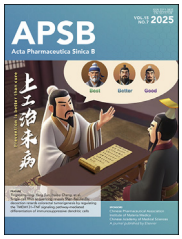
Acta Pharmaceutica Sinica B 2025;15(7):3355−3371

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

Acta Pharmaceutica Sinica B

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



REVIEW

Artificial intelligence and anti-cancer drugs' response

Xinrui Long[a,b,c,d,e,f,†], Kai Sun[a,c,d,e,f,†], Sicen Lai[a,b,c,d,e,f], Yuancheng Liu[a,c,d,e,f], Juan Su[a,c,d,e,f], Wangqing Chen[a,c,d,e,f], Ruhan Liu[a,c,d,e,f], Xiaoyu He[a,c,d,e,f], Shuang Zhao[a,c,d,e,f,*], Kai Huang[a,c,d,e,f,*]

[a]*Department of Dermatology, Xiangya Hospital, Central South University, Changsha 410028, China*
[b]*Xiangya School of Medicine, Central South University, Changsha 410031, China*
[c]*Furong Laboratory, Central South University, Changsha 410028, China*
[d]*National Engineering Research Center of Personalized Diagnostic and Therapeutic Technology, Changsha 410028, China*
[e]*Hunan Key Laboratory of Skin Cancer and Psoriasis, Xiangya Hospital, Central South University, Changsha 410028, China*
[f]*National Clinical Research Center for Geriatric Disorders, Xiangya Hospital, Central South University, Changsha 410028, China*

Received 29 January 2025; received in revised form 23 April 2025; accepted 2 May 2025

KEY WORDS

Drug resistance;
Anti-cancer drugs;
Artificial intelligence;
Multi-omics;
Precision medication

Abstract Drug resistance is one of the key factors affecting the effectiveness of cancer treatment methods, including chemotherapy, radiotherapy, and immunotherapy. Its occurrence is related to factors such as mRNA expression and methylation within cancer cells. If drug resistance in patients can be accurately identified early, doctors can devise more effective treatment plans, which is of great significance for improving patients' survival rates and quality of life. Cancer drug resistance prediction based on artificial intelligence (AI) technology has emerged as a current research hotspot, demonstrating promising application prospects in guiding clinical individualized and precise medication for cancer patients. This review aims to comprehensively summarize the research progress in utilizing AI algorithms to analyze multi-omics data including genomics, transcriptomics, epigenomics, proteomics, metabolomics, radiomics, and histopathology, for predicting cancer drug resistance. It provides a detailed exposition of the

*Corresponding authors.
E-mail addresses: kaiserhuang@csu.edu.cn (Kai Huang), shuangxy@csu.edu.cn (Shuang Zhao).
[†]These authors made equal contributions to this work.
Peer review under the responsibility of Chinese Pharmaceutical Association and Institute of Materia Medica, Chinese Academy of Medical Sciences.

https://doi.org/10.1016/j.apsb.2025.05.009

processes involved in data processing and model construction, examines the current challenges faced in this field and future development directions, with the aim of better advancing the progress of precision medicine.

1. Introduction

The occurrence of cancer has become one of the main obstacles to human life extension, which is second most common disease in non-communicable diseases[1,2]. The main hallmarks of cancer comprise activating invasion and metastasis, reprogramming cellular metabolism, and avoiding immune destruction[3]. Hence, cancer cells are likely to proliferate constantly and escape the immunologic surveillance, transferring from primary organ to metastatic organ[4]. Although the application of chemotherapy, targeted drugs and immune checkpoint inhibitors has significantly improved the rate of inhibiting tumor growth, they all face a common problem called drug resistance[5]. For example, in lung cancer, the EGFR T790M mutation is the most common mechanism of resistance to EGFR tyrosine kinase inhibitors[6]. The chronic myeloid leukemia patients treated by imatinib were eventually observed allosteric T315I mutation which leads to bad prognosis[7].

Drug resistance can be classified into adaptive drug resistance and acquired drug resistance. Some genetic mutations in cancer cells cause adaptive drug resistance[8]. Acquired drug resistance means that with the continuous progress of treatment, the good therapeutic efficacy continues to reduce, which may be related to the proliferation of pre-existing drug-resistant subclones after treatment[9]. Hence, both are significantly related to the composition of tumors and mutation of some genes in cancer cells, resulting in the demand for assessing drug sensitivity individually. For example, ARAF mutations lead to the resistance of RAF inhibitor belvarafenib in melanoma patients[10]. Activation of colorectal cancer ferroptosis by targeting LGR4 can overcome acquired resistance[11].

The application of artificial intelligence (AI) provided the support for omics data mining. A subclass of AI, known as machine learning, mainly refer to a kind of mathematical algorithms that can identify the features of a large amount of data through learning. It is widely used in the prediction of drug resistance. Nowadays, AI has been applied in many fields of biomedicine, which has shown great accuracy in identifying cancer subtypes, exploring the biomarkers of diseases and distinguishing diverse immunohistochemical scores[12,13]. Deep learning is a subset of machine learning with multiple sequential layers, which is especially suitable for training large datasets. Additionally, the application of AI is also used to identify the population, which is likely to show great sensitivity to the specific treatment[14]. For example, Johannet et al.[15] established a machine learning algorithm based on histology specimens with clinical data to predict the response of immune checkpoint inhibitors in advanced melanoma.

In this review, we primarily analyze and discuss the latest advancements in predicting cancer drug resistance based on AI algorithms and multi-omics data, encompassing various omics data types such as genomics, transcriptomics, epigenomics, proteomics, metabolomics, radiomics, and pathomics. In this section, we discussed the introduction and related research work. In Section 2, we discuss the application of AI in multi-omics data; in Section 3, we analyze the modeling process of AI; in Section 4, we delve into the cutting-edge research of AI in drug response prediction; and in Section 5, we analyze the existing obstacles and solutions.

2. The application of AI in multi-omics data

Due to differences in data sources and collection principles, multi-omics data exhibit diverse representation forms within AI models[16]. The rapid advancement of high-throughput sequencing technology has enabled the simultaneous synthesis and sequencing of countless short DNA templates ranging from hundreds to billions, significantly reducing both the time and cost required for sequencing. This remarkable progress has given rise to a vast amount of omics data, including genomics, transcriptomics, and epigenomics data[17]. Specifically, genomics revolves around the in-depth study of an organism's complete genome sequence, with the goal of understanding various changes, including gene mutations and copy number variations. It is primarily presented in the form of numerical matrices or sequences, encompassing variation information such as single nucleotide polymorphisms (SNP), insertions, or deletions, as well as information on gene expression levels. However, transcriptomics primarily focuses on the expression levels of genes under different conditions, and it also uses formatted sequence data for AI model analysis to interpret information such as differentially expressed genes and transcription factor binding sites. Simultaneously, epigenomics specializes in exploring diverse epigenetic modifications present in the genome, such as DNA methylation and histone modifications[18], which mostly exist in the form of numerical matrices or sequences. Furthermore, proteomics concerns itself with the types, quantities, and interactions of proteins, and its data representation is quite diverse, encompassing peptide mass fingerprinting, protein expression profiles (typically presented in the form of numerical matrices), and protein interaction networks (usually depicted in graphical or network diagrams). Furthermore, proteomics and metabolomics primarily obtain their respective data through mass spectrometry, a method used to determine molecular mass and structure. Proteomics focuses on the types, quantities, and interactions of proteins, with diverse data representation forms such as peptide mass fingerprinting, protein expression profiles (typically presented in the form of numerical matrices), and protein interaction networks (usually depicted in graphical or network diagrams). On the other hand, metabolomics is concerned with the types, concentrations, and changes of metabolites within organisms. In terms of data representation, common spectra include base peak chromatograms and total ion current chromatograms. Additionally, the multi-omics data

discussed in this review also encompasses radiomics and histopathomics, which involve extracting and mining features from medical images, primarily existing in the form of images but with varying dimensions and sizes. The sources and processing workflows of the aforementioned omics data are illustrated in Fig. 1.

Multi-omics data plays a crucial role in advancing the development of personalized and precision medicine. Currently, there are numerous publicly available multi-omics database resources and this review has compiled information on some representative public omics database resources in Table 1. Furthermore, the rapid development of AI and advancements in high-performance computing equipment have provided support for the analysis of exponentially growing multi-omics data. Currently, numerous studies are exploring the feasibility of utilizing AI technology for drug resistance prediction based on multi-omics data. This section will delve into the application of AI in multi-omics data, and summarize some representative studies based on cell line omics data and clinical omics data in Table 2[19-51] and Table 3[15,52-68], respectively.

