Acta Pharmaceutica Sinica B 2025;15(1):133−150

Chinese Pharmaceutical Association
Institute of Materia Medica, Chinese Academy of Medical Sciences

Acta Pharmaceutica Sinica B

www.elsevier.com/locate/apsb
www.sciencedirect.com

ORIGINAL ARTICLE

GSFM: A genome-scale functional module transformation to represent drug efficacy for *in silico* drug discovery

Saisai Tian[a,b,*,†], Xuyang Liao[b,c,g,†], Wen Cao[a,†], Xinyi Wu[d], Zexi Chen[e], Jinyuan Lu[f], Qun Wang[g], Jinbo Zhang[h], Luonan Chen[e,*], Weidong Zhang[a,b,c,f,g,i,*]

[a]*Department of Phytochemistry, School of Pharmacy, Second Military Medical University, Shanghai 200433, China*
[b]*State Key Laboratory for Quality Ensurance and Sustainable Use of Dao-di Herbs, Institute of Medicinal Plant Development, Chinese Academy of Medical Sciences and Peking Union Medical College, Beijing 100193, China*
[c]*Pharmacy College, Fujian University of Traditional Chinese Medicine, Fuzhou 350122, China*
[d]*School of Pharmacy, Henan University, Kaifeng 475004, China*
[e]*Key Laboratory of Systems Biology, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, Shanghai 200031, China*
[f]*School of Pharmacy, Anhui University of Chinese Medicine, Hefei 230012, China*
[g]*Shanghai Frontiers Science Center of TCM Chemical Biology, Institute of Interdisciplinary Integrative Medicine Research, Shanghai University of Traditional Chinese Medicine, Shanghai 201203, China*
[h]*Department of pharmacy, Tianjin Rehabilitation Center of Joint Logistics Support Force, Tianjin 300110, China*
[i]*The Research Center for Traditional Chinese Medicine, Shanghai Institute of Infectious Diseases and Biosafety, Institute of Interdisciplinary Integrative Medicine Research, Shanghai University of Traditional Chinese Medicine, Shanghai 201203, China*

Received 13 May 2024; received in revised form 15 July 2024; accepted 12 August 2024

KEY WORDS

GSFM;
Data transformation strategy;
Multi-dimensions activities;

Abstract Pharmacotranscriptomic profiles, which capture drug-induced changes in gene expression, offer vast potential for computational drug discovery and are widely used in modern medicine. However, current computational approaches neglected the associations within gene−gene functional networks and unrevealed the systematic relationship between drug efficacy and the reversal effect. Here, we developed a new genome-scale functional module (GSFM) transformation framework to quantitatively evaluate drug efficacy for *in silico* drug discovery. GSFM employs four biologically interpretable quantifiers:

*Corresponding authors.
E-mail addresses: saisai_tian@foxmail.com (Saisai Tian), lnchen@sibcb.ac.cn (Luonan Chen), wdzhangy@hotmail.com (Weidong Zhang).
†These authors made equal contributions to this work.

Peer review under the responsibility of Chinese Pharmaceutical Association and Institute of Materia Medica, Chinese Academy of Medical Sciences.

https://doi.org/10.1016/j.apsb.2024.08.017

Drug efficacy;
In silico drug discovery

GSFM_Up, GSFM_Down, GSFM_ssGSEA, and GSFM_TF to comprehensively evaluate the multi-dimension activities of each functional module (FM) at gene-level, pathway-level, and transcriptional regulatory network-level. Through a data transformation strategy, GSFM effectively converts noisy and potentially unreliable gene expression data into a more dependable FM active matrix, significantly outperforming other methods in terms of both robustness and accuracy. Besides, we found a positive correlation between RS_{GSFM} and drug efficacy, suggesting that RS_{GSFM} could serve as representative measure of drug efficacy. Furthermore, we identified WYE-354, perhexiline, and NTNCB as candidate therapeutic agents for the treatment of breast-invasive carcinoma, lung adenocarcinoma, and castration-resistant prostate cancer, respectively. The results from *in vitro* and *in vivo* experiments have validated that all identified compounds exhibit potent anti-tumor effects, providing proof-of-concept for our computational approach.

1. Introduction

In silico-driven drug discovery has garnered considerable attention for its ability to expedite the drug development process while reducing time, labor, and financial expenditure[1]. Computational screening of compounds *via* gene expression profiles has proven to be a highly effective method in drug discovery, facilitating the targeting of specific biological processes and identifying potential therapeutic agents[2-4]. Numerous pharmacotranscriptomic perturbed databases, such as CMap[5], LINCS[6], and ITCM[7], have been established to accelerate drug discovery efforts. For instance, Rabben et al.[8] repurposed ivermectin for gastric cancer treatment using the CMap database, while Chen et al.[9] identified homoharringtonine as a potential liver cancer therapy through the LINCS database. Additionally, various computational algorithms leveraging gene expression data have been developed to support drug discovery[10-13]. Lamb et al.[10] introduced a pattern-matching algorithm that utilizing the Kolmogorov−Smirnov statistic to rank all drugs based on their connectivity scores. Cheng et al.[11] proposed the eXtreme Sum score (XSum) algorithm to quantify the reversal relationship between the drugs and diseases. These algorithms offer valuable tools for analyzing gene expression data and identifying potential drug candidates.

However, many of these strategies primarily focus on conventional signatures that rely solely on differentially expressed genes, overlooking the gene−gene functional network associations. Consequently, these approaches often suffer from reduced robustness and accuracy. From a systems/network perspective, a gene with functional abnormalities disrupts not only its direct partners within the human protein−protein interactome, but also affects a nearby subnetwork centered around that gene[14-16]. In essence, many genes interact with one another, participate in the same biological functions or signaling pathways, and form functional modules (FMs)[17]. In this study, pathways or biological processes are defined as FMs. Various module-based approaches have been developed to characterize FM activity in individuals or samples. One commonly used method is ssGSEA[18], which infers pathway activity for each sample based on expression data. Another approach is Viper[19] (virtual inference of protein−activity by enriched regulon analysis), which uses enriched regulon analysis to infer the activity of transcriptional regulons. Nonetheless, these methods only provide a single score for each module at a single level and may introduce noise, thereby failing to fully capture the biological complexity[20]. Moreover, the relationship between drug reversal capacity and drug efficacy has yet to be systematically explored.

This study introduces a transformative framework for drug discovery termed genome-scale functional module (GSFM)-based drug discovery. This framework integrates four biologically interpretable quantifiers (BIQs) to evaluate the overall activity of the FMs: GSFM_Up, GSFM_Down, GSFM_ssGSEA, and GSFM_TF. GSFM_Up and GSFM_Down measure gene-level activity within a FM, with GSFM_Up assessing the ratio of highly expressed genes (95th percentile, Z-score >1.96) and GSFM_Down evaluating the ratio of lowly expressed genes (5th percentile, Z-score <-1.96). Conversely, GSFM_ssGSEA quantifies the pathway-level activity for each FM by comparing the normalized difference between the empirical cumulative distribution function of gene expression ranks inside and outside the module. Furthermore, GSFM_TF estimates transcriptional regulatory network-level activity by calculating the weighted average expression level of transcription factors (TFs) within a FM. These BIQs provide a comprehensive assessment of each FM's activity from a systems/network perspective.

To systematically understand how these modules respond to drug perturbations, the genome-scale functional module-based reversal score (RS_{GSFM}) was developed to compare FM pattern of drug perturbations with the disease-induced FM pattern, characterizing the potential of drug efficacy. Notably, a positive correlation was observed between RS_{GSFM} and half-maximal inhibitory concentration (IC_{50}), a key measure of drug efficacy. Using the GSFM computational framework, WYE-354 with significant RS_{GSFM} as an effective candidate for breast-invasive carcinoma (BRCA). Additionally, perhexiline exhibited notable RS_{GSFM} for lung adenocarcinoma (LUAD), and NTNCB was identified as a potential treatment for castration-resistant prostate cancer (CRPC) based on its significant RS_{GSFM}. These findings underscore the potential of the GSFM framework in pinpointing candidate drugs for specific cancers. Subsequently, each identified compound underwent further evaluation for its effectiveness in inhibiting the growth of subcutaneous xenografts derived from different cancer cell lines in nude mice. This large-scale computational analysis demonstrates the feasibility and potential of GSFM as a valuable tool in drug discovery, showcasing its versatile and broadly applicable framework in the field of drug development.

2. Materials and methods

2.1. Data collection and processing

Drug-induced gene expression profiles were downloaded from the LINCS library using the signatureSearch R package[21]. This dataset includes gene expression profiles derived from treatments of 8140 compounds across 30 cell lines for 24 h at a concentration of 10 μmol/L, covering 12,328 genes, including 978 landmark genes and 11,350 additional inferred genes. Efficacy data for each compound (IC_{50}) were sourced from the ChEMBL database (version 27). Cell lines were mapped between the LINCS and ChEMBL databases using the cell line names, followed by manual inspection. Due to potential variability in IC_{50} values for a single compound across different studies, even within the same cell lines, we adopted the median as a summary measure for the IC_{50}. This study focuses on three cancer types—BRCA, LUAD, and CRPC—based on the availability of relevant compound gene expression profiles in LINCS and corresponding compound efficacy data in ChEMBL. The selected cell lines were MCF7, A549, and PC3, chosen for their high number of compounds with both efficacy and expression data in the ChEMBL and LINCS databases, respectively. Additionally, gene expression profiles of tumors and adjacent normal tissues in BRCA and LUAD were extracted from the TCGA database. Since TCGA-PRAD primarily includes primary prostate tumors, it is important to note that PC3 represents castration-resistant prostate cancer (CRPC), which may exhibit significant molecular differences compared to primary tumors. Therefore, CRPC datasets were collected from GSE70768. Multi-validation datasets from the GEO database were also collected to enhance the model's robustness, including GSE3744, GSE42568, and GSE109169 for BRCA; GSE10072, GSE43458, and GSE81089 for LUAD; and GSE80609, GSE33277, and GSE3325 for CRPC. Moreover, TF-target gene pairs were gathered based on prior research findings[22], which provided evidence from literature-curated resources, TF binding motifs on promoters, or ChIP-seq peaks. For subsequent analysis, high-confidence TF-target pairs (confidence levels A and B) were selected, following the criteria outlined by Garcia-Alonso et al.[22] (Supporting Information Table S1).

2.2. Selection of functional modules

A common set of criteria for module selection includes assessing the relevance of functionality for each module, ensuring broad coverage of functionalities across different modules, and promoting extensive coverage of genes throughout the modules. In the use cases described here, we define 50 cancer-related hallmark genesets as FMs[23,24], including 'Adipogenesis', 'Allograft rejection', 'Androgen response', 'Angiogenesis', 'Apical junction', 'Apical surface', 'Apoptosis', 'Bile acid metabolism', 'Cholesterol homeostasis', 'Coagulation', 'Complement', 'DNA repair', 'E2F targets', 'Epithelial mesenchymal transition', 'Estrogen response early', 'Estrogen response late', 'Fatty acid metabolism', 'G2M checkpoint', 'Glycolysis', 'Hedgehog signaling', 'Heme metabolism', 'Hypoxia', 'IL2 STAT5 signaling', 'IL6 JAK STA3 signaling', 'Inflammatory response', 'Interferon α response', 'Interferon gamma response', 'KRAS signaling DN', 'KRAS signaling UP', 'Mitotic spindle', 'Mtorc1 signaling', 'MYC targets V1, 'MYC targets V2', 'Myogenesis', 'Notch signaling', 'Oxidative phosphorylation', 'p53 pathway', 'Pancreas β cells', 'Peroxisome', 'PI3K AKT MTOR signaling', 'Protein secretion', 'Reactive oxygen species pathway', 'Spermatogenesis', 'TGF β signaling', 'TNFA signaling *via* NFKB', 'Unfolded protein response', 'UV response DN', 'UV response UP', 'WNT β catenin signaling', 'Xenobiotic metabolism' (Supporting Information Table S2). These FMs were subsequently annotated into six primary process categories: 'Metabolism', 'Genetic Information Processing', 'Environmental', 'Cellular Processes', 'Cellular component', and 'Organismal Systems'. More importantly, these FMs cover diverse cellular activities, and convey a specific biological state or process. Meanwhile, they also show mutual exclusivity across modules (Supporting Information Fig. S1A and S1B). It is important to note that the choice of modules can vary depending on the specific research question posed by the researcher.

2.3. Multi-levels GSFM transformation

Inspired by previous research[18,19,25], the GSFM algorithm was developed to transform the gene expression matrix into the GSFM active matrix. The algorithm's core involves utilizing four distinct methods to calculate the overall activity of each FM at multi-levels: GSFM_Up, GSFM_Down, GSFM_ssGSEA, and GSFM_TF.

2.3.1. Gene-level

Both GSFM_Up and GSFM_Down are employed to evaluate FM activities at the gene level by calculating the proportion of genes exhibiting significantly elevated or reduced expression ($|Z\text{-score}| > 1.96$) within a specific FM.

$$\text{GSFM\_Up} = \frac{N_{\text{up}}}{N_{\text{all}}} \tag{1}$$

N_{up} represents the number of up-regulated genes in a FM, and N_{all} denotes the total number of genes within that FM.

$$\text{GSFM\_Down} = -1 \times \frac{N_{\text{down}}}{N_{\text{all}}} \tag{2}$$

N_{down} represents the number of down-regulated genes in the FM.

