Research Square

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

Machine Learning-Based Identification of a Glycolysis- Related Prognostic Signature and CENPA in Endometrial Cancer

Hang Yu
The First Affiliated Hospital of Harbin Medical University
Ze Liu
The First Affiliated Hospital of Harbin Medical University
Yanjun Liu
Affiliated Hospital of Qingdao University
Xiushuang Zheng
The First Affiliated Hospital of Harbin Medical University
Liqun Wang
202301238@hrbmu.edu.cn
The First Affiliated Hospital of Harbin Medical University

Research Article

Keywords: Endometrial cancer, Glycolysis, CENPA, Machine learning, Prognostic signature, Single-cell RNA sequencing

Posted Date: November 13th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11456652/v1

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

Additional Declarations: No competing interests reported.

Article

Machine Learning-Based Identification of a Glycolysis-Related Prognostic Signature and CENPA in Endometrial Cancer

Hang Yu [1], Ze Liu [2], Yajun Liu [3], Xiushuang Zheng [4], and Liqun Wang [5]*

1 The First Affiliated Hospital of Harbin Medical University; yuhang219@126.com
2 The First Affiliated Hospital of Harbin Medical University; liuzetaishuaila@163.com
3 Department of Obstetrics, Affiliated Hospital of Qingdao University;yajunliu2025@163.com
4 The First Affiliated Hospital of Harbin Medical University;900282@hrbmu.edu.cn
5 The First Affiliated Hospital of Harbin Medical University; 202301238@hrbmu.edu.cn
* Correspondence: 202301238@hrbmu.edu.cn
Department of Obstetrics and Gynecology

Abstract

Background: Endometrial cancer (EC) is strongly associated with metabolic disorders, while glycolytic reprogramming contributes to tumor progression and adverse outcomes. This study aimed to identify glycolysis-related prognostic genes and develop a machine learning-based signature for EC risk stratification.**Methods:** Transcriptomic and clinical data were obtained from The Cancer Genome Atlas. Glycolysis-related genes were collected from KEGG, MSigDB, and Reactome. Differential expression and univariate Cox analyses were used to identify prognostic candidates. Multiple machine learning algorithms and combinations were evaluated to construct a prognostic model. Functional enrichment, immune infiltration, somatic mutation, tumor mutational burden, and drug sensitivity analyses were performed. Single-cell RNA sequencing datasets GSE179640 and GSE173682 were analyzed to characterize CENPA expression across cell types and its clinical relevance.**Results:** Among 5,990 differentially expressed genes, 99 overlapped with 290 glycolysis-related genes, and 16 were associated with overall survival. Among 117 model combinations, random survival forest plus generalized boosted model generated a seven-gene signature comprising CLDN9, CENPA, GPI, TFF3, FBP1, SPAG4, and AURKA. The 1-, 3-, and 5-year AUCs were 0.88, 0.81, and 0.84 in the training cohort and 0.73, 0.77, and 0.76 in the internal validation cohort, respectively. High-risk patients had poorer survival and distinct pathway, immune, and genomic features. Single-cell analysis showed heterogeneous CENPA expression, while elevated CENPA was associated with advanced stage and poorer survival.**Conclusions:** The glycolysis-related signature stratified EC patients by survival risk and was associated with distinct biological features. CENPA emerged as a candidate prognostic biomarker requiring independent external validation and experimental investigation.

Academic Editor: Firstname Lastname

Received: date
Revised: date
Accepted: date
Published: date

Keywords: Endometrial cancer; Glycolysis; CENPA; Machine learning; Prognostic signature; Single-cell RNA sequencing;

1. Introduction

EC is one of the most common gynecological malignancies in women, and its incidence and mortality have continued to increase in recent years[1], particularly in developed countries[2]. According to the GLOBOCAN database of the International Agency for Research on Cancer (IARC), approximately 420,242 new cases of EC and 97,704 EC-related deaths were reported worldwide in 2022[3]. Although patients with early-stage

EC generally have a favorable prognosis following surgical treatment, and considerable advances have been made in radiotherapy, chemotherapy, endocrine therapy, and immunotherapy, clinical outcomes remain unsatisfactory for patients with advanced, recurrent, or metastatic disease[4, 5]. Therefore, elucidating the molecular mechanisms underlying EC initiation and progression and identifying biomarkers and therapeutic targets with potential for clinical translation are of considerable importance.

Metabolic reprogramming is a hallmark of cancer, and aerobic glycolysis plays a pivotal role in sustaining tumor cell proliferation and malignant progression. Glycolysis not only provides a rapid source of energy but also generates metabolic intermediates required for the biosynthesis of nucleotides, lipids, and other macromolecules[6]. Moreover, glycolytic reprogramming can promote tumor progression through redox regulation, epigenetic modifications, and remodeling of the tumor microenvironment. Accumulating evidence suggests that aberrant glycolysis is closely associated with tumor invasion, therapeutic resistance, and poor prognosis in gynecological malignancies[7, 8], while glucose metabolic reprogramming represents one of the prominent metabolic alterations in EC.

Recent studies have further highlighted the critical role of glycolysis in EC development and progression. Metabolic parameters derived from 18F-fluorodeoxyglucose positron emission tomography/magnetic resonance imaging (18F-FDG PET/MRI) can reflect tumor metabolic activity in EC and have been associated with tumor risk and aggressive clinicopathological features, including lymphovascular space invasion[9]. At the molecular level, several regulatory networks, including E2F1/TROAP,STAT3/LDHB/MDH2,FBXW7/ETV6/GLUT1,and IGF2BP3/MUC1/HIF-1α/ENO1, have been reported to promote EC progression by enhancing glycolysis or reprogramming glucose metabolism[10-12]. In addition, lactate, a major product of glycolysis, can induce M2 macrophage polarization and contribute to the establishment of an immunosuppressive tumor microenvironment, thereby further facilitating EC progression[13]. Collectively, these findings indicate that glycolytic reprogramming not only contributes to the malignant phenotype of EC cells but also plays an important role in shaping the tumor microenvironment.

Based on these observations, targeting glucose metabolism and its regulatory networks has emerged as a potential therapeutic strategy for cancer. A variety of glycolysis-related molecules, including HK2, PFK1, LDHA, PK, G6PD, GLUT, and HIF-1α, have been extensively investigated and have demonstrated potential as therapeutic targets[14]. However, current studies in EC have largely focused on individual metabolic enzymes or specific signaling pathways, whereas the systematic integration of glycolysis-related genes for prognostic assessment remains limited. In particular, robust glycolysis-related molecular signatures for prognostic risk stratification in EC have yet to be fully established.

Therefore, in the present study, we sought to systematically identify glycolysis-related genes associated with the prognosis of EC and to develop a prognostic model by integrating multiple machine learning algorithms.We further characterized the biological and immune features associated with different risk groups and investigated key candidate genes using single-cell transcriptomic and clinicopathological analyses. Through these analyses, we aimed to establish a robust glycolysis-related prognostic signature for risk stratification, identify potential prognostic biomarkers, and provide further insights into glycolytic reprogramming and its potential clinical implications in EC.

2. Materials and Methods

2.1. Data Sources and Processing

Transcriptomic data and corresponding clinical information used in this study were obtained from TCGA database. Transcriptomic profiles of EC, comprising 539 EC tissue samples and 35 normal endometrial tissue samples, were downloaded for subsequent analyses. Glycolysis-related genes were collected from the KEGG, the Hallmark gene sets in the MSigDB, and the Reactome database. After integration and removal of duplicate genes, a total of 290 glycolysis-related genes were retained for further analysis. In addition, the GSE179640 and GSE173682 datasets were obtained from the Gene Expression Omnibus (GEO) database for scRNA-seq analysis and characterization of the cell type-specific expression patterns of key genes.

2.2. Differential Expression Analysis

Differential expression analysis between EC tissues and normal endometrial tissues was performed using the edgeR package in R. DEGs were identified based on fold change (FC) and false discovery rate(FDR), with the significance thresholds set at $|\log2FC| > 1$ and FDR < 0.05. The EC-related DEGs were then intersected with the 290 glycolysis-related genes to identify differentially expressed glycolysis-related genes in EC. These genes were subsequently used for prognostic analysis and machine learning-based model construction.

2.3. Identification of Prognostic Glycolysis-Related Genes

Clinical follow-up data from the TCGA(The Cancer Genome Atlas) cohort were integrated with gene expression profiles. Univariate Cox proportional hazards regression analysis was performed using the survival package in R to assess the associations between differentially expressed glycolysis-related genes and OS in patients with EC. Genes with $p < 0.05$ were considered significantly associated with OS and were subsequently included as candidate prognostic genes for machine learning-based model construction.

2.4. Construction and Validation of the Machine Learning-Based Prognostic Model

Based on the identified prognostic glycolysis-related genes, multiple machine learning algorithms were employed to construct prognostic models for EC, including CoxBoost, elastic net (Enet), survival support vector machine (survival-SVM), least absolute shrinkage and selection operator (LASSO), partial least squares regression for Cox (plsRcox), ridge regression, random survival forest (RSF), stepwise Cox regression (StepCox), supervised principal components (SuperPC), and generalized boosted regression model (GBM). Patients with EC in the TCGA cohort were randomly divided into a training cohort and an internal validation cohort at a ratio of 8:2. The training cohort was used for model development and parameter optimization, whereas the internal validation cohort was used to assess model generalizability.

