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.

Integrating machine learning and molecular docking analysis reveals the molecular network associated with tris(1,3-dichloro-2-propyl) phosphate in hepatocellular carcinoma

Yicheng Wan
ShunDe Hospital Guangzhou University of Chinese Medicine
Chuying Zhong
ShunDe Hospital Guangzhou University of Chinese Medicine
Mengqing Yang
ShunDe Hospital Guangzhou University of Chinese Medicine
Simin Liao
ShunDe Hospital Guangzhou University of Chinese Medicine
Dezhu Chen
chendezhu@gzucm.edu.cn
ShunDe Hospital Guangzhou University of Chinese Medicine

Research Article

Keywords: Hepatocellular carcinoma, tris(1,3-dichloro-2-propyl) phosphate, machine learning, network toxicology, molecular docking, bioinformatics

Posted Date: November 20th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11493238/v1

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

Additional Declarations: No competing interests reported.

Abstract

Hepatocellular carcinoma (HCC) is one of the most common primary liver malignancies worldwide and represents a major global health burden. Increasing evidence suggests that environmental pollutants may contribute to hepatocarcinogenesis. Tris(1,3-dichloro-2-propyl) phosphate (TDCPP) is a widely used organophosphate flame retardant. Current studies have clearly demonstrated that TDCPP possesses genotoxicity, hepatotoxicity, and carcinogenic potential. However, the molecular mechanisms linking TDCPP exposure to HCC remain unclear. The aim of this study was to explore the potential toxic targets and molecular mechanisms of TDCPP-associated HCC using an integrative bioinformatics approach. Transcriptomic datasets were obtained from the Gene Expression Omnibus (GEO) database. Differential expression analysis and weighted gene co-expression network analysis (WGCNA) were performed to identify HCC-related genes, while TDCPP targets were predicted using ChEMBL, SwissTargetPrediction, and PharmMapper. Intersection analysis, functional enrichment analysis, machine learning, SHapley Additive exPlanations (SHAP), and molecular docking were subsequently conducted. The results identified 230 potential TDCPP-related HCC targets. Enrichment analyses indicated involvement in pathways associated with cell cycle regulation, DNA replication, p53 signaling, FoxO signaling, and cellular senescence. Nine core genes (COL15A1, NDC80, LCAT, MOGAT2, BDH2, CRHBP, CHST4, NAT2, and CYP2C19) were identified with strong diagnostic performance (AUC > 0.85). Molecular docking suggested stable binding interactions between TDCPP and several core proteins. These findings provide crucial insights for further exploration of the molecular mechanisms of TDCPP-induced liver cancer.

Introduction

Hepatocellular carcinoma (HCC) is one of the most common primary malignant tumors of the liver. Its incidence ranks sixth among all cancers worldwide and it is also the third leading cause of cancer-related death, thus representing a major global public health challenge (1). According to statistics, approximately 860,000 new liver cancer cases and 750,000 related deaths occur globally each year (2). The occurrence of HCC is usually associated with multiple risk factors, including chronic hepatitis B or hepatitis C virus infection, long-term alcohol consumption, metabolic-associated fatty liver disease, and other chronic liver injury factors (3). In recent years, with the aggravation of environmental pollution, increasing attention has been paid to the potential role of environmental chemicals in the development and progression of liver cancer. Among them, tris(1,3-dichloro-2-propyl) phosphate (TDCPP) is an organophosphate flame retardant widely used in plastics, furniture, electronic products, and textile materials. Due to its widespread presence in the environment and bioaccumulation characteristics, it has attracted considerable attention (4). It has been listed as a “reasonably anticipated human carcinogen” in the *Report on Carcinogens* (15th edition, 2021) released by the National Toxicology Program (NTP) (5). Because of its associated health risks, including neurotoxicity and endocrine disruption (6), its application in children’s products has been restricted in the United States and Canada.

Recent studies indicate that TDCPP may influence cancer-related biological processes through multiple molecular mechanisms. In vitro studies have shown that TDCPP exposure can disrupt cellular energy

metabolism and xenobiotic metabolism pathways, and inhibit DNA replication and cell proliferation-related signaling, thereby disturbing cellular homeostasis (7). In zebrafish models, TDCPP exposure can induce hepatic inflammatory responses, apoptosis, and structural liver damage, suggesting clear hepatotoxic effects (8). Further studies using a *kras* transgenic zebrafish liver cancer model confirmed that TDCPP can promote liver tumor progression, and its mechanism is closely associated with enhanced inflammatory responses and activation of tumor-related signaling pathways (9). These findings collectively suggest that TDCPP may not merely act as a simple hepatotoxic exposure factor but may also promote malignant biological processes related to liver cancer cells through inflammation activation, metabolic imbalance, and remodeling of tumor-related signaling pathways.

The emergence of network toxicology provides a transformative research framework for elucidating complex toxin-host interactions through the integration of multi-omics data, dynamic network modeling, and computational structure-activity analysis (10). Previous studies have used cell viability assays, Comet assay DNA damage analysis, and transcriptomic analyses to verify the role of TDCPP in inducing hepatocellular carcinoma (11, 12). However, these studies have not explored the dynamic interactions among genes in depth. To address this limitation, the present study adopted a triangulated systems toxicology strategy. Multi-omics data were integrated through topological analysis and functional enrichment analysis to reconstruct TDCPP-driven HCC pathways, and machine learning methods were used to prioritize key hub molecules. Meanwhile, integrated molecular docking was performed to verify target binding and their interaction thermodynamics. By elucidating the multi-scale carcinogenic network of TDCPP, this study aims to identify druggable targets that may serve as candidates for HCC chemoprevention and precision therapy, thereby addressing critical unmet needs in high-risk populations.

Materials and Methods

Acquisition of HCC Target Molecules

In this study, five HCC transcriptomic datasets were retrieved from the GEO database, including GSE45627, GSE74656, GSE54236, GSE135631, and GSE76427. Specifically, GSE45627 and GSE74656 were designated as the discovery cohort, whereas GSE54236, GSE135631, and GSE76427 served as validation cohorts. To reduce batch effects, a multi-stage normalization workflow was implemented: Surrogate Variable Analysis (SVA): the SVA package was used to model and correct potential confounding factors in the discovery cohort. ComBat correction: residual batch differences were further corrected using a parametric empirical Bayes framework. PCA after correction showed improved clustering of samples from different batches in the reduced-dimensional space, indicating successful data integration. The complete analysis workflow is shown in Fig. 1.

Acquisition of TDCPP Chemical Information and Targets

TDCPP was characterized through multi-database integration. Its physicochemical properties and biological parameters were systematically retrieved from PubMed, whereas the standard two-dimensional structural descriptor (SMILES: C(C(CCl)OP(= O)(OC(CCl)CCl)OC(CCl)CCl)Cl) was obtained from the PubChem database. Target prediction was performed using three complementary strategies—ChEMBL for ligand-receptor interaction analysis, SwissTargetPrediction (http://www.swisstargetprediction.ch) for chemogenomics-based target prediction, and PharmMapper (http://lilab-ecust.cn/pharmmapper) for three-dimensional pharmacophore matching-based prediction. All predicted targets were restricted to the Homo sapiens proteome.

Differential Gene Expression Analysis

Transcriptomic data were analyzed using the limma package. Differentially expressed genes (DEGs) were identified based on the criteria of FDR-adjusted $P < 0.05$ and $|\log2FC| > 0.585$ (1.5-fold change). The results were visualized using ggplot2.

Weighted Gene Co-expression Network Analysis (WGCNA)

A scale-free co-expression network was constructed using the WGCNA package: Sample QC: outlier samples were removed through hierarchical clustering. Soft Thresholding: the optimal power value was determined using the dynamic tree cutting method (scale-free $R^2 > 0.85$). Module Detection: hierarchical clustering based on the topological overlap matrix (TOM) was performed with parameters minModuleSize = 30 and mergeCutHeight = 0.25. Module-Trait Association: correlations between module eigengenes and phenotypes were calculated (Pearson $|r| > 0.5$, $P < 0.05$). Hub Gene Identification: hub genes were identified based on intramodular connectivity (kME > 0.8).

Identification of TDCPP-associated Disease Targets

Intersection analysis between DEGs/WGCNA hub genes and predicted TDCPP targets was performed to identify core HCC targets, and the results were visualized using a Venn diagram.