2.3.2. Pathway-level

The GSFM_ssGSEA[18] calculated the FM activity for each sample at the pathway level, by quantifying the normalized difference between the empirical cumulative distribution function (ECDF) of gene expression ranks within and outside the FM.

a. The inclusion of a gene within gene set FM can be demonstrated by Eq. (3):

$$P_{\text{FM}}^{\text{w}} = \sum_{r_j \in \text{FM}, j \le i} \frac{|r_j|^{\alpha}}{\sum_{r_j \in \text{FM}} |r_j|^{\alpha}} \tag{3}$$

b. The exclusion of a gene from set FM can be demonstrated by Eq. (4):

$$P_{N_{\text{FM}}} = \sum_{r_j \notin \text{FM}, j \le i} \frac{1}{(N - N_{\text{FM}})} \tag{4}$$

c. Calculate the ssGSEA for each gene and determine the ssGSEA for this gene set based on the maximum difference (selecting the ssGSEA with the highest absolute value) using Eq. (5):

$$\text{GSFM\_ssGSEA} = \sum_{i=1}^{N} \left[P_{\text{FM}}^{\text{w}} - P_{N_{\text{FM}}} \right] \tag{5}$$

where P_{FM}^{w} is the ECDF of genes within FMs, $P_{N_{\text{FM}}}$ is the ECDF of the remaining genes, r_j represents the genes within an expression profile, FM signifies the selected FMs, j denotes the rank of the gene, i indicates the total count of gene ranks within the expression matrix, N represents the total number of genes in the expression matrix, and N_{FM} stands for the total number of genes encompassed within the selected FMs. Index (α) is the weight, typically set to 1.

2.3.3. Transcriptional regulatory network-level

The GSFM_TF score assesses FM activity at the transcriptional regulatory network-level by calculating the average expression of TFs whose target genes show significant enrichment within each FM. Due to the tissue- or cell type-specific nature of transcriptional regulatory networks, correlations between TFs and their target genes were initially evaluated. Pairs exhibiting a significant correlation in a particular tumor type were retained (a relatively loose threshold: $|r| > 0.2$, $P < 0.05$). Subsequently, a one-tailed Fisher exact test was then conducted on the list of TF-target gene pairs to determine if the target genes of a TF are significantly enriched within the FM. Only the TFs demonstrating significant enrichment in a module were selected as the TFs associated with that FM. The calculation of TF regulation weight was conducted using Eqs. (6)−(8):

$$\text{Ratio}_{\text{Module}_i} = \frac{N_{\text{TF}_i}}{N_{\text{TF}_s}} \tag{6}$$

$$\text{Ratio}_{\text{TF}_i} = \frac{N_{\text{TF}_i}}{N_{\text{TF}}} \tag{7}$$

$$\text{Weight}_i = \text{Ratio}_{\text{Module}_i} \times \text{Ratio}_{\text{TF}_i} \tag{8}$$

where N_{TF_i} denotes the number of target genes for TF_i in one FM, N_{TF_s} signifies the number of target genes that have been regulated by all the TF_s, and N_{TF} represents the number of target genes regulated by TF_i across all selected FMs.

Weights within each module were standardized to ensure the total TF regulation weights sum to 1. The GSFM_TF score was calculated by aggregating the normalized weights, with each weight was multiplied by the corresponding expression level of the TFs, as Eq. (9):

$$\text{GSFM\_TF} = \frac{1}{N} \sum_{i=1}^{N} \text{weight}_i \times \text{Exp}(\text{TF}_i) \tag{9}$$

where N represents the number of significantly regulated TF_s in the FM, and $\text{Exp}(\text{TF}_i)$ is the expression level of the TF_s in the FM.

2.4. Benchmarks for distinct matching methods

The retrieval performance of four distinct matching methods—CMap[10], XSum[11], ZhangScore[12], and RS[13]—was systematically benchmarked. The specific features and scoring schemes of these methods are described as follows:

The CMap method separates disease signatures into two FM sets—upregulated and downregulated—disregarding the magnitude of differential expression. Initially, the drug GSFM activity matrix is used as a reference to compute the maximum deviation (MD)-based ES for the upregulated FM set (es_{up}) and the downregulated FM set (es_{down}). The CMap score is then defined as follows:

a. If es_{up} and es_{down} have the same algebraic sign then $\text{CMap}_{\text{GSFM}} = 0$.
b. $\text{CMap}_{\text{GSFM}} = \text{es}_{\text{up}} - \text{es}_{\text{down}}$.

The XSum method, akin to CMap, independently addresses up- and down-regulated FMs separately. The sums of the change values in drug FMs relative to upregulated disease FMs (sum_{up}) and downregulated disease FMs (sum_{down}) are first calculated. The XSum score is subsequently defined as Eq. (10):

$$\text{XSum}_{\text{GSFM}} = \text{sum}_{\text{up}} - \text{sum}_{\text{down}} \tag{10}$$

The ZhangScore method ranks FMs in a reference profile by the absolute value of their expression value. In the context of a disease or drug GSFM active matrix, R is defined as the complete replicate-consensus FM of the drug, while S represents the union of the upregulated and downregulated FM. The ZhangScore is then defined as Eq. (11):

$$\text{ZhangScore}_{\text{GSFM}} = \frac{\sum_{i=1}^{n} R_{(\text{FM}_i)} S_{(\text{FM}_i)}}{\sum_{i=1}^{n} (N - i + 1)} \tag{11}$$

where FM_i resents the i^{th} FM in R, $S_{(\text{FM}_i)}$ is 1 for up-regulated FMs or −1 for down-regulated FMs, and $R_{(\text{FM}_i)}$ is this FM's signed rank in R. n is the length of S, and N is the length of R.

The RS method was adapted from the RGES developed in previous studies[13].

(1) Given a ranked list of drug-related FMs L with n FM and a disease-related FMs S with m FMs.
(2) Given a vector V indicating the position (1, 2, …, n) of each module in L and sort the FMs in S in ascending order such that $V(i)$ is the position of FM i, where $i = 1, 2, \ldots, t$. Compute the following two values:

$$a = \max_{i=1}^{t} \left[\frac{i}{t} - \frac{V(i)}{n} \right] \tag{12}$$

$$b = \max_{i=1}^{t} \left[\frac{V(i)}{n} - \frac{i-1}{t} \right] \tag{13}$$

Then, for each up- and down-regulated disease FMs, we calculated a_{up}, a_{down}, b_{up}, and b_{down} using Eqs. (14)−(16):

$$\text{es}_{\text{up}} = \begin{cases} a_{\text{up}} & \text{if } a_{\text{up}} > b_{\text{up}} \\ -b_{\text{up}} & \text{if } a_{\text{up}} < b_{\text{up}} \end{cases} \tag{14}$$

$$\text{es}_{\text{down}} = \begin{cases} a_{\text{down}} & \text{if } a_{\text{down}} > b_{\text{down}} \\ -b_{\text{down}} & \text{if } a_{\text{down}} < b_{\text{down}} \end{cases} \tag{15}$$

$$\text{RS}_{\text{GSFM}} = \text{es}_{\text{up}} - \text{es}_{\text{down}} \tag{16}$$

In this context, es_{up} denotes the absolute enrichment score of an up-regulated FM list in a given profile, while es_{down} represents the absolute enrichment score of a down-regulated FM list in the same profile.

2.5. The performance evaluation of robustness and accuracy

The area under the receiver operating characteristic curve (ROC)[26] and the area under the precision-recall curve (PR)[27] were utilized to assess the performance of GSFM. In this research, ROC and PR were achieved using a rigorous 10-fold cross-validation framework. Briefly, the dataset was split into 10 subsets-9 subsets are used for training while the 10th subset is used for validation. Each subset served as the validation set once, therefore each individual was in the validation fold exactly once for each cross-validation run. Finally, the AUC of ROC and PR curves for each fold are averaged to produce a more robust estimate as our final ROC and PR. Meanwhile, enrichment analysis[28] was employed for further validation, using existing FDA-approved or investigational drugs specific to each individual cancer. Moreover, a novel indicator, accuracy score, was developed to qualify the performance.

$$\text{Accuracy score} = r \times (-\log(P_{\text{Enrichment}})) \tag{17}$$

where r is the correlation between RS_{GSFM} and $IC_{50,}$ and $P_{\text{Enrichment}}$ is derived from enrichment analysis based on FDA-approved or investigational drugs from clinicaltrials.gov for each individual cancer.

2.6. BIQ combination and FM size selection by drug retrieval performance

Considering that the BIQ combination and FM size are crucial factors influencing the performance of GSFM, a rigorous and robust analysis was conducted to determine the optimal BIQ combination and the appropriate FM size. In this study, we applied a novel benchmarking standard called ROC-based standard to evaluate the drug retrieval performance[26]. Drug response data for MCF7, A549, and PC3 cell lines were collected from the ChEMBL database. Compounds in LINCS and ChEMBL were mapped using the compound names, followed by manual inspection. Due to IC_{50} redundancy, the median IC_{50} of duplicate compounds was used to represent the compound's activity (Supporting Information Table S3). Compounds were categorized into effective ($IC_{50} < 10$ μmol/L) and ineffective ($IC_{50} > 10$ μmol/L) groups[13]. The retrieval performance was measured by assessing the ability to distinguish between effective and ineffective compounds using the ROC value, with higher ROC values indicating superior performance. Random combination experiments demonstrated the rationale for the current four BIQ combinations. All possible BIQ combinations and the maximum top x upregulated and downregulated FM sizes at 25, 50, 75, 100, 125, 150, 175, and 200 were extracted, defining them as FM-associated signatures to generate the ROC.

2.7. Mice and chemicals

Male and female BALB/c nude mice were obtained from Shanghai Leagene Biotechnology Co., Ltd. (Shanghai, China). All animals were kept and cared for in a pathogen-free environment, with unrestricted access to sterilized food and autoclaved water. The housing conditions included a 12-h light/dark cycle, maintaining a room temperature of 21 ± 2 °C, and humidity levels between 45% and 65%. Animal work was approved by the Second Military Medical University and animal studies were carried out in compliance with all federal and local institutional rules for the conduct of animal experiments. WYE-354 and perhexiline were purchased from MCE (MedChemExpress). NTNCB was purchased from TargetMol. These three compounds were dissolved in dimethyl sulfoxide (DMSO) at stock solutions of 10 mmol/L. All stock solutions were stored at −20 °C for subsequent use in cell-based experiments.

2.8. Cellular experiments

2.8.1. Cell culture

MCF7 and PC3 cells were purchased from Procell Life Science & Technology Co., Ltd., while MDA-MB-231, A549, NCI-H1975, and DU145 cells were obtained from Shanghai Institute of Cell Resource Center Life Science (Shanghai, China). A549, MDA-MB-231, and NCI-H1975 cells were cultured in DMEM supplemented with 10% FBS and 1% P/S at 37 °C in humidified atmosphere with 5% CO_2. MCF7 cells were cultured in MEM supplemented with 10 μg/mL Insulin, 10% FBS, and 1% P/S at 37 °C in humidified atmosphere with 5% CO_2. PC3 cells were cultured in F-12K supplemented with 10% FBS and 1% P/S at 37 °C in humidified atmosphere with 5% CO_2. DU145 cells were cultured in RPMI 1640 supplemented with 10% FBS and 1% P/S at 37 °C in humidified atmosphere with 5% CO_2.

2.8.2. Cell proliferation assay and colony formation assay

The effects of WYE-354, perhexiline, and NTNCB on cell proliferation were examined using the Cell Counting Kit-8 (CCK-8) assay. MCF7, MDA-MB-231, A549, NCI-H1975, PC3, and DU145 cells (100 μL) were seeded into 96-well plates at a density of 5000 cells/well. After allowing the cells to adhere overnight, the medium was aspirated. The cells in the 96-wells were incubated with 100 μL of different compound concentrations (0, 1.25, 2.5, 5.0, 10.0, 30.0, and 50.0 μmol/L) for 72 h. Subsequently, the cells were incubated with 10 μL of CCK-8 solution at 37 °C for 0.5−2 h. Optical density values at 450 and 620 nm were measured using an enzyme marker. The IC_{50} values were determined from a dose-response curve. Dose-response curves were determined using GraphPad Prism 9.0.0 software. For the colony formation assay, each of the BRCA, LUAD, and CRPC cell line was seeded in six-well plates, with 600 cells in 2 mL of media per well. After 8 days, cells were stained with a Crystal Violet Staining Solution (C0121, Beyotime).

2.9. In vivo antitumor study

2.9.1. Xenograft mouse model and drug treatment

To generate subcutaneous xenografts, 5×10^6 MDA-MB-231, A549, and PC3 cells were suspended in 100 μL of Dulbecco's PBS (Invitrogen) and injected subcutaneously near the right forelimb of 5-week-old female/male Balb/c mice. The mice were randomly divided into five groups ($n = 6$ per group). One group received intraperitoneal injections of 0.2 mL paclitaxel (10 mg/kg, positive control) every 3 days for 15 days. Treatment groups were administered intraperitoneal injection of 0.2 mL of WYE-354, perhexiline, and NTNCB at doses of 1, 3, and 10 mg/kg daily for 15 days. The vehicle control group received an equivalent volume of the vehicle solution without the active compounds. Tumor sizes were measured with a caliper before each treatment time point, and the tumor volumes were calculated by Eq. (18):

$$V = \frac{\left(W^2 \times L\right)}{2} \tag{18}$$

where V presents volume, L presents tumor length, W presents tumor width.

