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.

Integrative transcriptomics and molecular simulations identify IRS1 and PRKCQ as translational biomarkers for EDC-associated PCOS risk stratification

Xin Liu
Qilu Hospital of Shandong University
Lei Zhang
The First Affiliated Hospital of Shandong First Medical University & Shandong Provincial QianFoShan Hospital
Sai Han
Qilu Hospital of Shandong University
Jia Liu
Qilu Hospital, Shandong University
Zhihong Wu
Qilu Hospital, Shandong University
Li Gong
gongli@qiluhospital.com
Qilu Hospital, Shandong University

Article

Keywords: Polycystic ovary syndrome, Endocrine-disrupting chemicals, Machine learning, molecular docking

Posted Date: April 29th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9227802/v1

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

Additional Declarations: No competing interests reported.

Integrative transcriptomics and molecular simulations identify IRS1 and PRKCQ as translational biomarkers for EDC-associated PCOS risk stratification

Xin Liu[1], Lei Zhang[2], Sai Han[3], Jia Liu[4], Zhihong Wu[5], Li Gong[5].

1. Department of Ultrasound , Qilu Hospital of Shandong University ,Jinan, China

2. Department of Neurosurgery, The First Affiliated Hospital of Shandong First Medical University & Shandong Provincial QianFoShan Hospital, Shandong Medicine and Health Key Laboratory of Neurosurgery, Jinan, China

3. Department of Obstetrics and Gynecology, Qilu Hospital of Shandong University, Jinan, China

4. Laboratory of Basic Medical Sciences, Qilu Hospital, Cheeloo College of Medicine, Shandong University, Jinan, China

5. Department of Traditional Chinese Medicine, Qilu Hospital, Cheeloo College of Medicine, Shandong University, Jinan, China

Adress for correspondence:

Li Gong, Department of Traditional Chinese Medicine, Qilu Hospital, Cheeloo College of Medicine, Shandong University, 107, Wen Hua Xi Road, Jinan, Shandong Province, 250012, P. R. China.

E-mail:gongli@qiluhospital.com

Phone: +86-18560087337

Abstract

Background: Endocrine-disrupting chemicals (EDCs) are linked to polycystic ovary syndrome (PCOS), but their transcriptional mechanisms remain unclear.

Methods: Microarray datasets were analyzed to identify DEGs, with GSE10946 as the

training cohort and the merged external datasets GSE106724, GSE34526, and GSE98595 as the validation cohort. EDC-related targets were intersected with DEGs to obtain EDC-associated DEGs. Protein-protein interaction network construction and machine learning modeling were employed to identify hub genes. The diagnostic performance of hub genes was evaluated using receiver operating characteristic analysis, and a nomogram was constructed based on the key predictors. Gene set enrichment analysis (GSEA), immune cell infiltration analysis, molecular docking, and molecular dynamics simulations were conducted to explore potential mechanisms.

Results: A total of 63 EDC-related DEGs were screened, displaying multi-target regulatory features and enrichment in DNA replication/repair, cell cycle, and p53 signaling pathways. 13 hub genes were identified, among which insulin receptor substrate 1 (IRS1) and protein kinase C theta (PRKCQ) showed consistent upregulation in PCOS samples and demonstrated good diagnostic performance. An IRS1-PRKCQ nomogram achieved an area under the receiver operating characteristic curve (AUC) of 0.773 with good calibration and clinical net benefit. GSEA and immune analyses suggested associations with metabolic/replicative programs and immune infiltration patterns. Bisphenol A showed favorable docking to IRS1/PRKCQ, and molecular dynamics simulations supported a stable IRS1-bisphenol A complex.

Conclusions: IRS1 and PRKCQ represent EDC-associated biomarkers with potential clinical utility for PCOS risk stratification and mechanistic insight into EDC-related dysregulation.

Keywords: Polycystic ovary syndrome; Endocrine-disrupting chemicals; Machine

learning; molecular docking

Introduction

Polycystic ovary syndrome (PCOS) is the most common endocrine disorder affecting women globally, impacting approximately 6–20% of women of reproductive age[1]. This condition is closely characterized by clinical manifestations such as menstrual irregularities, ovulatory dysfunction, hyperandrogenemia, insulin resistance, and infertility[2,3]. In the long term, PCOS significantly increases the risk of metabolic and certain malignancies, particularly endometrial cancer , including type 2 diabetes, cardiovascular diseases, and endometrial cancer[4,5]. The pathogenesis of PCOS is multifactorial, with current research primarily focusing on genetic susceptibility, epigenetic modifications, gut microbiota dysbiosis, immune and inflammatory mechanisms[6-8]. Current treatments for PCOS primarily aim to alleviate symptoms and manage metabolic complications, with common strategies including oral contraceptives, ovulation-inducing drugs, insulin sensitizers, and lifestyle interventions[9,10]. However, most of these interventions target single pathological phenotypes, fail to fundamentally halt disease progression, and exhibit significant variability in efficacy across different phenotypes. Therefore, further exploration of PCOS pathogenesis and identification of novel therapeutic targets hold significant clinical importance.

Endocrine-disrupting chemicals (EDCs) constitute a class of exogenous compounds that interfere with the physiological functions of endogenous hormones by altering their synthesis, secretion, transport, metabolism, receptor binding, or

elimination[11]. Representative EDCs include bisphenol A (a phenolic compound), phthalates (plasticizers), and parabens (preservatives), all of which exert their deleterious effects primarily through their ability to mimic or antagonize steroid hormones via high-affinity binding to nuclear and membrane receptors, thereby disrupting critical hormonal signaling cascades[12,13]. Studies have shown that the levels of BPA in the plasma, urine, ovarian fluid of PCOS patients are significantly elevated, and are positively correlated with insulin resistance, polycystic ovary morphology, and hepatic steatosis; BPA and phthalates exacerbate the progression of PCOS through mechanisms such as oxidative stress, inflammation, and gut microbiota dysbiosis[14,15]. Environmental Disrupting Chemicals (EDCs) induce hypothalamic neuroinflammation, disrupting the pulsatile GnRH secretion rhythm and causing an imbalance in the LH/FSH ratio, which promotes ovarian hyperandrogenism[16]. Population studies have shown that women with high exposure to EDCs exhibit an increased risk of developing PCOS, and BPA may exacerbate PCOS-related hormonal disturbances by inhibiting ovarian steroidogenic enzyme activity[17,18]. Although existing research suggests a close association between EDCs and PCOS, the specific mechanisms, regulatory targets, and molecular pathways involved in EDCs' effects on PCOS remain unclear.

This study aims to identify key differentially expressed genes (DEGs) associated with EDCs in PCOS by integrating transcriptomic data with the EDC database. Core hub genes were screened using protein-protein interaction (PPI) networks and multiple machine learning algorithms. Diagnostic models were constructed and their predictive performance was evaluated. Gene set enrichment analysis (GSEA) and immune

infiltration analysis were performed to characterize the biological functions and immune landscape associated with the key genes. Molecular docking and molecular dynamics simulations were further conducted to investigate the binding interactions between EDCs and target proteins. The study reveals the molecular characteristics of EDC-related genes in PCOS, provides a diagnostic prediction tool: IRS1 – PRKCQ nomogram, and elucidates the potential mechanism by which EDCs (e.g., BPA) may regulate PCOS through IRS1.

Materials and methods

Data acquisition and study design

Gene expression datasets were retrieved from the Gene Expression Omnibus (GEO; https://www.ncbi.nlm.nih.gov/geo/). The training dataset GSE10946, generated on the GPL570 platform, comprised microarray profiles of follicular theca cells from patients with polycystic ovary syndrome (PCOS, n = 12) and normal controls (n = 11) Three independent GEO datasets, including GSE106724 (GPL21096; granulosa cells, PCOS/control = 6/6), GSE34526 (GPL570; granulosa cells, PCOS/control = 7/3), and GSE98595 (GPL6244; granulosa cells, PCOS/control = 10/6), were employed as external validation cohorts. In addition, endocrine-disrupting chemical (EDC) information were curated from a recently published study[19], yielding 15 compounds, including bisphenols, parabens, and phthalate metabolites (bisphenol A, bisphenol F, ethyl paraben, methyl paraben, n-butyl paraben, n-propyl paraben, mono-benzyl phthalate, mono-(2-ethyl-5-carboxypentyl) phthalate, mono-(2-ethyl-5-hydroxyhexyl)

phthalate, mono-(2-ethylhexyl) phthalate, mono-(2-ethyl-5-oxohexyl) phthalate, mono-ethyl phthalate, mono-iso-butyl phthalate, mono-methyl phthalate, and mono-n-butyl phthalate), which were used for subsequent compound-target interaction analyses to explore potential regulatory links between EDC exposure and PCOS-associated genes.

Identification of EDC-related differentially expressed genes (DEGs)

DEGs between patients with PCOS and normal controls were identified in the training dataset GSE10946 using the limma package in R software (version 4.2.2). Genes with a p-value < 0.05 and |log2 foldchange (FC)| > 0.5 were defined as DEGs. To identify genes potentially associated with EDCs, EDC-related target genes were retrieved from four public databases, including the Comparative Toxicogenomics Database (CTD; http://ctdbase.org/), ChEMBL (https://www.ebi.ac.uk/chembl/), SwissTargetPrediction (http://www.swisstargetprediction.ch/), and PharmMapper (http://www.lilab-ecust.cn/pharmmapper/). Subsequently, the intersection between EDC-related genes and the identified DEGs was determined using the ggvenn R package, yielding EDC-associated DEGs. The interaction network between EDC-related compounds and EDC-associated DEGs was visualized using Cytoscape software (version 3.10.3).

Functional enrichment analysis of EDC-associated DEGs

Functional enrichment analyses were performed to explore the biological functions and signaling pathways associated with the identified EDC-associated DEGs. Gene Ontology (GO) enrichment analyses, including biological process (BP), cellular

component (CC), and molecular function (MF) categories, as well as Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway enrichment analysis, were conducted using the clusterProfiler R package. Enrichment significance was evaluated using a hypergeometric test with multiple testing correction by the Benjamini-Hochberg method. The top 10 enriched GO terms and KEGG pathways were visualized using the ggplot2 package.

