Drug Response Prediction using Deep Learning Models and Single-Cell Perturbation Data across Pan Cancer Cell Lines

Phil Yao *

Newport High School, Newport, USA

* Corresponding Author Email: yaop2023@outlook.com

Abstract. Cancer remains a leading cause of death. Effective treatments remain limited, because of complications during therapies that contribute to poor survival outcomes. Developing new drugs can be expensive and time-consuming. One promising approach is drug repurposing by identifying potent treatments from existing drugs. This study aims at developing an efficient and effective framework for drug efficacy prediction. To achieve this goal, this study first processed publicly available drug perturbation cancer cell line datasets to enhance their usefulness for drug repurposing to cancer, in which this paper integrated perturbed pan cancer cell line gene expression data with their drug responses. This study integrated 10 selected cancer cell line data covering 8 cancers from public databases including CMAP, CTRP, CCLE, and GDSC. Secondly, this study applied advanced machine learning methods, particularly deep learning to learn useful features from these data, and transfer learning to make the predictive models generalized and interpretable. With deep learning and these datasets, this study developed a highly accurate deep neural network framework (DeepDrugCancer) that combines the drug molecular structures and gene expression data of cancer cell lines to predict the drug responsiveness. DeepDrugCancer achieved excellent performance with over 90% AUC test score across multiple cell lines. By integration with scRNA-seq data of breast cancer, predicted drugs such as aminocaproic acid, pancuronium bromide, and pseudopelletierine showed high cell-type specificity. Finally, this study conducted *in-vitro* (wet lab) experiments to validate the viability of DeepDrugCancer. The model predicted Mifanertinib to be sensitive on breast cancer cell lines and the *in-vitro* experiments confirmed the trend of such effects. The experiments have demonstrated that this framework efficiently utilizes drug-induced molecular signatures to predict therapeutic responses across certain cancers. By linking transcriptomic perturbations to drug responses, this study advances drug repurposing to expanded therapeutic options for cancer treatments as well as to precision oncology.

Keywords: Drugs repurpose, machine learning, transcriptomics, cancer.

1. Introduction

Cancer remains among the most prevalent and life-threatening diseases globally, with millions of new cases diagnosed each year [1]. Despite remarkable advances in early detection and therapeutic strategies, significant challenges to effective clinical management remain [2] because of substantial variability in patient responses to treatment. This heterogeneity arises from numerous factors, including genetic variations, epigenetic modifications, and environmental influences, underscoring the urgent need for developing personalized medicines. Tailoring treatments to individual patients can improve therapeutic efficacy, reduce adverse effects, and significantly lessen both physical and emotional burdens on patients [3]. However, this procedure is costly and time-consuming at this moment. An alternative yet promising approach is drug repurposing, with goals to find effective treatments using existing drugs even though those drugs were not originally designed for cancers [4].

Key challenges for drug repurposing include effective methods to predict drug responses. Recently, artificial intelligence (AI), particularly deep learning (DL), has emerged as a powerful tool for addressing the complexities of drug response prediction, for example, in oncology [5]. Deep learning methods can efficiently analyze vast, high-dimensional datasets spanning genomic, transcriptomic, and proteomic information, thereby revolutionizing traditional drug sensitivity assessment [6]. Additionally, machine learning techniques such as ensemble modeling have been successfully employed to identify predictive biomarkers, uncover novel therapeutic targets, and optimize drug

combinations, paving the way toward more precise and individualized cancer therapies [7]. In parallel, release of databases such as the Cancer Therapeutics Response Portal (CTRP), Genomics of Drug Sensitivity in Cancer (GDSC), Connectivity Map (CMAP), Cancer Cell Line Encyclopedia (CCLE), and PharmacoGx provide comprehensive molecular profiles and drug response data [8-12]. Leveraging these extensive resources, researchers have applied advanced AI methodologies, including transfer learning, graph-based neural networks, and multi-omics integration frameworks [13]. Transfer learning enables knowledge transfer across diverse datasets and disease contexts, mitigating challenges associated with limited annotated data [14, 15]. Graph-based models effectively capture intricate interactions among biological entities, such as genes, proteins, and drugs, facilitating discovery of novel molecular relationships and therapeutic pathways [16]. Meanwhile, multi-omics integration frameworks synthesize diverse molecular layers, offering a comprehensive understanding of tumor biology and significantly enhancing the predictive capabilities of AI models [17].

Despite ongoing advances and continuous improvement in predictive model performance, significant challenges remain. Issues such as inconsistent drug nomenclature across databases, varied experimental dosages, incomplete or missing data points, and limited annotation of certain cell lines hinder the clinical applicability and reproducibility of these AI-driven models. Addressing these challenges necessitates close collaboration among researchers, clinicians, and computational scientists to develop robust, interpretable, and clinically actionable predictive tools.

This study aims to explore and predict drug responses across pan-cancer cell lines through perturbation biology approaches [18, 19]. By integrating and harmonizing disparate datasets and employing state-of-the-art computational methodologies, this study aims to overcome current limitations related to data inconsistency and model generalizability. Ultimately, this research contributes to the development of clinically relevant AI-driven models, supporting the advancement of drug repurposing applied for precision oncology.

2. Materials and Methods

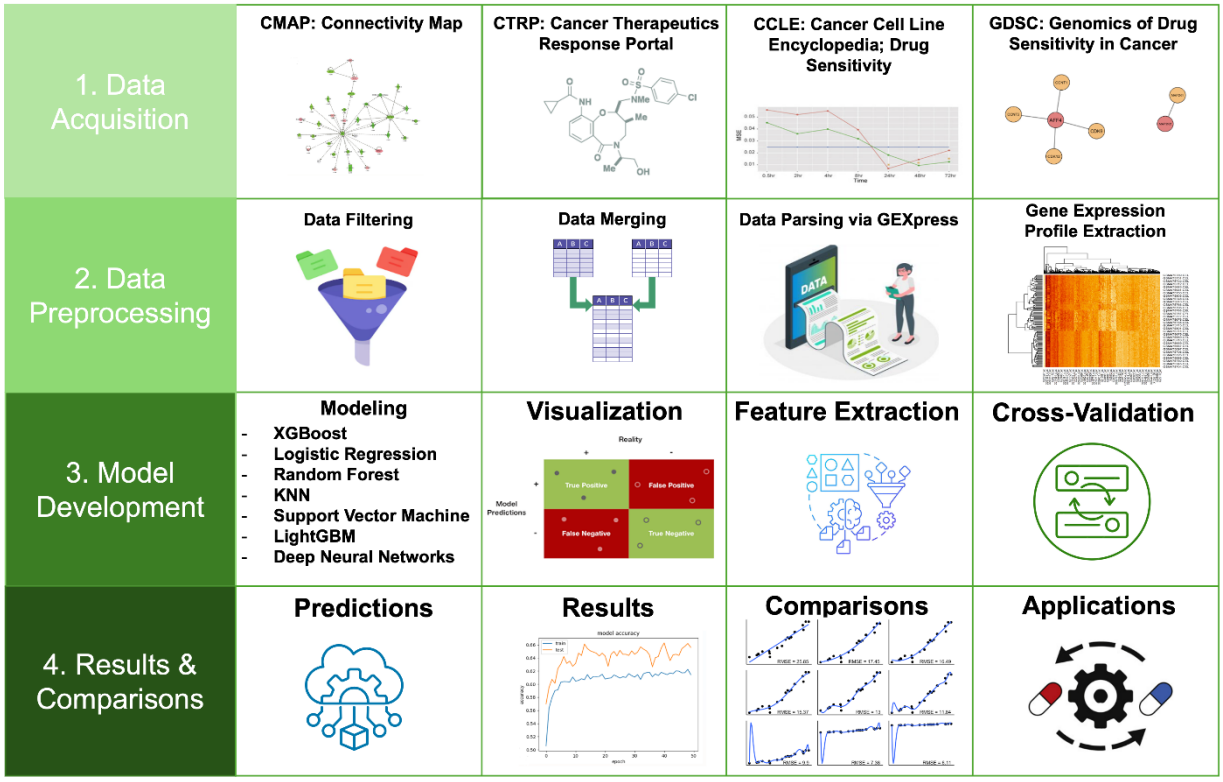

Figure 1. Project workflow for the design and evaluation of DeepDrugCancer.

(1) Acquiring data through scouring CMAP, CTRP, CCLE, and GDSC datasets. However, the design of DeepDrugCancer focused more heavily on CMAP and GDSC datasets. (2) Data

preprocessing through filtering, merging, and parsing CMAP and GDSC datasets, overcoming obstacles such as drug nomenclature. Additionally, gene expression profile extraction was a crucial step to obtaining molecular features to train the model. (3) Model development through applying and comparing various classification techniques, as well as cross-validation as an extra evaluation step. (4) Elicited results allowed for comparisons among classification models and cell lines, allowing for further predictions and applications to real world clinical trials and additional experimentations.

