Computers in Biology and Medicine 174 (2024) 108433

Contents lists available at ScienceDirect

Computers in Biology and Medicine

journal homepage: www.elsevier.com/locate/compbiomed

Exploratory drug discovery in breast cancer patients: A multimodal deep learning approach to identify novel drug candidates targeting RTK signaling

Anush Karampuri, Sunitha Kundur, Shyam Perugu *

Department of Biotechnology, National Institute of Technology, Warangal, 500604, India

ARTICLE INFO

Keywords:
RTK signaling
Multimodal deep learning
Breast cancer
Cell modal passports
GDSC

ABSTRACT

Breast cancer, a highly formidable and diverse malignancy predominantly affecting women globally, poses a significant threat due to its intricate genetic variability, rendering it challenging to diagnose accurately. Various therapies such as immunotherapy, radiotherapy, and diverse chemotherapy approaches like drug repurposing and combination therapy are widely used depending on cancer subtype and metastasis severity. Our study revolves around an innovative drug discovery strategy targeting potential drug candidates specific to RTK signalling, a prominently targeted receptor class in cancer. To accomplish this, we have developed a multimodal deep neural network (MM-DNN) based QSAR model integrating omics datasets to elucidate genomic, proteomic expression data, and drug responses, validated rigorously. The results showcase an R^2 value of 0.917 and an RMSE value of 0.312, affirming the model's commendable predictive capabilities. Structural analogs of drug molecules specific to RTK signalling were sourced from the PubChem database, followed by meticulous screening to eliminate dissimilar compounds. Leveraging the MM-DNN-based QSAR model, we predicted the biological activity of these molecules, subsequently clustering them into three distinct groups. Feature importance analysis was performed. Consequently, we successfully identified prime drug candidates tailored for each potential downstream regulatory protein within the RTK signalling pathway. This method makes the early stages of drug development faster by removing inactive compounds, providing a hopeful path in combating breast cancer.

1. Introduction

Breast cancer, the leading, highly aggressive, and heterogeneous cancer in women globally, presents distinct subtypes with varied molecular features [1,2]. In 2023, Krishnan et al. reported a 5-year age-standardized relative survival rate of 41.9 % for breast cancer patients in India, with 37.8 % mortality within five years among the 14, 148 cases studied [3], while in the United States, it accounted for 31 % of new cases and 15 % of female deaths [1]. Based on ER (Estrogen Receptors), PR (Progesterone Receptors), and HER2 (Human Epidermal Growth Factor Receptors), breast cancer was classified into four subtypes: Luminal A (Low Grade, ER+/PR+, HER-, low Ki67), Luminal B (ER+/PR+, HER + or HER-, high Ki67), TNBC (Triple Negative Breast Cancer) (ER-/PR- and HER2-), and HER2-enriched [4]. At the molecular level, signalling pathways such as RTK (Receptor Tyrosine Kinases), PI3K/MTOR, and ERK-MAPK are few among the targeted pathways of breast cancer [5], as shown in Supplementary Fig. 1 (sourced from GDSC – Genomics of Drug Sensitivity in Cancer). Exploratory data analysis on the drug response data from GDSC (Genomics of drug sensitivity in cancer) unveiled that among the 24 pathways within the dataset, RTK (Receptor Tyrosine Kinase) signalling emerged as the most targeted. Please refer to Supplementary Fig. 1. Various therapies, including radiotherapy [6], immunotherapy [7], adjuvant and neo-adjuvant therapy [8], endocrine therapy [9], and chemotherapy [10–12] are currently utilized based on the patient's condition and cancer stage [2]. Chemotherapy [13] and drug repurposing, utilizing FDA-approved drugs such as raloxifene [14] clopidogrel [15], alkylating agents [16, 17], and CDK 4/6 inhibitors [18], demonstrate potential in breast cancer treatment, expediting drug development by leveraging existing regulatory approvals and targeting specific molecular pathways for enhanced efficacy.

Aberrant signaling pathways, particularly Receptor Tyrosine Kinases (RTKs), crucial for cellular functions such as proliferation, differentiation, and survival, undergo ligand-induced dimerization, activating

* Corresponding author.
E-mail address: shyamperugu@nitw.ac.in (S. Perugu).

https://doi.org/10.1016/j.compbiomed.2024.108433
Received 1 February 2024; Received in revised form 4 April 2024; Accepted 7 April 2024
Available online 16 April 2024

downstream signaling cascades regulating gene expression and cell cycle progression [19]. Dysregulated RTK signaling in cancer, with overexpression of EGFR, VEGFR, PDGFR, FGFR, and IGFR, drives uncontrolled proliferation and metastasis by creating a pre-metastatic niche in the tumor microenvironment [20,21].

RTK signaling drives vasculogenic mimicry, fostering cancer cells' endothelial-like transformation and intra-tumoral blood vessel formation [22,23]. Cross-talk between Wnt and Growth Factor Receptor signaling induces therapeutic resistance and enhances tumor invasion by disrupting cell-to-cell adhesion [24]. Investigating this aberrant signaling mechanism links breast cancer's molecular intricacies to potential drug design avenues.

37 FDA-approved drugs target RTK signaling, yet cancer cell complexities, involving exosomes and miRNAs, lead to resistance [25]. Upregulated miR-505, miR-128, and miR-145 induce doxorubicin resistance, while dysregulated miR-345 and miR-7 cause cisplatin resistance. MiR34a, miR-100, and miR-30c implicate paclitaxel and multidrug resistance [26–28]. Exosomes containing HER2, like SK-BR-3 and BT-474, hinder trastuzumab [29], while PI3K/AKT/mTOR pathway activation links to PARP inhibitor resistance [25,30–32]. Multi-drug resistance-associated proteins and miRNA miR-128 are implicated in doxorubicin resistance [33]. These findings emphasize the urgent need for new, effective breast cancer therapeutics, replacing ineffective drugs.

Employing a multi-modal deep learning approach, a robust and efficient paradigm, pivotal for our exploratory drug discovery methodology, diverges from traditional methods by incorporating data from various modalities and repositories, thereby enhancing predictive capabilities [34]. In line with this, a multi-modal deep learning-based QSAR model was developed to integrate information from 52 breast cancer cell lines and their molecular profiling data. Please Refer to Supplementary Table 8 and Supplementary Fig. 4. This comprehensive dataset encompassed gene expression, proteomic, copy number, fusion protein, CRISPR knock-out, mutational, and drug response datasets. Please Refer to Supplementary Table 7. This data-driven approach correlates the structural features of drugs with the biological response and establishes a quantitative relationship [35–37]. Using multiple molecular profiling datasets in this way improves the model's predictive ability to generalize across a wide range of drugs.

To our knowledge, prior efforts in drug repurposing for breast cancer have predominantly been classification-based rather than pathway-specific, often limited to a few cell lines, among many. We propose a novel deep learning-based QSAR model utilizing 52 breast cancer cell lines' molecular profiling data, focusing on signaling pathway specificity. Our MM-DNN QSAR model integrates diverse molecular datasets, optimizes hyperparameters using Hyperopt, and screens 37 FDA-approved molecules for structural similarities from PubChem (refer to Fig. 1). Rigorous predictions expedite the discovery of effective interventions, particularly for cancers with dysregulated RTK signaling, overcoming observed treatment resistance. Employing a multimodal deep learning approach, our work demonstrates scientific rigor and novelty in exploratory drug discovery research.

