**Discover** Oncology

Analysis

# Computational analysis of DEHP's oncogenic role in colorectal cancer

**Zhou Zhu[1] · Jian Qin[2] · Chungang He[1] · Shuangyou Wang[1] · Yaolin Lu[1] · Shuai Wang[1] · Xiaogang Zhong[1]**

Received: 3 March 2025 / Accepted: 7 May 2025
Published online: 21 May 2025


**Abstract**

**Background** Colorectal cancer (CRC) remains a leading cause of cancer-related mortality globally, with many patients diagnosed at advanced stages. Current treatments, including surgery, chemotherapy, and targeted therapies, face limitations due to tumor metastasis and chemoresistance. Di-(2-ethylhexyl) phthalate (DEHP), a widely-used plasticizer, has been linked to various cancers, including CRC, through mechanisms such as metabolic reprogramming and inflammation. However, the direct relationship between DEHP and CRC requires further elucidation.

**Methods** We integrated transcriptomic data from TCGA-COADREAD (41 normal and 476 cancer tissues) and GEO datasets (GSE32323 and GSE21510) to identify differentially expressed genes (DEGs) using the `limma` package. We predicted DEHP molecular targets via SwissTargetPrediction and ChEMBL databases and constructed a protein–protein interaction (PPI) network using STRING. Machine learning methods, including LASSO regression, SVM, and Random Forest, identified key genes. SHAP analysis and ssGSEA were employed to evaluate gene importance and immune cell infiltration, respectively. Molecular docking experiments assessed the binding affinity of DEHP with key proteins.

**Results** Differential expression analysis identified 86 common genes involved in pathways such as PI3K-Akt and p53 signaling. The PPI network highlighted 14 candidate genes, with machine learning methods narrowing down to three key genes: CDK1, CDK4, and BCL2. SHAP analysis showed CDK1 and CDK4 as top contributors, while ssGSEA revealed significant correlations between these genes and immune cell infiltration. Molecular docking experiments demonstrated strong binding affinities of DEHP with BCL2 (– 8.7 kcal/mol), CDK1 (– 7.8 kcal/mol), and CDK4 (– 6.8 kcal/mol).

**Conclusion** This study provides comprehensive insights into the oncogenic mechanisms of DEHP in CRC, identifying key genes and pathways that may serve as potential therapeutic targets. Our findings highlight the need for further investigation into DEHP's role in CRC and its potential as a target for prevention and treatment strategies.

**Keywords** Colorectal cancer · DEHP · Machine learning · SHAP analysis · Molecular docking · Immune cell infiltration

## 1 Introduction

Colorectal cancer (CRC) is one of the most common malignant tumors worldwide, with both incidence and mortality rates remaining high [1, 2]. According to the latest statistical data, the incidence of colorectal cancer ranks among the top in China's malignant tumors, and its mortality rate is also relatively high [3, 4]. Despite continuous progress in early screening and diagnostic techniques, the majority of patients are diagnosed at an advanced stage, which increases the

Zhou Zhu and Jian Qin have contributed equally to this paper.

✉ Xiaogang Zhong, xiaogangzhong@163.com | [1]Department of Colorectal and Anal Surgery, The People's Hospital of Guangxi Zhuang Autonomous Region, Nanning 530021, Guangxi, China. [2]Department of Radiation Oncology, The People's Hospital of Guangxi Zhuang Autonomous Region, Nanning 530021, Guangxi, China.

difficulty of treatment and leads to poor prognosis [5, 6]. Currently, the main treatment methods for colorectal cancer include surgery, chemotherapy, targeted therapy, and immunotherapy [7–9]. However, these methods still have limitations in terms of therapeutic efficacy [10, 11]. For example, although surgical resection is the primary treatment, patients with advanced cancer often cannot undergo radical surgery due to tumor metastasis [12, 13]. Moreover, chemotherapy resistance and tumor recurrence severely affect patients' long-term survival [14–16]. Therefore, in-depth research into the pathogenesis of colorectal cancer and the search for new therapeutic targets and preventive strategies remain urgent issues that need to be addressed.

Di-(2-ethylhexyl) phthalate (DEHP) is a widely used plasticizer, commonly found in food packaging, medical devices, and daily consumer products [17–19]. In recent years, the potential health hazards of DEHP have attracted widespread attention, especially its association with various cancers [20–22]. Studies have shown that exposure to DEHP is significantly related to the risk of developing several types of cancer, including liver cancer [23, 24], pancreatic cancer [25, 26], bladder cancer [27], and cancers of the reproductive system [28–30]. For example, DEHP can promote cell proliferation and enhance cell survival by activating inflammatory signaling pathways in liver cancer cells [31]. In pancreatic cancer, DEHP exposure can increase the characteristics of cancer stem cells, leading to enhanced radiotherapy resistance [32]. Additionally, DEHP has been found to promote chemoresistance in colorectal cancer cells through metabolic reprogramming, reducing the sensitivity to chemotherapeutic drugs [33]. In colorectal cancer, DEHP exposure can lead to changes in cell surface glycosylation, enhancing the characteristics of cancer stem cells and thereby promoting tumor progression [34]. Although studies have revealed the potential oncogenic mechanisms of DEHP in various cancers, its direct relationship with colorectal cancer still needs further investigation.

Given the potential carcinogenicity of DEHP in multiple cancers and its oncogenic effects in colorectal cancer, in-depth research into the direct relationship between DEHP and colorectal cancer is of great scientific and clinical significance. On the one hand, this will help elucidate the specific mechanisms by which DEHP contributes to the development and progression of colorectal cancer, providing new evidence for the relationship between environmental factors and cancer. On the other hand, clarifying the mechanisms of DEHP's action may offer new targets and strategies for the prevention and treatment of colorectal cancer. For example, reducing DEHP exposure or developing interventions targeting its mechanisms of action could potentially lower the incidence of colorectal cancer and improve patient prognosis. Therefore, this study not only helps fill the gaps in current research but may also provide new ideas for the tertiary prevention of colorectal cancer.

## 2 Materials and methods

Figure 1 Flow chart of this study.

### 2.1 Acquisition of transcriptomic data and DEHP molecular targets

In this study, we collected colorectal cancer transcriptomic data (TCGA-COADREAD) from The Cancer Genome Atlas (TCGA) database, which included 41 normal colorectal tissue transcriptomes and 476 colorectal cancer transcriptomes. Additionally, we collected data from the GEO datasets GSE32323 and GSE21510. GSE32323 is a gene expression dataset of the transcriptomic study type, containing 17 pairs of colorectal cancer patient samples (both cancerous and non-cancerous tissues). GSE21510 is also a gene expression dataset of the transcriptomic study type, focusing on gene expression differences in colorectal cancer. This dataset includes 148 samples, comprising 25 non-cancerous tissue samples and 123 colorectal cancer patient samples. By merging these two datasets, we obtained a new combined dataset, including 42 normal samples and 140 colorectal cancer patients. These data were used to analyze colorectal cancer-related gene expression patterns and differences. We obtained the SMILES number of DEHP from the PubChem (https://pubchem.ncbi.nlm.nih.gov/) database and predicted its molecular targets using the SwissTargetPrediction (http://www.swisstargetprediction.ch/) and ChEMBL (https://www.ebi.ac.uk/chembl/) databases. These targets are proteins or genes that DEHP may interact with. Finally, we removed duplicates and merged the DEHP molecular targets to obtain a more accurate set of targets. We also obtained colorectal cancer-related disease targets from the GeneCards (https://www.genecards.org/) and OMIM (https://www.omim.org/) databases, which are known genes or proteins associated with the development and progression of colorectal cancer.