The design of the DeepDrugCancer framework consisted of four key phases (Fig. 1).

Phase 1 **Data Acquisition:**

This phase involved multiple publicly available datasets including but not limited to Connectivity Map (CMAP), Genomics of Drug Sensitivity in Cancer (GDSC), Cancer Therapeutics Response Portal (CTRP), and Cancer Cell Line Encyclopedia (CCLE), utilized as suggested by other papers in the field (see Introduction for references), while the primary datasets were CMAP and GDSC.

Phase 2 **Data Preprocessing:**

This phase aimed at streamlining model development and cleaning raw data. This process included filtering, merging, and parsing data via GEXpress to curate gene expression profiles of perturbed cancer cell lines exposed to different drugs.

Phase 3 **Model Development:**

This phase first conducted pretraining to learn embeddings. These embeddings mapped discrete molecular and gene features into their corresponding continuous valued representation feature vectors. On top of the projected features, various classification techniques were implemented and evaluated using Half-maximal Inhibitory Concentration (IC50) and Area under Drug Dosage Curve (AUC) metrics. To avoid overfitting, the models were subjected to cross-validation, confirming that both feature extraction and modeling could apply to unseen data. The optimal model was selected based on its cross-validation results. An initial attempt using a variational auto-encoder (VAE) for pretraining yielded poor performances, which led to a hypothesis that significant features for classification differ from those optimal for feature reconstruction (with VAE). Therefore, a classification approach using supervised learning was adopted.

Phase 4 **Results and Analysis:**

The performance of DeepDrugCancer was analyzed by comparing its 10 classifier candidates. Finally, DeepDrugCancer's predictions of some potential drugs for breast cancer, e.g., Mifanertinib, for repurposing were elicited and identified**.** This study conducted wet lab *in-vitro* analyses to investigate the drug response on breast cancer cell line (MCF-7) cell viability. This framework was further explored using scRNA-seq data of breast cancer data to investigate the cell type specificity of drug-related gene expression.

2.1. Dataset Acquisition

Drug molecular structure information, gene meta information, and gene expression data were downloaded from Connectivity Map (CMAP) that consists of 978 genes in 158 cell lines across 26 diseases, as well as DMSO control and drug treatment files in gctx format (level 3; https://clue.io/data/CMAP2020#LINCS2020). The drug response data across these cell lines were obtained in Genomics of Drug Sensitivity in Cancer (GDSC) website (https://www.cancerrxgene.org/downloads/anova). A subset of the top 10 cell lines across 8 diseases were selected, containing 1,710,340 instances of drug response data. For each drug treatment cell line, gene expression data were filtered to include only cell lines treated with a compound (treatment) or DMSO (control) for 24 hours, and a standardized drug dosage of 10 μM. Control and treatment expression data set for each cell line were then generated by parsing the control and treatment data with 978 landmark genes. The final dataset consisted of 41,236 (40,847 in treatment, 389 in control) experiments, involving 7,975 drugs for 10 cell lines.

2.2. Data Preprocessing - Drug Response and Molecular Feature Generation

A critical bottleneck in pharmacogenomics is the lack of standardized drug nomenclature. We utilized the PharmacoGx computational platform to create a mapping key based on canonical SMILES (Simplified Molecular Input Line Entry System) strings, resolving inconsistencies between CMAP and GDSC identifiers.

To facilitate correlation analysis with GDSC's gene expression profiles, molecular drug fingerprints were extracted using descriptors and a data frame with canonical SMILES was created [20]. The associated metadata dataset consisted of canonical SMILES, compound information, and experimental details. To extract sample information from each individual cell line, control restraints of compound control and DMSO control were imposed to produce two new datasets, DF-μM and DF-DMSO. Subsequently, the GDSC data (GDSC1 & 2; pan-cancer tissue analysis) were incorporated to obtain drug response information, specifically IC50 values, for 7,975 drugs across the 10 selected cell lines [21]. The target label of drug response is binary, with 1 corresponding to low IC50 (logarithmic, high drug response) values and 0 corresponding to high IC50 (logarithmic, low drug response) values [22].

The data was partitioned into 70% for training, 15% for validation, and 15% for test [23]. Features were normalized to mean zero and standard deviation of one. The resulting data are high dimensional. For example, MCF7 cell line dataset has 122 samples each with 1,190 features (212 molecular features and 978 gene features).

2.3. Model Development

2.3.1. Pretraining

Here, the goal was to pre-train a neural network. A deep learning architecture was developed to predict drug responses using gene expression profiles and molecular features of cell lines [24]. This architecture consists of three blocks.

The first block is an embedding layer that performs a linear projection of 1,190-dimensional input feature vectors to a 128-dimensional space. Each element in the input feature vector is binary, corresponding to whether a particular feature is present. The second block contains two encoding layers including a further processing of these 128-dimensional feature vectors with a dropout rate of 0.1 for regularization. The third block conducts feature classification: it projects the 128-dimensional feature vectors with another linear operation to new 128-dimensional feature vectors, followed by ReLU and a subsequent dropout layer with rate 0.5. Finally, a linear operation projects the 128-dimensional feature vectors to single-value scalar outputs.

The model was trained with learning rate 1e-4 and weight decay of 1e-3. A scheduler reduced learning rate by half if performance on a validation set was not decreasing after 5 epochs. To mitigate the effects of a small original training data size, e.g., 84 original samples, training samples were bootstrapped to 140 sampled instances. Mini-batch size was set to 128 for the validation, test, and resampled training datasets. Each training cycle had 200 epochs. In each epoch, gradients were initialized to 0 for all the model's layers. Models took the input molecular and gene expression features and then made predictions. These predictions were not normalized and therefore a Binary Cross Entropy (BCE) with Logits was applied for loss computation. Model parameters were updated with gradients from the back propagated losses at each layer. Performance metrics were the AUC and BCE loss, and they were evaluated on the train and validation sets throughout model optimization. Following training, the model was evaluated on test set using the same metrics. Other metrics like accuracy, precision, recall, and F1 were acknowledged throughout model optimization.

2.3.2. Transfer Learning

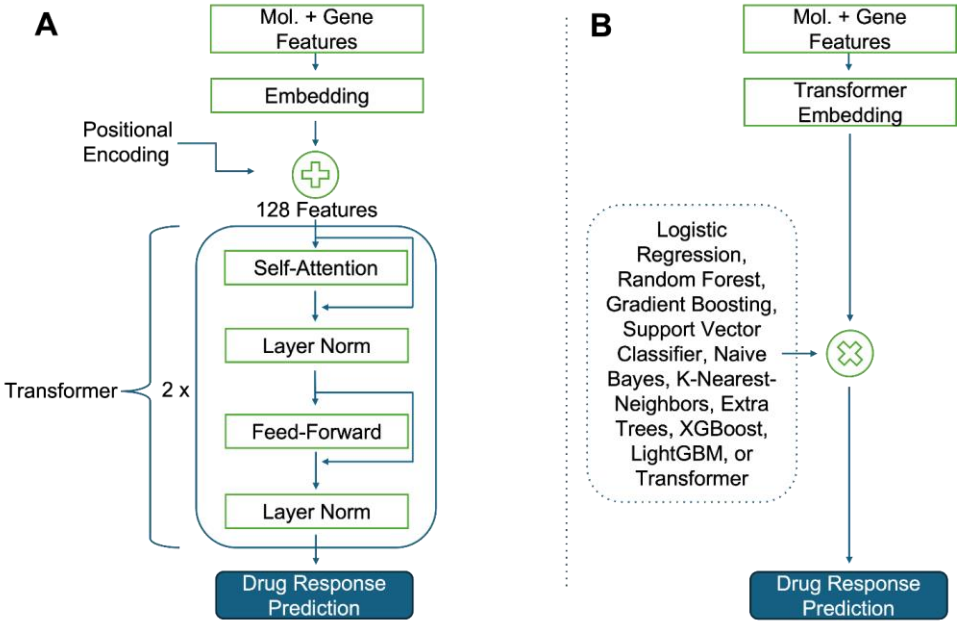

Figure 2. An overview of the DeepDrugCancer architecture.

(A) Model illustration for pretraining. It consists of an embedding layer, followed by a Transformer with two layers of encoders. The right figure illustrates DeepDrugCancer used for predictions. It is a composition of the pre-trained embeddings with a variety of classifiers, trained on top of the embeddings.