2. Methods

2.1. Data collection

The study sourced data from GDSC for drug-related information and Cell Model Passports for molecular profiling data. Cell Model Passports

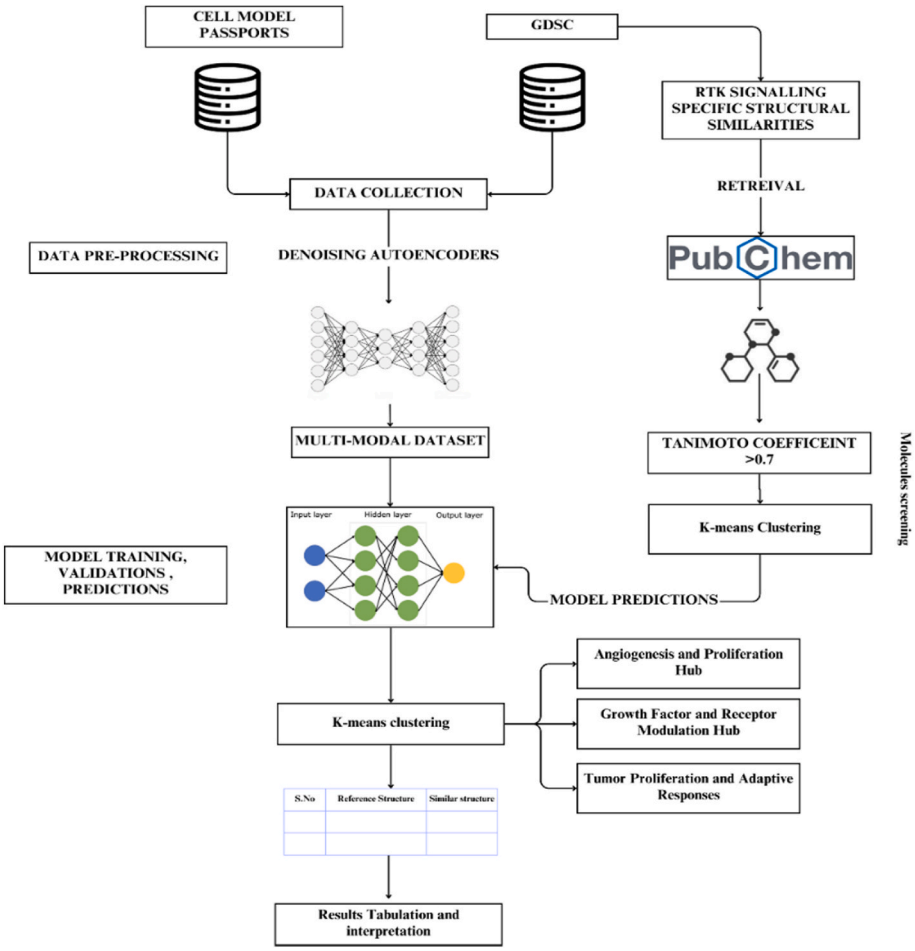

Fig. 1. Schematic diagram representing the pipeline of the research work.

provided diverse data types with an emphasis on recent updates. Details on dose-response data from GDSC and molecular profiling data from cell modal passport were available in Supplementary Tables 1 and 7 The dissemination of molecular profiling datasets is depicted in Supplementary Fig. 3.

The quantitative representation of molecules' 2D and 3D properties involved calculating Molecular Descriptors through the PADELpy Library in Python v3.12.0. Preliminary preprocessing included converting structures to 3D format, eliminating salts, and standardizing tautomers before descriptor calculation. This process yielded a comprehensive collection of 1444 2D and 421 3D descriptors, covering physiochemical, topological, electronic, geometric, and QSAR descriptors for each molecule examined.

2.2. Data pre-processing

The retrieved datasets underwent a preprocessing pipeline which included filtering based on unique identity (Please refer to Supplementary Table 8), handling of missing values using mean imputation through Simple Imputer, and encoding using MultiColumnLabelEncoder. Standardization of the datasets was achieved using the StandardScaler module in scikit-learn (Python v3.12.0), preparing them for Regression-based Deep Learning predictions of dose responses. Please refer to Fig. 1 for the workflow.

2.3. MM-DNN model training and validation

Utilizing autoencoders, a subset of Artificial Neural Networks designed for unsupervised learning, we compress input data into a lower-dimensional latent space through encoding and decoding

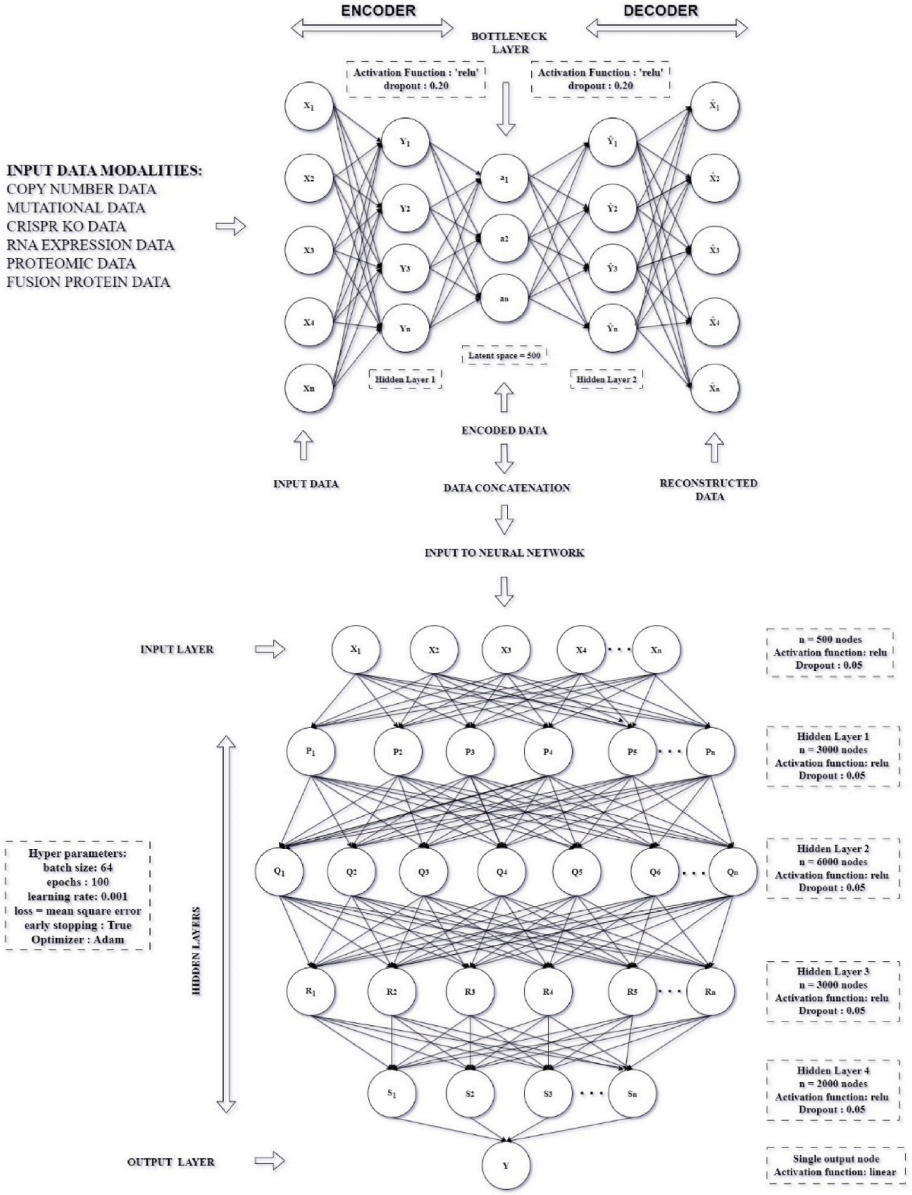

Fig. 2. MM-DNN Model Architecture: The schematic diagram illustrates the architecture of the MM-DNN model. It comprises an Autoencoder followed by a neural network. Each data modality undergoes individual processing through the Autoencoder, with the number of hidden layers adjusted based on dataset complexity. Dimensionality reduction reduces each dataset to 500 features. This procedure is iterated across all modalities, and the resulting reduced-dimensional datasets are concatenated for input into the neural network. Hyperparameters are fine-tuned using the Hyperopt optimization tool, with all optimized parameters specified in the diagram.