Functional Enrichment Analysis

The clusterProfiler package was used to perform GO enrichment analysis (biological process, cellular component, molecular function) and KEGG pathway analysis ($P < 0.05$) to explore the potential mechanisms of TDCPP in HCC pathogenesis.

Core Gene Screening Based on Machine Learning

To systematically identify TDCPP-related HCC diagnostic biomarkers, we constructed an integrated machine learning prediction framework. Using expression profile data from the training cohort, ten

classical machine learning algorithms—Lasso, SVM, RF, glmBoost, Stepglm, Ridge, Enet, GBM, LDA, XGBoost, and NaiveBayes—were applied to construct 127 predictive models. Hyperparameter optimization was performed using 5-fold cross-validation, and stratified sampling was applied to divide the dataset into a training set and an internal validation set. Model performance was evaluated based on the area under the receiver operating characteristic curve (AUC), accuracy, and F1-score. Subsequently, a stacking ensemble learning strategy was used to integrate predictions from the optimal individual models. High-confidence models (AUC > 0.9) were selected, and feature genes were ranked according to their frequency of occurrence in the models to identify candidate core genes. Finally, gene expression patterns were visualized using the pheatmap package.

Model Interpretation

Given the inherent “black box” characteristics of machine learning models, the SHAP algorithm was used to quantify the contribution of each feature to the prediction results. This method assigns a SHAP value to each feature, enabling interpretable evaluation of the influence of individual features on model predictions.

Molecular Docking Analysis

To validate the interactions between TDCPP and the identified core genes, molecular docking simulations were performed. Ligand structures in SDF format and the three-dimensional models of core target proteins were obtained from PubChem and UniProt, respectively. Protein structures were preprocessed by removing water molecules and adding hydrogen atoms, whereas ligands were geometrically optimized using the MMFF force field. The docking grid was centered on predicted active sites and adjusted according to ligand size and predicted binding modes. All docking procedures were performed using AutoDock Vina, and the results were analyzed and visualized using Discovery Studio 2019.

Results

Identification of TDCPP Target Proteins

The molecular structure of TDCPP was obtained from the PubChem database (Fig. 2A). Three complementary databases were used to systematically predict the potential biological targets of TDCPP—ChEMBL (curated bioactive molecules), PharmMapper (reverse pharmacophore mapping), and SwissTargetPrediction (ligand-based target prediction). After data integration and removal of duplicate entries, a total of 5,474 unique potential targets were identified (Fig. 2B).

Figure 2 **Identification of TDCPP target proteins**

1. (A) Chemical structure of TDCPP.
2. (B) Target prediction using ChEMBL, PharmMapper, and SwissTargetPrediction.

Identification of HCC-related Target Genes

To minimize batch effects, we merged the GSE45267 and GSE74656 datasets and comprehensively normalized the gene expression matrix. PCA showed that the data distribution improved after normalization, and the normalized datasets exhibited clearer clustering patterns. Differential expression analysis identified 16,127 genes that were significantly altered in HCC. These expression changes were visualized using volcano plots and heatmaps (Fig. 3C-D). In WGCNA, we first determined the optimal soft-thresholding power (β) to ensure that the network topology exhibited scale-free characteristics. Power values ranging from 1–20 were systematically evaluated, and the results indicated that $\beta = 5$ was the minimum value required to achieve the scale-free topology criterion ($R^2 \geq 0.8$). Using this parameter, we constructed a TOM and performed hierarchical clustering to identify co-expression modules. This analysis produced eight distinct gene modules, each visualized using different colors (Fig. 3E). Module-trait relationship analysis identified specific modules that were significantly associated with HCC ($P < 0.05$) (Fig. 3F). By integrating DEGs from conventional analysis with WGCNA module genes (after removing duplicate entries), we ultimately identified 2,425 HCC-related genes.

1. (A) PCA scatter plot showing clear separation between the GSE22097 and GSE38736 datasets before batch correction, indicating the presence of batch effects.
2. (B) PCA scatter plot after batch correction showing the integration of the GSE22097 and GSE38736 datasets, indicating reduced batch effects.
3. (C) Volcano plot showing DEGs based on logFC and statistical significance. Red dots indicate upregulated genes, blue dots indicate downregulated genes, and gray dots indicate non-significant genes.
4. (D) Heatmap showing the expression patterns of DEGs across samples. Red indicates upregulation, and blue indicates downregulation.
5. (E) Gene dendrogram generated by WGCNA based on hierarchical clustering of gene co-expression. Module colors in the lower panel represent different gene modules.
6. (F) Module-trait relationship heatmap showing correlations between modules identified by WGCNA and sample traits (control vs experimental). Values in the boxes represent correlation coefficients and p-values.
7. (G) Venn diagram showing the overlap between DEGs (purple) and WGCNA modules (yellow). The overlapping region represents genes identified by both methods.

Identification of TDCPP-associated HCC Disease Targets

Intersection analysis between TDCPP-targeted proteins and HCC-related genes identified 230 potential key targets involved in TDCPP-induced hepatocarcinogenesis (Fig. 4A-B). Functional characterization based on GO and KEGG enrichment analyses (Fig. 4C-D) revealed comprehensive molecular

mechanisms. GO functional annotation analysis indicated that TDCPP caused biological perturbations at multiple levels (Fig. 4A). In terms of biological processes, the targets were significantly enriched in the regulation of cell cycle transitions—such as mitotic cell cycle phase transition, G2/M phase transition, and DNA replication, suggesting that TDCPP may promote abnormal proliferation by disrupting the precise regulation of cell division. In terms of cellular components, the targets were mainly associated with chromosomal regions (such as telomeric regions), protein kinase complexes, and DNA polymerase complexes, indicating that TDCPP may primarily act on molecular machinery responsible for maintaining chromosomal stability and DNA synthesis. At the molecular function level, the targets were significantly enriched in kinase activity—including serine/threonine/tyrosine kinase activity, histone kinase activity, kinase regulator activity, and monosaccharide binding, suggesting that TDCPP may drive HCC progression by influencing kinase signaling cascades and metabolic reprogramming. KEGG pathway enrichment analysis further revealed the complex regulatory network underlying TDCPP-induced hepatocarcinogenesis (Fig. 4B). The results showed that the targets were mainly enriched in classical oncogenic signaling pathways—such as the p53 signaling pathway, FoxO signaling pathway, and cellular senescence—as well as the cell cycle and multiple cancer types, including prostate cancer, bladder cancer, small cell lung cancer, melanoma, glioma, and leukemia. These pathways share common mechanisms in tumor initiation and progression. In addition, the pathways were significantly enriched in stress-response pathways—including endocrine resistance, AGE-RAGE signaling, and fluid shear stress and atherosclerosis—suggesting that TDCPP may promote liver cancer progression through mechanisms involving cellular stress induction, metabolic reprogramming, and therapeutic resistance. Collectively, these findings indicate that TDCPP-induced hepatocarcinogenesis involves dysregulation of cell cycle checkpoints, increased chromosomal instability, and abnormal modulation of signaling pathways such as p53 and FoxO, thereby conferring proliferative advantages to tumor cells. Meanwhile, enrichment of kinase activity regulation and metabolism-related functions suggests enhanced metabolic adaptability in TDCPP-associated tumors.

1. (A) Venn diagram comparing genes associated with TDCPP exposure (purple) and HCC (yellow), showing 230 overlapping genes (3.2%).
2. (B) PPI network visualizing interactions among the overlapping genes. Red nodes indicate upregulated genes, blue nodes indicate downregulated genes, and edges represent predicted interactions.
3. (C) GO enrichment analysis showing overlapping genes involved in biological processes, cellular components, and molecular functions. The x-axis represents gene counts, and the color gradient indicates adjusted p-values (deeper red indicates higher significance).
4. (D) KEGG pathway analysis showing pathways enriched by the overlapping genes. The x-axis represents gene ratio, dot size indicates gene counts, and the color gradient represents adjusted p-values (red indicates higher significance).