Protein-protein interaction (PPI) network construction and candidate gene identification

The EDC-associated DEGs were uploaded to the Search Tool for the Retrieval of Interacting Genes/Proteins (STRING) database (https://string-db.org/), with the organism restricted to Homo sapiens and the minimum required interaction score set to a combined score > 0.15. The resulting interaction network was imported into Cytoscape software for visualization. To identify candidate genes within the PPI network, the cytoHubba plugin was applied, and three topological algorithms, including Degree, Betweenness, and Closeness, were used independently to rank genes according to their network importance. The top 20 genes ranked by each algorithm were intersected using the ggvenn R package, and the overlapping genes were defined as candidate genes for downstream analyses.

Machine learning model construction to identify hub genes

To construct predictive models based on candidate genes, a total of 12 machine learning algorithms were implemented in R software by integrating multiple R packages, including randomForestSRC, plsRglm, gbm, e1071, xgboost, mboost, BART,

and snowfall. The applied algorithms comprised least absolute shrinkage and selection operator (LASSO), Ridge regression, Elastic Net (Enet), stepwise generalized linear model (Stepglm), support vector machine (SVM), gradient boosting with component-wise linear model (glmBoost), linear discriminant analysis (LDA), partial least squares regression for generalized linear model (plsRglm), Random Forest (RF), generalized boosted regression models (GBM), extreme gradient boosting (XGBoost), and Naive Bayes. For Elastic Net models, different mixing parameters (α = 0.1–0.9) were tested to optimize the balance between L1 and L2 regularization. In addition, several hybrid models combining feature selection and classification algorithms (such as LASSO + LDA, glmBoost + Enet, LASSO + XGBoost) were further evaluated. Model training and tuning were performed in the training dataset and the diagnostic performance of each model and combined modeling strategies was assessed by calculating the area under the receiver operating characteristic curve (AUC). The trained models were subsequently applied to an independent validation cohort, which was constructed by merging the GSE106724, GSE34526, and GSE98595 datasets and correcting for batch effects, to assess model generalizability. Model performance across the training and validation datasets was summarized and visualized as a heatmap using the pheatmap R package, with higher AUC values indicating superior discriminative ability.

Hub gene expression analysis and diagnostic performance evaluation

The expression levels of identified hub genes were compared between PCOS patients and normal controls in both the merged validation cohort and the training

dataset. Gene expression values were extracted from normalized microarray data, and between-group differences were visualized using boxplots generated with the ggplot2 R package. Statistical comparisons between PCOS and control groups were performed using the Wilcoxon rank-sum test, and statistical significance was defined as $p < 0.05$. Receiver operating characteristic (ROC) curve analysis was performed using the pROC R package to evaluate the diagnostic performance of hub genes. ROC curves were generated in both the training dataset and the merged validation cohort, and the AUC was calculated to quantify discriminative ability. An AUC value closer to 1.0 indicated stronger diagnostic performance.

Nomogram construction and clinical performance evaluation

A nomogram based on the hub genes insulin receptor substrate 1 (IRS1) and protein kinase C theta (PRKCQ) was developed using multivariable logistic regression with the rms R package to predict disease risk. Hub genes were incorporated as continuous variables, and each predictor was assigned a weighted score proportional to its regression coefficient. The total score was calculated by summing the individual points for each predictor, and the corresponding predicted probability was obtained according to the total points scale. The discriminative ability of the nomogram model was evaluated using ROC curve analysis. Calibration of the nomogram was assessed by comparing predicted probabilities with observed outcomes using a calibration curve. Internal validation was performed using bootstrap resampling with 1,000 repetitions to reduce overfitting. The Hosmer-Lemeshow goodness-of-fit test was applied to evaluate the agreement between predicted and observed risks, and a non-significant p value

indicated good calibration performance. To evaluate the potential clinical utility of the nomogram, decision curve analysis (DCA) was conducted using the rmda R package. Net benefit was calculated across a range of threshold probabilities to compare the clinical value of the nomogram with individual predictors and default strategies ("treat all" or "treat none"). A higher net benefit across a wider range of threshold probabilities indicated superior clinical usefulness of the model.

Gene set enrichment analysis (GSEA)

To explore the biological functions associated with the hub genes IRS1 and PRKCQ, GSEA was performed using the clusterProfiler R package. The curated KEGG gene sets were obtained from the Molecular Signatures Database (MSigDB, https://www.gsea-msigdb.org/) using the c2.cp.kegg.v7.2.symbols.gmt reference gene set. In the training dataset GSE10946, Spearman correlation coefficients were calculated between each hub gene and all other genes using the psych R package. Genes were ranked according to the correlation coefficients in descending order, and GSEA was conducted separately for IRS1 and PRKCQ. Enriched pathways with a normalized enrichment score ($|NES| > 1$, false discovery rate (FDR) < 0.25, and nominal p value < 0.05 were considered statistically significant. The top five significantly enriched pathways were visualized using enrichment plots generated by the enrichplot R package.

Immune infiltration correlation analysis

To investigate the immunological relevance of IRS1 and PRKCQ, immune cell infiltration levels were estimated using the ssGSEA algorithm. Spearman correlation analysis was performed to assess the associations between hub gene expression levels

and immune cell infiltration scores. Correlation coefficients and corresponding p-values were calculated using the psych package, and the results were visualized as a heatmap using the pheatmap R package.

Molecular docking analysis

The molecular structures of candidate compounds were retrieved from the PubChem database (https://pubchem.ncbi.nlm.nih.gov/) in three-dimensional (3D) format. Protein structures corresponding to IRS1 and PRKCQ were obtained from the Protein Data Bank (PDB) (https://www.rcsb.org/). Molecular docking was performed using the CB-DOCK2 web server (http://183.56.231.194:8001/cb-dock2/). The molecular interactions between ligands and target proteins, including hydrogen bonds, hydrophobic interactions, and key residue contacts, were visualized and analyzed using PyMOL software.

Molecular dynamics simulation

Molecular dynamics simulations were performed using GROMACS (version 2024.4) to evaluate the stability and dynamic behavior of protein-ligand complexes. The AMBER99SB-ILDN force field was applied for protein parameters, while the TIP3P (three-point transferable intermolecular potential) water model was used to solvate the system. Each protein-ligand complex was placed in a cubic simulation box, ensuring a minimum distance of 1.0 nm between the protein surface and the box boundary. The system was then solvated and neutralized by adding appropriate counterions. Energy minimization was conducted using the steepest descent algorithm to remove unfavorable steric contacts. Subsequently, the system underwent

equilibration in two phases: a constant number, volume, and temperature ensemble (NVT) followed by a constant number, pressure, and temperature ensemble (NPT). During both equilibration phases, temperature coupling was achieved using the V-rescale thermostat with a reference temperature of 300 K and a time step of 2 fs. Each equilibration phase was carried out for 100 ps to ensure system stability. After equilibration, a production molecular dynamics simulation was performed for 100 ns under periodic boundary conditions. Trajectory analyses were conducted using built-in GROMACS tools to assess the dynamic stability of the complexes. The root mean square deviation (RMSD) of the protein backbone, root mean square fluctuation (RMSF) of individual residues, total energy of the system, and the number of hydrogen bonds between the protein and ligand were calculated throughout the simulation. Protein–ligand complexes with predicted binding affinities lower than −4 kcal/mol were selected for molecular dynamics simulations.

Statistical analysis

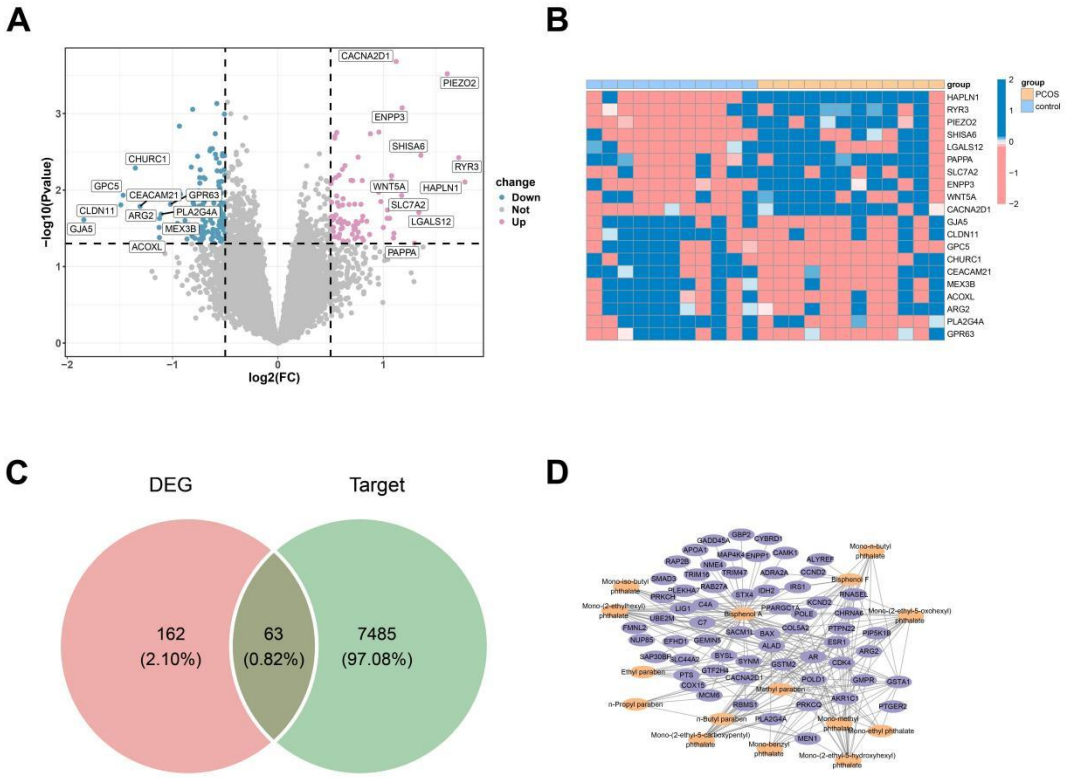
All statistical analyses were performed using R software (version 4.2.2). Between-group comparisons were conducted using the Wilcoxon rank-sum test, and correlation analyses were performed using Spearman's rank correlation coefficient. Multiple hypothesis testing was corrected using the Benjamini–Hochberg false discovery rate (FDR) method. All tests were two-sided, and $P < 0.05$ was considered statistically significant unless otherwise specified.