processes. This capability enables tasks such as dimensionality reduction, data transformation, and noise reduction. Consequently, we can enhance data preprocessing across diverse domains, including the concatenation of transformed datasets into a unified multimodal molecular profiling dataset [34]. Please Refer to Fig. 2.

t-SNE with perplexity 18, learning rate 185, and 500 iterations selected features from the drug response dataset. Denoising autoencoder processing then refined the molecular descriptors, adding noise and concatenating with molecular profiling data for enhanced predictions. Utilizing a Deep Neural Network architecture, we applied Hyperopt with Bayesian optimization for hyperparameter tuning to enhance our QSAR model's performance. Please refer to Table 1 and Fig. 2.

Systematically developed deep neural network-based QSAR models started with drug response data and progressively increased the data complexity to mimic biological system complexity. Subsequent iterations expanded by incorporating additional modalities. As data complexity heightened, a discernible trend emerged with diminishing validation scores, indicative of enhanced generalizability in the MM-DNN model. Please refer to Supplementary Tables 2 and 3 Fig. 3 illustrate the training and validation curves of the MM-DNN developed with Adam Optimizer. Validation metrics such as root mean square error (RMSE), mean square error (MSE), mean absolute error (MAE), Coefficient of Determination (R2), Coefficient of Determination for cross-validation (Q2), and 10-fold cross-validation were employed to assess the model's predictive ability and generalizability. The validation metrics outcomes have been Represented in Fig. 5. 10-Fold cross-validation metrics were represented in Fig. 6 and Supplementary Table 4.

Table 1
Optimized Hyperparameters Employed for MM-DNN Model Development. This table presents the hyperparameters optimized using Hyperopt for the development of the MM-DNN (Multi-Modal Deep Neural Network) model. *t*-SNE was employed for feature selection based on drug molecular descriptors, while autoencoder reduced dimensionality and facilitated data concatenation for model training, thereby simplifying data complexity. Deep neural network architecture was utilized for training the model on preprocessed datasets.

S. No	Algorithm	Purpose	Optimized Hyper-Parameters
1	t-Distributed Stochastic Neighbor Embedding	Feature Selection	perplexity 18, learning rate 185, and 500 iterations
2	Deep Neural Network	Model Training	Optimizer: Adam, Activation Function: 'ReLu', Drop out layer = 0.05, Loss Function: MSE (Mean Square Error), Epochs = 150, Batch size of 64, Learning Rate = 0.001, Early Stopping with Patience of 10, Validation Split: 0.25, verbose = 1, Restore best weights = True, Input layer with 500 nodes, four hidden layers with 3000, 6000, 3000, 2000 nodes and single output node with 'Linear' Activation function.
3	Standard Autoencoders	Dimensionality reduction	Optimizer: RMS prop, Activation Function: 'ReLu', Gaussian Noise with Standard deviation of 0.01, Loss Function: MSE (Mean Square Error), Epochs = 100, Batch size of 25, Learning Rate = 0.01, Early Stopping with Patience of 10, Validation Split: 0.3, Bottle Neck layer: 500 features, Input layers varied based on the dataset.

2.4. Structural similarities retrieval from PubChem

Notably, 37 drugs were found to be specific for various target proteins involved in RTK signaling, underscoring the substantial role of RTK Signaling in cancer survival. Please refer to Supplementary Fig. 2. The structural similarities of each drug targeting RTK Signaling were retrieved from the PubChem repository.

Canonical SMILES (Simplified Molecular Input Line Entry System) IDs for all structural similarities were extracted using the RDKit Library in Python v3.12.0. The Tanimoto coefficient served as the primary metric for comparing the reference molecule with its structural similarities, with a fixed threshold of 0.7. Pairs with a Tanimoto coefficient exceeding 0.7 were considered, while those below this threshold were disregarded. The filtered molecules for each structural similarity underwent K-means clustering, resulting in Similar and Dissimilar clusters. The reference molecule cluster was designated as the Similar Molecules cluster, while the other was considered Dissimilar. Molecules falling under a similar Molecule cluster were subsequently chosen for further investigation. A summary of the molecule filtering process is provided in Supplementary Table 5. Molecular descriptors for the final filtered molecules were calculated using the PADELpy library in Python v3.12.0.

2.5. Drug response predictions of screened structural similarities

Drug responses were predicted using the MM-DNN-based QSAR model, and the molecules were sorted based on the predicted IC50 (half maximal inhibitory concentration) values. The structural similarity with the top drug response is represented in Table 2. The 37 reference molecules were clustered into three groups using the K-means method and labelled as the Angiogenesis and Proliferation Hub, Growth Factor and Receptor Modulation Hub, and Tumor Proliferation and Adaptive Responses Hub based on their downstream regulatory proteins. Please refer to Supplementary Table 6. Feature importance analysis was performed for each cluster using an Extra gradient boosting algorithm and the results are presented in Supplementary Fig. 8.

3. Results

3.1. Model training and validation

As mentioned in the methodology section, molecular profiling and dose-response datasets have undergone pre-processing. Autoencoders were used to reduce the complexity to a latent dimension of 500 features. Supplementary Fig. 6 displays the training and loss curves of the MM-DNN model employing the SGD (Stochastic Gradient Descent) optimizer, yielding suboptimal validation metrics and loss. Consequently, we opted for the Adam optimizer for constructing the MM-DNN model. Hyperopt, a Bayesian optimization-based hyperparameter tuning technique, was employed to optimize the model parameters, as depicted in Supplementary Fig. 5 and detailed in the accompanying optimization curves and ranges. Subsequently, the model's fit to the preprocessed multimodal dataset was assessed through regression plots, outlined in Fig. 4, ensuring accuracy and absence of overfitting.

The trained model was validated using a test set and validation set. MM-DNN-based QSAR Model has shown an impressive Coefficient of determination (R2) value of 0.917, Root Mean Square Error value of 0.312, Mean Square Error value of 0.183, and Mean Absolute Error value of 0.289 for the test set. Please refer to Supplementary Table 2 and Fig. 5A. The validation set resulted in a promising R2 Value of 0.920, RMSE of 0.335, MSE of 0.197, MAE of 0.216, and Q2 (coefficient of determination of validation set) of 0.920. Please refer to Supplementary Table 3 and Fig. 5B. The test set, validation set, 10-fold validation (Fig. 6 and Supplementary Table 4), and training loss curve outcomes (Fig. 3) collectively underscore the efficacy and reliability of the developed MM-DNN QSAR Model.