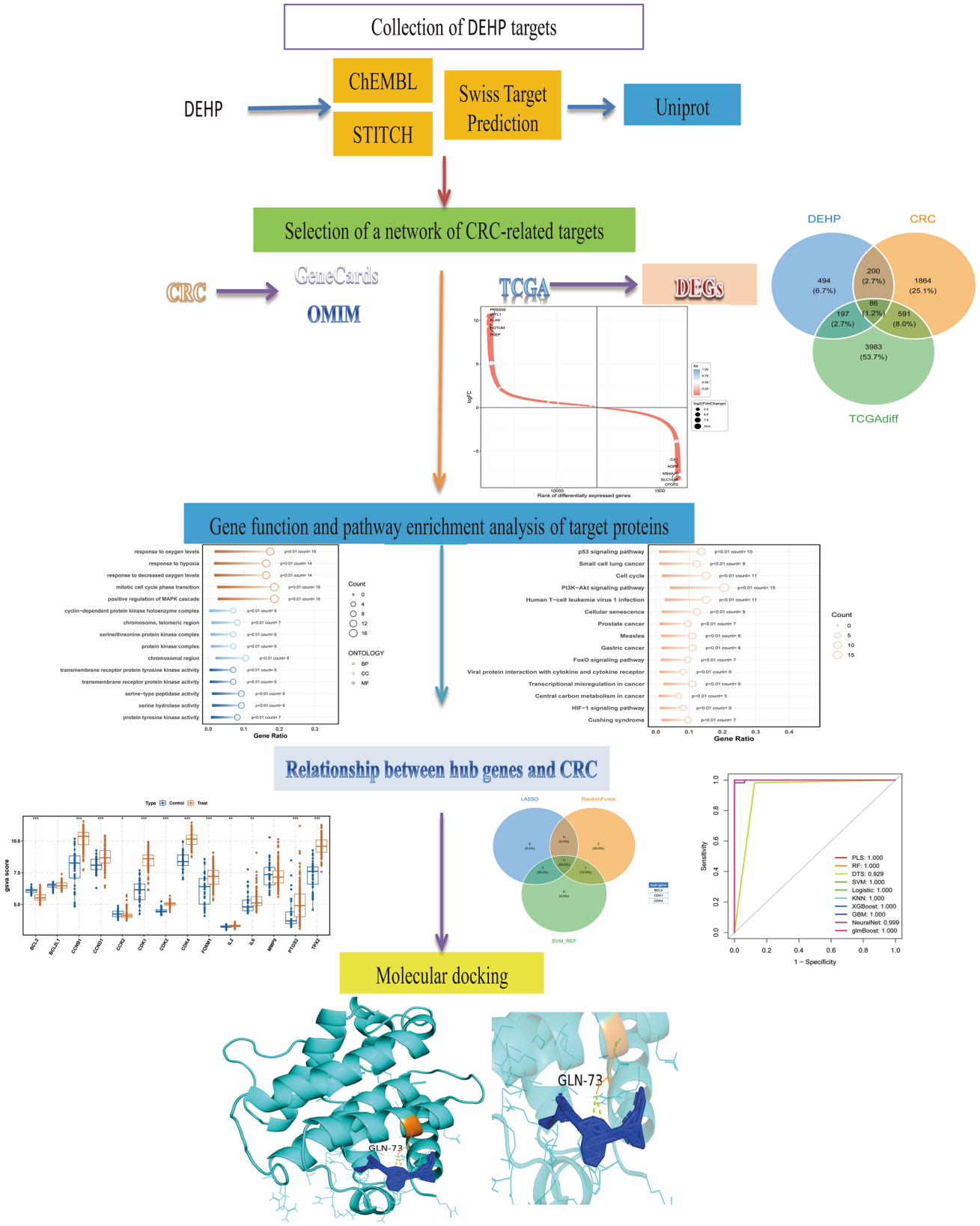



**Fig. 1** Flowchart of this study

### 2.2 Differential gene analysis in TCGA and target intersection analysis

In this study, we used the `limma` package in R to perform differential expression analysis on colorectal cancer and normal tissue samples from the TCGA database. During the analysis, we first obtained the transcriptomic data of colorectal cancer from the TCGA database, including 41 normal tissue samples and 476 cancer tissue samples. To ensure the accuracy of the analysis, we preprocessed the data, including background correction and normalization. Subsequently, we used the `lmFit` function in the `limma` package to fit a linear model and performed Bayesian statistical analysis using the `eBayes` function. When screening for differentially expressed genes, we set the significance threshold to $p < 0.05$ and combined it with $|\log2(\text{Fold Change})| > 1$ as the screening criteria. This criterion aims to identify genes with significant expression differences between colorectal cancer and normal tissues, providing a basis for further functional analysis. Finally, we used the `ggplot2` package to create a scatter plot showing the distribution of differentially expressed genes. In the scatter plot, significantly upregulated and downregulated genes are marked in different colors, intuitively displaying the trends in gene expression changes. Additionally, we used the `VennDiagram` R package to analyze the common genes among disease targets, differentially expressed genes from the TCGA database, and DEHP targets. By creating a Venn diagram, we clearly displayed the intersection results of these three gene sets.

### 2.3 GO and KEGG analysis

To further explore the potential roles of the intersecting genes in biological functions and signaling pathways, we conducted Gene Ontology (GO) analysis and Kyoto Encyclopedia of Genes and Genomes (KEGG) enrichment analysis in this study. The GO analysis mainly covers three aspects: biological processes, cellular components, and molecular functions, which are used to annotate the functions of genes in cellular biology processes. The KEGG analysis focuses on the signaling pathways in which genes are involved, revealing their potential mechanisms in disease development. During the analysis, we used the clusterProfiler R package for enrichment analysis, setting the significance threshold to $p < 0.05$ and $q < 0.05$ to identify significantly enriched GO terms and KEGG pathways. To more intuitively display the enrichment results, we further used the ggplot2 R package to create bubble plots. The bubble plots display the descriptions of enriched items on the x-axis, the number of enriched genes on the y-axis, and the size and color of the bubbles represent the number of genes and the significance of the p-values, respectively. This visualization method clearly presents the enrichment of intersecting genes in different biological functions and signaling pathways.

### 2.4 Construction of protein–protein interaction (PPI) network using STRING online database

To further explore the interactions among the 86 common genes at the protein level, we used the STRING online database (https://cn.string-db.org/) to construct a PPI network in this study. The STRING database is a widely used platform for functional protein association networks, integrating protein interaction information from various sources, including experimental validation, literature mining, and gene neighborhood analysis, and providing high-confidence protein interaction networks. When constructing the PPI network, we input the 86 common genes and set the confidence threshold to 0.8 to ensure that the interactions included in the network are highly credible. A confidence threshold of 0.8 means that only protein interactions supported by multiple pieces of evidence and with high confidence will be included in the final network. Through this setting, we were able to identify biologically significant protein interactions, providing a basis for further functional analysis.

### 2.5 Network visualization with cytoscape

To further analyze and visualize the PPI network, we imported the PPI network data exported from the STRING database into the Cytoscape software. In Cytoscape, we first imported the network data file exported from STRING and used the degree sorted circle layout algorithm to optimize the network arrangement. Next, we used the cytohubba plugin in Cytoscape to screen for sub-networks. In this study, we used three methods: Degree, Maximal Clique Centrality (MCC), and Stress. The Degree method assesses the importance of nodes based on their number of connections. The MCC method screens for key nodes by identifying the centrality of the largest cliques, while the Stress method evaluates the importance of nodes based on their stress values in the network. To ensure the reliability of the screening results, we set the screening criteria to the top 20 genes for each method, or the actual number if less than 20. This process not only

helped us identify key genes with significant roles in the PPI network but also provided an important set of candidate genes for further functional analysis and experimental validation.

### 2.6 Machine learning for key gene screening

To identify key genes associated with colorectal cancer, we used multiple machine learning methods to analyze the combined GEO datasets in this study. Firstly, we used the wilcox.test to verify the expression of the aforementioned gene in the GEO dataset. Subsequently, we performed LASSO regression analysis using the `glmnet` package, determining the optimal regularization parameter λ through cross-validation to screen for key genes. We then used the 'e1071' package for Support Vector Machine (SVM) analysis, selecting the model with 7 genes, which achieved an accuracy rate as high as 0.983. Additionally, we used the `randomForest` package for Random Forest analysis, screening for genes with importance greater than 2. Finally, we obtained key genes by taking the intersection of the results from the three methods using the `VENN` tool. To rigorously evaluate the stability and generalizability of our machine learning models, we employed a tenfold cross-validation strategy. Specifically, the combined GEO dataset was randomly partitioned into ten subsets, with seven subsets used for model training and the remaining subset for validation, repeating this process iteratively. Model performance was systematically assessed using metrics including accuracy, sensitivity, specificity, and the area under the receiver operating characteristic curve (AUC-ROC). The final models selected demonstrated consistently high performance, achieving accuracy up to 0.929, which underscores the reliability of our identified key genes.

### 2.7 SHAP model construction

To further understand the mechanisms by which key genes act in colorectal cancer, we constructed a SHapley Additive exPlanations (SHAP) model using the kernelshap, ggplot2, ranger, and shapviz R packages in this study. The SHAP model calculates the SHAP values for each gene's contribution to model predictions, helping to interpret the model's output results. First, we randomly divided the GEO dataset into training and validation sets in a 7:3 ratio. Based on the training set, we built a Random Forest model using the ranger package and calculated the SHAP values using the kernelshap package. Kernelshap efficiently implements the Kernel SHAP algorithm, assigning SHAP values to each feature in the model to quantify its importance to model predictions. During the calculation, we set the background dataset size to 200 and used exact calculation mode to ensure the accuracy of the results. We obtained ROC curves, beeswarm plots, feature dependence plots, and feature attribution waterfall plots.

### 2.8 Immune cell infiltration analysis

To comprehensively assess the immune cell infiltration in the tumor microenvironment of colorectal cancer patients, we used the Single Sample Gene Set Enrichment Analysis (ssGSEA) method, implemented through the GSVA R package in this study. SsGSEA is a gene set enrichment analysis method for individual samples, capable of evaluating the infiltration degree of different immune cells in tumor tissues. Specifically, we used the GSVA package to analyze the colorectal cancer transcriptomic data from the GEO database, calculating the enrichment scores of each immune cell in each sample. Subsequently, to explore the correlation between key genes and immune cell infiltration, we preprocessed and normalized the gene expression data using the limma package and combined it with the ggplot2 package to create correlation charts.