2.9.2. Hematoxylin-eosin (H&E) staining

The excised mouse subcutaneous xenograft tumor tissues underwent initial fixation in a 4% paraformaldehyde solution, followed by standard paraffin embedding and sectioning procedures. Subsequently, the paraffin-embedded sections were then sequentially immersed in Environmentally Friendly Dewaxing Transparent Liquid I for 20 min, followed by Liquid II for another 20 min. This was succeeded by a series of dehydration steps in anhydrous ethanol I for 5 min, anhydrous ethanol II for another 5 min, and finally, 75% ethanol for 5 min, after which they were rinsed in tap water. The sections were then prepared for staining by treating them with an HD constant staining pre-treatment solution for 1 min, followed by immersion in hematoxylin solution for 3−5 min and rinsing with tap water. The sections were further processed with hematoxylin differentiation solution, then rinsed, and treated with hematoxylin bluing solution, followed by another rinse with tap water. The prepared sections were then dried and examined under a microscope.

2.9.3. Immunohistochemistry

Subcutaneous graft tumor tissue blocks derived from mice, fixed in 4% paraformaldehyde and embedded in paraffin, were utilized. Immunohistochemical staining was performed using primary antibodies and HRP-conjugated secondary antibodies. Specifically, the anti-Ki67 antibody (1:500, Servicebio, Cat#: GB121499) and HRP-conjugated goat anti-rabbit IgG (H+L) (1:200, Servicebio, Cat#: GB23303) were employed.

2.9.4. TUNEL assay

The TUNEL (Terminal deoxynucleotidyl Transferase dUTP Nick End Labeling) assay, a well-established method for detecting apoptosis, was utilized to evaluate apoptotic cells in mouse tissue samples. Utilizing TUNEL Assay Kit from Servicebio (Cat#: G1501), apoptosis detection was visualized with a fluorescence microscope, where DAPI-induced blue fluorescence was excited at 330−380 nm and emitted at 420 nm, while FITC-induced green fluorescence was excited at 465−495 nm and emitted at 515−555 nm.

2.10. Statistical analyses

All experiments were repeated at least three times. Statistical differences between the control vehicle and treatment groups were calculated using one-way analysis of variance (ANOVA). Significance was determined with a threshold of P values < 0.05, with statistical significance denoted as $^{*}P < 0.05$, $^{**}P < 0.01$, and $^{***}P < 0.001$ for extreme statistical significance. All statistical analyses for validation results were performed using GraphPad Prism 9.0.0 software, while computational analyses were conducted in R version 4.1.1 and Python 3.9.7.

3. Results

3.1. Transformation framework of GSFM

Here, we devised a novel FM-based transformation framework, GSFM, for *in silico* drug discovery. Fig. 1 provides an overview of this framework. Initially, the gene expression matrix is converted into a GSFM active matrix using four BIQs: GSFM_Up, GSFM_Down, GSFM_ssGSEA, and GSFM_TF, which calculate overall activity for each FM at multi-levels (Fig. 1A). In the use cases presented here, 50 cancer-related hallmark genesets were collected as FMs, encompassing diverse cellular functional activities, including cellular processes, metabolism, and genetic information processing et al. All functional modules exhibit mutual exclusivity (Fig. S1A and S1B, Tables S1 and S2). Secondly, GSFM optimization was achieved by refining drug prediction algorithms, BIQ combinations, and adjusting FM size. Simultaneously, the relationship between drug reversal capacity and efficacy was elucidated (Fig. 1B). GSFM demonstrated strong correlations and outperformed traditional differential gene analysis (Gene) method and Viper methods[19]. Various performance metrics, such as correlation, ROC, PR, enrichment analysis, and accuracy score, underscored its robustness and accuracy. The ultimate step involves GSFM-based experimental validation of nominated drug candidates and mechanistic observations (Fig. 1C). This research focused specifically on three types of cancers: BRCA, LUAD, and CRPC, selected based on the availability of relevant compound gene expression profiles in LINCS and corresponding compound efficacy data in ChEMBL. RS_{GSFM} was employed to prioritize compounds, identifying top hits that have not been previously studied in each cancer animal model: WYE-354 (BRCA, RS_{GSFM}: −0.92), perhexiline (LUAD, RS_{GSFM}: −0.85), and NTNCB (CRPC, RS_{GSFM}: −0.89). These compounds were further verified using *in vitro* and *in vivo* tests to ensure their credibility.

3.2. Construct the multi-levels GSFM active matrix and determine the FM size

Clinical samples of BRCA and LUAD were sourced from TCGA, and CRPC from GSE70768 for subsequent analysis. The gene expression matrix was transformed into a GSFM active matrix following the established framework (Fig. S1C). A total of 200 FMs activities were calculated for each cancer type. Different FM signatures ($DFMS_{tumor}$) were generated using the limma package[29]. Similarly, the drug expression matrix was converted into a GSFM active matrix, resulting in the construction of $DFMS_{drug}$. The RS algorithm was employed to calculate the RS_{GSFM} value, characterizing the potential of drug efficacy. A systematic performance comparison was conducted with other drug prediction algorithms—CMap, XSum, and ZhangScore. Findings indicated that the RS algorithm within our computational framework surpasses the other three algorithms (Fig. S1D−S1F). Given the significant impact of BIQs and FM sizes on drug retrieval performance, a rigorous and robust analysis identified the optimal of BIQ combinations and FM sizes using the ROC-based benchmarking standards. All possible combinations ($n = 15$) were compared, revealing that the combination utilizing all four quantifiers consistently outperformed the other fourteen combinations across nearly all candidate FM sizes (Fig. S1G−S1I). Performance scores of the four quantifiers were also assessed at specific FM sizes (25, 50, 75, 100, 125, 150, 175, and 200), suggesting that $x = 125$ is a suitable condition where the performance score reached a steady state. Consequently, four quantifiers and $x = 125$ were used as criteria for the subsequent analysis (Fig. S1J−S1L). Furthermore, analyses demonstrated that the GSFM outperformed the Viper and Gene methods (Fig. S1J−S1L).

3.3. RS_{GSFM} correlates with drug efficacy in each individual cancer and shows superiority

The potential association between the RS_{GSFM} value of a compound and its effectiveness in the same cell line was assessed.

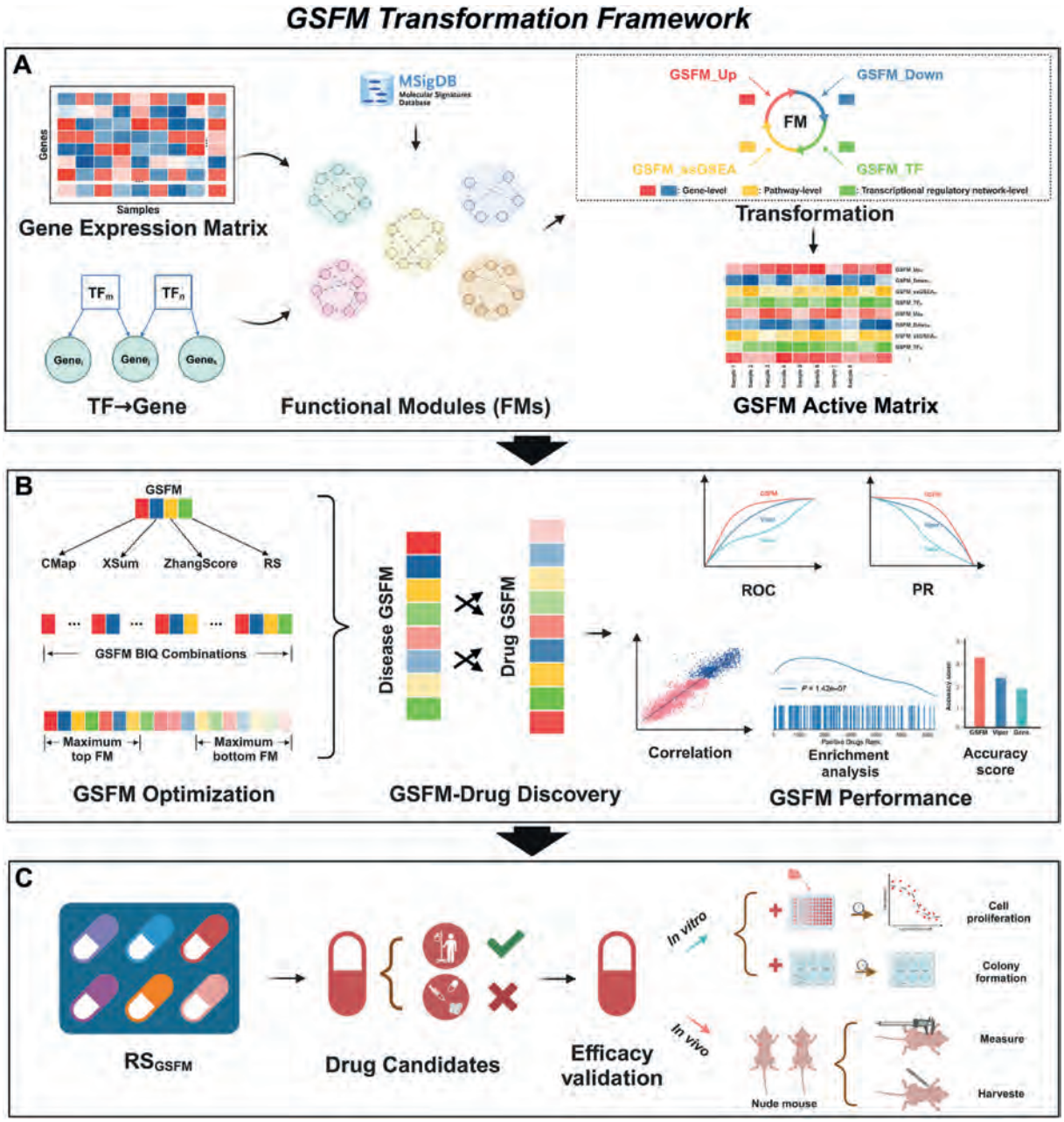


Figure 1 Overview of the GSFM transformation framework. (A) Transforming the gene expression matrix into GSFM active matrix involves employing four BIQs: GSFM_Up, GSFM_Down, GSFM_ssGSEA, and GSFM_TF. (B) ROC-based standard to assess drug retrieval performance through the integration of four prediction algorithms, the selection of BIQ combinations, and the determination of FM size to optimize the GSFM. Simultaneously, diverse performance metrics, including correlation, ROC, PR, enrichment analysis, and accuracy score, were employed to demonstrate its robustness and accuracy. (C) Experimental study to validate drug candidates in each individual cancer. Created with BioRender.com.

Fig. 2 illustrates that RS_{GSFM} exhibits a strong correlation with drug efficacy across all three cancer cell types (BRCA: $r = 0.61$, $P = 4.40e\text{-}22$; LUAD: $r = 0.62$, $P = 7.66e\text{-}11$; CRPC: $r = 0.60$, $P = 5.23e\text{-}08$; Fig. 2A–C). Compounds were subsequently categorized into functionally effective and ineffective groups (Table S3). The functionally effective compounds exhibited significantly lower RS_{GSFM} values across all three cancers (BRCA: $P = 1.46e\text{-}18$, LUAD: $P = 1.31e\text{-}06$, CRPC: $P = 1.90e\text{-}09$, Student's t-test, Fig. 2A–C). RS_{Gene} and RS_{Viper} were calculated similarly to RS_{GSFM}, and both showed correlations with drug efficacy. However, the GSFM algorithm revealed a stronger correlation and a more pronounced significance between RS_{GSFM} and drug efficacy compared to the other two methods (Fig. 2A–I).

