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.

Integrated in silico network toxicology, machine learning and molecular simulations to predict candidate colorectal cancer-related targets of three novel non-phthalate plasticizers

Wenzhu Lou
Ningbo Hospital of Integrated Traditional Chinese and Western Medicine
Wei Ren
Ningbo Hospital of Integrated Traditional Chinese and Western Medicine
Bangsheng Chen
cbs113@126.com

Ningbo Hospital of Integrated Traditional Chinese and Western Medicine

Research Article

Keywords: Colorectal cancer, Non-phthalate plasticizers, Network toxicology, Machine learning, Molecular docking, Molecular dynamics simulation

Posted Date: November 6th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11311223/v1

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

Additional Declarations: No competing interests reported.

Integrated *in silico* network toxicology, machine learning and molecular simulations to predict candidate colorectal cancer-related targets of three novel non-phthalate plasticizers

Wenzhu Lou [1], Wei Ren[1], Bangsheng Chen [2*]

1 Department of General Practice, Ningbo Hospital of Integrated Traditional Chinese and Western Medicine, No. 998, Qianhe Road, Yinzhou District, Ningbo, Zhejiang 315100, China

2 Emergency Medical Center, Ningbo Hospital of Integrated Traditional Chinese and Western Medicine, No. 998, Qianhe Road, Yinzhou District, Ningbo, Zhejiang 315100, China

[*]Corresponding author:

Bangsheng Chen

Emergency Medical Center, Ningbo Hospital of Integrated Traditional Chinese and Western Medicine, No. 998, Qianhe Road, Yinzhou District, Ningbo, Zhejiang 315100, China

Tel: +86 15968091821

E-mail: cbs113@126.com

ORCID: https://orcid.org/0000-0002-6428-8248

ORCIDs: Wenzhu Lou: https://orcid.org/0009-0009-4019-9089

Other author's email:

Wenzhu Lou, 793314264@qq.com

Wei Ren, renwei1121@163.com

Abstract

Colorectal cancer (CRC) remains a major global health burden, and chronic exposure to environmental xenobiotics is increasingly implicated in colorectal tumourigenesis. Alternative non-phthalate plasticizers (DOTP, DINCH and TMCH) have been widely adopted as safer substitutes for conventional phthalates, yet their colorectal carcinogenic potential and underlying molecular mechanisms remain poorly defined.

We therefore implemented an integrative computational toxicology workflow combining carcinogenicity prediction, compound-target mining, functional enrichment, weighted gene co-expression network analysis (WGCNA), large-scale machine-learning screening, multi-omics validation, immune-infiltration analysis, molecular docking and 100 ns all-atom molecular dynamics simulation. TMCH exhibited the highest predicted carcinogenic probability, and 79 genes shared between plasticizer exposure and CRC pathogenesis were identified. Among 118 machine-learning pipelines, a Random Forest plus linear discriminant analysis model using 20 of the 79 genes was selected on parsimony grounds, and SHapley Additive exPlanations prioritised five core targets: *ABCG2*, *CDK1*, *CA7*, *MET* and *TDP2*. Multi-cohort transcriptomics, immunohistochemistry and single-cell profiling confirmed their dysregulated expression with distinct cell-type tropism, whereas CIBERSORT and ESTIMATE analyses linked these genes to specific immune-cell subsets and to the global immune-stromal microenvironment. All five genes achieved strong diagnostic performance but showed divergent prognostic associations. Docking predicted favourable binding of DOTP-ABCG2 and DINCH-CDK1, whereas TMCH bound weakly to all five proteins despite its high predicted risk; molecular dynamics simulations supported the structural stability of the DOTP-ABCG2 complex. Collectively, this in silico study provides testable hypotheses rather than causal proof and nominates five candidate targets for experimental toxicological validation.

Keywords: Colorectal cancer, Non-phthalate plasticizers, Network toxicology, Machine learning, Molecular docking, Molecular dynamics simulation

1. Introduction

Colorectal cancer (CRC) represents one of the most prevalent malignancies worldwide and constitutes a substantial global public-health burden, ranking third in cancer incidence and second in cancer-associated mortality globally[1]. Despite continuous advances in endoscopic screening, surgical resection and combinatorial chemotherapeutic regimens, a considerable proportion of CRC patients are diagnosed at advanced metastatic stages, accompanied by poor therapeutic response and dismal clinical outcomes[2]. In addition to well-established risk factors including genetic predisposition, dietary patterns, age and inflammatory bowel disease, accumulating epidemiological and toxicological evidence indicates that chronic exposure to environmental xenobiotics serves as a modifiable risk contributor for colorectal tumourigenesis[3]. Therefore, identifying environmental carcinogenic hazards and deciphering their molecular mechanisms is critical for understanding CRC etiology, facilitating early-stage risk stratification and informing preventive public-health strategies.

Traditional phthalate plasticizers such as di-(2-ethylhexyl) phthalate (DEHP) have been ubiquitously applied in consumer commodities, medical devices and food-contact materials. Due to well-documented endocrine-disrupting, metabolic-perturbing and carcinogenic hazards, many jurisdictions have imposed stricter regulatory restrictions on phthalate usage in consumer products[4, 5]. To meet industrial demands, a suite of alternative non-phthalate plasticizers, including di-(2-ethylhexyl) terephthalate (DOTP), di-isononyl cyclohexane-1,2-dicarboxylate (DINCH), and 1,1,5,5-tetramethylcarbohydrazide (TMCH), have been widely promoted as safer substitutes and have gradually entered large-scale industrial application[6, 7].

Nevertheless, mounting emerging toxicological observations challenge the general assumption that these alternative plasticizers are biologically inert. Animal and cellular studies indicate DINCH can disrupt lipid homeostasis and elicit pro-inflammatory responses in liver, mammary-gland and immune-cell models, though existing toxicological evidence remains partly inconsistent between laboratory mechanistic studies and regulatory risk evaluations[8]. For DOTP, existing ecotoxicological experiments have demonstrated its capacity to trigger reactive oxygen species-mediated oxidative stress even at relatively low concentrations, which may serve as an initiating driver for cellular damage and oncogenic transformation, although robust mammalian-relevant toxicological data remain limited[9]. Furthermore, in silico carcinogenicity prediction suggests TMCH bears relatively high carcinogenic potential. Human populations are chronically exposed to these compounds via ingestion, dermal contact and inhalation. Even low-dose chronic exposure may disturb intracellular signalling cascades and remodel tissue microenvironments, potentially promoting oncogenic transformation[10]. Currently, however, mechanistic insights linking DOTP, DINCH and TMCH exposure to colorectal carcinogenesis remain fragmentary. It remains poorly characterised which molecular targets and downstream signalling cascades are perturbed upon plasticizer exposure, and how these molecular events contribute to CRC initiation and progression. Conventional toxicological assessments heavily rely on low-throughput cellular and animal experiments, which cannot comprehensively map compound-target-disease interaction networks, creating a critical knowledge gap for hazard identification and risk evaluation of alternative plasticizers.

Network toxicology provides a system-oriented computational paradigm to unravel complex toxicant-disease associations by integrating compound target prediction, multi-omics transcriptome resources and biological pathway enrichment, which enables hypothesis generation for toxic mechanisms without heavy wet-lab screening work[11]. Nevertheless, conventional network-toxicology workflows often generate large candidate gene sets, making it difficult to prioritise high-confidence core mediators among hundreds of predicted targets. Machine-learning approaches combined with SHapley Additive exPlanations (SHAP) interpretation have demonstrated remarkable capacity to distill compact high-impact molecular signatures from multi-omics datasets, simultaneously alleviating subjective manual selection and

enabling mechanistic interpretability of predictive models, which has been successfully applied in exploring toxicant-driven glioma and DEHP-associated Sjögren's syndrome molecular events[12, 13]. Despite these methodological advances, few studies have integrated network toxicology and interpretable machine-learning to systematically dissect the potential carcinogenic molecular landscape of alternative non-phthalate plasticizers in CRC.

In the present study, we constructed an integrative computational workflow combining carcinogenicity prediction, compound target mining, transcriptome-based disease-gene identification, weighted gene co-expression network analysis (WGCNA), protein-protein interaction (PPI) network, multi-algorithmic machine-learning screening, multi-omics validation and molecular docking simulation to probe the potential molecular mechanisms underlying alternative-plasticizer-associated colorectal carcinogenesis. We aimed to (i) screen overlapping genes mediating biological crosstalk between DOTP/DINCH/TMCH exposure and CRC pathogenesis; (ii) utilise comprehensive machine-learning pipelines coupled with SHAP interpretation to prioritise core molecular targets with diagnostic potential; (iii) characterise the transcriptomic, single-cell and immune-infiltration patterns of these core genes; (iv) employ molecular docking to provide structural clues for direct compound-protein binding interactions. This work provides computational hypotheses linking alternative-plasticizer exposure to CRC and offers candidate molecular biomarkers for future experimental toxicological validation and population-based risk assessment. The overall workflow of this study is illustrated in **Fig. 1.**

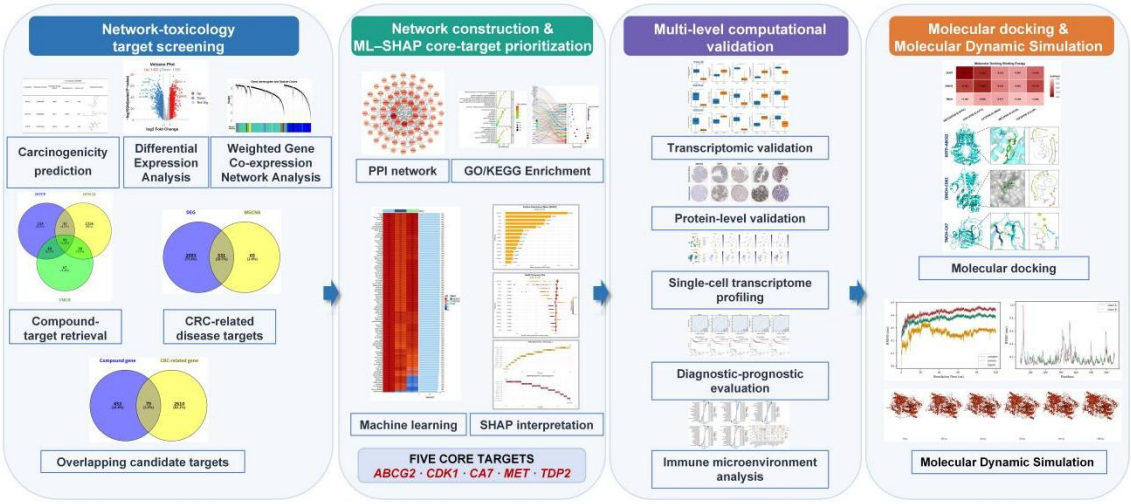


Fig. 1 Flowchart of the Study

2. Materials and Methods

2.1. Carcinogenicity Evaluation and Target Prediction of Alternative Plasticizers

The carcinogenic potentials of three alternative plasticizers: DOTP, DINCH, and TMCH were evaluated using the ADMETlab 3.0 (https://admetmesh.scbdd.com/) and ProTox-3.0 databases (https://tox.charite.de/protox3/). Their 2D chemical structures were retrieved from the PubChem database (https://pubchem.ncbi.nlm.nih.gov/). Putative protein targets for each compound were predicted using SwissTargetPrediction (https://swisstargetprediction.ch), TargetNET (http://targetnet.scbdd.com), and PharmMapper (https://lilab-ecust.cn/pharmmapper/), restricting the species to *Homo sapiens*. For SwissTargetPrediction and TargetNET, targets with a probability > 0 were retained; for PharmMapper, targets with a normalised fit score > 0 were retained. After removing duplicate entries, the remaining targets were compiled as a unified non-redundant compound-related target set for subsequent analyses.

For carcinogenicity prediction, ADMETlab 3.0 adopts the directed-message-passing neural network (DMPNN) model, where the output value represents the probability of belonging to the carcinogenic class; compounds with predicted carcinogenic probability >0.5 were regarded as potential carcinogens. ProTox-3.0 implements an integrated prediction framework combining molecular similarity comparison and ensemble machine-learning algorithms. To reduce single-model bias from individual *in silico* carcinogenicity tools, we adopted a transparent dual-source decision-making strategy for carcinogenic hazard screening. Compounds with predicted carcinogenic probability >0.5 in both ADMETlab 3.0 and ProTox-3.0 were retained for downstream network-toxicology analysis. This dual-threshold rule was pre-defined before analysis, with no post-hoc adjustment of cut-off values according to output results. Probability values from the two platforms were treated as independent computational hazard alerts rather than quantitative evidence for real-world carcinogenic potency. Compounds exhibiting carcinogenic probability >0.5 in both platforms were selected for subsequent target-disease interaction analysis.

2.2. Acquisition of CRC-Related Genes and Differential Expression Analysis

Two CRC transcriptomic datasets were obtained from the GEO database (https://www.ncbi.nlm.nih.gov/geo/): GSE21510, comprising 25 normal and 123 tumour samples, and GSE9348, comprising 12 normal and 70 tumour samples. Cross-dataset batch effects were corrected using the ComBat algorithm implemented in the “sva” R package. Principal component analysis (PCA) was performed before and after batch correction to visualise the correction efficacy. Differentially expressed genes (DEGs) between CRC and normal tissues were identified using the “limma” R package, with thresholds set at $|\log_2(\text{fold change})| \geq 1$ and adjusted $p < 0.05$. Volcano plots were generated using the “ggplot2” R package. Heatmaps were generated using

the “pheatmap” R package to display the expression patterns of the top 20 most significantly up-regulated and top 20 most significantly down-regulated genes.

2.3. Weighted Gene Co-expression Network Analysis (WGCNA)

To identify CRC phenotype-related modules, WGCNA was constructed on the merged training cohort using the “WGCNA” R package. The optimal soft-thresholding power was selected based on scale-free topology fit. Modules were defined using dynamic tree cutting, and the module most significantly correlated with the CRC status was selected. The genes within this key module were merged with the DEGs to establish a comprehensive set of CRC-related disease genes.