The embeddings obtained from the pre-trained model were taken to represent observed elements in the input feature vector. These input observations therefore were projected linearly into a latent space, which served as the input for nine traditional classifiers: Logistic Regression, Random Forest, Gradient Boosting, Support Vector Classifier, Naive Bayes, K-Nearest-Neighbors, Extra Trees, XGBoost, and LightGBM. They classify the projected features into binary outputs. The use of linear projections enhances both explainability and computing efficiency of the proposed framework.

In summary, this study has presented a complex neural network architecture. Its parameters were pretrained, and the pre-trained embedding layer provided a foundational representation of the molecular and differential gene expression features, upon which highly interpretable traditional classifiers were used to generate predictions (Fig. 2). This integrated architecture and its training mechanism constitute the DeepDrugCancer framework.

2.4. In-Vitro Validation of DeepDrugCancer Drug Response Predictions

Using DeepDrugCancer, this study identified Mifanertinib as a potentially sensitive drug on MCF7 cell lines using molecular structure and DeepDrugCancer model, suggesting it is a promising drug candidate for preclinical testing for breast cancer. To test this theory, this study performed *in-vitro* validation of the predicted drug response by wet lab experimentation. Logistically, Mifanertinib has undergone limited clinical testing for antineoplastic usage (which is at least less risk than no tests) but is commercially available.

To ascertain the viability of DeepDrugCancer and this prediction, a series of *in-vitro* analyses were performed. They consisted of four key components: extraction of cell RNA for transcriptomic sequencing, Western Blotting (WB), Cell Counting Kit-8 (CCK-8), and Annexin V treatment with propidium iodide (PI). These components respectively achieved detection of gene expression changes at the whole transcriptome level, confirmation of if changes in the transcription layer are also reflected at the protein function level, elicitation of the cells' population's viability after treatment with distinct dosages of the drug- namely functional effects, and obtaining photos to determine the cell-populations'

apoptotic nature after treatment with the drug that support target specificity. This *in-vitro* analysis via wet lab experimentation with Mifanertinib is a necessity to validate the DeepDrugCancer framework. DeepDrugCancer also identified other drugs that had public literature and DrugBank supporting DeepDrugCancer's predictions to be effective on trained cell lines (see Discussions).

2.4.1. Transcriptomic Sequencing

Total RNA was extracted from MCF7-k and MCF7 cells using TRIzol™ Reagent (Invitrogen, USA) by manufacturer protocol. Cells were briefly lysed in TRIzol™, followed by chloroform extraction and phase separation with centrifugation (12,000 g, 15 mins). The aqueous phase containing RNA was transferred to new tubes and precipitated with isopropanol. After centrifugation (12,000 g, 10 mins), pellets were combined with 75% ethanol, air-dried, and resuspended in DEPC water. RNA concentration was measured by spectrophotometer, and samples were kept at −70°C until further use. RNA samples were transcriptomic sequenced (see Results) externally (Azenta, Suzhou).

2.4.2. Western Blotting

Proteins were extracted from MCF7 cells treated with Mifanertinib. Same amounts of protein were loaded onto an SDS-PAGE gel aside a pre-stained protein ladder, separated at 80 V for 15 min followed by 120 V for 45 min, and transferred onto blotting paper via wet transfer system (300 A, 1h). Membranes were combined with TBST (Tris-buffered saline and 0.1% Tween-20) and blocked with a blocking buffer. Blotting papers were stored overnight at 4ºC with primary antibodies against β-actin for loading control, washed, then stored with anti-rabbit IgG, HRP-conjugated secondary antibodies overnight at 4ºC; blocking is significant to reduce the common problem of high background. Finally, protein bands were visualized with the Super ECL detection reagent (Thermo Fisher Scientific) and imaged via chemiluminescence detection machine.

2.4.3. Cell Viability Assay (CCK-8)

MCF7-k (knocked out) and MCF7 cells' viability were assessed using the Cell Counting Kit-8 (CCK-8; DOJINDO Laboratories). MCF7-k and MCF7 cells were pipetted into 30 wells at 100 µL per well and treated with Mifanertinib at 0.1 µM, 1 µM, and 10 µM. Lapatinib (1 µM) was included as positive control, while untreated wells received 0.1% (v/v) DMSO. After 24h of incubation at 37°C, 1 µL of CCK-8 reagent was added to each well. Plates were incubated at 37°C for 65 mins, then absorbance was measured with a Cytation 3 Cell Imaging Multi-Mode Reader (BioTek Instruments).

2.4.4. Annexin V Treatment

Apoptotic nature of cell lines was assessed leveraging Annexin V and propidium iodide (PI) staining (Thermo Fischer Scientific). MCF7-k and MCF7 cells were pipetted into 30 wells at 100 µL per well and treated with Mifanertinib (0.1 µM, 1 µM, 10 µM) and Lapatinib (1 µM) for 24 h in incubator. Untreated wells received 0.1% (v/v) DMSO. After treatment, cells were stained with Annexin V and PI diluted with PBS. Stained cells were then incubated for 25 mins at room temperature. Fluorescence imaging in the dark was performed to evaluate apoptosis with captured Annexin V (max emission: ~530 nm) and PI (max emission: ~636 nm) signals.

2.5. Integration with scRNA-seq Data to Investigate the Cell Type Specificity of Drug Related Gene Expression

DeepDrugCancer could be possibly used in the more heterogeneous transcriptomic landscape of clinical tumors, even though it might be trained for drug sensitivity prediction from cell line–based data. To evaluate this translational potential of DeepDrugCancer, this study applied single-cell RNA sequencing (scRNA-seq) data from patient-derived breast cancer samples with scDrugAct database [25, 26]. This approach enabled predicted drug sensitivities from the cell line–based model to show cell type specificities [27].

3. Results

3.1. Exploratory Data Analysis

The initial analysis of the integrated CMAP-GDSC dataset highlighted a significant heterogeneity in data availability across different cancer models. While lung (A549) and colon (HT29) cancer lines were well-represented, MCF-7 stood out as a particularly rich resource for breast cancer research. It ranked among the top cell lines for both the number of transcriptional experiments in CMAP and the diversity of pharmacological screenings in GDSC (see Figure 3.). This high density of high-quality, overlapping data provided a robust foundation for training the DeepDrugCancer model, minimizing the "sparse data" problem often encountered in pharmacogenomics and justifying the selection of MCF-7 as the primary model for downstream validation.

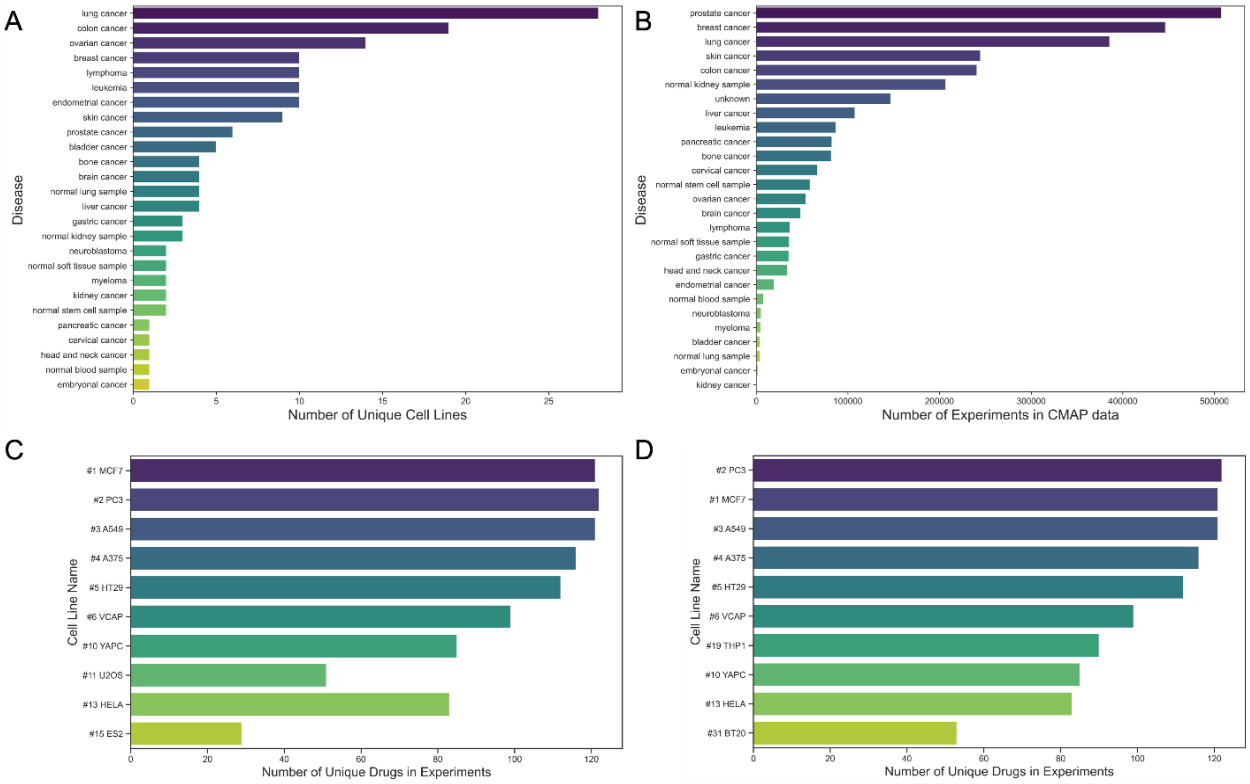

Figure 3. Overview of CMAP dataset and overlap with GDSC dataset.