To further evaluate the performance of GSFM, the following performance metrics were utilized: ROC, PR, enrichment analysis, and accuracy score. GSFM demonstrated superior performance compared to other methods (BRCA: ROC = 0.840, PR = 0.832; LUAD: ROC = 0.825, PR = 0.894; CRPC: ROC = 0.910, PR = 0.951, Fig. 3A–C). A collection of FDA-approved or

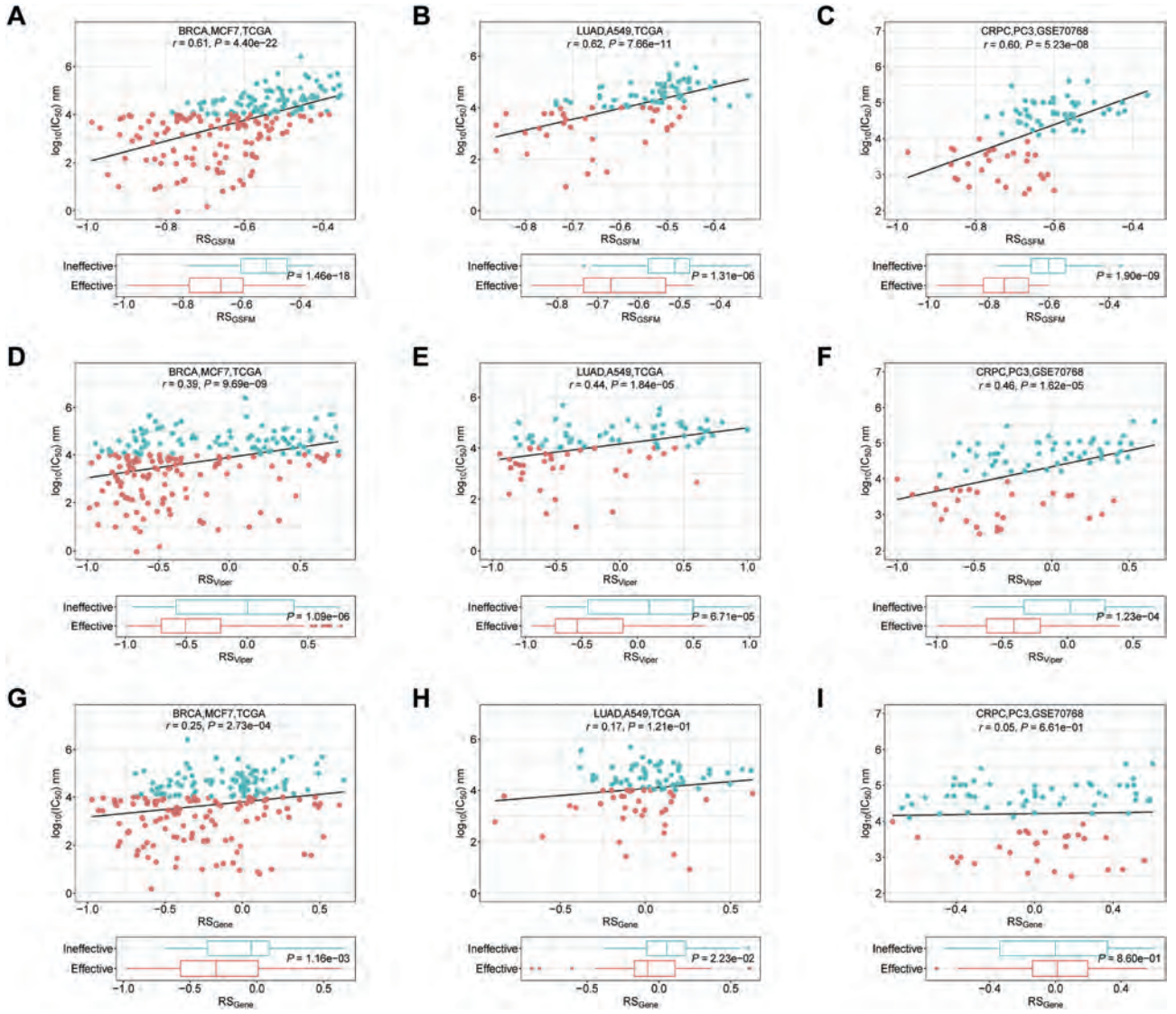


Figure 2 RS_{GSFM} represents drug efficacy in individual cancer cell lines. (A) The relationship between RS_{GSFM} and drug efficacy in MCF7 for BRCA. (B) The relationship between RS_{GSFM} and drug efficacy in A549 for LUAD. (C) The relationship between RS_{GSFM} and drug efficacy in PC3 for CRPC. (D) The relationship between RS_{Viper} and drug efficacy in MCF7 for BRCA. (E) The relationship between RS_{Viper} and drug efficacy in A549 for LUAD. (F) The relationship between RS_{Viper} and drug efficacy in PC3 for CRPC. (G) The relationship between RS_{Gene} and drug efficacy in MCF7 for BRCA. (H) The relationship between RS_{Gene} and drug efficacy in A549 for LUAD. (I) The relationship between RS_{Gene} and drug efficacy in PC3 for CRPC. Each dot represents one compound. Median IC_{50} was used when one compound has multiple IC_{50}s from different studies. ANOVA and Spearman correlation were used to measure correlation between RS and drug efficacy. Boxplots illustrate RS difference between effective compounds ($IC_{50} < 10$ μmol/L) and ineffective compounds ($IC_{50} > 10$ μmol/L).

investigational drugs for each individual cancer sourced from clinicaltrials.gov (Supporting Information Table S4). Subsequently, a comprehensive survey of clinical trials was then conducted, employing enrichment analysis within each specific cancer treatment domain. Notably, a significant enrichment of these drugs was observed at the top of the prediction results (BRCA: P = 4.52e-07; LUAD: P = 5.99e-03; CRPC: P = 5.15e-08, Fig. 3D−F, and Supporting Information Fig. S2A−S2F), further validating the analysis. An accuracy score was established to comprehensively evaluate the GSFM model and ensure a fair comparison. As illustrated in Fig. 3G−I, the GSFM method outperformed that of the Gene and Viper methods in terms of accuracy score. In summary, the GSFM algorithm proves to be highly effective for *in silico* drug discovery, surpassing other methods across all evaluated aspects.

3.4. *Evaluation of performance on validation datasets*

To further ensure the stability and generalizability of the GSFM computational framework, multiple independent cohorts were collected from the GEO database for external validation. Using the same computational formula, the expression data for BRCA (GSE3744, GSE42568, and GSE109169), LUAD (GSE10072, GSE43458, and GSE81089), and CRPC (GSE80609, GSE33277, and GSE3325) were converted into GSFM active matrix for each dataset. The details of each dataset are shown in Supporting

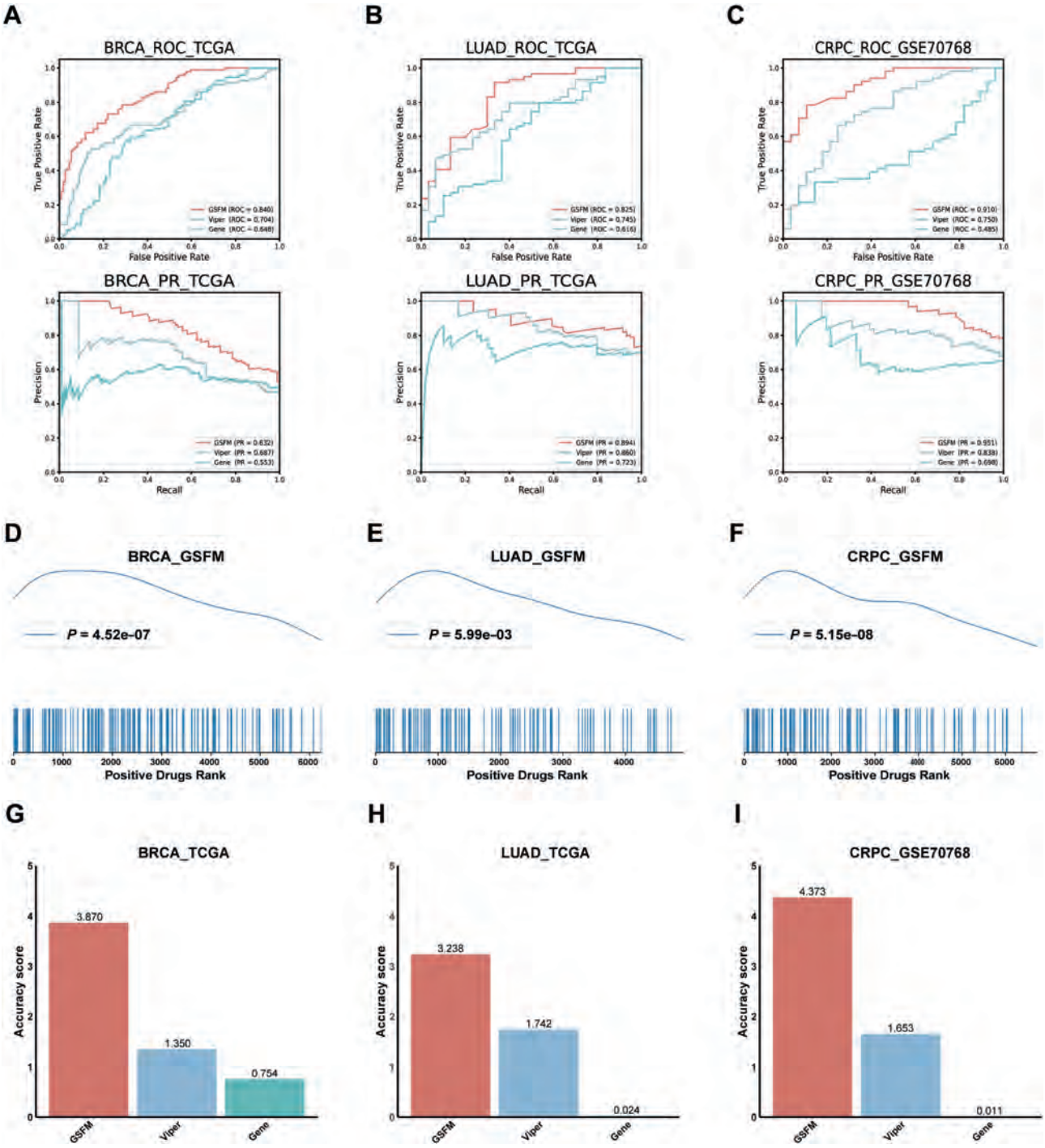
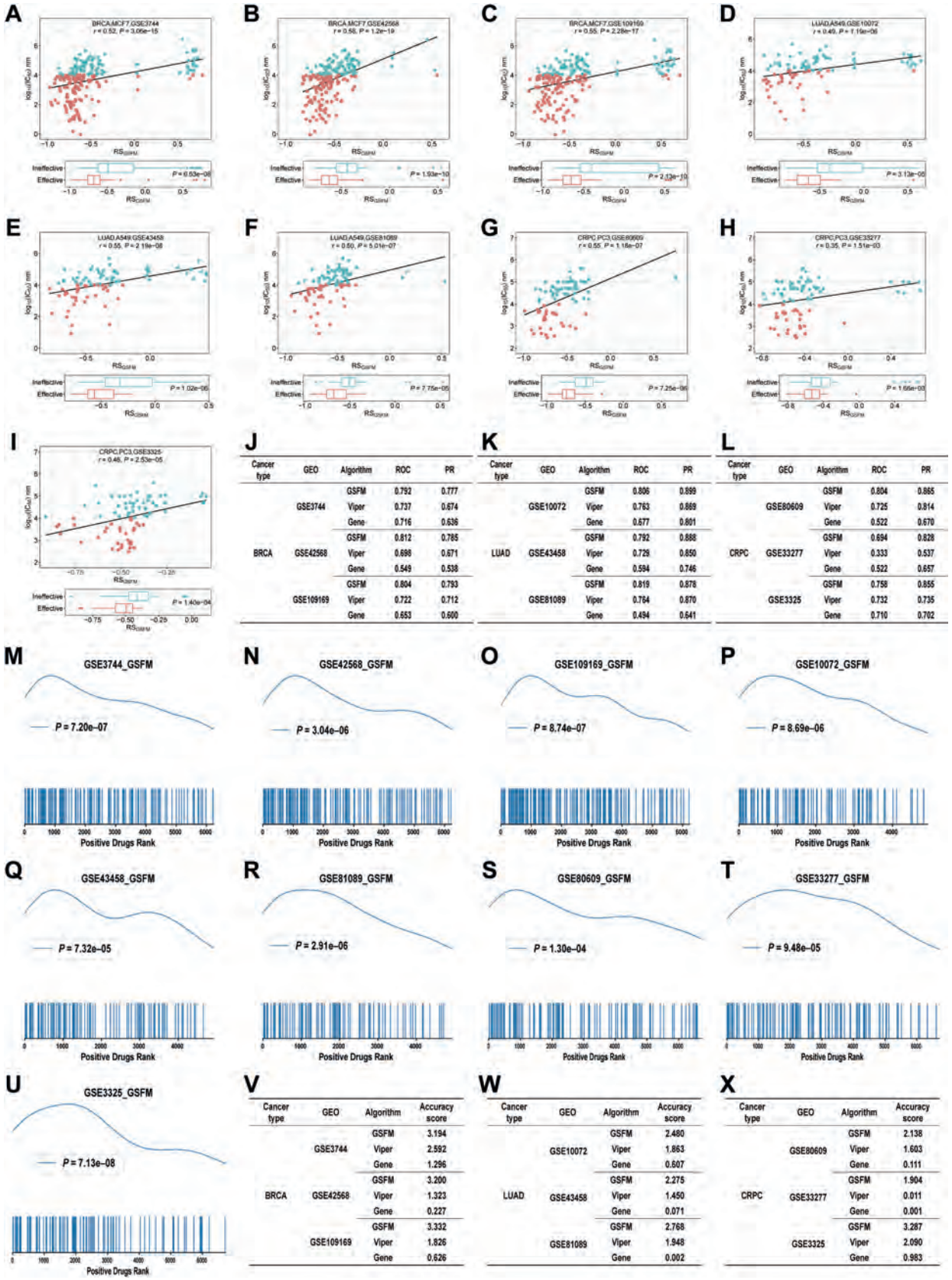


Figure 3 Evaluation metrics for GSFM performance. (A) ROC and PR for BRCA. (B) ROC and PR for LUAD. (C) ROC and PR for CRPC. (D) Enrichment analysis for BRCA in GSFM. (E) Enrichment analysis for LUAD in GSFM. (F) Enrichment analysis for CRPC in GSFM. (G) Accuracy score for BRCA. (H) Accuracy score for LUAD. (I) Accuracy score for CRPC.