Results

Identification of EDC-related DEGs in PCOS

Based on the GSE10946 training dataset, differential expression analysis between PCOS patients and controls was performed using the limma package. With the thresholds set at $p < 0.05$ and $|\log_2 FC| > 0.5$, a total of 225 DEGs were identified, including 137 downregulated and 88 upregulated genes in the PCOS group. The volcano plot depicted the overall distribution of DEGs based on fold change and statistical significance, with the top 10 upregulated and top 10 downregulated genes specifically highlighted (Figure 1A). Simultaneously, a heatmap of these selected genes demonstrated a clear separation between PCOS and control samples (Figure 1B). To investigate the potential involvement of EDCs in PCOS, EDC-related target genes were collected from four databases, including CTD, ChEMBL, Swiss Target Prediction, and PharmMapper, and their union yielded 7,548 EDC-associated target genes. Intersection of these targets with the identified DEGs resulted in 63 overlapping genes, which were defined as EDC-related DEGs (Figure 1C). Based on the interactions between these genes and EDC-associated compounds, a drug-target interaction network was constructed, comprising 78 nodes and 196 edges, including 63 target gene nodes and 15 compound nodes (Figure 1D), highlighting the multi-target regulatory characteristics of EDCs.

Fig.1

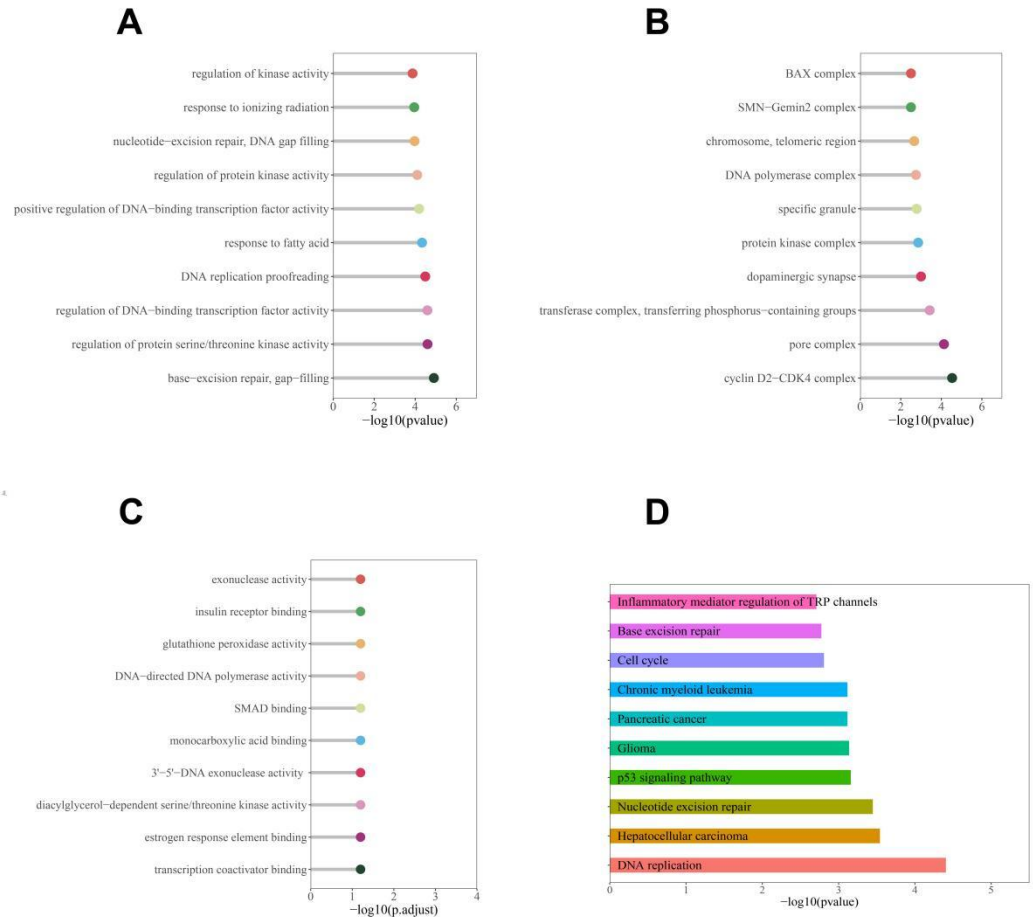


Functional enrichment analysis of EDC-related DEGs

To elucidate the biological functions and pathways associated with the identified EDC-related DEGs, GO and KEGG enrichment analyses were performed. GO enrichment analysis yielded a total of 569 significantly enriched GO terms, including 385 BP, 30 CC, and 154 MF. In the BP category, EDC-related DEGs were mainly enriched in base-excision repair, regulation of protein serine/threonine kinase activity, and regulation of DNA-binding transcription factor activity (Figure S1A). CC enrichment analysis indicated that these genes were predominantly associated with cyclin D2-CDK4 complex, pore complex, and transferase complex (Figure S1B). MF analysis revealed significant enrichment in transcription coactivator binding, estrogen

response element binding, and diacylglycerol-dependent serine/threonine kinase activity (Figure S1C). Furthermore, KEGG pathway analysis identified 45 significantly enriched pathways, with prominent enrichment in DNA replication, nucleotide excision repair, base excision repair, cell cycle, p53 signaling pathway, and several cancer-related pathways, including hepatocellular carcinoma, glioma, pancreatic cancer, and chronic myeloid leukemia (Figure S1D).

Fig.S1

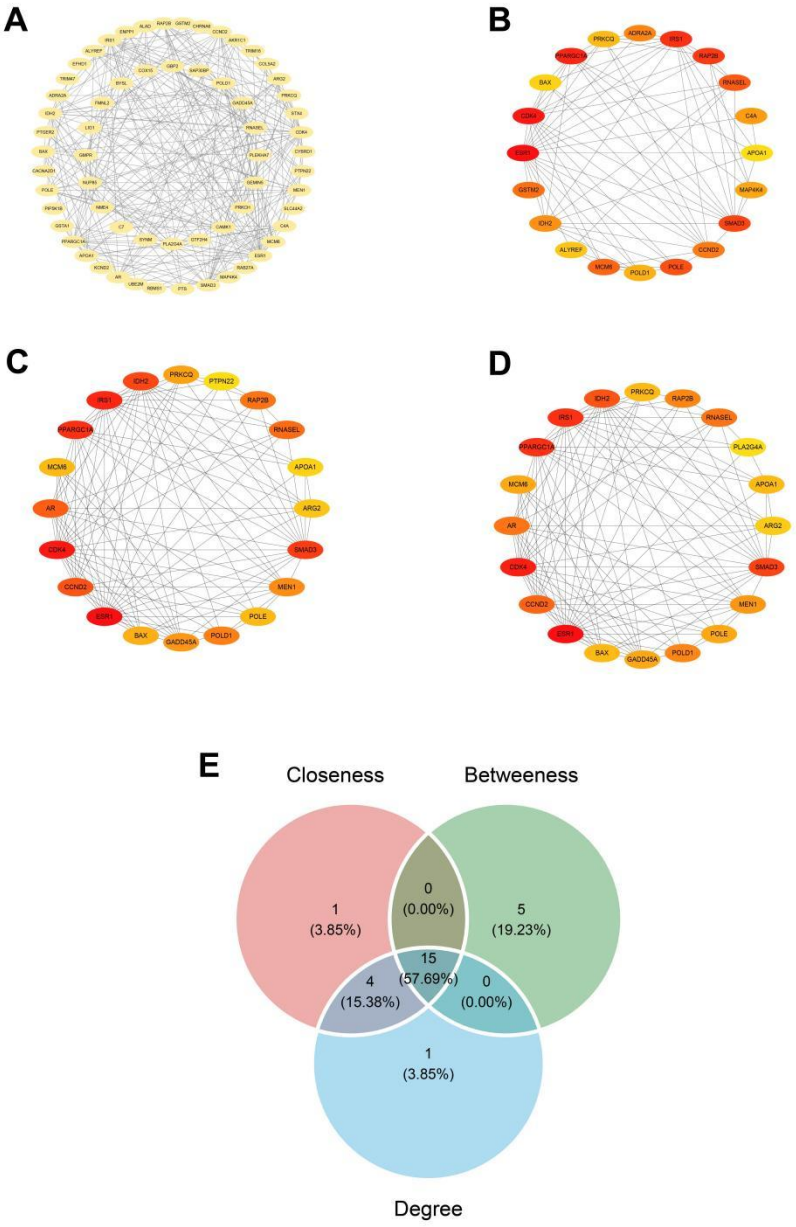


PPI network construction and screening of candidate genes

To investigate the interaction landscape among the EDC-related DEGs, a PPI network was constructed using the STRING database with a minimum confidence score

of >0.15. The resulting PPI network was visualized in Cytoscape and comprised 62 nodes and 267 edges (Figure 2A). Candidate genes were identified using the cytoHubba plugin in Cytoscape. Betweenness, Closeness, and Degree algorithms were applied to rank genes by network centrality, and the top 20 genes from each method were intersected using the ggvenn package, yielding 15 overlapping genes for further analysis (Figures 2B-2E).