Identification of Core Genes in TDCPP-induced HCC

Through comprehensive machine learning analysis of the 230 candidate targets, we constructed 127 predictive models to identify core genes associated with TDCPP-related HCC. The Lasso-RF ensemble model demonstrated the best performance in both the training and validation phases, achieving the highest accuracy (Fig. 5A), thereby identifying the following nine key genes—COL15A1, NDC80, LCAT, MOGAT2, BDH2, CRHBP, CHST4, NAT2, and CYP2C19. The diagnostic potential of these core genes was confirmed through ROC curve analysis (AUC > 0.85, Fig. 5B), and their differential expression patterns in HCC tissues were visualized using volcano plots (Fig. 5C). SHAP interpretability analysis revealed differences in their functional contributions—LCAT (SHAP value = 0.134) emerged as the most influential predictive factor (Fig. 5D).

1. (A) Model performance comparison. The heatmap displays AUC values of different models across datasets. Rows represent models, and the right column shows AUC values (higher values indicate better performance). Colors represent dataset sources.
2. (B) ROC curves of key genes (COL15A1, NDC80, LCAT, MOGAT2, BDH2, CRHBP, CHST4, NAT2, and CYP2C19). The x-axis represents the false positive rate, and the y-axis represents sensitivity. AUC indicates predictive performance.
3. (C) Volcano plot showing DEGs. The x-axis represents logFC, and the y-axis represents −log10(p-value). Red indicates upregulation, blue indicates downregulation, and key genes are labeled.
4. (D) Feature importance ranking. Bar plot showing the ranking of top genes based on feature importance. Wider bars indicate greater contribution to the model.

Molecular Docking Validation of TDCPP-Core Gene Interactions

To validate the potential binding interactions between TDCPP and the identified core genes (COL15A1, NDC80, LCAT, MOGAT2, BDH2, CRHBP, CHST4, NAT2, CYP2C19), we performed comprehensive molecular docking analysis. The results showed that TDCPP could spontaneously bind to all nine target proteins, with binding energies consistently lower than 0 kcal/mol, and strong binding affinities were observed with BDH2, CHST4, CYP2C19, and NAT2 (Table 1), suggesting stable and spontaneous molecular interactions. According to established criteria in molecular docking studies—binding energy < 0 kcal/mol indicates spontaneous binding capability, whereas binding energy < -5.0 kcal/mol indicates strong binding affinity. Visualization of the binding conformations (Fig. 6) provided structural evidence supporting direct molecular interactions between TDCPP and the core HCC-related targets identified through our machine learning analysis.

Table 1
Binding energy between TDCPP and target proteins in molecular docking analysis.

Ligand	Receptor	Binding Energy (kcal/mol)
TDCPP	BDH2	−5.7
TDCPP	CHST4	−6.3
TDCPP	COL15A1	−3.8
TDCPP	CRHBP	−3.7
TDCPP	CYP2C19	−5.4
TDCPP	LCAT	−4.5
TDCPP	MOGAT2	−3.7
TDCPP	NAT2	−5.5
TDCPP	NDC80	−3.7

1. Docking result of TDCPP with BDH2.
2. (B) Docking result of TDCPP with CHST4.
3. (C) Docking result of TDCPP with COL15A1.
4. (D) Docking result of TDCPP with CRHBP.
5. (E) Docking result of TDCPP with CYP2C19.
6. (F) Docking result of TDCPP with LCAT.
7. (G) Docking result of TDCPP with MOGAT2.
8. (H) Docking result of TDCPP with NAT2.
9. (I) Docking result of TDCPP with NDC80.

Discussion

Through integrated analysis of multi-database target prediction results, transcriptomic data, and machine learning and bioinformatics approaches, we systematically identified potential molecular targets associated with TDCPP exposure and the development and progression of HCC. Our network toxicology and bioinformatics analyses identified a total of 230 candidate genes, among which nine core genes—COL15A1, NDC80, LCAT, MOGAT2, BDH2, CRHBP, CHST4, NAT2, and CYP2C19—were considered to play important roles in the mechanisms of TDCPP-related liver carcinogenesis. Differential expression analysis revealed that these core genes exhibited significant abnormal expression patterns in HCC tissues. Machine learning models further confirmed that this gene signature has high diagnostic value. ROC analysis showed that all core genes had AUC values greater than 0.85, while SHAP interpretability analysis further highlighted LCAT (SHAP value = 0.134) as the most influential predictive factor. Molecular docking simulations further supported the biological relevance of these findings. The results

showed that TDCPP could spontaneously bind to all core proteins (binding energies < 0 kcal/mol) and exhibited strong binding affinities with BDH2, CHST4, CYP2C19, and NAT2 (binding energies < -5 kcal/mol). Collectively, these results indicate that the core genes identified in this study may represent important molecular targets through which TDCPP contributes to the development and progression of HCC.

As a widely used organophosphate flame retardant, TDCPP has potential carcinogenic-related toxicity and may participate in tumorigenesis through multiple molecular mechanisms. In the HepG2 hepatocyte model, several organophosphate flame retardants—including TDCIPP—can affect cellular phenotype and function, manifested by alterations in mitochondrial status, disruption of lipid homeostasis, and abnormal organelle function, suggesting that these compounds may exert adverse biological effects by disturbing hepatocellular homeostasis (13). In addition, TDCIPP can induce significant hepatotoxic responses in rats, manifested by pathological alterations in liver tissues and abnormal expression of genes associated with hepatic metabolism, indicating that this compound can disrupt liver physiological function and induce hepatocellular injury (14). Moreover, TDCPP has been shown to induce genome-wide DNA hypomethylation, suggesting that it may alter gene expression patterns through epigenetic regulation (15). Meanwhile, another flame retardant—tris(1,3-dichloro-2-propyl) phosphate—exhibits clear mutagenic activity after metabolic activation in liver homogenates, and its metabolites also display mutagenic effects (16). These lines of evidence collectively suggest that TDCPP may participate in molecular processes related to HCC through multiple mechanisms—including oxidative stress, DNA damage, inflammatory activation, and epigenetic dysregulation.

GO and KEGG enrichment analyses indicated that TDCPP-related targets were mainly enriched in key biological processes such as cell cycle regulation, DNA replication, the p53 signaling pathway, the FoxO signaling pathway, and cellular senescence. Among these processes, abnormal DNA damage responses and dysregulated cell cycle control are considered important molecular events in hepatocarcinogenesis, promoting abnormal cell proliferation and tumor progression (17). As a classical tumor suppressor pathway, the p53 signaling pathway plays a central role in regulating DNA repair, cell cycle arrest, and apoptosis, and its dysfunction has been recognized as a key molecular characteristic of HCC (18). Meanwhile, the FoxO signaling pathway participates in the regulation of cellular metabolism, oxidative stress responses, and cell survival processes, and it plays an important regulatory role in tumor development and progression (19). In addition, cellular senescence-related processes exert dual effects in tumorigenesis—on one hand inducing cell cycle arrest and exerting tumor-suppressive functions, and on the other hand potentially promoting tumor progression by altering the tumor microenvironment (20). These results suggest that TDCPP may participate in HCC-related molecular processes by disrupting cell cycle regulation and stress-response signaling networks.

Among the core genes identified in this study, multiple molecules play important roles in key biological processes such as cell cycle regulation, metabolic homeostasis, and maintenance of the tumor microenvironment. Among them, NDC80 is an important component of the kinetochore complex and plays a central role in the stable attachment between chromosomes and spindle microtubules as well as

in the mitotic process. Its abnormal expression has been confirmed to lead to chromosomal instability and promote tumor cell proliferation and tumor progression (21, 22). In contrast to cell cycle regulation, some core genes mainly participate in metabolic reprogramming processes. For example, LCAT is a key enzyme involved in cholesterol esterification and plays an important role in lipid transport and the maintenance of cholesterol homeostasis. Alterations in its expression have been considered closely associated with tumor-related lipid metabolic reprogramming (23). Similarly, MOGAT2 participates in triglyceride synthesis and lipid absorption processes and may support rapid tumor cell growth by promoting lipid accumulation and energy supply (24). In terms of redox regulation, BDH2 is closely associated with cellular iron metabolism and metabolic homeostasis, and its expression changes may influence cellular oxidative stress levels and regulate cell survival and apoptosis (25). In addition, the metabolic processes of environmental chemicals in vivo may also influence chemical carcinogenic effects, in which the detoxification-related enzyme NAT2 and the cytochrome P450 family member CYP2C19 play important roles in xenobiotic metabolism and may influence toxic effects by regulating the biotransformation and metabolic activity of chemical substances (26, 27). Meanwhile, COL15A1, as an important component of the extracellular matrix, plays an essential role in regulating the tumor microenvironment, cell adhesion, and angiogenesis, and alterations in its expression have been associated with tumor invasion and metastasis (28). In addition, genes such as CHST4 and CRHBP are also involved in immune regulation and cellular signal transduction processes and may play potential roles in maintaining tissue microenvironment homeostasis and cellular stress responses (29, 30). Overall, the coordinated involvement of these core genes in multiple biological processes—including cell cycle regulation, metabolic reprogramming, xenobiotic metabolism, and tumor microenvironment regulation—further reveals the complex molecular network underlying TDCPP-related hepatocarcinogenesis and provides important clues for subsequent mechanistic studies and targeted interventions.

