability and conformational dynamics of the *COL6A1*–BRD-K25737009 complex over 50 ns MD simulation. (A) RMSD trajectory and (B) per-residue RMSF map, evaluating global convergence and local flexibility, respectively. (C) Rg profile and (D) SASA time series, reflecting compactness and solvent exposure of the protein structure. RMSD, root mean square deviation; RMSF, root mean square fluctuation; Rg, radius of gyration; SASA, solvent-accessible surface area.

was not an independent test set. Although predictor centering and scaling were estimated from GSE51981 and applied to GSE23339, this did not remove platform or tissue effects. The code used caret random search with tuneLength = 5, but the saved model object and model$results output were unavailable; therefore, the exact sampled tuning grids and final hyperparameters could not be recovered, limiting exact reproducibility. Third, *COL6A1* lacked cross-dataset expression support, whereas *ERN2* and *IL34* showed stronger exploratory support. Fourth, immune-cell comparisons and gene-cell correlations used unadjusted nominal p values across multiple tests and may include false-positive findings. Fifth, the candidates lack qPCR, immunohistochemical, cellular, or animal validation. Finally, L1000CDS2 matching, docking, and 50-ns MD are separate computational analyses; they neither establish that a prioritized compound directly targets *COL6A1* nor demonstrate therapeutic efficacy. Prospective studies using tissue-matched sampling, locked model specifications, multiplicity-aware analyses, and functional experiments are required [36].

Despite these limitations, the study provides a hypothesis-generating framework for angiogenesis-associated transcriptional alterations in eutopic endometrium. The strongest exploratory cross-dataset support was observed for *ERN2* and *IL34*, whereas *COL6A1* remains a discovery-stage candidate. The classifier, immune-deconvolution, compound-matching, docking, and molecular dynamics findings require independent confirmation.

Conclusions

Integrated transcriptomic, machine-learning, and co-expression analyses identified angiogenesis-associated candidate signatures in the eutopic endometrium of patients with endometriosis, with stronger exploratory cross-dataset support for *ERN2* and *IL34* than for *COL6A1*. The RF result in GSE23339 represents selection-influenced exploratory cross-dataset discrimination rather than independent validation, and the unadjusted immune analyses and independent *COL6A1* structural analyses remain hypothesis-generating.

Larger tissue-matched cohorts with prespecified preprocessing and models, multiplicity-aware analyses, and functional experiments are needed before diagnostic or therapeutic conclusions can be drawn.

Availability of Data and Materials

Publicly available datasets GSE51981 and GSE23339 were analyzed in this study. These raw datasets can be accessed from the GEO database. Further processed datasets generated during this study are available from the corresponding author on reasonable request.

Author Contributions

Conceptualization, XMY, XJY, OJ; Methodology, XMY; Software, XMY; Formal Analysis, XMY; Resources, OJ, XMY; Data Curation, XMY; Writing — Original Draft, XMY; Writing - Review & Editing, XMY, OJ, XJY; Visualization, XMY; Supervision, XJY; Project Administration, XMY. All authors read and approved the final manuscript. All authors have participated sufficiently in the work to take public responsibility for appropriate portions of the content and agreed to be accountable for all aspects of the work in ensuring that questions related to its accuracy or integrity.

Ethics Approval and Consent to Participate

Not applicable.

Acknowledgment

Not applicable.

Funding

This work was supported by the Key Discipline Construction Fund for Endometriosis of Jiaxing Maternity and Child Health Care Hospital (no grant number assigned) and the Health Science and Technology Program of Zhejiang Province, funded by the Health Commission of Zhejiang Province (Grant No. 2025HY1231).

Conflict of Interest

The authors declare no conflict of interest.

Supplementary Material

Supplementary material associated with this article can be found, in the online version, at https://doi.org/10.24976/Discov.Med.202638214.277.

References

[1] Liang L, Min L, Liu J, Liu Y, Cheng W. Gut microbiota dysbiosis in endometriosis: mechanistic insights and gut microbiota-targeted therapeutic strategies. Frontiers in Microbiology. 2026; 17: 1776574. https://doi.org/10.3389/fmicb.2026.1776574

[2] Nanda A, K T, Banerjee P, Dutta M, Wangdi T, Sharma P, et al. Cytokines, Angiogenesis, and Extracellular Matrix Degradation are Augmented by Oxidative Stress in Endometriosis. Annals of Laboratory Medicine. 2020; 40: 390–397. https://doi.org/10.3343/alm.2020.40.5.390

[3] Zou X, Wang X, Xiong B, Wang M, Yang D, Lin Z, et al. Deciphering immune-inflammatory dysregulation in the endometriotic microenvironment: insights from single-cell omics and artificial intelligence. Frontiers in Immunology. 2026; 17: 1848643. https://doi.org/10.3389/fimmu.2026.1848643