Using the Mime package in R, the aforementioned machine learning algorithms were integrated in various combinations, resulting in a total of 117 candidate models[15]. Model training and hyperparameter optimization were performed in the training cohort using cross-validation, and model performance was subsequently evaluated in the internal validation cohort. The concordance index (C-index) was used to compare the overall discriminatory performance of the 117 candidate models, and the model showing favorable performance across the training and internal validation cohorts was selected for subsequent analysis. Time-dependent receiver operating characteristic (ROC) curve analysis was further performed to evaluate the predictive performance of the selected model at predefined time points.

Candidate genes were further ranked and selected according to their variable importance scores derived from the optimal model, ultimately identifying the key genes constituting the glycolysis-related prognostic signature for EC. Individual risk scores were calculated based on the optimal model, and patients were stratified into high- and low-risk groups using the median risk score derived from the training cohort as the cutoff. The

association between the resulting prognostic signature and patient survival outcomes was subsequently evaluated.

2.5. Functional Enrichment Analysis

To investigate the biological processes(BP) and molecular pathways potentially associated with the glycolysis-related prognostic signature, Gene Ontology (GO) and KEGG pathway enrichment analyses were performed using the cluster Profiler package in R[16].GO enrichment analysis encompassed three categories: BP, cellular component (CC), and molecular function (MF). In addition, gene set enrichment analysis (GSEA) was performed to characterize differences in biological pathways between the high- and low-risk groups, using the Hallmark gene sets from the MSigDB as the reference gene set. Pathways or gene sets with $p < 0.05$ were considered statistically significant.

2.6. Immune Cell Infiltration Analysis

The CIBERSORT algorithm was used to characterize immune cell infiltration in EC tissues and to compare differences in the immune microenvironment between the high- and low-risk groups[17]. CIBERSORT was applied to estimate the relative abundances of 22 immune cell subsets, including different subtypes of T cells, B cells, and other immune cell populations. Furthermore, the associations between risk scores and the infiltration levels of individual immune cell types were assessed to explore the potential relationship between the glycolysis-related prognostic signature and the immune microenvironment of EC.

2.7. Somatic Mutation and TMB Analysis

Somatic mutation data in Mutation Annotation Format (MAF) for patients with uterine corpus endometrial carcinoma (UCEC) were obtained from the TCGA database. TMB was calculated as the number of nonsynonymous somatic mutations per megabase of the coding region[18]. The maftools package in R was used to summarize, analyze, and visualize the somatic mutation profiles[19]. Oncoplots were generated separately for the high- and low-risk groups to characterize and compare the distributions of frequently mutated genes between the two groups.

2.8. Drug Sensitivity Prediction

Potential sensitivity to candidate anticancer agents in patients with EC was predicted based on data from the Genomics of Drug Sensitivity in Cancer (GDSC) database[20]. Drug sensitivity was predicted using the oncoPredict package in R based on pharmacogenomic data from the GDSC database.The half-maximal inhibitory concentration (IC50) of each drug was estimated based on the transcriptomic expression profiles of EC samples, with a lower predicted IC50 value indicating greater predicted drug sensitivity. Predicted IC50 values were compared between the high- and low-risk groups. In addition, Spearman correlation analysis was performed to evaluate the association between risk scores and predicted drug sensitivity.

2.9. Single-Cell RNA Sequencing Analysis

ScRNA-seq data from the GSE179640 and GSE173682 datasets were analyzed using the Seurat package in R. Data processing was performed according to a standard scRNA-seq analysis workflow, including quality control, normalization, identification of highly variable genes, dimensionality reduction, cell clustering, and visualization. Batch effects across different samples were corrected using the Harmony algorithm. Following dimensionality reduction and clustering, Uniform Manifold Approximation and Projection (UMAP) was used to visualize the cellular landscape. Cell types were annotated based on the expression of canonical marker genes with reference to the PanglaoDB and Cell Marker 2.0 databases. Subsequently, cell populations were further subjected to subpopulation analysis to identify distinct cellular subclusters.

2.10. Association of CENPA Expression with Clinicopathological Characteristics and Survival Outcomes

The associations between CENPA expression and clinicopathological characteristics were evaluated using data from the TCGA-UCEC cohort. The optimal cutoff value for CENPA expression was determined using the surv_cutpoint function in the survminer package based on OS data. Patients were subsequently stratified into high- and low-expression groups according to this cutoff. Kaplan–Meier survival curves were generated, and differences in OS between the two groups were evaluated using the log-rank test. Patients were subsequently stratified into high- and low-CENPA expression groups according to the optimal cutoff, and the association between CENPA expression and patient survival outcomes was evaluated.

2.11. Statistical Analysis

All statistical analyses were performed using R software. Differences between two groups were assessed using the Wilcoxon rank-sum test or Student's t-test, as appropriate. Correlations between continuous variables were evaluated using Spearman or Pearson correlation analysis, as appropriate. Survival curves were generated using the Kaplan–Meier method, and differences in survival between groups were assessed using the log-rank test. Cox proportional hazards regression analysis was performed to evaluate the associations between candidate genes and survival outcomes. Unless otherwise specified, all statistical tests were two-sided, and $p < 0.05$ was considered statistically significant.

3. Results

3.1. Identification of Differentially Expressed and Prognostic Glycolysis-Related Genes

Comparison of transcriptomic profiles between EC and normal endometrial tissues identified a total of 5,990 DEGs(Figure 1a,b). A total of 290 glycolysis-related genes were obtained by integrating gene sets from the KEGG, the Hallmark gene sets in the MSigDB, and the Reactome database, followed by removal of duplicate genes. Intersection of the 5,990 DEGs with the 290 glycolysis-related genes identified 99 differentially expressed glycolysis-related genes (Figure 1c).

Univariate Cox proportional hazards regression analysis further identified 16 candidate glycolysis-related genes significantly associated with OS in patients with EC ($p < 0.05$; Figure 1d). Comparison of the expression levels of these 16 prognostic glycolysis-related genes between EC and normal endometrial tissues revealed marked differences for most candidate genes (Figure 1e).

Among the 16 prognostic genes, 10 genes, including DEPDC1, LHX9, CENPA, HMMR, PFKM, IRS2, CLDN9, GPI, PSMC4, and AURKA, had hazard ratios (HRs) > 1, indicating that higher expression levels of these genes were associated with an increased risk of death. Further integration of the differential expression patterns revealed that AURKA, CENPA, CLDN9, DEPDC1, GPI, HMMR, and PSMC4 were upregulated in EC tissues and had HRs > 1, suggesting that their elevated expression may be associated with poorer prognosis in patients with EC.