Fig.2

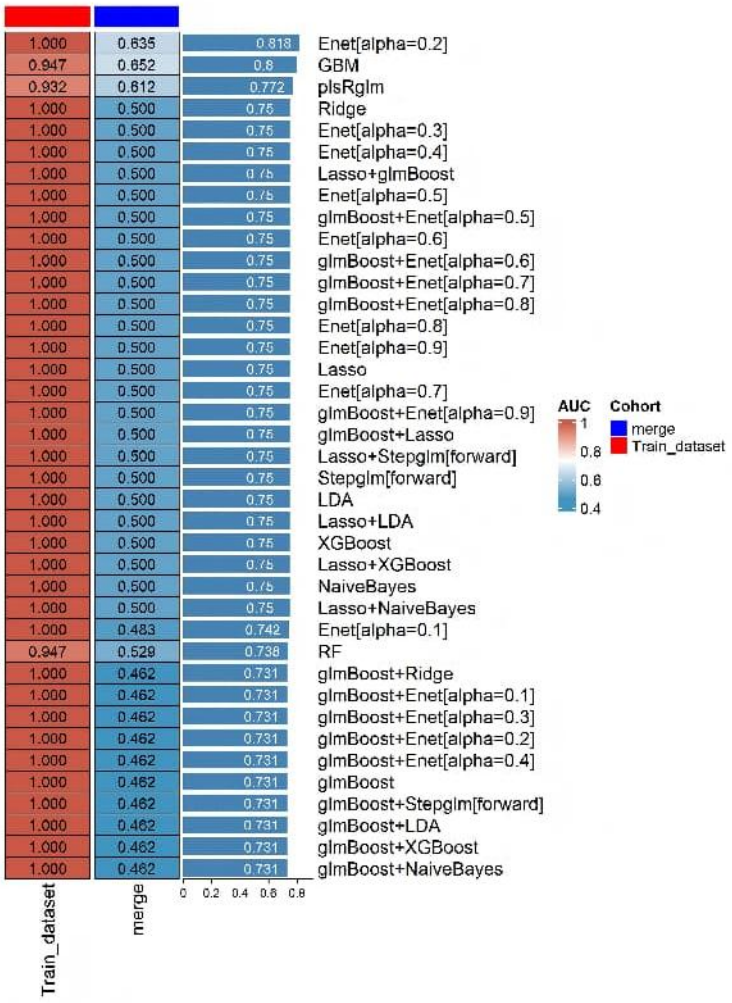


Machine learning-based identification of EDC-related hub DEGs

To further identify EDC-related hub DEGs, an integrative machine learning framework was applied to the candidate genes. A total of 12 machine learning algorithms, including LASSO, Ridge, Enet, Stepglm, SVM, glmBoost, LDA, plsRglm, RF, GBM, XGBoost, and Naive Bayes, were implemented. The GSE10996 dataset was used as the training cohort, while GSE106724, GSE34526, and GSE98595 were

merged and batch-corrected to serve as the validation cohort. The Elastic Net model with α = 0.2 (Enet[alpha = 0.2]) demonstrated the optimal predictive performance. Based on this optimal model, 13 hub genes were ultimately identified, including estrogen receptor 1 (ESR1), cyclin-dependent kinase 4 (CDK4), IRS1, peroxisome proliferator-activated receptor gamma coactivator 1 alpha (PPARGC1A), SMAD family member 3 (SMAD3), isocitrate dehydrogenase (NADP(+)) 2, mitochondrial (IDH2), cyclin D2 (CCND2), ribonuclease L (RNASEL), RAP2B, member of RAS oncogene family (RAP2B), DNA polymerase delta 1, catalytic subunit (POLD1), PRKCQ, minichromosome maintenance complex component 6 (MCM6), and apolipoprotein A1 (APOA1) (Figure 3).

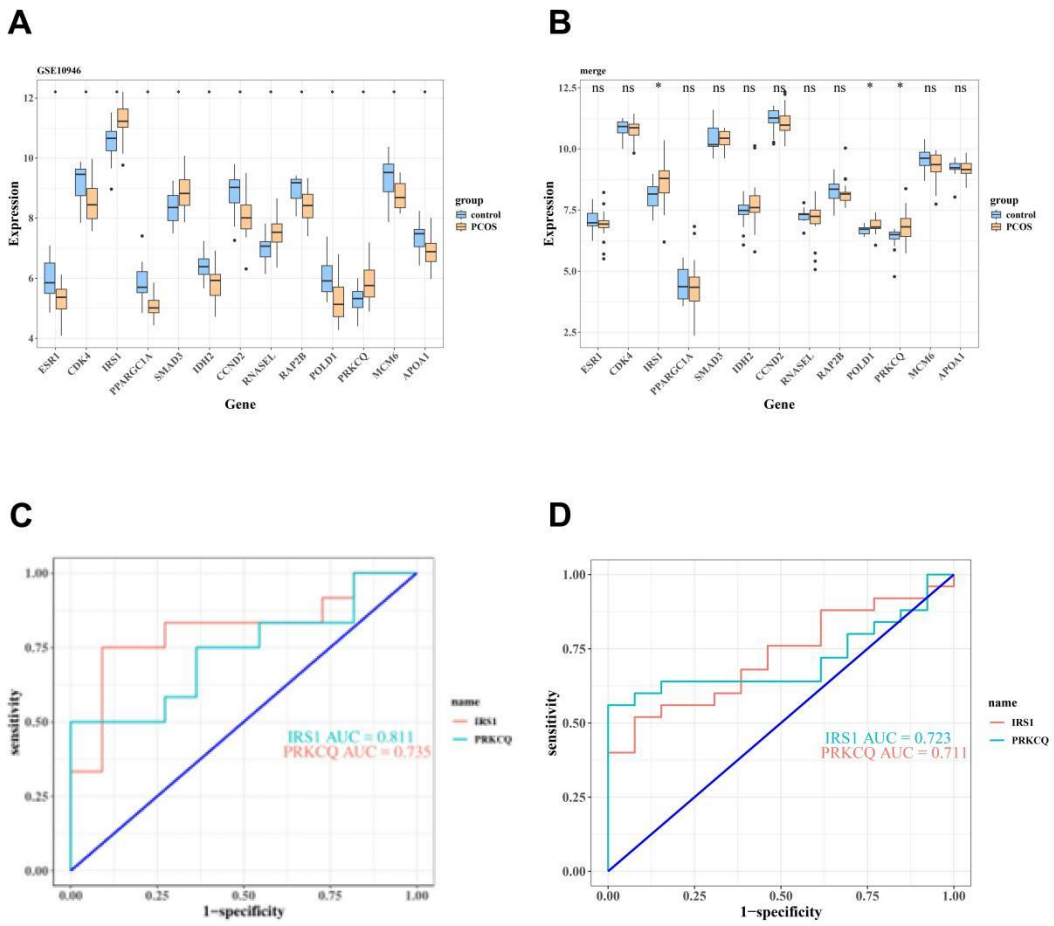
Fig.3



Validation of hub gene expression and diagnostic performance in PCOS

To clarify the expression patterns of the hub genes between PCOS and control groups, their expression levels were evaluated in the training dataset GSE10946 and in a merged validation cohort generated by integrating the GSE106724, GSE34526, and GSE98595 datasets. The results revealed that IRS1 and PRKCQ were significantly upregulated in PCOS compared with control samples in both the training and merged validation cohorts (Figures 4A and 4B). Subsequently, ROC curve analyses were performed for IRS1 and PRKCQ. The results demonstrated that both genes achieved AUC values greater than 0.7 in the training and validation cohorts, indicating good discriminatory ability between PCOS and control samples (Figures 4C and 4D).

Fig.4

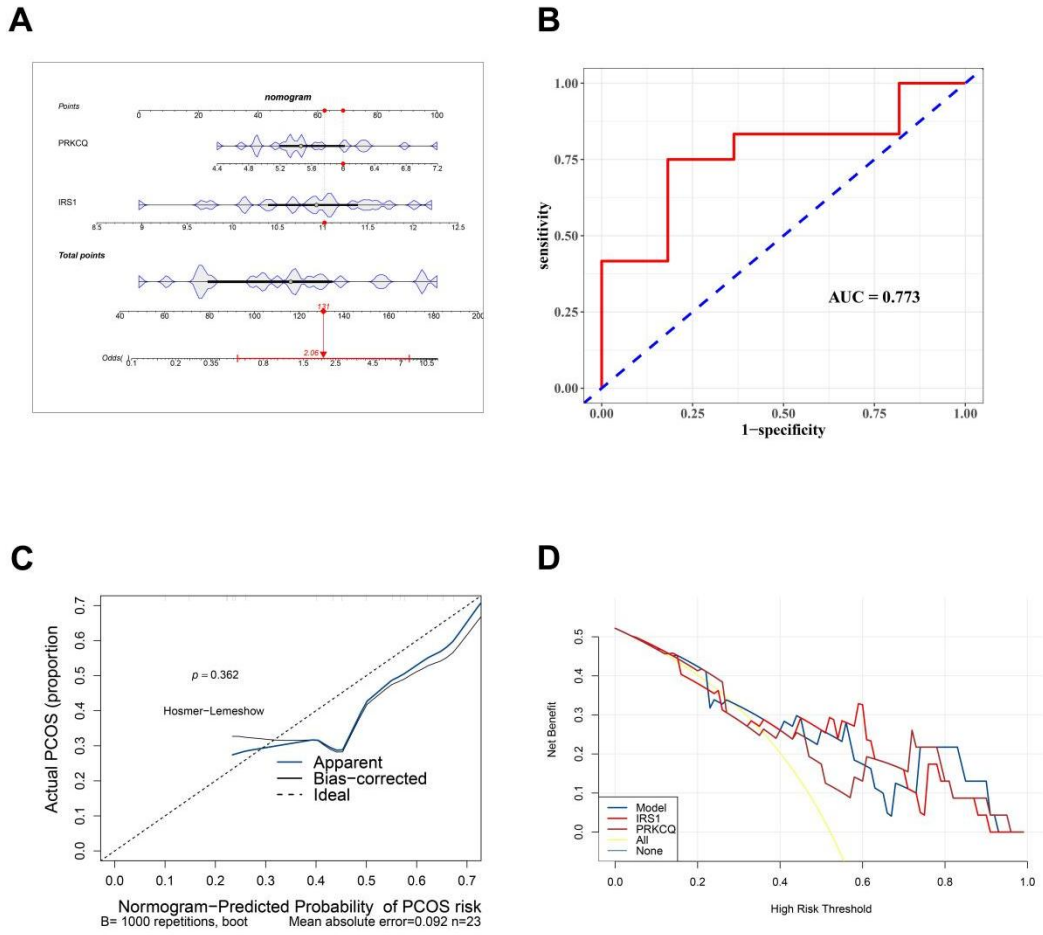


Construction and validation of an IRS1-PRKCQ-based nomogram for PCOS risk prediction

Based on the two hub genes, IRS1 and PRKCQ, a predictive nomogram was constructed to estimate the individual risk of PCOS. Each variable was assigned a weighted score proportional to its contribution to disease risk, and the total score corresponded to the predicted probability of PCOS, enabling convenient clinical application (Figure 5A). The nomogram achieved an AUC of 0.773, indicating good discrimination between PCOS patients and controls (Figure 5B). Model calibration was

assessed by calibration curve analysis with 1,000 bootstrap resamples. The predicted probabilities closely aligned with the observed outcomes, with the bias-corrected curve approaching the ideal reference line. The Hosmer–Lemeshow test yielded a non-significant result ($p = 0.362$), further supporting good model calibration and consistency (Figure 5C). To evaluate the potential clinical utility of the nomogram, DCA was performed. The nomogram provided a higher net benefit across a wide range of threshold probabilities compared with strategies based on PRKCQ or IRS1 alone, as well as the "treat-all" and "treat-none" approaches (Figure 5D). Collectively, these findings demonstrate that the IRS1-PRKCQ-based nomogram exhibits predictive accuracy, good calibration, and favorable clinical utility, supporting its potential value as an individualized risk assessment tool for PCOS.

