OncoTargets and Therapy

Dovepress
Taylor & Francis Group

Open Access Full Text Article

ORIGINAL RESEARCH

Integrated Network Pharmacology, Machine Learning and Experimental Validation to Identify the Key Targets and Compounds of *TiaoShenGongJian* for the Treatment of Breast Cancer

Huiyan Ying[1], Weikaixin Kong[1,2], Xiangwei Xu[3]

[1]Institute for Molecular Medicine Finland (FIMM), Hilife, University of Helsinki, Helsinki, Finland; [2]Department of Molecular and Cellular Pharmacology, School of Pharmaceutical Sciences, Peking University Health Science Center, Beijing, People's Republic of China; [3]Affiliated Yongkang First People's Hospital and School of Pharmaceutical Sciences, Hangzhou Medical College, Hangzhou, Zhejiang, People's Republic of China

Correspondence: Xiangwei Xu, Yongkang First People's Hospital and School of Pharmaceutical Sciences, Hangzhou Medical College, Hangzhou, Zhejiang, 321300, People's Republic of China, Tel +86 15858830343, Email xuxiangwei@hmc.edu.cn; Huiyan Ying, Institute for Molecular Medicine Finland (FIMM), HiLIFE, University of Helsinki, Helsinki, 00290, Finland, Tel +358 468489596, Email huiyan.ying@helsinki.fi

Background: TiaoShenGongJian (TSGJ) decoction, a traditional Chinese medicine for breast cancer, has unknown active compounds, targets, and mechanisms. This study identifies TSGJ's key targets and compounds for breast cancer treatment through network pharmacology, machine learning, and experimental validation.

Methods: Bioactive components and targets of TSGJ were identified from the TCMSP database, and breast cancer-related targets from GeneCards, PharmGkb, and RNA-seq datasets. Intersection of these targets revealed therapeutic targets of TSGJ. PPI analysis was performed via STRING, and machine learning methods (SVM, RF, GLM, XGBoost) identified key targets, validated by GSE70905, GSE70947, GSE22820, and TCGA-BRCA datasets. Pathway analyses and molecular docking were performed. TSGJ and core compounds' effectiveness was confirmed by MTT and RT-qPCR assays.

Results: 160 common targets of TSGJ were identified, with 30 hub targets from PPI analysis. Five predictive targets (HIF1A, CASP8, FOS, EGFR, PPARG) were screened via SVM. Their diagnostic, biomarker, immune, and clinical values were validated. Quercetin, luteolin, and baicalein were identified as core components. Molecular docking confirmed their strong affinities with predicted targets. These compounds modulated key targets and induced cytotoxicity in breast cancer cell lines in a similar way as TSGJ.

Conclusion: This study reveals the main active components and targets of TSGJ against breast cancer, supporting its potential for breast cancer prevention and treatment.

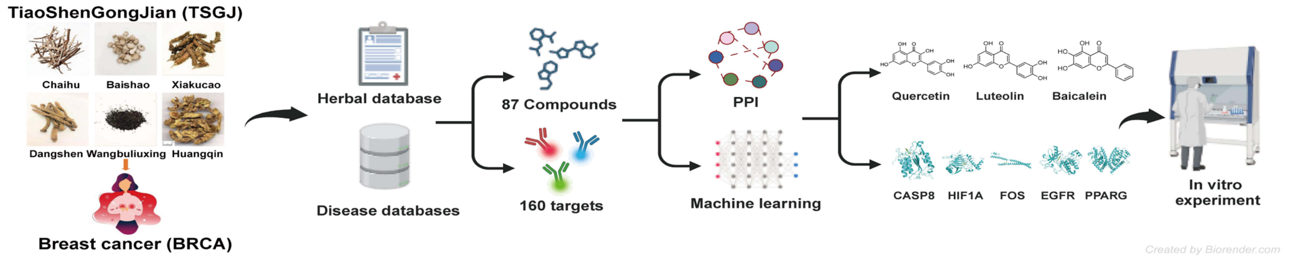
Keywords: TiaoShenGongJian decoction, traditional Chinese medicine, breast cancer, network pharmacology, machine learning, molecular mechanisms

Introduction

Breast cancer, a life-threatening disease with limited therapeutic potential, ranks among the most prevalent malignant tumors affecting women and is the most frequently diagnosed cancer, as well as the leading cause of cancer-related mortality in women.[1,2] Currently, radiotherapy, chemotherapy, and surgery remain the primary treatments for breast cancer.[3] However, for patients with advanced cancer, characterized by metastasis or recurrence post-surgery, conventional treatments are effective in only 25–35% of cases.[4] Cytotoxic chemotherapy drugs, such as taxanes, capecitabine, or anthracyclines, continue to be the mainstay of systemic treatment for most patients, yet they achieve favorable outcomes

Received: 29 August 2024
Accepted: 24 December 2024
Published: 16 January 2025

Graphical Abstract



in only about one-third of cases.[5,6] Chemotherapy drugs are nonselective, causing the destruction of both normal and cancer cells, which results in significant side effects.[7]

Recently, complementary and alternative therapies with minimal side effects, particularly traditional Chinese medicines (TCMs), have been increasingly studied as adjuvant therapies for cancer patients.[8] With a history spanning over 2000 years, TCM offers a well-developed theoretical framework for the diagnosis, prevention, and treatment of various ailments. TCM provides essential guidelines for treating numerous complex and challenging conditions.[9] In cancer treatment, TCM has become an indispensable adjunctive therapy due to its natural ingredients, wide availability, affordability, and minimal side effects, especially for patients who cannot tolerate Western medications.[10] TiaoShenGongJian (TSGJ) decoction, primarily composed of six herbs: Chaihu (*Bupleurum chinense* DC)., Baishao (*Paeonia lactiflora* Pall)., Xiakucao (*Prunella vulgaris* L)., Dangshen (*Codonopsis pilosula* (Franch). Nannf)., Wangbuliuxing (*Gypsophila vaccaria* (L). Sm)., and Huangqin (*Scutellaria baicalensis* Georgi), was originally proposed by Liu Shaowu, the founder of the theory of three parts and six diseases and a prominent TCM expert.[11] These herbs has been widely used to treat various advanced cancers, including breast cancer, nasopharyngeal carcinoma, esophageal cancer, and lymphoma.[12–16] However, its inherent mechanisms remain unknown, as comprehensive or systematic analyses of the signaling pathways and key targets involved in the action of TSGJ are lacking. As a traditional medicinal formula, TSGJ possesses multiple bioactive components, targets, and regulatory pathways. Therefore, traditional single-target pharmacological models are insufficient for in-depth studies.[17]

Network pharmacology has emerged as a powerful discipline to address the limitations of traditional single-target pharmacological models, particularly in the context of TCM's complex bioactive components, multiple targets, and intricate regulatory pathways. This rapidly growing field aims to reveal the intricate connections between targets, compounds, and diseases from a comprehensive viewpoint, making it highly suitable for studying the multifaceted nature of TCM.[18] It has proven particularly effective in investigating the "multiple-components and multiple-targets" mechanism of TCM.[19] Significant advancements have been made in understanding the role of TCM in preventing and treating breast cancer.[20–22] However, network pharmacology approaches may fall short in capturing the dynamic nature of disease progression and processing large datasets efficiently. To address these limitations, the integration of other approaches has become essential.

Machine learning, a key aspect of artificial intelligence, is essential for handling complex biological data, aiding in disease diagnosis,[23] and drug screening.[24] By incorporating machine learning algorithms, analytical models can achieve enhanced prediction accuracy and provide detailed analyses of disease targets.[25] The integration of network pharmacology with machine learning offers distinct advantages in identifying key targets and elucidating the pharmacological mechanisms of TCM in treating diverse diseases.[26,27] This combined approach provides a promising framework for quantifying relationships between TCM components and diseases, offering a novel route for exploring TCM mechanisms. Consequently, such methodologies are invaluable for the comprehensive and systematic analysis required to understand the treatment mechanisms of traditional formulas like TSGJ, thereby addressing the deficiencies of conventional single-target pharmacological models.

In our study, we used network pharmacology to investigate the bioactive ingredients and targets of TSGJ for treating breast cancer. We identified potential targets and analyzed them using machine learning algorithms for gene analysis and feature selection. The best algorithm helped pinpoint key diagnostic targets of TSGJ, enhancing our understanding of its core compounds and their role in breast cancer. Additionally, molecular docking provided insights into the binding interactions of these compounds with the targets. Finally, the cellular effect of the TSGJ and core compounds was validated through MTT, RT-qPCR. This workflow is illustrated in Figure 1.

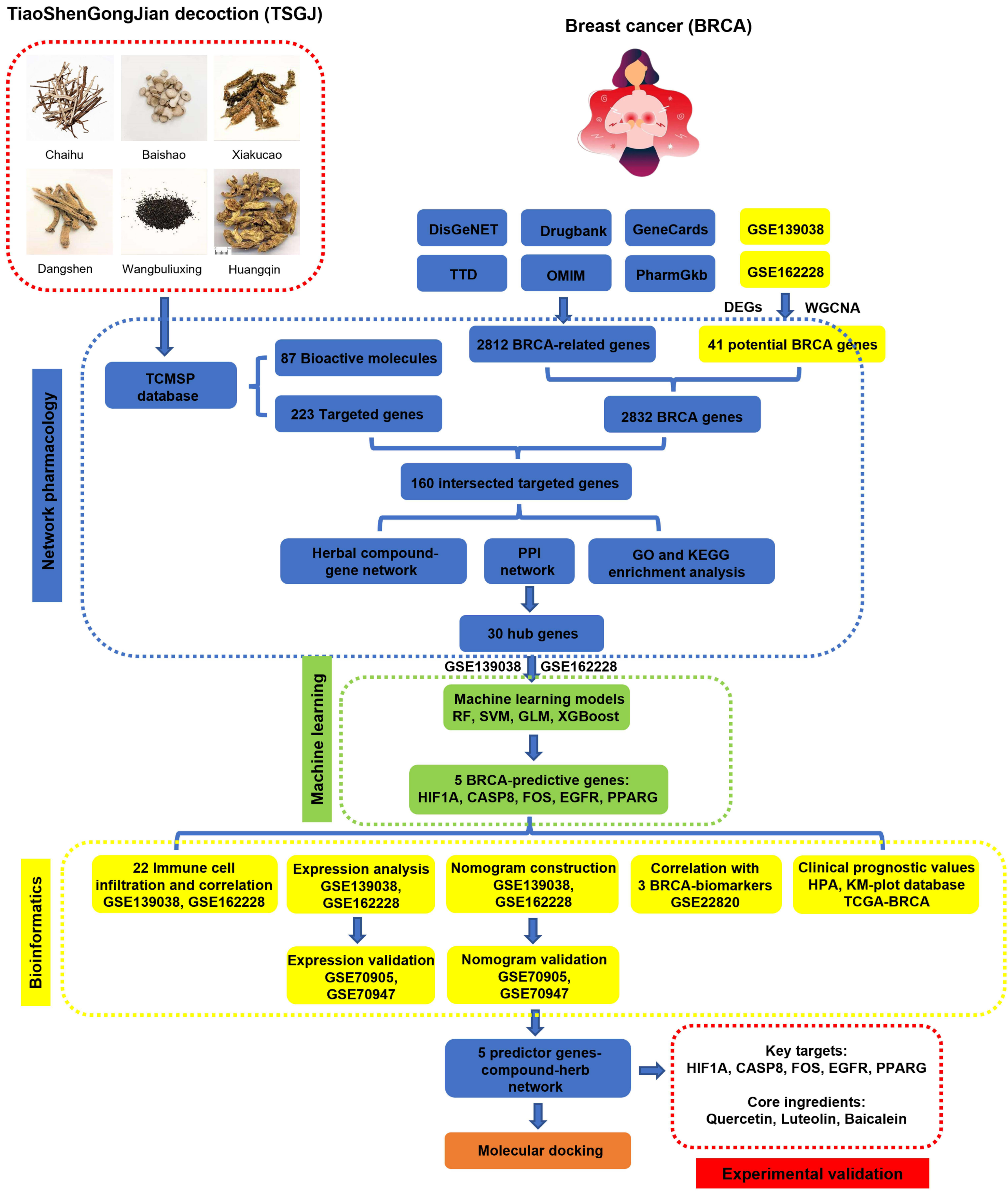

TiaoShenGongJian decoction (TSGJ)
Chaihu
Baishao
Xiakucao
Dangshen
Wangbuliuxing
Huangqin
Breast cancer (BRCA)
DisGeNET
Drugbank
GeneCards
GSE139038
TTD
OMIM
PharmGkb
GSE162228
DEGs
WGCNA
Network pharmacology
TCMSP database
87 Bioactive molecules
223 Targeted genes
2812 BRCA-related genes
41 potential BRCA genes
2832 BRCA genes
160 intersected targeted genes
Herbal compound-gene network
PPI network
GO and KEGG enrichment analysis
30 hub genes
GSE139038
GSE162228
Machine learning
Machine learning models RF, SVM, GLM, XGBoost
5 BRCA-predictive genes: HIF1A, CASP8, FOS, EGFR, PPARG
Bioinformatics
22 Immune cell infiltration and correlation GSE139038, GSE162228
Expression analysis GSE139038, GSE162228
Nomogram construction GSE139038, GSE162228
Correlation with 3 BRCA-biomarkers GSE22820
Clinical prognostic values HPA, KM-plot database TCGA-BRCA
Expression validation GSE70905, GSE70947
Nomogram validation GSE70905, GSE70947
5 predictor genes-compound-herb network
Molecular docking
Key targets: HIF1A, CASP8, FOS, EGFR, PPARG
Core ingredients: Quercetin, Luteolin, Baicalein
Experimental validation

Figure 1 A visual workflow outlining the analysis process employed in this study.

https://doi.org/10.2147/OTT.S486300

Materials and Methods

Screening of Herbal Bioactive Compounds and Targets in TSGJ Decoction

The bioactive compounds and corresponding targets associated with TSGJ decoction were collected from the Traditional Chinese Medicine Systems Pharmacology Database (TCMSP, https://tcmsp-e.com/index.php) with the following filter parameters: oral bioavailability (OB ≥ 30%) and drug likeness (DL ≥ 0.18). Then, the UniProt database (https://www.uniprot.org/) was used to convert the obtained target names into standard gene symbol IDs.[28]

Screening of Breast Cancer-Related Targets

In the initial step, we utilized the keyword "breast cancer" to retrieve the relevant targets associated with breast cancer from various well-known medical databases, including Genecards (http://www.genecards.org/), PharmGkb (https://www.pharmgkb.org/), DisGeNET (https://www.disgenet.org/browser/0/1/0/C0678222/), OMIM (https://omim.org/), Drugbank (https://www.drugbank.ca/), and TTD (https://db.idrblab.net/ttd/). The thresholds were, respectively set at relevance score >10 in Genecards database[29] and Score gda >0.1 in DisGeNET database[30] to obtain a more focused and manageable set of disease targets for the subsequent research.