Figure 1

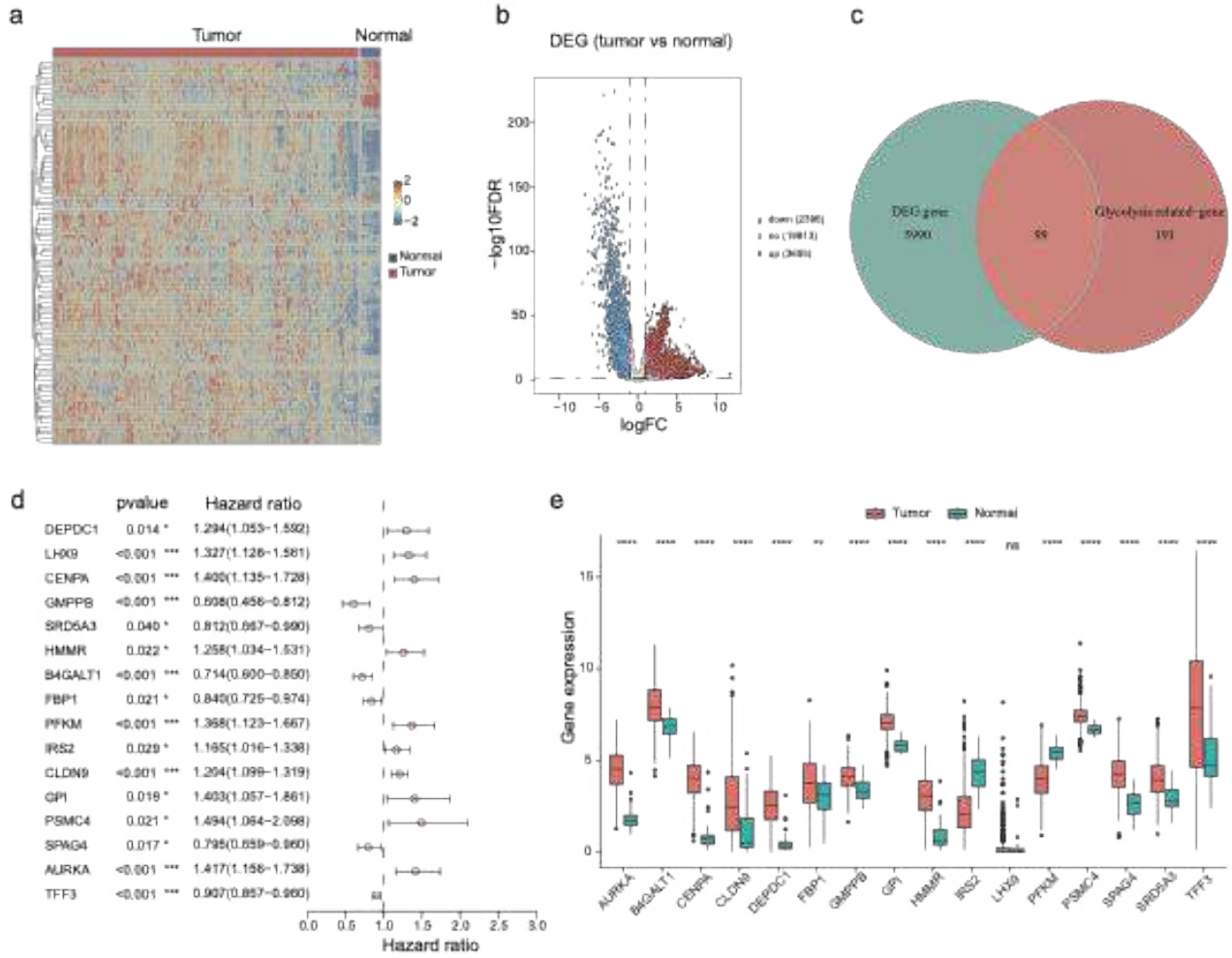


Figure 1. Identification of differentially expressed glycolysis-related genes and prognostic genes in EC.(**a**) Heatmap showing the expression patterns of DEGs between EC and normal endometrial tissues in the TCGA cohort. (**b**) Volcano plot of DEGs between EC and normal tissues. Red dots indicate upregulated genes, blue dots indicate downregulated genes, and gray dots indicate genes without significant differential expression. DEGs were defined as $|\log2FC| > 1$ and FDR < 0.05. (**c**) Venn diagram showing the intersection between EC-associated DEGs and glycolysis-related genes, identifying 99 differentially expressed glycolysis-related genes. (**d**) Forest plot of univariate Cox regression analysis showing 16 glycolysis-related genes significantly associated with OS in patients with EC. HRs and 95% confidence intervals (CIs) are shown. (**e**) Comparison of the expression levels of the 16 prognosis-associated glycolysis-related genes between EC and normal endometrial tissues.

3.2. Construction and Validation of a Machine Learning-Based Prognostic Model for EC

Based on the previously identified prognostic glycolysis-related genes, multiple machine learning algorithms and their combinations were employed to construct prognostic models for EC. The concordance indices (C-indices) of 117 candidate machine learning models were compared between the training cohort (Dataset 1) and the internal validation cohort (Dataset 2). Among these models, the integrated random survival forest

and generalized boosted model (RSF+GBM) demonstrated favorable overall predictive performance and was therefore selected as the optimal model (Figure 2a). Based on variable importance derived from the optimal model, seven key prognostic genes were further selected, including CLDN9, CENPA, GPI, TFF3, FBP1, SPAG4, and AURKA.

The prognostic performance of the RSF+GBM model was further evaluated using time-dependent ROC curve analysis. In the training cohort, the areas under the ROC curves (AUCs) for predicting 1-, 3-, and 5-year OS were 0.88, 0.81, and 0.84, respectively. In the internal validation cohort, the corresponding AUCs were 0.73, 0.77, and 0.76, respectively (Figure 2b), indicating that the model maintained good discriminatory performance across different time points, with relatively stable predictive performance in the internal validation cohort. Consistently, C-index analysis yielded values of 0.82 and 0.72 in the training and internal validation cohorts, respectively (Figure 2c). Specifically, the time-dependent ROC curves for 5-year OS yielded AUCs of 0.844 and 0.759 in the training and internal validation cohorts, respectively (Figure 2d), further supporting the ability of the model to discriminate long-term survival outcomes in patients with EC.

Risk scores were subsequently calculated using the RSF+GBM model, and patients were stratified into high- and low-risk groups. Kaplan–Meier survival analysis demonstrated significantly poorer OS in the high-risk group than in the low-risk group in both the training and internal validation cohorts (Figure 2e). In the training cohort, patients in the high-risk group had a significantly higher risk of death than those in the low-risk group(HR=7.98,95%CI:4.33–14.71, $p < 0.001$).A consistent survival difference was observed in the internal validation cohort (HR=4.81, $p < 0.001$). These findings indicate that the glycolysis-related prognostic model effectively stratified patients with EC into groups with distinct survival outcomes.

To further improve individualized prognostic assessment, age, tumor stage, and the model-derived risk score were integrated to construct a prognostic nomogram for patients with EC (Figure 2f).The nomogram enabled individualized estimation of 1-, 3-, and 5-year OS probabilities based on the risk score and clinical characteristics. Calibration curves were subsequently used to assess the agreement between nomogram-predicted and observed survival probabilities. The calibration curve for 1-year OS showed close agreement between the predicted and observed survival probabilities, indicating favorable calibration performance of the nomogram (Figure 2g).

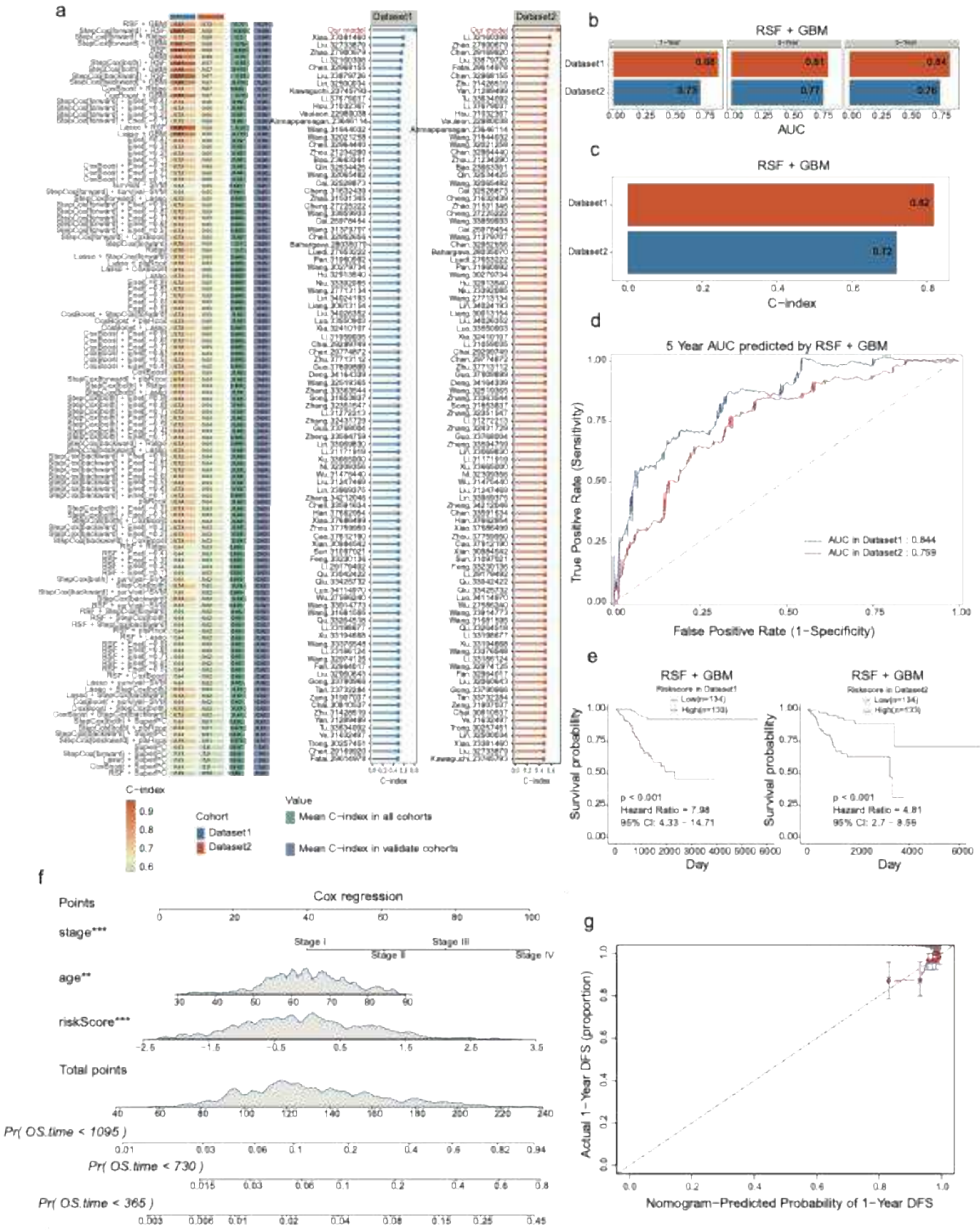


Figure 2. Construction and validation of the machine learning-based prognostic model for EC. (**a**) Comparison of the predictive performance of 117 machine-learning model combinations based

on the C-index in the training (Dataset 1) and validation (Dataset 2) cohorts. The random survival forest combined with generalized boosted regression model (RSF+GBM) showed favorable overall performance and was selected as the optimal model. (**b**) Time-dependent area under the AUC values of the RSF+GBM model for predicting 1-, 3-, and 5-year OS in Dataset 1 and Dataset 2. (**c**) Comparison of the C-index of the RSF+GBM model between the two datasets. (**d**) Time-dependent ROC curves evaluating the 5-year OS predictive performance of the RSF + GBM model in Dataset 1 and Dataset 2. (**e**) Kaplan–Meier survival curves comparing OS between the high- and low-risk groups stratified by the RSF+GBM-derived risk score in Dataset 1 and Dataset 2. (**f**) Nomogram incorporating age, tumor stage, and the model-derived risk score for predicting 1-, 3-, and 5-year OS in patients with EC. (**g**) Calibration curve evaluating the agreement between nomogram-predicted and observed 1-year OS probabilities.