Different types of omics data provide information across distinct dimensions. For instance, genomics data can reveal the overall architecture and composition of the genome, offering profound insights into life's microscopic mechanisms. In contrast, radiomics or pathomics extract quantitative feature information from medical images, primarily capturing organismal information from macroscopic perspectives. The strategic combination of diverse omics data streams can significantly enhance AI model performance. However, this does not imply that increasing the number of omics data types automatically improves outcomes. In other words, the decision to use single-type or multi-type omics data—and the specific selection of omics modalities—should be guided by the research objectives. When studies focus on specific biological processes or molecular layers, single-type omics data suffice. For example, if the goal is to understand the impact of gene expression levels on disease pathogenesis, and high-quality transcriptomics data are available, models built solely on transcriptomics can effectively analyze the relationship between expression variations and disease phenotypes.

However, if the objective of the task is to study complex diseases from a systems biology perspective, or to explore unknown biological mechanisms and discover new biomarkers, multi-omics integration becomes advantageous. For instance, in cardiovascular research, single omics layers often provide partial information, whereas integrated analysis can uncover complex interactions between genetic variations, expression patterns, proteomic profiles, and metabolic shifts. This holistic approach enables more accurate disease risk prediction, subtype classification, and therapeutic guidance. Similarly, when investigating aging-related biomarkers, combined transcriptomics, proteomics, and metabolomics analyses may reveal cross-layer molecular signatures—potential key regulators or novel biomarkers in aging processes.

Effective integration of heterogeneous omics data requires sophisticated fusion strategies to mitigate cross-modal discrepancies. Common multimodal fusion approaches include: Domain Alignment Techniques, which employ learned mappings to project diverse data modalities into a unified feature space; Ensemble Learning Methods, where independent models process each modality before integrating predictions to enhance overall performance; Multimodal Encoding Architectures, featuring modality-specific branches for feature extraction followed by fusion operations like concatenation, summation, or multiplication. These

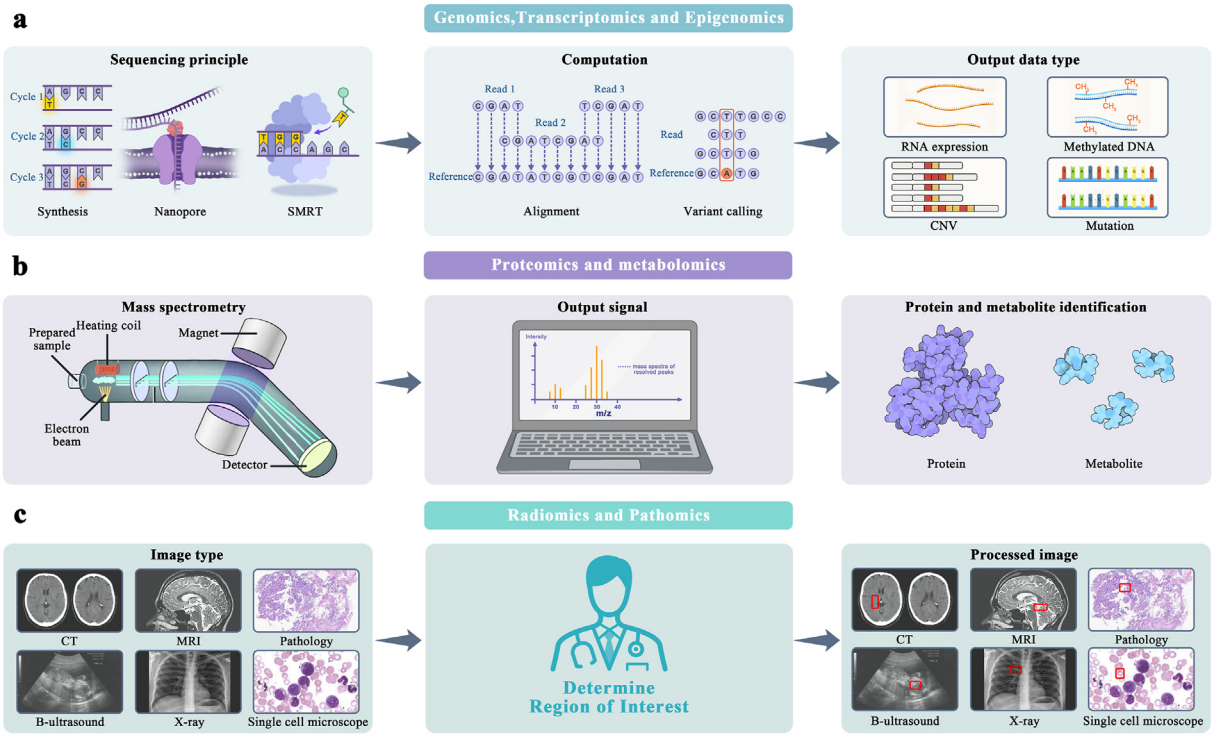


Figure 1 Omics data sources and processing procedures. (a) The genomic, transcriptomic, and epigenomic data obtained from sequencing technologies. (b) The proteomics and metabolomics data obtained from mass spectrometry technology. (c) The processing procedures of radiomics and pathomics data.

Table 1 Public omics database resources.

Database name	Data type	Description	Access link
The cancer genome Atlas (TCGA)	Genomics Transcriptomics Epigenomics Proteomics Histopathomics	Comprehensive multi-omics data from various cancer types, which can be used to explore mechanisms of drug resistance.	https://www.cancer.gov/tcga
Genomics of drug sensitivity in cancer (GDSC)	Genomics Drug data	Linking genomic features to drug response in cancer cell lines, focusing on resistance and sensitivity profiles.	https://www.cancerrxgene.org
Cancer cell line Encyclopedia (CCLE)	Genomics Transcriptomics Proteomics Drug data	Providing omics profiles of cancer cell lines, integrated with drug resistance and sensitivity data.	https://portals.broadinstitute.org/ccle
DepMap	Genomics Transcriptomics Proteome Drug sensitivity	Offering functional dependency data to identify cancer vulnerabilities, including resistance mechanisms.	https://depmap.org
PharmacoDB	Genomics Drug sensitivity	Aggregating multiple drug screening studies (e.g., GDSC, CCLE) into a single platform for drug resistance research.	https://pharmacodb.ca
OncoKB	Genomics Drug data	Curating database linking cancer mutations to drug resistance and potential therapeutic options.	https://www.oncokb.org
NCBI GEO (gene expression Omnibus)	Transcriptomics	Repository for gene expression studies, including data sets relevant to drug resistance mechanisms.	https://www.ncbi.nlm.nih.gov/geo
Metabolomics Workbench	Metabolomics	Providing metabolomics data relevant to cancer and drug resistance, including pathway insights.	https://www.metabolomicsworkbench.org
ProteomicsDB	Proteomics	Offering proteomics data that can be used to explore mechanisms of drug resistance in cancer.	https://www.proteomicsdb.org
cBioPortal	Multi-omics	Interactive platform for exploring multi-omics and clinical data related to drug resistance.	https://www.cbioportal.org

methodologies effectively address data heterogeneity and improve model performance. However, it is important to acknowledge that they inevitably increase computational complexity. Therefore, when selecting data modalities, a critical consideration lies in balancing data utility against computational demands—a trade-off that necessitates careful evaluation.

2.1. Applications in genomics, transcriptomics and epigenomics

The application of high-throughput sequencing technology has made it easier for researchers to obtain genomic, transcriptomic, and epigenomic data. These three are closely related to molecular medicine. Transcriptomics is most commonly used in the construction of drug resistance prediction models[69], often serving as standalone input data and frequently analyzed in conjunction with genomics or epigenomics. However, the use of genomics or epigenomics alone is less common.