Information Table S5. RS_{GSFM}, RS_{Gene}, and RS_{Viper} were qualified for each dataset, with RS_{GSFM} showing a higher correlation with drug efficacy across all cancers (Fig. 4A−I, and Fig. S2G−S2X). The higher ROC and PR values in the GSFM model also indicated superior performance of this computational framework (Fig. 4J−L, and Supporting Information Fig. S3). Additionally, a significant enrichment of potential drugs being tested in clinical trials for cancer treatment was observed enrichment at the top of our predictions in each dataset (Fig. 4M−U, and Supporting Information Fig. S4). Moreover, the accuracy scores demonstrated that the performance of the GSFM method outperformed traditional different gene expression analysis and Viper methods in all datasets (Fig. 4V−X, and Supporting Information Fig. S5). These findings validate the reliability and generalizability of the GSFM model, enhancing confidence in its robustness and broader applicability.



Cancer type	GEO	Algorithm	ROC	PR
BRCA	GSE3744	GSFM	0.792	0.777
		Viper	0.737	0.674
		Gene	0.716	0.636
	GSE42568	GSFM	0.812	0.785
		Viper	0.698	0.671
		Gene	0.549	0.538
	GSE109169	GSFM	0.804	0.793
		Viper	0.722	0.712
		Gene	0.653	0.600

Cancer type	GEO	Algorithm	ROC	PR
LUAD	GSE10072	GSFM	0.806	0.899
		Viper	0.763	0.869
		Gene	0.677	0.801
	GSE43458	GSFM	0.792	0.888
		Viper	0.729	0.850
		Gene	0.594	0.746
	GSE81089	GSFM	0.819	0.878
		Viper	0.764	0.870
		Gene	0.494	0.641

Cancer type	GEO	Algorithm	ROC	PR
CRPC	GSE80609	GSFM	0.804	0.865
		Viper	0.725	0.814
		Gene	0.522	0.670
	GSE33277	GSFM	0.694	0.828
		Viper	0.333	0.537
		Gene	0.522	0.657
	GSE3325	GSFM	0.758	0.855
		Viper	0.732	0.735
		Gene	0.710	0.702

Cancer type	GEO	Algorithm	Accuracy score
BRCA	GSE3744	GSFM	3.194
		Viper	2.592
		Gene	1.296
	GSE42568	GSFM	3.200
		Viper	1.323
		Gene	0.227
	GSE109169	GSFM	3.332
		Viper	1.826
		Gene	0.626

Cancer type	GEO	Algorithm	Accuracy score
LUAD	GSE10072	GSFM	2.480
		Viper	1.863
		Gene	0.607
	GSE43458	GSFM	2.275
		Viper	1.450
		Gene	0.071
	GSE81089	GSFM	2.768
		Viper	1.948
		Gene	0.002

Cancer type	GEO	Algorithm	Accuracy score
CRPC	GSE80609	GSFM	2.138
		Viper	1.603
		Gene	0.111
	GSE33277	GSFM	1.904
		Viper	0.011
		Gene	0.001
	GSE3325	GSFM	3.287
		Viper	2.090
		Gene	0.983

Figure 4 Evaluation of performance on validation datasets. (A−I) Relationship between drug efficacy and RS_{GSFM} across diverse cancer GEO datasets. (J−L) ROC and PR evaluation across distinct cancer GEO datasets. (M−U) Enrichment analysis highlighting GSFM in various cancer GEO datasets. (V−X) Accuracy score across different cancer GEO datasets.

3.5. *In vitro validation of drug hits in each individual cancer*

Observing a significant correlation between RS_{GSFM} and drug efficacy, this approach identified candidate compounds with high reversal potency for BRCA, LUAD, and CRPC. Although recent advancements in therapies such as surgery, radiation, and chemotherapy, have been approved, many yield only marginal survival benefits[30]. Thus, more effective therapeutic treatments are critically needed. The Drug Repurposing Hub (version: 3/24/2020), a carefully curated and annotated collection of FDA-approved drugs, clinical trial drugs, and pre-clinical tool compounds, was selected for the user-defined library (Supporting Information Table S6). Extensive data mining along and literature review were performed, focusing on the drug ranking based on RS_{GSFM}. Interestingly, many top-ranked compounds exhibited biological activity against BRCA, LUAD, and CRPC, with the top 50 compounds presented in Fig. 5A−C. Several of these compounds, already indicated for cancer treatment, have advanced to different stages of clinical verification, which lend credibility to our predictions. Three top-ranked candidate compounds were selected for further investigation: WYE-354 (RS_{GSFM}: −0.92) for BRCA, perhexiline (RS_{GSFM}: −0.85) for LUAD, and NTNCB (RS_{GSFM}: −0.89) for CRPC (Fig. 5D−F). WYE-354 and perhexiline have demonstrated anti-tumor activities in multiple cancer cell lines, including those of breast and lung cancer[31-33]. However, the effect of WYE-354 for BRCA and perhexiline for LUAD in animal models remains unstudied. NTNCB against CRPC has not been previously explored in the context of CRPC. As these compounds emerged as top hits from unbiased virtual screening, and are predicted to reverse the functional module expression of each cancer derived from clinical patient samples, they were included for further investigation to validate their therapeutic potential. Additionally, lower-ranked compounds such as dantron (Rank: 1571) for BRCA, RO-15-4513 (Rank: 997) for LUAD, and bromfenac (Rank: 1258) for CRPC were selected as a negative set for experimental assessment. Paclitaxel, a primary clinical drug extensively used to treat these three cancers, was included as a positive control[34-36].

In vitro experiments were initially conducted to investigate the anti-cancer activities of these compounds on six human cancer cell lines: MDA-MB-231 and MCF7 for BRCA, A549 and NCI-H1975 for LUAD, and PC3 and DU145 for CRPC. Considering the drug properties and stability of perhexiline and NTNCB, subsequent experiments validated the use of perhexiline maleate and NTNCB hydrochloride due to their strong drug absorption and bioavailability. The three compounds were cultured with the corresponding cancer cells at various concentrations for 72 h. All three compounds effectively suppressed cancer cell viability in a concentration-dependent manner. IC_{50} values were calculated to determine compounds' effectiveness against the specific cancer cell lines: WYE-354 exhibited an IC_{50} of 0.644 μmol/L for MCF7 and 1.908 μmol/L for MDA-MB-231; Perhexiline showcased an IC_{50} of 4.629 μmol/L for A549 and 6.768 μmol/L for NCI-H1975; NTNCB displayed an IC_{50} of 6.005 μmol/L for PC3 and 8.303 μmol/L for DU145 (Fig. 5G−I). Meanwhile, the positive control paclitaxel exhibited stronger cell sensitivity to these cell models at 10 μmol/L, validating its use as a positive control. None of the negative control molecules exhibited anti-cancer activities at 10 μmol/L (Fig. 5G−I). These findings suggest that the candidate compounds have significant potential to reverse cancer patterns and demonstrate efficacy *in vitro*, highlighting the predictive capability of GSFM (Fig. 5J−L). Additionally, these compounds also demonstrated notable inhibitory effects on the corresponding tumor cells (Fig. 5M−O). The clonogenic assay revealed concentration-dependent inhibition of colony formation by the three compounds in each cancer cell line. At 1 μmol/L, WYE-354 significantly decreased colony formation in MDA-MB-231 and MCF7 cells by 70.0% ($P < 0.001$) and 63.2% ($P < 0.001$), respectively (Fig. 5P). Perhexiline's suppressive effect on the colony formation of A549 and NCI-H1975 cells was clearly concentration-dependency. At 5 and 10 μmol/L, inhibition rates were 69.1% ($P < 0.001$), 89.6% ($P < 0.001$) for A549 cells, and 17.7% ($P < 0.01$) and 94.8% ($P < 0.001$) for NCI-H1975 cells, respectively (Fig. 5Q). NTNCB showed reductions of 72.2% ($P < 0.001$) and 84.7% ($P < 0.001$) for PC3 cells, and 65.7% ($P < 0.001$) and 93.4% ($P < 0.001$) for DU145 cells at concentrations of 10 and 20 μmol/L, respectively (Fig. 5R). There results emphasize the efficacy of these candidates in inhibiting the growth of particular specific cancer cell lines, indicating their promising therapeutic potential *in vitro* validation. This provides robust support for advancing further research and exploring potential clinical applications.

3.6. *In vivo validation of drug hits in each individual cancer*

To evaluate the anti-cancer effectiveness of WYE-354, perhexiline, and NTNCB *in vivo*, a study was conducted using xenograft tumor models generated from MDA-MB-231, A549, and PC3 cancer cells in nude mice (Fig. 6A). Paclitaxel was included as a positive control. The results demonstrated significant tumor growth suppression following treatment with WYE-354, perhexiline, and NTNCB (Fig. 6B−D). Compared with the vehicle group, WYE-354 at doses of 1, 3, and 10 mg/kg exhibited tumor weight inhibition rates of 28.3%, 55.4%, and 76.6% (Fig. 6E), and tumor volume inhibition rates of 30.7%, 60.0%, and 79.5%, respectively (Fig. 6E). Perhexiline at doses of 1, 3, and 10 mg/kg exhibited tumor weight inhibition rates of 23.9%, 45.8%, and 53.9% (Fig. 6F), and tumor volume inhibition rates of 21.2%, 43.4%, and 59.8%, respectively (Fig. 6F). NTNCB at doses of 1, 3, and 10 mg/kg exhibited tumor weight inhibition rates of 31.3%, 48.5%, and 67.6% (Fig. 6G), and tumor volume inhibition rates of 29.3%, 52.7% and 67.4% for, respectively (Fig. 6G). The average body weight of mice treated with WYE-354, perhexiline, and NTNCB did not show significant differences compared to the vehicle group (Supporting Information Fig. S6A−S6C).

The anti-tumor effects of WYE-354, perhexiline, and NTNCB were further investigated using pathological analysis (H&E, Ki67, and TUNEL assays). H&E staining was employed to examine the physiological morphology of mouse tumors, revealing a significant increase in the necrotic area within tumor tissues of xenografts treated with WYE-354 (10 mg/kg), perhexiline (10 mg/kg), and NTNCB (10 mg/kg) (Fig. 6H−J and Fig. S6D−S6F). The biomarker Ki67 was used to estimate tumor proliferation, showing a significant reduction in tumor cell proliferation following treatment with WYE−354 (10 mg/kg), perhexiline (10 mg/kg), and NTNCB (10 mg/kg) (Fig. 6H−J and Fig. S6D−S6I). Apoptosis in tumor cells was assessed using a TUNEL detection kit, where the nuclei and fragmented DNA of tumor cells were labeled with blue and green fluorescence, respectively. TUNEL analysis revealed that treatment with WYE-354 (10 mg/kg), perhexiline (10 mg/kg), and NTNCB (10 mg/kg) significantly enhanced apoptosis in the tumor tissues (Fig. 6K−M, and Fig. S6J−S6L). The observed changes in tumor volume were consistent with the aforementioned findings. Additionally, there

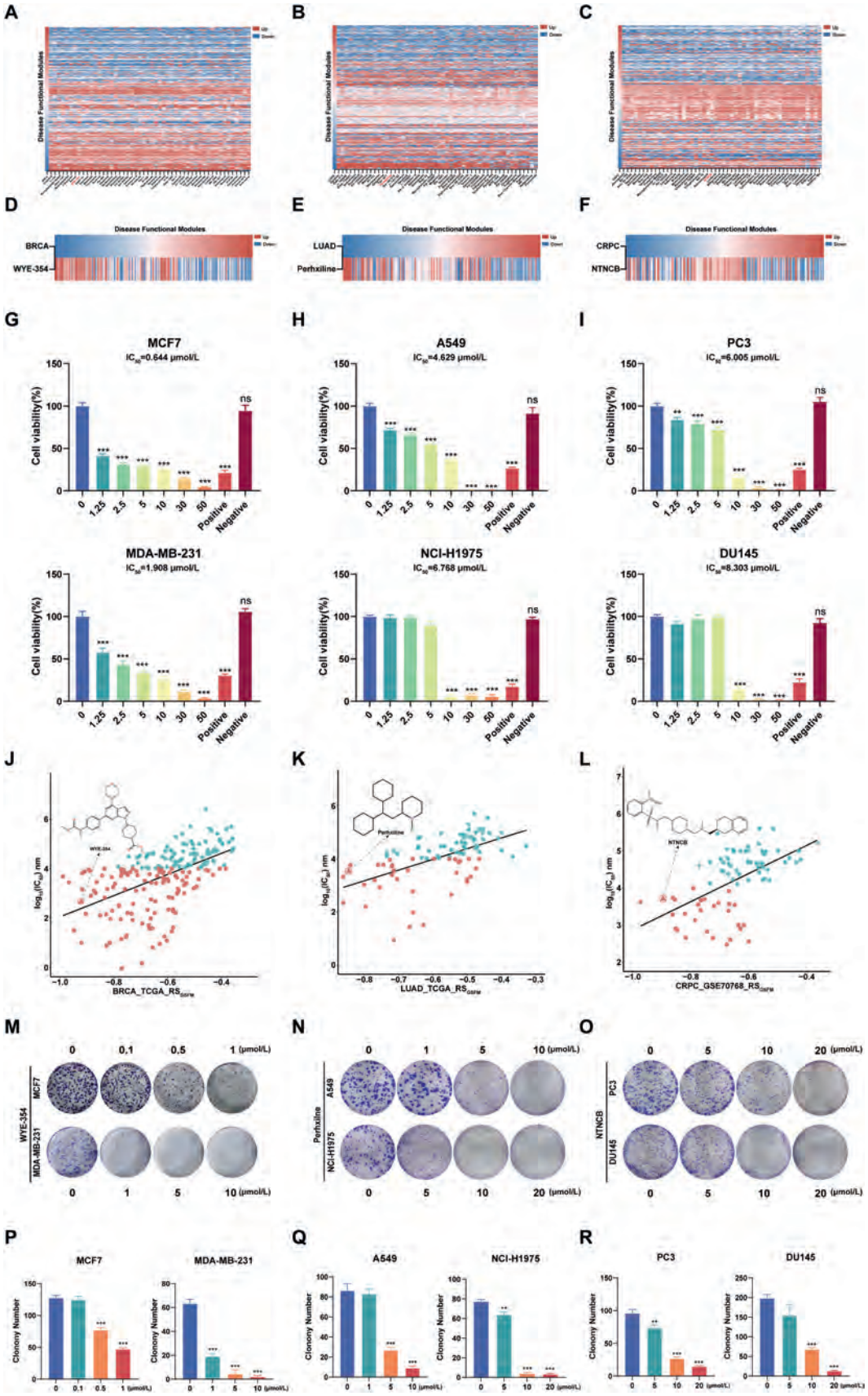


Figure 5 *In vitro* validation of drug hits in each individual cancer. (A–C) Reversal relationship between each cancer GSFM pattern and compound GSFM pattern, including BRCA (A), LUAD (B), and CRPC (C). The first column represents the disease GSFM pattern ranked by *Z*