(A) Number of unique cell lines in diseases in CMAP dataset. The top 3 are lung cancer, colon cancer, and ovarian cancer. (B) Number of experiments for each disease in CMAP dataset. Prostate cancer, breast cancer, and lung cancer have the most conducted experiments. (C) The number to the left of each cell line indicates its rank in the CMAP dataset for most experiments, and thus most data abundance. Number of unique drugs used in the experiments with the perturbed gene expression profiles from the GDSC dataset for each cell line presented in both CMAP and GDSC datasets. The selected 10 cell lines (in both datasets) with the highest CMAP experiment data are displayed. (D) Number of unique drugs used in the experiments with the perturbed gene expression profiles from the GDSC dataset for each cell line present in both CMAP and GDSC datasets. The top 10 cell lines (in both datasets) with the highest number of unique GDSC drugs tested are shown.

Index	Disease Name	Cell Line Name	Experiment Counts	Total Drug Count (in CMAP)	Drug Count (w/ GDSC)	% GDSC Drugs with IC50 < 0
1	Prostate Cancer	PC3	296688	8703	122	16.4%
2	Breast Cancer	MCF7	302664	7975	121	22.3%
3	Lung Cancer	A549	261333	6742	121	24.8%
4	Skin Cancer	A375	231121	4740	116	31.0%
5	Colon Cancer	HT29	220704	4109	112	25.9%
6	Prostate Cancer	VCAP	205266	6290	99	19.2%
7	Leukemia	THP1	34298	1250	90	16.7%
8	Pancreatic Cancer	YAPC	82732	1691	85	10.6%
9	Cervical Cancer	HELA	66746	1645	83	21.7%
10	Breast Cancer	BT20	8788	185	53	11.3%

Table 1. Top 10 Cell Lines in CMAP with most GDSC drugs tested.

Additional information includes experiment counts in CMAP, Total drugs tested with CMAP, total drugs tested with GDSC, and percent GDSC tested drugs that induce response on respective cell lines (logarithmic IC50 < 0). These are not the 10 selected cell lines DeepDrugCancer trained on.

The above analysis revealed large variations of drugs, diseases, and experiments distributions across different cell lines (Table 1). Model performances were generally stronger for cell lines with greater experimental data and drug data abundance. This posed a challenge for training models for DeepDrugCancer, as it aimed to cover up to 122 distinct drugs and hundreds of thousands of experiments.

3.2. Model Performance and Evaluation

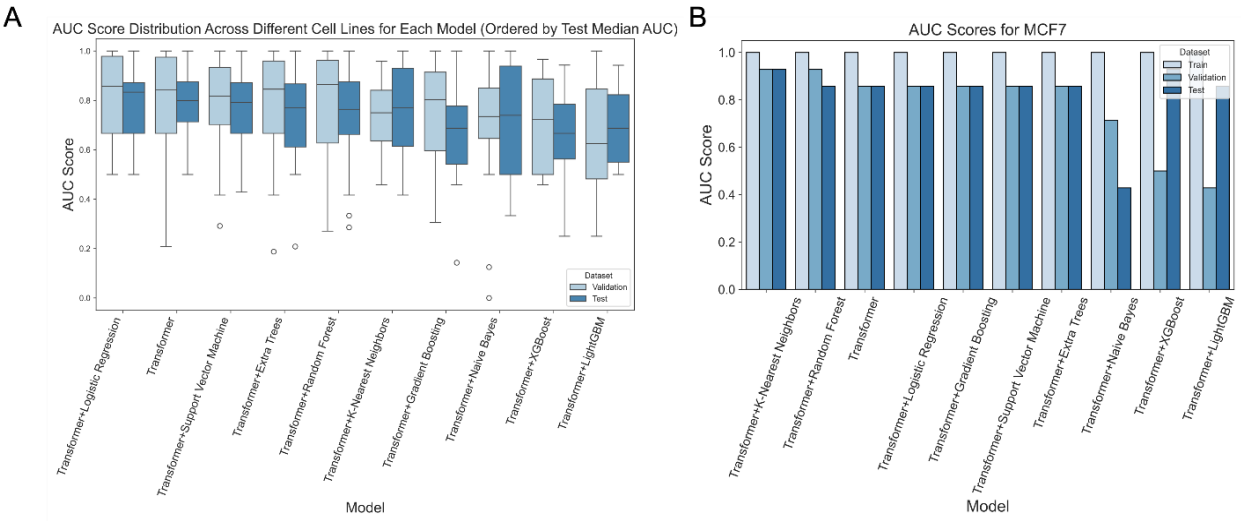

Figure 4. The Transformer + classifiers refer to the fact that the Transformer was used for pretraining the embedding, albeit the Transformer embedding and Classification is the actual applied framework for predicting drug response. The Transformer-only model above refers to using the embedding and Transformer for prediction.

(A) AUC score distribution across different cell lines for each model, ordered by test median AUC. Transformer + Logistic Regression had the best performance with mean AUC of 82.5% across all cell lines. (B) AUC scores for different classifiers on MCF7 cell line, ranked with validation score. KNN performed best with Transformer + Logistic Regression performance tied for third.

I measured cell line responsiveness, defined as a binary classification task on the logarithmic IC50 value [28]. High drug sensitivity (1) corresponds to a negative logarithmic IC50 value and low drug sensitivity (0) corresponds to a positive logarithmic IC50 value. DeepDrugCancer was used to predict

drug sensitivity on the test set for the selected 10 cell lines. Also, Transformer and nine traditional classification models across all cell lines (Fig. 4A) used in DeepDrugCancer were compared, and this study observed that Logistic Regression classifier performed the best. It attained an average test AUC score of 82.5% across the 10 cell lines.

Further comparison of model performances within MCF7 show Logistic Regression classifier to have third best performance (Fig. 4B). This study posits that the model's simplicity reduces its tendency to overfit on the limited datasets compared to other more complex classifiers. Furthermore, using simple algorithms like Logistic Regression enhances the interpretability of the results. DeepDrugCancer therefore adopted the Transformer for pretraining and Transformer Embedding + Logistic Regression, denoted as Transformer + Logistic Regression in the following, for classification, as this combination yielded the highest AUC score in classifying all cell line responsiveness, in comparison to other classifiers.

3.3. Model Interpretability Analysis

The model has an attractive property of interpretability as it uses a simple Logistic Regression classifier on learned embedding features. To evaluate this property and validate the model's biological relevance, this study conducted an interpretability analysis during exploration on a known drug Vorinostat, a drug used to treat T-Cell Lymphoma. The analysis aimed to reveal the nuanced relationships between the drug and a cell line's gene expression. Out of 978 genes, 55 were upregulated after treatment with Vorinostat (Fig. 5A). To understand the functional implications on these genes, genes that were associated with Vorinostat's effects on cell lines were analyzed and enriched for functional pathways. The odds ratio of disease pathway responsiveness to Vorinostat were calculated through KEGG pathway analysis (Fig. 5B), following established methodologies [29, 30]. Finally, to quantify the contribution of individual genes to the model's predictions of drug responsiveness, SHAP (SHapley Additive exPlanations) values were computed for metrics including IC50 and AUC (Fig. 5C and Fig. 5D) on the top 20 genes with the highest sensitivity to Vorinostat [31].