Most studies that build predictive models based on cell line data primarily source their data from public databases, such as GDSC and CCLE. Researchers commonly opt for transcriptomic data as the foundational dataset. In high throughput gene expression data, the dimension of features d is much larger than the number of samples n, which makes it difficult to construct an optimal classifier. To address this issue, various studies employed different approaches in data processing. Mourragui et al.[70] proposed a transfer learning algorithm named TRANSACT, capable of calculating biologically active pathways that are similar between cell lines and human tumors. Transcriptomic data was downloaded from GDSC and TCGA, but subjected to TRANSACT's nonlinear principal component analysis (PCA) processing and geometric comparisons. Ultimately, they built an Elastic Net (EN) prediction model based on the generated consensus feature score. This model demonstrated promising predictive results on clinical datasets, addressing the practical limitations of cell line data predictive models. Kim et al.[19] also conducted research on the model construction which is mainly based on analyzing active biological pathway after inputting transcriptomics data. The pathway dysregulation scores for each individual sample point were calculated using Pathifier algorithm which is designed to quantify the degree of pathway abnormality. Similarly, Liu et al.[20] utilized transcriptomic data and cancer signaling pathway information to establish a disease-specific driver network. Subsequently, they developed the NBSBM model based on this network. Some studies reduced the dimensionality of input data by standardizing the input criteria. Parca et al.[21] reduced input data by selecting genes that exhibited correlation with both gene expression levels and IC_{50}. Subsequently, they built models based on the expression levels of these selected genes. Similarly, Wiesweg et al.[52] restricted the gene expression data for model construction to immune-related datasets.

Since the purpose of anticancer drugs is to interfere with the dynamics and adaptive mechanisms underlying disease pathology, changes in gene expression after drug perturbation will partially reflect the potential mechanisms leading to drug response[71]. Therefore, some studies collected gene expression data post drug perturbation. For instance, Mannheimer et al.[34] compared the predictive performance of models constructed based on the disturbed gene expression data of 2, 6 and 24 h, and found that perturbed gene expression at 24 h is the best predictor of drug response. Tan et al.[23] used sensitivity signatures and drug activity

Table 2 The biological information overview of researches utilizing cell line data for the construction of AI Models.

Study	Input data source	Omics type	Input data
Kim et al. [19]	GSE; CCLE; CGP	Transcriptomics	PDS based on mRNA EXP
Liu et al. [20]	Prostate cancer cell lines; GDSC	Transcriptomics	Genes EXP
Parca et al.[21]	GDSC	Transcriptomics	mRNA EXP; drug-specific genes EXP
Mannheimer et al.[22]	GSE; NCI-60	Transcriptomics	Perturbed mRNA EXP by drugs
Tan et al.[23]	GDSC; CCLE; LINCS	Transcriptomics	Cell line sensitivity signature
Ma et al.[24]	GDSC; CCLE	Transcriptomics; genomics	mRNA EXP, CNV
Tang et al.[25]	GDSC; CCLE	Transcriptomics; genomics	mRNA EXP; MUT; CNV
Lee et al.[26]	CCLE; CGP; TCGA	Transcriptomics; genomics; epigenomics	MUT; expression hubs; known regulators; CNV; MET
El-Deredy et al.[27]	Gliomas cell	Metabolomics	Metabolic profiles by proton NMR spectrum
Liu et al.[28]	CRC cells	Metabolomics	Metabolomics data
Liu et al.[29]	CML cells	Metabolomics	Metabolomics data
Ma et al.[30]	NCI-60	Proteomics	Protein EXP
Mian et al.[31]	Breast cancer cell	Proteomics	Protein EXP
Yanagisawa et al.[32]	Colorectal cancer cell	Pathomics	Cell images
Steyaert et al.[33]	GDSC; Novartis PDXE	Transcriptomics	mRNA EXP
Majumdar et al.[34]	CCLE; CTRP v2	Transcriptomics	mRNA EXP
Sharma et al.[35]	GDSC; CCLE, LINCS	Transcriptomics	mRNA EXP
Naulaerts et al.[36]	GDSC	Transcriptomics; genomics; epigenomics	SNV; mRNA EXP; MET
Choi et al.[37]	GDSC; CCLE	Transcriptomics	mRNA EXP
Bazgir et al.[38]	NCI-60; GDSC	Transcriptomics	mRNA EXP
Sakellaropoulos et al.[39]	GDSC; CCLP	Transcriptomics	mRNA EXP
Chiu et al.[40]	CCLE; GDSC; TCGA	Transcriptomics; genomics	mRNA EXP; MUT
Carroll et al.[41]	MCLP; CCLE; CTRP	Proteomics	Protein EXP
Wang et al.[42]	GDSC; CCLE; CTRP	Transcriptomics	LncRNA EXP
Tan et al.[43]	CCLE; GDSC	Transcriptomics	mRNA EXP
Dorman et al.[44]	Cell lines from [45]	Transcriptomics; genomics	mRNA EXP; MUT; CNV
Menden et al.[46]	GDSC	Genomics	CNV; MUT; MSI
Chen et al.[47]	GDSC; CCLE	Transcriptomics	RNA EXP
Lee et al.[48]	GDSC; TCGA	Genomics; epigenomics	MUT; CNV; MET
Sun et al.[49]	\	Genomics; transcriptomics; proteomics	MUT; mRNA EXP; protein EXP
Jia et al.[50]	GDSC; TCGA	Transcriptomics	mRNA EXP
Wang et al.[51]	GDSC; CCLE	Transcriptomics	mRNA EXP

CHEM: Chemical structure molecular fingerprints; CNA: Copy-number alterations; CML: Chronic myelo-genous leukemia; DAS: Drug Activity Signatures; EXP: Expression; MET: Methylation; LN: Nature log; MUT: Mutation; NCRNA: Non-coding RNA; PDS: Pathway dysregulation scores; SCMS: Single cell mass spectrometry; SMILES: Simplified molecular input line entry system; TSS: Tissue sensitivity signature; LINCS: Library of integrated network-based cellular signatures.

signatures to predict the activity of a given drug against a specific cell line. Sensitivity signature is a representation of the gene expression changes that lead to cell death in a certain cell line. Drug activity signature is a representation of the "mode of action" for a drug, summarizing the cumulative differential gene expression upon treatment on cell lines.

The combined analysis of multi-omics data allows for a more comprehensive understanding of the molecular biological features of drug resistance. Many studies integrate transcriptomics with genomics or combine them with epigenomics to achieve a more holistic approach. DualGCN consists of two graph convolutional networks (GCN) that encode the chemical structures of drugs and the omics data of biological samples for predicting cancer drug response[24]. Tang et al.[25] proposed a pathway-based drug sensitivity prediction model called PathDSP, which integrates chemical structure information with the enrichment of cancer signaling pathways across drug-associated genes, gene expression, mutations, and copy number variation data to predict drug responses in cancer drug genomics datasets. Lee et al.[26] present a computational method called MERGE, which identifies reliable gene expression markers for drug sensitivity by integrating multi-omics prior information related to each gene's potential to drive cancer. Nam et al.[53] developed a machine learning model to predict temozolomide response for primary IDH-wt glioblastoma patients, which combines transcriptomics, genomics and epigenetics data.

High-quality data serves as the foundation for developing high-performance AI models. Although high-throughput sequencing technology has undergone rapid development in recent years, it is crucial to acknowledge that variations in high-throughput sequencing technology across different medical centers and sequencing platforms often lead to heterogeneity and noise in biological data, even when derived from the same omics layer. For

Table 3 The biological information overview of researches utilizing clinical data for the construction of AI Models.

Study	Omics type	Cancer type	Drug	Input data
Johannet et al.[15]	Pathomics	Melanoma	Immune checkpoint inhibitors	WSI; clinical data
Wiesweg et al.[52]	Transcriptomics	NSCLC	PD-L1 inhibitors	Immune-related mRNA EXP
Nam et al.[53]	Transcriptomics; genomics; epigenomics	Glioblastoma	Temozolomide	mRNA EXP; SNV; methylation
Yu et al.[54]	Proteomics	Ovarian serous carcinoma	Platinum	Protein EXP
Qureshi et al.[55]	Proteomics	NSCLC	Gefitinib and erlotinib	Clinical data; geometrical characteristics of the drug-target binding site; binding free energy
Yang et al.[56]	Radiomics	LARC	NCRT	ADC images
Xiong et al.[57]	Radiomics	Breast cancer	NCRT	MRI; clinical data
Yi et al.[58]	Radiomics; genomics	Ovarian	Platinum	CT images; clinicopathological data; SNP
Buttarelli et al.[59]	Transcriptomics	Ovarian	Platinum	mRNA EXP
Abraham et al.[60]	Transcriptomics	mCRC	FOLFOX	mRNA EXP
Lu et al.[61]	Transcriptomics	CRC	FOLFOX	mRNA EXP
Dercle et al.[62]	Radiomics	CRC	FOLFIRI in combination with cetuximab	CT images
Ruffalo et al.[63]	Transcriptomics	Breast cancer	Aromatase inhibitors	mRNA EXP; clinical data
Herold et al.[64]	Transcriptomics; genomics	AML	Cytarabine- and anthracycline-based induction treatment	mRNA EXP; mutation
Gampenrieder et al.[65]	Transcriptomics; epigenomics	Metastatic breast cancer	Bevacizumab	Methylation
Guo et al.[66]	Genomics	Cervical cancer	Platinum-based NRCT	24 SNPs in four genes
Nasimian et al.[67]	Transcriptomics	AML	Sorafenib	mRNA EXP
Kanda et al.[68]	Radiomics	Pancreatic ductal adenocarcinoma	5-Fluorouracil; gemcitabine and paclitaxel	CT images