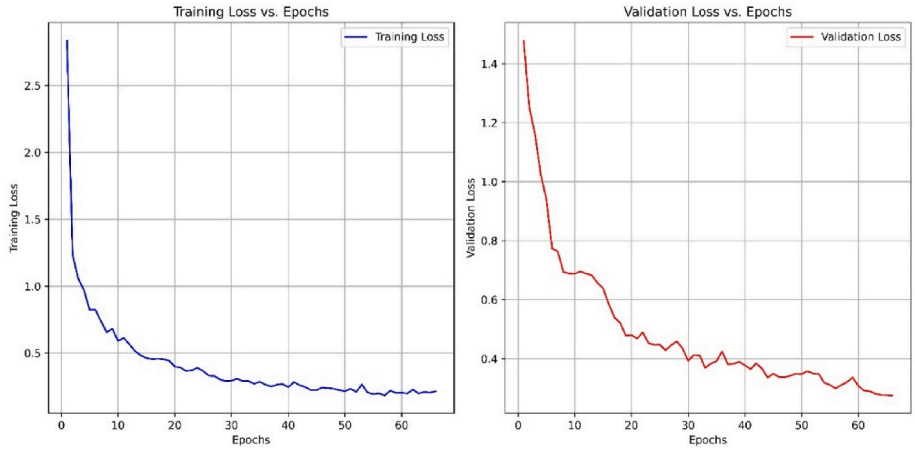

Fig. 3. Training and Validation Loss Curves: Illustrating the Reduction in Loss Over Epochs. This figure depicts the training and validation loss curves of the MM-DNN model developed using optimized hyperparameters. The curves demonstrate the progressive decrease in loss with increasing epochs. The inclusion of early stopping functionality allowed the model to terminate training prematurely, mitigating the risk of overfitting and loss escalation.

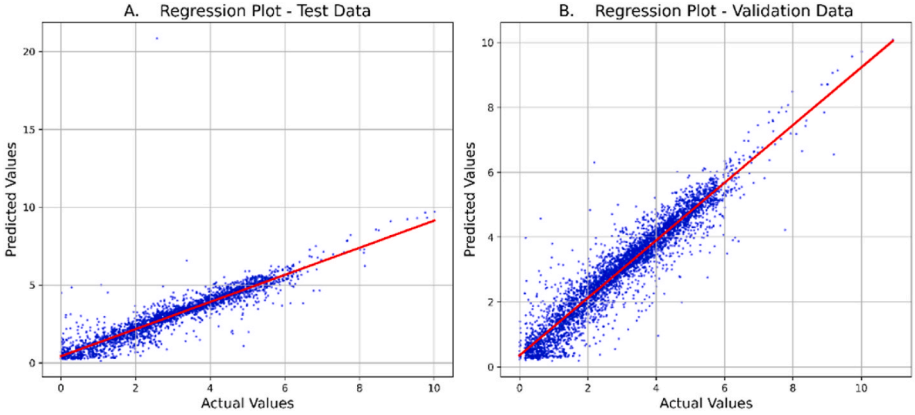

Fig. 4. Regression plots displaying the predictive performance assessment of the model on both the test and validation sets. These plots provide a visual evaluation of how well the model's predictions align with the actual values, aiding in the assessment of its predictive accuracy and generalization capabilities.

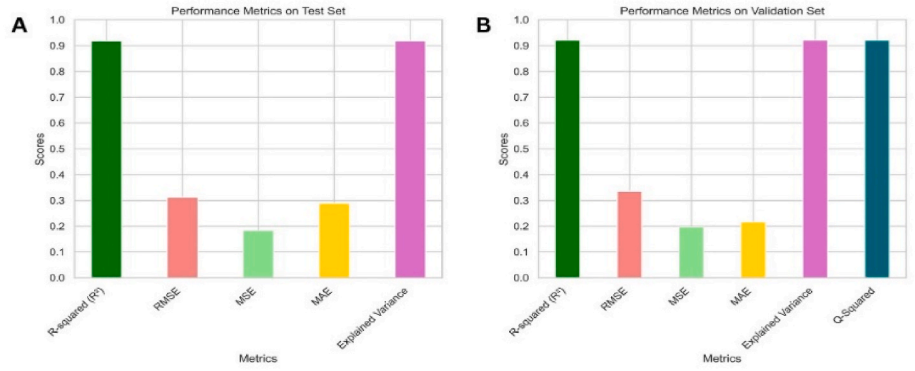

Fig. 5. Bar plots representing Performance validation metrics of MM-DNN Model. A: Test Set, B: Validation Set.

3.2. Structural similarities screening

Exploratory data analysis of drug response data revealed that the RTK signaling pathway is the most targeted, with 37 drugs focusing on different downstream regulators of RTK Signalling. The results of the analysis are presented in Supplementary Figs. 1, 2, 3, and 4 and Supplementary Tables 7 and 8 Molecules were screened based on similarity, and the molecule reduction in the count is represented in Supplementary Table 5. Table 2 presents the reference drug and its structural similarity to the most effective drug response.

The Angiogenesis and Proliferation Hub cluster predominantly involves crucial receptors for angiogenesis and cell proliferation pathways, such as FGFRs, VEGFRs, MET, PDGFRs, and EGFRs, driving blood vessel formation and cellular growth, indicative of a strong association with tumor angiogenesis and proliferation. In contrast, the Growth Factor and Receptor Modulation Hub cluster focuses on growth factor signaling, highlighting targets like NTRKs, MET, VEGFRs, and PDGFRs, crucial for regulating cell proliferation, survival, and differentiation in

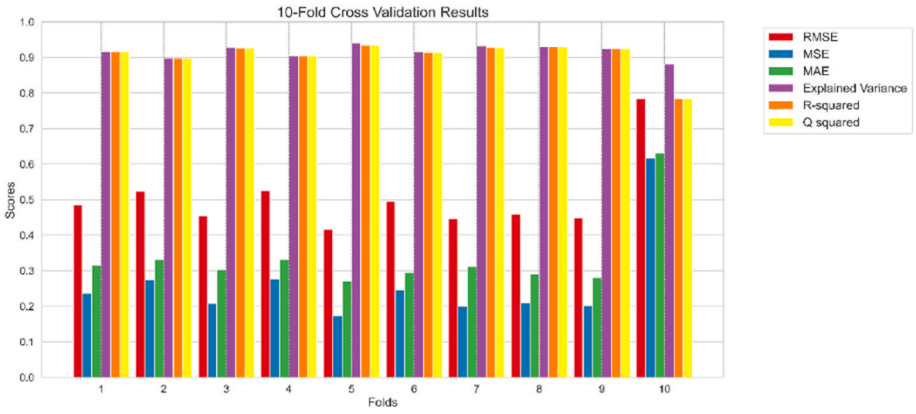

Fig. 6. Bar plot representing the 10-fold cross-validation results of MM-DNN-based QSAR Model.

cancer development. Lastly, the Tumor Proliferation and Adaptive Responses cluster comprises targets frequently dysregulated in cancer cells, including EGFRs, ERBBs, KIT, PDGFRs, and ALK, driving uncontrolled proliferation, survival, and therapy resistance.

The primary objective of conducting these clustering analyses was to gain a comprehensive understanding of the specific pharmacophoric features associated with particular drugs grouped within each cluster. This endeavour aimed to streamline subsequent stages of drug development by focusing solely on the most critical pharmacophoric features identified. Please refer to Supplementary Fig. 8. This approach opens up possibilities for the development of drugs capable of targeting a wide range of receptors within each hub.