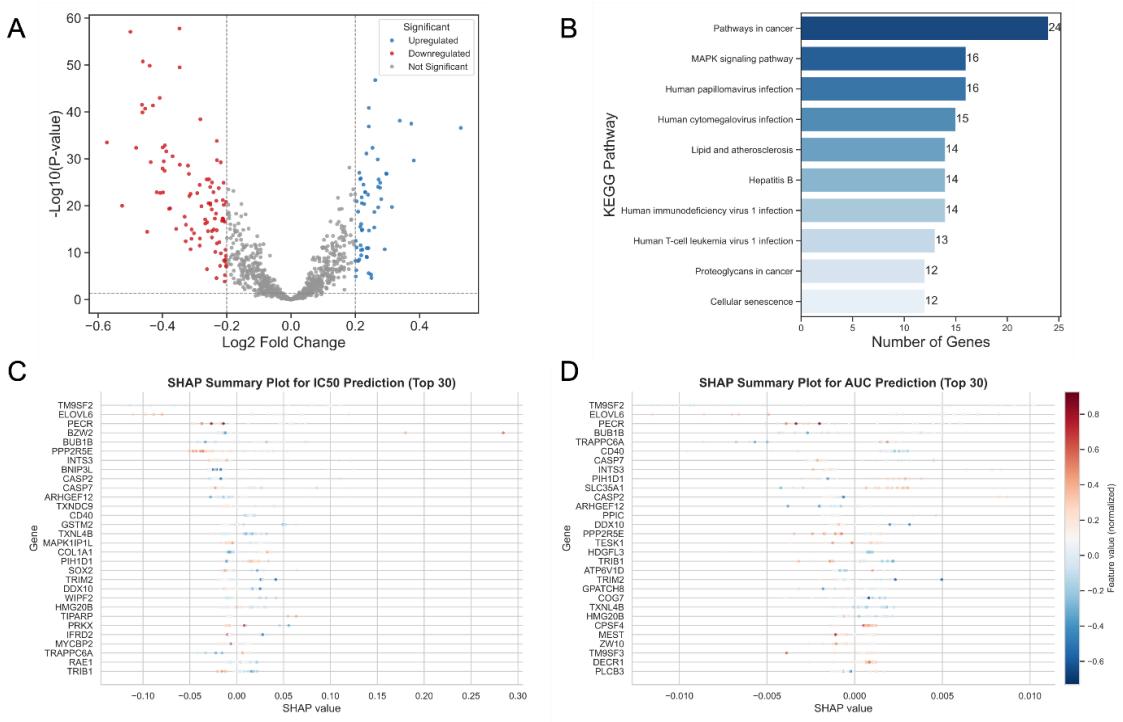

Figure 5. Overview of Vorinostat's effect on landmark genes and cell lines.

(A) Volcano plot identifying upregulated and downregulated genes in response to treatment with Vorinostat. Log2 Fold Change is used. (B) Odds ratio of 10 distinct enriched KEGG metabolic pathways after treatment with Vorinostat. (C) SHAP summary plot for IC50 prediction on various

cell lines after treatment with Vorinostat. (D) SHAP summary plot for AUC prediction on various cell lines after treatment with Vorinostat.

3.4. In-Vitro Experiment Validation - HER2 Inhibition by Mifanertinib Suppresses Tumor Cell Growth and Promotes Apoptosis

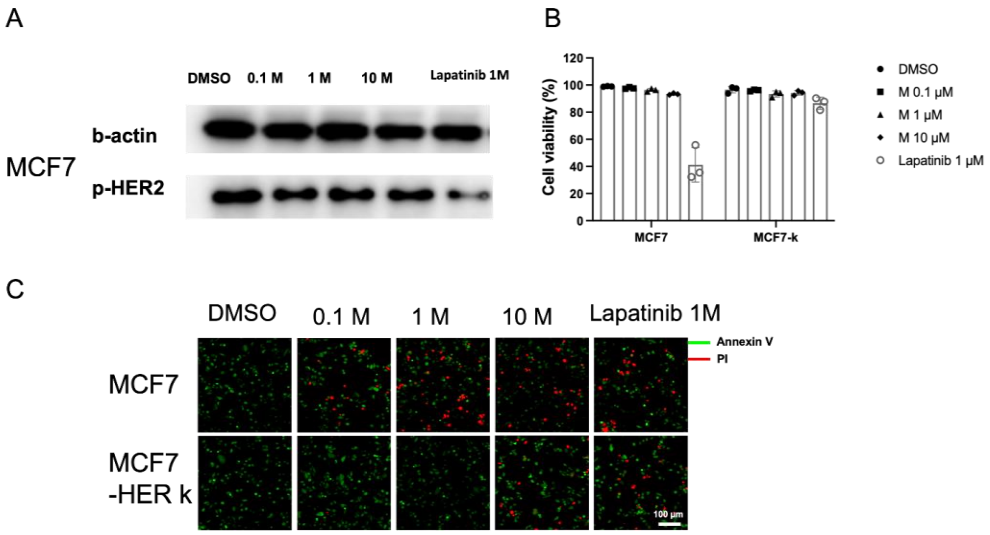

Figure 6. Results from wet lab *in-vitro* experiments with Mifanertinib on MCF7 and MCF7-k cell lines.

(A) Western blot results from protein samples with various drug treatment conditions as columns and antibodies used as rows. (B) % cell viability for CCK8 cell viability experiments with DMSO control, low, medium, and high dosage Mifanertinib treatment, and Lapatinib positive control on MCF7 and MCF7-k cells. (C) Annexin V and PI Treatment experiment with DMSO, Mifanertinib, and Erlotinib treatment dosages on MCF7 and MCF7-HER2-k cell images for apoptosis.

Mifanertinib selectively inhibited HER2 phosphorylation in MCF7 cells in a dose-dependent manner without affecting total HER2 protein levels, similar to the effects of Lapatinib (Fig. 6A). Knockdown controls showed no significant differences, suggesting the effects were drug-specific rather than gene interaction–dependent. Functionally, Mifanertinib treatment reduced cell viability (Fig. 6B) and increased apoptosis, as evidenced by Annexin V/PI staining (Fig. 6C). These results demonstrate that Mifanertinib acts as a HER2 tyrosine kinase inhibitor, translating biochemical suppression into strong anti-proliferative and pro-apoptotic effects. Notably, the extent of HER2 phosphorylation inhibition correlated with phenotypic changes, supporting a direct mechanistic link. Compared with single-target inhibitors, the dual activity of Mifanertinib on EGFR and HER2 suggests a broader therapeutic window, particularly in tumors with compensatory signaling between these pathways. Collectively, these findings highlight Mifanertinib's repurposing potential and validate DeepDrugCancer as an effective model for predicting drug efficacy.

Pathway enrichment analysis revealed distinct transcriptional responses between Mifanertinib- and Lapatinib-treated MCF7 cells. Mifanertinib (0.1 μM) significantly perturbed innate immune and inflammatory signaling, including Toll-like receptor, NOD-like receptor, and NF-κB pathways, as well as nucleotide excision repair and multiple virus-associated pathways (Fig. 7A). In contrast, Lapatinib (1 μM) preferentially affected Wnt and MAPK signaling, PPAR signaling, and adhesion- and metabolism-related processes (Fig. 7B). These differences suggest that although both drugs converge on HER2 inhibition, Mifanertinib elicits broader transcriptional rewiring of immune-related pathways, whereas Lapatinib primarily impacts canonical oncogenic signaling. This divergence highlights potential mechanistic distinctions between the two inhibitors and suggests that Mifanertinib may exert additional immunomodulatory effects beyond HER2 kinase blockade. This finding shows that Mifanertinib compares favorably with Lapatinib, a drug actively used to treat

advanced metastatic (already spread) HER2-receptor positive breast cancer [32] and thus contributes significant insight to the potential usage of Mifanertinib for breast cancer treatment.

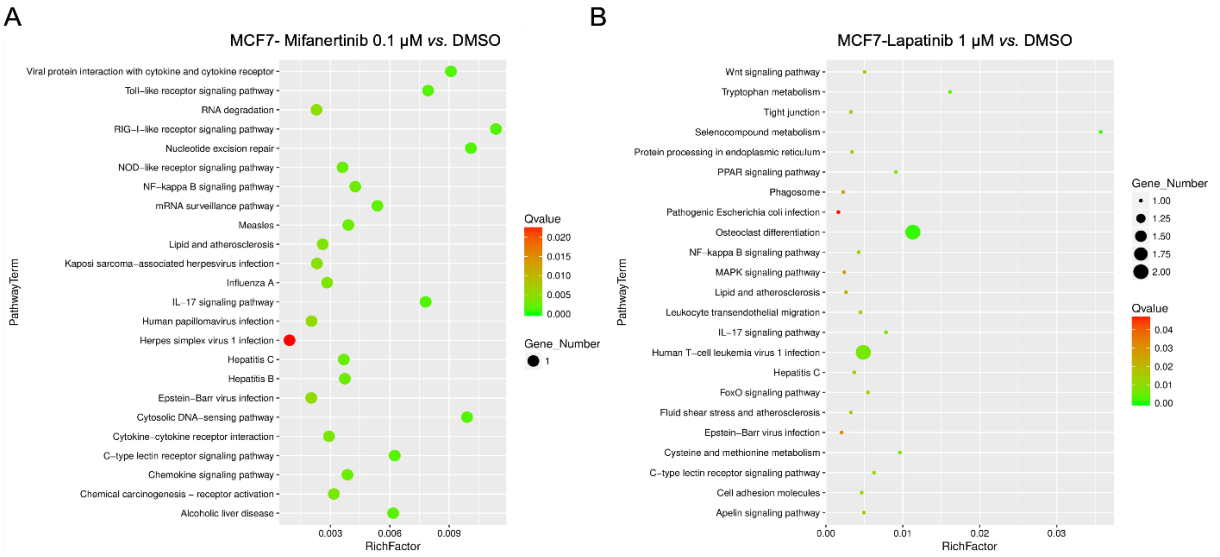

Figure 7. KEGG pathway enrichment analysis of MCF7 cells treated with Mifanertinib or Lapatinib.