AML: acute myeloid leukemia; CRC: colorectal cancer; EXP: expression; FOLFOX: 5-FU, leucovorin and oxaliplatin; LARC: locally advanced rectal cancer; NCRT: neoadjuvant chemoradiotherapy; NSCLC: non-small cell lung carcinoma.

instance, the Illumina platform primarily relies on the Sequencing By Synthesis principle, whereas the PacBio platform employs Single Molecule Real-Time sequencing technology. These differing sequencing technologies result in discrepancies in base call quality, sequencing error rates, sequencing coverage, and sequencing depth. Therefore, standardized preprocessing before use is vital to ensure data consistency. Specifically, data cleaning can be performed to remove low-quality sequences, adapter sequences, contaminated sequences, and other imperfections from sequencing data, thereby reducing noise and errors. Furthermore, imputation techniques can be adopted to fill in missing values in sequencing data, minimizing their impact on data analysis. Additionally, data standardization or normalization can be applied to map data onto the same statistical distribution, such as the standard normal distribution. Moreover, in the process of standardizing genomic data, international standards and guidelines, such as ISO/TS 24420:2023, should be followed to further enhance data consistency.

2.2. *Applications in metabolomics and proteomics*

The fundamental principle of mass spectrometry is to ionize molecules or atomic ions in the sample to be analyzed. Subsequently, under the influence of an external electric field, these ions are separated and detected based on their mass-to-charge ratio (m/z). This technique plays a crucial role in acquiring data for proteomics and metabolomics.

To the best of our knowledge, El-Deredy et al.[27] were the first to study metabolomics using proton nuclear magnetic resonance spectroscopy. They subsequently applied this approach to predict cellular responses to the chemotherapeutic drug 1-(2-chloroethyl)-3-cyclohexyl-1-nitrosourea *in vitro*. Liu et al.[28] employed single-probe SCMS technology to measure the metabolomic profiles of live cancer cells (HCT-116) with varying levels of chemotherapy-induced drug resistance (*i.e.*, none, low, and high). Based on these data, they constructed a series of machine learning models, including Random Forest (RF), Artificial Neural Network (ANN), and Penalized Logistic Regression (LR), to predict the degree of drug resistance in individual cells. Furthermore, Liu et al.[29] have demonstrated for the first time that the combination of single-cell mass spectrometry and machine learning methods can achieve rapid and reliable prediction of unknown single-cell phenotypes based on their metabolomic profiles. This approach holds promise for clinical application in predicting drug-resistant phenotypes prior to chemotherapy.

Proteomic analysis allows for the quantification of protein expression levels, thereby providing a more direct answer to functional and pharmacological questions. Ma et al.[30] collected protein expression level data from the NCI-60 and used reverse-phase protein lysate microarrays to perform proteomic detection. Protein samples were automatically spotted onto the chips and then measured with antibodies. They selected important protein predictors by RF and Relief and 118 anticancer agents were chosen by using protein expression profiles in the 60 untreated cell lines. Mian et al.[31] conducted their study based on data from breast cancer cells they cultured themselves. The analysis technique employed was surface-enhanced laser desorption/ionization

mass spectrometry, which allowed the combination and retention of molecular subgroups for proteomic analysis. This approach was integrated with an ANN algorithm to ultimately identify proteomic patterns associated with either control/drug treatment for sensitive cell lines or response/nonresponse for sensitive/resistant cell lines, respectively, for two widely used chemotherapeutic drugs, doxorubicin and paclitaxel.

Furthermore, in terms of drug resistance prediction at the patient level, Yu et al.[54] analyzed the proteomic characteristics of 130 patients with ovarian serous carcinoma. They digested the proteins using trypsin and separated the peptides by basic reversed-phase liquid chromatography. Subsequently, they employed high-resolution and high-accuracy Orbitrap mass spectrometry for proteomic and phosphoproteomic analyses. Finally, protein quantification was conducted based on ion intensity. Based on these proteomic expression data, they constructed various machine learning models, including the Lasso Cox model, RF, Support Vector Machine (SVM), Bagging, and Naive Bayes, to predict patients' responses to platinum-based drugs. Experimental results confirmed that the proteomic characteristics of patients with ovarian serous carcinoma can not only predict responses to platinum-based drugs but also provide insights into the biological processes that affect the therapeutic efficacy of these drugs.

Qureshi et al.[55] aimed to predict the response of patients with EGFR-mutant lung cancer to EGFR-TKIs. The dataset used in their research was unique and included features such as demographic and clinical information, geometric properties of drug-target binding sites, and binding free energy from molecular dynamics (MD) simulations of drug-target complexes. MD simulations were employed to model the unique mutation status of each patient, extracting molecular-level geometric features. These geometric features encompassed the number of convex atoms on the interaction surface, the matching rate of surface atoms, dynamic distances between the center of drug molecules and binding site residues throughout the entire trajectory, and the number of hydrogen bonds, which differs from others that construct data at the quantitative level of proteins. Instead, it focuses on the molecular structure, dynamic properties, and interactions of proteins with other molecules. This approach emphasizes a more microscopic perspective to predict the affinity of proteins for drugs.

In summary, significant progress has been made in using AI to analyze proteomic and metabolomic data for predicting drug resistance. These studies have contributed to a deeper understanding of the complex molecular landscape governing resistance mechanisms. With the continuous advancement of mass spectrometry technology and the sharing of sequencing data, this field is poised to better demonstrate its clinical application advantages.

2.3. *Applications in radiomics and pathomics*

Image analysis is also a significant application field of AI, with image analysis algorithms represented by Convolutional Neural Networks (CNNs) being widely used in the medical domain. Unlike genomics and transcriptomics, which require sequencing, the acquisition of medical images is more straightforward. Common medical images typically include macroscopic images such as CT, MRI, and ultrasound, as well as microscopic images like pathological sections or single-cell images observed under a microscope. These medical images have also been extensively applied to tasks related to drug resistance research. For instance, Yang et al.[56] and Xiong et al.[57] both utilized MRI data to predict the response of cancer patients to neoadjuvant chemoradiotherapy. The former focused on the drug sensitivity of patients with locally advanced rectal cancer, while the latter investigated the drug sensitivity of breast cancer patients. The process of radiomic feature extraction in both studies was quite similar, involving image annotation by radiologists unaware of the pathological results to determine the Region of Interest. However, they adopted different approaches in feature selection. Specifically, Yang et al.[56] performed the Wilcoxon rank sum test and then used Least Absolute Shrinkage and Selection Operator (Lasso) to select the most predictive features. Xiong et al.[57] applied RF to measure the Gini importance of each parameter and identified the top 10 important parameters as biomarkers for predicting treatment resistance, which were used for model construction. Meanwhile, to facilitate individualized assessment, Xiong et al.[57] integrated the Ki67 index and HER2 status with radiomic features to construct a combined prediction model.

Johannet et al.[15] developed two DCNN classifiers based on pathological slides of melanoma: the Segmentation Classifier and the Response Classifier. The Segmentation Classifier identifies tumor, lymphocyte, and stromal tissue regions, eliminating the need for manual annotation of the Regions of Interest by physicians, thereby significantly saving time and reducing labor costs. The Response Classifier predicts the responsiveness of melanoma patients to immune checkpoint inhibitors based on pathological images. This model overcomes limitations associated with time and spatial heterogeneity that impede PD-L1 performance, demonstrating consistent results across different scanner devices analyzing Whole Slide Images. This has crucial implications for clinical practice. Yanagisawa et al.[32] employed cancer cell images from two colorectal cancer cell lines to construct a CNN model, predicting the drug resistance of colorectal cancer cells to the antitumor drugs 5-fluorouracil and trifluridine. The model can be applied to identify circulating tumor cells (CTCs), streamlining the CTC sequence analysis process, which can facilitate a more convenient determination of drug resistance in patients and present promising clinical applications.