2.4. PPI Network Construction and Functional Enrichment

To identify biologically plausible molecular links between plasticizer exposure and colorectal carcinogenesis, we intersected the plasticizer-associated targets with the CRC-related gene set using the Venny 2.1.0 online tool (https://bioinfogp.cnb.csic.es/tools/venny/index.html). The overlapping targets were considered potential pathogenic mediators. A PPI network was constructed using the STRING database (https://cn.string-db.org/) with a confidence score ≥ 0.4, and the network was visualised using Cytoscape software (v3.10.3). Gene Ontology (GO) biological processes and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway enrichment analyses were performed using the DAVID database (https://david.ncifcrf.gov/), with an adjusted $p < 0.05$ considered statistically significant. The resulting enrichment data were subsequently visualised as GO bubble plots and KEGG Sankey-bubble diagrams using the Weishengxin online platform (https://www.bioinformatics.com.cn/).

2.5. Machine Learning-Based Core Target Screening and SHAP Interpretation

To prioritise core targets among the 79 genes shared between the plasticizer-associated and CRC-related gene sets, 118 machine-learning pipelines were constructed from twelve algorithms: Naive Bayes, least absolute shrinkage and selection operator (Lasso), random forest (RF), gradient boosting generalised linear model (glmBoost), support vector machine (SVM), Ridge regression, elastic net (Enet), partial least squares generalised linear regression (plsRglm), gradient boosting machine (GBM), linear discriminant analysis (LDA), extreme gradient boosting (XGBoost) and stepwise generalised linear models (StepGLM). These comprised 22 single-algorithm models and 96 two-stage pipelines.

Each two-stage pipeline coupled one of eight feature-selection strategies (Lasso, Ridge, elastic net at alpha = 0.5, 0.1 and 0.9, random forest, glmBoost and bidirectional StepGLM) with one of the twelve classifiers. Feature selection was performed inside each cross-validation fold on the training partition only, so that held-out samples never contributed to the definition of the gene subset. Eighty-four of

the 96 two-stage pipelines reduced the gene set (median 20 genes, range 5–66), the remainder retaining all 79 genes. Model hyperparameters were pre-specified rather than optimised by grid search, so that no information from the held-out cohorts could influence model construction. Regularised regression models (Lasso, Ridge and elastic net) used the cross-validated lambda.min obtained within the training cohort; support vector machines used the default radial kernel with cost = 1 and gamma = 1/p (p = number of features); random forests used the package default of 1,000 trees with a minimum terminal node size of 5; gradient boosting used 5,000 trees with interaction depth 3 and shrinkage 0.001, the number of trees being fixed by 5-fold internal cross-validation; XGBoost used max_depth = 3 and eta = 0.01 with the number of boosting rounds selected by 5-fold internal cross-validation; glmBoost used 100 boosting iterations with nu = 0.1; linear discriminant analysis and Naive Bayes used their default parameterisation; and plsRglm used the number of components selected by its internal cross-validation. The complete model grid, the per-pipeline gene sets and all per-cohort metrics are provided in Supplementary Data S1.

Six cohorts were analysed. The training cohort comprised two merged and batch-corrected datasets (GSE21510 and GSE9348; 230 samples, 37 normal and 193 CRC). Five independent external cohorts were used for validation and were not involved in model fitting: GSE24514 (49 samples; 15 normal and 34 CRC), GSE32323 (34; 17 and 17), TCGA-CRC (431; 51 and 380), GSE44076 (148; 50 and 98) and GSE113513 (28; 14 and 14 paired tumour and adjacent non-cancerous tissue from the same patients). TCGA-CRC expression data, comprising the colon adenocarcinoma (TCGA-COAD) and rectum adenocarcinoma (TCGA-READ) cohorts, were obtained as the combined COADREAD matrix from the UCSC Xena hub; all cohorts were restricted to the same 79 genes.

Expression values were Z-score normalised, and Gaussian noise (standard deviation 0.01) was added to the training matrix. Stratified 10-fold cross-validation with inverse class-frequency weighting was performed within the training cohort, with fixed random seeds at every step. Nine metrics were recorded for every pipeline in every cohort: AUC with its DeLong 95% confidence interval, PR-AUC (trapezoidal integration), F1-score, accuracy, sensitivity, specificity, precision, Matthews correlation coefficient and Brier score (mean squared error of the predicted probabilities), with F1 computed at the Youden-optimal threshold.

Pipelines were ranked by the mean AUC across the six cohorts, and the optimal model was selected from the leading group on the basis of parsimony and stability of the retained gene set.

SHAP values were computed with the kernelshap R package, using the predicted probability of the CRC class as the explained output and 100 randomly drawn training samples as the background (tolerance 0.05, maximum five iterations). Importance bar, summary beeswarm, dependence, force and waterfall plots were generated; genes were ranked by mean absolute SHAP value, and the five highest-ranked genes were

designated core targets. For the single-sample force and waterfall plots, one representative CRC and one representative control sample were selected from the training cohort as the samples whose predicted probability was closest to the median of their own class, so that these plots illustrate a typical rather than an extreme prediction.

2.6. Multi-omics and Single-Cell Expression Validation

To validate the expression patterns of the core targets identified by machine learning, transcriptomic, protein-level, and single-cell analyses were performed. Gene-level validation was conducted using the TCGA-CRC cohort and two independent external GEO datasets (GSE24514 and GSE32323). The expression levels of the core genes were compared between CRC and normal tissues using the Wilcoxon rank-sum test, with $p < 0.05$ considered statistically significant. Box plots were generated to visualise the expression distributions of the core genes.

Protein-level validation was performed using immunohistochemistry (IHC) images from the Human Protein Atlas (HPA) database (https://www.proteinatlas.org/). IHC staining images of the core target proteins in normal colorectal tissues and CRC tissues were retrieved and compared to evaluate whether protein expression changes were consistent with the transcriptomic findings.

For single-cell resolution, the single-cell RNA sequencing datasets CRC_GSE166555 and CRC_EMTAB8107 were analysed using the TISCH2 portal (http://tisch.comp-genomics.org/). The average expression levels (log(TPM/10+1)) of the core genes across different cell types were visualised using heatmaps. UMAP dimensionality reduction plots were generated to display the distribution and cell-type-specific expression patterns of the core genes. Major cell compartments (malignant, immune, and stromal cells) and finer subpopulations were annotated to resolve the expression specificity.

2.7. Immune Infiltration Analysis

The Cell-type Identification by Estimating Relative Subsets of RNA Transcripts (CIBERSORT) algorithm was used to estimate the relative abundances of 22 immune cell types in the TCGA-CRC cohort. Spearman correlation analysis was performed to assess the associations between core-gene expression and the abundance of each immune cell population, with $p < 0.05$ considered statistically significant. The results were visualised as lollipop plots for each of the five core genes and as a correlation heatmap covering all five genes and all 22 cell types. The Estimation of STromal and Immune cells in MAlignant Tumour tissues using Expression data (ESTIMATE) algorithm was additionally used to compute StromalScore, ImmuneScore and ESTIMATEScore for each TCGA-CRC sample, and Spearman correlation was applied to relate core-gene expression to these three metrics.

2.8. Diagnostic and Prognostic Assessment

Diagnostic receiver operating characteristic (ROC) curves were plotted using the "pROC" R package to evaluate the discriminatory ability of the core genes in distinguishing CRC from normal tissues. The area under the curve (AUC) was calculated as a quantitative metric of diagnostic performance. Prognostic values were evaluated by Kaplan-Meier (KM) survival curves and log-rank tests via the "survival" R package, with data derived from the KMplot database (https://kmplot.com/analysis/). A p-value < 0.05 was considered statistically significant.

2.9. Molecular Docking Analysis

Molecular docking was performed to evaluate the binding affinities between the three plasticizers and the core target proteins. The three-dimensional crystal structures of the target proteins were retrieved from the Protein Data Bank database (https://www.rcsb.org/), and the chemical structures of the compounds were obtained from the PubChem database. Preprocessing of proteins and ligands, including the removal of water molecules and the addition of hydrogen atoms, was carried out prior to docking simulations. Docking calculations were conducted using AutoDock Vina v1.2.7, and the binding energies (kcal/mol) were calculated to quantify the interaction strengths. The calculated binding energies were systematically summarised as a heatmap using the "pheatmap" R package for comparative analysis across different compound–protein pairs.

To visualise the detailed binding modes, three-dimensional interaction conformations were generated using PyMOL v2.3.0. For systematic characterisation of non-covalent interactions, the docked complexes were submitted to the Protein–Ligand Interaction Profiler (PLIP) web server (https://plip-tool.biotec.tu-dresden.de/plip-web/plip/), which identifies hydrogen bonds, hydrophobic contacts, π-stacking, π-cation interactions, salt bridges, and halogen bonds[14]. Two-dimensional interaction diagrams were generated using Maestro v12.8, with interaction types annotated based on the PLIP analysis results.

2.10 Molecular Dynamics Simulation

To further validate the dynamic stability of the top-ranked DOTP-ABCG2 complex from molecular docking, 100 ns all-atom molecular dynamics (MD) simulation was performed with the GROMACS[15] 2025.3 package. The initial coordinate was derived from the optimal docking conformation of *ABCG2* bound to DOTP. The Amber99sb-ildn force field was adopted for the protein[16], and the DOTP ligand topology was generated by ACPYPE with the GAFF force-field parameters[17]. The complex was solvated in a cubic TIP3P water box, and counter-ions were added to neutralize the system and adjust salt concentration to 0.15 M to mimic physiological conditions.

The system was first subjected to steepest-descent energy minimisation to resolve steric clashes. Sequential NVT (2 ns, 300 K, V-rescale thermostat) and NPT (2 ns,

1 bar, Parrinello-Rahman barostat) equilibration were carried out. Covalent bonds involving hydrogen atoms were constrained via the LINCS algorithm, and long-range electrostatics were treated using the Particle-Mesh Ewald (PME) method. Unrestricted 100 ns production simulation was run with a 2 fs timestep; trajectories were saved every 10 ps for post-processing.

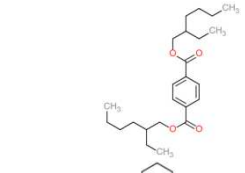
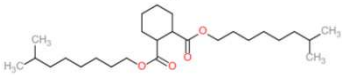
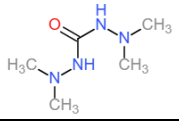
After removing periodic-boundary artifacts and trajectory fitting on protein Cα atoms, RMSD, RMSF, radius of gyration (Rg), buried SASA and interaction-energy components were computed with built-in GROMACS utilities. The free-energy landscape (FEL) was constructed using RMSD and Rg as collective variables. Representative structural snapshots were extracted at 20-ns intervals and visualised in PyMOL v2.3.0.

3. Results

3.1. Carcinogenic Probability and Target Profiles of Plasticizers

The carcinogenic probabilities of DOTP, DINCH, and TMCH were assessed by ADMETlab 3.0 and ProTox-3.0, and the corresponding predicted values are summarised in Table 1. According to the pre-defined threshold (carcinogenic probability > 0.5 for potential carcinogen), all three alternative plasticizers yielded carcinogenic probabilities above 0.5 across both prediction platforms, which supported their inclusion for the subsequent network-toxicology-based mechanism exploration. Among them, TMCH achieved the highest carcinogenic probability of 0.915 in ADMETlab 3.0, while DOTP and DINCH showed intermediate carcinogenic risk estimates in the two prediction systems. After merging prediction outputs from SwissTargetPrediction, TargetNET, and PharmMapper and removing duplicate entries, a total of 534 non-redundant putative human protein targets linked to these three plasticizers were obtained, which constituted the input dataset for downstream network toxicology analysis.

Table 1. Chemical information and carcinogenicity prediction of three alternative plasticizers.

Compound	Molecular Formula	Molecular Weight (g/mol)	Carcinogenic probability		Chemical structure
			ADMETlab 3.0	ProTox-3.0	
DOTP	$C_{24}H_{38}O_4$	390.6	0.592	0.83	
DINCH	$C_{26}H_{48}O_4$	424.7	0.606	0.77	
TMCH	$C_5H_{14}N_4O$	146.19	0.915	0.64	

Note: DOTP, Di-(2-ethylhexyl) terephthalate; DINCH, Di-isononyl cyclohexane-1,2-dicarboxylate; TMCH, 1,1,5,5-Tetramethylcarbohydrazide. Carcinogenicity probabilities were predicted using ADMETlab 3.0 (DMPNN-based model) and ProTox-3.0 (ensemble machine-learning combined with molecular similarity). Probability >0.5 indicates potential carcinogenic risk.

3.2. Identification of CRC-Related DEGs and WGCNA Key Module

PCA demonstrated substantial elimination of batch effects after ComBat correction (Fig. 2a-b). Differential expression analysis yielded 2,624 DEGs, including 1,422 up-regulated and 1,202 down-regulated genes (Fig. 2c). The top 20 most significantly up- and down-regulated genes are shown in Fig. 2d.

To identify CRC-associated co-expression modules, WGCNA was performed on the merged training cohort. The optimal soft-thresholding power was determined to be 5, at which the scale-free topology fit index (R^2) exceeded 0.85 (Fig. 2e). Hierarchical clustering resulted in 10 distinct co-expression modules, which were colour-coded for visualisation (Fig. 2f). Module–trait correlation analysis revealed that the brown module exhibited the strongest positive correlation with the CRC phenotype ($r = 0.89$, $P = 4.92 \times 10^{-78}$; Fig. 2g). Furthermore, module membership (MM) of genes in the brown module was highly correlated with gene significance (GS) for CRC (cor = 0.964, P = 0.000; Fig. 2h), indicating that genes within this module are closely associated with the disease phenotype. The brown module contained 596 genes, which were then merged with the DEGs to establish a comprehensive set of 2,689 CRC-related genes (Fig. 2i).