Second, potential genes associated with breast cancer were screened in the Gene Expression Omnibus (GEO, https://www.ncbi.nlm.nih.gov/geo/) database based on bioinformatics approaches. Here, we utilized the GSE139038 (24 normal samples, 41 tumor samples) and GSE162228 (24 normal samples, 109 tumor samples) datasets to perform differentially expressed gene analysis and weighted gene coexpression network analysis (WGCNA). The expression data were normalized and merged by the "sva" package in R language. A | log2 (fold change) | > 1 and adjusted p value < 0.05 were used to screen for DEGs in the GEO datasets through the use of the "limma" package in R software. Furthermore, the "WGCNA" package in R was used to identify the main gene modules associated with breast cancer in the GEO datasets. A similar matrix was formed based on the gene expression data within the GEO datasets. By setting an appropriate soft power be β value, the topological overlap matrix (TOM) and TOM-based dissimilarity were obtained to profile the gene coexpression modules with a minimum of 60 genes in each module. To identify hub genes within each module, we set the criteria of module membership (MM > 0.8) and gene significance (GS > 0.5).[31] Finally, the hub genes from the WGCNA were intersected with the differentially expressed genes to obtain genes with high potential for association with breast cancer. All the targets from the above databases were combined to identify breast cancer-related targets.

TSGJ Therapeutic Targets and protein–protein Interaction (PPI) Network

A Venn diagram of the TSGJ targets and the breast cancer-related targets was drawn using the "Venn" package in R. The protein family class information of the intersected targets was queried from DisGeNET (https://www.disgenet.org/) and then presented as a pie chart. To better understand the interactions between bioactive compounds and disease targets, we used Cytoscape 3.8.0 software to construct a comprehensive network showing the connections between herbal compounds and their corresponding disease targets. Additionally, the intersecting targets of herbal compounds and breast cancer were imported into the STRING platform (https://string-db.org/) to construct a protein–protein interaction (PPI) network by setting confidence scores > 0.4, setting the protein type as "Homo sapiens", and hiding disconnected nodes. The PPI network was then downloaded and uploaded to Cytoscape for topological analysis.

The hub genes were chosen systematically by evaluating various network centrality measures, including degree (DC), eigenvector (EC), betweenness (BC), closeness (CC), LAC, and network (NC), through the CytoNCA plugin in Cytoscape. These measures assessed the significance of nodes (genes) within the PPI network, allowing us to identify the most influential genes as core candidates. Genes with higher centrality values, representing higher connectivity and importance in the network, were considered hub genes.[32]

Construction of Machine Learning Models to Screen Disease-Predictive Genes

Four machine learning models, namely, extreme gradient boosting (XGBoost), support vector machine (SVM), random forest (RF) and generalized linear model (GLM), were established to select the key predictive genes of breast cancer within the hub genes of the PPI network. XGBoost is an ensemble machine learning method based on gradient boosting that handles various types of data both in classification and regression problems.[33] SVM, a powerful supervised algorithm, aims to determine the hyperplane that effectively separates the data points in the feature space to ensure optimal model performance and robust results.[34] RF is another ensemble method that operates by building decision trees to make predictions.[35] The GLM generalizes multiple linear regression models, and its interpretability is significantly enhanced, particularly when forward feature selection is employed to build the model.[36] First, the four packages "caret", "randomforest", "dalex" and "xgboost" in R were used to build the above models. Then, the GSE139038 and GSE162228 datasets were randomly split into a training set and a test set at a proportion of 3:1. The models were fit in the training set by 5-fold cross-validation under default parameters. After that, the model accuracy was evaluated in the test set and could be visualized by the residual distribution and box line. The ROC curves of the four models were also plotted using the "pROC" package in R. Consequently, the permutation-based feature importance of the optimal model was used to screen out the top 5 key predictive genes related to breast cancer.

Gene Expression and Diagnostic Analysis

The expression of the 5 key genes was visualized by a "limma" boxplot in R to show the difference between normal samples and tumor samples from the GSE139038 and GSE162228 datasets. The ROC curves were plotted to evaluate the diagnostic sensitivity and specificity of each gene. Furthermore, the analysis was also performed based on two other external validation datasets: the GSE70905 (47 normal samples, 47 tumor samples) and GSE70947 (148 normal samples, 148 tumor samples) datasets.

Nomogram Model Construction and Disease-Related Biomarker Correlation

The five genes from the optimal machine learning model were regarded as predictors for constructing a nomogram model using the "rms" R package for evaluating the occurrence of breast cancer. Every predictor could obtain a score based on the expression level to sum the total score to represent the risk of the disease. The predictive power of the nomogram model was evaluated using calibration curves and decision curve analysis (DCA). In addition, the ROC curves of the model were drawn to assess the accuracy of the prediction within two external GEO databases.[37] Moreover, we included an additional external dataset, GSE22820 (10 normal samples, 176 tumor samples), to determine the associations between the predictor genes and 3 well-established biomarkers of breast cancer, namely, ER (ESR1), PR (PGR), and HER2 (ERBB2), which have been extensively studied and reported in the literature.[38] Spearman correlation analysis was implemented on the expression data, and statistical significance was considered at a threshold of $p < 0.05$.

Assessment of the Correlations Between the Immune Cell Content and Predictor Gene Expression

The "CIBERSORT" R package was used to analyze the levels of infiltrating immune cells in breast cancer samples and normal tissue samples. A violin plot was subsequently constructed to visualize the differences between these two groups. To further identify the correlations between the 5 predictor genes and immune cells, correlation coefficients were calculated and are presented in a heatmap.

Evaluation of Key Protein Expression and Clinical Prognostic Correlations

We used two public databases, namely, the HPA (https://www.proteinatlas.org/) and Kaplan–Meier plotter (https://kmplot.com) databases, to preliminarily validate the significance of the 5 identified key targets in breast cancer. The immunohistochemical images of the 5 key target proteins expressed in normal tissue or breast cancer samples were downloaded from the HPA database to represent protein expression differences. To further explore the effect of the key targets on the prognosis of breast cancer patients, we analyzed the correlation between overall survival and the mRNA expression data of 2976 patients through the Kaplan–Meier plotter database. In addition, mRNA expression data and corresponding clinical information for breast cancer patients were obtained from the TCGA database (https://portal.gdc.

https://doi.org/10.2147/OTT.S486300

cancer.gov/). The clinical prognostic impact of the key targets on three main prognostic factors, namely, age, tumor stage and TNM stage, was evaluated within the TCGA-BRCA cohort.

GO and KEGG Enrichment Analysis

To identify the gene functions and signaling pathways of the compound-disease targets, enrichment analysis based on the Gene Ontology (GO) database (https://geneontology.org/) and Kyoto Encyclopedia of Genes and Genomes (KEGG) database (https://www.kegg.jp/) was carried out by using several R packages, including "ggplot2", "org.Hs.eg.db", "clusterProfiler", "enrichplot", and "pathview". The GO analysis revealed three significant aspects, namely, biological processes (BP), cell components (CC), and molecular functions (MF),and the top 5 main terms in each aspect were collected. KEGG was adopted for the biological interpretation and systematic analysis of the genes to localize them to signaling pathways. The top 20 terms with adjusted p values less than 0.05 were considered. Furthermore, the top 10 GO-BP and KEGG pathways related to the 5 key genes were selected to profile their relationships in the Sankey plot with the "ggsankey" R package.

Network of Key Gene–Compound–Herb Interactions and Molecular Docking

First, the 5 key gene–compound–herb network was visualized through a Sankey diagram to visualize the interrelationship between the bioactive compounds of the herbs and the key genes. Molecular docking was then employed to precisely determine the conformation of compounds within the specific binding site of the targets and assess their binding affinity. The molecular docking procedure was mainly divided into 3 parts: (1) Preparation of the proteins: In this study, the crystal structures of the 5 key targets were downloaded from the PDB database (https://www.rcsb.org/) in the PDB format and then imported into PyMOL software to remove solvents and organics. Subsequently, the missing hydrogens were added to the proteins using AutoDock Tools 1.5.6 software and exported in PDBQT format. (2) Preparation of ligands: The 2D structures of the bioactive compounds were obtained from the PubChem database (https://pubchem.ncbi.nlm.nih.gov/) in the sdf format. Then, Chem3D software was applied to perform energy minimization in the MM2 force field, and the results were saved in 3D mol2 format. The PDBQT format was converted to AutoDock Tools 1.5.6 software. (3) Docking and calculation: First, the appropriate mating pockets covering the whole proteins for blind docking were constructed with the "grid box" operation in AutoDock Tools 1.5.6 software. The configuration of the mating box was saved for subsequent docking. The rigid docking of the ligands and proteins was performed with Vina software, and the binding energy was calculated based on the Lamarckian genetic algorithm.[39] After that, the optimal conformation of the ligand–protein complex with the highest binding affinity was visualized by PyMOL software and the interactions between them were analyzed by Discovery Studio software.

Determination of Compound Concentrations in TSGJ by HPLC

The TSGJ dry powder (100 mg, provided by Xinrong Biotechnology Co., Ltd.) was dissolved in 50 mL of methanol using a 40-minute sonication to ensure complete dissolution. The resulting solution was centrifuged, and the supernatant was filtered through a 0.22 μm syringe filter. For high-performance liquid chromatography (HPLC) analysis, 10 μL of the filtrate was injected into the system. An appropriate amount of quercetin, luteolin, and baicalein standards was accurately weighed and individually dissolved or mixed in 5 mL of methanol. These solutions were prepared and injected following the same procedure. HPLC detection was performed using an Agilent C18 column (4.6 × 250 mm, 3 μm) with acetonitrile and 0.1% phosphoric acid aqueous solution as mobile phases A and B, respectively. Gradient elution was applied with a program transitioning from 20% to 42% mobile phase A and 80% to 58% mobile phase B over 35 minutes. UV detection was conducted at a wavelength of 280 nm. The column temperature was maintained at 35°C, and the flow rate was set to 0.8 mL/min.

Cell Culture and Chemical Reagents

The human breast cancer cell lines MDA-MB-231 (ATCC) were cultured with RPMI1640 containing 10% FBS, 2 mm L-glutamine, and 1% Penicillin (10000 U/mL)-Streptomycin (10000 μg/mL) at 37 °C with 5% CO_2. MCF-10A cells (ATCC) were incubated with DMEM/F12 medium including 10% FBS, 2 mm L-glutamine, 10 μg/mL human insulin and

1% Penicillin (10000 U/mL)-Streptomycin (10000 μg/mL) at 37 °C with 5% CO_2. Quercetin (CAS No: 117–39-5, ≥ 98% purity), Luteolin (CAS No: 491–70-3, ≥ 98% purity), Baicalein (CAS No: 491–67-8, ≥ 98% purity) were purchased from Sigma Aldrich and then dissolved in 0.5% dimethyl sulfoxide (DMSO) for drug administration. The composite solution of the three compounds (QLB solution) was prepared by mixing quercetin, luteolin, and baicalein at ratio of 1.5:1.5:1, to match their respective levels in the TSGJ solution.