[4] Chen S, Liu Y, Zhong Z, Wei C, Liu Y, Zhu X. Peritoneal immune microenvironment of endometriosis: Role and therapeutic perspectives. Frontiers in Immunology. 2023; 14: 1134663. https://doi.org/10.3389/fimmu.2023.1134663

[5] Wu H, Wang W, Zhu W, Ma S. Gut Microbiota and Inflammatory Profiles in Endometriosis: A Retrospective Study Based on 16S rRNA Sequencing. Clinical and Experimental Obstetrics & Gynecology. 2025; 52: 44257. https://doi.org/10.31083/CEOG44257

[6] Symons LK, Miller JE, Kay VR, Marks RM, Liblik K, Koti M, et al. The Immunopathophysiology of Endometriosis. Trends in Molecular Medicine. 2018; 24: 748–762. https://doi.org/10.1016/j.molmed.2018.07.004

[7] Neagu AN, Whitham D, Bruno P, Morrissiey H, Darie CA, Darie CC. Omics-Based Investigations of Breast Cancer. Molecules (Basel, Switzerland). 2023; 28: 4768. https://doi.org/10.3390/molecules28124768

[8] Yao M, Wang R, Cui R, Lou F, Chen Z. Identification of diabetes-related signatures as prognostic and therapeutic biomarkers in colon cancer. Discover Oncology. 2025; 16: 1604. https://doi.org/10.1007/s12672-025-03476-1

[9] Tamaresis JS, Irwin JC, Goldfien GA, Rabban JT, Burney RO, Nezhat C, et al. Molecular classification of endometriosis and disease stage using high-dimensional genomic data. Endocrinology. 2014; 155: 4986–4999. https://doi.org/10.1210/en.2014-1490

[10] Hawkins SM, Creighton CJ, Han DY, Zariff A, Anderson ML, Gunaratne PH, et al. Functional microRNA involved in endometriosis. Molecular Endocrinology. 2011; 25: 821–832. https://doi.org/10.1210/me.2010-0371

[11] Hu R, Chen F, Yu X, Li Z, Li Y, Feng S, et al. Construction and validation of a prognostic model of angiogenesis-related genes in multiple myeloma. BMC Cancer. 2024; 24: 1269. https://doi.org/10.1186/s12885-024-13024-9

[12] Li S, Gao K, Yao D. Comprehensive Analysis of angiogenesis associated genes and tumor microenvironment infiltration characterization in cervical cancer. Heliyon. 2024; 10: e33277. https://doi.org/10.1016/j.heliyon.2024.e33277

[13] Zhao Y, Li H, Ma X, Meng X, Tang Q. Identification of biomarkers related to angiogenesis in myocardial ischemia-reperfusion injury and prediction of potential drugs. PloS One. 2024; 19: e0300790. https://doi.org/10.1371/journal.pone.0300790

[14] Chen B, Han Y, Sheng S, Deng J, Vasquez E, Yau V, et al. An angiogenesis-associated gene-based signature predicting prognosis and immunotherapy efficacy of head and neck squamous cell carcinoma patients. Journal of Cancer Research and Clinical Oncology. 2024; 150: 91. https://doi.org/10.1007/s00432-024-05606-8

[15] Kang J, Jiang J, Xiang X, Zhang Y, Tang J, Li L. Identification of a new gene signature for prognostic evaluation in cervical cancer: based on cuproptosis-associated angiogenesis and multi-omics analysis. Cancer Cell International. 2024; 24: 23. https://doi.org/10.1186/s12935-023-03189-x

[16] Zheng C, Chen X, Ke Y, Xu X, Wu C, Jiang L. Constructing models for Crohn's disease diagnosis and prediction of infliximab non-response based on angiogenesis-related genes. Frontiers in Immunology. 2024; 15: 1239496. https://doi.org/10.3389/fimmu.2024.1239496

[17] Wang J, Cui X, Weng Y, Wei J, Chen X, Wang P, et al. Application of an angiogenesis-related genes risk model in lung adenocarcinoma prognosis and immunotherapy. Frontiers in Genetics. 2023; 14: 1092968. https://doi.org/10.3389/fgene.2023.1092968

[18] Langfelder P, Horvath S. WGCNA: an R package for weighted correlation network analysis. BMC Bioinformatics. 2008; 9: 559. https://doi.org/10.1186/1471-2105-9-559

[19] Varadi M, Anyango S, Deshpande M, Nair S, Natassia C, Yordanova G, et al. AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. Nucleic Acids Research. 2022; 50: D439–D444. https://doi.org/10.1093/nar/gkab1061

[20] Trott O, Olson AJ. AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. Journal of Computational Chemistry. 2010; 31: 455–461. https://doi.org/10.1002/jcc.21334

[21] Liu Y, Yang X, Gan J, Chen S, Xiao ZX, Cao Y. CB-Dock2: improved protein-ligand blind docking by integrating cavity detection, docking and homologous template fitting. Nucleic Acids Research. 2022; 50: W159–W164. https://doi.org/10.1093/nar/gkac394