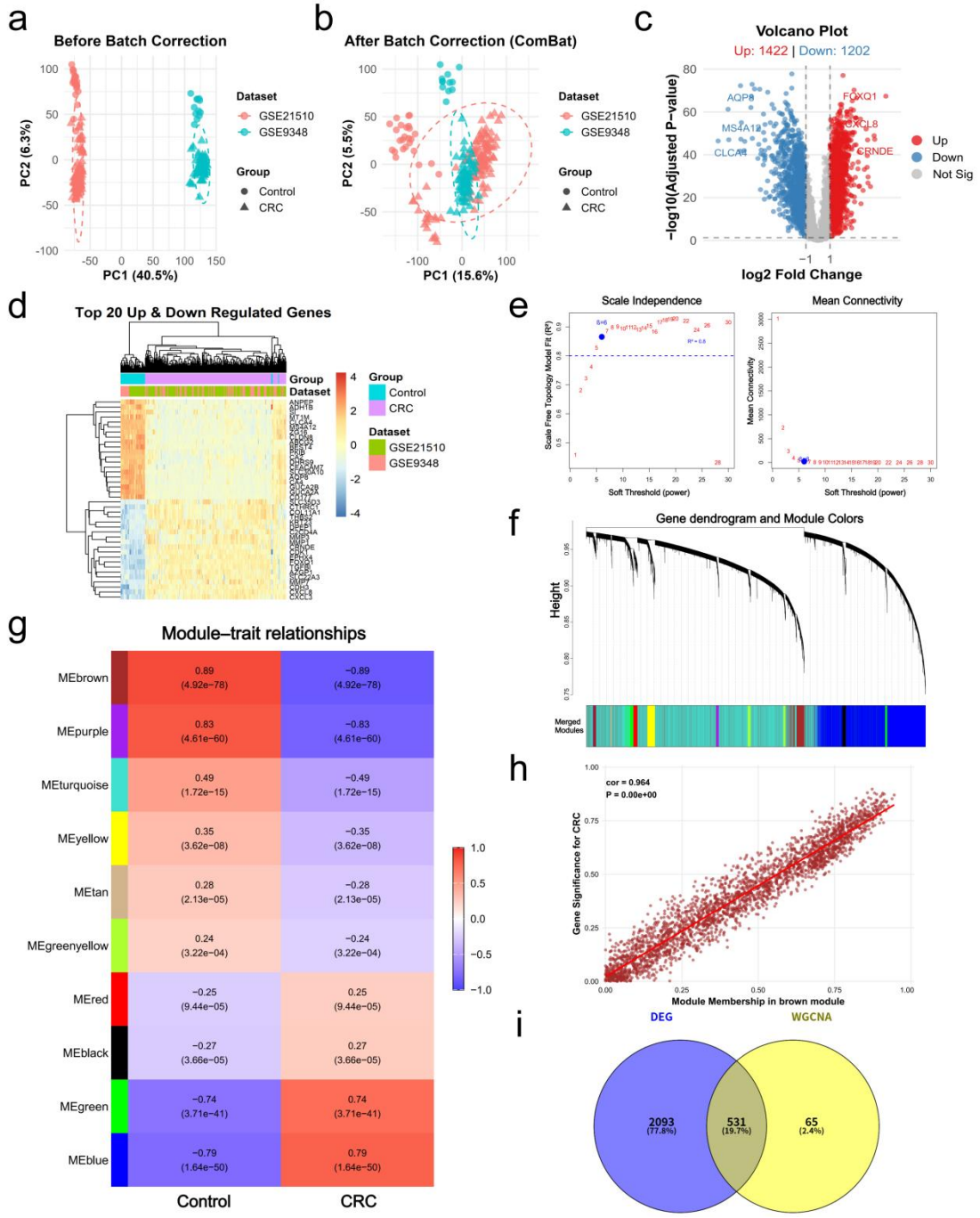


Fig. 2 Identification of CRC-related target genes. (a) PCA scatter plot before batch correction, showing clear separation between the GSE21510 and GSE9348 datasets, indicating batch effects. (b) PCA scatter plot after batch correction, demonstrating successful integration of the two datasets and marked reduction of batch effects. (c) Volcano plot displaying DEGs based on log_2FC and statistical significance; red dots indicate up-regulation, blue dots down-regulation, and gray dots non-significant genes. (d) Heatmap illustrating the expression patterns of the top 20 up- and down-regulated DEGs across samples. (e) Scale independence analysis for soft-thresholding power selection in WGCNA; the left panel displays the scale-free fit index (R^2) as a function of soft-threshold powers, and the right panel shows mean connectivity. (f) Gene dendrogram and module identification via hierarchical clustering; the colour bar beneath the dendrogram represents distinct gene

modules. (g) Module–trait relationship heatmap depicting correlations (with P-values) between WGCNA-derived modules and CRC phenotype. (h) Scatter plot illustrating the relationship between brown module membership and gene significance; a highly significant positive correlation was observed. (i) Venn diagram displaying DEGs, brown module genes, and their intersection. A total of 2,689 non-redundant genes were identified as CRC-related disease genes

3.3. PPI Network Construction and Functional Enrichment Analysis

Venn diagram analysis revealed the distribution of predicted targets among the three alternative plasticizers DOTP, DINCH, and TMCH, with 84 targets shared by all three compounds (Fig. 3a). Intersection of the 534 compound-associated targets with the 2,689 CRC-related genes identified 79 overlapping targets, which were considered potential pathogenic mediators of alternative plasticizer-induced CRC (Fig. 3b). To explore the interactions among these targets, a protein-protein interaction (PPI) network was constructed, which revealed a complex interaction landscape with several hub nodes exhibiting high connectivity (Fig. 3c), where node colour intensity and size are proportional to the degree value, with darker and larger nodes indicating higher degree centrality.

Functional enrichment analysis was then performed to elucidate the biological roles of the 79 overlapping targets. For GO enrichment (Fig.3d), the dominant biological-process terms included positive regulation of cell population proliferation, proteolysis, apoptotic-process-related signalling, and multiple transcriptional regulatory processes, together with G protein-coupled receptor signalling. At the molecular-function and cellular-component levels, enriched terms pointed to nuclear-receptor activity, peptidase and protein-kinase activity, as well as protein localization to cytosol, extracellular exosome and cell surface. KEGG pathway enrichment (Fig.3e) revealed significant enrichment in cell cycle, PPAR signalling pathway, pathways in cancer, cellular senescence and p53 signalling pathway. Collectively, these enrichment results indicate that the overlapping targets may drive colorectal malignant transformation by regulating cell proliferation, apoptosis, proteolytic remodelling and multiple oncogenic signalling cascades, although enrichment analysis alone cannot establish the directionality or the causal ordering of these effects.

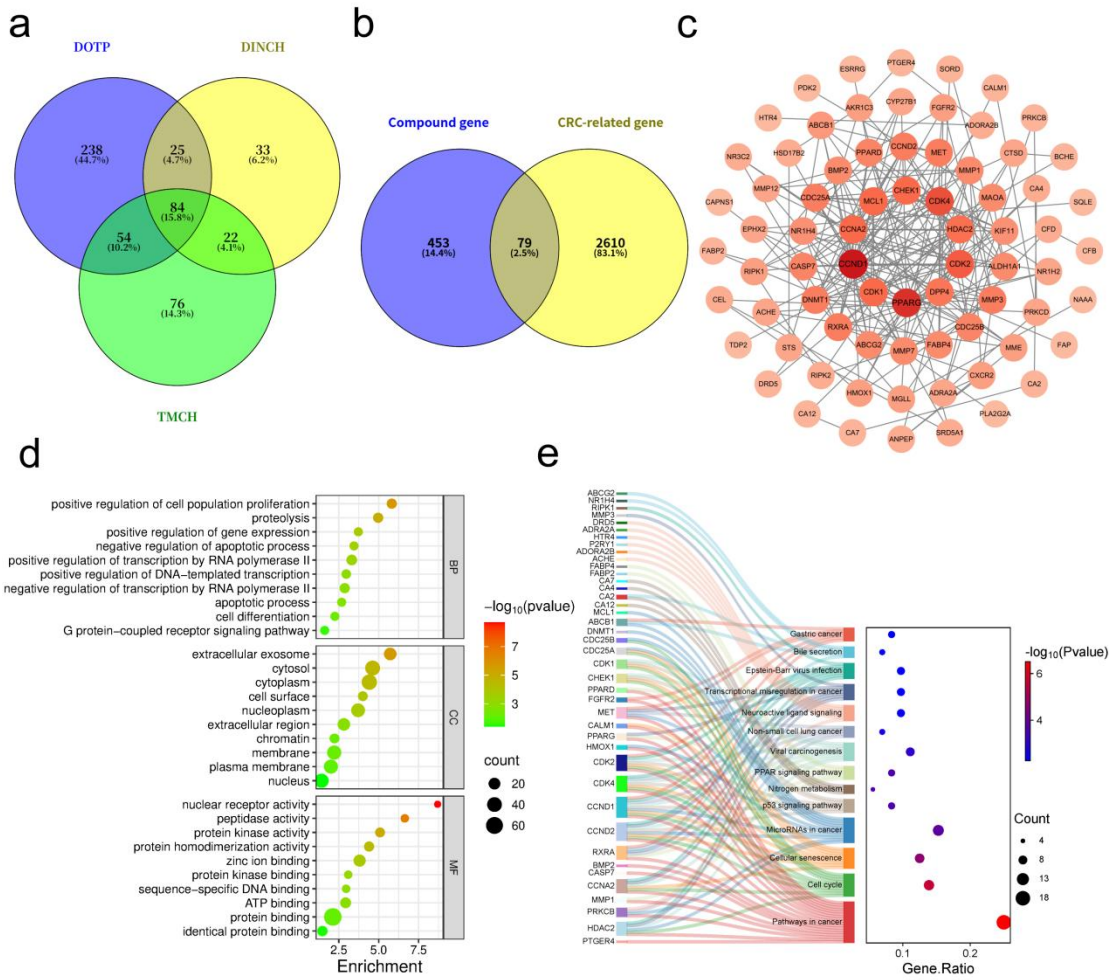


Fig. 3 Identification of overlapping targets and functional enrichment analysis. (a) Venn diagram of the three alternative plasticizer-associated targets. (b) Venn diagram of the intersection between plasticizer targets and CRC-related genes. (c) PPI network of the 79 overlapping targets. (d) GO enrichment analysis of alternative plasticizers-colorectal cancer targets. (e) KEGG enrichment analysis of alternative plasticizers-colorectal cancer targets

3.4. Performance of Machine Learning Models and Core Target Identification

All 118 pipelines converged and were successfully validated in all six cohorts (1,180 model-fold fits). The AUC values of the 50 highest-ranked pipelines are shown in Fig. 4a and those of the complete grid in Supplementary Fig. S3; the AUC of every pipeline in every cohort is given in Supplementary Table S1 and the complete metric set in Supplementary Data S1.

Discrimination was high in every cohort and the ranking of pipelines was correspondingly compressed: 65 of the 118 pipelines reached an AUC of exactly 1.000 in both GSE44076 and GSE113513, and the 15 highest-ranked pipelines differed by less than 0.005 in mean AUC. The reported model was therefore selected

for parsimony rather than nominal performance. RF+LDA achieved a mean AUC of 0.9938 from 20 genes and ranked ninth of 118, whereas the nominally top-ranked pipeline (Enet alpha = 0.1 + plsRglm) required 66 of the 79 genes to gain 0.0045 in mean AUC.

The performance of RF+LDA in each cohort is given in Table 2. Discrimination was essentially complete in GSE32323, GSE44076 and GSE113513 (AUC 1.000), and remained very high in TCGA-CRC (0.9994; DeLong 95% CI 0.9984–1.000) and GSE24514 (0.9637; 0.8926–1.000). PR-AUC ranged from 0.969 to 1.000 and the F1-score from 0.986 to 1.000. The lowest-performing cohort, GSE24514, also had the fewest controls (n = 15) and therefore the widest confidence interval.

Table 2. Performance of the reported RF+LDA model across the six cohorts.

Cohort	n (normal/CRC)	AUC (95% CI)	PR-AUC	F1	Sensitivity	Specificity
Training	230 (37/193)	1.000 (1.000–1.000)	1.000	1.000	1.000	1.000
GSE24514	49 (15/34)	0.964 (0.893–1.000)	0.969	0.986	1.000	0.933
GSE32323	34 (17/17)	1.000 (1.000–1.000)	1.000	1.000	1.000	1.000
TCGA-CRC	431 (51/380)	0.999 (0.998–1.000)	1.000	0.997	0.995	1.000
GSE44076	148 (50/98)	1.000 (1.000–1.000)	1.000	1.000	1.000	1.000
GSE113513	28 (14/14)	1.000 (1.000–1.000)	1.000	1.000	1.000	1.000

Note: AUC is given with its DeLong 95% confidence interval; F1 was computed at the Youden-optimal threshold.

SHAP analysis of the RF+LDA model ranked *ABCG2*, *CDK1*, *CA7*, *MET* and *TDP2* as the five highest-contributing genes, with mean absolute SHAP values of 0.0597, 0.0469, 0.0433, 0.0358 and 0.0261, respectively (Fig. 4b, 4c; Supplementary Table S2). No other gene exceeded a mean absolute SHAP value of 0.018, and these five genes were therefore designated the core targets for downstream validation. Dependence plots showed that increasing expression of *CDK1* and *MET* raised SHAP values, shifting predictions towards the CRC phenotype, whereas increasing expression of *ABCG2*, *CA7* and *TDP2* lowered them (Supplementary Fig. S1). Receiver operating characteristic curves of the reported RF+LDA model in each of the six cohorts are shown in Supplementary Fig. S2. A force plot for a representative CRC sample is shown in Fig. 4d and waterfall plots for one CRC and one control sample in Fig. 4e.

The near-ceiling AUC values should be interpreted with care (Supplementary Table S4). Random 79-gene panels processed through the identical pipeline reached mean AUC values of 0.997 in the training cohort, 1.000 in TCGA-CRC, 0.998 in GSE44076 and 0.976 in GSE113513; the median single-gene AUC of the 79 panel genes was 0.939, compared with 0.674 for 2,000 randomly drawn genes, indicating that the panel is enriched for discriminative genes by construction. An unsupervised classifier based on the first principal component alone also separated the two classes with an AUC of 1.000, and a leave-one-series-out scheme, in which the model was trained on two of the three source datasets and tested on the third, reached an AUC of 0.960. In addition, the merged training cohort carries a batch-class association (one of

its three source datasets contains no controls, and membership of the source dataset alone predicts the class label with an AUC of 0.789), so the training AUC is reported for completeness only. Discrimination in the independent cohorts therefore indicates that the panel carries genuine tumour-versus-normal signal rather than that a particular algorithm or gene set is superior.

Of the two additional cohorts, GSE44076 provides the better-resolved external validation, its 75th percentile DeLong confidence interval width being 0.0023 versus 0.0936 in GSE113513; GSE113513 contributes a paired design that removes between-individual variation and is reported as a sensitivity analysis (Supplementary Fig. S4, Supplementary Table S3).

Collectively, the machine-learning and SHAP analyses converged on five high-priority core targets - *ABCG2*, *CDK1*, *CA7*, *MET* and *TDP2* - that may mediate the oncogenic association between alternative-plasticizer exposure and colorectal carcinogenesis.