candidate compounds showed no evident toxicity on major organs, including the heart, liver, spleen, lungs, and kidney (Fig. 6N–P, and Fig. S6M–S6O). Overall, the results strongly suggest that WYE-354, perhexiline, and NTNCB effectively inhibit tumor growth both *in vivo* and *in vitro*. Notably, these compounds had no discernible effect on the average body weight of treated mice compared to the control group and did not produce significant host toxicity.

3.7. Multi-level activities of functional modules reveal mechanisms of action for each candidate

To elucidate the therapeutic mechanisms of these compounds against cancer, an extensive comparative analysis was performed between the drug's GSFM active matrix and the disease-specific GSFM active matrix. For WYE-354 against BRCA, the GSFM active matrix indicated a significant upregulation in Hypoxia and WNT β catenin signaling at both gene and pathway levels, alongside a concurrent downregulation of E2F targets and G2M checkpoint across gene, pathway, and transcriptional regulatory network levels (Fig. 7A). These functional modules potentially enhance breast cancer outcomes by impairing migration and invasion and affecting apoptosis or proliferation[37-40]. For perhexiline against LUAD, the GSFM active matrix demonstrated a marked elevation in Cholesterol homeostasis, juxtaposed with a decline in MYC targets V2, E2F targets, and G2M checkpoint at least two levels (Fig. 7B), which could mitigate lung cancer risk and influence apoptosis or proliferation of lung cancer cells[41-43]. For NTNCB against CRPC, the GSFM active matrix showed a notable increase in Cholesterol homeostasis, accompanied by a decrease in E2F targets, G2M checkpoint, and DNA repair subsequent to NTNCB treatment at multiple levels (Fig. 7C). These functional modules could affect the proliferation, migration, and invasion of prostate cancer cells[44-47].

The GSFM transformation framework also elucidates the critical roles of TFs in regulating various disease states. As shown in Fig. 7D–F, TFs exhibited significant differential expression following drug treatment, with their target genes also significantly enriched in differentially regulated FMs. In BRCA, higher expression of FOXM1, which decreases after WYE-354 treatment, was particularly notable within the G2M checkpoint module and E2F targets modules (Fig. 7D). Previous studies have demonstrated that FOXM1 suppression in cancer cells results in reduced cell proliferation and diminished invasive and migratory capabilities[48]. This aligns with the observed *in vivo* and *in vitro* experimental effects of WYE-354 in breast cancer treatment. Interestingly, FOXO1, FOXO3, and FOXO4 showed upregulation following WYE-354 treatment. These transcription factors have known tumor-suppressive effects upon activation, inhibiting both tumor growth and metastasis[49]. Additionally, the activation of TFDP1 may facilitate tumor progression, so its downregulation could inhibit tumor growth[50]. In LUAD, perhexiline treatment resulted in decreased FOXM1 expression (Fig. 7E), corroborating its role in inhibiting lung cancer cell proliferation[51]. Moreover, previous studies have shown that FOXA1 siRNA causes G0/G1 phase cell cycle arrest and reduced the invasion, migration, and proliferation abilities of the A549 cells[52]. Thus, perhexiline treatment may promote apoptosis and inhibit the proliferation of A549 cells by downregulating FOXA1 expression. Furthermore, inhibited KLF6 transcription could facilitated NSCLC growth[53]. Therefore, perhexiline may inhibit lung cancer growth by activating KLF6. In CRPC, NTNCB treatment resulted in downregulated PAX2 expression (Fig. 7F), consistent with findings suggesting that PAX2 knockdown inhibits prostate tumor growth[54]. NTNCB also decreased the expression of PAX4 and PAX6, both known to suppress prostate cancer[55,56]. Additionally, NTNCB could activate GATA6, thereby restraining prostate cancer cell proliferation, metastasis, and tumor growth[57]. In summary, WYE-354's exhibits potential anti-BRCA effects by modulating TFs such as FOXM1, TFDP1, FOXO1, FOXO3, and FOXO4; perhexiline's impacts LUAD growth and proliferation through its control over FOXM1, FOXA1, and KLF6; NTNCB's regulates PAX2, PAX4, PAX6, and GATA6, suggesting its therapeutic potential in inhibiting CRPC development. Analyzing these factors using GSFM provides profound insights into their therapeutic mechanisms, particularly their roles in modulating TFs and essential functional modules.

3.8. Data availability

GSFM is implemented in Python and R, which can be accessed at https://github.com/LXY1201/GSFM. Additional data supporting the conclusions of this study are available upon request from the corresponding author.

4. Discussion

The precision medicine initiative seeks to unveil new drugs tailored to diseases defined at the molecular level[58]. The conventional target-based drug discovery approach, focusing on disrupting individual targets, faces challenges such as inadequate drug efficacy, drug resistance, and off-target effects[59]. Recently, the identification of drugs capable of reversing disease gene expression has emerged as a promising alternative approach, and demonstrated potential in drug discovery[9,13,60]. However, most gene expression-based approaches primarily rely on differentially expressed genes, often neglecting gene–gene functional networks.

score. The remaining columns represent GSFM pattern change after treatment with different compounds. Compounds (highlighted in red) were selected based on their $RS_{GSFM,}$ and novelty within BRCA, LUAD, and CRPC, respectively. (D) Reversal relationship between BRCA GSFM pattern and WYE-354 GSFM pattern. (E) Reversal relationship between LUAD GSFM pattern and perhexiline GSFM pattern. (F) Reversal relationship between CRPC GSFM pattern and NTNCB GSFM pattern. (G–I) Drug efficacy in six cancer lines, including MDA-MB-231 (G, positive= paclitaxel, and negative= dantron, 10 μmol/L), MCF7 (G, positive= paclitaxel, and negative= dantron, 10 μmol/L), A549 (H, positive= paclitaxel, and negative= RO-15-4513, 10 μmol/L), NCI-H1975 (H, positive= paclitaxel, and negative= RO-15-4513, 10 μmol/L), PC3 (I, positive= paclitaxel, and negative= bromfenac, 10 μmol/L), and DU145 (I, positive= paclitaxel, and negative= bromfenac, 10 μmol/L), measured by cell proliferation assays after 72 h treatment. (J–L) The correlation between drug efficacy and RS_{GSFM} in each individual cancer, including BRCA (J), LUAD (K), and CRPC (L). (M, P) WYE-354 demonstrated inhibition of colony formation in MCF7 and MDA-MB-231 cells. (N, Q) Perhexiline exhibited inhibition of colony formation in A549 and NCI-H1975 cells. (O, R) NTNCB showed inhibition of colony formation in PC3 and DU145 cells. Data represents mean ± standard error of the mean (SEM) from three independent biological replicates. $*P < 0.05$; $**P < 0.01$; $***P < 0.001$.

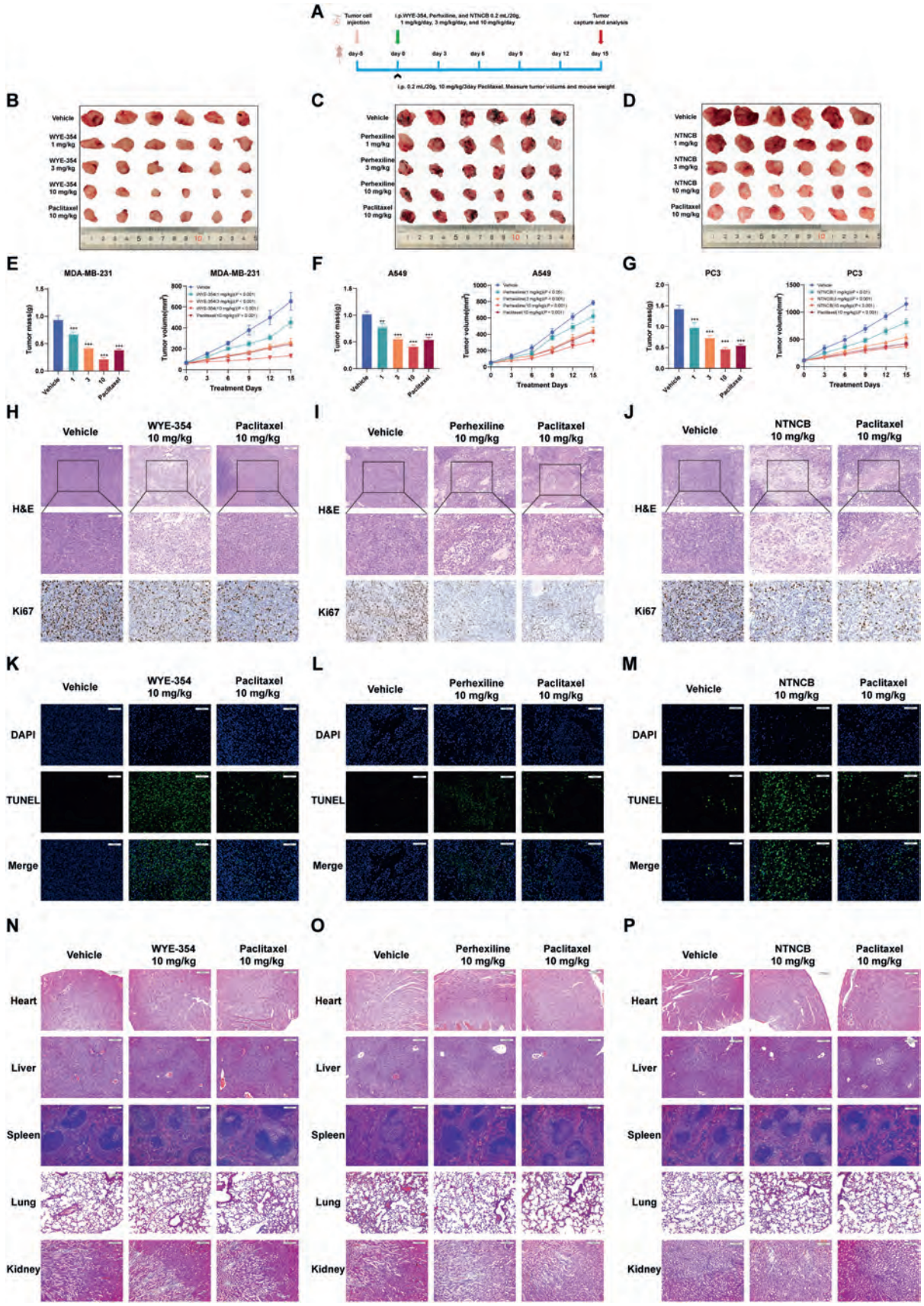


Figure 6 *In vivo* validation of drug hits in each individual cancer. (A) Tumor xenograft mouse model and treatment strategy. Mice ($n = 6$ per group) were utilized for the xenograft mouse model (blue syringe). Candidate drugs were administered intraperitoneally at various concentrations