Supplementary Fig. 8 displays the top 15 molecular descriptors identified as significant for the drugs in each cluster, determined through extra gradient boosting-based feature selection. Together, these clusters offer valuable insights into the diverse molecular mechanisms underlying tumor progression and provide potential therapeutic targets for further investigation. Top hits and their structural references and IUPAC (International Union of Pure and Applied Chemistry) nomenclature were represented in Supplementary Table 10, emphasizing key functional groups and unique elements through colour-coded distinctions, facilitating easy comparison.

4. Discussion

We developed a robust multimodal deep-learning QSAR model by integrating GDSC drug response and cancer cell passport molecular profiling data, achieving exceptional performance with an R2 of 0.917 and an RMSE of 0.312 on the test set, surpassing previous drug repurposing endeavours. Our primary objective centred on identifying novel drug candidates, particularly within the RTK signaling pathway, departing from conventional methodologies [38]. Adhering to QSAR principles [35,37], we meticulously scrutinized unannotated structural analogs of RTK signaling-specific drugs to uncover promising alternatives for drug-resistant cancer cell lines [29]. This involved systematic screening of PubChem for unannotated chemical molecules exhibiting structural similarities, leveraging our meticulously trained MM-DNN-based QSAR model for predictive analyses.

The application of deep learning in biomedical research offers groundbreaking solutions to various challenges. By integrating protein encoding strategies with deep learning algorithms, researchers have significantly improved the accuracy and control of false discovery rates in protein function annotation, surpassing traditional methods [39]. Employing convolutional neural networks (CNN) alongside diverse protein encoding strategies, deep learning enables the effective annotation of type IV bacterial secretion system effectors (T4SEs), thus aiding in the suppression of infections [40]. Furthermore, deep learning facilitates the development of task-specific encoding algorithms for RNA interactions, enhancing our ability to identify RNA interactions in physiological and pathological processes [41]. Additionally, the utilization of deep learning in profiling protein-protein interaction (PPI) sites through frameworks like EnsemPPIS enables accurate proteome-wide profiling, showcasing the broad applicability of deep learning in this domain [42]. In the realm of protein function annotation, deep learning-driven strategies like AnnoPRO effectively address the long-tail problem, offering superior performance and advancing biological sciences research [43]. These findings underscore the transformative potential of deep learning in revolutionizing biomedical research practices.

Researchers have extensively applied multi-modal deep learning approaches for various cancer-related predictions, such as survival in non-small cell lung carcinoma [44], classification-based predictions in breast cancer therapeutics [45], and pathological complete responses to neoadjuvant chemotherapy [46] Additionally, machine learning and deep learning algorithms have been leveraged for drug repurposing studies, with models integrating data from differentially expressed genes and Drug Bank [38], kernel-based pharmacogenomics datasets, and TCGA data [47], as well as gene expression data to predict the effectiveness of FDA-approved drugs in treating breast cancer [48]. The cited studies offer a robust foundation for our research, showcasing the efficacy of multi-modal deep learning in cancer prediction and the utilization of machine learning for exploratory drug discovery.

We prioritized the utilization of a multimodal deep neural network due to its efficiency in integrating diverse molecular profiling datasets, thereby enhancing our comprehension of intricate relationships among these datasets and drug response data, consequently bolstering the model's predictive capacity. Traditional machine learning algorithms, such as Extragradient boosting, support vector machines with radial basis function kernel, and random forest, struggled to achieve satisfactory validation metrics and accurately predict test data due to the data's complexity and high dimensionality across multiple profiling datasets. Please Refer to Supplementary Fig. 7. Our experimentation revealed that these methods were insufficient for handling the complexity and heterogeneity of the data. Thus, we turned to deep learning as a more effective approach to address these challenges.

Our approach, targeting specific pathways for novel drug candidate studies, shows robustness and significant potential in advancing early-stage drug development. However, there are notable limitations. To enhance predictive robustness, incorporating 3D molecular descriptors like COMFA and COMSIA can be instrumental, in capturing steric and electrostatic interactions in 3D lattice [49–52]. Additionally, leveraging image-based modalities through convolutional neural networks, particularly with tissue section images, can offer valuable insights into complex interactions [53]. Target protein-based pharmacophore modeling provides a molecular framework for precision targeting within protein clusters. Integrating data on protein molecular features using

Table 2
Tabular column representing the top Structural similarities of each Drug molecule.