(A) Pathway enrichment in MCF7 cells treated with Mifanertinib (0.1 μM) versus DMSO control. Enriched pathways included immune signaling cascades (e.g., Toll-like receptor, RIG-I-like receptor, NOD-like receptor, NF-κB signaling), DNA damage repair (nucleotide excision repair), and viral infection–related pathways. Dot size corresponds to the number of enriched genes, while color represents q-values. (B) Pathway enrichment in MCF7 cells treated with Lapatinib (1 μM) versus DMSO. Lapatinib treatment was associated with distinct pathway signatures, including Wnt signaling, MAPK signaling, PPAR signaling, and adhesion- and metabolism-related pathways.

3.5. Results from Integration with scRNA-seq

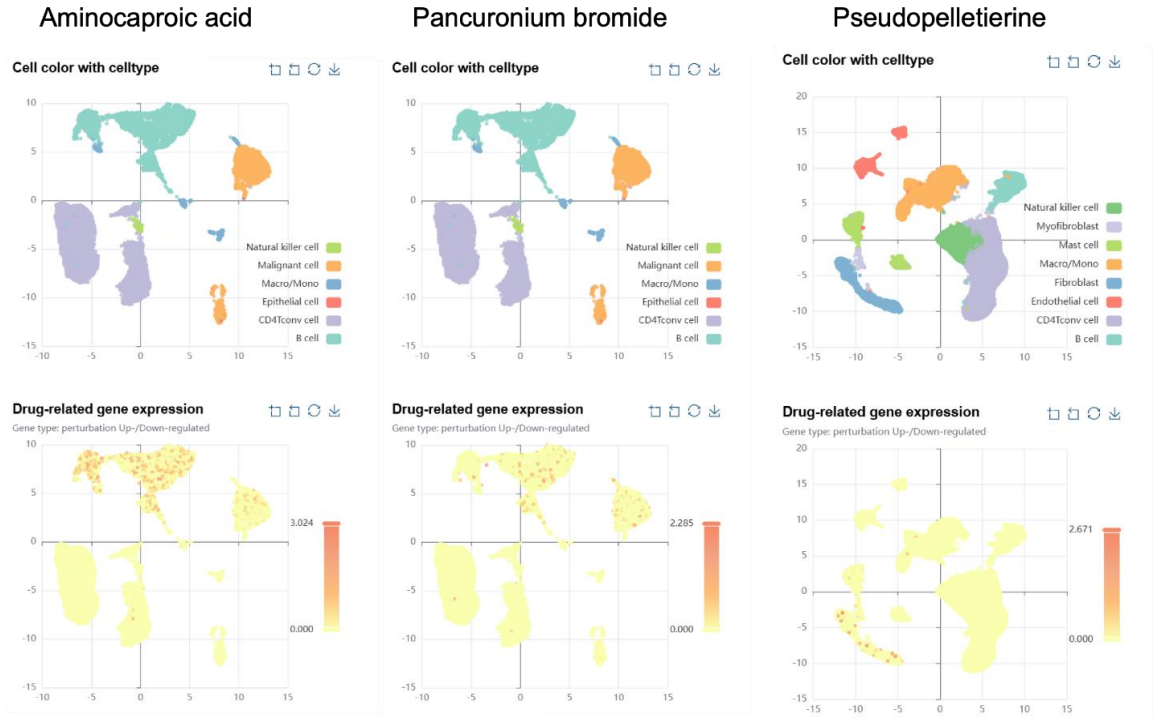

Figure 8. Single-cell validation of drug response predictions in patient-derived breast cancer samples.

UMAP projections of patient-derived scRNA-seq data showing cell type annotations (top row) and drug-related gene expression patterns (bottom row). Distinct transcriptional perturbations were observed in malignant epithelial cell clusters, while stromal and immune populations exhibited minimal responses. These results demonstrate that drug-related gene expression is cell type–specific, highlighting intra-tumor heterogeneity in drug sensitivity at single-cell resolution.

To evaluate DeepDrugCancer predictions in a clinically relevant context, this study integrated single-cell RNA sequencing (scRNA-seq) data from patient-derived breast cancer samples with scDrugAct website [26] (Fig. 8). The drug-related gene expression mapped across major cell populations revealed that transcriptional perturbations were not uniformly distributed but instead showed clear cell type specificity. In particular, malignant epithelial subpopulations exhibited stronger drug-related expression changes compared to stromal and immune cells, indicating that therapeutic effects were concentrated within tumor compartments. This intra-tumor heterogeneity underscores the cell-type specific drug responses in breast cancer. Collectively, these results highlight that DeepDrugCancer not only predicts drug efficacy but can also point out drugs with cell population–specific effectiveness, supporting the model's potential utility in precision oncology. This cross-context validation underscores the potential for using cell line trained models to infer patient specific drug responses at single-cell resolution, thereby supporting future precision oncology applications [33].

4. Discussions

4.1. Mifanertinib: A New Weapon in the Breast Cancer Arsenal

The identification of Mifanertinib as a potent agent for breast cancer is the most consequential outcome of this study. Originally developed as a third-generation EGFR inhibitor for non-small cell lung cancer (NSCLC), specifically to target resistance mutations like T790M, its utility in breast cancer has been largely unexplored.

Our data indicates that Mifanertinib acts as a potent HER2 inhibitor in breast cancer cells. While MCF-7 cells are clinically categorized as HER2-low or HER2-negative , they are known to express functional HER2 receptors that engage in crosstalk with the Estrogen Receptor (ER) pathway, often driving resistance to endocrine therapies like tamoxifen. By effectively blocking HER2 phosphorylation, Mifanertinib disrupts this resistance axis. This positions it as a promising candidate for HER2-low breast cancers, a massive patient population that has recently become the focus of intense clinical interest following the success of antibody-drug conjugates like T-DXd.

Furthermore, the drug's unique ability to activate NF-κB and TLR pathways suggests it operates via a dual mechanism: direct oncogene inhibition coupled with immune stimulation. This immunomodulatory profile distinguishes it from "pure" kinase inhibitors and suggests potential synergy with immune checkpoint blockade, a hypothesis that warrants immediate investigation in immunocompetent animal models.

4.2. Cross-Validation with Clinical Trials

In the *in-vitro* wet lab experiments, this study chose Mifanertinib as the drug to be tested. However, this study came across other drugs that the model predicted as effective (see Supplementary Details for list) backed by public literature. For example, DeepDrugCancer classified Ceritinib as high sensitivity on MCF7 cell line. This study cross-validated this prediction with public literature, which showed that Ceritinib was a "novel triple negative breast cancer therapeutic agent" [34]. Furthermore, Ceritinib acts as a protein kinase inhibitor and can trigger immunogenic cell death via on-target effects [35, 36].

Another example is Purmorphamine, a drug in its exploration phase (at the time of this study) that DeepDrugCancer predicted to induce responsiveness on the MCF7 cell line [37]. MCF7 is a well-studied cell line with nearly 50 years of research [38], which corresponds to greater resources to be utilized by researchers, such as single-cell data for breast cancer. For example, in the CMAP dataset,

MCF7 stood out in experiments conducted and in the GDSC dataset, MCF7 stood out in drugs tested. MCF7 is therefore more suitable for this study and likely yields better results than if an alternative cell line with less data was used.

4.3. The Power of Transformers in Pharmacogenomics

Methodologically, this study validates the superiority of Transformer-based embeddings for pharmacogenomic tasks. Traditional descriptors often fail to capture the context of molecular features—how a specific chemical substructure behaves differently depending on the global topology of the molecule. Transformers, via self-attention, naturally encode this context. Our results show that these "context-aware" embeddings linearize the complex landscape of drug response, enabling even simple classifiers like Logistic Regression to achieve high accuracy. This has profound implications for AI interpretability: by shifting the complexity to the feature extractor (the Transformer), we can use simpler, more transparent models for the final decision, facilitating the "trust" required for clinical adoption.