### 2.9 Molecular docking

To investigate the binding capacity of DEHP with key genes, we obtained the corresponding protein molecular structures of these genes from the Protein Data Bank (PDB) database in this study. Protein structures for the key targets were obtained from the PDB. We searched for structures determined by X-ray crystallography and selected those with high resolution (typically better than 2.5 Å) to ensure accurate atomic coordinates. When multiple structures were available, the one with the highest resolution and most complete conformation was chosen. The selected structures were then processed to remove non-essential molecules (e.g., water, co-crystallized ligands) and to model any missing residues, as needed. Finally, the prepared structures were visualized using PyMOL software (version 2.5.0) prior to conducting molecular docking experiments with AutoDock software. Subsequently, we performed molecular docking experiments using AutoDock software (version 4.2.6). During the docking process, we set the binding energy threshold to – 6.0 kcal/mol, a commonly used standard in molecular docking experiments, to screen for ligand-receptor complexes with strong

**Fig. 2** **A** Scatter plot of differentially expressed genes. The scatter plot displays the distribution of differentially expressed genes between ▶ colorectal cancer patients and controls in the TCGA database. **B** Venn diagram of gene intersections. The Venn diagram shows the intersection results among disease targets, differentially expressed genes from the TCGA database, and DEHP targets. The overlap indicates the final selection of 86 common genes. **C** Lollipop plot of GO enrichment analysis. The lollipop plot displays the results of GO enrichment analysis, showing significantly enriched biological processes such as positive regulation of the MAPK cascade, protein kinase complex, and transmembrane receptor protein tyrosine kinase activity. **D** Lollipop plot of KEGG enrichment analysis. The lollipop plot displays the results of KEGG enrichment analysis, showing significantly enriched signaling pathways such as the PI3K – Akt signaling pathway and p53 signaling pathway

binding capabilities. Finally, we obtained the binding energies of the key genes with DEHP, identifying key proteins that can bind DEHP effectively.

### 2.10 Statistical analysis

The statistical analysis in this study was primarily conducted using the R language and various bioinformatics tools. First, we used the `limma` package to perform differential analysis on gene expression data to screen for significantly different genes. Next, we conducted GO and KEGG enrichment analysis using the `clusterProfiler` package to reveal the functional and pathway associations of genes. Additionally, we combined the `randomForest`, `xgboost`, and `caret` packages for machine learning analysis to screen for key genes and assess their contributions to the model using the SHAP model. Finally, we analyzed immune cell infiltration using the `ssGSEA` package and verified the binding capacity of key genes with DEHP through molecular docking experiments using AutoDock software.

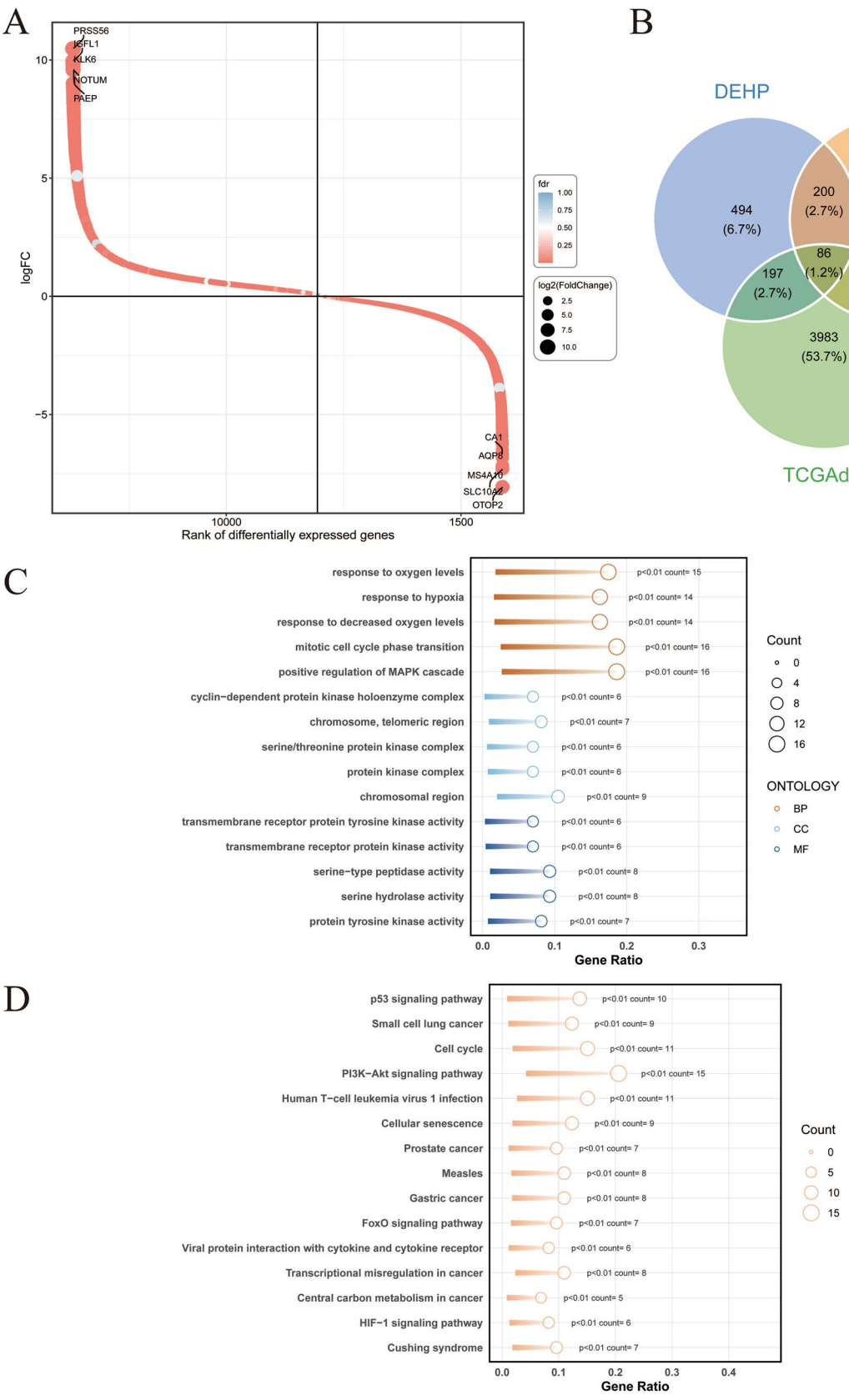
## 3 Results

### 3.1 Identification of 86 common genes and key pathways

We first used the `limma` package to perform differential expression analysis on the gene expression data of colorectal cancer patients and controls from the TCGA database in this study. The distribution of differentially expressed genes was intuitively displayed through a scatter plot (Fig. 2A). Subsequently, we conducted intersection analysis among disease targets, differentially expressed genes from the TCGA database, and DEHP targets, eventually screening out 86 common genes (Fig. 2B). We then performed GO and KEGG enrichment analysis on these genes. Figure 2C shows that these genes are primarily involved in biological processes that facilitate the positive regulation of the MAPK cascade. Specifically, several gene products act as key intermediaries by enhancing receptor protein tyrosine kinase activity, which in turn triggers adaptor proteins and downstream phosphorylation events. Moreover, some of these proteins serve as scaffolds within the protein kinase complex, ensuring proper assembly and stabilization of signaling components. Together, these roles contribute to an amplified and sustained activation of the MAPK pathway, ultimately promoting cellular processes such as proliferation and differentiation.. The KEGG enrichment analysis indicated that these common targets were mainly involved in signaling pathways such as the PI3K – Akt signaling pathway and p53 signaling pathway (Fig. 2D). These analysis results suggest that DEHP may participate in the development of colorectal cancer by affecting these key biological processes and signaling pathways.

### 3.2 PPI network construction and selection of 14 candidate genes

To screen for genes more closely related to colorectal cancer, we constructed a PPI network using the STRING database in this study, setting the confidence threshold to 0.8 to identify interactions among genes. The constructed network contained 86 nodes and 129 edges (Fig. 3A). Subsequently, we used the 'cytohubba' plugin in Cytoscape software to perform topological analysis on the network using three methods: Degree, MCC, and stress. We screened out the top 20 key genes for each method (Figs. 3B–D). Finally, by taking the intersection of the three methods, we obtained 14 candidate genes, namely BCL2, BCL2L1, CCNB1, CCND1, CCR2, CDK1, CDK2, CDK4, FOXM1, IL2, IL6, MMP9, PTGS2, and TPX2 (Fig. 3E).


A
PRSS56
KSFL1
KLK6
NOTUM
PAEP
CA1
AQP8
MS4A10
SLC10A2
OTOP2
logFC
Rank of differentially expressed genes
fdr
log2(FoldChange)
B
DEHP
CRC
TCGAdiff
494 (6.7%)
200 (2.7%)
1864 (25.1%)
86 (1.2%)
197 (2.7%)
591 (8.0%)
3983 (53.7%)
C
response to oxygen levels
response to hypoxia
response to decreased oxygen levels
mitotic cell cycle phase transition
positive regulation of MAPK cascade
cyclin−dependent protein kinase holoenzyme complex
chromosome, telomeric region
serine/threonine protein kinase complex
protein kinase complex
chromosomal region
transmembrane receptor protein tyrosine kinase activity
transmembrane receptor protein kinase activity
serine−type peptidase activity
serine hydrolase activity
protein tyrosine kinase activity
Gene Ratio
Count
ONTOLOGY
BP
CC
MF
D
p53 signaling pathway
Small cell lung cancer
Cell cycle
PI3K−Akt signaling pathway
Human T−cell leukemia virus 1 infection
Cellular senescence
Prostate cancer
Measles
Gastric cancer
FoxO signaling pathway
Viral protein interaction with cytokine and cytokine receptor
Transcriptional misregulation in cancer
Central carbon metabolism in cancer
HIF−1 signaling pathway
Cushing syndrome
Gene Ratio
Count