Cell Viability Assay

The MDA-MB-231 cells and MCF-10A cells (100 μL, 5×10^3 cells) were seeded on a 96-well plate and were respectively treated with different concentrations of TSGJ solutions (0, 0.5, 1, 1.5, 2, 2.5, 3, 4 mg/mL), and equivalent composite QLB solution (0, 15, 30, 45, 60, 75, 90, 120 μg/mL) for 48 h. The 0.5% DMSO was taken to dissolve all the drugs and 0.5% DMSO as a control group to subtract the effect of DMSO. Then, the culture medium was replaced by 100 μL MTT (0.5 mg/mL in PBS) and kept culture for additional 4h. After that, the MTT solution was removed and 100 μL DMSO was added to completely dissolve the formazan crystals. The absorbance intensity was detected by Varioskan Flash plate reader (Thermo Fisher Scientific, USA) at 570 nm. All the experiments were performed with 6 duplicates in 3 independent times.

RT-qPCR Analysis

The MDA-MB-231 cells and MCF-10A cells (2×10^5 cells) were seeded on a 6-well plate and respectively treated with TSGJ solutions (2 mg/mL), and equivalent composite QLB solution (60 μg/mL) for 48 h. Subsequently, total RNA was extracted from the cells using a Trizol kit and converted into cDNA with a reverse transcription kit. The expression of the associated genes was detected using the SYBR kit and normalized to GAPDH (endogenous control). Non-treated MDA-MB-231 cells and MCF-10A cells were also measured. The primers for each gene are listed in Supplementary Table 1.

Results

Screening of TSGJ Bioactive Compounds and Targets

The herbal composition of the TSGJ decoction and their detailed information are listed in Table 1. By searching each herb in the TCMSP database, bioactive compounds with OB ≥ 30% and DL ≥ 0.18 were obtained. A total of 87 TSGJ compounds were collected, 17 from *Chaihu*, 13 from *Baishao*, 11 from *Xiakucao*, 21 from *Dangshen*, 4 from *Wangbuliuxing*, and 36 from *Huangqin* (Supplementary Table 2), and an upset plot was drawn to further illustrate the composition of the herb compounds (Supplementary Figure 1). Then, the 223 corresponding targets of TSGJ associated with the compounds were identified, and their gene symbol IDs were obtained from the UniProt database (Supplementary Table 3).

Screening of Breast Cancer-Related Targets

First, we identified thousands of targets associated with breast cancer that satisfied the screening criteria from 6 databases, namely, DisGeNET (672), DrugBank (193), GeneCards (2229), OMIM (31), PharmGkb (295), and TTD (105).

Furthermore, potential breast cancer-related targets could also be identified through bioinformatics approaches. The cutoff criteria were a | log2 (fold change) | > 1 and an adjusted p value < 0.05. We identified 487 DEGs in the tumor samples, including 148 upregulated and 339 downregulated genes (Figure 2A and B). In addition, WGCNA was

Table 1 Detailed Information on the Herbs Included in the Decoction

Chinese name	Latin name	English name	Amount (g)	Place of origin
Chaihu	*Bupleurum chinense* DC.	Bupleurum Root	15	Gansu, China
Baishao	*Paeonia lactiflora* Pall.	White Peony Root	30	Anhui, China
Xiakucao	*Prunella vulgaris* L.	Common Selfheal Fruit-Spike	30	Henan, China
Dangshen	*Codonopsis pilosula* (Franch.) Nannf.	Codonopsis	30	Shanxi, China
Wangbuliuxing	*Gypsophila vaccaria* (L.) Sm.	Cowherb Seed	15	Hebei, China
Huangqin	*Scutellaria baicalensis* Georgi	Chinese Skullcap	15	Shanxi, China

https://doi.org/10.2147/OTT.S486300

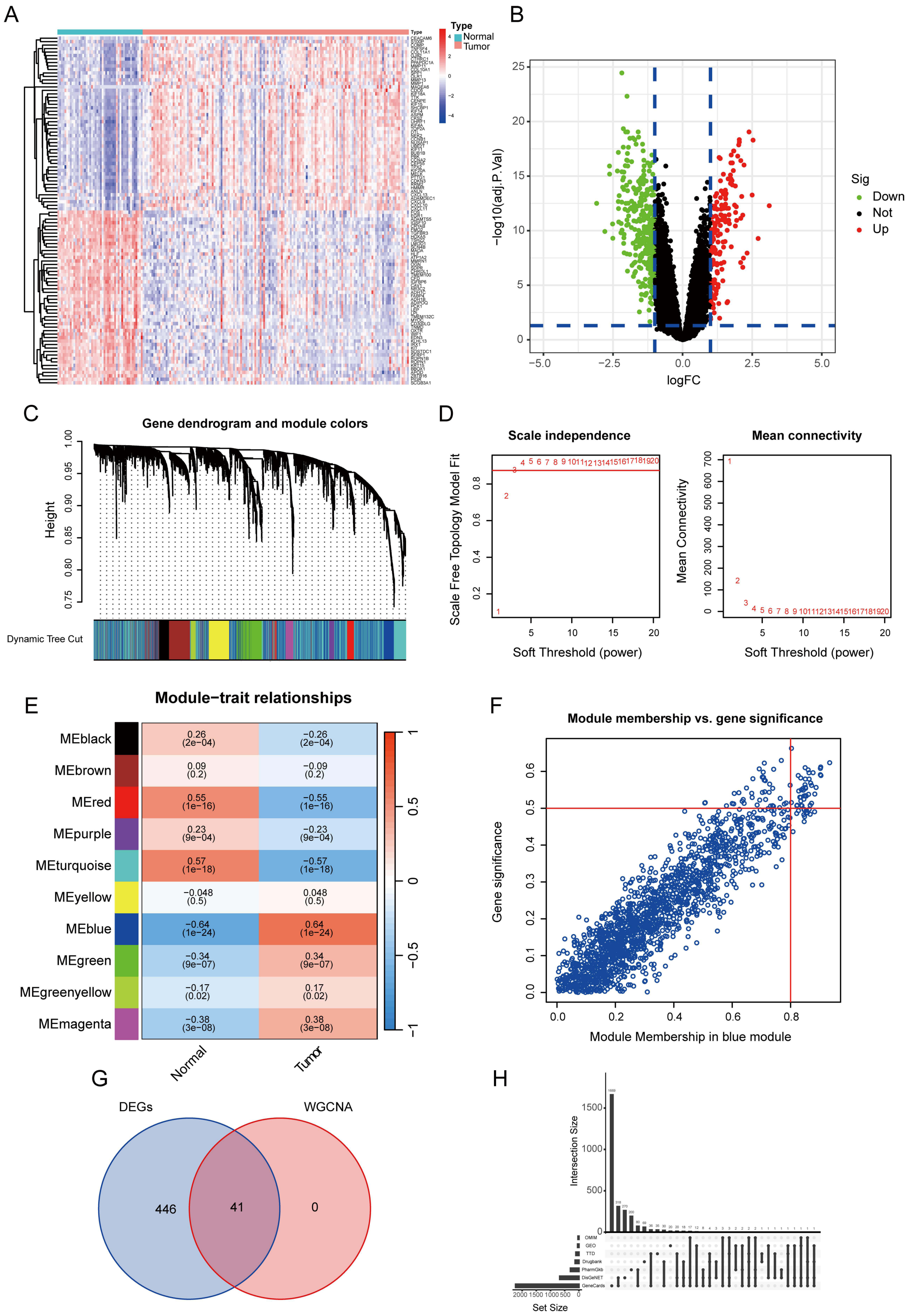


Figure 2 Identification of breast cancer-related targets. (**A**) Heatmap of DEGs. The top 50 upregulated and downregulated genes are shown. (**B**) Volcano plot of DEGs. (**C**) Clustering gene dendrogram based on the dissimilarity measurement. (**D**) Analysis of the scale-free topology model fit for various soft power values. (**E**) The relationship map between the modules and clinical traits. (**F**) Scatter plot of the genes in the blue module. (**G**) Venn diagram showing the genes that overlapped between DEGs and significant genes according to WGCNA. (**H**) Upset plot showing all the breast cancer-related genes from various databases.

conducted to identify the critical gene modules highly related to breast cancer. Distinct gene modules were classified according to the gene cooperative expression data. The genes exhibiting highly coordinated expression were assigned to the same module (Figure 2C). We obtained an appropriate soft power β value of 3 on the basis of scale independence and mean connectivity (Figure 2D) and then divided 5712 genes into 10 modules with at least 60 genes in each module. Subsequently, an analysis of the similarity and adjacency between gene modules and clinical traits (normal and tumor) was performed. Finally, the blue module (1679 genes) demonstrated the strongest correlation ($r = \pm 0.64$, $p = 1e\text{-}24$) with breast cancer (Figure 2E). Considering the defined cutoff standards (MM > 0.8 and GS > 0.5), 41 genes were identified as candidate hub genes significantly enriched in the blue module (Figure 2F). Next, we performed an intersection of the DEGs and the hub genes from WGCNA, which led to the identification of 41 overlapping genes. These genes were considered relevant pathogenic genes in the breast cancer samples extracted from the GEO database (Figure 2G). Subsequently, by merging them with the target genes screened from the above 6 databases, we obtained a union set containing 2832 nonrepetitive targets closely associated with breast cancer (Figure 2H).

Analysis of the Therapeutic Targets of TSGJ and the Associated PPI Network

By intersecting the 2832 breast cancer-related targets and the 223 herbal compound targets of TSGJ, a total of 160 common targets were identified as therapeutic targets of TSGJ for breast cancer (Figure 3A). Through the DisGeNet database, we classified the therapeutic targets into different types, and we found that the target attributes were diverse, indicating that TSGJ exerts multilayer modulatory effects on breast cancer.[32] Among them, enzymes and kinases were the most common (Figure 3B). Then, an herbal compound-gene network was constructed to determine the associations between the targets and associated bioactive compounds of TSGJ (Figure 3C). The network consisted of 222 nodes, representing a total of 66 bioactive compounds and 160 targets. The edges in the network establish meaningful connections between these nodes, indicating the relationships that exist between the bioactive compounds and their associated targets. The 66 bioactive compounds were derived from 6 herbs containing TSGJ, each of which was labeled with a different color.

The PPI network was constructed by introducing the above targets into the STRING database, and the hub genes were selected through network topology analysis on the basis of 6 parameters (DC, EC, BC, CC, LAC, and NC; see Methods). With the CytoNCA plugin, three clusters of target modules were constructed. First, the PPI network of the 160 targets consisted of 160 nodes and 3422 interaction edges (Figure 3D). By filtering the nodes with 6 parameters greater than the average values, 61 targets with 1363 interaction edges were identified (Figure 3E). After further filtering with the same method, 30 targets were screened out as the hub genes with high with high connectivity and significance in the network (Figure 3F).

Construction of Machine Learning Models and Screening of Disease-Predictive Genes

Here, to further identify disease-predictive genes with high diagnostic value, four machine learning models (XGBoost, SVM, RF and GLM) were built based on the expression data of the 30 PPI hub genes in GEO datasets. The residual distributions of the four models were analyzed and plotted through the “DALEX” package. The SVM model had the lowest root mean square residual, and it also presented a relatively lower residual distribution (Figure 4A and B). Subsequently, gene importance analysis was conducted for the four methods, and the top 5 important genes were ranked according to feature importance (Figure 4C). And their feature importance values in each model were listed in Supplementary Table 4. Next, receiver operating characteristic (ROC) curves for the four models were generated using 5-fold cross-validation to comprehensively evaluate their discriminative performance. The SVM model displayed the highest performance, with an area under the curve (AUC) of 0.948, followed by RF at 0.943, XGBoost at 0.902, and GLM at 0.867 (Figure 4D). In summary, the SVM model was considered the optimal model for identifying disease with the highest accuracy. The five most important genes (HIF1A, CASP8, FOS, EGFR, and PPARG) were screened through the SVM model for use as predictors of breast cancer and were utilized to better illustrate the correlations among herbal compounds, predictor genes and breast cancer for subsequent analysis.