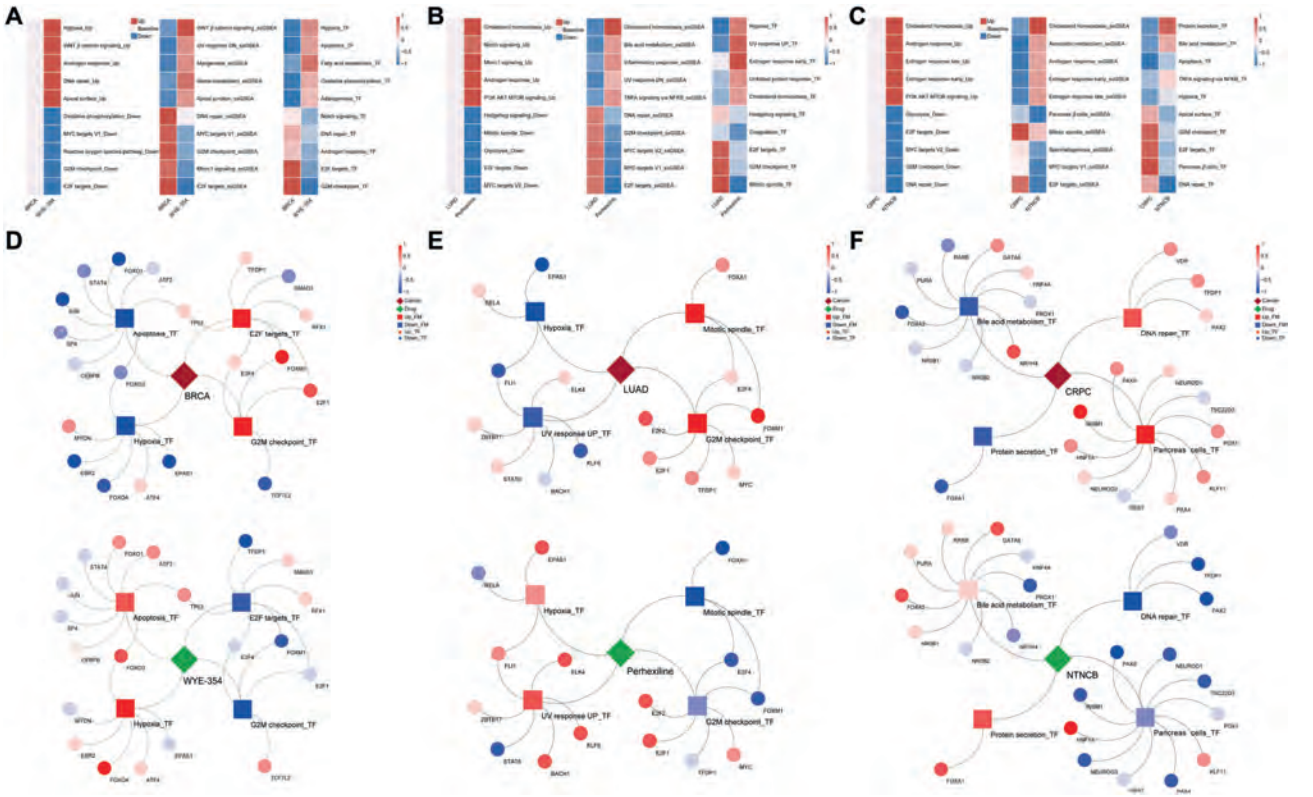


Figure 7 Mechanistic prediction of drug candidates. (A) The heatmap depicts the top ten reversals of FM (gene, pathway, and transcriptional regulatory network levels) in BRCA and treatment with WYE-354. (B) The heatmap depicts the top ten reversals of FM (gene, pathway, and transcriptional regulatory network levels) in LUAD and treatment with perhexiline. (C) The heatmap depicts the top ten reversals of FM (gene, pathway, and transcriptional regulatory network levels) in CRPC and treatment with NTNCB. (D) The network illustrates four FM_TF within the context of BRCA and treatment with WYE-354. This visualization showcases the interconnectedness and roles of transcription factors influenced by WYE-354 in BRCA. (E) The network illustrates four FM_TF within the context of LUAD and treatment with perhexiline. This visualization showcases the interconnectedness and roles of transcription factors influenced by perhexiline in LUAD. (F) The network illustrates four FM_TF within the context of CRPC and treatment with NTNCB. This visualization showcases the interconnectedness and roles of transcription factors influenced by NTNCB in CRPC.

This limitation reduces robustness and accuracy. Additionally, a systematic and in-depth examination of the relationship between reversal potency and drug efficacy remains lacking.

Here, we propose a genome-scale functional module-based transformation strategy (GSFM), representing a new generation in drug discovery systems. From a system perspective, this

daily for 15 days (1, 3, and 10 mg/kg. Green arrow). Paclitaxel was administered intraperitoneally every 3 days for 15 days (10 mg/kg. Black arrow). (B) The photographs of tumors derived from naked mice in representative MDA-MB-231 tumors *in vivo*. (C) The photographs of tumors derived from naked mice in representative A549 tumors *in vivo*. (D) The photographs of tumors derived from naked mice in representative PC3 tumors *in vivo*. (E) Quantitative analysis of the MDA-MB-231 cells tumor mass and tumor volume. (F) Quantitative analysis of the A549 cells tumor mass and tumor volume. (G) Quantitative analysis of the PC3 cells tumor mass and tumor volume. (H) Results of H&E and Ki67 staining in representative MDA-MB-231 cells tumor sections of vehicle, WYE-354 (10 mg/kg), and paclitaxel (10 mg/kg) treated naked mice. Scale bar = 200 μm (H&E up). Scale bar = 100 μm (H&E down and Ki67). (I) Results of H&E and Ki67 staining in representative A549 cells tumor sections of vehicle, perhexiline (10 mg/kg), and paclitaxel (10 mg/kg) treated naked mice. Scale bar = 200 μm (H&E up). Scale bar = 100 μm (H&E down and Ki67). (J) Results of H&E and Ki67 staining in representative PC3 cells tumor sections of vehicle, NTNCB (10 mg/kg), and paclitaxel (10 mg/kg) treated naked mice. Scale bar = 200 μm (H&E up). Scale bar = 100 μm (H&E down and Ki67). (K) Results of TUNEL staining in representative MDA-MB-231 cells tumor sections of vehicle, WYE-354 (10 mg/kg), and paclitaxel (10 mg/kg) treated naked mice, blue—nuclei and green—TUNEL. Scale bar = 100 μm. (L) Results of TUNEL staining in representative A549 cells tumor sections of vehicle, perhexiline (10 mg/kg), and paclitaxel (10 mg/kg) treated naked mice, blue—nuclei and green—TUNEL. Scale bar = 100 μm. (M) Results of TUNEL staining in representative PC3 cells tumor sections of vehicle, NTNCB (10 mg/kg), and paclitaxel (10 mg/kg) treated naked mice, blue—nuclei and green—TUNEL. Scale bar = 100 μm. (N) H&E staining results of five main organs from MDA-MB-231 xenograft model mice (vehicle, WYE-354 (10 mg/kg), and paclitaxel (10 mg/kg)). Scale bar = 100 μm. (O) H&E staining results of five main organs from A549 xenograft model mice (vehicle, perhexiline (10 mg/kg), and paclitaxel (10 mg/kg)). Scale bar = 100 μm. (P) H&E staining results of five main organs from PC3 xenograft model mice (vehicle, NTNCB (10 mg/kg), and paclitaxel (10 mg/kg)). Scale bar = 100 μm. $^{*}P < 0.05$; $^{**}P < 0.01$; $^{***}P < 0.001$.

computational framework analyzes multi-level activities, including gene-level, pathway-level, and transcriptional regulatory network-level, transforming high-dimensional data into low-dimensional active data. Compared to other module-based scoring approaches like ssGSEA and Viper, GSFM enhances biological interpretability by offering multiple activity levels for a single module. Using FMs as transcriptional features, GSFM successfully reveals the correlation between reversal potency and drug efficacy. The optimal BIQ combination and FM size, critical for GSFM's performance, were identified through robust analysis. The combination of four biologically interpretable quantifiers consistently outperformed others, with $x = 125$ determined as the optimal parameter. Systematic performance comparisons with other drug prediction algorithms, including CMap, XSum, and ZhangScore, showed that the RS algorithm in GSFM outperforms these alternatives. Specifically, the advantages of RS algorithm in our GSFM computational framework include: (1) it is not restricted by the drug/disease differential expression FMs and uses all FMs to enhance method reliability[13]; (2) it focuses on the reversal relationship between the disease and drug candidates, yielding superior better in drug response prediction[9]; (3) it shows a positive correlation with drug efficacy[13]; (4) it consumes fewer computer resources and user-friendly. Overall, The RS method has its value for drug discovery, especially when computational resources and biological knowledge are limited. This large-scale computational approach effectively predicts potential new drug candidates for various cancer types. Case studies on BRCA, LUAD, and CRPC illustrate the correlation between drug reversal potency (RS_{GSFM} of drug) and drug efficacy. The positive correlation observed between RS_{GSFM} and IC_{50} suggests that combining disease expression obtained from clinical samples and drug expression profiled *in vitro* can predict drug efficacy. Notably, RS_{GSFM} encapsulates the molecular characteristics of clinical samples, distinguishing it from high-throughput screening technologies that primarily evaluate drug activity in specific cell lines. WYE-354, perhexiline, and NTNCB were identified as potential treatments for BRCA, LUAD, and CRPC, respectively. Their therapeutic efficacy was confirmed in experimental models using subcutaneous xenograft tumors, validating the accuracy of the initial predictions. Importantly, our proposed method is not limited to anti-tumor drug screening and it can also be applied to other disorders. The main changes required are the input data, which must be tailored to the type of disease under study. The key steps include: users need to provide functional modules that are strongly related to the disease of interest; users need to collect expression matrix for the disease as well as the corresponding normal control samples. Then, these data should then be input into the GSFM model for candidate drug screening. Users only need to prepare the aforementioned files and follow our documentation (https://github.com/LXY1201/GSFM) to input the data into our GSFM algorithm, and they will be able to obtain the corresponding results.

Besides, it is crucial to recognize that the choice of modules can differ depending on the particular research question being explored. For different studies, users must collect modules strongly related to diseases based on background knowledge. Of course, users can also define their own modules of interest. These modules should cover diverse cellular activities and also show mutual exclusivity across modules. It is essential to ensure that the selected modules are highly relevant to the research subject and biologically interpretable. While selecting of modules from different sources to lead to slight variations in the results, as long as the disease-related modules are well-aligned with the disease under study, the top-ranked candidate drugs will remain consistent. We also recommend that users input modules from different sources and use our GSFM model for drug prediction. By conducting a comprehensive analysis, such as intersecting the drug outputs from different sources, the results can be made more robust. However, certain limitations of this work must be acknowledged. Firstly, accurately defining the functional modules poses challenges when applying GSFM to new datasets without prior knowledge. Secondly, additional experimental and clinical studies are essential to validate and further elucidate the mechanisms of the identified candidates in each individual cancer. Thirdly, our method may not be applicable to certain rare diseases when normal samples are unavailable. Despite these constraints, the GSFM transformer strategy effectively qualifies multi-level activities of FMs. Surely, when normal samples are missing, we also recommend users combine the usage of the GTEx database, which covers nearly 54 types of normal tissue samples. Furthermore, the LINCS dataset includes nearly 20 tumors and 158 cell lines, so the GSFM computational framework can be used for a wide range of anti-tumor drug screenings. For cancer types not represented in LINCS, we recommend using the cell line most similar to the target cell line as a data source for drug screening.

In conclusion, the GSFM framework offers a novel approach to *in silico* drug discovery and serves as a valuable tool for assessing drug efficacy. Distinguished by its modularization, reproducibility, robustness, and flexibility, this methodology holds substantial promise for immediate application in preventive and personalized medicine.

5. Conclusions

In this study, we proposed GSFM, a novel genome-scale functional module transformation framework to quantitatively evaluate drug efficacy for *in silico* drug discovery. Unlike previous methods, GSFM transformation framework could comprehensively evaluate the activities of each FM at gene-level, pathway-level and transcriptional regulatory network-level using four biologically interpretable quantifiers: GSFM_Up, GSFM_Down, GSFM_ssGSEA, and GSFM_TF. Moreover, the ablation study confirmed the validity of incorporating the four quantifiers into the framework. Using GSFM, we could effectively convert the noisy and seemingly "unreliable" gene expression data into more "reliable" functional module data for drug discovery. Meanwhile, we systematically analyzed the model performance in anti-cancer drug prediction, and the results showed that GSFM achieved excellent performance on multi-datasets. Besides, the candidate drugs were validated through *in vitro* and *in vivo* experiments, providing proof-of-concept for our computational approach. Therefore, this work not only showcases the feasibility and potential of GSFM as a valuable tool in drug discovery but also highlights its versatile and widely applicable framework within the realm of drug development.

Acknowledgments

This work was funded by the National Key Research and Development Program of China (2022YFC3502000), the National Natural Science Foundation of China (82141203, 82274172, 82430119), Shanghai Municipal Science and Technology Major Project (ZD2021CY001), Key project at central government level: The ability establishment of sustainable use for valuable Chinese

medicine resources (2060302), Innovation Team and Talents Cultivation Program of National Administration of Traditional Chinese Medicine (ZYYCXTDD-202004), the support of Wild Goose Array Project, Zhengzhou Center of PLAJLSF, the Chenguang Program of Shanghai Education Development Foundation and Shanghai Municipal Education Commission (23CGA45, Saisai Tian) and Tianjin Health Research Project (TJWJ2024QN100).

Author contributions

Weidong Zhang, Luonan Chen, and Saisai Tian: conceptualization, original draft, methodology, review and editing, funding acquisition, and supervision. Saisai Tian, Xuyang Liao, and Wen Cao analyzed the data, carried out the experiments, generated the figures, and wrote the paper. Xinyi Wu, Zexi Chen, Jinyuan Lu, Qun Wang, and Jinbo Zhang participated in part of the experiments. All authors read and approved the final manuscript.

Conflicts of interest

The authors declare that they have no competing interests.