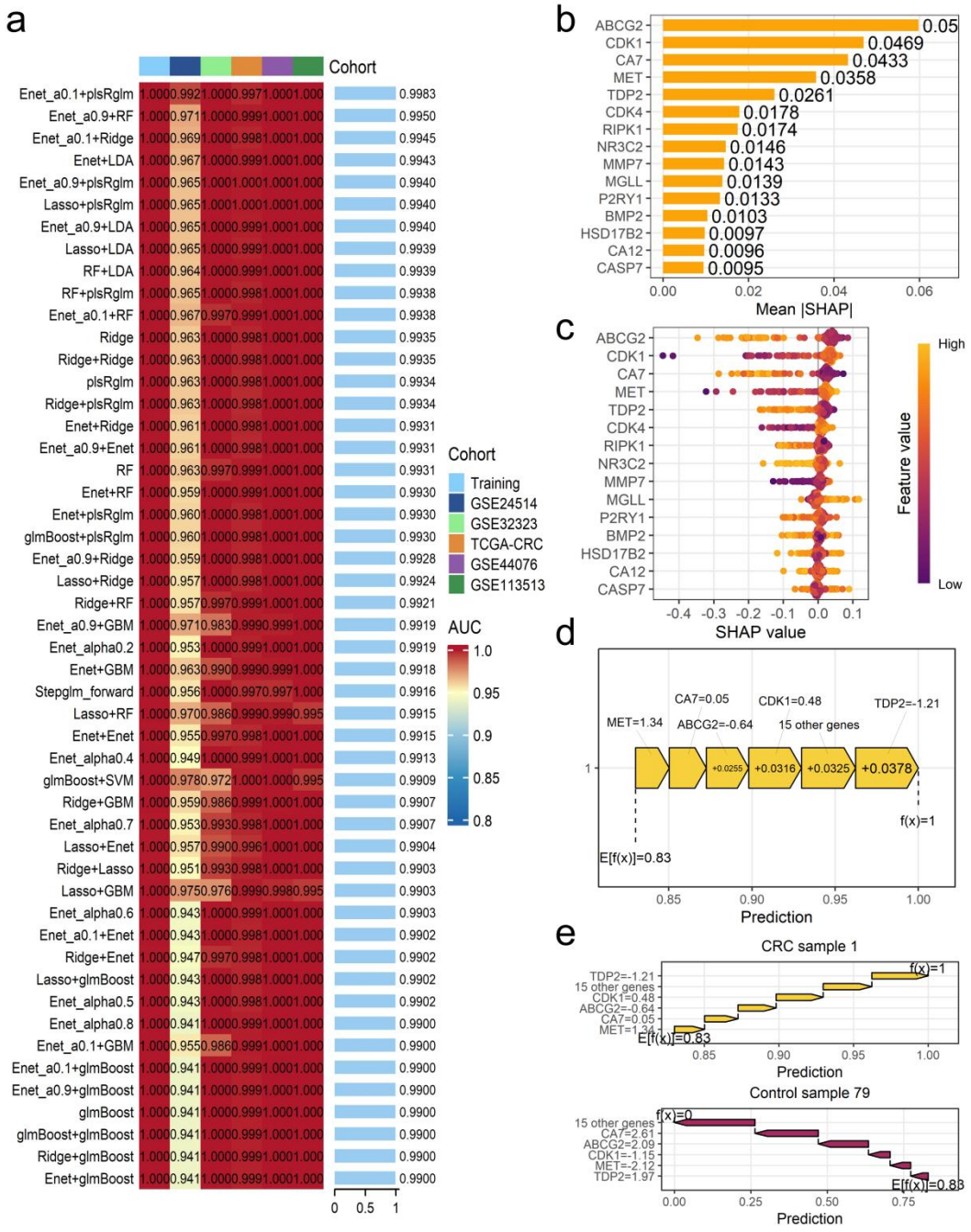


Fig. 4 Machine learning-based core target screening and SHAP interpretation. (a) Heatmap of the area under the receiver operating characteristic curve (AUC) for the 50 highest-ranked of the 118 machine-learning pipelines across the training cohort and five independent external validation cohorts (GSE24514, GSE32323, TCGA-CRC, GSE44076 and GSE113513). Rows are pipelines, columns are cohorts, and each cell gives the AUC to three decimal places. The cohort key sits to the right of the heatmap with the AUC colour bar beneath it (scale 0.80–1.00), and the horizontal bars give the mean AUC across the six cohorts. (b) Feature importance for the reported RF+LDA model, ranked by mean absolute SHAP value; longer bars indicate a greater contribution to the model output. (c) SHAP summary beeswarm plot for the same model. Each point represents one sample; its horizontal position is the SHAP value of that gene and the colour encodes the Z-score-normalised expression

(orange, high; purple, low). Positive SHAP values shift the prediction towards the colorectal cancer phenotype. (d) SHAP force plot decomposing the prediction for a representative colorectal cancer sample. Each arrow represents the SHAP contribution of one gene, with its length proportional to the magnitude of that contribution, and the numbers above the arrows give the corresponding Z-score-normalised expression values; the contributions of the 15 remaining model genes are pooled into a single arrow labelled 'other genes'. E[f(x)] is the expected model output and f(x) the predicted probability for this sample. (e) SHAP waterfall plots for one representative colorectal cancer sample (upper) and one representative control sample (lower). Bars show the stepwise accumulation of SHAP values from the expected model output E[f(x)] to the prediction f(x), with the five core targets shown individually and the remaining 15 model genes pooled into a single term

3.5. Expression and Protein-Level Validation of Core Targets

To externally validate the transcriptional dysregulation of five SHAP-prioritised core targets identified from machine-learning analysis, we compared their mRNA expression levels between CRC tumour tissues and normal colorectal mucosa across three independent transcriptomic cohorts, including TCGA-CRC, GSE24514 and GSE32323 (Fig. 5a-c). Consistent expression alterations were observed for all five genes across these independent datasets. Among them, *CDK1* and *MET* were significantly up-regulated in CRC tumour samples relative to normal controls in all three cohorts ($p < 0.001$). As critical mediators governing cell-cycle progression and oncogenic receptor tyrosine kinase signalling, their elevated transcript levels are consistent with the proliferative phenotype of colorectal malignant lesions.

In contrast, *ABCG2*, *CA7*, and *TDP2* exhibited marked down-regulation in tumour tissues compared with normal colorectal mucosa across each of the three validation cohorts ($p < 0.001$). *ABCG2*, encoding an important xenobiotic efflux transporter, *CA7*, a carbonic anhydrase responsible for intracellular and extracellular pH homeostasis, and *TDP2*, involved in DNA double-strand break repair, all showed sustained reduction in tumour specimens, implying the loss of protective physiological functions during colorectal malignant transformation.

To further determine whether these transcript-level perturbations can be recapitulated at the protein level, we retrieved IHC staining images of human colorectal tissues from the HPA database (Fig. 5d). In normal colonic epithelium, *ABCG2*, *CA7* and *TDP2* displayed moderate-to-strong immunoreactivity, whereas their protein staining intensity was obviously weakened in CRC tissues. Conversely, *CDK1* and *MET* presented faint staining signals in normal mucosal samples, but showed abundant positive protein expression in CRC tumour cells. Overall, the IHC protein-level observations were highly concordant with mRNA-expression patterns derived from bulk transcriptomic cohorts.

Taken together, multi-cohort transcriptomic comparison together with HPA protein evidence demonstrated robust and reproducible dysregulation of *ABCG2*, *CDK1*, *CA7*,

MET, and *TDP2* in CRC, providing independent multi-omics support for the core-gene list obtained from the RF+LDA-SHAP machine-learning pipeline. To further resolve which cellular compartments drive these bulk-level expression changes, we subsequently performed single-cell transcriptomic profiling analysis.

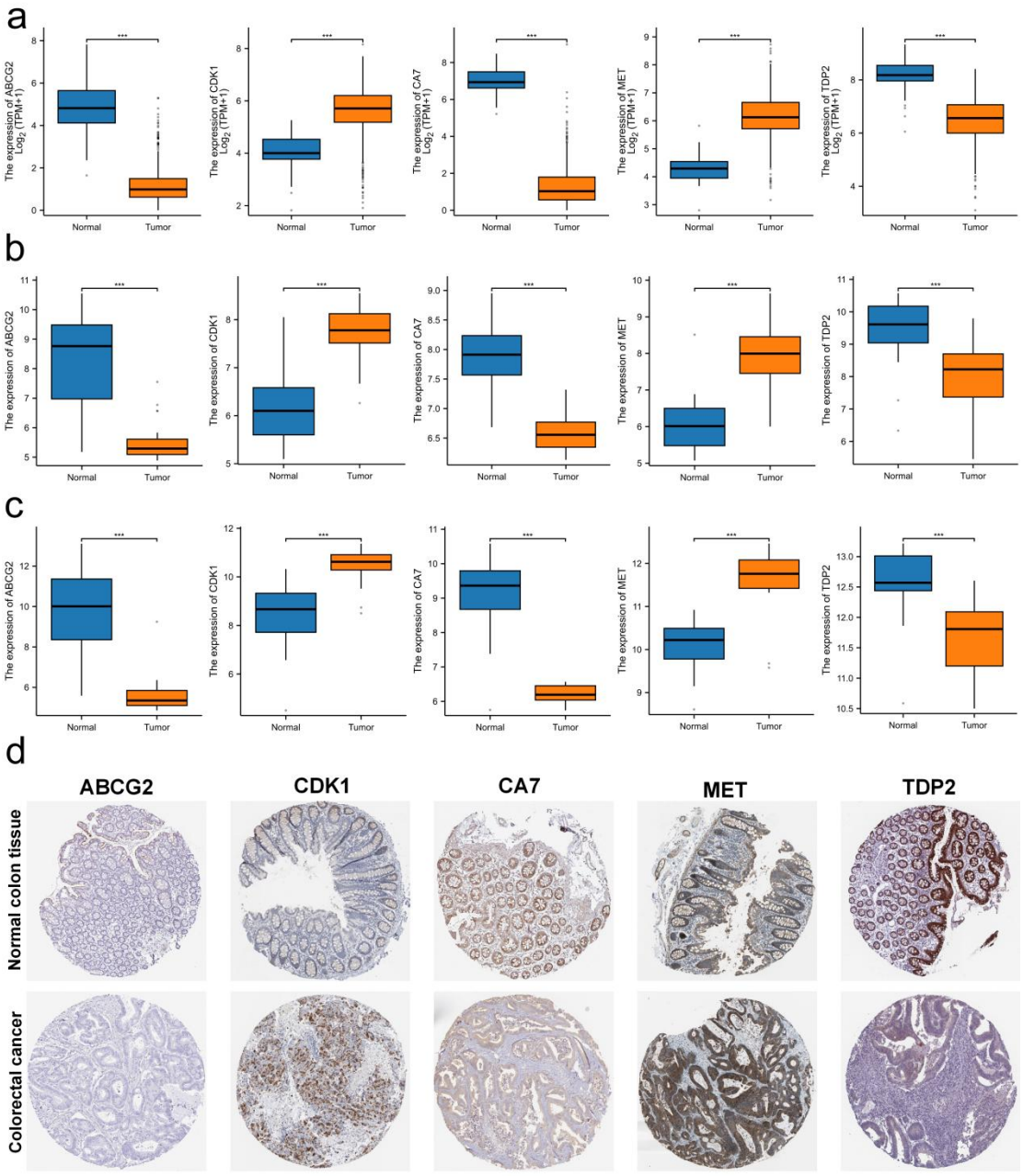


Fig. 5 Expression validation of the five core genes at both transcriptomic and protein levels. (a–c) Box plots showing the mRNA expression levels of *ABCG2*, *CDK1*, *CA7*, *MET*, and *TDP2* in CRC tissues versus normal colorectal controls from the TCGA-CRC cohort (a), GSE24514 (b), and GSE32323 (c) datasets. Statistically significant differences between tumour and normal groups are annotated; *p < 0.05, **p < 0.01, ***p < 0.001. (d) Immunohistochemistry (IHC) representative images of five core target proteins in normal colon epithelium and colorectal cancer tissues obtained from the HPA database. Protein-level expression alterations are consistent with transcriptomic dysregulation observed in bulk RNA-seq cohorts

3.6. Single-Cell Expression Landscape of Core Targets

To further dissect the cell-type-specific expression patterns of the five core targets identified by machine learning, we interrogated two independent CRC single-cell transcriptomic datasets, GSE166555 and EMTAB8107. The averaged gene expression levels across immune cells, malignant epithelial cells, and stromal cell compartments were first quantified and visualised in heatmaps (Fig. 6a-b). In the GSE166555 cohort, *TDP2* exhibited the highest mean expression in stromal cells (log(TPM/10+1) = 2.12), followed by malignant cells (1.11), while *ABCG2* and *CA7* showed preferential expression within stromal compartments. *CDK1* and *MET* were predominantly enriched in malignant cells. Consistently, the EMTAB8107 dataset also revealed that *TDP2* had markedly elevated expression in both malignant and stromal cell populations, with the highest value observed in stromal cells (0.46). *CDK1* maintained prominent expression in malignant cells across both datasets, supporting its malignant cell-specific signature.

Subsequently, UMAP dimensional reduction plots were generated to visualise the spatial distribution of core gene expression at single-cell resolution (Fig. 6c–f). The cell lineage annotation UMAPs of EMTAB8107 (Fig. 6c) and GSE166555 (Fig. 6e) defined major cell populations, including malignant epithelial cells, diverse immune subsets, fibroblasts and endothelial stromal cells. The corresponding cell compartment classification UMAPs further grouped cells into immune, malignant and stromal categories (Fig. 6d, 6f). Gene expression overlay on UMAP demonstrated that *CDK1* and *TDP2* displayed broad and strong expression signals within malignant cell clusters in both datasets, consistent with the compartment-level averaged expression results. *CA7* was mainly detected in subsets of malignant epithelial cells, with limited expression in immune and stromal compartments. *MET* showed moderate expression localized to malignant cell populations in both CRC cohorts. *ABCG2* exhibited relatively low overall expression, with sparse positive signals scattered across stromal and malignant cell subpopulations.