In conclusion, radiomics and pathomics data collection is often more convenient than other omics, such as transcriptomics and proteomics. It holds greater clinical value than predictive models built on resistance data at the cellular level. However, in comparison, efficiently extracting image features can be more challenging.

3. The modeling processes of AI

The scientists believed that computers could be used to accomplish tasks such as decision-making, language interpretation and visual perception, thus proposing the concept of AI[72]. Samuel proposed the concept of machine learning and described it as a method of programming to enable computers to learn rules from experience[73]. Machine learning is an important branch of AI, primarily encompassing traditional machine learning algorithms and deep learning algorithms represented by neural networks. Both of these types of technologies have been extensively studied in the field of drug resistance. Therefore, the AI discussed in this paper mainly focuses on the application of these two types of technologies.

This section focuses on the modeling process of AI algorithms in the paper, including Data collection, Algorithm selection, Data preprocessing, Model training, and Performance evaluation, as shown in Fig. 2.

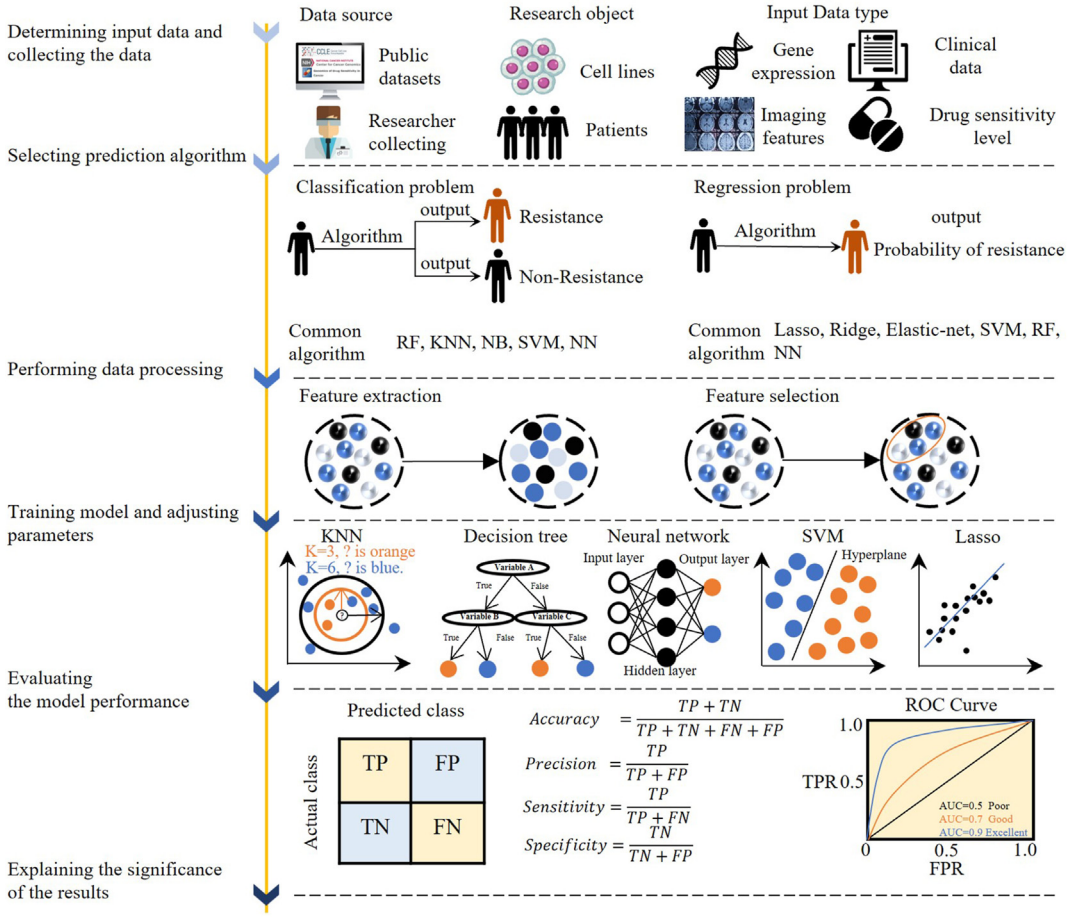


Figure 2 The main construction process of drug resistance predicted algorithm.

3.1. *Data collection*

The research on predicting drug sensitivity using AI algorithms can be divided into two categories: using AI algorithms to predict drug resistance in patients and using AI algorithms to predict drug resistance of cell lines to one or multiple drugs.

In the first condition, the data are usually from patients with the same cancer who use the same drug. The most common type of input data is mRNA expression data[52,59,60,74]. In addition to mRNA expression data, the common input data include methylation level[65], gene mutations[64], protein expression[54] and the features of CT or MRI[27,56,57]. To increase the accuracy of model, the clinical data such as Tumor Node Metastasis stage, age and so on were also often input as predictive factors combined with other omics data.

In the second condition, the data for constructing algorithms mainly is from the public database, such as Genomics of Drug Sensitivity in Cancer (GDSC)[75], Cancer Cell Line Encyclopedia (CCLE)[76], Cancer Therapeutics Response Portal (CTRP)[77] and NCI-60[78]. The screening of NCI-60 anticancer drugs identifies compounds with growth inhibitory or toxic effects on specific tumor types, including cell line genomic, transcriptomic and epigenomic, proteomics and metabonomics data. In addition to NCI-60, proteomics and metabolomics data are also available in the CCLE. CCLE and GDSC contain more abundant types of cell lines than NCI-60, which is more convenient for finding anticancer drug sensitivity biomarkers. However, the data included in CCLE and GDSC are obtained before the treatment. Comparing with CCLE and GDSC, Connectivity map[79] and the Library of Integrated Network-Based Cellular Signatures[80] were constructed by perturbation-driven gene expression, which can find compounds with similar perturbation-driven gene expression and is beneficial to drug repurposing and the exploration of drug targets. The data from xenografts or patients are also used to verify or improve the algorithm in some of research.

3.2. *Algorithm selection*

Selecting an appropriate algorithm is crucial for achieving high robustness and accuracy performance. The choice of algorithm primarily depends on factors such as the type of data being processed, the size of the dataset, and the type of task. The existing mainstream algorithms are mainly divided into two categories: traditional machine learning algorithms and deep learning algorithms represented by artificial neural networks. Traditional machine learning algorithms include logistic regression, linear regression, decision tree (DT), K-nearest neighbors (KNN), and SVM, among others. These algorithms are capable of handling various data types such as images, sequences, or multimodal data, and they have significant advantages, particularly in the analysis of genomic and transcriptomic sequencing data. They have been widely applied in tasks related to predicting anticancer drug responses. For example, Sinha et al.[81] developed a precision oncology prediction tool named “PERCEPTION” based on EN

regression model, which can accurately capture the development of drug resistance in lung cancer patients undergoing tyrosine kinase inhibitor therapy. Traditional machine learning algorithms are characterized by strong interpretability, fast analysis speeds, and low computational resource requirements. However, they have the drawback of relying on expert knowledge for manual feature selection, which is a relatively complex process. Additionally, their performance and generalization capabilities are limited.

Deep learning methods, represented by artificial neural networks, such as CNNs, recurrent neural networks (RNNs), and transformers, have also demonstrated significant potential for application in cancer drug resistance prediction tasks. Specifically, CNN is a classic visual processing model that can autonomously learn and extract both local and global features from images, which are then utilized for various tasks such as classification and prediction. RNN is a neural network architecture specifically tailored for processing and predicting sequential data, rendering it highly suitable for applications involving sequence data in fields like genomics and transcriptomics. Transformer is currently a popular neural network architecture based on the attention mechanism. It has demonstrated superior performance compared to architectures such as CNN and RNN in both sequence data and image data domains, making it a mainstream algorithm for analyzing multi-omics data. Compared to traditional machine learning algorithms, deep learning boasts advantages such as automatic feature extraction, end-to-end learning, and strong generalization capabilities. However, it also has drawbacks including high data requirements, substantial computational resource consumption, and poor interpretability.

In existing research, it is common to use a combination of multiple algorithms, which can be CNN combined with RNN or Transformer, or Transformer combined with K-nearest neighbors, etc. This approach aims to enhance model performance by integrating the advantages of different algorithms. For instance, Karampuri et al.[82] developed a drug sensitivity and resistance prediction model named ResisenseNet by integrating CNN and RNN, which can effectively learn long-range and temporal patterns from amino acid sequences and transcription factor data, achieving a prediction accuracy of 0.9794.