https://doi.org/10.2147/OTT.S486300

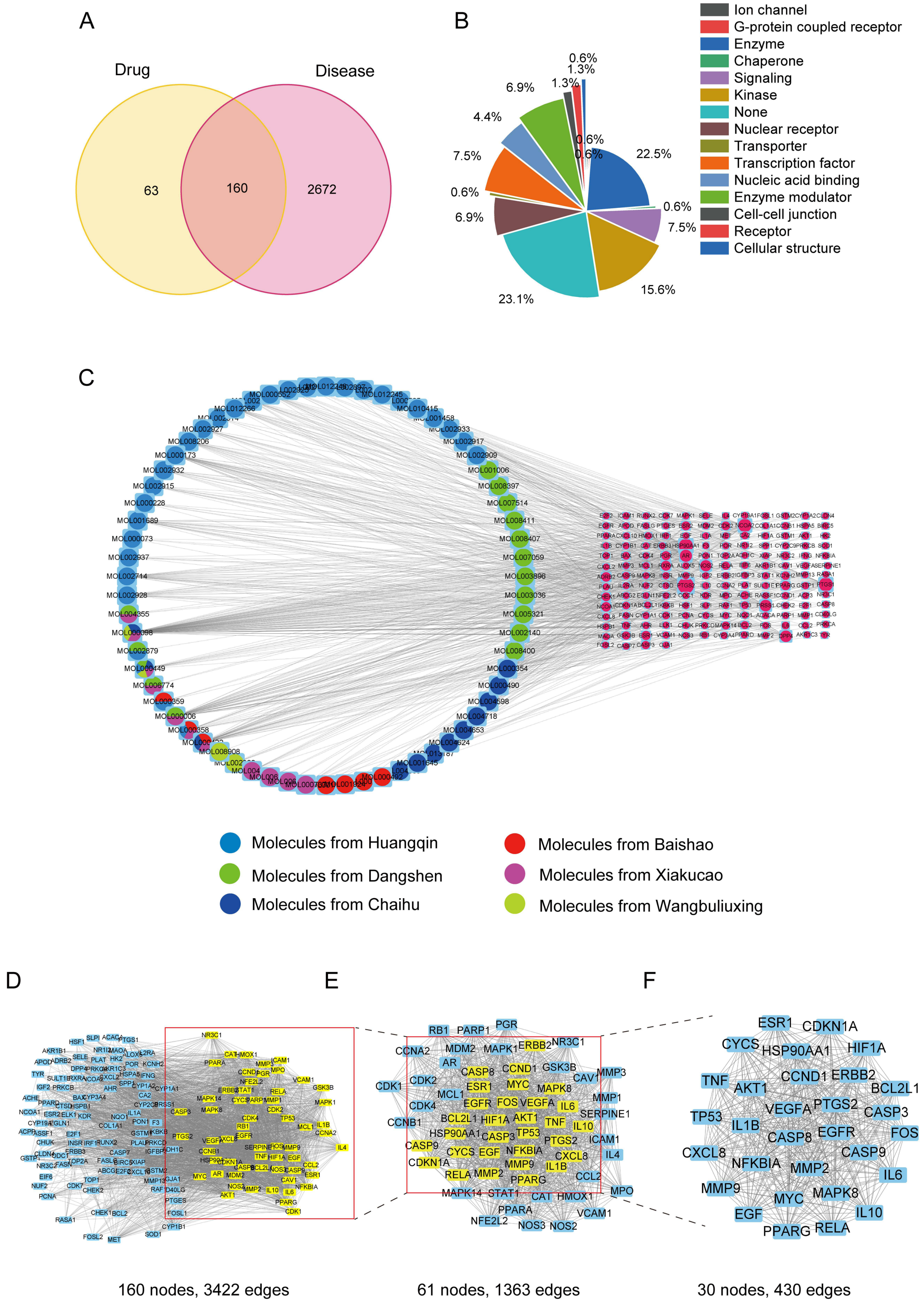


Figure 3 The therapeutic targets of TSGJ. (**A**) Venn diagram of common targets between TSGJ and breast cancer. (**B**) Pie diagram showing the gene classes of the common targets. (**C**) The herbal compound–gene network. The left circles represent the 66 bioactive compounds from the 6 herbs, and the right circles represent the associated genes. Clusters of the PPI network Clusters of the PPI network. (**D**) Cluster 1 with 160 nodes and 3422 interaction edges. (**E**) Cluster 2 with 61 nodes and 1363 interaction edges. (**F**) Cluster 3 with 30 nodes and 430 interaction edges.

https://doi.org/10.2147/OTT.S486300

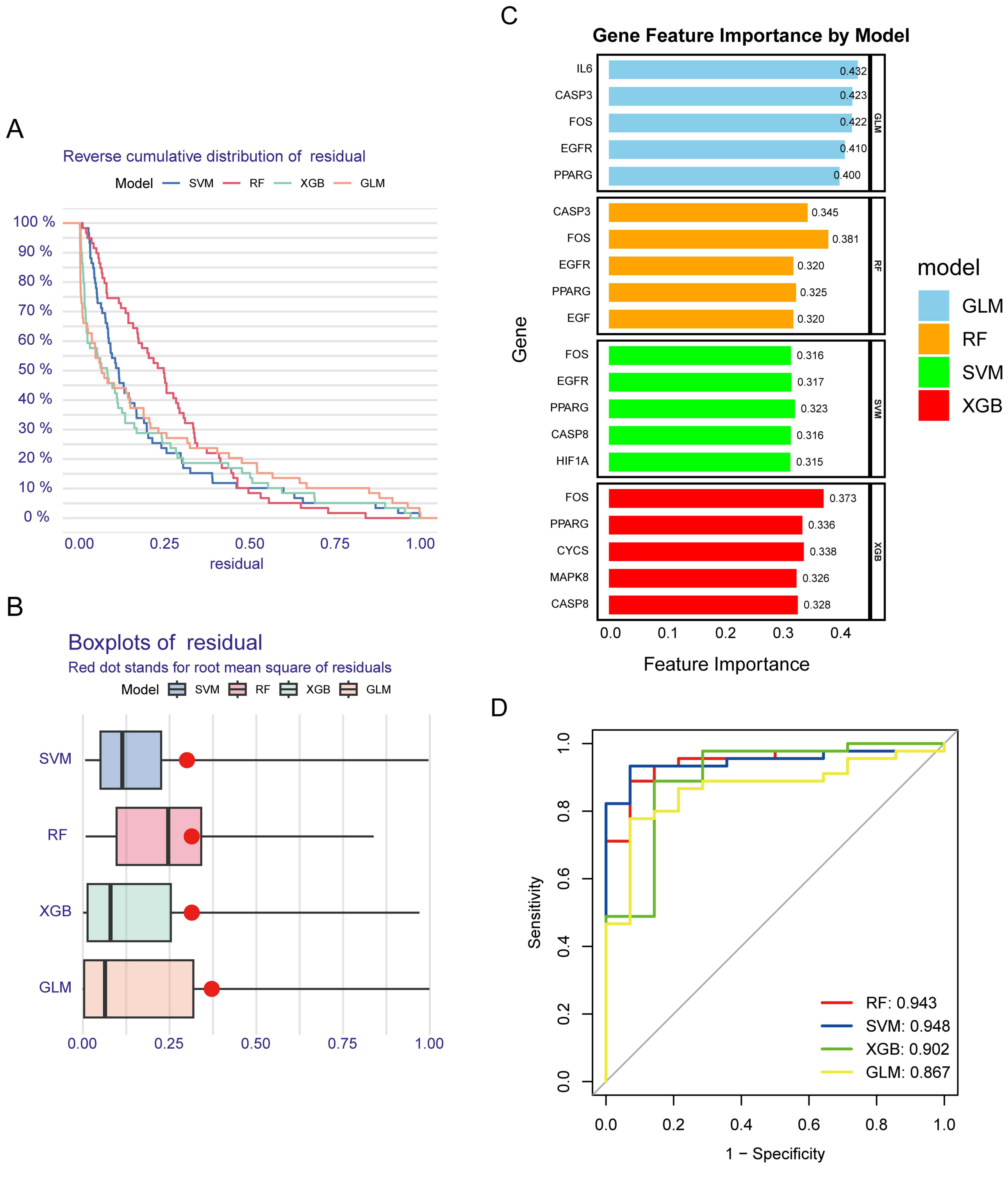


Figure 4 Construction of four machine learning models using GSE139038 and GSE162228 (XGBoost, SVM, RF and GLM) to identify breast cancer disease-predictive genes. (**A**) Reverse cumulative distribution of residuals for the four models. (**B**) Boxplots of the residuals for the four models. The red dots indicate the root mean square of the residuals. (**C**) The top 5 significant genes in the four models. (**D**) ROC curves of the four models using 5-fold cross-validation in the test set.

Expression and Validation Analysis of the Predictor Genes

To better understand the expression of the five predictor genes in the normal and breast cancer samples, we first analyzed the expression levels of each gene in the GEO datasets for the machine learning models. The results showed that all five genes were differentially expressed in the tumor samples, with a statistically significant p value < 0.001. Among them,

https://doi.org/10.2147/OTT.S486300

CASP8 and HIF1A were upregulated in the tumor samples, while FOS, EGFR, and PPARG were downregulated. Furthermore, the ROC curves indicated that FOS exhibited the highest diagnostic accuracy, with an AUC of 0.823. Moreover, the remaining four genes had AUC values above 0.650, making them equally promising predictors (Figure 5A). To more accurately profile the expression levels of the five genes, we also used external GEO datasets (GSE70905 and GSE70947) for validation. The expression data within these two datasets were normalized and merged by the "limma" R package, after which the gene expression data were analyzed. Although there was no notable difference in the expression of HIF1A between normal and tumor samples, the trends in the expression of the remaining four genes were highly consistent with the significant differences in AUC values, confirming the diagnostic value of these genes (Figure 5B).

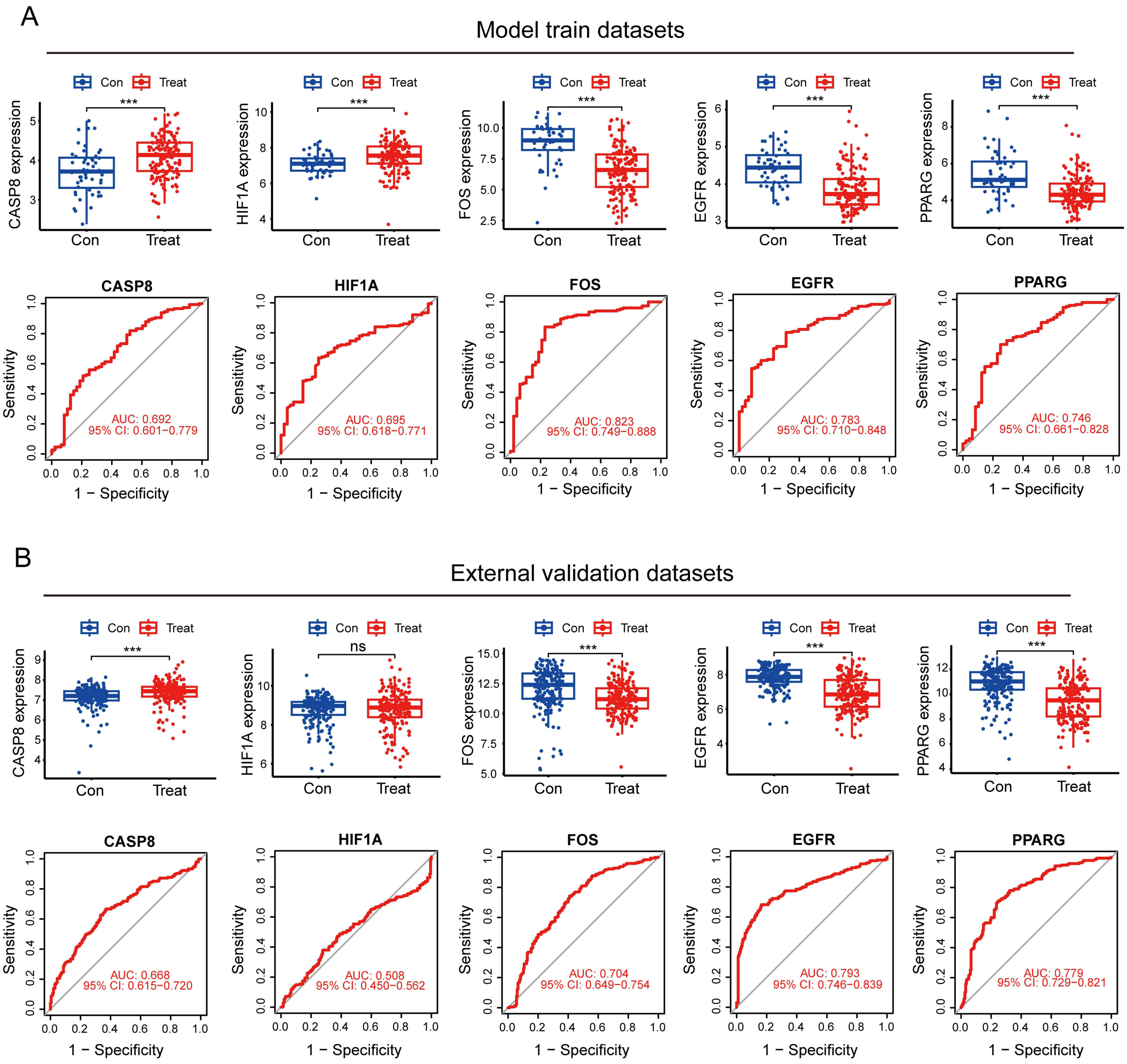


Figure 5 Expression and validation analysis of the five predictor genes. (**A**) Differences in expression levels and ROC curves in the GSE139038 and GSE162228 datasets. (**B**) Differences in expression levels and ROC curves in the GSE70905 and GSE70947 datasets. Con: normal samples, Treat: breast cancer samples.

We Constructed a Nomogram and Explored the Correlation of Gene Expression with 3 Existing Biomarkers

To make the prediction model as simple as possible, we constructed an SVM model and corresponding nomogram model with the five predictor genes to estimate the incidence of breast cancer (Figure 6A). By testing the specific expression values, a separate score interval was assigned to each gene. The sum of the scores for the five genes could be used to predict the risk of breast cancer. The nomogram provided a clearer image to illustrate the effects of multiple genes that act together in disease, and the gene expression levels were intuitively linked to the risk of disease occurrence using numerical values. Then, the predictive effectiveness of the nomogram model was evaluated by calibration curve analysis and DCA in the test set. Through the calibration curve, we observed that the margin between the solid and dashed lines was very small, indicating that the error between disease prediction risk and actual risk was minimal (Figure 6B). The red line for the model genes in the DCA graph was far from the rest of the lines, which confirmed the high accuracy of the model and can serve as valuable support for making clinical treatment decisions (Figure 6C). Additionally, the models validated on the external GEO datasets (GSE70905 and GSE70947) both exhibited satisfactory performance, with high AUC values of 0.811 in the GSE70905 dataset and 0.898 in the GSE70947 dataset, again suggesting that our diagnostic model was equally effective in distinguishing between normal and breast cancer patients (Figure 6D and E).