3.3. Potential Biological Mechanisms Associated with the Risk Score

To further investigate the biological differences between patients with EC stratified into different risk groups, differential expression analysis was performed between the high- and low-risk groups. A total of 1,502 DEGs were identified, including 1,036 upregulated and 466 downregulated genes (Figure 3a). GO enrichment analysis was subsequently performed based on these DEGs. In the BP category, the DEGs were primarily enriched in intermediate filament organization, intermediate filament-based processes, keratinization, keratinocyte differentiation, and epidermal cell differentiation. In the cellular component (CC) category, the enriched terms were mainly associated with intermediate filaments and the intermediate filament cytoskeleton. In the MF category, the DEGs were predominantly enriched in ion channel activity, transmembrane transporter activity, protease inhibitor activity, and hormone activity (Figure 3b). These findings suggest that the two risk groups may differ in cytoskeletal remodeling, epithelial differentiation, and transmembrane transport processes.

KEGG pathway enrichment analysis further revealed that the DEGs were significantly enriched in multiple hormone regulation- and metabolism-related pathways, including hormone signaling-related pathways, retinol metabolism, fat digestion and absorption, steroid hormone biosynthesis, calcium signaling, cAMP signaling, thyroid hormone synthesis, estrogen signaling, PPAR signaling, and cholesterol metabolism (Figure 3c). These findings indicate that risk stratification was accompanied by alterations in hormone signaling, lipid metabolism, and other metabolic processes, suggesting that these biological processes may be associated with the malignant phenotypes of EC reflected by the glycolysis-related prognostic signature.

GSEA was further performed to characterize differences in pathway activity between the high- and low-risk groups. The results showed positive enrichment of several canonical cancer-related gene sets, including the G2M checkpoint, E2F targets, and KRAS-related signaling, whereas WNT/β-catenin-related signaling exhibited negative enrichment (Figure 3d). Collectively, these findings suggest that the two risk groups differ not only in metabolic and hormone-related processes but also in cell cycle regulation, proliferation-associated transcriptional programs, and oncogenic signaling pathways, providing further biological insights into the potential link between the glycolysis-related prognostic signature and clinical outcomes in EC.

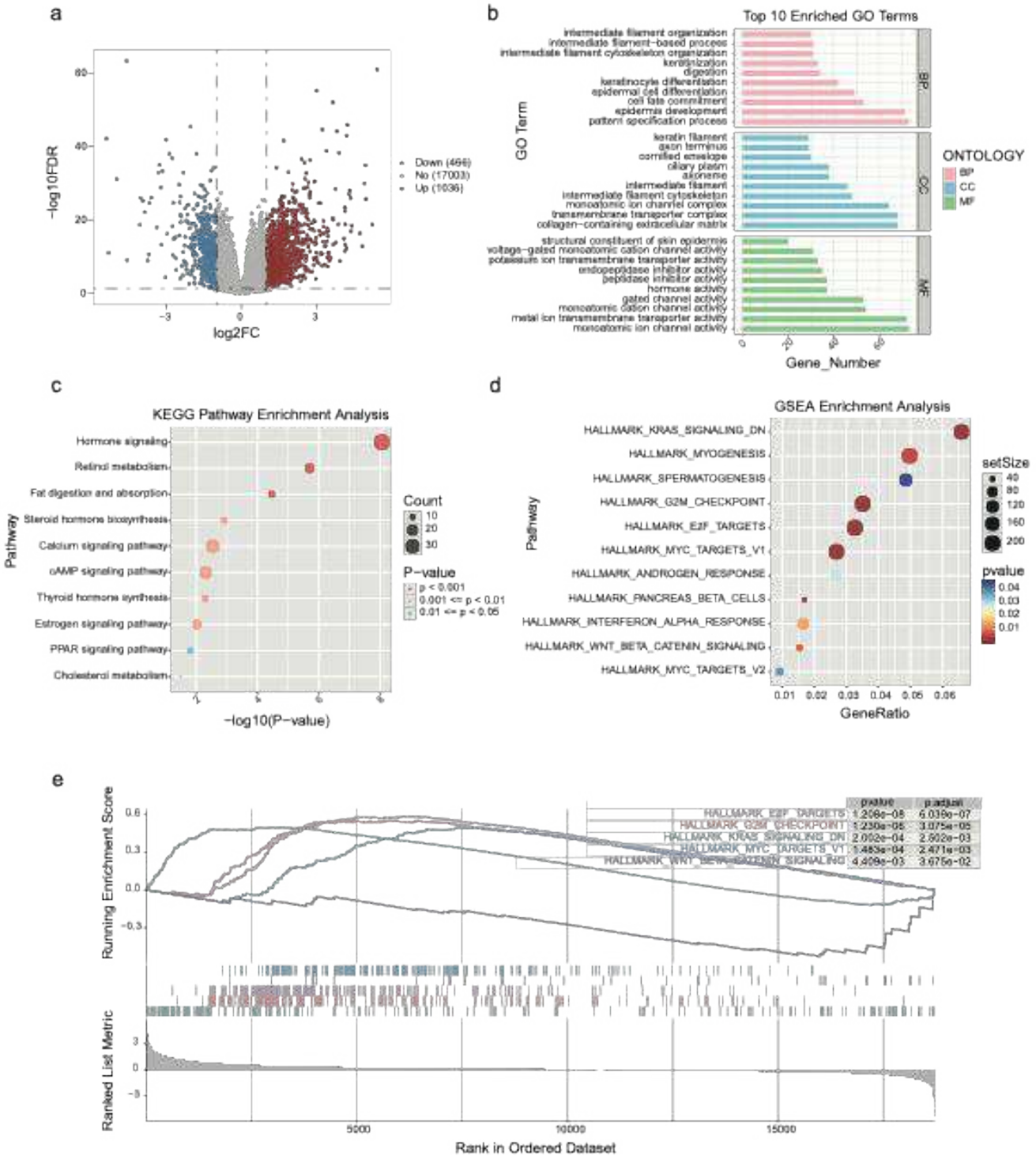


Figure 3. Differential expression and functional enrichment analyses associated with the glycolysis-related prognostic risk signature in EC. (**a**) Volcano plot showing DEGs between the high- and low-risk groups. Red dots represent upregulated genes, blue dots represent downregulated genes, and gray dots represent genes without significant differential expression. (**b**) GO enrichment analysis of the DEGs. The top enriched terms in BP, cellular component (CC), and MF categories are shown. (**c**) KEGG pathway enrichment analysis of the DEGs. Bubble size represents the number of enriched genes, and bubble color indicates the statistical significance of enrichment. (**d**) GSEA based on the Hallmark gene sets from the MSigDB. The bubble plot shows

representative enriched Hallmark pathways, with bubble size indicating gene-set size and color indicating the corresponding *P* value; representative enrichment curves are shown on the right.

3.4. Association Between the Glycolysis-Related Risk Score and Immune Cell Infiltration

To further investigate the association between the glycolysis-related prognostic signature and the tumor immune microenvironment in EC, the relative infiltration levels of 22 immune cell subsets were estimated using the CIBERSORT algorithm. Considerable heterogeneity in immune cell composition was observed across EC samples (Figure 4a). Comparison of immune cell infiltration between the high- and low-risk groups revealed significantly lower infiltration levels of CD8+T cells and regulatory T cells (Tregs) in the high-risk group (Figure 4b), indicating distinct immune infiltration profiles between patients with different risk stratifications.

Correlations among the different immune cell subsets were subsequently evaluated. CD8+T cell infiltration showed a strong positive correlation with activated CD4+ memory T cells and a strong negative correlation with M0 macrophages. Although both CD8+T cells and Tregs exhibited reduced infiltration in the high-risk group, the correlation between these two cell populations was relatively weak, suggesting that their alterations within the tumor immune microenvironment may not occur in parallel (Figure 4c).

Further correlation analysis revealed that the risk score was negatively correlated with the infiltration levels of CD8+T cells and Tregs, with Tregs showing a relatively pronounced negative correlation. In contrast, activated myeloid dendritic cells were positively correlated with the risk score, whereas resting myeloid dendritic cells were negatively correlated with the risk score. In addition, naïve B cells, plasma cells, γδ T cells, M1 macrophages, eosinophils, and resting mast cells showed varying degrees of positive correlation with the risk score (Figure 4d). Collectively, these findings indicate that the glycolysis-related risk score is associated with distinct patterns of immune cell infiltration in EC, suggesting that patients in different risk groups may exhibit distinct tumor immune microenvironmental characteristics.