4.4. Limitations and Future Directions

While promising, this study is not without limitations. Cell line models, including MCF-7, are imperfect proxies for clinical tumors, lacking the full 3D architecture and metabolic heterogeneity of patient tissue. Although our single-cell validation mitigates this, future work must incorporate patient-derived organoids (PDOs) and spatial transcriptomics to fully capture the nuances of the TME. Additionally, while we validated HER2 inhibition, the dual-kinase nature of Mifanertinib (EGFR/HER2) requires further dissection to determine the exact contribution of EGFR inhibition to the observed phenotype in breast cancer cells.

5. Conclusions

DeepDrugCancer represents a significant leap forward in AI-driven drug discovery. By harmonizing massive datasets and leveraging the representational power of Transformer models, we have built a robust engine for identifying high-value drug repurposing candidates. The discovery and validation of Mifanertinib as a dual-action (anti-HER2/immunomodulatory) agent for breast cancer provides a compelling proof-of-concept. Coupled with the identification of unconventional candidates like aminocaproic acid and pancuronium, this study underscores the vast, untapped potential within the existing pharmacopeia. As we move towards the era of precision oncology, frameworks like DeepDrugCancer will be instrumental in delivering rapid, cost-effective, and personalized therapeutic solutions to patients

References

[1] Ahmad, F. B.; Anderson, R. N. The Leading Causes of Death in the US for 2020. *JAMA* **2021**, *325* (18), 1829. https://doi.org/10.1001/jama.2021.5469.

[2] Jaehde, U.; Liekweg, A.; Simons, S.; Westfeld, M. Minimising Treatment-Associated Risks in Systemic Cancer Therapy. *Pharm. World Sci.* **2008**, *30* (2), 161–168. https://doi.org/10.1007/s11096-007-9157-4.

[3] Tanoli, Z.; Vähä-Koskela, M.; Aittokallio, T. Artificial Intelligence, Machine Learning, and Drug Repurposing in Cancer. *Expert Opin. Drug Discov.* **2021**, *16* (9), 977–989. https://doi.org/10.1080/17460441.2021.1883585.

[4] Saranraj, K.; Kiran, P. U. Drug Repurposing: Clinical Practices and Regulatory Pathways. *Perspect. Clin. Res.* **2025**, *16* (2), 61–68. https://doi.org/10.4103/picr.picr_70_24.

[5] Failli, M.; Demir, S.; Del Río-Álvarez, Á.; Carrillo-Reixach, J.; Royo, L.; Domingo-Sàbat, M.; Childs, M.; Maibach, R.; Alaggio, R.; Czauderna, P.; Morland, B.; Branchereau, S.; Cairo, S.; Kappler, R.; Armengol, C.; Di Bernardo, D. Computational Drug Prediction in Hepatoblastoma by Integrating Pan-Cancer Transcriptomics with Pharmacological Response. *Hepatology* **2024**, *80* (1), 55–68. https://doi.org/10.1097/HEP.0000000000000601.

[6] Issa, N. T.; Stathias, V.; Schürer, S.; Dakshanamurthy, S. Machine and Deep Learning Approaches for Cancer Drug Repurposing. *Semin. Cancer Biol.* **2021**, *68*, 132–142. https://doi.org/10.1016/j.semcancer.2019.12.011.

[7] Al Khzem, A. H.; Gomaa, M. S.; Alturki, M. S.; Tawfeeq, N.; Sarafroz, M.; Alonaizi, S. M.; Al Faran, A.; Alrumaihi, L. A.; Alansari, F. A.; Alghamdi, A. A. Drug Repurposing for Cancer Treatment: A Comprehensive Review. *Int. J. Mol. Sci.* **2024**, *25* (22), 12441. https://doi.org/10.3390/ijms252212441.

[8] Baptista, D.; Ferreira, P. G.; Rocha, M. Deep Learning for Drug Response Prediction in Cancer. *Brief. Bioinform.* **2021**, *22* (1), 360–379. https://doi.org/10.1093/bib/bbz171.

[9] Yang, W.; Soares, J.; Greninger, P.; Edelman, E. J.; Lightfoot, H.; Forbes, S.; Bindal, N.; Beare, D.; Smith, J. A.; Thompson, I. R.; Ramaswamy, S.; Futreal, P. A.; Haber, D. A.; Stratton, M. R.; Benes, C.; McDermott, U.; Garnett, M. J. Genomics of Drug Sensitivity in Cancer (GDSC): A Resource for Therapeutic Biomarker Discovery in Cancer Cells. *Nucleic Acids Res.* **2012**, *41* (D1), D955–D961. https://doi.org/10.1093/nar/gks1111.

[10] Lamb, J.; Crawford, E. D.; Peck, D.; Modell, J. W.; Blat, I. C.; Wrobel, M. J.; Lerner, J.; Brunet, J.-P.; Subramanian, A.; Ross, K. N.; Reich, M.; Hieronymus, H.; Wei, G.; Armstrong, S. A.; Haggarty, S. J.; Clemons, P. A.; Wei, R.; Carr, S. A.; Lander, E. S.; Golub, T. R. The Connectivity Map: Using Gene-Expression Signatures to Connect Small Molecules, Genes, and Disease. *Science* **2006**, *313* (5795), 1929–1935. https://doi.org/10.1126/science.1132939.

[11] Musa, A.; Ghoraie, L. S.; Zhang, S.-D.; Galzko, G.; Yli-Harja, O.; Dehmer, M.; Haibe-Kains, B.; Emmert-Streib, F. A Review of Connectivity Map and Computational Approaches in Pharmacogenomics. *Brief. Bioinform.* **2017**, bbw112. https://doi.org/10.1093/bib/bbw112.

[12] Barretina, J.; Caponigro, G.; Stransky, N.; Venkatesan, K.; Margolin, A. A.; Kim, S.; J.Wilson, C.; Lehár, J.; Kryukov, G. V.; Sonkin, D.; Reddy, A.; Liu, M.; Murray, L.; Berger, M. F.; Monahan, J. E.; Morais, P.; Meltzer, J.; Korejwa, A.; Jané-Valbuena, J.; Mapa, F. A.; Thibault, J.; Bric-Furlong, E.; Raman, P.; Shipway, A.; Engels, I. H.; Cheng, J.; Yu, G. K.; Yu, J.; Aspesi, P.; De Silva, M.; Jagtap, K.; Jones, M. D.; Wang, L.; Hatton, C.; Palescandolo, E.; Gupta, S.; Mahan, S.; Sougnez, C.; Onofrio, R. C.; Liefeld, T.; MacConaill, L.; Winckler, W.; Reich, M.; Li, N.; Mesirov, J. P.; Gabriel, S. B.; Getz, G.; Ardlie, K.; Chan, V.; Myer, V. E.; Weber, B. L.; Porter, J.; Warmuth, M.; Finan, P.; Harris, J. L.; Meyerson, M.; Golub, T. R.; Morrissey, M. P.; Sellers, W. R.; Schlegel, R.; Garraway, L. A. Addendum: The Cancer Cell Line Encyclopedia Enables Predictive Modelling of Anticancer Drug Sensitivity. *Nature* **2012**, *492* (7428), 290–290. https://doi.org/10.1038/nature11735.

[13] Partin, A.; Brettin, T. S.; Zhu, Y.; Narykov, O.; Clyde, A.; Overbeek, J.; Stevens, R. L. Deep Learning Methods for Drug Response Prediction in Cancer: Predominant and Emerging Trends. *Front. Med.* **2023**, *10*, 1086097. https://doi.org/10.3389/fmed.2023.1086097.

[14] Jafari, M.; Tao, X.; Barua, P.; Tan, R.-S.; Acharya, U. R. Application of Transfer Learning for Biomedical Signals: A Comprehensive Review of the Last Decade (2014–2024). *Inf. Fusion* **2025**, *118*, 102982. https://doi.org/10.1016/j.inffus.2025.102982.

[15] Zhao, Z.; Alzubaidi, L.; Zhang, J.; Duan, Y.; Gu, Y. A Comparison Review of Transfer Learning and Self-Supervised Learning: Definitions, Applications, Advantages and Limitations. *Expert Syst. Appl.* **2024**, *242*, 122807. https://doi.org/10.1016/j.eswa.2023.122807.