Breast cancer diagnosis and characterization require the assessment of biomarker proteins present inside or on the surface of tumor cells. The 3 most typical prognostic biomarkers in breast cancer are ER (ESR1), PR (PGR), and HER2 (ERBB2).[38,40] The expression levels of these biomarkers can significantly impact treatment approaches for individuals with breast cancer in clinical settings.[41] Therefore, we also employed another external GSE22820 dataset to determine the associations between the 5 predictor genes and the 3 biomarkers of breast cancer. We concluded that HIF1A and EGFR were negatively correlated with ESR1 (HIF1A, R = −0.39; EGFR, R = −0.65), PGR (HIF1A, R = −0.36; EGFR, R = −0.53), and ERBB2 (HIF1A, R = −0.21; EGFR, R = −0.42), while CASP8, FOS and PPARG were weakly positively correlated with ESR1 expression (CASP8, R = 0.15; FOS, R = 0.17; PPARG, R = 0.13); and PGR (FOS, R = 0.28; PPARG, R = 0.18) (Supplementary Figure 2A-C). The results provided evidence of the exceptional pathological diagnostic value of the 5 predictor genes.

Correlations Between Immune Cell Infiltration and Predictor Genes

Immunotherapy is revolutionizing the treatment of breast cancer by boosting one's own immune system to attack cancer cells more efficiently. Certain immune-modulating drugs can significantly improve the survival rate of cancer patients. Understanding the tumor immune microenvironment and intricate interaction networks is crucial for the development of immunotherapy.[42] In this study, we applied the CIBERSORT algorithm to compare immune cell variations between normal and breast cancer samples. The relative proportions of 22 immune cell infiltrates in the normal control and tumor treatment groups were calculated within the GSE139038 and GSE162228 datasets (Supplementary Figure 3A) and are presented in a boxplot (Supplementary Figure 3B). The results revealed that the infiltration of memory B cells, activated memory T cells, follicular helper T cells, gamma delta T cells, M0 macrophages, M1 macrophages, and neutrophils significantly increased, while the infiltration of naive B cells, plasma cells, CD8 T cells, resting NK cells, and monocytes exhibited the opposite trend. Alterations in the immune microenvironment could be a significant factor contributing to the onset of breast cancer, suggesting new insights for tumor immunotherapy. Analysis of the correlation between the 5 predictor genes and immune cells was also conducted to understand their immunoregulatory functions. The genes were related to up to 16 out of 22 immune cells, including naive B cells, resting NK cells (positively related), neutrophils, and follicular helper T cells (negatively related) (Supplementary Figure 3C). These correlation results suggested that the 5 genes also play a crucial role in regulating the tumor microenvironment of breast cancer patients.

Protein Expression and Clinical Prognostic Value of the Predictor Genes

The protein expression data for normal and breast cancer tissues were downloaded from the HPA database in the form of immunohistochemical images. The images of the 5 identified predictor genes further confirmed the differential expression levels

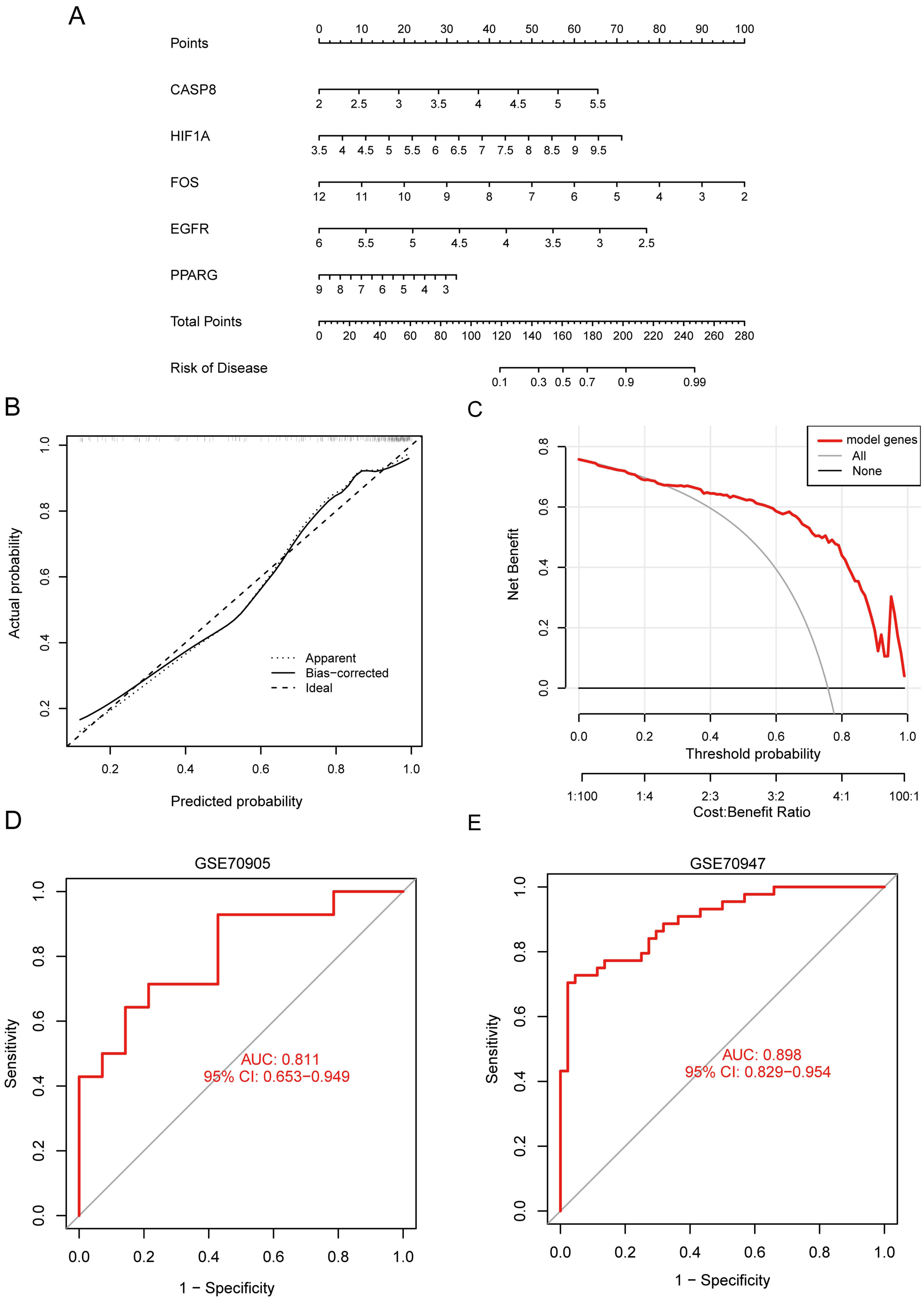


Figure 6 Construction and validation of the nomogram model. (**A**) Nomogram constructed based on the five predictor genes to predict the risk of breast cancer. Analysis of the predictive efficiency of the nomogram model. (**B**) Calibration curve. (**C**) DCA diagram. ROC curves to assess the model accuracy based on GEO datasets. (**D**) GSE70905. (**E**) GSE70947.

between the normal and breast cancer samples; moreover, the expression levels of CASP8 and HIF1A were obviously greater in the tumor samples than in the normal samples, and the expression levels of FOS, EGFR, and PPARG were slightly lower (Supplementary Figure 4A). This result was consistent with the mRNA expression data shown in Figure 7. To assess the prognostic values of the 5 predictor genes, the Kaplan–Meier plotter database and TCGA database were used to analyze the relationships between gene expression and survival rate and disease progression, respectively. By evaluating the impacts of the expression of these 5 genes on the survival rate of 2976 patients via the KM plot database, we observed that increased expression of CASP8, HIF1A, and EGFR and decreased expression of FOS and PPARG led to a poor survival prognosis for breast cancer patients (Supplementary Figure 4B). For a more in-depth analysis, we investigated the relationships between the mRNA expression of 5 predictor genes and various clinicopathological features, including age, tumor stage and TNM stage, in the TCGA database. The TCGA-BRCA cohort for breast cancer encompasses 1085 clinical records of patients with breast cancer. Gene expression was strongly related to prognosis, among which CASP8 could affect tumor stage and T stage; FOS was related to both tumor stage and TNM stage; EGFR was related to age and T stage; and PPARG was related to age, tumor stage, and TN stage (Supplementary Figure 5A-E). Collectively, the data presented above strongly indicated that the key predictor genes identified through integrative network pharmacology and machine learning approaches play a crucial role in the onset and progression of breast cancer. This preliminary confirmation underscores the reliability of the predictive outcomes derived from the approaches. In subsequent steps, we could further determine the main herbal compounds through constructing an interaction network of the 5 predictor genes and their corresponding compounds while exploring how they are involved in the development of breast cancer.

Enrichment Analysis

To explore the biological functions of the relevant target genes and the involved disease-related signaling pathways, we performed Gene Ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) analyses on the 160 genes at the intersection of TSGJ and breast cancer genes and screened genes for which $p < 0.05$ was considered to indicate a significant difference. Among them, GO-enriched terms were mainly related to reactive oxygen species, response to oxidative stress, membrane rafts, membrane microdomains, DNA-binding transcription factor binding, and RNA polymerase II-specific DNA-binding transcription factor binding, which indicated that TGSJ could affect cell proliferation,

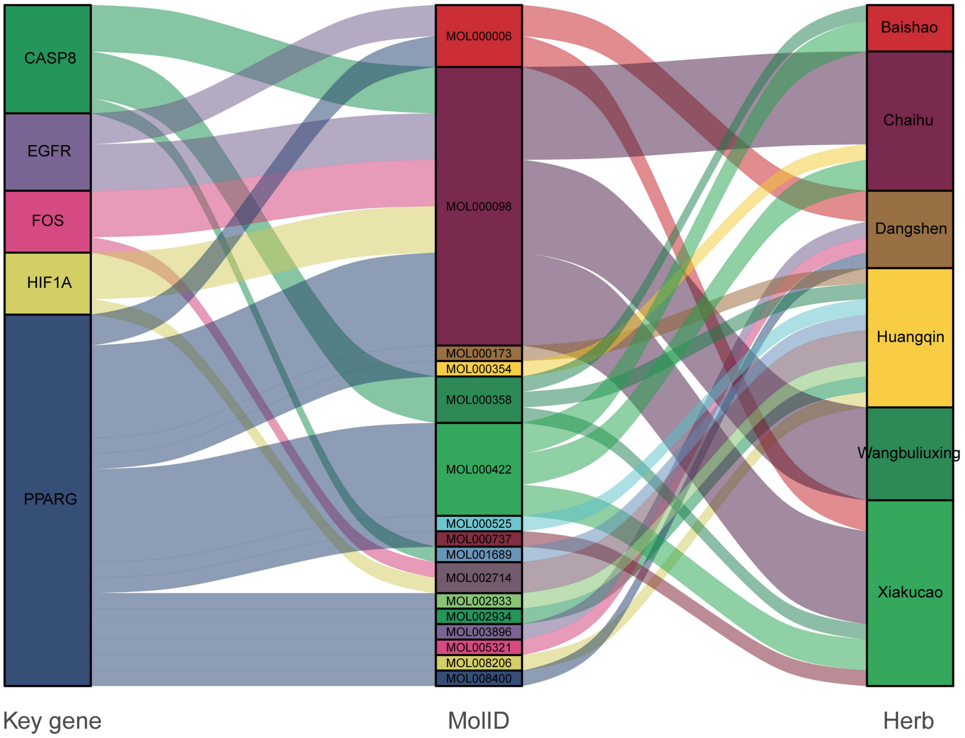


Figure 7 The Sankey plot of the 5 predictor genes-bioactive compounds-herbs network. Left column represents the identified 5 predictor genes, middle column represents their corresponding bioactive compounds, and the right column represents the herb origins.

metabolism, differentiation, and survival by regulating ROS and participated in cell activities by affecting key proteins, DNA, and RNases (Supplementary Figure 6A). The level of ROS is associated with the uncontrolled proliferation and high metabolic rate of cancer cells and is related to maintaining tumor phenotypes.[43] A total of 151 out of 160 genes were mapped to 174 KEGG pathways, and Supplementary Figure 6B illustrates the top 15 pathways that exhibited significant enrichment. The IL-17 signaling pathway has a crucial impact on the development and progression of breast cancer. Targeting this pathway or its specific downstream mediators using chemotherapy drugs or small interfering RNA (siRNA) interference represents a promising and innovative therapeutic approach for inhibiting this disease.[44] The TNF signaling pathway can induce the activation of multiple cancer-related pathways, such as the PI3K/Akt and AKT/NF-κB signaling pathways, thus regulating the progression and metastasis of breast cancer.[45,46]

In addition, we constructed a relationship plot of the 5 identified predictor genes with their corresponding top 10 enriched GO BP and KEGG terms (Supplementary Figure 7). Multiple pathways, such as the MAPK signaling pathway, apoptosis pathway, HIF-1 signaling pathway, IL-17 signaling pathway, PD-L1 expression pathway, PD-1 checkpoint pathway in cancer, PI3K-Akt signaling pathway, PPAR signaling pathway, and TNF signaling pathway, can affect cell survival, apoptosis, proliferation, etc. Regulated activity of these pathways is often associated with cancer progression or drug resistance.[47–49] We found that the 5 predictor gene-enriched signaling pathways were more closely related to the progression of breast cancer, which further proved that the 5 identified predictor genes were more likely to play significant roles in the prevention and treatment of breast cancer by TGSJ.