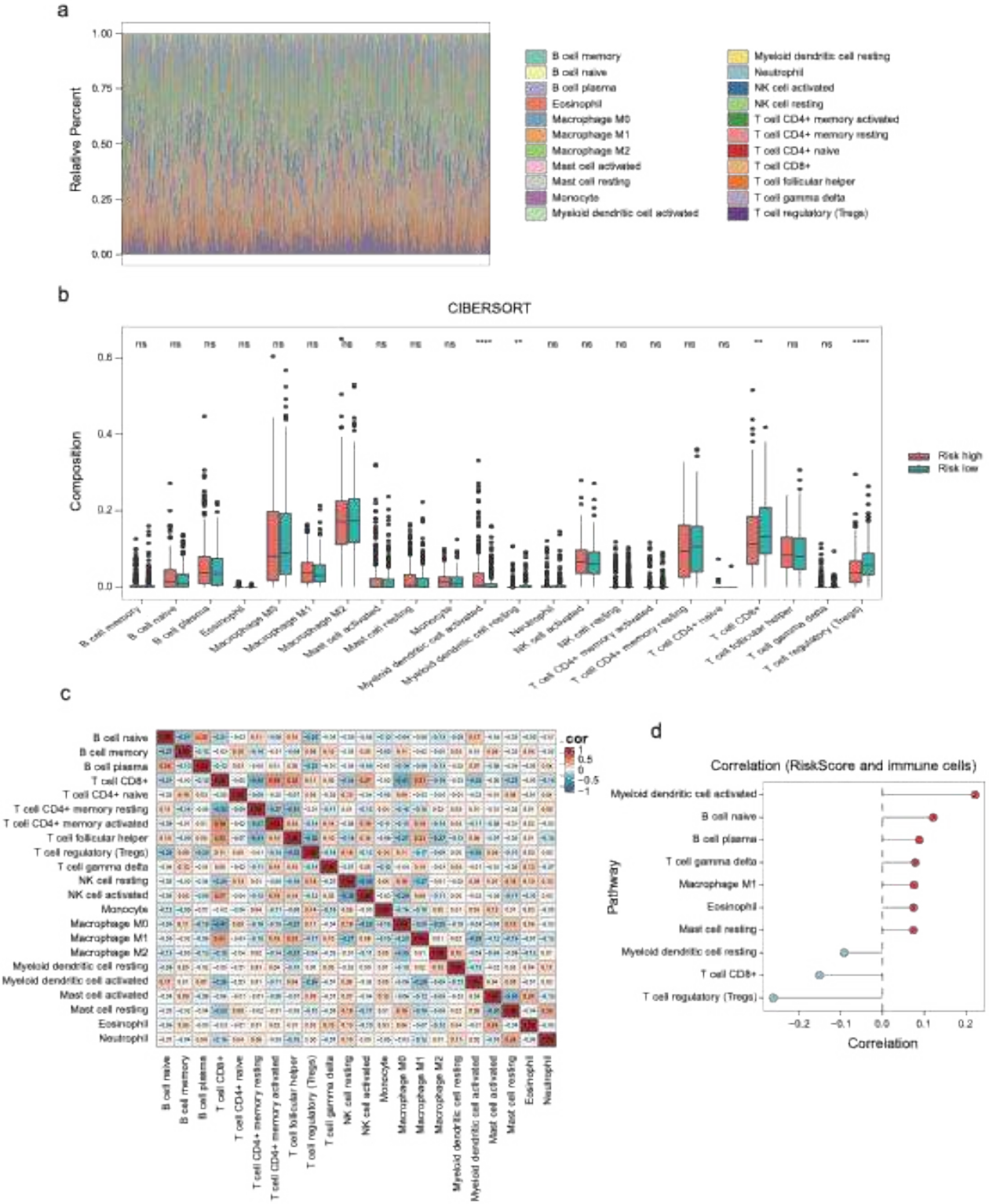


Figure 4. Immune infiltration landscape and its association with the glycolysis-related risk score in EC.(**a**) Relative proportions of 22 tumor-infiltrating immune cell subsets estimated using the CIBERSORT algorithm in individual EC samples. Different colors represent distinct immune cell populations. (**b**) Comparison of the relative abundance of 22 immune cell subsets between the high- and low-risk groups. Statistical significance is indicated above each comparison. (**c**) Correlation heatmap showing pairwise relationships among the 22 immune cell subsets. Red indicates positive correlations, whereas blue indicates negative correlations; the intensity of the color

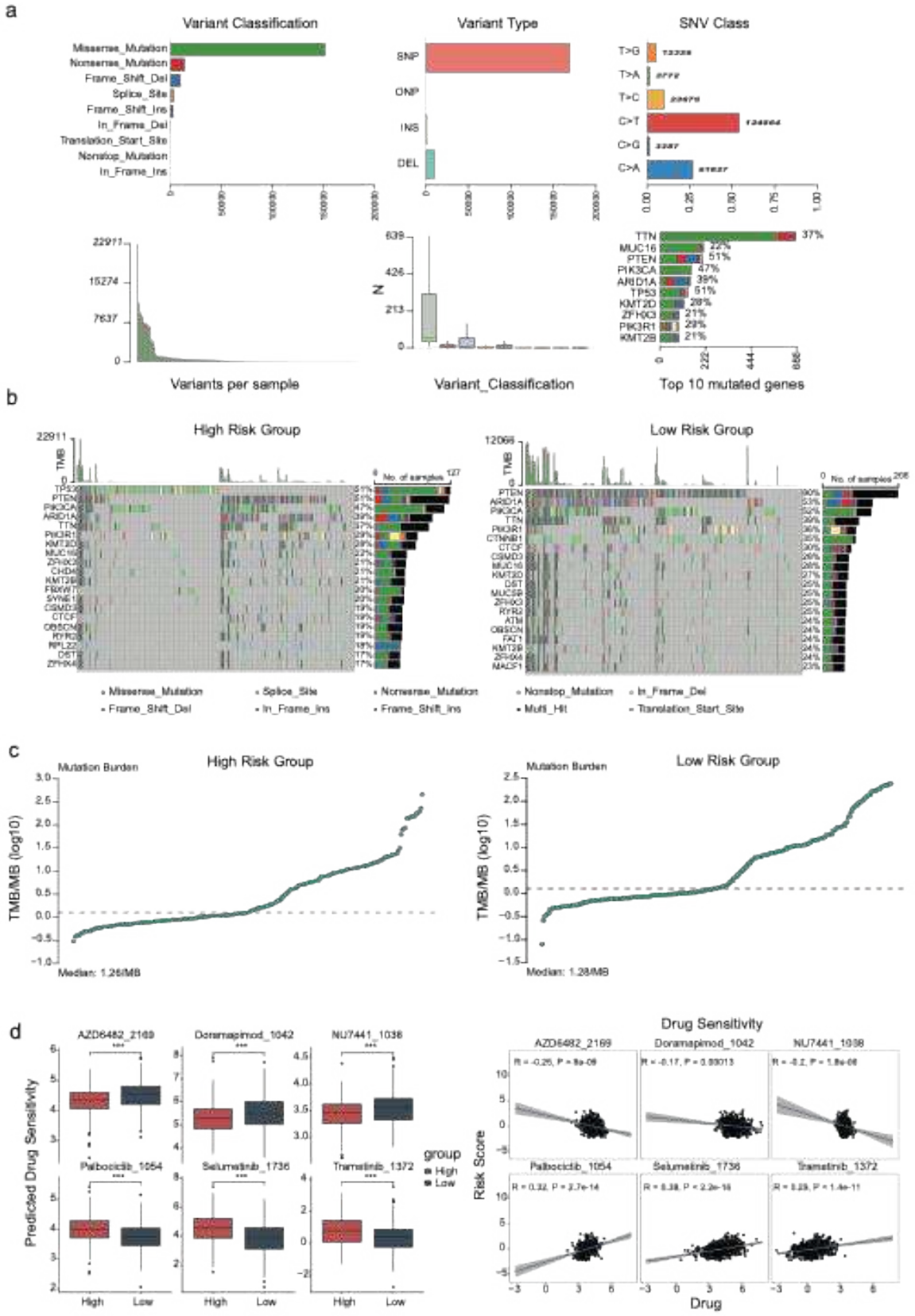
reflects the strength of the correlation. (**d**)Correlation between the glycolysis-related risk score and the abundance of significantly associated immune cell subsets. Red dots indicate positive correlations with the risk score, whereas blue dots indicate negative correlations.

3.5. Genomic Alterations and Predicted Drug Sensitivity Across Risk Groups

To further characterize the genomic features of patients with EC across different risk strata, somatic mutation profiles and TMB were analyzed using TCGA somatic mutation data. Across the overall cohort, missense mutations were the most common type of somatic alteration, followed by nonsense mutations, frameshift deletions, splice-site mutations, and frameshift insertions. Among single-nucleotide variants (SNVs), C > T transitions represented the predominant nucleotide substitution pattern (Figure 5a).

Comparison of somatic mutation profiles between the high- and low-risk groups revealed broadly similar overall mutation types and distribution patterns, although differences were observed in the mutation frequencies of several frequently mutated genes. TP53 was the most frequently mutated gene in the high-risk group, with a mutation frequency of approximately 51%, whereas PTEN was the most frequently mutated gene in the low-risk group, with a mutation frequency of approximately 80% (Figure 5b). These findings suggest that the two risk groups may exhibit distinct mutation patterns involving key cancer-associated genes, although the biological relationship between these genomic alterations and risk stratification requires further investigation. TMB analysis showed comparable overall mutation burdens between the two groups, with median TMB values of approximately 1.26 and 1.28 mutations/Mb in the high- and low-risk groups, respectively, and no apparent overall difference was observed (Figure 5c).