S. No	Reference Drug Name	Structural Similarity	Predicted IC50 value (μmol)
1	Alectinib	3-amino-6,6-dimethyl-9-[4-(oxetan-3-yl) piperazin-1-yl]-5H-benzo[b]carbazol-11-one	0.758705
2	Amuvatinib	3-(1,3-Benzodioxol-5-yl)-1-[4-([1] benzofuro[3,2-*d*]pyrimidin-4-yl) piperazin-1-yl]propane-1-thione; methanesulfonic acid	1.804693
3	AST-1306	[1-[[4-[3-Chloro-4-[(3-fluorophenyl) methoxy]anilino]quinazolin-6-yl] amino]-2-(2-fluorobutoxycarbonylamino)-1-oxopropan-2-yl] acetate	0.509847
4	Axitinib	*N*-[5-(hydroxyamino)-5-oxopentyl]-2-[[3-(2-pyridin-2-ylethenyl)-1H-indazol-6-yl]sulfanyl]benzamide	1.010532
5	AZD1332	5-Chloro-N2-[(1 S)-1-(5-fluoropyridin-2-yl)ethyl]-N4-(5-isopropoxy-1H-pyrazol-3-yl)pyrimidine-2,4-diamine phosphate	1.946533
6	AZD4547	1-chloro-N,N,2-trimethylprop-1-en-1-amine; N-[5-[2-(3,5-dimethoxyphenyl) ethyl]-1H-pyrazol-3-yl]-3-methoxybenzamide	1.443234
7	AZD6094	3-(Imidazo[1,2-*b*]pyridazin-6-ylmethyl)-5-(1-methylpyrazol-4-yl)triazolo[4,5-*b*] pyrazine	2.287226
8	AZD8931	6-[4-(3-chloroanilino)-7-methoxyquinazolin-6-yl]oxy-*N*-hydroxyhexanamide	0.66
9	BIBF-1120	methyl 3-[*N*-[4-[(3-aminopropylamino) methyl]phenyl]-*C*-phenylcarbonimidoyl]-2-hydroxy-1H-indole-6-carboxylate	1.385818
10	BMS-754807	*N*-[1-[4-[(5-cyclopropyl-1H-pyrazol-3-yl)amino]pyrrolo[2,1-*f*][1,2,4]triazin-2-yl]piperidin-4-yl]acetamide	1.224906
11	Cabozantinib	4-(6,7-dimethoxyquinolin-4-yl)oxy-*N*-[3-(4-fluorophenyl)propyl]aniline	1.231128
12	CI-1033	(Z)-*N*-[4-(3-chloro-4-fluoroanilino)-7-(difluoromethoxy)quinazolin-6-yl]-4-(dimethylamino)but-2-enamide	1.073408
13	Crizotinib	3-Phenylmethoxy-5-(1-piperidin-4-ylpyrazol-4-yl)pyridin-2-amine	1.647109
14	Dasatinib	*N*-(2-chloro-6-methylphenyl)-2-[[6-[4-(2-hydroxyethyl)piperazin-1-yl]-2-methylpyrimidin-4-yl]amino]-1,3-thiazole-5-carboxamide; sulfuric acid	1.062769
15	Foretinib	*N*-[4-[7-(3-bromopropoxy)-6-methoxyquinolin-4-yl]oxy-3-fluorophenyl]-1-formylcyclopropane-1-carboxamide	0.802407
16	GSK 1904529 A	*N*-(2,6-difluorophenyl)-3-{3-[2-({5-methyl-2-(methyloxy)-4-[4-(methylsulfonyl)-1-piperazinyl]phenyl} amino)-4 pyrimidinyl]imidazo[1,2-*a*] pyridin-2-yl}benzamide	1.062241
17	GW2580	2,4-Diamino-5-[3,4-dimethoxy-5-(2-propynyloxy)benzyl]pyrimidine	1.320202
18	GW441756	(3Z)-3-ethylidene-1H-pyrrolo[3,2-*b*] pyridin-2-one; 1-methylindole	1.021878
19	JNJ38877605	6-[Difluoro-(6-iodo-[1,2,4]triazolo[4,3-*b*]pyridazin-3-yl)methyl]quinoline	0.908253
20	Kobe 2602	1-(4-Ethylphenyl)-3-[2-nitro-4-(trifluoromethyl)anilino]thiourea	1.096508
21	lapatinib	Methyl 6-[[5-[4-[3-chloro-4-[(3-fluorophenyl)methoxy]anilino] quinazolin-6-yl]furan-2-yl] methylamino]hexanoate	0.739111
22	Linifanib	1-[4-(3-amino-1H-indazol-4-yl)phenyl]-3-(2-fluoro-5-methylphenyl)-1-(2-hydroxyethyl)urea	1.442908
23	Masitinib	4-(2-aminophenyl)-1-*N*-[4-methyl-3-[(4-pyridin-3-yl-1,3-thiazol-2-yl)amino] phenyl]cyclohexa-1,5-diene-1,4-dicarboxamide	1.882307
24	Motesanib	*N*-(4-pentylphenyl)-2-(pyridin-4-ylmethylamino)pyridine-3-carboxamide	0.988261
25	NVP-TAE684	5-chloro-2-*N*-[2-methoxy-4-(4-piperazin-4-ium-1-ylpiperidin-1-yl) phenyl]-4-*N*-(2-propan-2-ylsulfonylphenyl)pyrimidine-2,4-diamine; 2,2,2-trifluoroacetate	0.907894
26	OSI-930	3-(pyridin-4-ylmethylamino)-*N*-[4-(trifluoromethoxy)phenyl]thiophene-2-carboxamide	1.790038
27	Pazopanib	3-[4-[[4-[(2,3-dimethylindazol-6-yl)-methylamino]pyrimidin-2-yl]amino] phenyl]-*N*-methylpropane-1-sulfinamide	1.30786
28	PD173074	*tert*-butyl *N*-[*N*'-[2-[amino-[(Z)-2-amino-3-[[7-(tert-butylcarbamoylamino)-6-(3,5-dimethoxyphenyl)pyrido[2,3-*d*] pyrimidin-2-yl]amino]prop-1-enyl] amino]ethyl]-*N*-[(2-methylpropan-2-yl) oxycarbonyl]carbamimidoyl]carbamate	1.104513
29	PF-00299804	*N*-[4-(3-chloro-4-fluoroanilino)-7-(difluoromethoxy)quinazolin-6-yl] acetamide	1.065045
30	PHA-665752	5-[5-(2,6-Dichloro-phenylmethanesulfonyl)-2-oxo-1,2-dihydro-indol-(3Z)-ylidenemethyl]-2,4-dimethyl-1H-pyrrole-3-carboxylic acid amide	1.004722
31	Quizartinib	1-[4-(6-Methoxyimidazo[2,1-b][1,3] benzothiazol-2-yl)phenyl]-3-(3-methyl-1,2-oxazol-5-yl)urea	1.197091
32	SB505124	4-(1,3-benzodioxol-5-yl)-5-(6-methylpyridin-2-yl)-1H-imidazole-2-amine	1.402854
33	SB52334	Ethane; 2-methylpyridine-3-carboxylic acid	1.944159
34	Sorafenib	4-[4-[[[[4-Chloro-3-(trifluoromethyl) phenyl]amino]carbonyl]amino]-3-fluorophenoxy]-*N*-methyl-2-pyridinecarboxamide hydrate	2.10531
35	SU11274	*tert*-butyl 4-[5-[(Z)-[5-[(3-chlorophenyl)-methylsulfamoyl]-2-oxo-1H-indol-3-ylidene]methyl]-2,4-dimethyl-1H-pyrrole-3-carbonyl] piperazine-1-carboxylate	1.006986
36	Sunitinib	*N*-[2-(diethylamino)ethyl]-5-[1-(5-fluoro-2-oxo-1,3-dihydroindol-3-yl) propyl]-2,4-dimethyl-1H-pyrrole-3-carboxamide	1.526341
37	Tivozanib	1-[4-(6,7-Dimethoxyquinolin-4-yl) oxyphenyl]-3-(5-methyl-1,2-oxazol-3-yl) urea	1.260934

Natural Language Processing and one-dimensional convolutional neural networks can illuminate intricate ligand-target interactions [54]. The non-uniformity in dataset contents limits the model's potential, which could be addressed by increasing database entries and utilizing data from other patient-specific databases like TCGA, INTEDE, DRESIS, and ADCdb. Please refer to Supplementary Table 9. While our model achieved a performance metric of 0.917, there is room for improvement by leveraging advanced deep neural network models like Recurrent Neural Networks (RNNs) and Transformers for more effective predictions. While predictive models have significant scope in early-stage drug development, their translational impact on clinical practice underscores the importance of conducting in vitro and in vivo studies aligning with insilico findings.

In the field of drug discovery, the acquisition of additional cell lines and chemical compounds for extensive research carries significant

financial implications. The investigation of numerous chemical compounds to identify promising molecules is resource-intensive. Consequently, the importance of predictive models developed through Machine Learning (ML) and Deep Learning (DL) becomes evident in streamlining this process. These models efficiently filter out molecules with minimal or no desired biological response, offering cost-effective solutions. In 2024, databases like PubChem boast over 116 million compounds, and ChemEMBL contains 2.4 million compounds, providing a vast pool for potential discoveries. Our MM-DNN-based QSAR model emerges as a robust drug exploratory tool, uniquely capable of targeting specific signalling pathways. Furthermore, by integrating insights from databases such as INTEDE [55], DRESIS [56], DrugMAP [57], ADCdb [58], TheMarker [59], TTD [60] and VARIDT [61] (Please refer to Supplementary Table 9), our approach enables comprehensive exploration and discovery of potential inhibitors, thereby advancing drug discovery efforts and contributing to precision medicine initiatives. This capability allows us to leverage drugs initially intended for other purposes and unannotated drugs, presenting a strategic advantage in the drug development screening process. Machine Learning, Deep Learning, and Molecular Docking play vital roles in early drug development, reducing time and narrowing down molecules. Although further in vitro and in vivo studies are essential in later stages, our identified molecules undergo rigorous screening for their pharmacokinetic, dosage, and side effect profiles, advancing them for subsequent studies.

Data availability

We obtained the drug response data from the Genomics of Drug Sensitivity in Cancer (GDSC) database, accessible at https://www.cancerrxgene.org/. The molecular profiling data for the cancer cell lines was sourced from the Cell Model Passports database, available at https://cellmodelpassports.sanger.ac.uk/.

CRediT authorship contribution statement

Anush Karampuri: Writing – review & editing, Writing – original draft, Visualization, Validation, Methodology, Investigation, Data curation, Conceptualization. **Sunitha Kundur:** Writing – review & editing, Resources, Project administration, Investigation, Conceptualization. **Shyam Perugu:** Writing – review & editing, Software, Resources, Investigation, Formal analysis, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

We express our gratitude to the National Institute of Technology, Warangal for their valuable support in our research work.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.compbiomed.2024.108433.