This study has several advantages. First, by integrating network toxicology, bioinformatics analysis, machine learning, and molecular docking approaches, this study constructed a molecular interaction network of TDCPP-related hepatocarcinogenesis at the systems level, providing a new research perspective for understanding the relationship between environmental chemicals and tumor development. Second, by combining differential expression analysis and WGCNA co-expression network analysis to screen HCC-related genes and intersecting them with TDCPP targets predicted from multiple databases, the reliability of identifying potential key targets was improved. Third, machine learning models were further used to screen core genes with high diagnostic value, and SHAP interpretability analysis was applied to evaluate their contributions, thereby enhancing the robustness and interpretability of the model results. Finally, molecular docking analysis verified the potential stable binding capacity between TDCPP and the core proteins, providing structural-level evidence for potential targets and offering a theoretical basis for further mechanistic studies and targeted interventions.

However, this study still has several limitations. First, this study was mainly based on analyses of public databases and computational biology methods, and differences in sample sources and potential biases among datasets may influence the generalizability of the results. Second, the identified core genes and

their associated molecular networks were mainly derived from bioinformatics prediction and machine learning screening, and their biological functions still require further validation through in vitro and in vivo experiments. Third, molecular docking analysis only predicts the potential binding capacity between TDCPP and target proteins at the structural level, and its functional effects in real biological environments remain to be further investigated. Future studies integrating multi-omics data and functional experimental validation are needed to further elucidate the specific molecular mechanisms of TDCPP in hepatocellular carcinogenesis.

Conclusion

This study systematically revealed the potential molecular network through which TDCPP may participate in the development and progression of HCC by integrating bioinformatics analysis, machine learning, and molecular docking approaches, and identified multiple key genes and related signaling pathways. The results indicate that TDCPP may participate in hepatocellular carcinogenesis by affecting cell cycle regulation, metabolic pathways, and tumor-related signaling pathways, thereby providing new insights for further investigation into its molecular mechanisms and potential targeted interventions.

Funding Declaration

This work was supported by the Traditional Chinese Medicine Specialty Construction Project in Foshan's 14th Five-Year Plan (TSZKJS25); The Specialized Medical Training Project in Foshan's 14th Five-Year Plan(FSPY145053); Nationally Recognized Wang Jie Traditional Chinese Medicine Inheritance Studio, Shunde, Foshan, China.

Data availability

The specific accession number of the GEO dataset is GSE45627,GSE74656,GSE54236,GSE135631 and GSE76427. The ChEMBL dataset originates from https://www.ebi.ac.uk/chembl/search_results/Tris(1%2C3-dichloro-2-propyl)%20phosphate. The PubChem dataset originates from https://pubchem.ncbi.nlm.nih.gov/compound/26177. The UniProt dataset originates from https://www.uniprot.org/uniprotkb?query=(proteome:UP000005640). The SwissTargetPrediction dataset originates from http://swisstargetprediction.ch/, with the SMILE C(C(CCl)OP(= O)(OC(CCl)CCl)OC(CCl)CCl)Cl. PharmMapper and PubMed: details are provided in the data availability file.

Abbreviations

HCC
Hepatocellular carcinoma
TDCPP
Tris(1,3-dichloro-2-propyl) phosphate
GEO

Gene Expression Omnibus
SVA
Surrogate variable analysis
PCA
Principal component analysis
DEGs
Differentially expressed genes
WGCNA
Weighted gene co-expression network analysis
TOM
Topological overlap matrix
GO
Gene Ontology
KEGG
Kyoto Encyclopedia of Genes and Genomes
ROC
Receiver operating characteristic
AUC
Area under the curve
SHAP
SHapley Additive exPlanations
SMILES
Simplified molecular-input line-entry system
RF
Random forest
SVM
Support vector machine
GBM
Gradient boosting machine
LDA
Linear discriminant analysis
Enet
Elastic net
GLM
Generalized linear model
BDH2
3-hydroxybutyrate dehydrogenase 2
NDC80
Nuclear division cycle 80 kinetochore complex component
LCAT, Lecithin cholesterol acyltransferase

MOGAT2
Monoacylglycerol O-acyltransferase 2
COL15A1
Collagen type XV alpha 1 chain
CHST4
Carbohydrate sulfotransferase 4
CRHBP
Corticotropin-releasing hormone binding protein
NAT2
N-acetyltransferase 2
CYP2C19
Cytochrome P450 family 2 subfamily C member 19.

Declarations

Declaration of competing interest

The authors declare that there are no known financial interests or personal relationships that might influence the impartiality of the work described in this paper.

Ethics, Consent to Participate, and Consent to Publish declarations

Not applicable.

Clinical Trial Number

Not applicable.

Author Contribution

Yicheng Wan: Writing - original draft, Formal analysis, Data curation, Conceptualization. Chuying Zhong: Writing - review and editing, Methodology, Conceptualization. Mengqing Yang: Software, Resources. Simin Liao: Visualization, Validation, Methodology, Data curation. Dezhu Chen: Visualization, Validation, Supervision, Funding acquisition, Data curation.

References

1. Rumgay H, Arnold M, Ferlay J, Lesi O, Cabasag CJ, Vignat J, Laversanne M, McGlynn KA, Soerjomataram I. Global burden of primary liver cancer in 2020 and predictions to 2040. J Hepatol. 2022;77:1598–606. https://doi.org/10.1016/j.jhep.2022.08.021.
2. Li Q, Ding C, Cao M, Yang F, Yan X, He S, Cao M, Zhang S, Teng Y, Tan N, Wang J, Xia C, Chen W. Global epidemiology of liver cancer 2022: An emphasis on geographic disparities. Chin Med J. 2024;137:2334–42. https://doi.org/10.1097/CM9.0000000000003264.
3. Singh SP, Madke T, Chand P. Global epidemiology of hepatocellular carcinoma. J Clin Exp Hepatol. 2025;15:102446. https://doi.org/10.1016/j.jceh.2024.102446.
4. Stapleton HM, Klosterhaus S, Keller A, Ferguson PL, van Bergen S, Cooper E, Webster TF, Blum A. Identification of flame retardants in polyurethane foam collected from baby products. Environ Sci Technol. 2011;45:5323–31. https://doi.org/10.1021/es2007462.
5. National Toxicology, Program. 2021. 15th report on carcinogens. Report on Carcinogens 15, roc15. https://doi.org/10.22427/NTP-OTHER-1003
6. Wang C, Chen H, Li H, Yu J, Wang X, Liu Y. Review of emerging contaminant tris(1,3-dichloro-2-propyl)phosphate: Environmental occurrence, exposure, and risks to organisms and human health. Environ Int. 2020;143:105946. https://doi.org/10.1016/j.envint.2020.105946.
7. Zhang J, Williams TD, Chipman JK, Viant MR. Defensive and adverse energy-related molecular responses precede tris (1,3-dichloro-2-propyl) phosphate cytotoxicity. J Appl Toxicol. 2016;36:649–58. https://doi.org/10.1002/jat.3194.
8. Liu C, Su G, Giesy JP, Letcher RJ, Li G, Agrawal I, Li J, Yu L, Wang J, Gong Z. Acute exposure to tris(1,3-dichloro-2-propyl) phosphate (TDCPP) causes hepatic inflammation and leads to hepatotoxicity in zebrafish. Sci Rep. 2016;6:19045. https://doi.org/10.1038/srep19045.
9. Chen S, Gong Z, Letcher RJ, Liu C. Promotion effect of liver tumor progression in male kras transgenic zebrafish induced by tris (1,3-dichloro-2-propyl) phosphate. Ecotoxicol Environ Saf. 2020;191:110220. https://doi.org/10.1016/j.ecoenv.2020.110220.
10. Sturla SJ, Boobis AR, FitzGerald RE, Hoeng J, Kavlock RJ, Schirmer K, Whelan M, Wilks MF, Peitsch MC. Systems toxicology: From basic research to risk assessment. Chem Res Toxicol. 2014;27:314–29. https://doi.org/10.1021/tx400410s.
11. Saquib Q, Al-Salem AM, Siddiqui MA, Ansari SM, Zhang X, Al-Khedhairy AA. Organophosphorus flame retardant TDCPP displays genotoxic and carcinogenic risks in human liver cells. Cells. 2022;11:195. https://doi.org/10.3390/cells11020195.
12. Chen H, Wang P, Du Z, Wang G, Gao S. Oxidative stress, cell cycle arrest, DNA damage and apoptosis in adult zebrafish induced by tris(1,3-dichloro-2-propyl) phosphate. Aquat Toxicol. 2018;194:37–45. https://doi.org/10.1016/j.aquatox.2017.11.001.
13. Yu D, Hales BF, Robaire B. Organophosphate ester flame retardants and plasticizers affect the phenotype and function of HepG2 liver cells. Toxicol Sci. 2024;199:261–75. https://doi.org/10.1093/toxsci/kfae034.