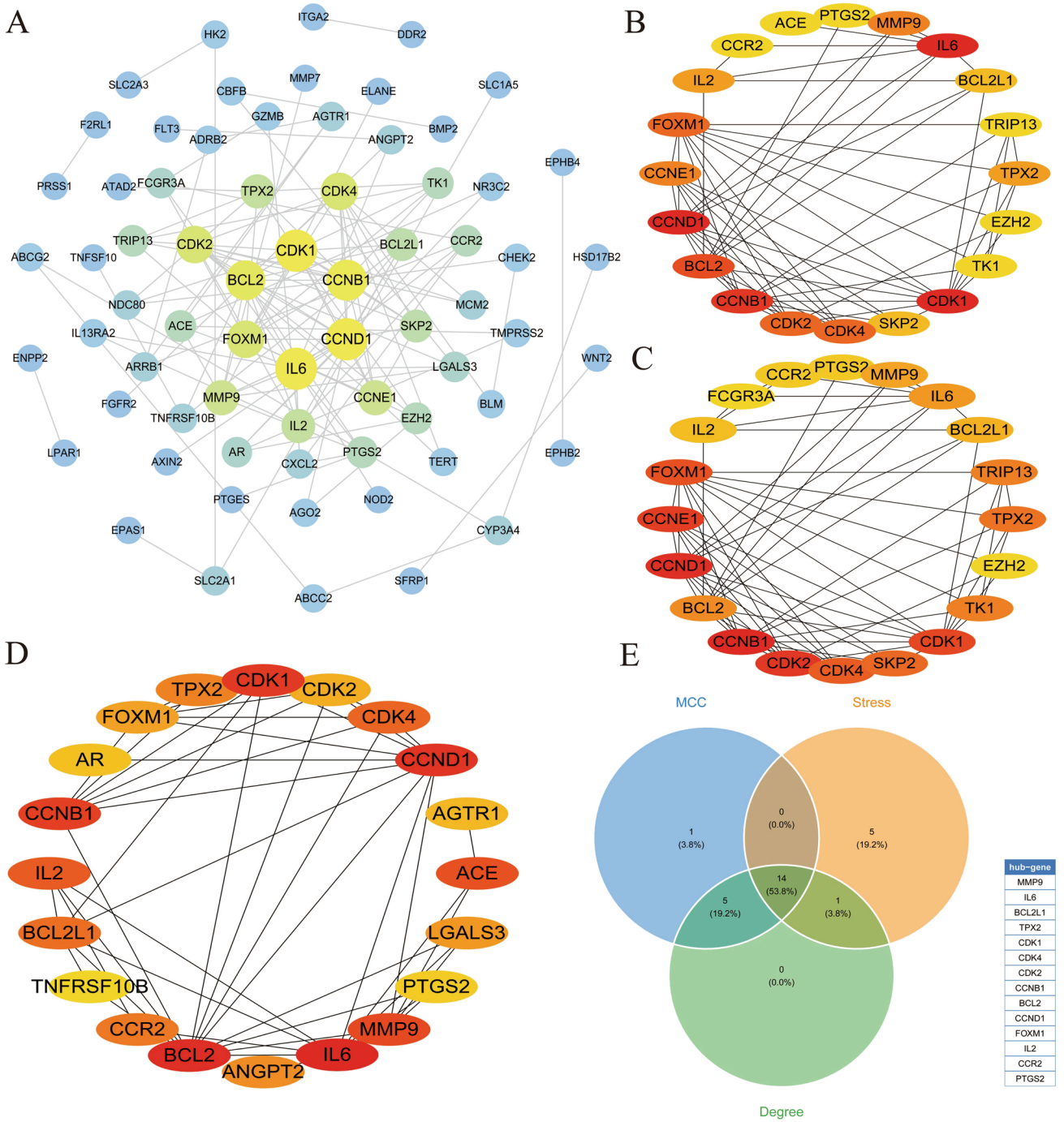



**Fig. 3** **A** Visualization of the PPI network. The figure shows the PPI network constructed based on the STRING database, containing 86 nodes (genes) and 129 edges (interactions between genes). The size of the nodes indicates the degree of connectivity, and the color indicates the importance of the genes. **B**–**D** Key gene screening results using different topological analysis methods. **B**, 2**C**, and 2**D** display the top 20 key genes selected using the Degree, MCC, and stress methods, respectively. **E** Intersection of screening results from three methods. The Venn diagram shows the intersection of the screening results from the Degree, MCC, and stress methods, ultimately identifying 14 candidate genes

### 3.3 Machine learning identifies three key genes: CDK1, CDK4, and BCL2

To refine the initial list of 14 candidate genes derived from the PPI network, we implemented a multi-step filtering strategy that integrated both statistical significance and interaction strength. First, we compared the expression levels of the 14 candidate genes between colorectal cancer (CRC) tissues and controls using the Wilcoxon test. Two genes, MMP9 and BCL2L1, were excluded because they did not exhibit statistically significant differences (Fig. 4A). Subsequently, we evaluated the remaining genes using a Support Vector Machine (SVM) model. The SVM analysis showed that when incorporating between 2 and 7 genes, the model achieved an accuracy of up to 0.983, and the model with 7 genes was chosen for further analysis (Figs. 4B, C). Next, LASSO regression analysis was performed to reduce redundancy and prioritize genes that contributed most strongly to differentiating between CRC and normal samples (Fig. 4D). In parallel, a Random Forest analysis was conducted to assess gene importance. In this analysis, only genes with an importance score greater than 2 were retained; notably, CDK1, CCNB1, and CDK4 emerged as the most significant (Figs. 4E, F). Finally, by intersecting the results of the SVM, LASSO, and Random Forest analyses, we identified three key genes—CDK1, CDK4, and BCL2—that consistently demonstrated robust predictive power and functional relevance (Fig. 4G). In summary, other candidate genes were excluded because they either did not meet the statistical significance criteria in the differential expression analysis, contributed weakly to the predictive models, or exhibited insufficient evidence of strong interactions in the PPI network. This rigorous, multi-method approach ensured that only genes with both high statistical significance and strong interaction profiles were retained for further investigation.

### 3.4 SHAP analysis highlights CDK1 and CDK4 as top contributors

To gain insights into how each key gene influences the model's predictions, we constructed a SHAP (SHapley Additive exPlanations) model. SHAP values offer a unified measure of feature importance by quantifying each gene's contribution to the prediction outcome. In our analysis, the GEO dataset was randomly split into training (70%) and validation (30%) sets, and the Decision Tree (DTS) model—selected based on its robust performance (AUC of 0.929)—was used to compute SHAP values (Fig. 5A). Figure 5B presents a bar chart summarizing the average absolute SHAP values for CDK1, CDK4, and BCL2. Here, CDK1 and CDK4 show higher values, indicating that variations in their expression have a stronger impact on predicting colorectal cancer status compared to BCL2. The beeswarm plot (Fig. 5C) further illustrates the distribution of SHAP values across all samples for each gene. This plot reveals that as the expression levels of CDK1 and CDK4 change, their corresponding SHAP values vary significantly—highlighting their influential roles in the model's decision-making process. Additionally, the variable dependence plot (Fig. 5D) explicitly shows that increases in the expression levels of CDK1 and CDK4 correspond to higher SHAP values. This trend reinforces the notion that higher expression of these genes is associated with a greater contribution to the risk prediction for colorectal cancer. Finally, the waterfall plot (Fig. 5E) offers an individual-level explanation by breaking down how each gene's SHAP value cumulatively drives the final prediction for specific samples. Collectively, these SHAP plots provide a comprehensive interpretation of our model's inner workings, clearly emphasizing the dominant roles of CDK1 and CDK4 in colorectal cancer prediction. This detailed explanation should help readers unfamiliar with SHAP techniques to better understand the contribution of each gene to our computational analysis.

### 3.5 Immune cell infiltration and significant correlations with key genes

We used the ssGSEA method to analyze the immune cell infiltration in colorectal cancer patients and controls from the GEO database in this study. The results showed significant differences in most immune cells among different patients. For example, the infiltration levels of active CD4 T cells, active CD8 T cells, memory CD8 T cells, type 2 T helper cells, and natural cells were higher in colorectal cancer patients (Fig. 6A, B, $P < 0.0001$). Subsequently, we analyzed the correlation between the three key genes (CDK1, CDK4, and BCL2) and immune cells. The results showed that these genes were significantly correlated with multiple immune cells. Specifically, BCL2 was significantly correlated with Activated B cells, Effector memory CD8 T cells, Immature B cells, Natural killer cells, Type 17 T helper cells, Type 1 T helper cells, MDSCs, and Central memory CD8 T cells (Fig. 6C, $P < 0.001$); CDK1 was significantly correlated with