References

[1] R.L. Siegel, K.D. Miller, N.S. Wagle, A. Jemal, Cancer statistics, 2023, CA A Cancer J. Clin. 73 (2023) 17–48.
[2] R. Hong, B. Xu, Breast cancer: an up-to-date review and future perspectives, Cancer Commun. 42 (2022) 913–936.
[3] K. Sathishkumar, J. Sankarapillai, A. Mathew, et al., Breast cancer survival in India across 11 geographic areas under the national cancer registry programme, Cancer (2024), https://doi.org/10.1002/cncr.35188.
[4] K.S. Johnson, E.F. Conant, M.S. Soo, Molecular subtypes of breast cancer: a review for breast radiologists, J Breast Imaging 3 (2021) 12–24.
[5] L. Vania, M. Hermawan Widyananda, V.D. Kharisma, A. Nur, M. Ansori, N. Maksimiuk, M. Derkho, A. Denisenko, N.I. Sumantri, A. Patera Nugraha, Anticancer activity prediction of garcinia mangostana l. Against her2-positive breast cancer through inhibiting egfr, her2 and igf1r protein : a bioinformatics study, Biochem. Cell. Arch. 21 (2021) 3313–3321.
[6] Q. Wu, S. Chen, W. Peng, D. Chen, Current perspectives on cell-assisted lipotransfer for breast cancer patients after radiotherapy, World J. Surg. Oncol. (2023), https://doi.org/10.1186/s12957-023-03010-z.
[7] B. Henriques, F. Mendes, D. Martins, Immunotherapy in breast cancer: when, how, and what challenges? Biomedicines (2021) https://doi.org/10.3390/biomedicines9111687.
[8] T. Shien, H. Iwata, Adjuvant and neoadjuvant therapy for breast cancer, Jpn. J. Clin. Oncol. 50 (2020) 225–229.
[9] M.M. Haque, K.V. Desai, Pathways to endocrine therapy resistance in breast cancer, Front. Endocrinol. (2019), https://doi.org/10.3389/fendo.2019.00573.
[10] P. Di Nardo, C. Lisanti, M. Garutti, S. Buriolla, M. Alberti, R. Mazzeo, F. Puglisi, Chemotherapy in patients with early breast cancer: clinical overview and management of long-term side effects, Expet Opin. Drug Saf. 21 (2022) 1341–1355.
[11] M. Dahan, M. Cortet, C. Lafon, F. Padilla, Combination of focused ultrasound, immunotherapy, and chemotherapy: new perspectives in breast cancer therapy, J. Ultrasound Med. 42 (2023) 559–573.
[12] A.F. Dibha, S. Wahyuningsih, A.N.M. Ansori, et al., Utilization of secondary metabolites in algae kappaphycus alvarezii as a breast cancer drug with a computational method, Phcog. J. 14 (2022) 536–543.
[13] S.M.A. Hamami, M. Fai, A.F. Aththar, et al., Nano transdermal delivery potential of fucoidan from sargassum sp. (Brown algae) as chemoprevention agent for breast cancer treatment, Phcog. J. 14 (2022) 789–795.
[14] A.S. Correia, F. Gärtner, N. Vale, Drug combination and repurposing for cancer therapy: the example of breast cancer, Heliyon (2021) e06120, https://doi.org/10.1016/j.heliyon.2021.e05948.
[15] A. Denslow, M. Świtalska, J. Jarosz, D. Papiernik, K. Porshneva, M. Nowak, J. Wietrzyk, Clopidogrel in a combined therapy with anticancer drugs—effect on tumor growth, metastasis, and treatment toxicity: studies in animal models, PLoS One (2017), https://doi.org/10.1371/journal.pone.0188740.
[16] K. Nakatsukasa, H. Koyama, Y. Oouchi, et al., Docetaxel and cyclophosphamide as neoadjuvant chemotherapy in HER2-negative primary breast cancer, Breast Cancer 24 (2017) 63–68.
[17] J.C. Singh, A. Mamtani, A. Barrio, et al., Pathologic complete response with neoadjuvant doxorubicin and cyclophosphamide followed by paclitaxel with trastuzumab and pertuzumab in patients with HER2-positive early stage breast cancer: a single center experience, Oncol. 22 (2017) 139–143.
[18] N.C. Turner, D.J. Slamon, J. Ro, et al., Overall survival with palbociclib and fulvestrant in advanced breast cancer, N. Engl. J. Med. 379 (2018) 1926–1936.
[19] E. Cordover, A. Minden, Signaling pathways downstream to receptor tyrosine kinases: targets for cancer treatment, J Cancer Metastasis Treat (2020), https://doi.org/10.20517/2394-4722.2020.101.
[20] C. Francavilla, C.S. Obrien, Fibroblast growth factor receptor signalling dysregulation and targeting in breast cancer, Open Biol (2022), https://doi.org/10.1098/rsob.210373.
[21] N. Bou Antoun, A.M. Chioni, Dysregulated signalling pathways driving anticancer drug resistance, Int. J. Mol. Sci. (2023), https://doi.org/10.3390/ijms241512222.
[22] Y. Hao, D. Baker, P Ten Dijke, TGF-β-mediated epithelial-mesenchymal transition and cancer metastasis, Int. J. Mol. Sci. (2019), https://doi.org/10.3390/ijms20112767.
[23] C.C. Lin, K.M. Suen, J. Lidster, J.E. Ladbury, The emerging role of receptor tyrosine kinase phase separation in cancer, Trends Cell Biol. (2023), https://doi.org/10.1016/j.tcb.2023.09.002.
[24] P. Merikhian, M.R. Eisavand, L. Farahmand, Triple-negative breast cancer: understanding Wnt signaling in drug resistance, Cancer Cell Int. (2021), https://doi.org/10.1186/s12935-021-02107-3.
[25] W.Z. Hu, C.L. Tan, Y.J. He, G.Q. Zhang, Y. Xu, J.H. Tang, Functional miRNAs in breast cancer drug resistance, OncoTargets Ther. 11 (2018) 1529–1541.
[26] N. Dastmalchi, R. Safaralizadeh, B. Baradaran, M. Hosseinpourfeizi, A. Baghbanzadeh, An update review of deregulated tumor suppressive microRNAs and their contribution in various molecular subtypes of breast cancer, Gene (2020), https://doi.org/10.1016/j.gene.2019.144301.
[27] M.H. Soheilifar, N. Masoudi-Khoram, S. Madadi, S. Nobari, H. Maadi, H. Keshmiri Neghab, R. Amini, M. Pishnamazi, Angioregulatory microRNAs in breast cancer: molecular mechanistic basis and implications for therapeutic strategies, J. Adv. Res. 37 (2022) 235–253.
[28] S. Afzal, M. Hassan, S. Ullah, H. Abbas, F. Tawakkal, M.A. Khan, Breast cancer; discovery of novel diagnostic biomarkers, drug resistance, and therapeutic implications, Front. Mol. Biosci. (2022), https://doi.org/10.3389/fmolb.2022.783450.
[29] X. Dong, X. Bai, J. Ni, H. Zhang, W. Duan, P. Graham, Y. Li, Exosomes and breast cancer drug resistance, Cell Death Dis. (2020), https://doi.org/10.1038/s41419-020-03189-z.
[30] Mei Y, Liao X, Zhu L, Yang H Overexpression of RSK4 reverses doxorubicin resistance in human breast cancer cells via PI3K/Akt signaling pathway. https://doi.org/10.1093/jb/mvaa009/5711292.
[31] C. Dong, J. Wu, Y. Chen, J. Nie, C. Chen, Activation of PI3K/AKT/mTOR pathway causes drug resistance in breast cancer, Front. Pharmacol. (2021), https://doi.org/10.3389/fphar.2021.628690.