[16] Gaudelet, T.; Day, B.; Jamasb, A. R.; Soman, J.; Regep, C.; Liu, G.; Hayter, J. B. R.; Vickers, R.; Roberts, C.; Tang, J.; Roblin, D.; Blundell, T. L.; Bronstein, M. M.; Taylor-King, J. P. Utilizing Graph Machine Learning within Drug Discovery and Development. *Brief. Bioinform.* **2021**, *22* (6), bbab159. https://doi.org/10.1093/bib/bbab159.

[17] Sartori, F.; Codicè, F.; Caranzano, I.; Rollo, C.; Birolo, G.; Fariselli, P.; Pancotti, C. A Comprehensive Review of Deep Learning Applications with Multi-Omics Data in Cancer Research. *Genes* **2025**, *16* (6), 648. https://doi.org/10.3390/genes16060648.

[18] Bunne, C.; Stark, S. G.; Gut, G.; Del Castillo, J. S.; Levesque, M.; Lehmann, K.-V.; Pelkmans, L.; Krause, A.; Rätsch, G. Learning Single-Cell Perturbation Responses Using Neural Optimal Transport. *Nat. Methods* **2023**, *20* (11), 1759–1768. https://doi.org/10.1038/s41592-023-01969-x.

[19] Nyman, E.; Stein, R. R.; Jing, X.; Wang, W.; Marks, B.; Zervantonakis, I. K.; Korkut, A.; Gauthier, N. P.; Sander, C. Perturbation Biology Links Temporal Protein Changes to Drug Responses in a Melanoma Cell Line. *PLOS Comput. Biol.* **2020**, *16* (7), e1007909. https://doi.org/10.1371/journal.pcbi.1007909.

[20] (20)An, X.; Chen, X.; Yi, D.; Li, H.; Guan, Y. Representation of Molecules for Drug Response Prediction. *Brief.* Bioinform. **2022**, *23* (1), bbab393. https://doi.org/10.1093/bib/bbab393.

[21] Zhang, X.; Cha, I.-H.; Kim, K.-Y. Use of a Combined Gene Expression Profile in Implementing a Drug Sensitivity Predictive Model for Breast Cancer. *Cancer Res. Treat.* **2017**, *49* (1), 116–128. https://doi.org/10.4143/crt.2016.085.

[22] Pozdeyev, N.; Yoo, M.; Mackie, R.; Schweppe, R. E.; Tan, A. C.; Haugen, B. R. Integrating Heterogeneous Drug Sensitivity Data from Cancer Pharmacogenomic Studies. *Oncotarget* **2016**, *7* (32), 51619–51625. https://doi.org/10.18632/oncotarget.10010.

[23] Joseph, V. R. Optimal Ratio for Data Splitting. *Stat. Anal. Data Min. ASA Data Sci. J.* **2022**, *15* (4), 531–538. https://doi.org/10.1002/sam.11583.

[24] Jiang, J.; Chen, L.; Ke, L.; Dou, B.; Zhang, C.; Feng, H.; Zhu, Y.; Qiu, H.; Zhang, B.; Wei, G.-W. A Review of Transformer Models in Drug Discovery and Beyond. *J. Pharm. Anal.* **2025**, *15* (6), 101081. https://doi.org/10.1016/j.jpha.2024.101081.

[25] Jiangping, H.; Lihui, L.; Jiekai, C.; Center for Cell Lineage and Atlas (CCLA), Bioland Laboratory (Guangzhou Regenerative Medicine and Health Guangdong Laboratory), Guangzhou 510320, China; Key Laboratory of Regenerative Biology of the Chinese Academy of Sciences and Guangdong Provincial Key Laboratory of Stem Cell and Regenerative Medicine, Guangzhou Institutes of Biomedicine and Health, Chinese Academy of Sciences, Guangzhou 510530, China. Practical Bioinformatics Pipelines for Single-Cell RNA-Seq Data Analysis. *Biophys. Rep.* **2022**, *8* (3), 158–169. https://doi.org/10.52601/bpr.2022.210041.

[26] Xu, Y.; Zhang, Y.; Song, K.; Liu, J.; Zhao, R.; Zhang, X.; Pei, L.; Li, M.; Chen, Z.; Zhang, C.; Wang, P.; Li, F. ScDrugAct: A Comprehensive Database to Dissect Tumor Microenvironment Cell Heterogeneity Contributing to Drug Action and Resistance across Human Cancers. *Nucleic Acids Res.* **2025**, *53* (D1), D1536–D1546. https://doi.org/10.1093/nar/gkae994.

[27] Gambardella, G.; Viscido, G.; Tumaini, B.; Isacchi, A.; Bosotti, R.; Di Bernardo, D. A Single-Cell Analysis of Breast Cancer Cell Lines to Study Tumour Heterogeneity and Drug Response. *Nat. Commun.* **2022**, *13* (1), 1714. https://doi.org/10.1038/s41467-022-29358-6.

[28] Ali, S.; Abuhmed, T.; El-Sappagh, S.; Muhammad, K.; Alonso-Moral, J. M.; Confalonieri, R.; Guidotti, R.; Del Ser, J.; Díaz-Rodríguez, N.; Herrera, F. Explainable Artificial Intelligence (XAI): What We Know and What Is Left to Attain Trustworthy Artificial Intelligence. *Inf. Fusion* **2023**, *99*, 101805. https://doi.org/10.1016/j.inffus.2023.101805.

[29] Schissler, A. G.; Aberasturi, D.; Kenost, C.; Lussier, Y. A. A Single-Subject Method to Detect Pathways Enriched With Alternatively Spliced Genes. *Front. Genet.* **2019**, *10*, 414. https://doi.org/10.3389/fgene.2019.00414.

[30] Chen, L.; Chu, C.; Lu, J.; Kong, X.; Huang, T.; Cai, Y.-D. Gene Ontology and KEGG Pathway Enrichment Analysis of a Drug Target-Based Classification System. *PLOS ONE* **2015**, *10* (5), e0126492. https://doi.org/10.1371/journal.pone.0126492.

[31] Ponce-Bobadilla, A. V.; Schmitt, V.; Maier, C. S.; Mensing, S.; Stodtmann, S. Practical Guide to SHAP Analysis: Explaining Supervised Machine Learning Model Predictions in Drug Development. *Clin. Transl. Sci.* **2024**, 17 (11), e70056. https://doi.org/10.1111/cts.70056.

[32] Opdam, F. L.; Guchelaar, H.-J.; Beijnen, J. H.; Schellens, J. H. M. Lapatinib for Advanced or Metastatic Breast Cancer. *The Oncologist* **2012**, *17* (4), 536–542. https://doi.org/10.1634/theoncologist.2011-0461.

[33] Ding, S.; Chen, X.; Shen, K. Single-cell RNA Sequencing in Breast Cancer: Understanding Tumor Heterogeneity and Paving Roads to Individualized Therapy. *Cancer Commun.* **2020**, *40* (8), 329–344. https://doi.org/10.1002/cac2.12078.

[34] Dong, S.; Yousefi, H.; Savage, I. V.; Okpechi, S. C.; Wright, M. K.; Matossian, M. D.; Collins-Burow, B. M.; Burow, M. E.; Alahari, S. K. Ceritinib Is a Novel Triple Negative Breast Cancer Therapeutic Agent. *Mol. Cancer* **2022**, *21* (1), 138. https://doi.org/10.1186/s12943-022-01601-0.

[35] Schäkel, L.; Mirza, S.; Winzer, R.; Lopez, V.; Idris, R.; Al-Hroub, H.; Pelletier, J.; Sévigny, J.; Tolosa, E.; Müller, C. E. Protein Kinase Inhibitor Ceritinib Blocks Ectonucleotidase CD39 – a Promising Target for Cancer Immunotherapy. *J. Immunother. Cancer* **2022**, *10* (8), e004660. https://doi.org/10.1136/jitc-2022-004660.

[36] Petrazzuolo, A.; Perez-Lanzon, M.; Liu, P.; Maiuri, M. C.; Kroemer, G. Crizotinib and Ceritinib Trigger Immunogenic Cell Death via On-Target Effects. *OncoImmunology* **2021**, *10* (1), 1973197. https://doi.org/10.1080/2162402X.2021.1973197.

[37] Qiu, S.; Chen, G.; Peng, J.; Liu, J.; Chen, J.; Wang, J.; Li, L.; Yang, K. LncRNA *EGOT* Decreases Breast Cancer Cell Viability and Migration via Inactivation of the Hedgehog Pathway. *FEBS Open Bio* **2020**, *10* (5), 817–826. https://doi.org/10.1002/2211-5463.12833.

[38] Comșa, Ș.; Cîmpean, A. M.; Raica, M. The Story of MCF-7 Breast Cancer Cell Line: 40 Years of Experience in Research. *Anticancer Res.* **2015**, *35* (6), 3147–3154.