14. Wang S, Hu X, Li X. Sub-chronic exposure to tris(1,3-dichloro-2-propyl) phosphate induces sex-dependent hepatotoxicity in rats. Environ Sci Pollut Res. 2019;26:33351–62. https://doi.org/10.1007/s11356-019-06383-5.
15. Volz DC, Leet JK, Chen A, Stapleton HM, Katiyar N, Kaundal R, Yu Y, Wang Y. Tris(1,3-dichloro-2-propyl)phosphate induces genome-wide hypomethylation within early zebrafish embryos. Environ Sci Technol. 2016;50:10255–63. https://doi.org/10.1021/acs.est.6b03656.
16. Gold MD, Blum A, Ames BN. Another flame retardant, tris-(1,3-dichloro-2-propyl)-phosphate, and its expected metabolites are mutagens. Science. 1978;200:785–7. https://doi.org/10.1126/science.347576.
17. Yang SF, Chang CW, Wei RJ, Shiue YL, Wang SN, Yeh YT. 2014. Involvement of DNA damage response pathways in hepatocellular carcinoma. BioMed Res. Int. 2014, 153867. https://doi.org/10.1155/2014/153867
18. Hussain SP, Schwank J, Staib F, Wang XW, Harris CC. TP53 mutations and hepatocellular carcinoma: Insights into the etiology and pathogenesis of liver cancer. Oncogene. 2007;26:2166–76. https://doi.org/10.1038/sj.onc.1210279.
19. Yang S, Pang L, Dai W, Wu S, Ren T, Duan Y, Zheng Y, Bi S, Zhang X, Kong J. Role of Forkhead Box O proteins in hepatocellular carcinoma biology and progression. Front Oncol. 2021;11:667730. https://doi.org/10.3389/fonc.2021.667730.
20. Cai X, Guillot A, Liu H. Cellular senescence in hepatocellular carcinoma: The passenger or the driver? Cells. 2022;12:132. https://doi.org/10.3390/cells12010132.
21. DeLuca JG, Musacchio A. Structural organization of the kinetochore–microtubule interface. Curr Opin Cell Biol. 2012;24:48–56. https://doi.org/10.1016/j.ceb.2011.11.003.
22. Yan X, Huang L, Liu L, Qin H, Song Z. Nuclear division cycle 80 promotes malignant progression and predicts clinical outcome in colorectal cancer. Cancer Med. 2018;7:420–32. https://doi.org/10.1002/cam4.1284.
23. Jonas A. Lecithin cholesterol acyltransferase. Biochim. Biophys Acta. 2000;1529:245–56. https://doi.org/10.1016/s1388-1981(00)00153-0.
24. Yen CL, Stone SJ, Koliwad S, Harris C, Farese RV. DGAT enzymes and triacylglycerol biosynthesis. J Lipid Res. 2008;49:2283–301. https://doi.org/10.1194/jlr.R800018-JLR200.
25. Yang WC, Lin SF, Wang SC, Tsai WC, Wu CC, Wu SC. The effects of human BDH2 on the cell cycle, differentiation, and apoptosis and associations with leukemia transformation in myelodysplastic syndrome. Int J Mol Sci. 2020;21:3033. https://doi.org/10.3390/ijms21093033.
26. Sim E, Abuhammad A, Ryan A. Arylamine N-acetyltransferases: From drug metabolism and pharmacogenetics to drug discovery. Br J Pharmacol. 2014;171:2705–25. https://doi.org/10.1111/bph.12598.
27. Zanger UM, Schwab M. Cytochrome P450 enzymes in drug metabolism: Regulation of gene expression, enzyme activities, and impact of genetic variation. Pharmacol Ther. 2013;138:103–41. https://doi.org/10.1016/j.pharmthera.2012.12.007.

28. Clementz AG, Harris A, Collagen XV. Exploring its structure and role within the tumor microenvironment. Mol Cancer Res 11, 1481–6. https://doi.org/10.1158/1541-7786.MCR-12-0662.
29. Zhang L, Fan Y, Wang X, Yang M, Wu X, Huang W, Lan J, Liao L, Huang W, Yuan L, Pan H, Wu Y, Chen L, Guan J. Carbohydrate sulfotransferase 4 inhibits the progression of hepatitis B virus-related hepatocellular carcinoma and is a potential prognostic marker in several tumors. Front Oncol. 2020;10:554331. https://doi.org/10.3389/fonc.2020.554331.
30. Chen B, Chen S, Wang X, Zhang J, Wang H, Li J, Zhang Z, Yu F, Kong W. A pan-cancer analysis uncovering the function of CRHBP in tumor immunity, prognosis and drug response: Especially its function in LIHC. Sci Rep. 2024;14:3112. https://doi.org/10.1038/s41598-024-52387-8.
31. National Research Council (US) Subcommittee on Flame-Retardant Chemicals. (2000). Toxicological Risks of Selected Flame-Retardant Chemicals. National Academies Press (US). https://doi.org/10.17226/9841
32. Kamata E, Naito K, Nakaji Y, Ogawa Y, Suzuki S, Kaneko T, Takada K, Kurokawa Y, Tobe M. [Acute and subacute toxicity studies of tris (1,3-dichloro-2-propyl) phosphate on mice]. Eisei Shikenjo Hokoku. 1989;(107):36–43. Japanese. PMID: 2636931.