Collectively, single-cell profiling revealed that the five core targets exhibited distinct cell-type expression tropism in CRC. *CDK1* and *TDP2* are highly expressed in malignant cells, whereas *ABCG2* and *CA7* are preferentially enriched in stromal compartments. *MET* is moderately expressed in malignant epithelial cells. These cell-type-resolved expression profiles provide cellular-level evidence supporting the involvement of these core genes in the initiation and progression of colorectal cancer, and further rationalize their association with the oncogenic effects induced by alternative plasticizer exposure.

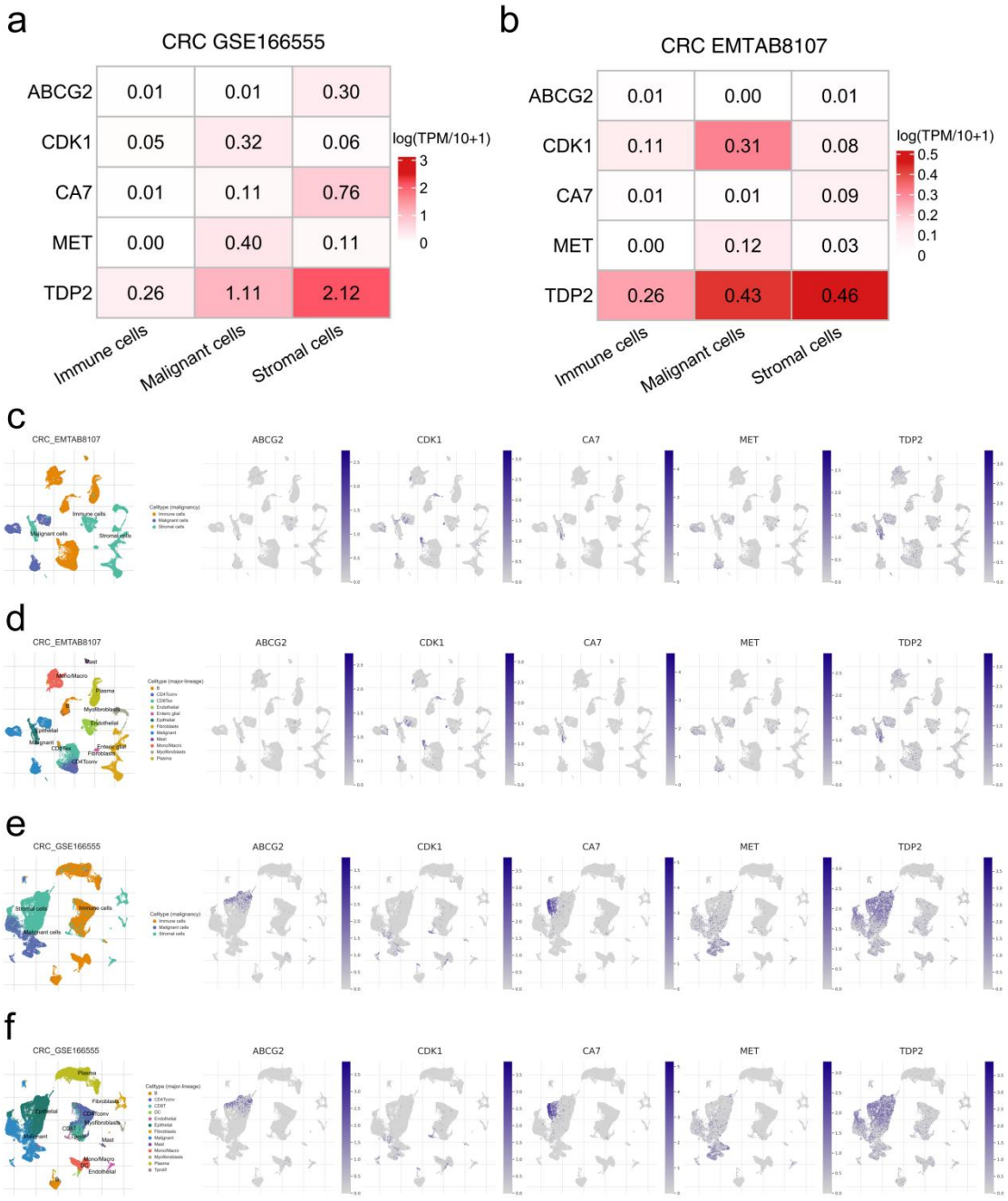


Fig. 6 Single-cell expression landscape of five core targets in two independent CRC single-cell datasets. (a, b) Heatmaps showing the average log(TPM/10+1) expression values of *ABCG2*, *CDK1*, *CA7*, *MET* and *TDP2* across immune cells, malignant cells and stromal cells in GSE166555 (a) and EMTAB8107 (b). (c, d) UMAP plots of the EMTAB8107 dataset. (c) UMAP coloured by annotated major cell lineages; overlaid UMAPs show the expression distribution of each core gene. (d) UMAP coloured by three major cell compartments (immune, malignant, stromal cells), with corresponding gene expression overlay plots. (e, f) UMAP plots of the GSE166555 dataset. (e) UMAP coloured by annotated major cell lineages; overlaid UMAPs display gene expression patterns. (f) UMAP coloured by immune, malignant and stromal compartments, with gene expression overlay for each core target. colour intensity represents normalised gene expression level

3.7. Immune Infiltration Analysis

To characterise the immune microenvironment associated with the five core targets, CIBERSORT was used to estimate the abundance of 22 immune cell types in the TCGA-CRC cohort, and Spearman correlations between core-gene expression and each cell population were calculated.

The five genes showed partly opposing correlation patterns. *ABCG2* expression was positively correlated with resting mast cells ($R = 0.246$, $p < 0.001$) and M2 macrophages ($R = 0.241$, $p < 0.001$), and negatively correlated with activated mast cells ($R = -0.200$, $p < 0.001$) and memory B cells ($R = -0.122$, $p < 0.001$) (Fig. 7a). *CDK1* displayed the strongest positive associations with activated CD4 memory T cells ($R = 0.308$, $p < 0.001$), M1 macrophages ($R = 0.167$, $p < 0.001$) and eosinophils ($R = 0.155$, $p < 0.001$), and showed the strongest negative association of the entire analysis with regulatory T cells ($R = -0.412$, $p < 0.001$), followed by M0 macrophages ($R = -0.163$, $p < 0.001$) (Fig. 7b). *CA7* was positively correlated with plasma cells ($R = 0.164$, $p < 0.001$) and regulatory T cells ($R = 0.160$, $p < 0.001$), and negatively correlated with M0 macrophages ($R = -0.169$, $p < 0.001$) (Fig. 7c). *MET* correlated positively with resting CD4 memory T cells ($R = 0.140$, $p < 0.001$) and activated dendritic cells ($R = 0.110$, $p < 0.01$), and negatively with regulatory T cells ($R = -0.189$, $p < 0.001$) (Fig. 7d). *TDP2* was positively correlated with resting CD4 memory T cells ($R = 0.257$, $p < 0.001$) and eosinophils ($R = 0.167$, $p < 0.001$), and negatively correlated with regulatory T cells ($R = -0.202$, $p < 0.001$) (Fig. 7e).

The correlation heatmap summarising all five genes against the 22 cell types confirmed these patterns and made the divergence between genes explicit (Fig. 7f). In particular, regulatory T cells were negatively correlated with *CDK1*, *MET* and *TDP2* but positively correlated with *CA7*, and resting CD4 memory T cells were positively correlated with both *MET* and *TDP2*. The association with activated CD4 memory T cells was positive for *CDK1* but negative for *CA7*, indicating that the core targets do not act on the immune microenvironment in a uniform direction.

The ESTIMATE analysis of global immune-stromal features is presented in the supplementary material. Collectively, these results indicate that *ABCG2*, *CDK1*, *CA7*, *MET* and *TDP2* are associated with remodelling of the colorectal tumour immune microenvironment, involving T cell subsets, macrophages, mast cells and dendritic cells, and may therefore provide a mechanistic link between alternative-plasticizer exposure and CRC progression(Supplementary Fig. S5).

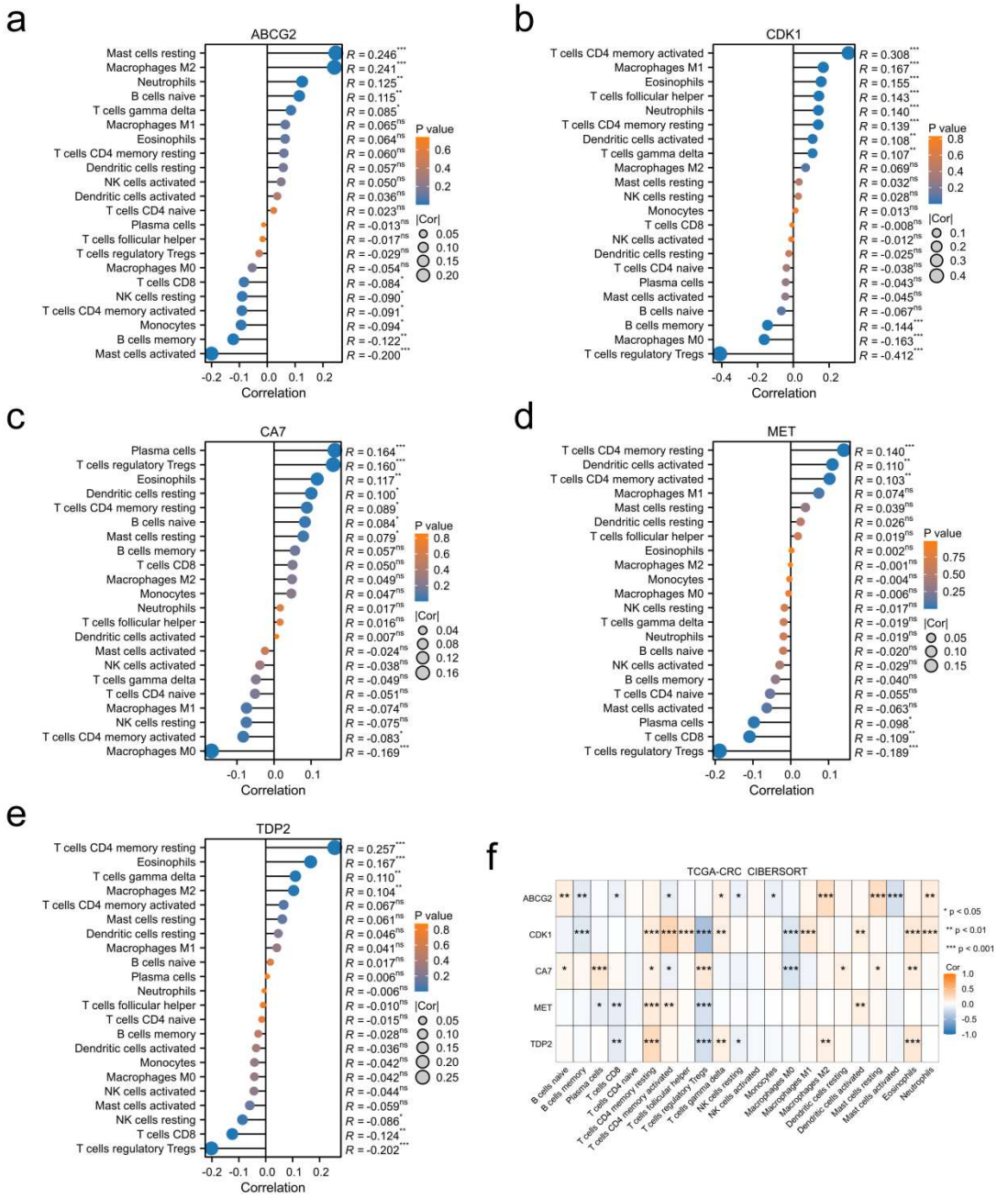


Fig. 7 Immune infiltration analysis of five core targets in TCGA-CRC cohort. (a-e) Lollipop plots showing the Spearman correlation between the expression of *ABCG2* (a), *CDK1* (b), *CA7* (c), *MET* (d) and *TDP2* (e) and the abundance of 22 immune cell types estimated by CIBERSORT. Dot colour encodes the p value (blue, low; orange, high) and dot size the absolute correlation coefficient; the correlation coefficient and its significance are given to the right of each row (ns, not significant; *p < 0.05, **p < 0.01, ***p < 0.001). Cell types are ordered by correlation within each panel. (f) Heatmap summarising the Spearman correlations between the five core genes and all 22 immune cell populations in TCGA-CRC cohort; colour denotes the correlation coefficient and asterisks indicate significance (*p < 0.05, **p < 0.01, ***p < 0.001)

3.8. Diagnostic and Prognostic Assessment

To further characterise the clinical relevance of the five machine-learning-identified core targets, we performed diagnostic ROC analysis and Kaplan-Meier prognostic survival evaluation. ROC curves were generated to quantify the capacity of each core gene to discriminate CRC tumour tissues from normal colorectal mucosa (Fig. 8a-e). All five genes achieved outstanding diagnostic discriminatory performance, with AUC values as follows: *ABCG2* (AUC = 0.985, 95% CI: 0.975–0.995), *CDK1* (AUC = 0.919, 95% CI: 0.896–0.942), *CA7* (AUC = 0.998, 95% CI: 0.995–1.000), *MET* (AUC = 0.971, 95% CI: 0.955–0.986), and *TDP2* (AUC = 0.946, 95% CI: 0.909–0.983).

Subsequently, Kaplan-Meier overall-survival analysis retrieved from the KMplot database was applied to evaluate the prognostic implication of core-gene expression in CRC patients (Fig. 8f-j). Elevated expression of *ABCG2* (HR = 1.45, 95% CI: 1.14–1.83, log-rank p = 0.0023) and *CA7* (HR = 1.62, 95% CI: 1.23–2.13, log-rank p = 0.00054) corresponded to significantly unfavourable overall survival. By contrast, high transcript abundance of *CDK1* (HR = 0.59, 95% CI: 0.48–0.73, log-rank p = 1.1×10^{-6}), *MET* (HR = 0.71, 95% CI: 0.55–0.91, log-rank p = 0.0064), and *TDP2* (HR = 0.62, 95% CI: 0.50–0.76, log-rank p = 6.4×10^{-6}) was associated with improved patient survival outcomes.

Notably, we observed an apparent paradox: although *CDK1* and *MET* are well-recognised oncogenic drivers in CRC, their high expression correlated with better overall survival in this cohort. Collectively, these results demonstrated that the five core genes possess excellent diagnostic potential for CRC, while exhibiting divergent prognostic associations, which warrants further mechanistic interpretation in the discussion section.