Key Gene–Compound–Herb Network Construction and Molecular Docking

A key gene-compound-herb network was constructed to visualize the correlations between the 5 predictor genes, bioactive compounds, and herbs to identify the major bioactive compounds. The Sankey plot revealed that the 5 predictor genes were related to 16 bioactive compounds from the 6 herbs (Figure 7). Detailed information on the corresponding 16 bioactive compounds is shown in Table 2. In particular, MOL000098 (quercetin) was associated with all 5 genes, MOL000006 (luteolin) and MOL002714 (baicalein) corresponded to 2 genes, while the other compounds were linked to only 1 gene each. Molecular docking was performed to calculate the binding energy of the 5 predictor genes and their related bioactive compounds. The crystal structures with improved resolutions of the 5 predictor genes were downloaded from the PDB database as follows: CASP8A (PDB ID: 3KJQ), HIF1A (PDB ID: 2ILM), FOS (PDB ID: 2WT7), EGFR (PDB ID: 5EM6), and PPARG (PDB ID: 3ADX). AutoDock Vina software was then used to perform the protein–ligand docking and calculate the binding energy. In general, the binding energy reflects the binding strength, with a decrease in energy signifying an enhanced binding capability.[32] A binding energy below 0 suggests the potential

Table 2 Detailed Information on the Bioactive Molecules Associated with Key Genes

MolID	MolName	PubChem CID	Chemical formula	Herbs of origin
MOL000006	Luteolin	5280445	$C_{15}H_{10}O_6$	Dangshen, Xiakucao
MOL000098	Quercetin	5280343	$C_{15}H_{10}O_7$	Chaihu, Wangbuliuxing, Xiakucao
MOL000173	Wogonin	5281703	$C_{16}H_{12}O_5$	Huangqin
MOL000354	Isorhamnetin	5281654	$C_{16}H_{12}O_7$	Chaihu
MOL000358	Beta-sitosterol	222284	$C_{29}H_{50}O$	Baishao, Huangqin, Xiakucao
MOL000422	Kaempferol	5280863	$C_{15}H_{10}O_6$	Baishao, Chaihu, Xiakucao
MOL000525	Norwogonin	5281674	$C_{15}H_{10}O_5$	Huangqin
MOL000737	Morin	5281670	$C_{15}H_{10}O_7$	Xiakucao
MOL001689	Acacetin	5280442	$C_{16}H_{12}O_5$	Huangqin
MOL002714	Baicalein	5281605	$C_{15}H_{10}O_5$	Huangqin
MOL002933	5,7,4'-Trihydroxy-8-methoxyflavone	5322078	$C_{16}H_{12}O_6$	Huangqin
MOL002934	Neobaicalein	124211	$C_{19}H_{18}O_8$	Huangqin
MOL003896	7-Methoxy-2-methyl isoflavone	354368	$C_{17}H_{14}O_3$	Dangshen
MOL005321	Frutinone A	441965	$C_{16}H_8O_4$	Dangshen
MOL008206	Moslosooflavone	188316	$C_{17}H_{14}O_5$	Huangqin
MOL008400	glycitein	5317750	$C_{16}H_{12}O_5$	Dangshen

for the ligand molecule to bind to the receptor target in a spontaneous manner, while a binding energy lower than −5 indicates favorable and strong binding affinity.[50,51] The docking scores were exhibited via heatmap, showing the binding energies of all 22 protein–ligand pairs were all below −5, indicating that the compounds could firmly bind to the protein (Figure 8A). Quercetin, luteolin and baicalein were identified as the core compounds of TSGJ as they had strong affinities with up to 2 key targets. Their binding configurations with associated targets were visualized by PyMOL

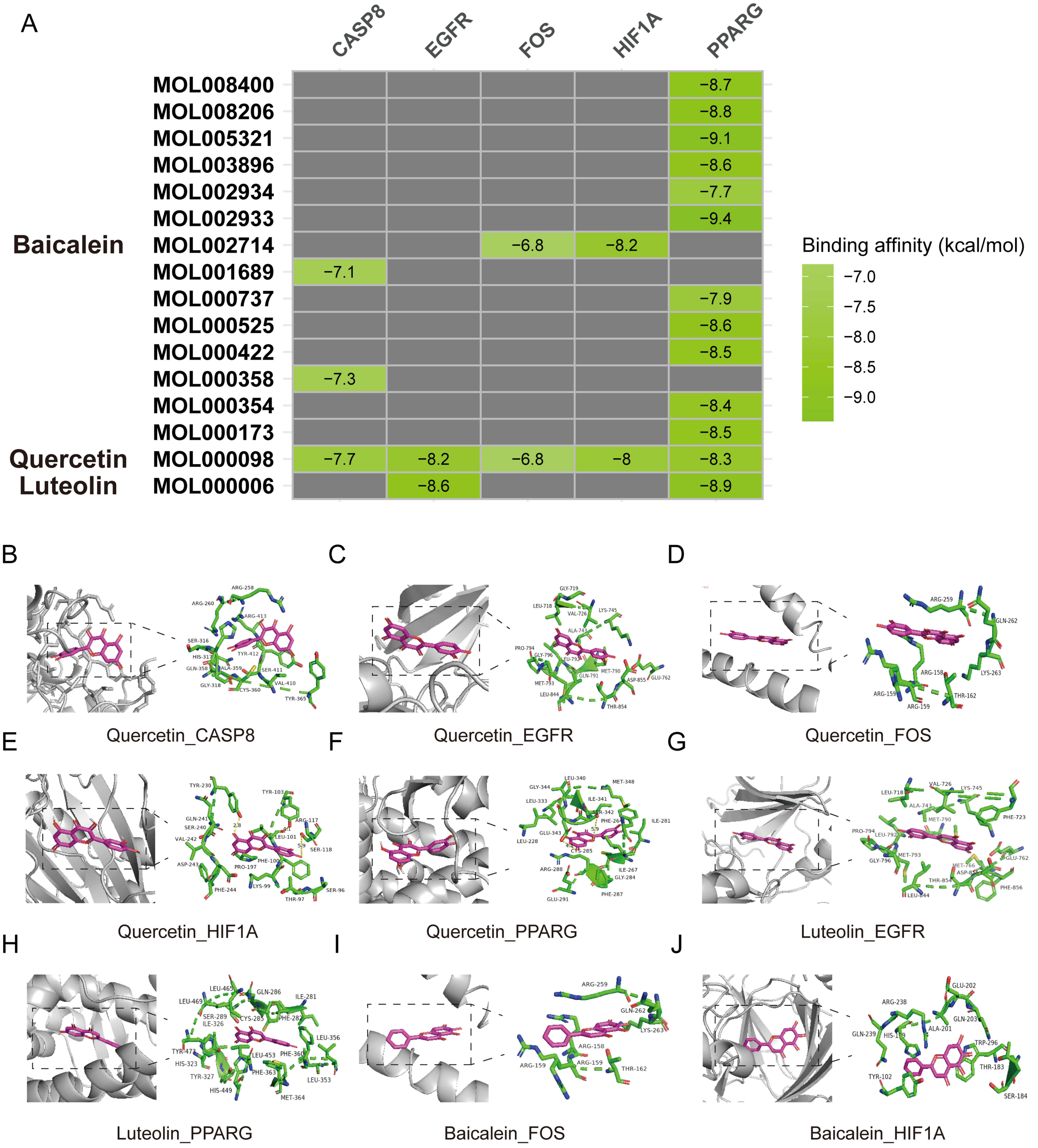


Figure 8 Molecular docking between the 5 predictor genes and their associated bioactive compounds. (**A**) A heatmap illustrates the binding affinities obtained from the molecular docking of 5 key targets and associated molecules. The core compounds which could bind to at least 2 targets were labelled as Baicalein, Quercetin, Luteolin. The conformation of the core compounds bound to their associated targets. (**B**) Quercetin to CASP8. (**C**) Quercetin to EGFR. (**D**) Quercetin to FOS. (**E**) Quercetin to HIF1A. (**F**) Quercetin to PPARG. (**G**) Luteolin to EGFR. (**H**) Luteolin to PPARG. (**I**) Baicalein to FOS. (**J**) Baicalein to HIF1A.

software (Figure 8B-8J). The compounds were found to exhibit multiple types of interactions with the proteins, primarily including Van der Waals, Hydrogen Bond, Pi-Alkyl, Pi-Sigma, and Pi-Pi Stacked interactions (Supplementary Figure 8A-I). And the core compounds were selected to perform cellular validation to detect their effects on the breast cancer cell viability and associated mRNA expressions.

Cellular Validation

The HPLC results confirmed that quercetin, luteolin, and baicalein are present at measurable concentrations in the TSGJ decoction (Supplementary Figure 9), further validating the rationality of the component analysis conducted via network pharmacology. Specifically, in a 2 mg/mL TSGJ solution, approximately 23 μg/mL of quercetin and luteolin, and 15 μg/mL of baicalein were detected (Supplementary Table 5). Subsequently, we conducted cell experiments to evaluate the anticancer effects using both the equivalent TSGJ solution and a composite solution of the three compounds (QLB solution) at matched concentrations. TSGJ solution and QLB solution exhibited significant cytotoxicity against MDA-MB-231 cells in a dose-dependent manner in a similar trend, indicating that the three compounds are the main active components against breast cancer of TSGJ (Figure 9A). At lower concentrations (TSGJ ≤ 2 mg/mL, QLB ≤ 60 μg/mL), the compounds showed minimal inhibitory effects on MCF-10A, while inhibiting MDA-MB-231 by over 50%. At higher concentrations, they mildly inhibited MCF-10A, but cell viability remained above 60%, indicating low toxicity to non-cancerous cells. These results suggest the compounds have selective activity, with limited effects on non-cancerous cells at therapeutic concentrations (Figure 9B). Further analysis of the mRNA expression levels revealed CASP8, EGFR, FOS, and HIF1A were highly expressed in MDA-MB-231, while PPARG was less expressed in the tumor cells. And their expression levels could be regulated by the TSGJ and QLB solutions (Figure 9C-G). These results indicated that the three identified compounds (quercetin, luteolin, and baicalein) could collectively modulate key targets and induce cytotoxicity, contributing to the therapeutic effect of TSGJ on breast cancer.

Discussion

Breast cancer (BRCA) is the most frequently diagnosed malignancy in women worldwide and is the leading cause of cancer-related death among women.[52] Due to limited treatment, its poor prognosis and mortality rates have been increasing.[53] TCM has gained traction as a complementary therapy for breast cancer and is known for reducing toxicity, improving prognosis, and enhancing survival. Its complex formulations act on multiple targets, leading to a broad range of effects through intricate molecular networks. Network pharmacology has been key in understanding TCM mechanisms. Employing machine learning to identify critical disease targets offers a forward-looking method for TCM research, enabling predictions from large-scale biological data.[54]

The therapeutic use of TSGJ decoction, a traditional TCM formula for breast cancer, has historically been noted. Despite its observed anticancer effects, its specific mechanisms and primary active components have remained largely unexplored. Our study employed network pharmacology to create a detailed biological network linking herbal compounds to disease targets and pathways. We used machine learning to identify key diagnostic targets for breast cancer and their corresponding active compounds, offering fresh insights into TCM's potential in disease prevention and treatment.

Drawing on the TCMSP database, we identified 87 active compounds from 6 herbs in the TSGJ concoction and mapped 223 related targets. Additionally, our analysis incorporated RNA-sequencing data from the GEO database (Table 3) and used DEG analysis and WGCNA to refine the identification of BRCA-related pathogenic genes. We identified 160 therapeutic targets for TSGJ in breast cancer, which is indicative of TCM's multifaceted approach to disease modulation.

A PPI network refers to a network composed of interactions among proteins within a living organism.[55] Biological processes within living organisms are often complex and rely on the close connection and collective action of multiple proteins.[56] Active ingredients are also likely to act on more than a single target. We imported 160 therapeutic targets into the STRING database to construct the PPI network, and through network topology analysis and filtering, we ultimately identified 30 core targets. These targets demonstrated a high degree value in the central network of the PPI network, establishing a dominant position. Consequently, they may play a crucial role in the anti-BRCA effect of TSGJ.