Figures

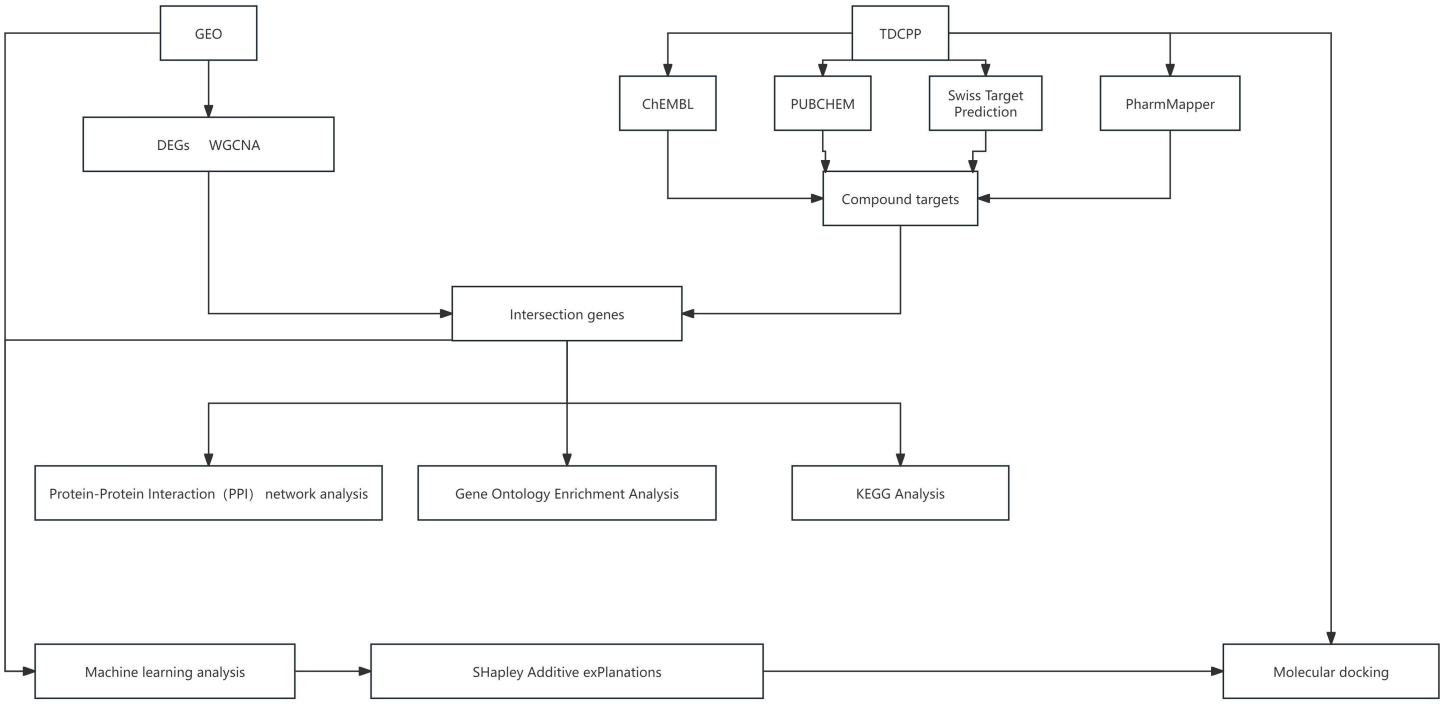


Figure 1

Workflow of the present study

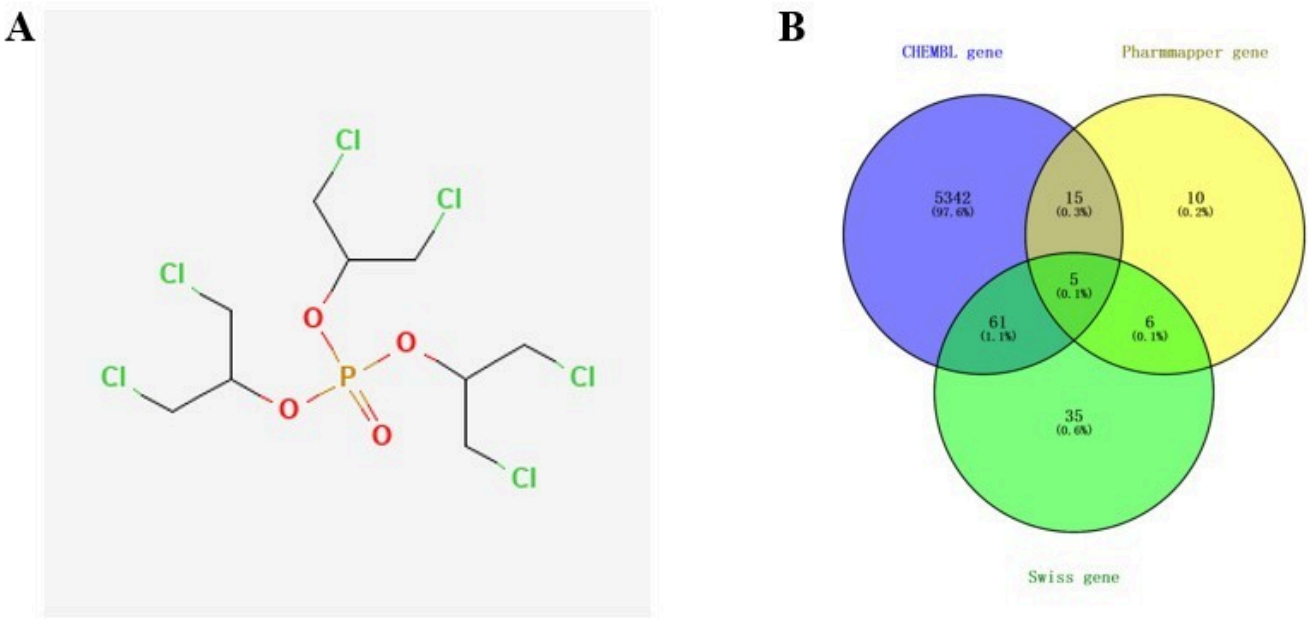


Figure 2

Identification of TDCPP target proteins

(A) Chemical structure of TDCPP.
(B) Target prediction using ChEMBL, PharmMapper, and SwissTargetPrediction.

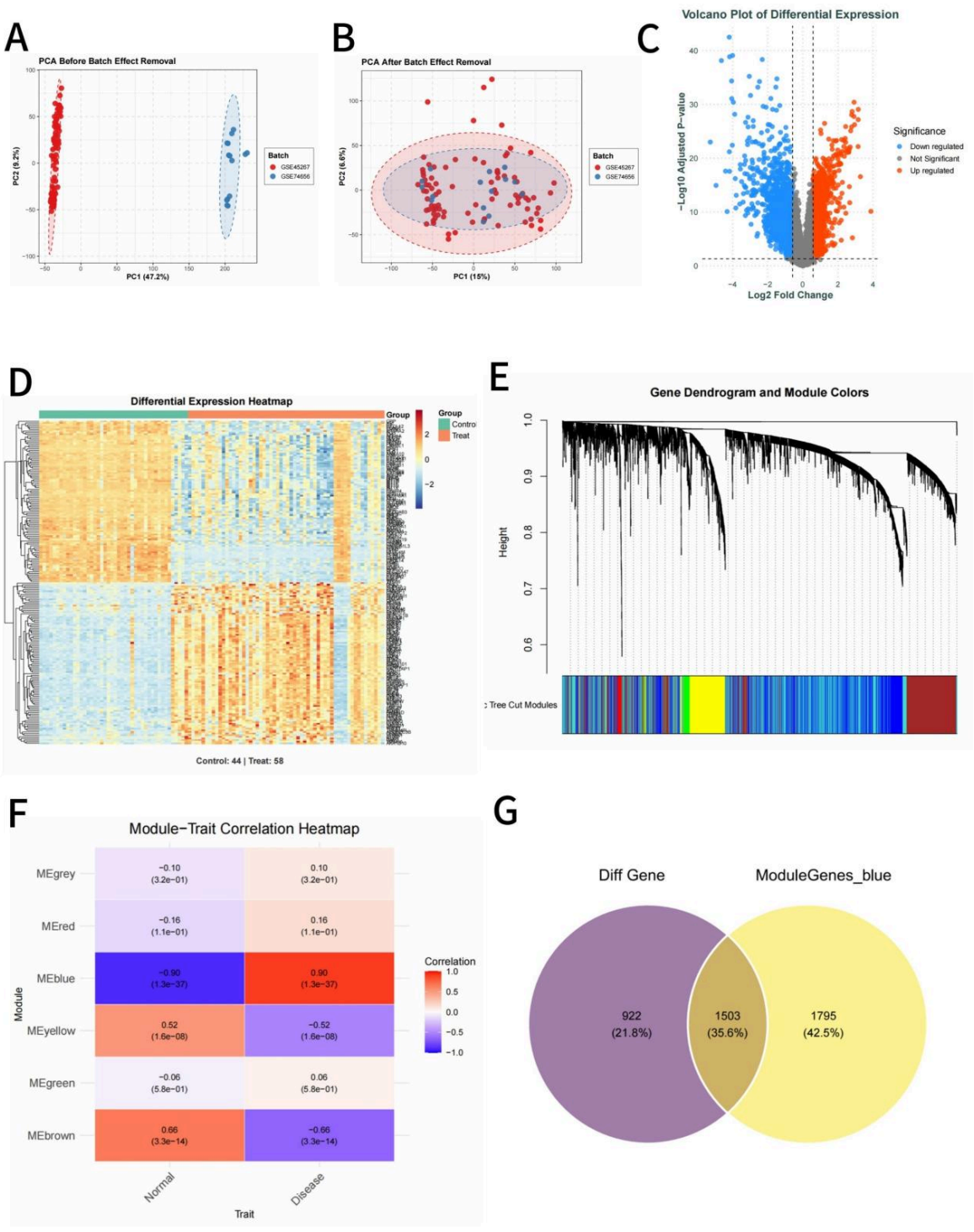


Figure 3

Identification of HCC-related target genes

(A) PCA scatter plot showing clear separation between the GSE22097 and GSE38736 datasets before batch correction, indicating the presence of batch effects.
(B) PCA scatter plot after batch correction showing the integration of the GSE22097 and GSE38736