3.3. *Data preprocessing*

Dimension reduction is often needed before model training, which can avoid the huge amount of calculation caused by dimension disaster. In fact, with the increase of the number of variables in fitting model, the measurement error tends to increase without increasing the accuracy of the model. and when the data dimension is too high relative to the data sample size, over-fitting is more likely to occur[83,84]. There are two ways to reduce dimensionality called feature extraction and feature selection.

After determining the input features, data cleaning should also be performed before training. In reality, training data often contains missing values, duplicates, and errors, so sorting and correcting the data is an essential part of the pre-training process. The handling of missing data can significantly affect the accuracy of the model. The simplest approach is to delete rows or columns with a large number of missing values, but this can reduce the number of samples or features, thereby impacting model fitting. There are three types of missing data: completely random missing, random deletion, and non-random deletion. When the third type of missing data occurs, it may introduce significant bias into the model. Therefore, when the data loss is less than 10%, we can adopt the following methods for processing: one method is to directly use the mean, mode, or median to fill in the missing data. Another method is to use algorithms such as KNN, RF, or others to estimate the missing values, allowing for the calculation of the range or actual distribution of the specific features. The advantage of this method is that it generates multiple estimates for the missing values. However, this uncertainty is also incorporated into the model building process[85].

3.4. *Model training*

The training of AI models refers to the process of using a data-driven approach to adjust the model's parameters (such as weights and biases in neural networks) to fit the distribution of the data, with the goal of minimizing prediction errors or losses, thereby enabling the model to make accurate judgments on new data. Specifically, the parameters of AI models, represented by neural networks, primarily consist of two components: hyperparameters and the model's weight parameters and bias parameters. During the training process, the model's weight parameters and bias parameters are adjusted by calculating the error between the predicted output of the model and the actual corresponding data results, in order to improve the mode's ability to fit the data. Furthermore, hyperparameters refer to factors such as the learning rate, batch size, number of training epochs, depth and width of the network, loss function, optimizer, etc., which are preset manually before the model training and are used to control the training process of the model. The specific settings of hyperparameters depend on the type and characteristics of the data and tasks. For instance, the choice of batch size requires comprehensive consideration of hardware resources and training efficiency. A larger batch size can improve training efficiency but may result in increased memory usage. Conversely, a smaller batch size allows the model to adapt more flexibly to the data distribution, albeit at the cost of potentially slower training speeds. The depth and width of the model determine its complexity and expressive ability. More complex models may possess stronger expressive capabilities but are prone to overfitting, whereas simpler models may suffer from underfitting. In practical applications, it is often necessary to conduct experiments with multiple sets of different hyperparameter values and observe the changes in model performance in order to find the optimal combination of hyperparameters.

3.5. *Performance evaluation*

The evaluation indexes are different in the classification task and the regression task. For the classification task, the common evaluation indicators are confusion matrix and ROC curve. The confusion matrix includes four parts: True positive, True negative, False negative, False positive. On this basis, it is extended some other evaluated index like Accuracy, Precision, Recall, Specificity, and F1 score.

Although Accuracy can reflect the performance of the model. However, if the sample distribution is extremely uneven, the performance of the model cannot be reflected by Accuracy alone. For example, when 90% of the cell lines are drug-resistant, if the model classifies all the cell lines as drug-resistant cell lines, the Accuracy reaches 90%, but the actual model classification ability is not good. The use of Precision can solve the problem, which can better reflect the classification effect of the model and identify the proportion of all the positive results. Moreover, there is a balance

between Precision and Recall. Improving Precision often requires the model to make positive predictions more cautiously, ensuring that most of the samples predicted as positive are correct. However, this may also lead to missing some actual positives, thereby reducing the recall rate. Similarly, increasing Recall rates also reduces Precision. The F1 score can be seen as a weighted average of the model's precision and recall, which simultaneously considers both the precision and recall of the classification model.

ROC curve is another commonly used performance evaluation method when evaluating classification problems. The ROC curve can reflect the prediction ability of the model, but there will be some limitations in the fitting of unbalanced data sets. Precision-Recall does not have the defect, which can reflect the relationship between Precision and Recall under different thresholds. It is convenient for researchers to measure the classification ability of the classifier in unevenly distributed data.

When evaluating the model of regression problems, we often use Mean Square Error (MSE), Mean Absolute Error (MAE), Root Mean Square Error (RMSE), Normalized Root Mean Square Error (NRMSE), Coefficient of determination (R^2), *Pearson* correlation coefficient (r) and Spearman's rank correlation coefficient (R_s).

4. AI for anti-cancer drug response prediction

AI algorithms can analyze large-scale biological data to assist researchers in identifying potential drug resistance biomarkers or directly predicting the responsiveness to anticancer drugs. For clinical doctors, AI algorithms can provide guidance for medication, improving the success rate of treatments. For pharmaceutical companies, traditional drug resistance testing often consumes a significant amount of time and resources, while AI algorithms can make predictions in a shorter time, reducing costs and waiting times. Additionally, AI can help researchers identify potential drug resistance biomarkers, guiding the development of new drugs. For patients, AI can assist doctors in accurately predicting a patient's resistance to different drugs based on the analysis of their gene expression data and other biological information, enabling personalized treatment plans. Through literature retrieval and screening out based on topic, we found that existing documents can be mainly categorized into two types. The first type directly outputs drug resistance results through AI algorithms. The second type uses AI algorithms to predict drug resistance by predicting resistance biomarkers such as proteins and mRNA indirectly.

4.1. AI for drug resistance prediction

Numerous studies have been conducted to predict responses to anti-cancer drugs based on either cell line or clinical data. AI models based on cell line data typically utilize diverse cell lines and drug information sourced from public databases. In contrast, AI models based on clinical data often rely on data from patients with a single type of cancer and information pertaining to a single drug. Tables 4 and 5 provide an overview of some studies based on cell line and clinical data, respectively.

4.1.1. AI models constructed based on cell line data

Most researchers use ensemble learning to complete the model construction, which are combined with several different machine learning algorithms. It can improve the prediction performance and improve the classification accuracy.

Majumdar et al.[34] developed a novelty k-means ensemble Support Vector Regression (SVR), which is a blend of supervised and unsupervised learning methods. The prediction of drug response was realized by the ensemble of clustering (PCA and k-means clustering) and local regression. The study split the low-dimensional data to k clusters and trained SVR models on each of the different clusters respectively. Finally, the best output is selected based on a scoring system related to the similarity of gene expression profiles. The k-means ensemble SVR had the lowest MSE, showing better performance than other common algorithms, including LR, Back Propagation Neural network, SVR and Quantile Regression Forest.

The ensemble learning algorithm developed by Sharma and Rani is an optimized algorithm based on rotating forest[35]. In the proposed technique, researchers obtained feature subsets from original data and all feature subsets are transformed linearly using PCA. The transformed feature subsets are input to the basic learner for training. The basic learners are DT, NN and extreme machine learning, which leads to better generalization and performance than that only including DT. Compared with Kernelized Bayesian Multi-task Learning KBMTL, EN, Scalable-Time Ridge Estimator by Averaging of Models, NN and Vanilla RF, the ensemble machine learning shows better predictive performance.

Tan et al.[43] proposed a staking method called ELDAP, which is slightly different from above ensemble learning. The staking method means individual predictions of a collection of models are introduced as inputs to a second-level learning algorithm is an extensively used technique for boosting prediction accuracy. In the study, Tan et al. blended EN, KBMTL, NN and Pair SVR through a stacking module. The staking method have lower MSE than other any baseline learner when training the data from GDSC and CCLE.

In addition to improving model prediction performance through ensemble learning, researchers often use NN to construct better models. NN simulates the working principle of the human brain and deep learning is a kind of machine learning based on deep neural networks (DNN). In essence, DNN is the optimization of NN.

Ma et al.[30] proposed a cancer drug response prediction model based on dual GCN, named DualGCN. This model consists of two modules, namely Drug-GCN and Bio-GCN, both of which use graph neural networks as their backbone. Specifically, Drug-GCN takes the chemical structure information of drugs as input, treating each drug as a graph where nodes represent drug atoms and edges indicate connections between atoms, thereby effectively extracting the intrinsic chemical properties of the drugs. Bio-GCN, on the other hand, takes the genetic features of cancer samples as input, treating each cancer sample as a graph where nodes represent proteins (genes) and edges represent interactions between proteins, thus effectively encoding the latent features of the genes. Finally, the encoded features from the two modules are concatenated and input into a multilayer perceptron to obtain the response of a given drug on a given cancer sample. DualGCN incorporates prior knowledge on cancer-related genes and protein−protein interactions while avoiding the use of large-scale single nucleotide variant data, and its performance is superior to most state-of-the-art methods.