Potential differences in drug sensitivity between the two risk groups were further evaluated based on the GDSC database, using predicted half-maximal inhibitory concentration (IC50) values as an indicator of drug response. The high-risk group exhibited lower predicted IC50 values for AZD6482, doramapimod, and NU7441 than the low-risk group. Consistently, risk scores were negatively correlated with the predicted IC50 values of these agents, suggesting that higher risk scores may be associated with greater predicted sensitivity to these drugs. In contrast, the high-risk group exhibited higher predicted IC50 values for Palbociclib, selumetinib, and trametinib, and risk scores were positively correlated with their predicted IC50 values, suggesting relatively lower predicted sensitivity to these agents among high-risk patients ($p < 0.05$; Figure 5d). Collectively, these findings indicate that the glycolysis-related risk stratification is associated with distinct genomic alteration patterns and may also capture differences in predicted drug responses, providing potential clues for further investigation of risk-adapted therapeutic strategies in EC.



409

Figure 5. Genomic mutation landscape and drug sensitivity associated with the glycolysis-related risk signature in EC.(**a**) Overview of the somatic mutation landscape in EC, including variant classifications, variant types, SNV classes, variants per sample, and the most frequently mutated 410 411 412

genes. (**b**) Oncoplots showing the distribution and frequency of somatic mutations in the high- and low-risk groups. Different colors represent distinct mutation types, and the bar plots indicate the mutation frequencies of individual genes. (**c**) Distribution of tumor mutation burden in patients from the high- and low-risk groups. (**d**) Comparison of predicted drug sensitivity between the high- and low-risk groups and correlation analysis between the glycolysis-related risk score and the estimated sensitivity to selected therapeutic agents. Boxplots show differences in predicted drug response between risk groups, whereas scatter plots illustrate the correlations between risk scores and drug sensitivity.

3.6. Prioritization of CENPA and Its Single-Cell Expression Characteristics

To further characterize the cell type-specific expression patterns of candidate prognostic genes within the EC tumor microenvironment, the GSE179640 and GSE173682 scRNA-seq datasets were obtained from the GEO database, comprising five EC samples and three normal endometrial samples. Following quality control, data integration, dimensionality reduction, clustering, and cell type annotation, the cellular composition and transcriptional characteristics of different cell populations were characterized at single-cell resolution. Uniform Manifold Approximation and Projection (UMAP) analysis revealed relatively distinct clusters corresponding to different cell types, highlighting the cellular heterogeneity of EC tissues. Further examination of CENPA expression revealed that it was expressed across multiple cell populations, with relatively high expression levels in specific cellular subpopulations (Figure 6a). These findings indicate heterogeneous CENPA expression across different cellular populations within the EC microenvironment.

The preceding machine learning analysis identified seven key prognostic genes, including CLDN9, CENPA, GPI, TFF3, FBP1, SPAG4, and AURKA. The ML.Corefeature.Prog.Screen function in the Mime package was subsequently used to evaluate the selection frequency of candidate genes across different machine learning model combinations. CENPA was frequently selected across multiple models and, together with genes such as CLDN9 and GPI, exhibited relatively robust feature selection across different modeling approaches (Figure 6b). Considering the machine learning results, its single-cell expression characteristics, and existing evidence regarding its biological relevance[21, 22], CENPA was selected as the primary candidate gene for subsequent analyses to further investigate its potential biological and clinical significance in EC.

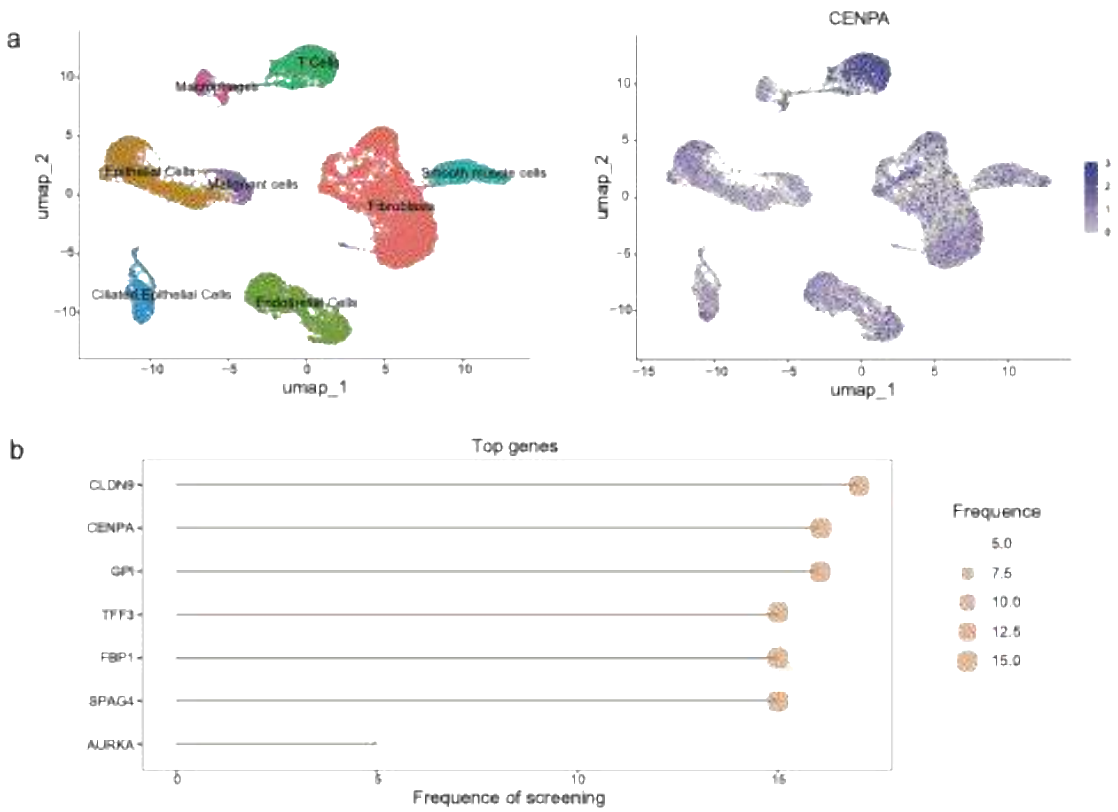


Figure 6. Single-cell characterization and further validation of CENPA in EC.(**a**) UMAP visualization of major cell populations identified by scRNA-seq analysis in EC tissues (left) and the cell-type-specific expression distribution of CENPA across the identified cell populations (right). Major cell types included epithelial cells, ciliated epithelial cells, T cells, macrophages, myometrial cells, fibroblasts, endothelial cells, and smooth muscle cells. (**b**) Screening frequency of the seven key prognostic genes, including CLDN9, CENPA, GPI, TFF3, FBP1, SPAG4, and AURKA, across the machine-learning model combinations.

3.7. Association of CENPA Expression with Clinicopathological Characteristics and Patient Prognosis

To further evaluate the clinical significance of CENPA, its associations with patient age, tumor stage, and OS were analyzed using the TCGA-UCEC cohort. Patients were stratified into four age groups: <50, 50–59, 60–69, and ≥70 years. CENPA expression differed across age groups, with statistically significant differences observed in some pairwise comparisons. However, no consistent increasing or decreasing trend in CENPA expression was observed with advancing age (Figure 7a), suggesting that the relationship between CENPA expression and age may be complex and does not exhibit a clear age-dependent pattern.

We next investigated the association between CENPA expression and tumor stage. CENPA expression was significantly higher in patients with advanced-stage EC (FIGO stages III–IV) than in those with early-stage disease (FIGO stages I–II) (p = 0.00013; Figure 7b). These findings demonstrate an association between elevated CENPA

expression and advanced tumor stage, suggesting that CENPA may be linked to more aggressive clinicopathological features of EC.

Patients were subsequently stratified into high- and low-CENPA expression groups according to the optimal cutoff value, and Kaplan–Meier survival analysis was performed to evaluate the association between CENPA expression and OS. Patients with high CENPA expression exhibited significantly poorer OS than those with low CENPA expression (log-rank p = 0.0013; Figure 7c). Together with the preceding univariate Cox regression results, these findings consistently demonstrate an association between elevated CENPA expression and unfavorable survival outcomes, supporting the potential utility of CENPA as a prognostic biomarker for EC.