datasets, indicating reduced batch effects.
(C) Volcano plot showing DEGs based on logFC and statistical significance. Red dots indicate upregulated genes, blue dots indicate downregulated genes, and gray dots indicate non-significant genes.
(D) Heatmap showing the expression patterns of DEGs across samples. Red indicates upregulation, and blue indicates downregulation.
(E) Gene dendrogram generated by WGCNA based on hierarchical clustering of gene co-expression. Module colors in the lower panel represent different gene modules.
(F) Module-trait relationship heatmap showing correlations between modules identified by WGCNA and sample traits (control vs experimental). Values in the boxes represent correlation coefficients and p-values.
(G) Venn diagram showing the overlap between DEGs (purple) and WGCNA modules (yellow). The overlapping region represents genes identified by both methods.

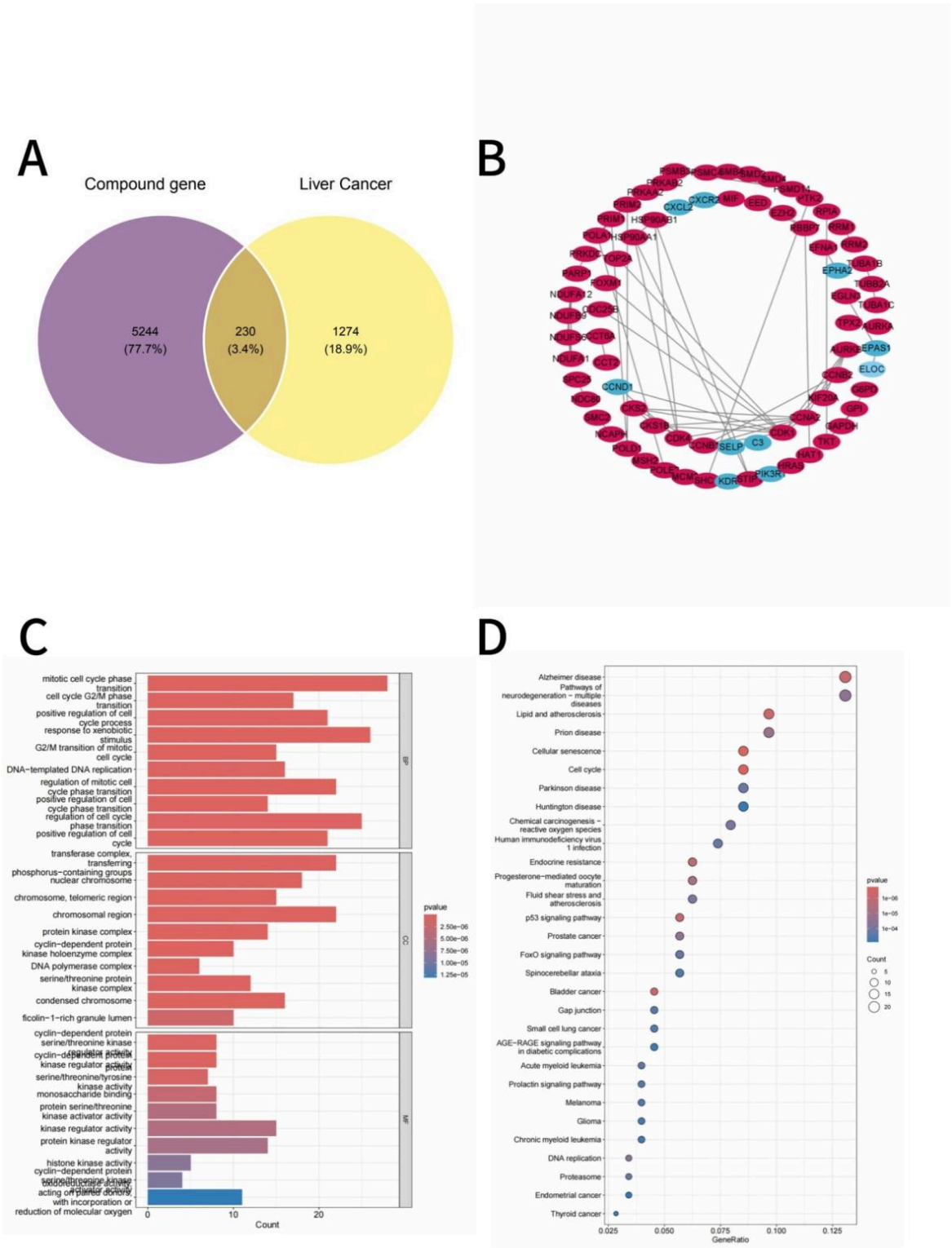
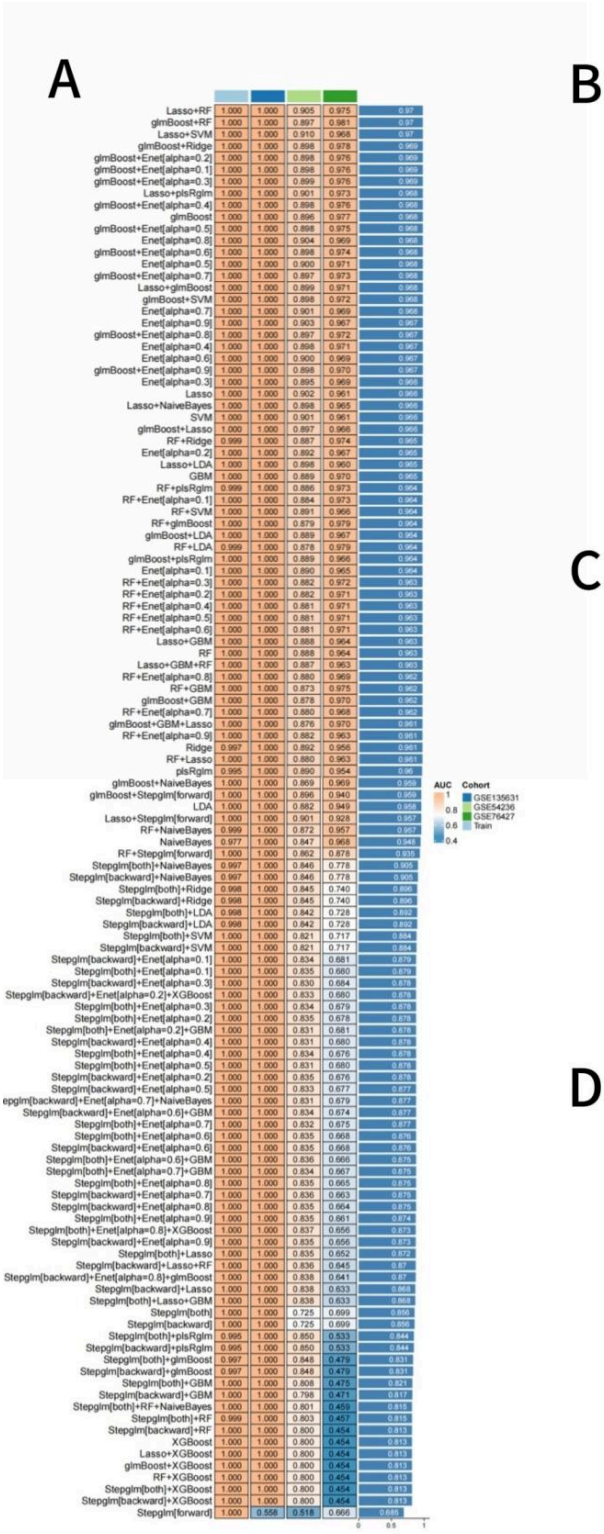