[32] Y. Yoshioka, K. Minoura, R.u. Takahashi, F. Takeshita, T. Taya, R. Horii, Y. Fukuoka, T. Kato, N. Kosaka, T. Ochiya, An integrative genomic analysis revealed the relevance of microRNA and gene expression for drug-resistance in human breast cancer cells, Mol. Cancer (2011), https://doi.org/10.1186/1476-4598-10-135.
[33] K. Jamialahmadi, F. Zahedipour, G. Karimi, The role of microRNAs on doxorubicin drug resistance in breast cancer, J. Pharm. Pharmacol. 73 (2021) 997–1006.
[34] G. Shtar, Multimodal machine learning for drug knowledge discovery, in: WSDM 2021 - Proceedings of the 14th ACM International Conference on Web Search and Data Mining, Association for Computing Machinery, Inc, 2021, pp. 1115–1116.
[35] P.S. Mahalakshmi, Y. Jahnavi, A review on QSAR studies, International Journal of Advances in Pharmacy and Biotechnology 6 (2020) 19–23.
[36] A. Bak, Two decades of 4d-qsar: a dying art or staging a comeback? Int. J. Mol. Sci. (2021) https://doi.org/10.3390/ijms22105212.
[37] S. Kausar, A.O. Falcao, An automated framework for QSAR model building, J. Cheminf. (2018), https://doi.org/10.1186/s13321-017-0256-5.
[38] C. Cui, X. Ding, D. Wang, L. Chen, F. Xiao, T. Xu, M. Zheng, X. Luo, H. Jiang, K. Chen, Drug repurposing against breast cancer by integrating drug-exposure expression profiles and drug-drug links based on graph neural network, Bioinformatics 37 (2021) 2930–2937.
[39] J. Hong, Y. Luo, Y. Zhang, J. Ying, W. Xue, T. Xie, L. Tao, F. Zhu, Protein functional annotation of simultaneously improved stability, accuracy and false discovery rate achieved by a sequence-based deep learning, Briefings Bioinf. 21 (2019) 1437–1447.
[40] J. Hong, Y. Luo, M. Mou, J. Fu, Y. Zhang, W. Xue, T. Xie, L. Tao, Y. Lou, F. Zhu, Convolutional neural network-based annotation of bacterial type IV secretion system effectors with enhanced accuracy and reduced false discovery, Briefings Bioinf. 21 (2020) 1825–1836.
[41] Y. Wang, Z. Pan, M. Mou, et al., A task-specific encoding algorithm for RNAs and RNA-associated interactions based on convolutional autoencoder, Nucleic Acids Res. (2023), https://doi.org/10.1093/nar/gkad929.
[42] M. Mou, Z. Pan, Z. Zhou, L. Zheng, H. Zhang, S. Shi, F. Li, X. Sun, F. Zhu, A transformer-based ensemble framework for the prediction of protein–protein interaction sites, Research (2023), https://doi.org/10.34133/research.0240.
[43] L. Zheng, S. Shi, M. Lu, et al., AnnoPRO: a strategy for protein function annotation based on multi-scale protein representation and a hybrid deep learning of dual-path encoding, Genome Biol. (2024), https://doi.org/10.1186/s13059-024-03166-1.
[44] J.G. Ellen, E. Jacob, N. Nikolaou, N. Markuzon, Autoencoder-based multimodal prediction of non-small cell lung cancer survival, Sci. Rep. (2023), https://doi.org/10.1038/s41598-023-42365-x.
[45] S. Kayikci, T.M. Khoshgoftaar, Breast cancer prediction using gated attentive multimodal deep learning, J Big Data (2023), https://doi.org/10.1186/s40537-023-00749-w.
[46] K.M. Boehm, P. Khosravi, R. Vanguri, J. Gao, S.P. Shah, Harnessing multimodal data integration to advance precision oncology, Nat. Rev. Cancer 22 (2022) 114–126.
[47] Y. Wang, Y. Yang, S. Chen, J. Wang, Deepdrk: a deep learning framework for drug repurposing through kernel-based multi-omics integration, Briefings Bioinf. (2021), https://doi.org/10.1093/bib/bbab048.
[48] N. Saberian, A. Peyvandipour, M. Donato, S. Ansari, S. Draghici, A new computational drug repurposing method using established disease-drug pair knowledge, Bioinformatics 35 (2019) 3672–3678.
[49] S. Alam, F. Khan, 3D-QSAR studies on Maslinic acid analogs for Anticancer activity against Breast Cancer cell line MCF-7, Sci. Rep. (2017), https://doi.org/10.1038/s41598-017-06131-0.
[50] A. Bak, Two decades of 4d-qsar: a dying art or staging a comeback? Int. J. Mol. Sci. (2021) https://doi.org/10.3390/ijms22105212.
[51] Mishra K, Jain SK, Pant R Rational Drug Design and Optimization of New Leads Using Modern Quantitative Structure-Activity Relationship (QSAR) Techniques.
[52] Sirvi S. Ram, A review on quantitative structure-activity and relationships (QSAR) methods, International Journal of Scientific Research and Management 10 (2022) 624–628.
[53] S. Chattopadhyay, A. Dey, P.K. Singh, D. Oliva, E. Cuevas, R. Sarkar, MTRRE-Net: a deep learning model for detection of breast cancer from histopathological images, Comput. Biol. Med. (2022), https://doi.org/10.1016/j.compbiomed.2022.106155.
[54] Y. Zhang, Y. Hu, N. Han, A. Yang, X. Liu, H. Cai, A survey of drug-target interaction and affinity prediction methods via graph neural networks, Comput. Biol. Med. (2023), https://doi.org/10.1016/j.compbiomed.2023.107136.
[55] J. Yin, F. Li, Y. Zhou, et al., INTEDE: interactome of drug-metabolizing enzymes, Nucleic Acids Res. 49 (2021) D1233–D1243.
[56] X. Sun, Y. Zhang, H. Li, et al., DRESIS: the first comprehensive landscape of drug resistance information, Nucleic Acids Res. 51 (2023) D1263–D1275.
[57] F. Li, J. Yin, M. Lu, et al., DrugMAP: molecular atlas and pharma-information of all drugs, Nucleic Acids Res. 51 (2023) D1288–D1299.
[58] L. Shen, X. Sun, Z. Chen, et al., ADCdb: the database of antibody–drug conjugates, Nucleic Acids Res. 52 (2024) D1097–D1109.
[59] Y. Zhang, Y. Zhou, Y. Zhou, et al., TheMarker: a comprehensive database of therapeutic biomarkers, Nucleic Acids Res. 52 (2024) D1450–D1464.
[60] Y. Zhou, Y. Zhang, D. Zhao, X. Yu, X. Shen, Y. Zhou, S. Wang, Y. Qiu, Y. Chen, F. Zhu, T TD: Ther apeutic Targ et D atabase describing tar get drugg ability inf ormation, Nucleic Acids Res. 52 (2024) D1465–D1477.
[61] J. Yin, W. Sun, F. Li, et al., Varidt 1.0: variability of drug transporter database, Nucleic Acids Res. 48 (2020) D1042–D1050.