Appendix A. Supporting information

Supporting information to this article can be found online at https://doi.org/10.1016/j.apsb.2024.08.017.

References

1. Shaker B, Ahmad S, Lee J, Jung C, Na D. In silico methods and tools for drug discovery. *Comput Biol Med* 2021;**137**:104851.
2. Alhumaydhi FA. Integrated computational approaches to screen gene expression data to determine key genes and therapeutic targets for type-2 diabetes mellitus. *Saudi J Biol Sci* 2022;**29**:3276−86.
3. Bates S. The role of gene expression profiling in drug discovery. *Curr Opin Pharmacol* 2011;**11**:549−56.
4. Szalai B, Veres DV. Application of perturbation gene expression profiles in drug discovery—from mechanism of action to quantitative modelling. *Front Syst Biol* 2023;**3**:1126044.
5. Lamb J, Crawford ED, Peck D, Modell JW, Blat IC, Wrobel MJ, et al. The connectivity map: using gene-expression signatures to connect small molecules, genes, and disease. *Science* 2006;**313**:1929−35.
6. Koleti A, Terryn R, Stathias V, Chung C, Cooper DJ, Turner JP, et al. Data portal for the library of integrated network-based cellular signatures (lincs) program: integrated access to diverse large-scale cellular perturbation response data. *Nucleic Acids Res* 2018;**46**: D558−66.
7. Tian S, Zhang J, Yuan S, Wang Q, Lv C, Wang J, et al. Exploring pharmacological active ingredients of traditional Chinese medicine by pharmacotranscriptomic map in ITCM. *Brief Bioinform* 2023;**24**: bbad027.
8. Rabben HL, Andersen GT, Ianevski A, Olsen MK, Kainov D, Grønbech JE, et al. Computational drug repositioning and experimental validation of ivermectin in treatment of gastric cancer. *Front Pharmacol* 2021;**12**:625991.
9. Yang C, Zhang H, Chen M, Wang S, Qian R, Zhang L, et al. A survey of optimal strategy for signature-based drug repositioning and an application to liver cancer. *Elife* 2022;**11**:e71880.
10. Lamb J. The connectivity map: a new tool for biomedical research. *Nat Rev Cancer* 2007;**7**:54−60.
11. Cheng J, Yang L, Kumar V, Agarwal P. Systematic evaluation of connectivity map for disease indications. *Genome Med* 2014;**6**:540.
12. Zhang SD, Gant TW. A simple and robust method for connecting small-molecule drugs using gene-expression signatures. *BMC Bioinf* 2008;**9**:258.
13. Chen B, Ma L, Paik H, Sirota M, Wei W, Chua MS, et al. Reversal of cancer gene expression correlates with drug efficacy and reveals therapeutic targets. *Nat Commun* 2017;**8**:16022.
14. Cui H, Srinivasan S, Korkin D. Enriching human interactome with functional mutations to detect high-impact network modules underlying complex diseases. *Genes* 2019;**10**:933.
15. Khan A, Rehman Z, Hashmi HF, Khan AA, Junaid M, Sayaf AM, et al. An integrated systems biology and network-based approaches to identify novel biomarkers in breast cancer cell lines using gene expression data. *Interdiscip Sci* 2020;**12**:155−68.
16. Silverman EK, Schmidt H, Anastasiadou E, Altucci L, Angelini M, Badimon L, et al. Molecular networks in network medicine: development and applications. *Wiley Interdiscip Rev Syst Biol Med* 2020; **12**:e1489.
17. Tornow S, Mewes HW. Functional modules by relating protein interaction networks and gene expression. *Nucleic Acids Res* 2003;**31**: 6283−9.
18. Hänzelmann S, Castelo R, Guinney J. GSVA: gene set variation analysis for microarray and RNA-seq data. *BMC Bioinf* 2013;**14**:7.
19. Alvarez MJ, Shen Y, Giorgi FM, Lachmann A, Ding BB, Ye BH, et al. Functional characterization of somatic mutations in cancer using network-based inference of protein activity. *Nat Genet* 2016;**48**: 838−47.
20. Choobdar S, Ahsen ME, Crawford J, Tomasoni M, Fang T, Lamparter D, et al. Assessment of network module identification across complex diseases. *Nat Methods* 2019;**16**:843−52.
21. Duan Y, Evans DS, Miller RA, Schork NJ, Cummings SR, Girke T. signatureSearch: environment for gene expression signature searching and functional interpretation. *Nucleic Acids Res* 2020;**48**:e124.
22. Garcia-Alonso L, Holland CH, Ibrahim MM, Turei D, Saez-Rodriguez J. Benchmark and integration of resources for the estimation of human transcription factor activities. *Genome Res* 2019;**29**: 1363−75.
23. Liberzon A, Birger C, Thorvaldsdóttir H, Ghandi M, Mesirov JP, Tamayo P. The molecular signatures database (msigdb) hallmark gene set collection. *Cell Syst* 2015;**1**:417−25.
24. Targonski CA, Shearer CA, Shealy BT, Smith MC, Feltus FA. Uncovering biomarker genes with enriched classification potential from Hallmark gene sets. *Sci Rep* 2019;**9**:9747.
25. Qin G, Knijnenburg TA, Gibbs DL, Moser R, Monnat Jr RJ, Kemp CJ, et al. A functional module states framework reveals transcriptional states for drug and target prediction. *Cell Rep* 2022;**38**:110269.
26. de Hond AAH, Steyerberg EW, van Calster B. Interpreting area under the receiver operating characteristic curve. *Lancet Digit Health* 2022; **4**:e853−5.
27. Cabot JH, Ross EG. Evaluating prediction model performance. *Surgery* 2023;**174**:723−6.
28. Chen EY, Tan CM, Kou Y, Duan Q, Wang Z, Meirelles GV, et al. Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool. *BMC Bioinf* 2013;**14**:128.
29. Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, et al. Limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res* 2015;**43**:e47.
30. Debela DT, Muzazu SG, Heraro KD, Ndalama MT, Mesele BW, Haile DC, et al. New approaches and procedures for cancer treatment: current perspectives. *SAGE Open Med* 2021;**9**:20503121211034366.
31. Yu K, Toral-Barza L, Shi C, Zhang WG, Lucas J, Shor B, et al. Biochemical, cellular, and *in vivo* activity of novel ATP-competitive and selective inhibitors of the mammalian target of rapamycin. *Cancer Res* 2009;**69**:6232−40.
32. Dhakal B, Tomita Y, Drew P, Price T, Maddern G, Smith E, et al. Perhexiline: old drug, new tricks? A summary of its anti-cancer effects. *Molecules* 2023;**28**:3624.
33. Rodríguez-Enríquez S, Hernández-Esquivel L, Marín-Hernández A, El Hafidi M, Gallardo-Pérez JC, Hernández-Reséndiz I, et al.

Mitochondrial free fatty acid β-oxidation supports oxidative phosphorylation and proliferation in cancer cells. *Int J Biochem Cel Biol* 2015;**65**:209−21.
34. Tolaney SM, Tarantino P, Graham N, Tayob N, Parè L, Villacampa G, et al. Adjuvant paclitaxel and trastuzumab for node-negative, HER2-positive breast cancer: final 10-year analysis of the open-label, single-arm, phase 2 APT trial. *Lancet Oncol* 2023;**24**:273−85.
35. Sugawara S, Lee JS, Kang JH, Kim HR, Inui N, Hida T, et al. Nivolumab with carboplatin, paclitaxel, and bevacizumab for first-line treatment of advanced nonsquamous non-small-cell lung cancer. *Ann Oncol* 2021;**32**:1137−47.
36. Ruiz de Porras V, Font A, Aytes A. Chemotherapy in metastatic castration-resistant prostate cancer: current scenario and future perspectives. *Cancer Lett* 2021;**523**:162−9.
37. Tutzauer J, Sjöström M, Holmberg E, Karlsson P, Killander F, Leeb-Lundberg LMF, et al. Breast cancer hypoxia in relation to prognosis and benefit from radiotherapy after breast-conserving surgery in a large, randomised trial with long-term follow-up. *Br J Cancer* 2022;**126**:1145−56.
38. Xu X, Zhang M, Xu F, Jiang S. Wnt signaling in breast cancer: biological mechanisms, challenges and opportunities. *Mol Cancer* 2020;**19**:165.
39. Oshi M, Takahashi H, Tokumaru Y, Yan L, Rashid OM, Nagahashi M, et al. The E2F pathway score as a predictive biomarker of response to neoadjuvant therapy in $ER^+/HER2^-$ breast cancer. *Cells* 2020;**9**:1643.
40. Ashwell S, Zabludoff S. DNA damage detection and repair pathways–recent advances with inhibitors of checkpoint kinases in cancer therapy. *Clin Cancer Res* 2008;**14**:4032−7.
41. Hoppstädter J, Dembek A, Höring M, Schymik HS, Dahlem C, Sultan A, et al. Dysregulation of cholesterol homeostasis in human lung cancer tissue and tumour-associated macrophages. *EBioMedicine* 2021;**72**:103578.
42. Dhanasekaran R, Deutzmann A, Mahauad-Fernandez WD, Hansen AS, Gouw AM, Felsher DW. The MYC oncogene−the grand orchestrator of cancer growth and immune evasion. *Nat Rev Clin Oncol* 2022;**19**:23−36.
43. Park SM, Choi EY, Bae DH, Sohn HA, Kim SY, Kim YJ. The LncRNA EPEL Promotes lung cancer cell proliferation through e2f target activation. *Cell Physiol Biochem* 2018;**45**:1270−83.
44. El-Kenawi A, Dominguez-Viqueira W, Liu M, Awasthi S, Abraham-Miranda J, Keske A, et al. Macrophage-Derived cholesterol contributes to therapeutic resistance in prostate cancer. *Cancer Res* 2021;**81**:5477−90.
45. Zhou Q, Wang C, Zhu Y, Wu Q, Jiang Y, Huang Y, et al. Key genes and pathways controlled by e2f1 in human castration-resistant prostate cancer cells. *Onco Targets Ther* 2019;**12**:8961−76.
46. Chao Y, Chen Y, Zheng W, Demanelis K, Liu Y, Connelly JA, et al. Synthetic lethal combination of CHK1 and WEE1 inhibition for treatment of castration-resistant prostate cancer. *Oncogene* 2024;**43**:789−803.
47. Mateo J, Boysen G, Barbieri CE, Bryant HE, Castro E, Nelson PS, et al. DNA repair in prostate cancer: biology and clinical implications. *Eur Urol* 2017;**71**:417−25.
48. Dey P, Wang A, Ziegler Y, Kim SH, El-Ashry D, Katzenellenbogen JA, et al. Suppression of tumor growth, metastasis, and signaling pathways by reducing foxm1 activity in triple negative breast cancer. *Cancers (Basel)* 2020;**12**:2677.
49. Jiramongkol Y, Lam EW. FOXO transcription factor family in cancer and metastasis. *Cancer Metastasis Rev* 2020;**39**:681−709.
50. Moscovich M, LeDoux MS, Xiao J, Rampon GL, Vemula SR, Rodriguez RL, et al. Dystonia, facial dysmorphism, intellectual disability and breast cancer associated with a chromosome 13q34 duplication and overexpression of TFDP1: case report. *BMC Med Genet* 2013;**14**:70.
51. Madhi H, Lee JS, Choi YE, Li Y, Kim MH, Choi Y, et al. FOXM1 inhibition enhances the therapeutic outcome of lung cancer immunotherapy by modulating PD-L1 expression and cell proliferation. *Adv Sci* 2022;**9**:e2202702.
52. Li J, Zhang S, Zhu L, Ma S. Role of transcription factor FOXA1 in non-small cell lung cancer. *Mol Med Rep* 2018;**17**:509−21.
53. Li F, Zhang Q, Gong Y, Yu J. The lncKLF6/KLF6 feedback loop regulates the growth of non-small cell lung cancer. *Am J Cancer Res* 2018;**8**:1427−39.
54. Ueda T, Ito S, Shiraishi T, Kulkarni P, Ueno A, Nakagawa H, et al. Hyper-expression of PAX2 in human metastatic prostate tumors and its role as a cancer promoter in an *in vitro* invasion model. *Prostate* 2013;**73**:1403−12.
55. Shyr CR, Tsai MY, Yeh S, Kang HY, Chang YC, Wong PL, et al. Tumor suppressor PAX6 functions as androgen receptor co-repressor to inhibit prostate cancer growth. *Prostate* 2010;**70**:190−9.
56. Thaper D, Vahid S, Zoubeidi A. Neural transcription factors in disease progression. *Adv Exp Med Biol* 2019;**1210**:437−62.
57. Li Y, Li H, Wei X. Long noncoding RNA LINC00261 suppresses prostate cancer tumorigenesis through upregulation of GATA6-mediated DKK3. *Cancer Cel Int* 2020;**20**:474.
58. Dugger SA, Platt A, Goldstein DB. Drug development in the era of precision medicine. *Nat Rev Drug Discov* 2018;**17**:183−96.
59. Sadri A. Is target-based drug discovery efficient? Discovery and "off-target" mechanisms of all drugs. *J Med Chem* 2023;**66**:12651−77.
60. Pham TH, Qiu Y, Liu J, Zimmer S, O'Neill E, Xie L, et al. Chemical-induced gene expression ranking and its application to pancreatic cancer drug repurposing. *Patterns (N Y)* 2022;**3**:100441.