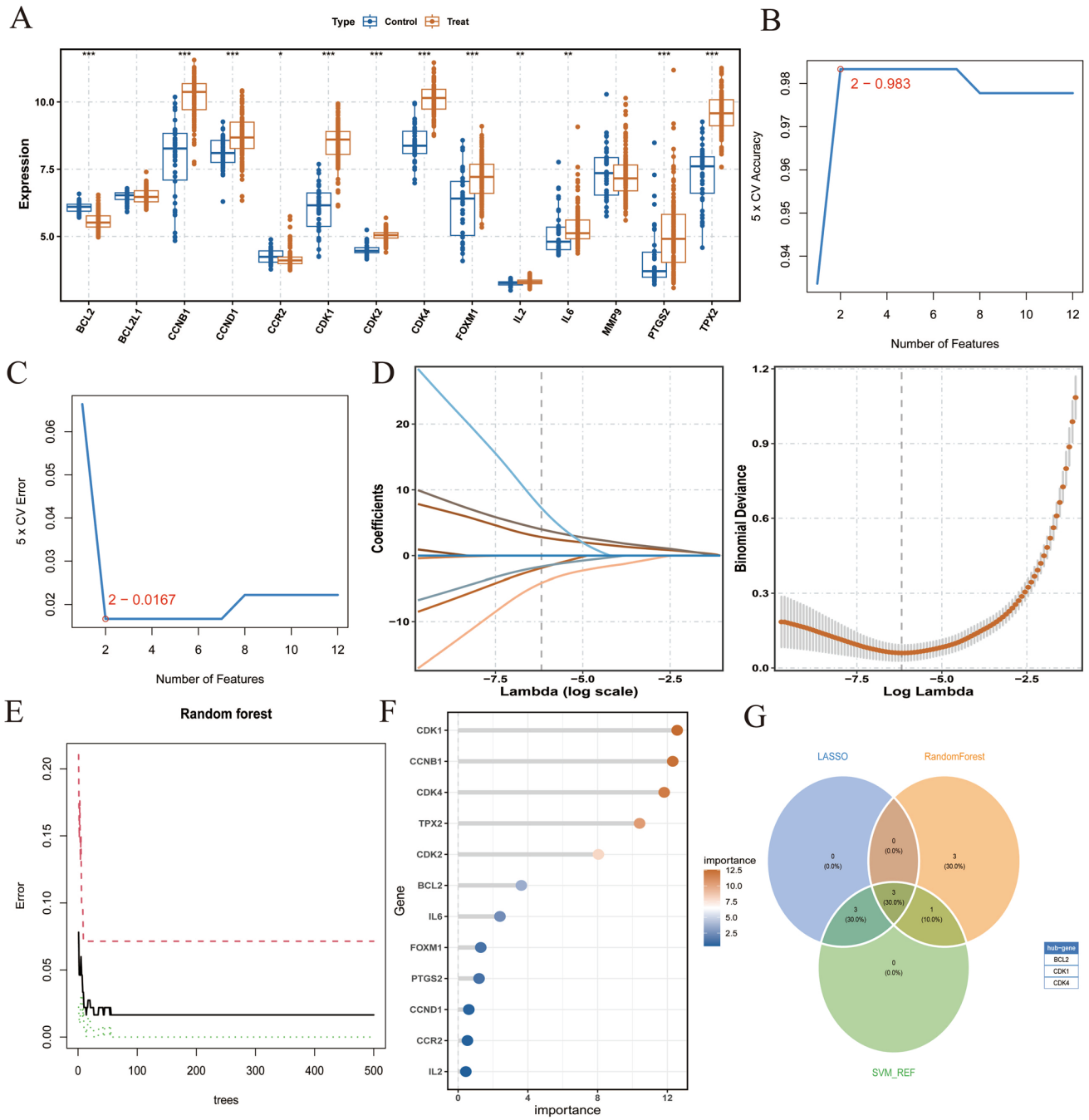



**Fig. 4** **A** Differential expression analysis of candidate genes. The bar chart displays the expression differences of the 14 candidate genes between the disease group and the control group. **B**, **C** Support Vector Machine analysis results. **B** and **C** display the results of Support Vector Machine (SVM) analysis. Figure 3C shows the model accuracy for different numbers of genes, and Fig. 3D shows the performance of the model with 7 genes. The figure illustrates the changes in model accuracy (y-axis) with the number of genes (x-axis), with the model with 7 genes selected for subsequent analysis. **D** LASSO regression analysis results. The figure shows the results of LASSO regression analysis, with the x-axis representing the logarithm of the regularization parameter λ and the y-axis representing the regression coefficients of each gene. The changes in gene coefficients with varying λ are shown to screen for key genes. **E** Random Forest analysis results. **E** displays the gene importance ranking, and Fig. 3F shows genes with importance greater than 2. **G** Intersection of screening results from three machine learning methods. The Venn diagram shows the intersection of the screening results from LASSO regression, Support Vector Machine, and Random Forest methods, ultimately identifying three key genes: CDK1, CDK4, and BCL2

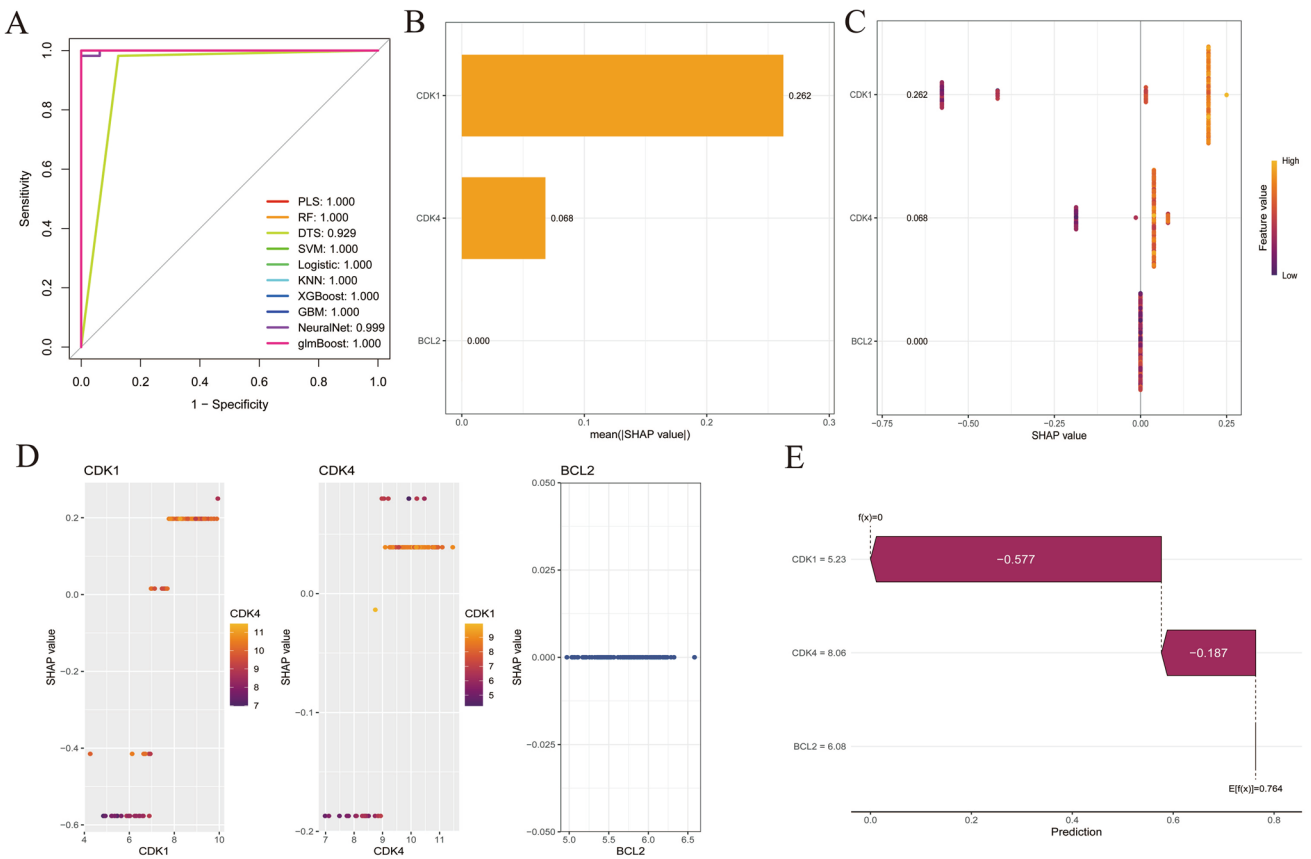



**Fig. 5** **A** ROC curve analysis. The figure displays the ROC curve analysis results for ten machine learning models based on the GEO dataset. **B**: Bar chart of key gene importance. The bar chart displays the importance scores of the three key genes: CDK1, CDK4, and BCL2. **C** Variable importance beeswarm plot. The beeswarm plot displays the importance scores of CDK1, CDK4, and BCL2, with color intensity indicating gene importance, further confirming the higher importance of CDK1 and CDK4. **D** Variable dependence plot. The figure displays the variable dependence plots for CDK1 and CDK4, with the x-axis representing gene expression levels and the y-axis representing SHAP values. The plots show that as gene expression levels increase, the importance of SHAP values also increases. **E** Variable attribution waterfall plot. The waterfall plot displays the SHAP value attribution for CDK1 and CDK4, showing each gene's contribution to model predictions in bar chart form, further supporting the importance of CDK1 and CDK4

Activated CD4 T cells, Type 2 T helper cells, Effector memory CD4 T cells, Gamma delta T cells, and Memory B cells (Fig. 6D, $P < 0.001$); CDK4 was significantly correlated with CD56 bright natural killer cells, Activated CD4 T cells, and Gamma delta T cells (Fig. 6E, $P < 0.001$).

### 3.6 Strong binding affinity of DEHP with key proteins

We investigated the binding capacity of DEHP with the three key gene proteins (BCL2, CDK1, and CDK4) through molecular docking experiments in this study. The results showed that the binding energies of DEHP with BCL2, CDK1, and CDK4 were – 8.7, – 7.8, and – 6.8 kcal/mol, respectively, indicating that DEHP can bind well with these proteins (Fig. 7A–C).