Fig.5



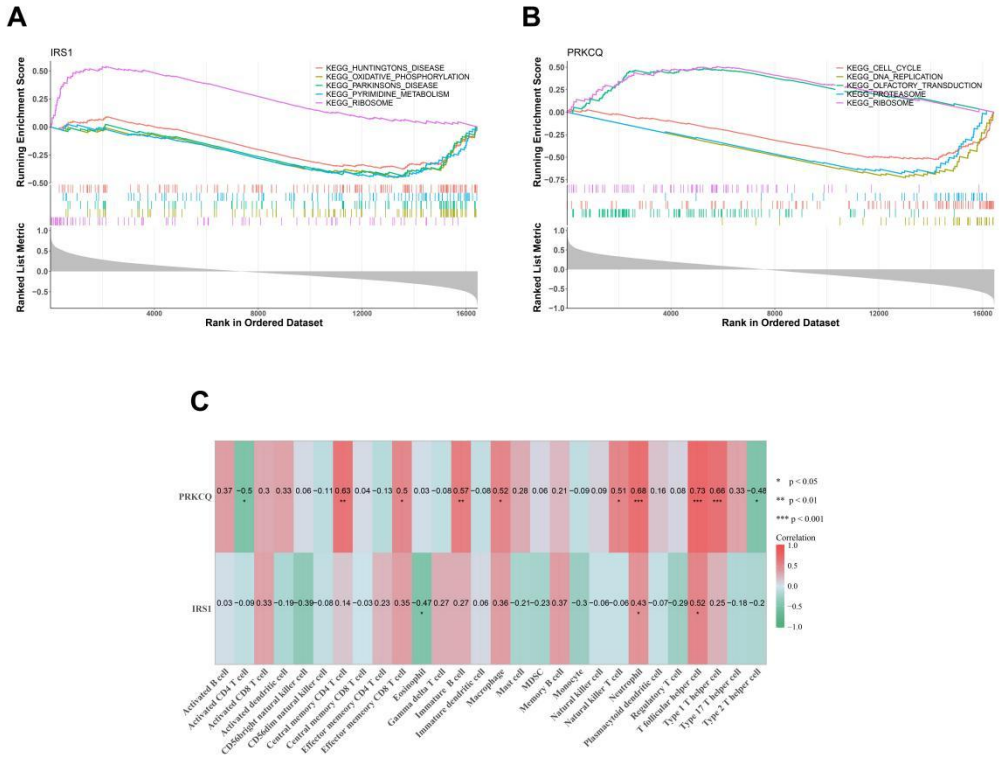
363

364 **Functional enrichment and immune correlation analysis of IRS1and PRKCQ**

365 To elucidate the biological functions and signaling pathways associated with IRS1
366 and PRKCQ, GSEA was performed using the MSigDB c2.cp.kegg.v7.2.symbols.gmt
367 gene set as the reference. For IRS1, a total of 43 significantly enriched KEGG pathways
368 were identified. The top five enriched pathways were ribosome, oxidative
369 phosphorylation, pyrimidine metabolism, Huntington's disease, and Parkinson's
370 disease (Figure 6A), indicating that IRS1-associated gene expression profiles are
371 closely linked to translational activity, energy metabolism, and DNA synthesis-related
372 processes. For PRKCQ, 61 significantly enriched KEGG pathways were detected. The
373 top five pathways included cell cycle, DNA replication, proteasome, ribosome, and

olfactory transduction pathways (Figure 6B), suggesting that PRKCQ may be involved in RNA processing, cell cycle regulation, and protein turnover. We applied the ssGSEA algorithm to quantify the infiltration characteristics of 28 immune cell subsets and subsequently analyzed the correlations between hub genes and immune cell populations (Figure 6C). The results showed that IRS1 was significantly negatively correlated with eosinophils ($r = -0.47$, $p < 0.05$), while showing positive correlations with neutrophils ($r = 0.43$, $p < 0.05$) and T follicular helper cells ($r = 0.52$, $p < 0.05$). PRKCQ was significantly associated with multiple immune cell types, including activated CD4 T cells ($r = -0.50$, $p < 0.05$), central memory CD4 T cells ($r = 0.63$, $p < 0.01$), effector memeory CD8 T cell ($r = 0.50$, $p < 0.05$), immature B cells ($r = 0.57$, $p < 0.01$), Macrophage ($r = 0.52$, $p < 0.05$), Natural killer T cell ($r = 0.51$, $p < 0.05$), Neutrophil ($r = 0.68$, $p < 0.001$), T follicular helper cells ($r = 0.73$, $p < 0.001$), type 1 T helper cells ($r = 0.66$, $p < 0.001$), and Type 2 T helper cell ($r = -0.48$, $p < 0.05$). These findings suggest that PRKCQ and IRS1 may contribute to immune microenvironment dysregulation in PCOS by modulating distinct immune cell infiltration patterns.

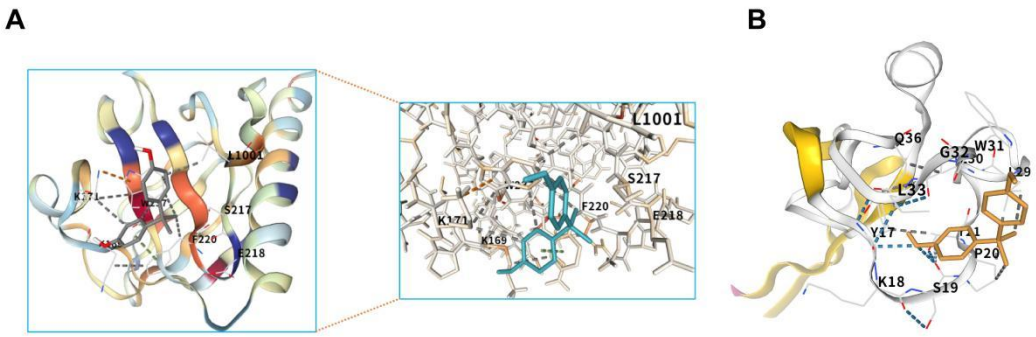
Fig.6



Molecular docking analysis of hub genes

Molecular docking analysis demonstrated that Bisphenol A (BPA) could stably bind to IRS1 and PRKCQ with binding affinities of −4.8 and −4.9 kcal/mol, respectively. The docking conformations showed that the ligand was well accommodated within the predicted binding pocket and interacted with multiple key residues through hydrogen bonds and hydrophobic interactions (Figures 7A and 7B).

Fig.7



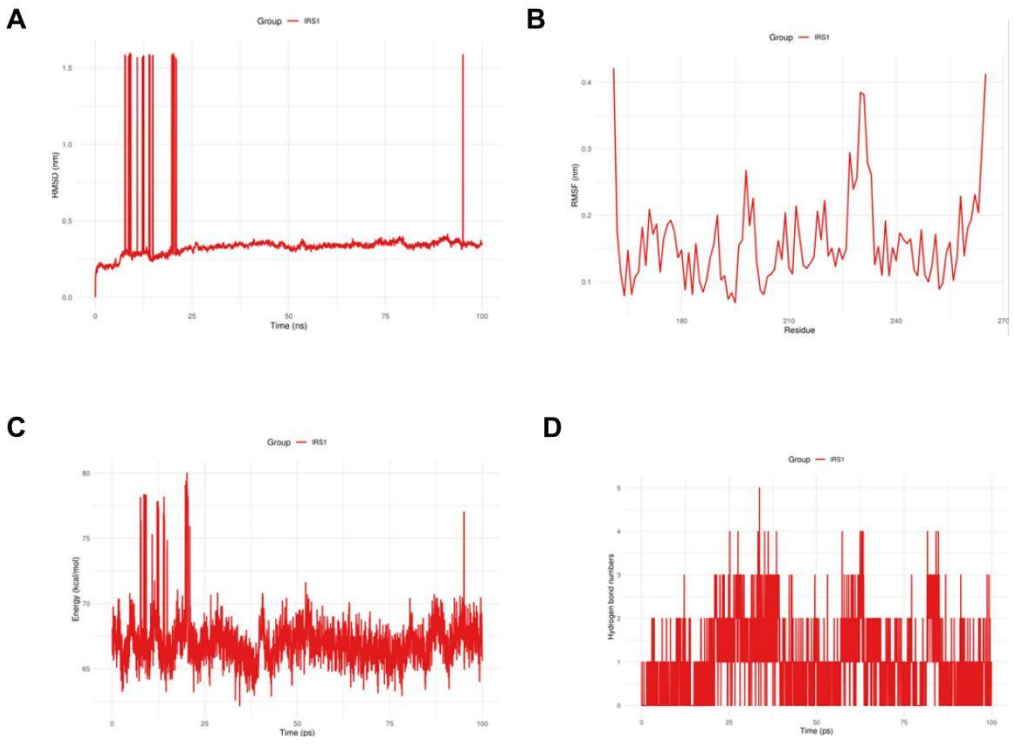
398

399 **Molecular dynamics simulation of the IRS1-Bisphenol A complex**

400 To further evaluate the structural stability of the IRS1-Bisphenol A complex,
401 molecular dynamics simulations were performed. The RMSD analysis showed that,
402 after an initial equilibration phase, the complex rapidly reached a stable conformation
403 and maintained a relatively constant RMSD throughout the simulation, indicating
404 overall structural stability (Figure 8A). RMSF analysis revealed that most IRS1
405 residues exhibited low flexibility (RMSF < 0.2 nm), with only limited regions
406 displaying higher fluctuations, primarily at the N-terminus, residues 230–240, and the
407 C-terminus (RMSF > 0.3 nm) (Figure 8B). The total energy profile of the IRS1-
408 Bisphenol A complex further supported system stability. Although relatively large
409 energy fluctuations were observed during the initial 0–25 ps of the simulation
410 (approximately 65–80 kcal/mol), the total energy stabilized between 65 and 72 kcal/mol
411 from 25 to 100 ps, with reduced fluctuation amplitude and no apparent energy drift
412 (Figure 8C). In addition, hydrogen bond analysis demonstrated that IRS1 and Bisphenol
413 A consistently formed 2–3 hydrogen bonds during the simulation, indicating persistent

intermolecular interactions. The dynamic equilibrium of hydrogen bond formation further corroborated the structural stability of the complex, in agreement with the RMSD and energy analyses (Figure 8D). Collectively, these results suggest that Bisphenol A forms a stable complex with IRS1 under simulated physiological conditions.