Tang et al.[25] constructed a pathway-based cancer drug sensitivity prediction model named PathDSP, which is built upon a fully connected neural network, taking two types of drug-related features and three types of cell line-related features as input, where the chemical data is processed using Morgan fingerprints,

Table 4 The overview of studies based on cell line data for drug response prediction.

Study	Backbone/Method	Task	Result
Mourragui et al.[70]	EN	Drug response prediction	\
Kim et al.[19]	EN	Drug resistance prediction	AUC = 1.00; AUPCR = 1.00
Liu et al.[20]	SB	Drug response prediction	AUC = 0.942, 0.791 (prostate cancer cell, breast cancer cell)
Mannheimer et al.[22]	SVR	Drug sensitivity prediction	r = 0.6074
Ma et al.[24]	GCN	Drug response prediction	RMSE = 1.079; r = 0.925; R_s = 0.907
Tang et al.[25]	FNN	Drug sensitivity prediction	RMSE = 0.35; MAE = 0.24; R^2 = 0.92; r = 0.96
Liu et al.[28]	RF; NN; LR	Drug resistance prediction	AUC = 0.998,0.997,0.991 (RF, LR, NN)
Ma et al.[30]	RF	Drug response prediction	\
Liu et al.[29]	RF; LR; NN	Drug resistance prediction	AUC = 0.9542, 0.9884, 0.9976 (RF, LR, NN)
Mian et al.[31]	NN	Drug sensitivity prediction	\
Yanagisawa et al.[32]	CNN	Drug resistance prediction	TPR = 0.68; TNR = 0.76; ACC = 0.72
Majumdar et al.[34]	Ensemble of k-means; PCA; SVR	Drug response prediction	MSE = 0.336, 0.307, 0.301,0.350, 0.450
Sharma et al.[35]	RF	Drug sensitivity prediction	MSE = 3.14, 0.414 (GDSC, CCLE)
Naulaerts et al.[36]	XGBoost	Drug sensitivity prediction	$R_s > 0.25$
Choi et al.[37]	NN and EN	Drug resistance prediction	AUC = 0.891,0.951; AUPRC = 0.932,0.958; ACC = 0.81,0.91 (GDSC, CCLE)
Bazgir et al.[38]	CNN	Drug sensitivity prediction	ACC = 0.754 (NCI-60); NRMSE = 0.414, r = 0.911 (GDSC)
Chiu et al.[40]	NN	Drug response prediction	MSE = 0.59(GDSC); MSE = 0.62(CCLE)
Carroll et al.[41]	O-PLS	Drug response prediction	RMSE = 0.09
Wang et al.[42]	EN	Drug response prediction	r = 0.59, 0.62 (CCLE, CTRP)
Tan et al.[43]	KMTL	Drug response prediction	MSE = 3.305, 0.501, 0.556 (GDSC, CCLE), NCI-DREAM)
Dorman et al.[44]	SVM	Drug sensitivity prediction	\
Menden et al.[46]	RF; NN	Drug sensitivity prediction	r = 0.85, 0.85; R^2 = 0.72, 0.72; RMSE = 0.83, 0.84 (NN, RF)
Chen et al.[47]	NN	Drug response prediction	AUC = 0.898; TNR = 0.899; PREC = 0.926; Recall = 0.899; F1 = 0.892
Lee et al.[48]	NN	Drug response prediction	AUC = 0.8939; AUPRC = 0.5906, ACC = 0.9137; PREC = 0.6409
Sun et al.[49]	EN	Drug sensitivity prediction	\
Jia et al.[50]	RF; NN; SVM	Drug response prediction	AUC = 0.875
Wang et al.[51]	GCN	Drug sensitivity prediction	R^2 = 0.863; AUC = 0.858

ACC: accuracy; KMTL: Kernelized Multi-Task Learning; EN: Elastic Net; SB: Sparse Bayesian; SVR: Pairwise Support Vector Regression; NN: Neural Network; RF: Random Forest; LR: Logistic Regression; CNN: Convolutional Neural Network; GCN: Graph convolutional networks; PCA: principal component analysis; O-PLS: orthogonal partial least squares; AUC: area under curve; AUPCR: area under the Precision-Recall curve; TNR: true negative rate; TPR: true positive rate; r: Pearlman Correlation Coefficient; MSE: mean-square error; R_s: Spearman correlation coefficient; RMSE: root mean square error; R^2: determination coefficient.

Table 5 The overview of studies based clinical data for drug response prediction.

Study	Backbone/Method	Task	Result
Nam et al.[53]	XGBoost	Drug response prediction	P = 1.12e−4, 3.63e−4 (PFS, OS)
Yu et al.[54]	RF; SVM; Bagging; NBM	Drug response prediction	\
Yang et al.[56]	RF	Drug resistance prediction	AUC = 0.83; TPR = 0.889; TNR = 0.928; ACC = 0.913
Xiong et al.[57]	MLR	Drug sensitivity prediction	AUC = 0.935
Johannet et al.[15]	CNN	Drug response prediction	AUC = 0.805
Yi et al.[58]	RF and SVM	Drug resistance prediction	AUC = 0.967
Buttarelli et al.[59]	RF	Drug response prediction	ACC = 0.93; Precision = 0.94
Abraham et al.[60]	RF; SVM; NN; etc.	Drug response prediction	HR = 0.629, 0.483; P = 0.04, 0.02 (FOLFOX, FOLFOXIRI)
Lu et al.[61]	SVM; RF; NN	Drug sensitivity prediction	AUC = 0.827, 0.877, 0.800 (SVM, RF, NN)
Dercle et al.[62]	RF	Drug response prediction	AUC = 0.80, 0.72; TPR = 0.80, 0.82; TNR = 0.78, 0.61 (FSHQ, FCSD)
Ruffalo et al.[63]	SVM, RF	Drug response prediction	AUC = 0.91
Herold et al.[64]	Lasso	Drug resistance prediction	ACC = 0.77
Nasimian et al.[67]	TabNet	Drug resistance prediction	ACC = 0.83; Precision = 0.84
Kanda et al.[68]	CNN	Drug sensitivity prediction	ACC = 0.934; AUC = 0.979

ACC: accuracy; HR: hazard ratio; MLR: multivariable logistic regression; NBM: Naive Bayes Model; GBM: Gradient Boosting Machine; KNN: K-Nearest Neighbor; DT: decision tree; Lasso: logistic regression; PFS: progression-free survival; OS: overall survival; TNR: true negative rate; TPR: true positive rate; AUC: area under curve.

while the cell line omics data is analyzed through pathway enrichment. In the experiments, the data was standardized using z-scores before being input into the model, and a ten-fold cross-validation was conducted. Additionally, early stopping was applied to avoid overfitting and ensure the robustness of the results. The experimental results demonstrate that PathDSP addresses the issues of high dimensionality and heterogeneity in the data by mapping the data used in the task onto curated pathways. Compared with the current state-of-the-art methods, this approach achieves better performance, enhances the model's generalization ability across different datasets, and offers explanatory power at the cancer pathway level beyond individual genes or mutations. Yanagisawa et al.[32] constructed a CNN based on the VGG16 model to predict drug response at the single-cell level. Interestingly, it focused on circulating tumor cells entering cancer patients' circulatory systems and realized the predictive goal by training models based on phase-contrast images of colorectal cell lines, DLD-1 and HCT-116. The accuracy of model is about 0.8. Since circulating tumor cells in the patient's blood can be extracted for predicting drug reactivity, the algorithm is suitable for clinical use.

Bazgir et al.[38] present a feature representation approach termed REFINED to arrange high-dimensional vectors in a compact image form conducible for CNN-based deep learning. Different from Yanagisawa et al., the training images are not obtained directly from microscope. The feature map was generated by minimizing the pairwise distance values following a Bayesian Metric Multidimensional Scaling Approach from high-dimensional vectors, making genomics and proteomics data can be used as the training data of CNN in the form of two-dimensional images. CNN-based deep learning has the advantages of automated feature extraction and high-accuracy prediction. Comparing with two alternative image generation methods: PCA representation and random projections and other machine learning algorithms: artificial NN, RF, SVR, logistic regression, REFINED-CNN show the best predictive performance in the NCI-60 cell lines.