## 4 Discussion

Colorectal cancer is one of the most common gastrointestinal malignancies, with an incidence rate that is increasing annually and a high mortality rate [35–37]. DEHP has been found to have a direct and close association with colorectal cancer [38, 39]. Investigating the relationship between the two from this perspective holds significant importance. Our study provides a comprehensive investigation into the potential oncogenic mechanisms of DEHP in CRC through integrative bioinformatics analysis and molecular docking experiments. Our integrative bioinformatics analysis and molecular docking experiments identified three pivotal genes—CDK1, CDK4, and BCL2—as key

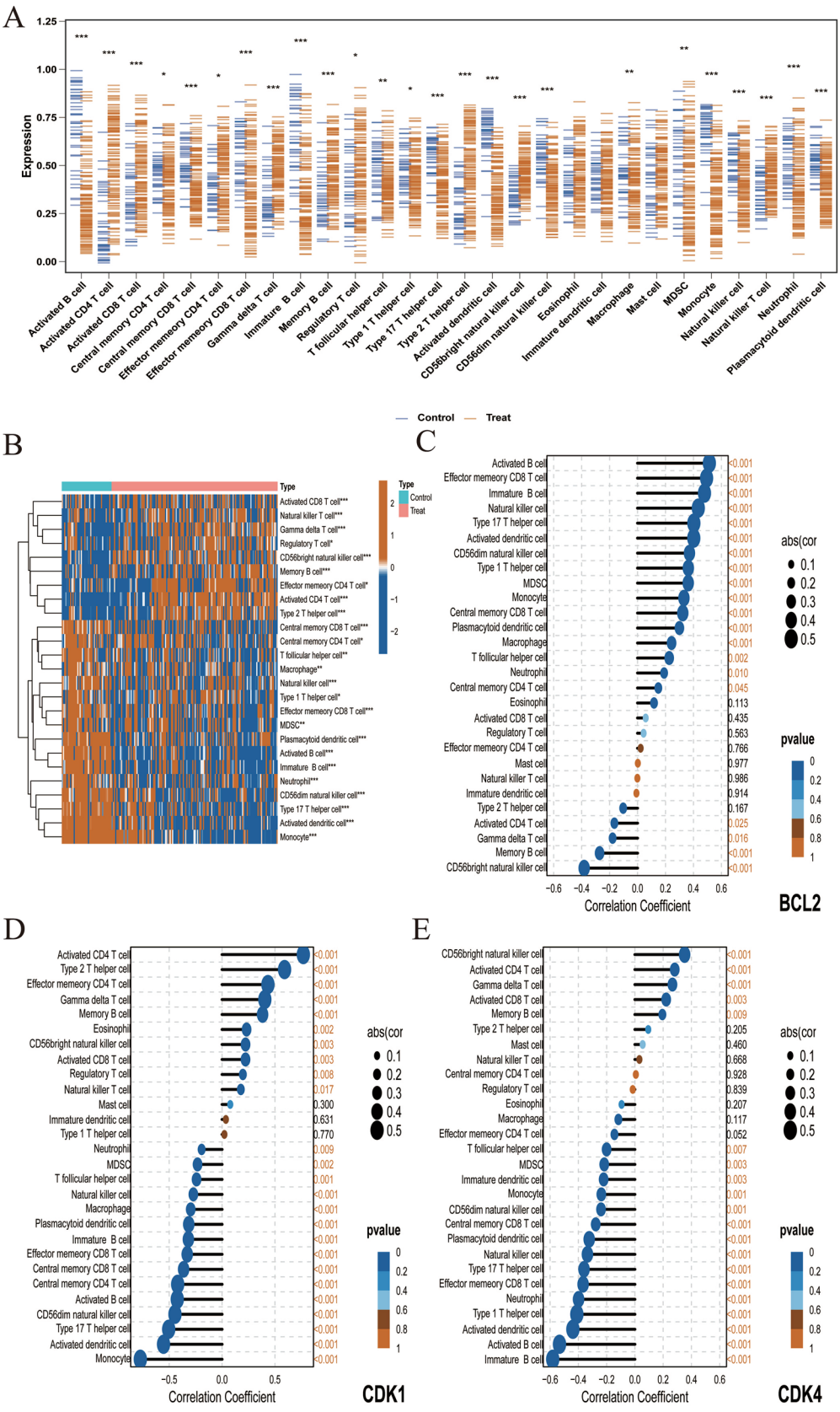



**Fig. 6** **A**-**B** Immune cell infiltration analysis. The bar chart displays the differences in immune cell infiltration between colorectal cancer patients and controls in the GEO database. The x-axis represents immune cell types, and the y-axis represents the degree of immune cell infiltration. The figure shows higher infiltration levels of active CD4 T cells, active CD8 T cells, memory CD8 T cells, type 2 T helper cells, and natural cells in colorectal cancer patients ($P < 0.0001$). **C**-**E** Correlation between key genes and immune cells. **C**, **D**, and **E** display the correlations between BCL2, CDK1, and CDK4 with immune cells, respectively. The x-axis represents immune cell types, and the y-axis represents correlation coefficients. Significant correlations are indicated with asterisks ($P < 0.001$)

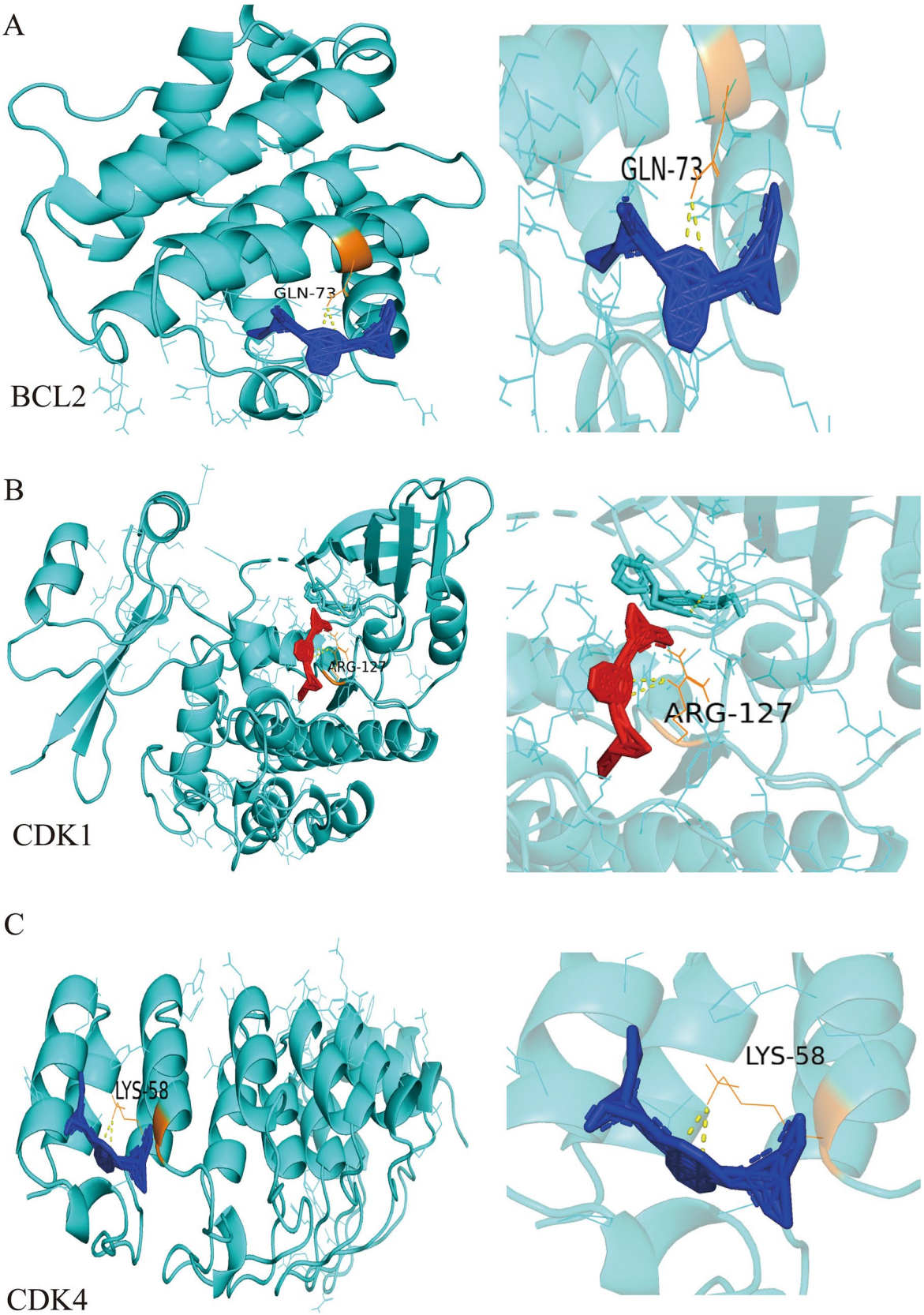



**Fig. 7** **A**, **B**, and **C** Molecular docking experiment results. The figure displays the binding energies of DEHP with BCL2, CDK1, and CDK4. The binding energies are -8.7, -7.8, and -6.8 kcal/mol, respectively, indicating that DEHP can bind well with these proteins

mediators of DEHP-induced colorectal cancer progression. By intersecting DEHP-associated targets with colorectal cancer gene expression profiles, we discovered 86 common genes, which were subsequently refined via protein–protein interaction network analysis and machine learning approaches. Notably, molecular docking revealed strong binding affinities between DEHP and these proteins (binding energies: CDK1 at − 7.8 kcal/mol, CDK4 at − 6.8 kcal/mol, and BCL2 at − 8.7 kcal/mol). These findings suggest that DEHP may contribute to colorectal oncogenesis by modulating critical pathways such as PI3K–Akt and p53, highlighting potential targets for therapeutic intervention.