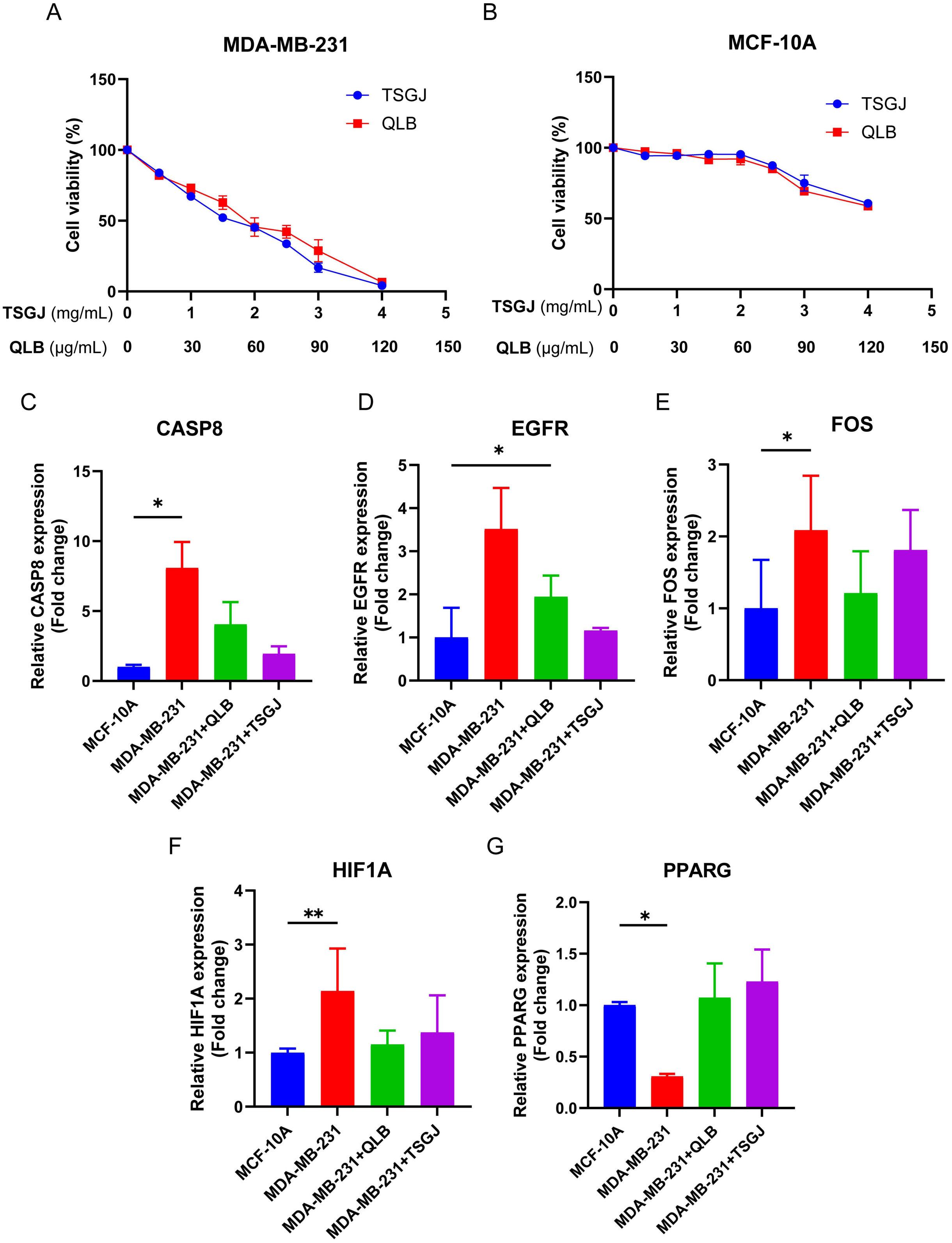


Figure 9 Cellular effects of the 3 mixed core compounds including Quercetin, Luteolin, Baicalein (QLB) and TCM formula (TSGJ). Cell viability after treated with different concentrations of QLB and TSGJ for 48h in (**A**) MDA-MB-231, (**B**) MCF-10A. RT-qPCR detection of the 5 key genes mRNA expressions treated with the QLB and TSGJ for 48h. (**C**) CASP8, (**D**) EGFR, (**E**) FOS, (**F**) HIF1A, (**G**) PPARG. A significance threshold was set at $p < 0.05$ to determine statistical significance, with * indicating $p < 0.05$ and ** indicating $p < 0.01$.

https://doi.org/10.2147/OTT.S486300

Table 3 Datasets Used in This Study

Datasets	Samples	Usage
GSE139038	24 normal, 41 tumor samples	DEGs, WGCNA, model training, immune infiltration analysis
GSE162228	24 normal, 109 tumor samples	
GSE70905	47 normal, 47 tumor samples	Nomo model validation, expression validation
GSE70947	148 normal, 148 tumor samples	
GSE22820	10 normal, 176 tumor samples	3 biomarkers correlation
TCGA-BRCA	112 normal, 1105 tumor samples	Clinical prognostic correlation

Nevertheless, the PPI network has limitations because it only includes densely connected proteins without delving into the associations between these proteins and the occurrence of disease.[57] Therefore, there is an urgent need for further data analysis and interpretation. Recently, machine learning algorithms based on RNA sequencing data have been progressively applied for the statistical analysis of disease signature genes and have shown excellent predictive results.[58–60] In this study, we established and compared four machine learning models (RF, SVM, GLM and XGBoost) with the expression profiles of hub genes from PPIs. The SVM model displayed the highest performance, with an AUC of 0.948, indicating that it has the highest predictive efficacy in the prediction of BRCA diagnostic genes. Subsequently, five important predictive genes (HIF1A, CASP8, FOS, EGFR, and PPARG) were screened through SVM, and we constructed a nomogram to estimate the occurrence of BRCA. HIF1A serves as a fundamental key regulator of hypoxia signaling and is associated with the brain metastasis of BRCA[61] and the early recurrence of ER-type BRCA.[62] CASP8, recognized as one of the initial low penetrance loci, has been linked to the susceptibility of BRCA in investigations focusing on candidate genes.[63] FOS can play a role in governing both human breast biology and the development and progression of cancer.[64] The signaling pathway involving EGFR has been extensively studied in the progression of cancer. Fundamental alterations in EGFR signaling occur during the invasion, dissemination, and eventual metastasis of BRCA.[1] PPARG is associated with poor prognosis in BRCA patients because it regulates tumor-infiltrating cells within the tumor microenvironment through various pathways.[31] Thus, these five disease-related genes have great potential as predictor genes for BRCA prediction, prevention, and treatment. The five genes were differentially expressed in the BRCA samples and exhibited high diagnostic accuracy. The nomogram model demonstrated exceptional predictive efficacy, underscoring its significance for clinical applications. In addition, as three biomarkers (ESR1, PGR, and ERBB2) were strongly associated with the occurrence, progression and prognosis of BRCA, we also performed correlation analysis between the biomarkers and five predictor genes with another external dataset. The results showed significant correlations, providing evidence of the exceptional pathological diagnostic value of the 5 predictor genes.

Furthermore, our research revealed a significant link between five key predictor genes and the immune system, with these genes influencing 16 out of 22 types of immune cells, suggesting a role for these genes in regulating the tumor immune environment during BRCA progression. Clinical prognostic evaluations also revealed that these genes impact the overall survival of BRCA patients by affecting various clinical features, making them reliable indicators of BRCA onset and progression. Although the pathway enrichment of 160 therapeutic targets appears complex, protein–protein interaction (PPI) analysis and support vector machine (SVM) model screening revealed that the pathways related to the top 5 predictor genes are closely related to BRCA progression. This finding underscores the potential importance of TSGJ in preventing and treating BRCA. Linking these genes to their corresponding active ingredients in TSGJ, particularly quercetin, luteolin, and baicalein, we constructed a network that enhances our understanding of the function of TSGJ in BRCA prevention and treatment. Molecular docking assays further confirmed the strong binding affinity between these components and the predictor genes, reinforcing their significance in the efficacy of TSGJ. The cytotoxic effects of quercetin, luteolin, and baicalein on breast cancer cells, as well as their regulatory effects on key targets, were also confirmed through MTT and RT-qPCR cell experiments. These findings demonstrate the therapeutic effects of TSGJ on breast cancer.

In summary, our study adopted a new strategy to synthesize network pharmacology and machine learning data through vast data from multiple sources, focusing on the 87 active ingredients of TSGJ decoction and their effects on 160 BRCA-related genes. The sheer volume of data presents a challenge in pinpointing crucial drug therapy mechanisms. By integrating PPI network analysis with machine learning, we identified 5 key genes linked to BRCA progression and identified the 3 most impactful components associated with these genes. This provides a new, comprehensive method for identifying the core components and mechanisms of TCMs for disease treatment, while experimental validation exhibited a more complete picture.

Conclusion

This study applied network pharmacology, machine learning, and experimental validation to identify key components, targets, and mechanisms of TSGJ decoction against breast cancer. Multiple analysis highlighted critical compounds (quercetin, luteolin, baicalein) and five key predictive targets (HIF1A, CASP8, FOS, EGFR, PPARG) involved in breast cancer progression, immune regulation, and prognosis. This research supports the therapeutic potential of TSGJ decoction in breast cancer prevention and treatment.

Data Sharing Statement

All data analyzed during this study are included in the websites mentioned above.

Ethics Approval and Consent to Participate

This study was approved by the Ethics Committee of Affiliated Yongkang First People's Hospital (Approval No. YKSDYRMYYEC2024-LW-HS-077-01) and adhered to the principles outlined in the Declaration of Helsinki. Informed consent was obtained from all participants prior to their inclusion in the study.

Consent for Publication

Not applicable.

Acknowledgments

We acknowledge Biorender (https://app.biorender.com/) for generation of the graphical abstract (Ying, H. (2025) https://BioRender.com/t36p701).

Funding

This study was supported by funding from the Jinhua City Science and Technology Project (No.2024-4-308), K. Albin Johansson Stiftelse Foundation (grant no. 2023-12), the Finnish Cultural Foundation (grant no. 00241240), the Maud Kuistila Memorial Foundation sr (grant no. 2024-0147B), and the China Scholarship Council (grant no. 2020 I78).

Disclosure

The authors have no competing interests to declare in this work.

References

1. Ali R, Wendt MK. The paradoxical functions of EGFR during breast cancer progression. *Signal Transduct Targeted Ther*. 2017;2(1):16042. doi:10.1038/sigtrans.2016.42
2. Arnold M, Morgan E, Rumgay H, et al. Current and future burden of breast cancer: global statistics for 2020 and 2040. *Breast*. 2022;66:15–23. doi:10.1016/j.breast.2022.08.010
3. Huang S, Parekh V, Waisman J, et al. Cutaneous metastasectomy: is there a role in breast cancer? A systematic review and overview of current treatment modalities. *J Surg Oncol*. 2022;126(2):217–238. doi:10.1002/jso.26870
4. Ni X, Han JQ, X. Y, et al. Percutaneous CT-guided microwave ablation as maintenance after first-line treatment for patients with advanced NSCLC. *Onco Targets Ther*. 2015;8:3227–3235. doi:10.2147/ott.S90528
5. Diana A, Carlino F, Franzese E, et al. Early triple negative breast cancer: conventional treatment and emerging therapeutic landscapes. *Cancers*. 2020;12(4):819. doi:10.3390/cancers12040819
6. Li -D-D, Tao Z-H, Wang B-Y, et al. Apatinib plus vinorelbine versus vinorelbine for metastatic triple-negative breast cancer who failed first/second-line treatment: the NAN trial. *Npj Breast Cancer*. 2022;8(1):110. doi:10.1038/s41523-022-00462-6
7. Ouyang J, Xie A, Zhou J, et al. Minimally invasive nanomedicine: nanotechnology in photo-/ultrasound-/radiation-/magnetism-mediated therapy and imaging. *Chem. Soc. Rev*. 2022;51(12):4996–5041. doi:10.1039/d1cs01148k
8. Leong F, Hua X, Wang M, et al. Quality standard of traditional Chinese medicines: comparison between European Pharmacopoeia and Chinese Pharmacopoeia and recent advances. *Chin Med*. 2020;15(1):76. doi:10.1186/s13020-020-00357-3