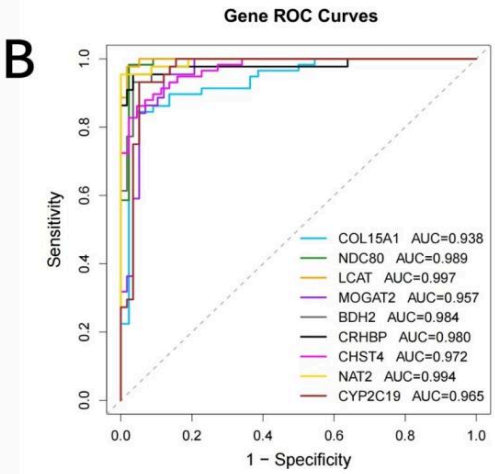
Figure 4

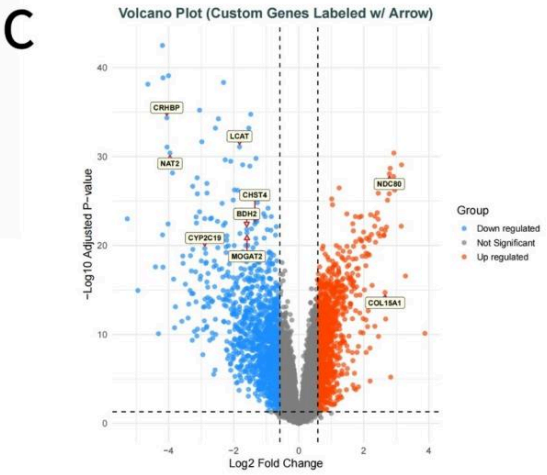
Identification of TDCPP-associated disease targets in HCC

(A) Venn diagram comparing genes associated with TDCPP exposure (purple) and HCC (yellow), showing 230 overlapping genes (3.2%).
(B) PPI network visualizing interactions among the overlapping genes. Red nodes indicate upregulated

genes, blue nodes indicate downregulated genes, and edges represent predicted interactions.
(C) GO enrichment analysis showing overlapping genes involved in biological processes, cellular components, and molecular functions. The x-axis represents gene counts, and the color gradient indicates adjusted p-values (deeper red indicates higher significance).
(D) KEGG pathway analysis showing pathways enriched by the overlapping genes. The x-axis represents gene ratio, dot size indicates gene counts, and the color gradient represents adjusted p-values (red indicates higher significance).







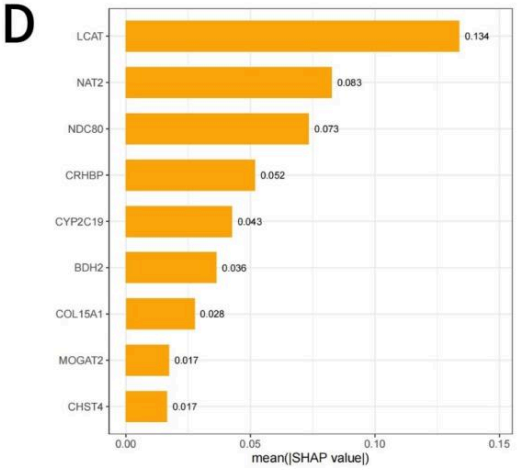


Figure 5

Machine learning-based identification and diagnostic validation of key genes in HCC

(A) Model performance comparison. The heatmap displays AUC values of different models across datasets. Rows represent models, and the right column shows AUC values (higher values indicate better performance). Colors represent dataset sources.
(B) ROC curves of key genes (COL15A1, NDC80, LCAT, MOGAT2, BDH2, CRHBP, CHST4, NAT2, and CYP2C19). The x-axis represents the false positive rate, and the y-axis represents sensitivity. AUC indicates predictive performance.
(C) Volcano plot showing DEGs. The x-axis represents logFC, and the y-axis represents −log10(p-value). Red indicates upregulation, blue indicates downregulation, and key genes are labeled.
(D) Feature importance ranking. Bar plot showing the ranking of top genes based on feature importance. Wider bars indicate greater contribution to the model.

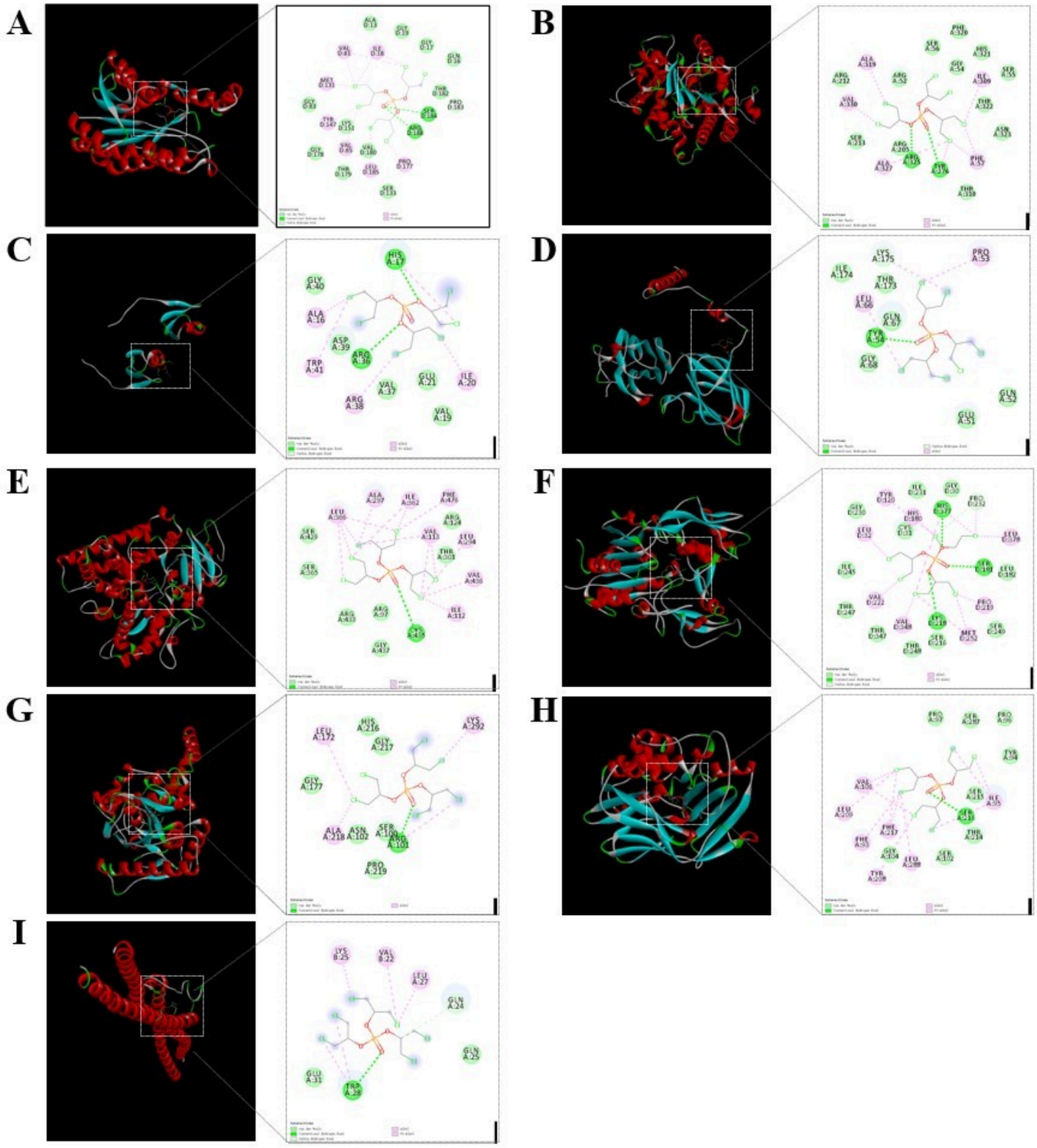


Figure 6

Molecular docking validation of TDCPP-core gene interactions

Docking result of TDCPP with BDH2.
(B) Docking result of TDCPP with CHST4.
(C) Docking result of TDCPP with COL15A1.

(D) Docking result of TDCPP with CRHBP.
(E) Docking result of TDCPP with CYP2C19.
(F) Docking result of TDCPP with LCAT.
(G) Docking result of TDCPP with MOGAT2.
(H) Docking result of TDCPP with NAT2.
(I) Docking result of TDCPP with NDC80.

Supplementary Files

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

- DataSource.zip