The identification of 86 common genes intersecting between DEHP targets, disease-associated genes, and differentially expressed genes from colorectal cancer patients highlights the complex interplay between environmental factors and genetic pathways in CRC. These genes are enriched in critical biological processes and signaling pathways, such as the MAPK cascade, PI3K – Akt signaling, and p53 pathways, which are well-known for their roles in cancer development [40–43]. Sheng et al. [44] demonstrated that dysregulation of these signaling cascades plays a critical role in driving tumor growth and metastasis in colorectal cancer, thereby reinforcing our findings on the importance of these pathways in CRC progression [44]. For instance, the SENP1-YBX1-AKT signaling axis has been shown to promote CRC progression through AKT phosphorylation, further emphasizing the significance of these pathways as therapeutic targets [44].

The construction of a PPI network and subsequent machine learning analysis identified CDK1, CDK4, and BCL2 as key genes in CRC. These genes have been implicated in various cancers, including CRC, and their identification in our study underscores their potential as therapeutic targets [45–47]. Notably, CDK1 and CDK4 emerged as top contributors based on SHAP analysis, highlighting their critical roles in CRC development. This finding is supported by previous research indicating that CDK1 is involved in cell cycle regulation and tumor progression.

Our analysis of immune cell infiltration revealed significant correlations between CDK1, CDK4, and BCL2 and various immune cells, suggesting that DEHP may influence the tumor microenvironment by modulating immune cell infiltration. This is consistent with emerging evidence that the tumor immune microenvironment plays a crucial role in CRC progression and response to therapy [48, 49]. For example, Rauth et al. have shown that immune-associated gene signatures can predict the progression of atherosclerotic plaques and response to immunotherapy, highlighting the importance of considering immune cell infiltration in cancer research [50].

Molecular docking experiments demonstrated strong binding affinities between DEHP and the key proteins, providing molecular evidence for DEHP's oncogenic potential in CRC. This finding is particularly relevant given the growing body of research on the role of environmental factors in cancer development. For instance, a recent study identified N-glycosylation modifications of Cathepsin D (CTSD) as a potential therapeutic target for CRC liver metastasis, further emphasizing the importance of understanding the molecular interactions driving CRC progression [51].

Our integrative bioinformatics analysis and molecular docking experiments identified three pivotal genes—CDK1, CDK4, and BCL2—as key mediators of DEHP-induced colorectal cancer progression. By intersecting DEHP-associated targets with colorectal cancer gene expression profiles, we discovered 86 common genes, which were subsequently refined via protein–protein interaction network analysis and machine learning approaches. Molecular docking revealed strong binding affinities between DEHP and these proteins (binding energies: CDK1 at – 7.8 kcal/mol, CDK4 at – 6.8 kcal/mol, and BCL2 at – 8.7 kcal/mol). These findings suggest DEHP may contribute to colorectal oncogenesis not only through influencing critical pathways such as PI3K–Akt and p53 but also by directly interacting at the molecular level with these key proteins. For instance, the direct binding of DEHP to BCL2 may enhance its anti-apoptotic function, facilitating tumor cell survival and proliferation. Similarly, DEHP's interaction with CDK1 and CDK4 proteins could directly activate their kinase activities, promoting cell cycle progression and unchecked proliferation.

Furthermore, our analysis highlighted significant correlations between CDK1, CDK4, BCL2, and various immune cells within the tumor microenvironment. This observation suggests that DEHP's oncogenic potential extends beyond direct intracellular interactions to modulate the immune microenvironment, possibly affecting tumor immune evasion and progression. DEHP binding to BCL2 might alter tumor cell immunogenicity, influencing the activity of effector T cells and natural killer cells, thereby facilitating immune evasion. Future experimental validation in vitro and in vivo is necessary to comprehensively elucidate the molecular mechanisms underlying DEHP's role in CRC progression.

In a word, our study elucidates the potential oncogenic mechanisms of DEHP in CRC by identifying key genes and pathways involved in its action. These findings not only provide new insights into the pathogenesis of CRC but also highlight potential therapeutic targets for future research. Future studies should focus on experimental validation of these targets and exploring the clinical applications of the identified pathways. Additionally, our results underscore the need for further investigation into the role of environmental factors like DEHP in cancer development and the importance of targeting the tumor microenvironment in therapeutic strategies.

However, our study is not without limitations. Firstly, the analysis relies on raw data sourced from databases, which may introduce sample size-related constraints. Additionally, despite the large sample size, our research has only conducted preliminary network toxicological analyses and has yet to validate or elucidate a more detailed mechanism of action through fundamental experiments. Future studies should focus on further verifying the underlying mechanisms.

## 5 Conclusion

In summary, our study employs an integrative approach combining transcriptomic analysis, machine learning, and molecular docking to elucidate the oncogenic mechanisms of DEHP in colorectal cancer. We identified 86 common genes from DEHP-associated and CRC-related datasets and refined this list to three key genes—CDK1, CDK4, and BCL2—using a robust protein–protein interaction network and advanced machine learning techniques. These genes are critically involved in pivotal signaling pathways, including PI3K–Akt and p53, which regulate cell proliferation, apoptosis, and tumor progression. Molecular docking experiments further confirmed strong binding affinities between DEHP and these target proteins, while immune cell infiltration analysis highlighted DEHP's potential role in modulating the tumor microenvironment. Collectively, our findings not only deepen the understanding of DEHP's contribution to colorectal oncogenesis but also identify promising therapeutic targets for future intervention strategies.

**Acknowledgements** No.

**Author contributions** Z.Z.: Investigation, Methodology, Validation, Writing. J.Q. and C.G.H.: Writing—Review & Editing. S.Y.W. and Y.L.L: Review & Editing. S.W: Resources, Supervision. X.G.Z: Funding Acquisition.

**Funding** This study was partially funded by the General Research Program of Guangxi Natural Science Foundation. (Grant No. 2023GXNSFAA026257).

**Data availability** All data generated or analyzed during this study can be obtained directly by contacting the corresponding author (X.G.Z.).

**Code availability** Further enquiries can be directed to the corresponding author.

## Declarations

**Ethics approval and consent to participate** The data used in this study were obtained from public databases, therefore no additional ethical certification was required.

**Consent for publication** Not applicable.

**Competing interests** The authors declare no competing interests.



## References

1. Patel SG, Dominitz JA. Screening for colorectal cancer. Ann Intern Med. 2024;177(4):Itc49–64.
2. Cañellas-Socias A, Sancho E, Batlle E. Mechanisms of metastatic colorectal cancer. Nat Rev Gastroenterol Hepatol. 2024;21(9):609–25.
3. Zheng RS, Chen R, Han BF, Wang SM, Li L, Sun KX, He J. Cancer incidence and mortality in China, 2022. Zhonghua Zhong Liu Za Zhi. 2024;46(3):221–31.
4. Sun Y, Zhang X, Hang D, Lau HC, Du J, Liu C, Yu J. Integrative plasma and fecal metabolomics identify functional metabolites in adenoma-colorectal cancer progression and as early diagnostic biomarkers. Cancer Cell. 2024;42(8):1386-1400.e8.
5. Abedizadeh R, Majidi F, Khorasani HR, Abedi H, Sabour D. Colorectal cancer: a comprehensive review of carcinogenesis, diagnosis, and novel strategies for classified treatments. Cancer Metastasis Rev. 2024;43(2):729–53.