9. Fu K, Xu M, Zhou Y, et al. The status quo and way forwards on the development of Tibetan medicine and the pharmacological research of Tibetan materia Medica. *Pharmacol Res*. 2020;155:104688. doi:10.1016/j.phrs.2020.104688
10. Dai X, Feng J, Chen Y, et al. Traditional Chinese Medicine in nonalcoholic fatty liver disease: molecular insights and therapeutic perspectives. *ChinMed*. 2021;16(1):68. doi:10.1186/s13020-021-00469-4
11. Shen H, Cao J, Zhang L, et al. Classification research of TCM pulse conditions based on multi-label voice analysis. *J Tradit Chin Med*. 2024;11 (2):172–179. doi:10.1016/j.jtcms.2024.03.008
12. Law BY-K, Mo J-F, Wong VK-W. Autophagic effects of Chaihu (dried roots of Bupleurum Chinense DC or Bupleurum scorzoneraefolium WILD). *Chin Med*. 2014;9(1):21. doi:10.1186/1749-8546-9-21
13. Xu T, Wang Q, Liu M. A network pharmacology approach to explore the potential mechanisms of huangqin-baishao herb pair in treatment of cancer. *Med Sci Monit*. 2020;26:e923199. doi:10.12659/msm.923199
14. Wang T, Fu X, Wang Z. Danshen formulae for cancer: a systematic review and meta-analysis of high-quality randomized controlled trials. *Surgery*. 2019;2019(1):2310639. doi:10.1155/2019/2310639
15. Guo J, Lou M-P, L-L. H, et al. Uncovering the pharmacological mechanism of the effects of the Banxia-Xiakucao Chinese Herb Pair on sleep disorder by a systems pharmacology approach. *Sci Rep*. 2020;10(1):20454. doi:10.1038/s41598-020-77431-1
16. Feng L, Zhang X, Hua H, et al. Vaccaria segetalis extract can inhibit angiogenesis J. *Asian Biomed*. 2017;6(5):683–692. doi:10.5372/1905-7415.0605.108
17. Casas AI, Hassan AA, Larsen SJ, et al. From single drug targets to synergistic network pharmacology in ischemic stroke. *Proc Natl Acad Sci*. 2019;116(14):7129–7136. doi:10.1073/pnas.1820799116
18. Wang Y, Yang SH, Zhong K, et al. Network pharmacology-based strategy for the investigation of the anti-obesity effects of an ethanolic extract of Zanthoxylum bungeanum maxim. *Front Pharmacol*. 2020:11. doi:10.3389/fphar.2020.572387.
19. Liu L, Zhang B, Zhou Z, et al. Integrated network pharmacology and experimental validation approach to investigate the mechanisms of radix rehmanniae praeparata - angelica sinensis - radix achyranthis bidentatae in treating knee osteoarthritis. *Drug Des Devel Ther*. 2024;18:1583–1602. doi:10.2147/dddt.S455006
20. He Y, Zheng X, Sit C, et al. Using association rules mining to explore pattern of Chinese medicinal formulae (prescription) in treating and preventing breast cancer recurrence and metastasis. *J Transl Med*. 2012;10(1):S12. doi:10.1186/1479-5876-10-S1-S12
21. Chan P-W, Chiu J-H, Huang N, et al. Influence of traditional chinese medicine on medical adherence and outcome in estrogen receptor (+) breast cancer patients in Taiwan: a real-world population-based cohort study. *Phytomedicine*. 2021;80:153365. doi:10.1016/j.phymed.2020.153365
22. Huang C-H, Chang H-P, S-Y. S, et al. Traditional Chinese medicine is associated with a decreased risk of heart failure in breast cancer patients receiving doxorubicin treatment. *J Ethnopharmacol*. 2019;229:15–21. doi:10.1016/j.jep.2018.09.030
23. Tran A, Walsh CJ, Batt J, et al. A machine learning-based clinical tool for diagnosing myopathy using multi-cohort microarray expression profiles. *J Transl Med*. 2020;18(1):454. doi:10.1186/s12967-020-02630-3
24. Zhu E, Shu X, Z. X, et al. Screening of immune-related secretory proteins linking chronic kidney disease with calcific aortic valve disease based on comprehensive bioinformatics analysis and machine learning. *J Transl Med*. 2023;21(1):359. doi:10.1186/s12967-023-04171-x
25. Jin M, Ren W, Zhang W, et al. Exploring the underlying mechanism of shenyankangfu tablet in the treatment of glomerulonephritis through network pharmacology, machine learning, molecular docking, and experimental validation. *Drug Des Devel Ther*. 2021;15:4585–4601. doi:10.2147/dddt.S333209
26. Ma S, Liu J, Li W, et al. Machine learning in TCM with natural products and molecules: current status and future perspectives. *ChinMed*. 2023;18 (1):43. doi:10.1186/s13020-023-00741-9
27. Yang J, Tian S, Zhao J, et al. Exploring the mechanism of TCM formulae in the treatment of different types of coronary heart disease by network pharmacology and machining learning. *Pharmacol Res*. 2020;159:105034. doi:10.1016/j.phrs.2020.105034
28. Gao Y, Ji W, M. L, et al. Systemic pharmacological verification of Guizhi Fuling decoction in treating endometriosis-associated pain. *J Ethnopharmacol*. 2022;297:115540. doi:10.1016/j.jep.2022.115540
29. Zhang Z, Kim BS, Han W, et al. Identifying oxidized lipid metabolism-related LncRNAs as prognostic biomarkers of head and neck squamous cell carcinoma. *J Personalized Med*. 2023;13(3):488. doi:10.3390/jpm13030488
30. Piñero J, Ramírez-Anguita JM, Saüch-Pitarch J, et al. The DisGeNET knowledge platform for disease genomics: 2019 update. *Nucleic Acids Res*. 2019;48(D1):D845–D855. doi:10.1093/nar/gkz1021
31. Zhang S, Niu Q, Tong L, et al. Identification of the susceptible genes and mechanism underlying the comorbid presence of coronary artery disease and rheumatoid arthritis: a network modularization analysis. *BMC Genomics*. 2023;24(1):411. doi:10.1186/s12864-023-09519-7
32. Shi S, Zhao S, Tian X, et al. Molecular and metabolic mechanisms of bufalin against lung adenocarcino ma: new and comprehensive evidences from network pharmacology, metabolomics and molecular biology experiment. *Comput. Biol. Med.* 2023;157:106777. doi:10.1016/j.compbiomed.2023.106777
33. Taşcı B, Omar A, Ayvaz S. Remaining useful lifetime prediction for predictive maintenance in manufacturing. *Comput Ind Eng*. 2023;184:109566. doi:10.1016/j.cie.2023.109566
34. Ding C, Bao TY, Huang HL. Quantum-inspired support vector machine. *IEEE Trans Neural Netw Learn Syst*. 2022;33(12):7210–7222. doi:10.1109/TNNLS.2021.3084467
35. Han S, Kim H, Lee Y-S. Double random forest. *Mach Learn*. 2020;109(8):1569–1586. doi:10.1007/s10994-020-05889-1
36. Song L, Langfelder P, Horvath S. Random generalized linear model: a highly accurate and interpretable ensemble predictor. *BMC Bioinf*. 2013;14 (1):5. doi:10.1186/1471-2105-14-5
37. Jia M, Li J, Zhang J, et al. Identification and validation of cuproptosis related genes and signature markers in bronchopulmonary dysplasia disease using bioinformatics analysis and machine learning. *BMC Med Inf Decis Making*. 2023;23(1):69. doi:10.1186/s12911-023-02163-x
38. Chen CJ, Chen TH, Lei J, et al. Correlation of ER, PR, and HER2 at the protein and mRNA levels in Asian patients with operable breast cancer. *Biosci Rep*. 2022;42(1). doi:10.1042/bsr20211706
39. Liu M, Zhao X, Ma Z, et al. Discovery of potential Q-marker of traditional Chinese medicine based on chemical profiling, chemometrics, network pharmacology, and molecular docking: centipeda minima as an example. *Phytochemical Anal*. 2022;33(8):1225–1234. doi:10.1002/pca.3173
40. López-Barajas IB, Muñoz A, Legerén M, et al. Evaluation of the conversion rate in Ki-67, estrogen receptor (ER), progesterone receptor (PR) and HER2 between primary breast cancer and relapse and their value as a prognostic factor. *Ann Oncol*. 2016:27:vi31. doi:10.1093/annonc/mdw363.51.

41. Gamble P, Jaroensri R, Wang H, et al. Determining breast cancer biomarker status and associated morphological features using deep learning. *Communicat Med*. 2021;1(1):14. doi:10.1038/s43856-021-00013-3
42. Liu Y, Hu Y, Xue J, et al. Advances in immunotherapy for triple-negative breast cancer. *Mol Cancer*. 2023;22(1):145. doi:10.1186/s12943-023-01850-7
43. Zhang J, Li X, Han X, et al. Targeting the Thioredoxin System for Cancer Therapy. *Trends Pharmacol Sci*. 2017;38(9):794–808. doi:10.1016/j.tips.2017.06.001
44. Alinejad V, Dolati S, Motallebnezhad M, et al. The role of IL17B-IL17RB signaling pathway in breast cancer. *Biomed Pharmacothe*. 2017;88:795–803. doi:10.1016/j.biopha.2017.01.120
45. Cruceriu D, Baldasici O, Balacescu O, et al. The dual role of tumor necrosis factor-alpha (TNF-α) in breast cancer: molecular insights and therapeutic approaches. *Cell Oncol*. 2020;43(1):1–18. doi:10.1007/s13402-019-00489-1
46. Mercogliano MF, Bruni S, Elizalde PV, et al. Tumor necrosis factor α blockade: an opportunity to tackle breast cancer. *Front Oncol*. 2020;10:584. doi:10.3389/fonc.2020.00584
47. Wang Y, Jia Z, Liang C, et al. MTSS1 curtails lung adenocarcinoma immune evasion by promoting AIP4-mediated PD-L1 monoubiquitination and lysosomal degradation. *Cell Discovery*. 2023;9(1):20. doi:10.1038/s41421-022-00507-x
48. Yang H, Geng Y-H, Wang P, et al. Extracellular ATP promotes breast cancer chemoresistance via HIF-1α signaling. *Cell Death Dis*. 2022;13 (3):199. doi:10.1038/s41419-022-04647-6
49. Zhong Y, Yu F, Yang L, et al. HOXD9/miR-451a/PSMB8 axis is implicated in the regulation of cell proliferation and metastasis via PI3K/AKT signaling pathway in human anaplastic thyroid carcinoma. *J Transl Med*. 2023;21(1):817. doi:10.1186/s12967-023-04538-0
50. Lv L, Wang X, Wu H. Assessment of palmitic acid toxicity to animal hearts and other major organs based on acute toxicity, network pharmacology, and molecular docking. *Comput Biol Med*. 2023;158:106899. doi:10.1016/j.compbiomed.2023.106899
51. Fu X, Ma B, Zhou M, et al. Network pharmacology integrated with experimental validation to explore the therapeutic role and potential mechanism of Epimedium for spinal cord injury. *Front Mol Neurosci*. 2023;16. doi:10.3389/fnmol.2023.1074703
52. Shan P, Yang F, Qi H, et al. Alteration of MDM2 by the small molecule YF438 exerts antitumor effects in triple-negative breast cancer. *Cancer Res*. 2021;81(15):4027–4040. doi:10.1158/0008-5472.Can-20-0922
53. Zhu J, Kong W, Huang L, et al. MLSP: a bioinformatics tool for predicting molecular subtypes and prognosis in patients with breast cancer. *Comput Struct Biotechnol J*. 2022;20:6412–6426. doi:10.1016/j.csbj.2022.11.017
54. Wen G, Wen P, Tang Z. Research on data mining method of TCM prescription based on machine learning. *J Phys Conf Ser*. 2021;1952(2):022033. doi:10.1088/1742-6596/1952/2/022033
55. Kotlyar M, Pastrello C, Sheahan N, et al. Integrated interactions database: tissue-specific view of the human and model organism interactomes. *Nucleic Acids Res*. 2015:44(D1):D536–D541. doi:10.1093/nar/gkv1115.
56. Wang A, Luan HH, Medzhitov R. An evolutionary perspective on immunometabolism. *Science*. 2019;363(6423):6423):eaar3932. doi:10.1126/science.aar3932
57. Tadaka S, Kinoshita K. NCMine: core-peripheral based functional module detection using near-clique mining. *Bioinformatics*. 2016;32 (22):3454–3460. doi:10.1093/bioinformatics/btw488
58. Li R, Zhu J, Zhong W-D, et al. Comprehensive evaluation of machine learning models and gene expression signatures for prostate cancer prognosis using large population cohorts. *Cancer Res*. 2022;82(9):1832–1843. doi:10.1158/0008-5472.Can-21-3074
59. Martínez BA, Shrotri S, Kingsmore KM, et al. Machine learning reveals distinct gene signature profiles in lesional and nonlesional regions of inflammatory skin diseases. *Sci Adv*. 2022;8(17):eabn4776. doi:10.1126/sciadv.abn4776
60. Wang J, Lu Y, Sun G, et al. Machine learning algorithms for a novel cuproptosis-related gene signature of diagnostic and immune infiltration in endometriosis. *Sci Rep*. 2023;13(1):21603. doi:10.1038/s41598-023-48990-w
61. Ebright RY, Zachariah MA, Micalizzi DS, et al. HIF1A signaling selectively supports proliferation of breast cancer in the brain. *Nat Commun*. 2020;11(1):6311. doi:10.1038/s41467-020-20144-w
62. Collin LJ, Maliniak ML, Cronin-Fenton DP, et al. Hypoxia-inducible factor-1α expression and breast cancer recurrence in a Danish population-based case control study. *Breast Cancer Res*. 2021;23(1):103. doi:10.1186/s13058-021-01480-1
63. Cox A, Dunning AM, Garcia-Closas M, et al. A common coding variant in CASP8 is associated with breast cancer risk. *Nat Genet*. 2007;39 (3):352–358. doi:10.1038/ng1981
64. Liu X, Bai F, Wang Y, et al. Loss of function of GATA3 regulates FRA1 and c-FOS to activate EMT and promote mammary tumorigenesis and metastasis. *Cell Death Dis*. 2023;14(6):370. doi:10.1038/s41419-023-05888-9