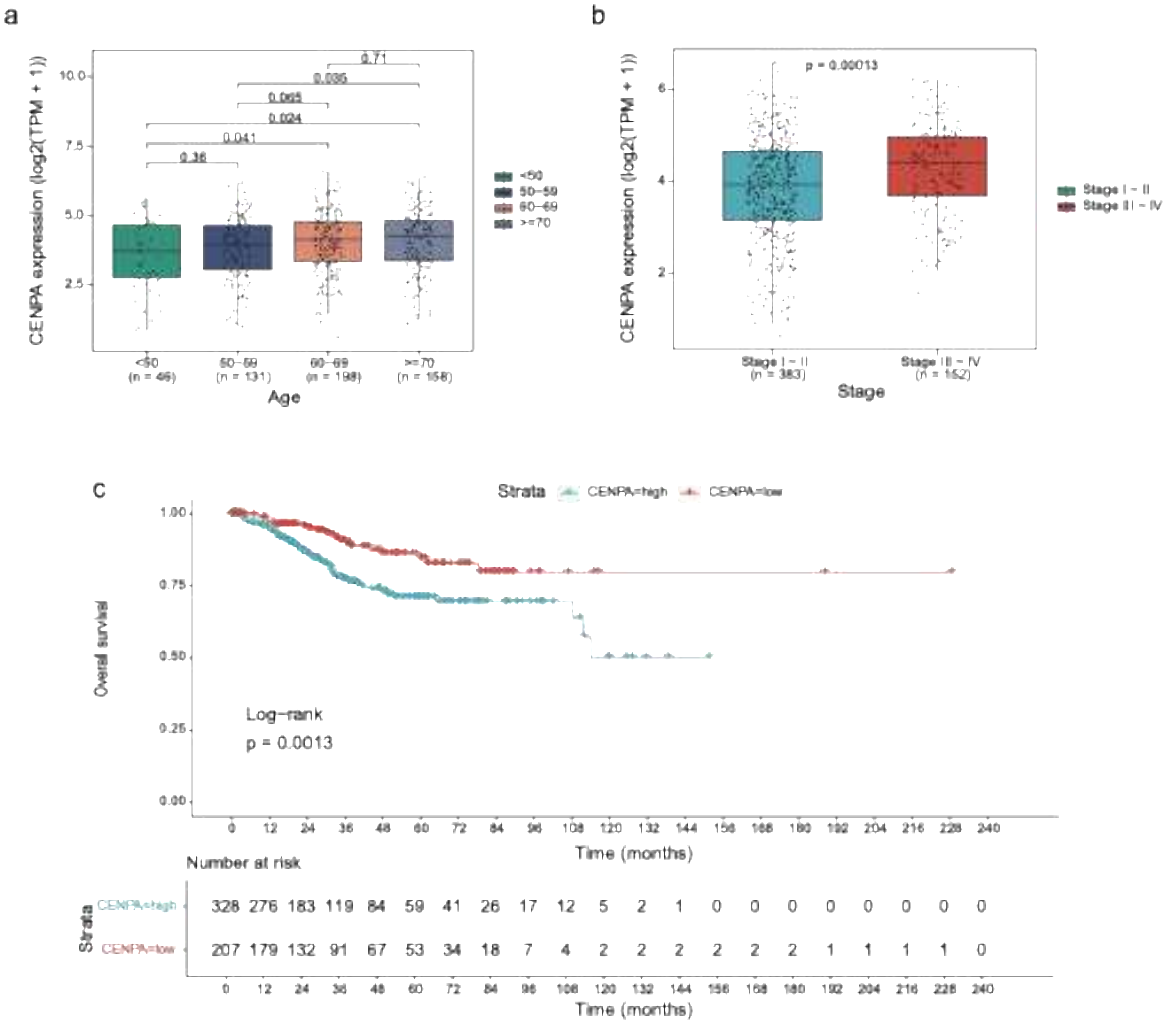


Figure 7. Association of CENPA expression with clinicopathological characteristics and prognosis in EC.(**a**) Comparison of CENPA expression among patients with EC in different age groups (<50, 50–59, 60–69, and ≥70 years). Pairwise *P* values are indicated above the boxplots. (**b**) Comparison of CENPA expression between patients with early-stage (stage I–II) and advanced-stage (stage III–IV) EC. CENPA expression was significantly higher in patients with stage III–IV disease than in those with stage I–II disease (p = 0.00013). (**c**) Kaplan–Meier survival analysis comparing OS between patients with high and low CENPA expression. Survival distributions differed significantly between the two groups (log-rank p = 0.0013).

4. Discussion

In this study, we integrated EC-related DEGs with glycolysis-related genes using transcriptomic data from the TCGA-UCEC cohort and developed a glycolysis-related prognostic model by combining multiple machine learning approaches. Seven key prognostic genes, including CLDN9, CENPA, GPI, TFF3, FBP1, SPAG4, and AURKA, were ultimately identified. The resulting RSF+GBM model demonstrated good discriminatory performance in both the training and internal validation cohorts and effectively stratified patients with EC into groups with distinct survival outcomes. Further functional enrichment, immune infiltration, genomic alteration, drug sensitivity, and single-cell transcriptomic analyses suggested that this risk signature was not only associated with patient survival but also reflected multidimensional biological differences involving cell proliferation, metabolic reprogramming, hormone regulation, and remodeling of the tumor microenvironment.

Metabolic reprogramming is a fundamental hallmark of cancer. Enhanced glycolysis provides rapidly proliferating tumor cells with both energy and biosynthetic intermediates and may also influence tumor progression through lactate accumulation, alterations in redox homeostasis, and epigenetic regulation. EC is closely associated with metabolic abnormalities, including obesity, diabetes mellitus, and insulin resistance, suggesting that altered glucose metabolism may play an important role in its development and progression. In the present study, the glycolysis-related risk signature was significantly associated with OS, further supporting a potential link between dysregulated glucose metabolism and aggressive biological behavior in EC. Importantly, the model developed in this study was designed to predict survival outcomes and facilitate prognostic risk stratification rather than to directly support the early diagnosis of EC. Therefore, its potential clinical utility requires further validation in independent prospective cohorts.

Functional enrichment analyses provided additional insights into the potential biological basis underlying the glycolysis-related risk signature. DEGs between the high- and low-risk groups were predominantly involved in intermediate filament organization, keratinization, epithelial cell differentiation, transmembrane transport, and ion channel-related processes, suggesting that cytoskeletal remodeling and alterations in epithelial phenotypes may accompany different prognostic risk states. KEGG analysis further revealed enrichment of pathways associated with hormone regulation, retinol metabolism, lipid metabolism, steroid hormone biosynthesis, and PPAR signaling. Given the pronounced hormone responsiveness of endometrial tissue, these findings raise the possibility of crosstalk between glycolysis-related metabolic alterations, hormone signaling, and lipid metabolic networks in EC. GSEA further revealed positive enrichment of G2M checkpoint, E2F target, and KRAS-related gene sets, whereas WNT/β-catenin-related signaling showed negative enrichment. Both G2M checkpoint and E2F-related programs are closely involved in cell-cycle progression and proliferative activity. Thus, enhanced cell cycle-associated transcriptional programs may partially account for the poorer survival outcomes observed in patients with higher risk scores. Nevertheless, these enrichment analyses reveal associations rather than causal relationships and do not establish that glycolytic reprogramming directly drives alterations in these pathways.

Distinct immune infiltration patterns were also observed between the two risk groups. CD8+T-cell infiltration was significantly reduced in the high-risk group and was negatively correlated with the risk score, suggesting that the high-risk state may be accompanied by reduced cytotoxic antitumor immune activity. Interestingly, Treg infiltration was also decreased in the high-risk group and negatively correlated with the risk score, indicating that the immune differences associated with risk stratification cannot be simply characterized as conventional immunosuppression. In addition, activated dendritic cells were positively correlated with the risk score, whereas resting dendritic cells

showed a negative correlation, suggesting alterations in immune cell composition and functional states across different risk groups. Taken together, these findings are more consistent with remodeling of the tumor immune microenvironment than with a unidirectional increase in immunosuppression. Glycolytic metabolites, particularly lactate, have been implicated in the regulation of T cells, macrophages, and dendritic cells. It is therefore plausible that metabolic reprogramming and immune microenvironmental alterations interact in EC; however, this hypothesis requires further functional validation.

At the genomic level, the high- and low-risk groups exhibited broadly comparable overall mutation profiles and TMB levels, although differences were observed in the frequencies of several commonly mutated cancer-associated genes. TP53 mutations were more frequent in the high-risk group, whereas PTEN mutations were more prevalent in the low-risk group. Aberrant TP53 has been associated with genomic instability and aggressive tumor phenotypes, whereas loss or dysfunction of PTEN is closely linked to dysregulation of the PI3K/AKT signaling pathway. These findings suggest that the glycolysis-related risk signature may, at least in part, reflect the biological heterogeneity of EC arising from distinct molecular backgrounds. However, because the independent or causal relationships between individual mutations and the risk score were not specifically evaluated in the present study, TP53 or PTEN alterations should not be interpreted as direct determinants of risk stratification. Moreover, the absence of a marked difference in overall TMB between the two groups suggests that the prognostic information captured by the glycolysis-related risk model may not be primarily driven by the overall somatic mutation burden.

Drug sensitivity prediction further suggested potential differences in therapeutic response across risk groups. The high-risk group exhibited lower predicted IC50 values for AZD6482, doramapimod, and NU7441, but higher predicted IC50 values for Palbociclib, selumetinib, and trametinib. These findings suggest that risk score-based stratification may potentially provide information relevant to differential drug response. However, these results were computational predictions derived from GDSC data and transcriptomic profiles and therefore cannot be considered equivalent to actual clinical responses to these agents. Their potential therapeutic relevance will require validation in cellular and animal models, as well as in real-world or prospective treatment cohorts.

Among the seven key prognostic genes identified in the present study, CENPA was further prioritized as a candidate gene for detailed investigation. CENPA is a centromere-specific histone H3 variant that plays an essential role in centromere identity, accurate chromosome segregation, and mitosis. Aberrant CENPA expression has been reported in multiple malignancies and has frequently been associated with increased cellular proliferation, chromosomal instability, and unfavorable clinical outcomes. In the present study, CENPA was upregulated in EC tissues, and elevated CENPA expression was significantly associated with advanced tumor stage and poorer OS, consistent with the HR > 1 observed in the preceding univariate Cox regression analysis. Single-cell transcriptomic analysis further demonstrated that CENPA was expressed across multiple cellular populations, with relatively high expression in specific subpopulations, Single-cell transcriptomic analysis further demonstrated heterogeneous CENPA expression across multiple cellular populations within the EC microenvironment. Because CENPA expression was not restricted to a specific cell population, however, the current findings do not support its interpretation as a malignant cell-specific marker in EC. Nevertheless, because CENPA expression was not restricted to a specific cell population, the current findings do not support its interpretation as an EC malignant cell-specific marker.