[22] Van Der Spoel D, Lindahl E, Hess B, Groenhof G, Mark AE, Berendsen HJC. GROMACS: fast, flexible, and free. Journal of Computational Chemistry. 2005; 26: 1701–1718. https://doi.org/10.1002/jcc.20291

[23] Lamandé SR, Bateman JF. Collagen VI disorders: Insights on form and function in the extracellular matrix and beyond. Matrix Biology: Journal of the International Society for Matrix Biology. 2018; 71-72: 348–367. https://doi.org/10.1016/j.matbio.2017.12.008

[24] Cloots E, Simpson MS, De Nolf C, Lencer WI, Janssens S, Grey MJ. Evolution and function of the epithelial cell-specific ER stress sensor IRE1β. Mucosal Immunology. 2021; 14: 1235–1246. https://doi.org/10.1038/s41385-021-00412-8

[25] García-López D, Zaragoza-Ojeda M, Eguía-Aguilar P, Arenas-Huertero F. Endoplasmic Reticulum Stress in Gliomas: Exploiting a Dual-Effect Dysfunction through Chemical Pharmaceutical Compounds and Natural Derivatives for Therapeutical Uses. International Journal of Molecular Sciences. 2024; 25: 4078. https://doi.org/10.3390/ijms25074078

[26] Hiltbrunner S, Mannarino L, Kirschner MB, Opitz I, Rigutto A, Laure A, et al. Tumor Immune Microenvironment and Genetic Alterations in Mesothelioma. Frontiers in Oncology. 2021; 11: 660039. https://doi.org/10.3389/fonc.2021.660039

[27] Bézie S, Freuchet A, Sérazin C, Salama A, Vimond N, Anegon I, et al. IL-34 Actions on FOXP3+ Tregs and CD14+ Monocytes Control Human Graft Rejection. Frontiers in Immunology. 2020; 11: 1496. https://doi.org/10.3389/fimmu.2020.01496

[28] Boulakirba S, Pfeifer A, Mhaidly R, Obba S, Goulard M, Schmitt T, et al. IL-34 and CSF-1 display an equivalent macrophage differentiation ability but a different polarization potential. Scientific Reports. 2018; 8: 256. https://doi.org/10.1038/s41598-017-18433-4

[29] Williams L, Layton T, Yang N, Feldmann M, Nanchahal J. Collagen VI as a driver and disease biomarker in human fibrosis. The FEBS Journal. 2022; 289: 3603–3629. https://doi.org/10.1111/febs.16039

[30] Cook CJ, Wiggin N, Fogg KC. Characterizing the Extracellular Matrix Transcriptome of Endometriosis. Reproductive Sciences (Thousand Oaks, Calif.). 2024; 31: 413–429. https://doi.org/10.1007/s43032-023-01359-w

[31] Lin K, Ma J, Peng Y, Sun M, Xu K, Wu R, et al. Autocrine Production of Interleukin-34 Promotes the Development of Endometriosis through CSF1R/JAK3/STAT6 signaling. Scientific Reports. 2019; 9: 16781. https://doi.org/10.1038/s41598-019-52741-1

[32] Kobayashi H, Imanaka S. Understanding the molecular mechanisms of macrophage polarization and metabolic reprogramming in endometriosis: A narrative review. Reproductive Medicine and Biology. 2022; 21: e12488. https://doi.org/10.1002/rmb2.12488

[33] Du J, Lv Z. Deciphering the Diagnostic and Natural Therapeutic Implications of Necrosis by Sodium Overload and NK Signatures in Endometriosis Patients. International Journal of Molecular Sciences. 2026; 27: 4535. https://doi.org/10.3390/ijms27104535

[34] Liang Q, Yang S, Mai M, Chen X, Zhu X. Mining phase separation-related diagnostic biomarkers for endometriosis through WGCNA and multiple machine learning techniques: a retrospective and nomogram study. Journal of Assisted Reproduction and Genetics. 2024; 41: 1433–1447. https://doi.org/10.1007/s10815-024-03079-9

[35] Jiang H, Zhang X, Wu Y, Zhang B, Wei J, Li J, et al. Bioinformatics identification and validation of biomarkers and infiltrating immune cells in endometriosis. Frontiers in Immunology. 2022; 13: 944683. https://doi.org/10.3389/fimmu.2022.944683

[36] Fassbender A, Rahmioglu N, Vitonis AF, Viganò P, Giudice LC, D'Hooghe TM, et al. World Endometriosis Research Foundation Endometriosis Phenome and Biobanking Harmonisation Project: IV. Tissue collection, processing, and storage in endometriosis research. Fertility and Sterility. 2014; 102: 1244–1253. https://doi.org/10.1016/j.fertnstert.2014.07.1209