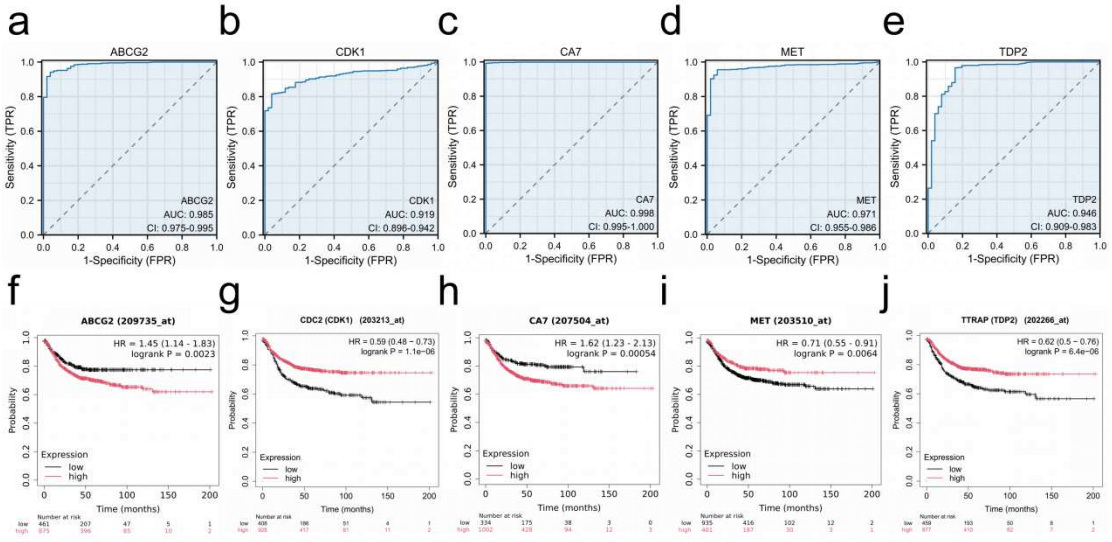


Fig. 8 Diagnostic and prognostic assessment of the five core genes in CRC. (a-e) Receiver operating characteristic (ROC) curves of *ABCG2* (a), *CDK1* (b), *CA7* (c), *MET* (d), and *TDP2* (e) for discriminating CRC from normal colorectal tissues, with

corresponding area under the curve (AUC) and 95% confidence intervals indicated. (f-j) Kaplan-Meier overall-survival curves stratified by high- versus low-gene expression of *ABCG2* (f), *CDK1* (g), *CA7* (h), *MET* (i), and *TDP2* (j) derived from the KMplot database. Hazard ratios (HR), 95% confidence intervals and log-rank p-values are annotated

3.9. Molecular Docking Analysis of Plasticizer–Target Interactions

To provide structural computational evidence supporting potential physical interactions between three alternative plasticizers and the five core target proteins identified in the above multi-omics analyses, molecular docking simulations were performed. The binding-energy matrix across all compound-protein pairs is summarised in the heatmap (Fig. 9a). Collectively, DOTP and DINCH yielded considerably stronger binding affinities toward most core proteins compared with TMCH. Among all pairs, the most favourable binding energy was observed for the DOTP-ABCG2 complex (−8.221 kcal/mol), followed by the DINCH-CDK1 complex (−7.608 kcal/mol). By comparison, TMCH exhibited relatively weak binding against all five core targets, with binding energies ranging only from −4.168 to −4.877 kcal/mol.

According to the magnitude of binding affinity, DOTP-ABCG2 and DINCH-CDK1 were chosen for detailed visual inspection of 2D and 3D ligand-protein interaction patterns. The TMCH-CA7 complex was retained as a representative weak-binding reference (Fig. 9b-d). In the DOTP-ABCG2 complex (−8.221 kcal/mol), DOTP was stably embedded inside the protein binding pocket; this interaction was predominantly stabilized by extensive hydrophobic contacts and π-stacking interactions (Fig. 9b). For the DINCH-CDK1 complex (−7.608 kcal/mol), the binding interface was mainly governed by hydrophobic forces, supplemented by π-alkyl contacts (Fig. 9c). On the contrary, the TMCH-CA7 complex (−4.877 kcal/mol) displayed a distinct interaction signature dominated by hydrogen bonds, accompanied by limited hydrophobic contacts, which is consistent with its comparatively weak binding strength (Fig. 9d).

It should be emphasised that docking outputs represent purely *in silico* structural predictions of putative binding propensity and cannot serve as direct proof of intracellular physical binding, nor can docking-derived binding energies infer functional activation or inhibition of target proteins. These simulation results provide structural hypotheses that DOTP and DINCH may exhibit binding propensity toward core target proteins mainly via hydrophobic and π-mediated interactions. Of note, although TMCH obtained the highest carcinogenic probability in earlier carcinogenicity prediction, it showed low binding affinity to these five core proteins, implying its potential oncogenic risk may be mediated through molecular routes other than high-affinity binding to this set of core targets (this remains a computational conjecture requiring experimental verification). Taken together, molecular docking furnishes testable structural hypotheses linking alternative-plasticizer exposure to CRC for subsequent wet-lab experimental validation.

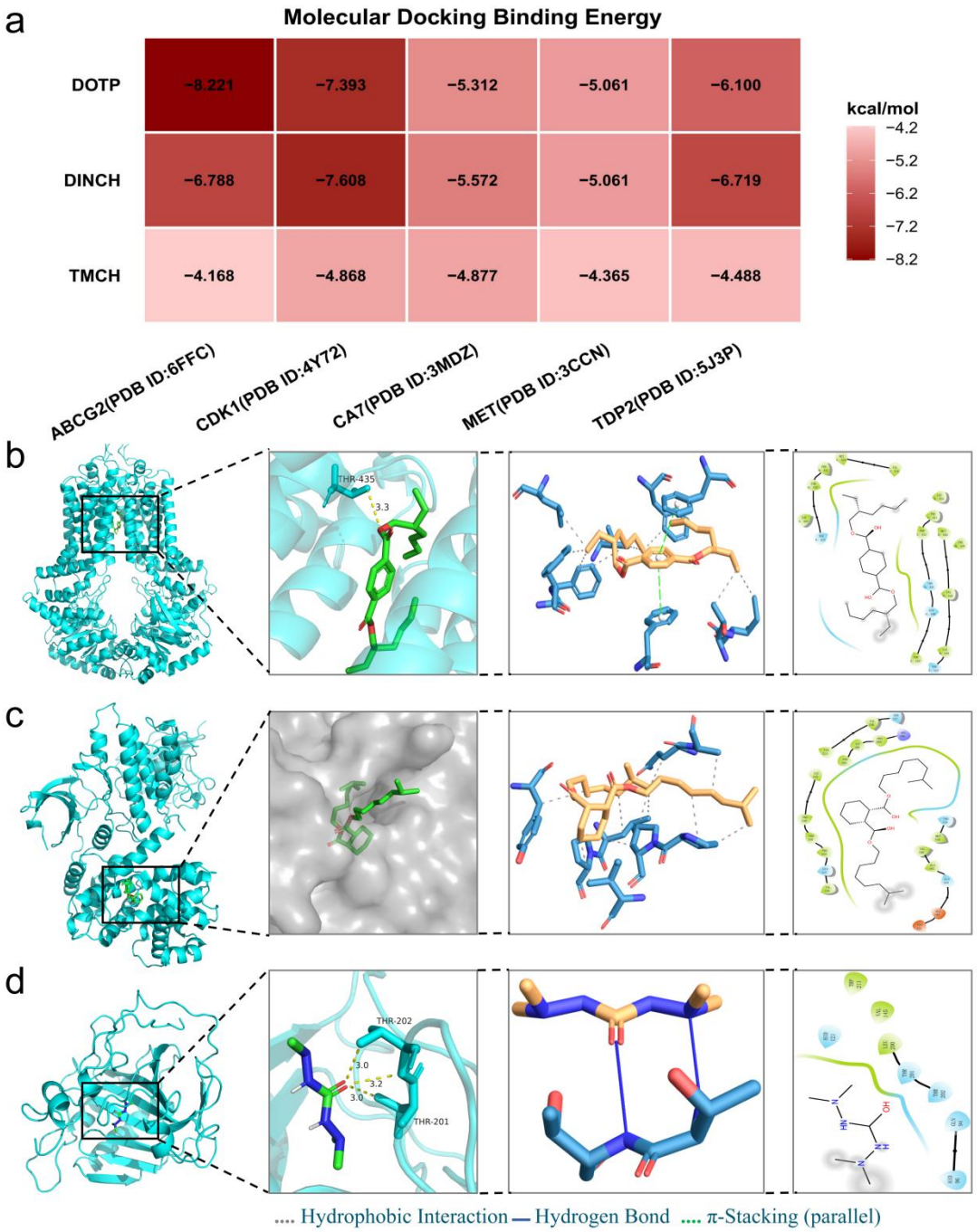


Fig. 9 Molecular docking analysis of plasticizer-target interactions. (a) Heatmap summarising binding energies (kcal/mol) for all pairwise combinations of three plasticizers (DOTP, DINCH, TMCH) and five core target proteins (*ABCG2*, *CDK1*, *CA7*, *MET*, *TDP2*); lower values correspond to stronger binding affinity. (b-d) Representative 3D and 2D interaction diagrams for selected complexes: (b) DOTP-ABCG2 (− 8.221 kcal/mol), (c) DINCH-CDK1 (− 7.608 kcal/mol), (d) TMCH-CA7 (−4.877 kcal/mol, weak-binding control). Hydrogen bonds, hydrophobic contacts and π-system-mediated interactions are annotated accordingly

3.10 Molecular Dynamics Simulation Results

To further validate the dynamic stability of the DOTP‑ABCG2 complex obtained from molecular docking, a 100 ns all-atom molecular dynamics simulation was performed. The RMSD profiles (Fig. 10a) showed that the backbone of *ABCG2* and the DOTP ligand underwent rapid conformational adjustment within the initial 20 ns. After 20 ns, the RMSD values of the protein and ligand fluctuated within a narrow range, indicating that the whole system reached a stable equilibrium state. No obvious dissociation of DOTP from the ABCG2 binding pocket was observed throughout the simulation.

The RMSF analysis (Fig. 10b) revealed the residue flexibility of chain A and chain B of *ABCG2*. Most residues within the binding pocket exhibited relatively low fluctuation values, while large fluctuations mainly appeared in flexible loop and terminal regions, demonstrating that the binding of DOTP restricted the local motion of pocket residues. The buried SASA (Fig. 10c) and Rg (Fig. 10d) remained relatively stable after equilibration, which suggested that the DOTP‑ABCG2 complex maintained a compact conformation without large-scale protein unfolding during simulation.

The time-dependent binding energy components (Fig. 10e) demonstrated that van der Waals interactions dominated the protein‑ligand binding, while electrostatic interactions contributed weakly to the total binding energy. The FEL (Fig. 10f) was constructed using RMSD and Rg as reaction coordinates. The FEL map indicated that the complex was trapped in a low-free-energy basin, indicating a dominant stable conformation. The structural snapshots at different simulation time points (Fig. 10g) showed that DOTP persistently occupied the ABCG2 binding cavity during the entire 100 ns simulation.

Collectively, MD simulation results computationally supported that the predicted DOTP-ABCG2 complex could maintain dynamic structural stability under simulated physiological aqueous conditions, providing dynamic structural hypotheses to complement static molecular-docking outputs.

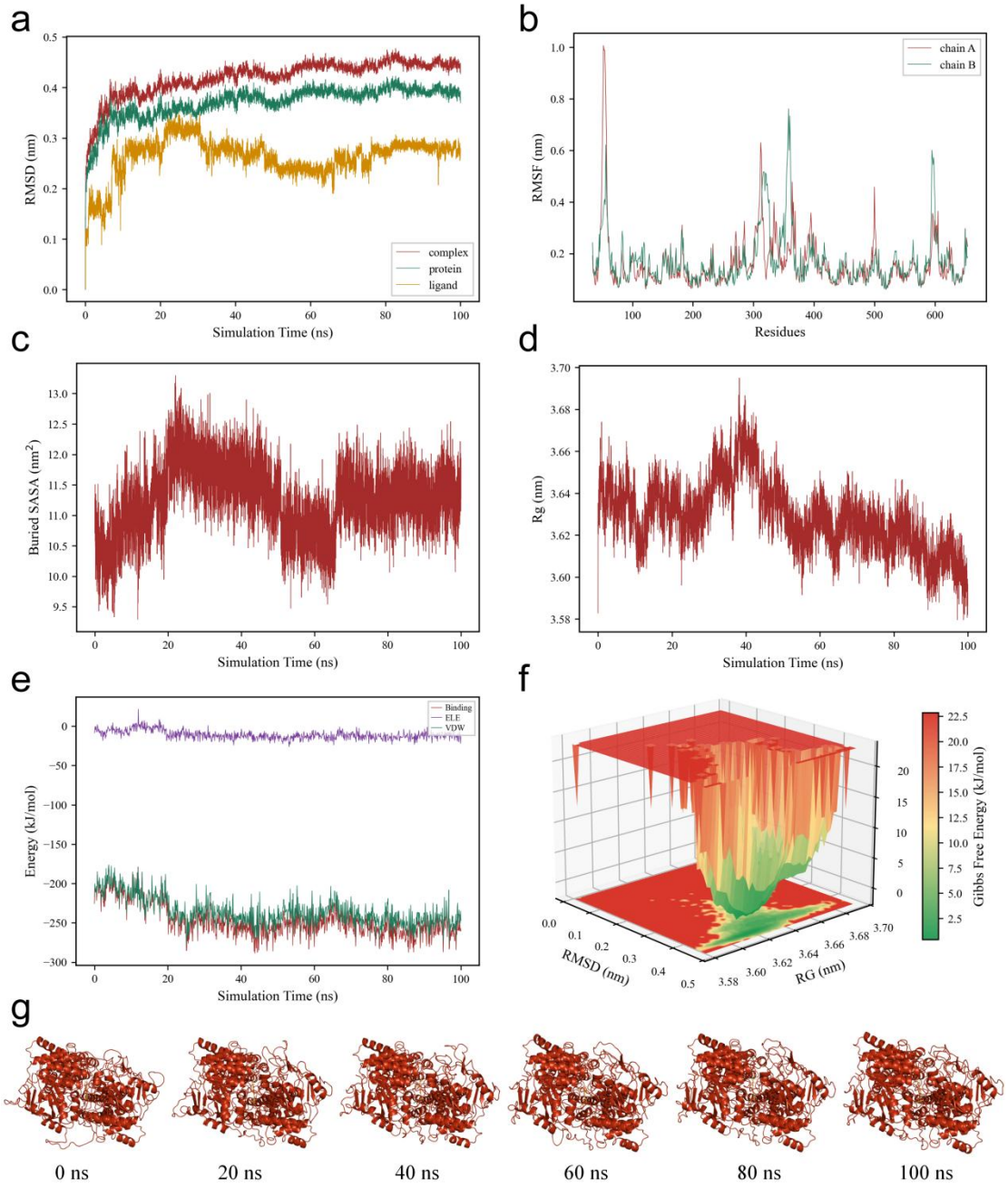


Fig. 10 Molecular dynamics simulation analyses of the DOTP-ABCG2 complex. (a) RMSD curves of the complex, ABCG2 protein backbone and DOTP ligand over the 100 ns simulation. (b) RMSF of per-residue Cα atoms for chain A and chain B of *ABCG2*. (c) Time evolution of buried SASA of the DOTP-ABCG2 complex. (d) Radius of gyration (Rg) of the complex. (e) Variation of total binding energy, electrostatic (ELE) and van der Waals (VDW) energy components during simulation. (f) Free energy landscape (FEL) constructed with RMSD and Rg as collective variables. The colour bar represents Gibbs free energy (kJ/mol). (g) Structural snapshots of the DOTP–ABCG2 complex at 0 ns, 20 ns, 40 ns, 60 ns, 80 ns and 100 ns

4. Discussion

CRC remains one of the most formidable malignancies globally, standing as the second leading cause of cancer-related mortality and accounting for nearly 916 000 deaths annually[18]. The disease imposes a considerable burden on patients and healthcare systems, characterised by high treatment costs, significant morbidity, and diminished quality of life. Despite advances in endoscopic screening and multimodal therapy, a substantial proportion of patients are diagnosed at advanced stages with limited therapeutic options[19]. Therefore, identifying modifiable environmental risk factors and actionable molecular targets is essential for improving early risk stratification and preventive intervention for CRC.