Importantly, the inclusion of CENPA within a glycolysis-related prognostic framework does not establish that CENPA directly regulates glycolysis. Rather, the present findings primarily support an association among CENPA, glycolysis-related

transcriptional characteristics, and unfavorable clinical outcomes in EC. Given its established role in cell-cycle progression and chromosome segregation, increased CENPA expression may accompany enhanced proliferative activity and the consequent increase in metabolic demand. Alternatively, CENPA and glycolytic programs may be coordinately regulated by shared upstream mechanisms, resulting in parallel changes during tumor progression. Further mechanistic studies are therefore required to determine whether a direct regulatory relationship exists between CENPA and glycolytic reprogramming in EC.

Several limitations of this study should be acknowledged. First, the prognostic model was primarily developed using the retrospective TCGA cohort and validated through an internal split of the same cohort. A completely independent external clinical cohort was not available, and the generalizability of the model therefore requires further evaluation. Second, although scRNA-seq datasets were integrated to characterize the cellular expression pattern of CENPA, heterogeneity in sample sources, sequencing platforms, and clinical characteristics across public datasets may have influenced the results. Third, several potentially relevant confounding factors, including treatment regimens, molecular subtypes, metabolic comorbidities, and other clinical variables, could not be comprehensively incorporated into the model. In addition, the immune infiltration, drug sensitivity, and pathway enrichment analyses were predominantly based on computational inference and lacked experimental validation. Finally, the present study did not directly demonstrate that CENPA participates in glycolytic reprogramming in EC. Its specific molecular functions and potential regulatory relationship with glucose metabolism therefore warrant further investigation using in vitro and in vivo experimental models.

5. Limitations

Several limitations of this study should be acknowledged. First, this study was based entirely on publicly available datasets, including TCGA and GEO, and the sample sizes of the included cohorts may limit the statistical power and generalizability of the findings. Second, differences in sample characteristics, sequencing platforms, and data processing methods across datasets may introduce potential heterogeneity and batch effects. Third, although functional enrichment analyses and computational drug sensitivity assessments provided insights into potential biological mechanisms and therapeutic implications, these findings remain exploratory and cannot establish causal relationships or confirm actual drug responses. Importantly, no experimental validation, such as quantitative real-time PCR, Western blotting, or in vitro functional assays, was performed in this study. Therefore, the findings should be interpreted with caution, and further validation using larger independent cohorts and experimental investigations is necessary to establish their robustness and biological relevance.

6.Conclusions

In conclusion, we developed and internally validated a machine learning-based glycolysis-related prognostic signature for EC that effectively stratified patients according to survival risk. The risk signature was associated with distinct metabolic, cell cycle-related, immune, and genomic characteristics. CENPA was further prioritized as a candidate prognostic biomarker, with elevated expression associated with advanced tumor stage and unfavorable survival outcomes. These findings provide a framework for further investigation of glycolysis-related prognostic heterogeneity in EC. Independent external validation and experimental studies are required to confirm the clinical utility of the prognostic signature and elucidate the biological role of CENPA in EC.

Author Contributions: HY and LZ: writing-original draft, methodology, formal analysis and investigation. ZL: sample collection, investigation. YJL、LQW and XSZ: writing-review and editing. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by The fundamental research funds for the provincial universities in heilongjiang province（2026，Liqun Wang）

Conflicts of Interest: The authors declare no conflicts of interest.

Data Availability Statement： The datasets generated during and analysed during the current study are available in the publicly available databases. Cancer genomic and clinical data were retrieved from The Cancer Genome Atlas (TCGA) database through the Genomic Data Commons (GDC) Data Portal (https://portal.gdc.cancer.gov/; study accession number: phs000178).

Gene expression data were obtained from the Gene Expression Omnibus (GEO) database (https://www.ncbi.nlm.nih.gov/geo/) under accession numbers GSE179640 (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE179640) and GSE173682 (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE173682).

Functional enrichment and pathway analyses were performed using the Kyoto Encyclopedia of Genes and Genomes (KEGG; https://www.kegg.jp/), the Molecular Signatures Database (MSigDB; https://www.gsea-msigdb.org/gsea/msigdb/), and Reactome (https://reactome.org/).

Drug sensitivity data were obtained from the Genomics of Drug Sensitivity in Cancer (GDSC) database (https://www.cancerrxgene.org/).

Ethical Approval and Accordance: Not applicable.

Abbreviations

The following abbreviations are used in this manuscript:

EC	Endometrial cancer
GEO	Gene Expression Omnibus
TCGA	The Cancer Genome Atlas
FDR	false discovery rate
C-index	concordance index
ROC	receiver operating characteristic
GSEA	gene set enrichment analysis

References

1. BANZ-JANSEN C, HELWEG L P, KALTSCHMIDT B. Endometrial Cancer Stem Cells: Where Do We Stand and Where Should We Go? [J]. Int J Mol Sci, 2022, 23(6).
2. CROSBIE E J, KITSON S J, MCALPINE J N, et al. Endometrial cancer [J]. Lancet, 2022, 399(10333): 1412–28.
3. BRAY F, LAVERSANNE M, SUNG H, et al. Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries [J]. CA Cancer J Clin, 2024, 74(3): 229–63.
4. HOW J A, JAZAERI A A, WESTIN S N, et al. Translating biological insights into improved management of endometrial cancer [J]. Nat Rev Clin Oncol, 2024, 21(11): 781–800.
5. DI DONATO V, GIANNINI A, BOGANI G. Recent Advances in Endometrial Cancer Management [M]. J Clin Med. Switzerland. 2023.

6. KIERANS S J, TAYLOR C T. Glycolysis: A multifaceted metabolic pathway and signaling hub [J]. J Biol Chem, 2024, 300(11): 107906.
7. KARKIA R, MACCARTHY G, PAYNE A, et al. The Association Between Metabolic Syndrome and the Risk of Endometrial Cancer in Pre- and Post-Menopausal Women: A UK Biobank Study [J]. J Clin Med, 2025, 14(3).
8. ZHANG W, LIN Y T, JIANG J W, et al. Mapping the expression and functional landscape of key enzymes in glucose metabolism within human gynecological tumors [J]. Curr Probl Cancer, 2025, 54: 101155.
9. IRONI G, MAPELLI P, BERGAMINI A, et al. Hybrid PET/MRI in Staging Endometrial Cancer: Diagnostic and Predictive Value in a Prospective Cohort [J]. Clin Nucl Med, 2022, 47(3): e221–e9.
10. SHEN L, WANG J, LI Y, et al. Transcription Factor STAT3-Activated LDHB Promotes Tumor Properties of Endometrial Cancer Cells by Inducing MDH2 Expression [J]. Mol Biotechnol, 2025, 67(2): 562–74.
11. FEI S, XU X, WANG T, et al. FBXW7 mutations reprogram glucose metabolism by activating the ETV6-GLUT1 axis [J]. Mol Cancer, 2026, 25(1).
12. JIANG R, LIN X, ZHENG J, et al. Integrated single-cell and bulk transcriptomics reveals an IGF2BP3-driven metabolic circuitry in endometrial cancer [J]. BMC Cancer, 2026, 26(1).
13. YANG S, MA Y, WU T, et al. Lactate transmission from hypoxic tumor cells promotes macrophage senescence and M2 polarization via the DNMT1-NHE7 axis to accelerate endometrial cancer progression [J]. Cell Death Dis, 2026, 17(1): 185.
14. MA J, YAO Z, MA L, et al. Glucose metabolism reprogramming in gynecologic malignant tumors [J]. J Cancer, 2024, 15(9): 2627–45.
15. LIU Z, LIU L, WENG S, et al. Machine learning-based integration develops an immune-derived lncRNA signature for improving outcomes in colorectal cancer [J]. Nat Commun, 2022, 13(1): 816.
16. WU T, HU E, XU S, et al. clusterProfiler 4.0: A universal enrichment tool for interpreting omics data [J]. Innovation (Camb), 2021, 2(3): 100141.
17. CHEN B, KHODADOUST M S, LIU C L, et al. Profiling Tumor Infiltrating Immune Cells with CIBERSORT [J]. Methods Mol Biol, 2018, 1711: 243–59.
18. CHAN T A, YARCHOAN M, JAFFEE E, et al. Development of tumor mutation burden as an immunotherapy biomarker: utility for the oncology clinic [J]. Ann Oncol, 2019, 30(1): 44–56.
19. MAYAKONDA A, LIN D C, ASSENOV Y, et al. Maftools: efficient and comprehensive analysis of somatic variants in cancer [J]. Genome Res, 2018, 28(11): 1747–56.
20. YANG W, SOARES J, GRENINGER P, et al. Genomics of Drug Sensitivity in Cancer (GDSC): a resource for therapeutic biomarker discovery in cancer cells [J]. Nucleic Acids Res, 2013, 41(Database issue): D955–61.
21. LI S, ZHANG Z, LI Z, et al. CENPA promotes glutamine metabolism and tumor progression by up-regulating SLC38A1 in endometrial cancer [J]. Cell Signal, 2024, 117: 111110.
22. WANG L, WU H, TANG X, et al. CENPA Promotes Endometrial Cancer Progression by Stabilizing YY1 and Enhancing Glycolytic Metabolism [J]. Front Biosci (Landmark Ed), 2026, 31(3): 49033.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.