6. Jiang C, Zhou Q, Yi K, Yuan Y, Xie X. Colorectal cancer initiation: understanding early-stage disease for intervention. Cancer Lett. 2024;589: 216831.
7. Gupta S, May FP, Kupfer SS, Murphy CC. Birth cohort colorectal cancer (CRC): implications for research and practice. Clin Gastroenterol Hepatol. 2024;22(3):455-469.e7.
8. Tjader NP, Toland AE. Immunotherapy for colorectal cancer: insight from inherited genetics. Trends Cancer. 2024;10(5):444–56.
9. Singh M, Morris VK, Bandey IN, Hong DS, Kopetz S. Advancements in combining targeted therapy and immunotherapy for colorectal cancer. Trends Cancer. 2024;10(7):598–609.
10. Dosunmu GT, Shergill A. Colorectal cancer: genetic underpinning and molecular therapeutics for precision medicine. Genes (Basel). 2024. https://doi.org/10.3390/genes15050538.
11. Giardina C, Kuo A, Nito K, Kurkcu S. Early onset colorectal cancer: cancer promotion in young tissue. Biochem Pharmacol. 2024;226: 116393.
12. Sun Z, Ma T, Huang Z, Lu J, Xu L, Wang Y, Xiao Y. Robot-assisted radical resection of colorectal cancer using the KangDuo surgical robot versus the da Vinci Xi robotic system: short-term outcomes of a multicentre randomised controlled noninferiority trial. Surg Endosc. 2024;38(4):1867–76.
13. Nakamura Y, Watanabe J, Akazawa N, Hirata K, Kataoka K, Yokota M, Oki E. ctDNA-based molecular residual disease and survival in resectable colorectal cancer. Nat Med. 2024;30(11):3272–83.
14. Zheng H, Liu J, Cheng Q, Zhang Q, Zhang Y, Jiang L, Chen Q. Targeted activation of ferroptosis in colorectal cancer via LGR4 targeting overcomes acquired drug resistance. Nat Cancer. 2024;5(4):572–89.
15. Shitara K, Muro K, Watanabe J, Yamazaki K, Ohori H, Shiozawa M, Yoshino T. Baseline ctDNA gene alterations as a biomarker of survival after panitumumab and chemotherapy in metastatic colorectal cancer. Nat Med. 2024;30(3):730–9.
16. Singh U, Kokkanti RR, Patnaik S. Beyond chemotherapy: Exploring 5-FU resistance and stemness in colorectal cancer. Eur J Pharmacol. 2025;991: 177294.
17. Li S, Gu X, Zhang M, Jiang Q, Xu T. Di (2-ethylhexyl) phthalate and polystyrene microplastics co-exposure caused oxidative stress to activate NF-κB/NLRP3 pathway aggravated pyroptosis and inflammation in mouse kidney. Sci Total Environ. 2024;926: 171817.
18. Wang X, Li D, Zheng X, Hong Y, Zhao J, Deng W, Wu S. Di-(2-ethylhexyl) phthalate induces ferroptosis in prepubertal mouse testes via the lipid metabolism pathway. Environ Toxicol. 2024;39(3):1747–58.
19. Zhang H, Liu D, Chen J. Di-2-ethylhexyl phthalate (DEHP) exposure increase female infertility. Reprod Toxicol. 2024;130: 108719.
20. Song P, Lv D, Yang L, Zhou J, Yan X, Liu Z, Dong Q. Di-(2-ethylhexyl) phthalate promotes benign prostatic hyperplasia through KIF11-Wnt/β-catenin signaling pathway. Ecotoxicol Environ Saf. 2024;281: 116602.
21. Tang L, Wang Y, Yan W, Zhang Z, Luo S, Wen Q, Xu Y. Exposure to di-2-ethylhexyl phthalate and breast neoplasm incidence: a cohort study. Sci Total Environ. 2024;926: 171819.
22. Zhu Y, Ma XY, Cui LG, Xu YR, Li CX, Talukder M, Li JL. Di (2-ethylhexyl) phthalate induced lipophagy-related renal ferroptosis in quail (Coturnix japonica). Sci Total Environ. 2024;919: 170724.
23. Wen Y, Rattan S, Flaws JA, Irudayaraj J. Multi and transgenerational epigenetic effects of di-(2-ethylhexyl) phthalate (DEHP) in liver. Toxicol Appl Pharmacol. 2020;402: 115123.
24. Lee CY, Suk FM, Twu YC, Liao YJ. Long-term exposure to low-dose Di-(2-ethylhexyl) phthalate impairs cholesterol metabolism in hepatic stellate cells and exacerbates liver librosis. Int J Environ Res Public Health. 2020. https://doi.org/10.3390/ijerph17113802.
25. Wang MC, Wang BF, Ren HT, Huang YQ, Jing C, Pan JY, Ma HB. Exposure to endocrine disruptor DEHP promotes the progression and radiotherapy resistance of pancreatic cancer cells by increasing BMI1 expression and properties of cancer stem cells. Ecotoxicol Environ Saf. 2024;283: 116970.
26. Zhao L, Zheng J, Qin J, Xu X, Liu X, Yang S, Dong R. Combined Astragalus, vitamin C, and vitamin E alleviate DEHP-induced oxidative stress and the decreased of insulin synthesis and secretion in INS-1 cells. Ecotoxicol Environ Saf. 2023;268: 115675.
27. Lin F, Zheng WC, Ke ZB, Chen DN, Xue YT, Lin YZ, Xu N. A comprehensive analysis-based study of Di-(2-ethylhexyl) phthalate (DEHP)-Environmental explanation of bladder cancer progression. Environ Pollut. 2025;367: 125625.
28. Xu K, Wang Y, Gao X, Wei Z, Han Q, Wang S, Chen M. Polystyrene microplastics and di-2-ethylhexyl phthalate co-exposure: Implications for female reproductive health. Environ Sci Ecotechnol. 2024;22: 100471.
29. Zhang Y, Li J, Shi W, Lu L, Zhou Q, Zhang H, Yin L. Di(2-ethylhexyl) phthalate induces reproductive toxicity and transgenerational reproductive aging in Caenorhabditis elegans. Environ Pollut. 2023;336: 122259.
30. Shi YQ, Fu GQ, Zhao J, Cheng SZ, Li Y, Yi LN, Zhang DY. Di(2-ethylhexyl)phthalate induces reproductive toxicity via JAZF1/TR4 pathway and oxidative stress in pubertal male rats. Toxicol Ind Health. 2019;35(3):228–38.
31. Shigano M, Takashima R, Satomoto K, Sales H, Harada R, Hamada S. Confirmation of Di(2-ethylhexyl) phthalate-induced micronuclei by repeated dose liver micronucleus assay: focus on evaluation of liver micronucleus assay in young rats. Genes Environ. 2024;46(1):17.
32. Li L, Wang F, Zhang J, Wang K, De X, Li L, Zhang Y. Typical phthalic acid esters induce apoptosis by regulating the PI3K/Akt/Bcl-2 signaling pathway in rat insulinoma cells. Ecotoxicol Environ Saf. 2021;208: 111461.
33. Chen HP, Pan MH, Chou YY, Sung C, Lee KH, Leung CM, Hsu PC. Effects of di(2-ethylhexyl)phthalate exposure on 1,2-dimethyhydrazine-induced colon tumor promotion in rats. Food Chem Toxicol. 2017;103:157–67.
34. Su WC, Tsai YC, Chang TK, Yin TC, Tsai HL, Huang CW, Wang JY. Correlations between urinary monoethylhexyl phthalate concentration in healthy individuals, individuals with colorectal adenomas, and individuals with colorectal cancer. J Agric Food Chem. 2021;69(25):7127–36.
35. Zielińska A, Włodarczyk M, Makaro A, Sałaga M, Fichna J. Management of pain in colorectal cancer patients. Crit Rev Oncol Hematol. 2021;157: 103122.
36. Li Q, Geng S, Luo H, Wang W, Mo YQ, Luo Q, Xu B. Signaling pathways involved in colorectal cancer: pathogenesis and targeted therapy. Signal Transduct Target Ther. 2024;9(1):266.
37. Gogoi P, Kaur G, Singh NK. Nanotechnology for colorectal cancer detection and treatment. World J Gastroenterol. 2022;28(46):6497–511.
38. Shih PC, Chen HP, Hsu CC, Lin CH, Ko CY, Hsueh CW, Lee YK. Long-term DEHP/MEHP exposure promotes colorectal cancer stemness associated with glycosylation alterations. Environ Pollut. 2023;327: 121476.

39. Wang G, Chen Q, Tian P, Wang L, Li X, Lee YK, Chen W. Gut microbiota dysbiosis might be responsible to different toxicity caused by Di-(2-ethylhexyl) phthalate exposure in murine rodents. Environ Pollut. 2020;261: 114164.
40. Brown BA, Myers PJ, Adair SJ, Pitarresi JR, Sah-Teli SK, Campbell LA, Lazzara MJ. A histone methylation-MAPK signaling axis drives durable epithelial-mesenchymal transition in hypoxic pancreatic cancer. Cancer Res. 2024;84(11):1764–80.
41. Bu L, Zhang Z, Chen J, Fan Y, Guo J, Su Y, Guo J. High-fat diet promotes liver tumorigenesis via palmitoylation and activation of AKT. Gut. 2024;73(7):1156–68.
42. Browne IM, André F, Chandarlapaty S, Carey LA, Turner NC. Optimal targeting of PI3K-AKT and mTOR in advanced oestrogen receptor-positive breast cancer. Lancet Oncol. 2024;25(4):e139–51.
43. Peuget S, Zhou X, Selivanova G. Translating p53-based therapies for cancer into the clinic. Nat Rev Cancer. 2024;24(3):192–215.
44. Sheng Z, Luo S, Huang L, et al. SENP1-mediated deSUMOylation of YBX1 promotes colorectal cancer development through the SENP1-YBX1-AKT signaling axis. Oncogene. 2025. https://doi.org/10.1038/s41388-025-03302-6.
45. Westaby D, Jiménez-Vacas JM, Figueiredo I, Rekowski J, Pettinger C, Gurel B, Sharp A. BCL2 expression is enriched in advanced prostate cancer with features of lineage plasticity. J Clin Invest. 2024. https://doi.org/10.1172/JCI179998.
46. Ci M, Zhao G, Li C, Liu R, Hu X, Pan J, Cui H. OTUD4 promotes the progression of glioblastoma by deubiquitinating CDK1 and activating MAPK signaling pathway. Cell Death Dis. 2024;15(3):179.
47. Kudo R, Safonov A, Jones C, Moiso E, Dry JR, Shao H, Chandarlapaty S. Long-term breast cancer response to CDK4/6 inhibition defined by TP53-mediated geroconversion. Cancer Cell. 2024;42(11):1919-1935.e9.
48. Garg P, Ramisetty SK, Raghu Subbalakshmi A, Krishna BM, Pareek S, Mohanty A, Singhal SS. Gynecological cancer tumor Microenvironment: Unveiling cellular complexity and therapeutic potential. Biochem Pharmacol. 2024;229:116498.
49. Khosravi GR, Mostafavi S, Bastan S, Ebrahimi N, Gharibvand RS, Eskandari N. Immunologic tumor microenvironment modulators for turning cold tumors hot. Cancer Commun (Lond). 2024;44(5):521–53.
50. Rauth S, Malafa M, Ponnusamy MP, Batra SK. Emerging trends in gastrointestinal cancer targeted therapies: harnessing tumor microenvironment, immune factors, and metabolomics insights. Gastroenterology. 2024;167(5):867–84.
51. Xiong N, Du Y, Huang C, Yan Q, Zhao L, Yang C, Shen Z. N-glycosylation modification of CTSD affects liver metastases in colorectal cancer. Adv Sci (Weinh). 2025;12(7): e2411740.