Fig.8



Discussion

This study integrated GEO data with EDC databases to explore the key EDC-related DEGs in PCOS. Utilizing PPI networks and machine learning, we identified 13 hub genes, among which IRS1 and PRKCQ exhibited stable upregulation in PCOS. We constructed and validated a diagnostic nomogram demonstrating good predictive

performance and potential clinical utility. Further exploration through GSEA, immune infiltration analysis, molecular docking, and molecular dynamics simulations uncovered potential pathogenic mechanisms. The results revealed that IRS1-related genes were significantly enriched in metabolic reprogramming pathways, including ribosome biogenesis, oxidative phosphorylation, and pyrimidine metabolism, while PRKCQ showed a strong association with multiple immune cell infiltration levels. Additionally, BPA formed stable binding interactions with both IRS1 and PRKCQ.

EDCs are linked to the onset of numerous diseases, including cardiovascular conditions, prostate disorders, pancreatic ailments, cancer, and others, posing grave threats to human health[20-22]. In particular, EDCs exert substantial effects on reproductive-related disorders, including recurrent miscarriage, precocious puberty, gestational diabetes mellitus, preeclampsia, and male asthenospermia[22-26]. Furthermore, studies have indicated that emotional and behavioral abnormalities in childhood may be linked to maternal exposure to phthalates (PAEs) during pregnancy[27]. The mechanisms underlying the reproductive toxicity of EDCs are highly complex and remain incompletely understood. Current hypotheses include the following: (1) EDCs interfere with ovarian hormone synthesis and hormone receptor expression, thereby affecting ovarian function and follicular development[28]; (2) EDCs impair mitochondrial function, causing oxidative damage to DNA, proteins, and lipids, which compromises oocyte quality and ovarian function[29]; (3) EDCs regulate epigenetic modifications to influence gene expression related to ovarian function, thereby affecting follicular development[30]. GO/KEGG analysis in our study revealed that EDC-associated DEGs

were significantly enriched in DNA repair, cell cycle regulation, and the p53 pathway. Consequently, we propose that EDC exposure may impair ovarian cell proliferative homeostasis, genomic stability, and metabolic-reproductive coupling.

Insulin receptor substrate 1 (IRS1) is intimately linked to hyperinsulinemia in PCOS[31-33]. Functioning as a key downstream adaptor protein for insulin receptors, IRS1 critically regulates the PI3K/AKT pathway[2] and modulatesluteinized granulosa cells[3]. Furthermore, chronic inflammation and metabolic stress may further drive the inhibitory phosphorylation of IRS1 by activating kinases like S6K and JNK, thereby exacerbating insulin resistance[34]. In this study, IRS1 was markedly upregulated across both the training and validation sets. Genes linked to IRS1 expression demonstrated enrichment in critical pathways, including ribosome biogenesis, oxidative phosphorylation, and pyrimidine metabolism. These findings underscore active metabolic reprogramming, robust energy production, and heightened cellular synthesis. Moreover, the AUC value of IRS1 exceeding 0.7 demonstrates a good discriminative capacity, solidifying its potential as a valuable molecular biomarker for predicting PCOS or assessing the impact of EDC exposure. Notably, EDCs such as BPA have been shown to interfere with IRS1 expression and phosphorylation[35]. Our molecular docking and molecular dynamics simulations suggested stable binding interactions. We hypothesize that by targeting IRS1, EDCs may exacerbate insulin signaling dysfunction, thereby disrupting ovarian function and energy metabolism coupling.

PRKCQ is a key regulator in T cell activation, participating in the NF-κB and AP-1 signaling pathways, and plays a central role in immune synapse formation and inflammatory cytokine production for immune synapse formation and inflammatory cytokine expression[36]. PCOS is widely recognized as a low-grade chronic inflammatory disorder characterized by elevated levels of IL-6, TNF-α, and CRP [37,38]. We hypothesize that PRKCQ may play a role in immune dysregulation-mediated ovarian dysfunction. The expression and enrichment analyses indicated that: PRKCQ is significantly upregulated in PCOS; GSEA enrichment occurs in the cell cycle, DNA replication, and proteasome pathway; it shows significant positive correlation with multiple immune cells (Tfh, memory T, $CD4^+$, B cells, macrophages, etc.); these findings suggest that PRKCQ not only participates in immune activation but may also influence cell proliferation and genomic stability Therefore, PRKCQ can serve as a potential biomarker for the immune subtype of PCOS; it holds promise for distinguishing patients with inflammation-dominant PCOS; and demonstrates good AUC performance (>0.7), indicating certain diagnostic capability. The model constructed in combination with IRS1 shows enhanced predictive efficacy. EDCs have been demonstrated to activate immune responses, inducing Th1/Th17 polarization and macrophage activation[39,40]. As a key molecule in T-cell activation, PRKCQ may serve as a potential target for EDC interference.

This study integrates environmental exposure, machine learning, and molecular simulation to identify the EDC-related regulatory network centered on IRS1 and PRKCQ, constructs an IRS1–PRKCQ nomogram model, and suggests potential stable

interactions between bisphenol A and IRS1, providing a new perspective for understanding the environmental driving mechanisms and molecular heterogeneity of PCOS.

Conclusions

Systematically, this research explored the transcriptional regulatory connection between environmental endocrine disruptors (EDCs) and polycystic ovary syndrome (PCOS) using an integrated bioinformatics and machine learning framework. Based on our findings, IRS1 and PRKCQ appear to represent potentially significant hub genes that may play central roles in the pathogenesis and progression of PCOS.

The combined IRS1-PRKCQ model demonstrated exceptional diagnostic performance and significant clinical application potential, underscoring the profound value of integrated immune-metabolic signatures for PCOS risk stratification. Overall, this work proposes a novel immune-metabolic framework linking environmental EDC exposure to PCOS disease heterogeneity and identifies promising biomarkers for individualized risk assessment. These findings establish a crucial theoretical foundation for subsequent experimental validation and offer fresh insights for developing environment-exposure-focused diagnostic strategies and mechanistic investigations into PCOS.

Limitations

There are still several limitations to our study that should be acknowledged. (1) Retrospective Analysis of Public Datasets: This study relied exclusively on publicly

available microarray datasets, which may introduce selection bias and limit the generalizability of findings. The lack of prospective, clinically annotated cohorts with detailed EDC exposure assessments restricts our ability to establish causal relationships between specific EDCs and PCOS pathogenesis. (2) Validation in Independent Clinical Populations: While machine learning models demonstrated promising diagnostic performance in silico, the IRS1-PRKCQ signature and nomogram require prospective validation in independent, ethnically diverse clinical cohorts with measured EDC exposure levels to confirm their clinical utility and generalizability across different populations. (3) In Vitro Confirmation of Mechanistic Insights: Although molecular docking and dynamics simulations suggested stable interactions between bisphenol A and hub proteins, these computational predictions lack experimental validation. Functional studies, such as gene knockdown/overexpression experiments and EDC exposure assays in PCOS-relevant cellular or animal models, are needed to substantiate the proposed mechanistic pathways. (4) Limited EDC Diversity in Target Analysis: The study primarily focused on bisphenol A as a representative EDC, while numerous other endocrine-disrupting chemicals (such as phthalates, dioxins, and persistent organic pollutants) also contribute to PCOS risk. The computational pipeline did not comprehensively evaluate the full spectrum of EDCs, potentially overlooking compound-specific transcriptional mechanisms and limiting the generalizability of our findings to broader environmental exposures. (5) Single-Cell Resolution Absence: The bulk transcriptomic analysis employed in this study may mask cellular heterogeneity within ovarian tissue and distinct immune cell populations. Critical cell-type-specific

gene expression patterns and EDC-responsive subpopulations could be obscured, as bulk analysis averages signals across diverse cell types. Single-cell RNA sequencing approaches would provide deeper insights into cell-specific EDC responses and precise mechanisms of IRS1/PRKCQ dysregulation in distinct ovarian and immune cell lineages.

Abbreviations

EDCs	Endocrine-disrupting chemicals
DEGs	Differentially expressed genes
PCOS	Polycystic ovary syndrome
GSEA	Gene set enrichment analysis
PPI	Protein-protein interaction
GO	Gene Ontology
BP	Biological process
CC	Cellular component
MF	Molecular function
KEGG	Kyoto Encyclopedia of Genes and Genomes
LASSO	Least absolute shrinkage and selection operator
Enet	Elastic Net
Stepglm	Stepwise generalized linear model
SVM	Support vector machine
glmBoost	Gradient boosting with component-wise linear model
LDA	Linear discriminant analysis
plsRglm	Partial least squares regression for generalized linear model
RF	Random Forest

GBM	Generalized boosted regression models
XGBoost	Extreme gradient boosting
DCA	Decision curve analysis
NES	Normalized enrichment score
FDR	False discovery rate
AUC	The area under the receiver operating characteristic curve
ROC	Receiver operating characteristic
PDB	Protein Data Bank
TIP3P	Three-point transferable intermolecular potential
NVT	Constant number, volume, and temperature ensemble
NPT	Constant number, pressure, and temperature ensemble
RMSD	The root mean square deviation
RMSF	The root mean square fluctuation
FDR	False discovery rate
BPA	Bisphenol A
IRS1	Insulin receptor substrate 1
PRKCQ	Protein kinase C theta

Acknowledgements

Not applicable.