Several previous studies have integrated network toxicology together with interpretable machine-learning workflows to unravel disease-relevant mechanisms triggered by environmental contaminants. Representative examples include the mechanistic exploration of bisphenol A-induced ischemic stroke and benzo[a]pyrene-associated periodontitis[20, 21]. Similar integrative workflows have also been applied to decipher environmental-contaminant-driven colorectal tumourigenesis[22]. Those reports mainly focused on legacy toxicants and single-compound exposure scenarios, and relied on relatively limited machine-learning pipelines for target screening. Compared with existing work, the present study possesses several distinct novelties. Methodologically, we integrated carcinogenicity prediction, compound-target mining, WGCNA, a large-scale screening of 118 machine-learning pipelines, SHAP interpretative modelling, multi-omics validation and molecular docking and molecular dynamics simulation, forming a comprehensive computational workflow for toxic-target hypothesis generation. Beyond static molecular docking predictions, molecular dynamics further supplied dynamic structural evidence for ligand-protein interaction, reducing the limitation of snapshot-based docking results. For scientific discovery, this is, to our knowledge, the first study systematically exploring colorectal carcinogenic risks posed by three widely-used alternative non-phthalate plasticizers (DOTP, DINCH, TMCH). Furthermore, we uncovered an intriguing paradox that TMCH exhibits the highest predicted carcinogenic hazard yet displays weak binding affinity against the five core protein targets identified herein. From a translational perspective, the five-gene signature achieved outstanding discriminatory performance for CRC, supporting its potential as a multi-gene diagnostic panel for further experimental validation.

Functional enrichment of the overlapping targets did not merely confirm their involvement in generic cancer-related categories; rather, it positioned these targets at the intersection of several CRC-driving programs. The enriched biological processes were dominated by proliferation, proteolysis, apoptotic regulation, and transcriptional control, whereas the molecular function landscape was characterised by nuclear receptor, peptidase, and protein kinase activities. This profile is mechanistically coherent with a model in which plasticizer exposure lowers the threshold for malignant transformation by simultaneously sustaining proliferative signalling and weakening apoptotic surveillance[23].

The enrichment of extracellular exosome and cell-surface components is particularly notable, as exosome-mediated intercellular communication is increasingly recognised

as a driver of CRC progression and microenvironment remodelling[24]. KEGG mapping further indicated that the core targets converge on cell-cycle control, p53 signalling, cellular senescence, PPAR signalling, and microRNA-related networks. These pathways are not independent; rather, they form an interconnected network in which cell-cycle deregulation, impaired senescence, metabolic reprogramming, and post-transcriptional regulation reinforce one another. Within this network, *CDK1* and *MET* may provide proliferative and receptor-kinase inputs, whereas *ABCG2*, *CA7*, and *TDP2* may contribute to detoxification, pH homeostasis, and genome maintenance, respectively. The coordinate dysregulation of these genes therefore suggests a multi-pronged perturbation through which alternative plasticizers could create a permissive microenvironment for colorectal carcinogenesis[25, 26]. However, because these inferences are derived from computational enrichment rather than functional experiments, they should be regarded as hypothesis-generating and require experimental validation.

Multi-cohort transcriptomic data together with HPA immunohistochemistry consistently demonstrated that *CDK1* and *MET* are up-regulated in CRC, whereas *ABCG2*, *CA7* and *TDP2* are down-regulated in tumour tissues. Single-cell profiling further resolved cell-type tropism: *CDK1* and *MET* are predominantly expressed in malignant epithelial cells; *ABCG2* and *CA7* are enriched in stromal compartments; *TDP2* is broadly expressed in both malignant and stromal populations. Beyond expression patterns, each core target carries biologically relevant functions closely connected to colorectal tumour biology. As a well-known xenobiotic efflux transporter (BCRP), reduced *ABCG2* expression in CRC may impair the clearance of exogenous compounds and additionally modulate chemotherapeutic sensitivity toward agents such as irinotecan, a first-line drug for colorectal cancer[27]. Our molecular dynamics simulation further illustrated that DOTP could stably occupy the ABCG2 binding pocket over 100 ns simulation, supporting the possibility that DOTP may interfere with the physiological function of *ABCG2* through direct physical binding. *CA7*, a carbonic anhydrase isoform, participates in intracellular and extracellular pH homeostasis; its down-regulation may facilitate the acidic extracellular microenvironment characteristic of the Warburg effect, which has been reported to promote tumour survival and invasion[28]. *TDP2* mediates DNA double-strand break repair, and its decreased expression could aggravate genomic instability[29], a phenotype frequently triggered by plasticizer-related oxidative stress. As key cell-cycle and oncogenic drivers, elevated *CDK1* and *MET* have been reported to promote malignant proliferation[30, 31], consistent with their enrichment within malignant epithelial cell clusters observed in single-cell datasets.

Immune-microenvironment analyses combining CIBERSORT and ESTIMATE further illustrated that these five genes modulate both specific immune-cell subsets and the global immune-stromal architecture of CRC. *CDK1* correlated positively with anti-tumour-associated immune populations and negatively with regulatory T-cell abundance, while *ABCG2* showed positive correlations with ImmuneScore and StromalScore. These multi-layer observations collectively suggest that plasticizer-linked molecular perturbations may drive CRC pathogenesis by altering intrinsic tumour-cell programmes as well as remodelling the surrounding tumour microenvironment.

Although all five core genes possessed strong diagnostic capacity for distinguishing CRC from normal mucosa, divergent survival associations were observed in KMplot-derived survival data. Notably, oncogenic drivers *CDK1* and *MET* showed favourable overall-survival correlations in this cohort despite their tumour-promoting roles. Several confounding factors may help reconcile this apparent paradox. First, survival outputs from KMplot are derived from median-dichotomized gene-expression values and lack adjustment for critical clinical covariates including tumour stage, microsatellite status and chemotherapy regimens[32, 33]. Second, high *CDK1* expression coincided with signatures of immune-activated tumour microenvironments with reduced Treg infiltration, implying these proliferative tumours may represent immunologically "hot" lesions more susceptible to anti-tumour immune surveillance[34]. Third, highly proliferative tumours driven by *CDK1*/*MET* may respond better to cytotoxic chemotherapy in the analysed cohorts[35]. We emphasise that these survival associations represent correlative computational outputs rather than causal conclusions. Even so, the robust diagnostic performance of this five-gene signature merits further translational exploration.

An important paradox emerged between carcinogenicity prediction and molecular-docking outputs: TMCH obtained the highest carcinogenic probability according to ADMETlab 3.0, yet exhibited weak binding affinities against all five core target proteins. It should be emphasised that this comparison is entirely derived from *in silico* tools, and does not reflect real-world toxicological data. This discrepancy highlights the distinct read-outs of different *in silico* toxicology tools. Carcinogenicity prediction generates structure-based hazard alerts, which may encompass indirect toxicity, metabolic activation and reactive-metabolite-mediated damage. In contrast, molecular docking simulates only static direct binding between parent compound and given protein crystal structures[36]. Molecular dynamics further complements docking by capturing dynamic conformational fluctuation of protein-ligand complexes, yet it is still limited to simulated aqueous conditions and static parent compound structures without considering metabolic transformation. Accordingly, the notion that TMCH-associated oncogenic risk may operate through molecular routes outside high-affinity interaction with this set of five core targets remains a computational conjecture. This observation reinforces that no single computational assay can fully characterise compound hazard; multi-modal computational evidence should be jointly interpreted for toxic-risk assessment.

Collectively, we propose an integrated computational hypothetical mechanistic model linking alternative-plasticizer exposure to colorectal carcinogenesis, which can be visualised in a dedicated schematic diagram. This mechanistic model is derived purely from multi-omics and structural computational outputs and has not been functionally validated by wet-lab experiments. Chronic intestinal exposure to DOTP, DINCH or TMCH is hypothesised to perturb the expression of five core mediators: up-regulation of oncogenic *CDK1* and *MET*, together with down-regulation of protective *ABCG2*, *CA7* and *TDP2*. These molecular disturbances are hypothesised to jointly drive cell-cycle dysregulation and apoptotic resistance within colonic epithelial cells, while simultaneously reshaping the immune-stromal tumour microenvironment. Structural-level evidence from docking and molecular dynamics supports that DOTP may directly interact with *ABCG2*, potentially disrupting its biological function. The

combined effects of malignant-cell-intrinsic defects and microenvironmental reprogramming ultimately facilitate the initiation and progression of colorectal cancer.

Compared with conventional network-toxicology pipelines that largely rely on PPI topology to filter hub genes[37], the present workflow leverages multi-omics data combined with large-scale machine-learning screening, independent external-cohort validation and SHAP interpretability. Such a multi-step analytical strategy objectively narrows hundreds of candidate targets down to a compact set of high-confidence core mediators and mitigates subjective manual-selection bias that frequently exists in conventional network-toxicology target filtering[38]. By adding molecular dynamics simulation, our workflow extends static docking results toward dynamic protein-ligand behaviour, and this modular analytical framework can be generalised to hazard-mechanism research for other under-studied environmental contaminants.

Several inherent limitations of this purely computational investigation need explicit acknowledgement. First, all results are computational predictions derived from public omics repositories, rather than direct experimental proof. Compound-target prediction databases may yield false-positive interactions; transcriptomic and PPI analyses capture correlation but not causality[39]. Second, public transcriptomic cohorts lack matched human biomonitoring data for internal DOTP/DINCH/TMCH exposure. Hence, the five-gene classifier distinguishes CRC versus normal tissue but cannot serve as a specific exposure biomarker for plasticizers. Third, ComBat batch correction cannot fully eliminate hidden confounders such as lifestyle, medication history and co-exposure to other environmental chemicals. Fourth, molecular docking only simulates static ligand-protein interactions and does not recapitulate compound metabolism, post-translational protein modification or subcellular compartmentalisation. Fifth, the molecular dynamics simulation was only performed for the DOTP-ABCG2 complex under ideal simulated physiological conditions; it cannot recapitulate intracellular metabolic modification or real-world complex cellular environment, and simulations for other compound-target pairs are absent in current study. Sixth, carcinogenicity assessment is entirely in silico, without supporting animal-experimental carcinogenicity assays. Seventh, and most relevant to the interpretation of the machine-learning results, tumour-versus-normal classification of public colorectal transcriptomic cohorts yields intrinsically high separation, because malignant and normal mucosa differ in thousands of transcripts; near-ceiling AUC values therefore overstate the specificity of any individual feature set or algorithm. The null-control analyses reported in section 3.4 (Supplementary Table S4) showed that randomly drawn 79-gene panels achieved comparable mean AUC values and that the first principal component alone separated the two classes, while the merged training cohort carries batch-class confounding such that dataset-of-origin information alone partially predicts the class label. High performance metrics should therefore be read as evidence that the prioritised genes carry genuine tumour-versus-normal signal, rather than as evidence of algorithmic superiority; this is why independent external cohorts, parsimony and cross-pipeline stability, rather than nominal AUC, were used to select the reported model.

To translate these computational findings, we propose a three-tier validation strategy: (i) *in vitro* low-dose exposure of CRC cell lines to assess core-gene expression and phenotypic changes; (ii) *in vivo* chronic exposure models to evaluate colorectal

pathological alterations; (iii) human cohort studies integrating plasticizer biomonitoring with multi-omics profiling.

From a public-health perspective, our computational evidence raises a hypothetical concern that alternative non-phthalate plasticizers are not necessarily biologically inert despite being marketed as safer substitutes for legacy phthalates[40, 41]. Given that the gastrointestinal tract constitutes a major human exposure portal, chronic low-dose intestinal exposure to DOTP, DINCH and TMCH deserves continued toxicological risk appraisal. Future work should prioritise the three-tier experimental roadmap outlined above, and additionally consider environmentally-relevant mixture exposure scenarios that better mirror real-world human contact with multiple plasticizer contaminants.

In conclusion, this integrative computational study systematically generated in silico hypotheses probing colorectal-carcinogenic hazards of three alternative non-phthalate plasticizers (DOTP, DINCH, TMCH). By combining network toxicology, WGCNA, large-scale machine-learning screening with SHAP interpretation, multi-omics validation and molecular docking and molecular dynamics simulation, we identified five core candidate targets (*ABCG2*, *CDK1*, *CA7*, *MET*, *TDP2*) hypothesised to potentially mediate plasticizer-associated CRC development. This work provides multi-level testable mechanistic hypotheses, as well as a candidate multi-gene diagnostic signature, laying a theoretical foundation for subsequent experimental toxicology and translational research.

Authors' contributions

Wenzhu Lou: Writing-original draft, Visualisation, Software, Methodology, Investigation, Formal analysis, Data curation.
Wei Ren: Data curation, Visualisation, Software, Investigation.
Bangsheng Chen: Conceptualization, Funding acquisition, Methodology, Project administration, Writing-review & editing.

Funding

This work was supported by the Zhejiang Provincial Health Commission Science and Technology Program (Grant No. 2025HY1074).

Competing interests

The authors have no conflict of interest.

Ethics approval and consent to participate

This study re-analysed publicly available, de-identified transcriptomic datasets deposited in GEO, TCGA and ArrayExpress. No new human samples were collected for this study. Therefore, ethics approval and consent to participate are not applicable.

Consent for publication

Not applicable.

Clinical trial number

Not applicable.

Data availability

Raw transcriptomic datasets analysed in this study are publicly available from the NCBI Gene Expression Omnibus (GSE21510, GSE9348, GSE24514, GSE32323, GSE44076, GSE113513, GSE166555), from ArrayExpress (EMTAB-8107), and from The Cancer Genome Atlas colorectal cohort (TCGA-COADREAD, accessed via the UCSC Xena hub). Source code and intermediate processed files for all bioinformatic analyses are available at GitHub (https://github.com/louwenzhu/colorectal-cancer-related-targets-of-three-novel-non-phthalate-plasticizers). No new raw sequencing data were generated in this study.

Acknowledgements

We would like to thank the researchers of the GSE21510, GSE9348, GSE24514, GSE32323, GSE44076, and GSE113513 datasets, and sample data from TCGA for their contribution, as well as all other individuals who contributed to this study.

References

1. Islam MS, Gopalan V, Lam AK, Shiddiky MJA. Current advances in detecting genetic and epigenetic biomarkers of colorectal cancer. Biosens Bioelectron. 2023;239:115611. https://doi.org/10.1016/j.bios.2023.115611

2. Carbone F, Spinelli A, Ciardiello D, Realis Luc M, de Pascale S, Bertani E, et al. Prognosis of early-onset versus late-onset sporadic colorectal cancer: systematic review and meta-analysis. Eur J Cancer. 2025;215:115172. https://doi.org/10.1016/j.ejca.2024.115172

3. Xu J, Qu J, Jin H, Mao W. Associations between microplastics in human feces and colorectal cancer risk. J Hazard Mater. 2025;495:139099. https://doi.org/10.1016/j.jhazmat.2025.139099

4. Shinohara N, Oguri T, Takagi M, Ueyama J, Isobe T. Evaluating the risk of phthalate and non-phthalate plasticizers in dust samples from 100 Japanese houses. Environ Int. 2024;183:108399. https://doi.org/10.1016/j.envint.2023.108399

5. Panneel L, Cleys P, Poma G, Ait Bamai Y, Jorens PG, Covaci A, et al. Ongoing exposure to endocrine disrupting phthalates and alternative plasticizers in neonatal intensive care unit patients. Environ Int. 2024;186:108605. https://doi.org/10.1016/j.envint.2024.108605

6. Mohamed DFMS, Tarafdar A, Lee SY, Oh HB, Kwon JH. Assessment of biodegradation and toxicity of alternative plasticizer di(2-ethylhexyl) terephthalate: impacts on microbial biofilms, metabolism, and reactive oxygen species-mediated stress response. Environ Pollut. 2024;355:124217. https://doi.org/10.1016/j.envpol.2024.124217

7. Ketema RM, Kasper-Sonnenberg M, Ait Bamai Y, Miyashita C, Koch HM, Pälmke C, et al. Exposure trends to the non-phthalate plasticizers DEHTP, DINCH, and DEHA in children from 2012 to 2017: the Hokkaido study. Environ Sci Technol. 2023;57(32):11926–36. https://doi.org/10.1021/acs.est.3c03172

8. Íñigo-Catalina L, Linillos-Pradillo B, Schlumpf M, Lichtensteiger W, Paredes SD, Rancan L, et al. DINCH exposure triggers inflammatory, oxidative, and apoptotic pathways in the liver of Long-Evans lactating rats and their offspring. Int J Mol Sci. 2024;25(23):13017. https://doi.org/10.3390/ijms252313017

9. Li X, Li M, Jiang N, Yao X, Wang Q, Lv H, et al. Evaluation of soil ecological health after exposure to environmentally relevant doses of di (2-ethylhexyl) phthalate: insights from toxicological studies of earthworms at different ecological niches. Environ Pollut. 2023;322:121204. https://doi.org/10.1016/j.envpol.2023.121204

10. Manankandayalage CP, Unruh DK, Perry R, Krempner C. 1,8-Dihydroxy naphthalene—a new building block for the self-assembly with boronic acids and 4,4′-bipyridine to stable host–guest complexes with aromatic hydrocarbons. Molecules. 2023;28(14):5394. https://doi.org/10.3390/molecules28145394

11. Liu H, Di B, Men X, Zhu J, Liu Y, Luo J, et al. An integrated network toxicology and multi-omics study identifies ENO1 as a candidate mediator in benzo[a]pyrene-related gastric cancer progression. Front Pharmacol. 2026;17:1857843. https://doi.org/10.3389/fphar.2026.1857843

12. Li R, Ren C, He L, Wang S, Guo G. Machine learning-integrated network toxicology uncovers glioma targets of DEHP. Front Toxicol. 2026;8:1771011. https://doi.org/10.3389/ftox.2026.1771011

13. Lili C, Zhongfu T, Ming L, Huang C. Network toxicology integrated with machine learning and SHAP analysis identifies overlapping immune signatures between di(2-ethylhexyl) phthalate (DEHP) and Sjögren's syndrome. BMC Pharmacol Toxicol. 2026;27(1):65. https://doi.org/10.1186/s40360-026-01119-x

14. Schake P, Bolz SN, Linnemann K, Schroeder M. PLIP 2025: introducing protein–protein interactions to the protein–ligand interaction profiler. Nucleic Acids Res. 2025;53(W1):W463–5. https://doi.org/10.1093/nar/gkaf361

15. Van Der Spoel D, Lindahl E, Hess B, Groenhof G, Mark AE, Berendsen HJC. GROMACS: fast, flexible, and free. J Comput Chem. 2005;26(16):1701–18. https://doi.org/10.1002/jcc.20291

16. Liu H, Tan Q, Han L, Huo S. Observations on AMBER force field performance under the conditions of low pH and high salt concentrations. J Phys Chem B. 2017;121(42):9838–47. https://doi.org/10.1021/acs.jpcb.7b07528

17. Kagami L, Wilter A, Diaz A, Vranken W. The ACPYPE web server for small-molecule MD topology generation. Bioinformatics. 2023;39(6):btad350. https://doi.org/10.1093/bioinformatics/btad350

18. Wang Y, Chen C, Xiao X, Kang Y, He J, Beloqui A, et al. Advanced oral therapies for colorectal cancer via nanomedicine and microbiota modulation. Adv Drug Deliv Rev. 2026;228:115741. https://doi.org/10.1016/j.addr.2025.115741

19. Fernández-Hernández S, Hidalgo-León MÁ, Lacalle-González C, Olivera-Salazar R, Ochieng' Otieno M, García-Foncillas J, et al. Dormancy in colorectal carcinoma: detection and therapeutic potential. Biomolecules. 2025;15(8):1119. https://doi.org/10.3390/biom15081119

20. Li Y, Chen X, Zhong P, Xing Y, Miao P. Exploring the pathophysiological relationship between bisphenol A exposure and ischemic stroke risk using network toxicology and machine learning. BMC Neurol. 2025;25(1):439. https://doi.org/10.1186/s12883-025-04460-6

21. Wen W, Li R, Zhuo P, Deng C, Zhang D. Integrated network toxicology, machine learning and molecular docking reveal the mechanism of benzopyrene-induced periodontitis. BMC Pharmacol Toxicol. 2025;26(1):118. https://doi.org/10.1186/s40360-025-00961-9

22. Su S, Liao Y, Wu B, Li Y, Lou F, Mei Z, et al. Novel environmental contaminant 6PPD-quinone promotes malignant phenotypes in colorectal cancer cells and identifies candidate response-associated genes. Chem Biol Interact. 2026;438:112307. https://doi.org/10.1016/j.cbi.2026.112307

23. Hu SW, Yeh MH, He YH, Wei YL, Cheng FJ, Yeh YL, et al. Co-targeting *CDK4*/6 enhances anti-cancer activity and alleviates immune-related adverse events of anti-PD-1 antibody for breast cancer. Cell Rep Med. 2025;6(11):102429. https://doi.org/10.1016/j.xcrm.2025.102429

24. Bhome R, Emaduddin M, James V, House LM, Thirdborough SM, Mellone M, et al. Epithelial to mesenchymal transition influences fibroblast phenotype in colorectal cancer by altering miR-200 levels in extracellular vesicles. J Extracell Vesicles. 2022;11(5):e12226. https://doi.org/10.1002/jev2.12226

25. Nastaranpour M, Suiwal S, Stachon T, Fries FN, Amini M, Seitz B, et al. miRNA expression profile in primary limbal epithelial cells of aniridia patients. Invest Ophthalmol Vis Sci. 2025;66(1):20. https://doi.org/10.1167/iovs.66.1.20

26. Ryu Y, Hwang JS, Bo Noh K, Park SH, Seo JH, Shin YJ. Adipose mesenchymal stem cell-derived exosomes promote the regeneration of corneal endothelium through ameliorating senescence. Invest Ophthalmol Vis Sci. 2023;64(13):29. https://doi.org/10.1167/iovs.64.13.29

27. de Groot RA, Reedijk D, Faucher Q, Mihăilă SM, Masereeuw R. Strategies for overcoming ABC transporter-mediated multidrug resistance in colorectal cancer. Am J Physiol Cell Physiol. 2025;329(3):C699–717. https://doi.org/10.1152/ajpcell.00412.2025

28. Garofalo FM, Terrazas KA, Giovannuzzi S, Bonifazi EL, Carta F, Supuran CT, et al. γ-aminobutyric methyl ester-derived sulfamides as carbonic anhydrase inhibitors. Bioorg Chem. 2025;164:108814. https://doi.org/10.1016/j.bioorg.2025.108814

29. Al Mahmud MR, Ishii K, Bernal-Lozano C, Delgado-Sainz I, Toi M, Akamatsu S, et al. *TDP2* suppresses genomic instability induced by androgens in the epithelial cells of prostate glands. Genes Cells. 2020;25(7):450–65. https://doi.org/10.1111/gtc.12770

30. Chen HD, Huang CS, Xu QC, Li F, Huang XT, Wang JQ, et al. Therapeutic targeting of CDK7 suppresses tumour progression in intrahepatic cholangiocarcinoma. Int J Biol Sci. 2020;16(7):1207–17. https://doi.org/10.7150/ijbs.39779

31. Li J, Tan G, Cai Y, Liu R, Xiong X, Gu B, et al. A novel apigenin derivative suppresses renal cell carcinoma via directly inhibiting wild-type and mutant *MET*. Biochem Pharmacol. 2021;190:114620. https://doi.org/10.1016/j.bcp.2021.114620

32. Reyes-González JM, Quiñones-Díaz BI, Santana Y, Báez-Vega PM, Soto D, Valiyeva F, et al. Downstream effectors of ILK in cisplatin-resistant ovarian cancer. Cancers (Basel). 2020;12(4):880. https://doi.org/10.3390/cancers12040880

33. Wang JJ, Siu MKY, Jiang YX, Leung THY, Chan DW, Wang HG, et al. A combination of glutaminase inhibitor 968 and PD-L1 blockade boosts the immune response against ovarian cancer. Biomolecules. 2021;11(12):1749. https://doi.org/10.3390/biom11121749

34. Li B, Wei P, Xie B, Zhong X, Wang X, Lou X, et al. Targeting tumour-intrinsic *CDK1*/cyclin B1 complex improves responses to immunotherapy in pancreatic cancer. Cancer Lett. 2026;648:218464. https://doi.org/10.1016/j.canlet.2026.218464

35. Castellanet O, Ahmad F, Vinik Y, Mills GB, Habermann B, Borg JP, et al. BCL-XL blockage in TNBC models confers vulnerability to inhibition of specific cell cycle regulators. Theranostics. 2021;11(19):9180–97. https://doi.org/10.7150/thno.60503

36. Mahmud S, Afrose S, Biswas S, Nagata A, Paul GK, Mita MA, et al. Plant-derived compounds effectively inhibit the main protease of SARS-CoV-2: an in silico approach. PLoS One. 2022;17(8):e0273341. https://doi.org/10.1371/journal.pone.0273341

37. Li H, Dong Y, Chen E, Li T. Mechanistic insights into N-nitrosamine compounds and food-based interventions: a network toxicology and molecular docking study. J Environ Sci Health B. 2025;60(12):577–91. https://doi.org/10.1080/03601234.2025.2608540

38. Yuan Y, Lin C, Zhang T, Song C, Sun Y, Tang H. Environmental PET-microplastic exposure and risk of non-alcoholic fatty liver disease: an integrated computational toxicology and multi-omics study. Naunyn Schmiedebergs Arch Pharmacol. 2026;399(8):11759–84. https://doi.org/10.1007/s00210-026-05115-0

39. Mombelli E, Pandard P. Evaluation of the OECD QSAR toolbox automatic workflow for the prediction of the acute toxicity of organic chemicals to fathead minnow. Regul Toxicol Pharmacol. 2021;122:104893. https://doi.org/10.1016/j.yrtph.2021.104893

40. Callaghan MA, Alatorre-Hinojosa S, Connors LT, Singh RD, Thompson JA. Plasticizers and cardiovascular health: role of adipose tissue dysfunction. Front Pharmacol. 2021;11:626448. https://doi.org/10.3389/fphar.2020.626448

41. Qadeer A, Anis M, Warner GR, Potts C, Giovanoulis G, Nasr S, et al. Global environmental and toxicological data of emerging plasticizers: current knowledge, regrettable substitution dilemma, green solution and future perspectives. Green Chem. 2024;26(10):5635–83. https://doi.org/10.1039/d3gc03428c

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryDataS1.xlsx
- Supplementarymaterials.docx