Author contributions

All authors developed the idea. Xin Liu: Writing-original draft, Visualization, Validation, Methodology, Investigation, Formal analysis, Data curation. Lei Zhang: Writing-original draft, software, Investigation, Data curation. Sai Han: Investigation, Formal analysis, Data curation. Jia Liu: Investigation, Formal analysis, Data curation.

Zhihong Wu: Investigation, Data curation, Funding acquisition. Li Gong: Writing-review & editing, Methodology, Funding acquisition， Conceptualization, Resources, Supervision, Validation, Visualization, Formal analysis, Data curation, Project administration. All authors read and approved the final manuscript.

Funding

This research is supported by grants from the China Postdoctoral Science Foundation (2019M652410); Natural Science Foundation of Shandong Province (ZR2022MH045).

Data availability

The datasets analyzed during the current study are available in the Gene Expression Omnibus (GEO) repository (https://www.ncbi.nlm.nih.gov/geo/) under accession numbers GSE10946, GSE106724, GSE34526, and GSE98595. Additional public resources used in this study include CTD (https://ctdbase.org/), ChEMBL (https://www.ebi.ac.uk/chembl/), SwissTargetPrediction (https://www.Swisstargetprediction.ch/), PharmMapper (https://www.lilab-ecust.cn/pharmmapper/), STRING (https://string-db.org/), PubChem (https://pubchem.ncbi.nlm.nih.gov/), the Protein Data Bank (PDB; https://www.rcsb.org/), and the Molecular Signatures Database (MSigDB; https://www.gsea-msigdb.org/gsea/msigdb). All data generated or analyzed during this study are available from the corresponding author upon reasonable request.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no conflicts of interest.

Clinical trial number

Not applicable.

References

1 Siddiqui, S., Mateen, S., Ahmad, R. & Moin, S. A brief insight into the etiology, genetics, and immunology of polycystic ovarian syndrome (PCOS). *Journal of assisted reproduction and genetics* **39**, 2439-2473, doi:10.1007/s10815-022-02625-7 (2022).

2 Singh, S. *et al.* Polycystic Ovary Syndrome: Etiology, Current Management, and Future Therapeutics. *Journal of clinical medicine* **12**, doi:10.3390/jcm12041454 (2023).

3 Naeem, I. *et al.* Polycystic ovarian syndrome a risk factor for non-communicable diseases: insights into recent research and prevention approaches. *Journal of ovarian research* **18**, 219, doi:10.1186/s13048-025-01741-z (2025).

4 Henney, A. E. *et al.* Risk of Type 2 Diabetes, MASLD and Cardiovascular Disease in People Living With Polycystic Ovary Syndrome. *The Journal of clinical endocrinology and metabolism* **110**, 1235-1246, doi:10.1210/clinem/dgae481 (2025).

5 Pililis, S. *et al.* The Cardiometabolic Risk in Women with Polycystic Ovarian Syndrome (PCOS): From Pathophysiology to Diagnosis and Treatment. *Medicina (Kaunas, Lithuania)* **60**, doi:10.3390/medicina60101656 (2024).

6 Rawat, K., Sandhu, A., Gautam, V., Saha, P. K. & Saha, L. Role of genomic DNA methylation in PCOS pathogenesis: a systematic review and meta-analysis involving case-controlled clinical studies. *Molecular human reproduction* **28**, doi:10.1093/molehr/gaac024 (2022).

7 Li, P. *et al.* Perturbations in gut microbiota composition in patients with polycystic ovary syndrome: a systematic review and meta-analysis. *BMC medicine* **21**, 302, doi:10.1186/s12916-023-02975-8 (2023).

8 Huang, X., Luo, X., Huang, S., Chen, X. & Qiu, L. Inhibition of FoxO1 alleviates polycystic ovarian syndrome by reducing inflammation and the immune response. *Functional & integrative genomics* **24**, 6, doi:10.1007/s10142-024-01284-4

(2024).
9 Gautam, R. *et al.* The Role of Lifestyle Interventions in PCOS Management: A Systematic Review. *Nutrients* **17**, doi:10.3390/nu17020310 (2025).
10 Melin, J. *et al.* Metformin and Combined Oral Contraceptive Pills in the Management of Polycystic Ovary Syndrome: A Systematic Review and Meta-analysis. *The Journal of clinical endocrinology and metabolism* **109**, e817-e836, doi:10.1210/clinem/dgad465 (2024).
11 Kahn, L. G., Philippat, C., Nakayama, S. F., Slama, R. & Trasande, L. Endocrine-disrupting chemicals: implications for human health. *The lancet. Diabetes & endocrinology* **8**, 703-718, doi:10.1016/s2213-8587(20)30129-7 (2020).
12 Yilmaz, B., Terekeci, H., Sandal, S. & Kelestimur, F. Endocrine disrupting chemicals: exposure, effects on human health, mechanism of action, models for testing and strategies for prevention. *Reviews in endocrine & metabolic disorders* **21**, 127-147, doi:10.1007/s11154-019-09521-z (2020).
13 Varticovski, L., Stavreva, D. A., McGowan, A., Raziuddin, R. & Hager, G. L. Endocrine disruptors of sex hormone activities. *Molecular and cellular endocrinology* **539**, 111415, doi:10.1016/j.mce.2021.111415 (2022).
14 Ozga, M. & Jurewicz, J. Environmental exposure to selected non-persistent endocrine disrupting chemicals and polycystic ovary syndrome: a systematic review. *International journal of occupational medicine and environmental health* **38**, 98-121, doi:10.13075/ijomeh.1896.02551 (2025).
15 Lemogne Robert, A., Oliveira Torres, E. & Jota Baptista, C. Effects of endocrine disruptive chemicals (EDCs) and therapeutic approaches to the polycystic ovary syndrome (PCOS): A current state-of-the-art. *Environmental toxicology and pharmacology* **120**, 104842, doi:10.1016/j.etap.2025.104842 (2025).
16 Stathori, G. *et al.* Endocrine-Disrupting Chemicals, Hypothalamic Inflammation and Reproductive Outcomes: A Review of the Literature. *International journal of molecular sciences* **25**, doi:10.3390/ijms252111344 (2024).
17 Srnovršnik, T., Virant-Klun, I. & Pinter, B. Polycystic Ovary Syndrome and Endocrine Disruptors (Bisphenols, Parabens, and Triclosan)-A Systematic Review. *Life (Basel, Switzerland)* **13**, doi:10.3390/life13010138 (2023).
18 Lazúrová, Z., Figurová, J., Hubková, B., Mašlanková, J. & Lazúrová, I. Urinary bisphenol A in women with polycystic ovary syndrome - a possible suppressive effect on steroidogenesis? *Hormone molecular biology and clinical investigation* **42**, 303-309, doi:10.1515/hmbci-2020-0032 (2021).
19 Lin, Y. *et al.* Novel insights into the causal relationship between endocrine-disrupting chemicals and breast cancer mediated by circulating metabolites. *Environmental pollution (Barking, Essex : 1987)* **375**, 126349, doi:10.1016/j.envpol.2025.126349 (2025).
20 Wu, D., Bing, S., Qiu, H., Wang, S. & Zhang, Y. Association Between Mixed Exposure to Endocrine-Disrupting Chemicals and Cardiovascular Health: Results from the 2003-2016 NHANES. *Cardiovascular toxicology* **26**, 7, doi:10.1007/s12012-025-10084-6 (2025).
21 Fajkić, A. *et al.* Ultra-Processed Diets and Endocrine Disruption, Explanation of

Missing Link in Rising Cancer Incidence Among Young Adults. *Cancers* **17**, doi:10.3390/cancers17132196 (2025).

22 Kim, K. The Role of Endocrine Disruption Chemical-Regulated Aryl Hydrocarbon Receptor Activity in the Pathogenesis of Pancreatic Diseases and Cancer. *International journal of molecular sciences* **25**, doi:10.3390/ijms25073818 (2024).

23 Jiang, T. T. *et al.* Association of co-exposure to EDCs in early pregnancy with threatened abortion: The mediation effect of progesterone. *Environment international* **207**, 110001, doi:10.1016/j.envint.2025.110001 (2026).

24 Yigit Yalcin, B., Bilik Oyman, G. & Kubat Uzum, A. Endocrine disruptors and pregnancy. *Turk J Med Sci* **55**, 1625-1634, doi:10.55730/1300-0144.6123 (2025).

25 Choe, Y. *et al.* Prenatal and childhood exposure to endocrine-disrupting chemicals and early thelarche in 8-year-old girls: A prospective study using Bayesian kernel regression. *Environmental research* **263**, 120056, doi:10.1016/j.envres.2024.120056 (2024).

26 Jorge, B. C. *et al.* Potential role of endocrine-disrupting chemicals in preeclampsia: A hypothesis based on integration of epidemiological and experimental evidence. *Reproductive toxicology (Elmsford, N.Y.)* **137**, 109029, doi:10.1016/j.reprotox.2025.109029 (2025).

27 Han, X. & Jin, X. The impact, mechanisms and prevention strategies of environmental endocrine disruptors on male reproductive health. *Frontiers in endocrinology* **16**, 1573526, doi:10.3389/fendo.2025.1573526 (2025).

28 Almeida-Toledano, L. *et al.* Effect of prenatal phthalate exposure on fetal development and maternal/neonatal health consequences: A systematic review. *The Science of the total environment* **950**, 175080, doi:10.1016/j.scitotenv.2024.175080 (2024).

29 Amar, S. *et al.* Bisphenol S Impaired Human Granulosa Cell Steroidogenesis in Vitro. *International journal of molecular sciences* **21**, doi:10.3390/ijms21051821 (2020).

30 Huang, M., Liu, S., Fu, L., Jiang, X. & Yang, M. Bisphenol A and its analogues bisphenol S, bisphenol F and bisphenol AF induce oxidative stress and biomacromolecular damage in human granulosa KGN cells. *Chemosphere* **253**, 126707, doi:10.1016/j.chemosphere.2020.126707 (2020).

31 Liang, Y. *et al.* Impact of endocrine disrupting chemicals (EDCs) on epigenetic regulation in the uterus: a narrative review. *Reproductive biology and endocrinology : RB&E* **23**, 80, doi:10.1186/s12958-025-01413-z (2025).

32 Lonardo, M. S. *et al.* Hypothalamic-Ovarian axis and Adiposity Relationship in Polycystic Ovary Syndrome: Physiopathology and Therapeutic Options for the Management of Metabolic and Inflammatory Aspects. *Current obesity reports* **13**, 51-70, doi:10.1007/s13679-023-00531-2 (2024).

33 Zhao, H., Zhang, J., Cheng, X., Nie, X. & He, B. Insulin resistance in polycystic ovary syndrome across various tissues: an updated review of pathogenesis, evaluation, and treatment. *Journal of ovarian research* **16**, 9, doi:10.1186/s13048-022-01091-0 (2023).

34 Ke, Z. H. *et al.* Human-relevant doses of bisphenol a induce insulin resistance by

downregulating Irs1 expression in skeletal muscle through estrogen receptor-α. *Ecotoxicology and environmental safety* **303**, 118791, doi:10.1016/j.ecoenv.2025.118791 (2025).
35 Kulsange, S. E. *et al.* SWATH-MS reveals that bisphenol A and its analogs regulate pathways leading to disruption in insulin signaling and fatty acid metabolism. *Food and chemical toxicology : an international journal published for the British Industrial Biological Research Association* **188**, 114667, doi:10.1016/j.fct.2024.114667 (2024).
36 Byerly, J. H., Port, E. R. & Irie, H. Y. PRKCQ inhibition enhances chemosensitivity of triple-negative breast cancer by regulating Bim. *Breast cancer research : BCR* **22**, 72, doi:10.1186/s13058-020-01302-w (2020).
37 Bansal, B. *et al.* A Systematic Review of Inflammatory Markers in Polycystic Ovary Syndrome (PCOS) and Meta-Analysis of Interleukin-6 (IL-6) in Case-Control Studies. *Cureus* **17**, e82350, doi:10.7759/cureus.82350 (2025).
38 Deng, H., Chen, Y., Xing, J., Zhang, N. & Xu, L. Systematic low-grade chronic inflammation and intrinsic mechanisms in polycystic ovary syndrome. *Frontiers in immunology* **15**, 1470283, doi:10.3389/fimmu.2024.1470283 (2024).
39 Zhao, B., Lu, Q., Chen, M. & Lash, G. E. Impact of endocrine disrupting chemicals on macrophages at the maternal-fetal interface. *Seminars in immunopathology* **47**, 29, doi:10.1007/s00281-025-01055-8 (2025).
40 Dong, Y., Gao, L., Sun, Q., Jia, L. & Liu, D. Increased levels of IL-17 and autoantibodies following Bisphenol A exposure were associated with activation of PI3K/AKT/mTOR pathway and abnormal autophagy in MRL/lpr mice. *Ecotoxicology and environmental safety* **255**, 114788, doi:10.1016/j.ecoenv.2023.114788 (2023).
41 Kanehisa, M., Sato, Y., Kawashima, M., Furumichi, M. & Tanabe, M. KEGG as a reference resource for gene and protein annotation. *Nucleic Acids Research* **44(D1):D457-D462**, doi:10.1093/nar/gkv1070 (2016).
42 Kanehisa, M. A.-O. X., Furumichi, M., Sato, Y., Matsuura, Y. & Ishiguro-Watanabe, M. KEGG: biological systems database as a model of the real world. *Nucleic Acids Research* **53(1):D672-D677**, doi:doi:10.1093/nar/gkae1020 (2025).

Figure Legends

Figure 1. Identification of endocrine-disrupting chemical (EDC)-related differentially expressed genes (DEGs) in polycystic ovary syndrome (PCOS) and construction of the compound-target interaction network. (A) Volcano plot showing the distribution of genes based on $\log_2$ fold change ($\log_2$ FC) and $-\log_{10}$(P value) in the training dataset GSE10946. Differential expression analysis was performed using the limma package, with genes meeting the thresholds of p value < 0.05 and $|\log_2 FC| > 0.5$ considered significant. Upregulated genes are shown in pink, downregulated genes are shown in

blue, and non-significant genes are shown in gray. (B) Heatmap illustrating the expression patterns of the top 10 upregulated and top 10 downregulated DEGs between the PCOS and control groups. Each column represents an individual sample, and each row represents a gene. (C) Venn diagram of the overlap between DEGs and predicted compound-related target genes. Candidate EDC-associated targets were collected by integrating target information from the Comparative Toxicogenomics Database (CTD), ChEMBL, SwissTargetPrediction, and PharmMapper databases. (D) Compound-target interaction network was constructed based on the overlapping genes. Nodes represent candidate compounds and target genes, and edges indicate predicted interactions between compounds and their corresponding targets. Purple nodes represent target genes, and orange nodes represent compound (drug) names.

Figure 2. Protein-protein interaction (PPI) network construction and candidate gene screening. (A) PPI network of EDC-related DEGs was constructed using the STRING database and visualized in Cytoscape. Nodes represent proteins encoded by genes, and edges indicate protein-protein interactions. (B-D) Subnetworks of the top-ranked genes were identified using three different topological algorithms implemented in the cytoHubba plugin: Betweenness (B), Closeness (C), and Degree (D). Node color intensity reflects the relative importance of each gene based on the corresponding algorithm, with darker colors indicating higher centrality scores. (E) Venn diagram showing the overlap of the top 20 genes identified by the Betweenness, Closeness, and Degree algorithms.

Figure 3. Performance comparison of machine learning models for hub gene selection. Heatmap of area under the receiver operating characteristic curve (AUC) values for multiple machine learning models and their combinations in the training dataset (GSE10996) and the merged validation cohort (GSE106724, GSE34526, and GSE98595).

Figure 4. Expression validation and diagnostic performance of hub genes. (A) Box plots of the expression levels of hub genes in the training dataset GSE10946 between PCOS and control groups. (B) Box plots showing the expression levels of the identified hub genes in the merged validation cohort (GSE106724, GSE34526, and GSE98595)

between PCOS and control groups. Statistical significance between groups is indicated above each gene ($p < 0.05$; ns, not significant). (C) Receiver operating characteristic (ROC) curve analysis was performed to assess the diagnostic accuracy of insulin receptor substrate 1 (IRS1) and protein kinase C theta (PRKCQ) in the training cohort (GSE10946). (D) The diagnostic efficacy of IRS1 and PRKCQ in the merged validation cohort was determined using ROC curves.

Figure 5. Construction and evaluation of the nomogram model based on hub genes. (A) Nomogram integrating IRS1 and PRKCQ for predicting disease risk. Each variable corresponds to a specific score, and the total points are calculated to estimate the individual probability of disease occurrence. (B) ROC curve analysis was performed to assess the discriminative ability of the nomogram model. (C) The concordance between predicted risks generated by the nomogram and actual observed outcomes was evaluated using a calibration curve. The dashed diagonal line represents ideal prediction, while the solid lines indicate the apparent and bias-corrected performance. (D) Decision curve analysis (DCA) was used to assess the clinical utility of the nomogram by quantifying the net benefit across a range of threshold probabilities, compared with single-gene models and default strategies (treat-all or treat-none).

Figure 6. Gene set enrichment analysis (GSEA) and immune infiltration analysis of hub genes. (A) GSEA plots showing the top five significantly enriched KEGG pathways associated with IRS1 in the training dataset. (B) GSEA plots showing the top five significantly enriched KEGG pathways associated with PRKCQ. Important enrichment pathways were identified under the criteria of |normalized enrichment score (NES)| > 1, false discovery rate (FDR) < 0.25, and $p < 0.05$. (C) Heatmap illustrating the correlation between the expression levels of IRS1 and PRKCQ and the relative abundance of 28 immune cell types estimated by the ssGSEA algorithm. Color intensity represents the Spearman correlation coefficient, with red indicating positive correlation and green indicating negative correlation. Statistical significance is annotated ($^{*}p < 0.05$, $^{**}p < 0.01$, $^{**}p < 0.001$).

Figure 7. Molecular docking analysis of hub genes with endocrine-disrupting chemicals. (A) Predicted binding conformation of the compound within the target protein pocket,

shown as an overall docking pose (left) and a magnified view of the binding interface (right). Key interacting amino acid residues involved in ligand binding are labeled, and noncovalent interactions are indicated. (B) Detailed visualization of the ligand-protein interaction pattern, highlighting the spatial arrangement of the ligand and surrounding residues within the binding cavity, as well as the putative hydrogen bonds and hydrophobic interactions stabilizing the complex.

Figure 8. Molecular dynamics simulation of the IRS1-Bisphenol A complex. (A) Root mean square deviation (RMSD) of the IRS1-Bisphenol A complex during the molecular dynamics simulation, reflecting the overall structural stability over time. (B) Root mean square fluctuation (RMSF) of IRS1 residues, indicating residue-level flexibility upon Bisphenol A binding. (C) Total energy of the IRS1-Bisphenol A system as a function of simulation time. (D) Time-dependent variation in the number of hydrogen bonds formed between IRS1 and Bisphenol A during the simulation.

Figure S1. Functional enrichment analysis of EDC-related DEGs. (A) Gene Ontology (GO) enrichment analysis for biological processes (BP). (B) GO enrichment analysis for cellular components (CC). (C) GO enrichment analysis for molecular functions (MF). (D) Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway enrichment analysis. All enrichment results were filtered under the threshold of p-value < 0.05, and the top 10 significantly enriched terms or pathways were selected for visualization. The horizontal axis represents $-\log_{10}$(p-value), and each dot or bar corresponds to an enriched functional term or pathway[41,42].