RefDNN was developed by Choi et al.[37] as a novel Reference Drug-based Neural Network model for prediction of anticancer drug response and identification of biomarkers related to drug resistance, which requires a pair consisting of the molecular structure data of a drug and gene expression profiles. RefDNN combines EN with DNN. The EN algorithm used for calculating high-dimensional gene expression data, which allows it to identify the biomarker of drug resistance. In performance comparison, the predictive performance of RefDNN is significantly better than the three traditional machine learning models, such as KNN, Multilayer Perceptron, EN. However, RefDNN does not exceed the two ensemble models XGBoost and RF in some metrics.

4.1.2. AI models constructed based on clinical data

The construction of AI algorithm for drug response based on clinical data is faced with the difficult problem of data collection. Existing studies usually need to collect patients' data like genome, epigenetics, clinical data. Because there is no large public database with various omics data and drug resistance information of patients, studies based on clinical data are usually faced with problems that the datasets for construction are small and single. RF[56,60,62,63], Lasso Regression[64], SVM[63,75] are frequently used in the prediction of drug resistance.

Yi et al.[58] include the radiomic texture features of CT images, clinicopathological features and SNP features of 102 OC patients in the study. The innovation of this method is that SVM was used to score the two weak regressors for regression, and the classification effect was better than that obtained by RF and Lasso alone. Moreover, combined with multi-omics data, the model had a higher AUC than that constructed only for gene-level characteristics or radiomic data.

Abraham et al.[60] developed and validated a molecular signature predictive of efficacy of oxaliplatin-based chemotherapy combined with bevacizumab in patients with mCRC. The algorithm used for prediction is deliberative analytics (DEAN) framework combing many Algorithms. Researchers take TTNT (time-to-next treatment) into training and use Real Word Evidence (RWE) including 412 patients receiving FOLFOX/bevacizumab therapy. It can distinguish FOLFOX-treated mCRC patients with increased benefit from those with decreased benefit.

Xiong et al.[57] proposed a combined model integrating MRI image features and independent clinical data of 125 breast cancer patients. First, for features of MRI images, Lasso regression was selected for screening. Backward step-wise selection following the Akaike's information criterion was applied to select the optimal variable dataset based on clinical characteristics. Finally, the remaining variables of clinical data and MRI images features were integrated into a multivariate logistic regression model. The AUC of combined models was higher than that of models constructed based solely on radiomic or clinical features, which is 0.986 in the training cohort and 0.935 in the validation cohort.

Liu et al.[20] construct a disease-specific driver signal network based on genomic data from TCGA patients with prostate cancer. Meanwhile, the drug resistance data of 16 prostate cancer cell lines to Dasatinib collected by Wang et al[86]. was used to construct a network-based sparse Bayesian machine (NBSBM) algorithm by combining a sparse Bayesian classifier with a Laplace graph. Then, researchers used NBSBM to predict resistance to Tamoxifen (GSE17705) in 103 breast cancer patients. Comparing the SVM based recursive feature elimination classifier (SVM-RRFE) and NBSVM, AUC of NBSBM reached almost 0.8 with better predictive performance.

Constrained by the size of the data, there are almost no studies that build neural network algorithm models based on clinical data. Researchers primarily improve algorithm performance by stacking several layers of AI algorithms or using multiple AI algorithms in parallel.

4.2. AI for drug resistance biomarker prediction

AI algorithms are used to predict resistance-related biomarkers, thereby indirectly predicting drug responsiveness. P-glycoprotein, MRP1 and BRCP are the most relevant ABC-transporters in conferring a multidrug-resistance phenotype to cancer cells[87]. Zhong et al.[88] proposed an integrated scheme for the simultaneous treatment of both feature selection and parameter based on SVM called GA-CG-SVM, which is used for identifying the substrates and non-substrates of BCRP and perform better than C4.5 DT and k-NN model. it presented 91.3% and 85.0% of overall accuracy in the training and test set, respectively[1]. Gantner et al.[89] also proposed a machine learning model for predicting BCRP substrates, which have no requirement of preprocessing of the predicted chemical structure. However, its overall accuracy is slightly lower with 82% in training set and 74.5% in test set. Estrada-Tejedor and Ecker used a modified algorithm based on Self-Organizing Maps called CSOM, an unsupervised neural network, to predict P-glycoprotein, MRP1 and BCRP substrates, which can solve the problem of processing imbalanced data[90].

AI algorithms can also be used to predict other molecular related to drug resistance. For example, to MicroRNA, it regulates gene expression in eukaryotes and prokaryotes and affect the anti-cancer drug response. PDSM-LGCN, a light graph convolution neural network for predicting drug sensitivity associated micro-RNAs, can calculate the relationship between microRNA and anticancer drugs sensitivity, which saves time and effort than traditional bioinformatics methods. The AUC and AUPRC is 0.8872 and 0.9026, respectively. The core principle is to obtain high-dimensional features of both parties and then reconstructed the correlation relationship according to the eigenvalues[91].

Additionally, many researchers tend to detect biomarkers related to drug resistance. Han et al.[92] applied SVM algorithm to the screening of serum biomarkers for predicting chemotherapy resistance in small cell lung cancer, and they identified *m/z* 10468Da as S100-A9 for separating the chemotherapy-resistance group from all patients treated by cisplatin plus etoposide. Also, there was a machine learning algorithm developed by Shannon et al.[93], aiming at identifying predictive molecular biomarkers for cisplatin chemosensitivity following surgical resection in ovarian cancer. They evaluated many algorithms and finally choose the GB with the highest accuracy. Four potential biomarkers were obtained. Different from them, Wu et al.[94] first determined clear cell renal cell carcinoma subtypes based on the expression of mRNA, miRNA and lncRNA by using unsupervised learning, then they detected the subtype specific biomarkers based on RF. Kim et al.[95] conducted a prospective phase II trial of ramucirumab plus paclitaxel as second-line treatment in 62 patients with metastatic gastric cancer and identified some transcriptomes, which are obviously upregulated in non-responders and associated with angiogenesis, chemotaxis, and VEGF signals, enrichments of metabolism and sonic hedgehog cellular pathways. Based on the analysis of somatic mutations on patients with stage III colorectal cancer, Lin et al. constructed a survival decision tree to divide patients into seven subtypes and found that mutations in DNA damage response genes represent promising biomarkers to assess the response of stage III colorectal cancer patients to oxaliplatin-based chemotherapy. The study considered the impact of tumor evolution through cancer evolutionary tree because the genetic mutation order or co-occurring genetic mutations are important in cancer biology[96].

These researches have boosted the detect for anti-cancer drug biomarkers among multiple cancer patients and conducted preliminary screening for biomarkers, which is benefit for the research on drug resistance mechanism and clinical medication. However, it is hard to interpret the result because the nature of AI is black box, and lacking the rigorous experiments to verify the drug resistance mechanism of biomarkers leads to less practical of the screening results.

4.3. *AI for compound−protein interaction prediction*

Accurate prediction of compound−protein interactions can enable researchers to rapidly screen for potential drug candidates and optimize drug structures, which is of great significance in advancing the development of personalized medicine. Traditional laboratory validation methods are not only time-consuming and labor-intensive but also costly. In recent years, significant progress has been made in the application of AI methods in this field. The DeepMind research team[97] has developed a model called AlphaFold3, which is designed to predict the interactions and structures between proteins and a wide range of other biomolecules. This model can predict the structures and interactions of almost all biomolecules, including but not limited to proteins, DNA, RNA, and small molecules, thereby presenting novel possibilities for drug design and disease treatment. Ni et al.[98] have developed a method based on knowledge graph technology to analyze compound−protein interactions from perturbation transcriptomics, which identifies ectonucleotide pyrophosphatase/phosphodiesterase-1 as the target for the unique anti-tumor immunotherapy effect of tankyrase inhibitor K-756 and discovers five novel candidate drugs targeting the emerging cancer therapeutic target aldehyde dehydrogenase 1B1, with a hit rate of up to 10.2%. Cao et al.[99] have proposed a novel geometric diffusion network based on protein surfaces, named SurfDock, which can generate reliable protein-ligand complex conformations with high accuracy and has been successfully applied to structure-based virtual screening. Tao et al.[100] proposed an innovative framework named BEACON, which enhances the accuracy of compound−protein interaction predictions by bridging chemical structures and conceptual knowledge. Extensive experiments conducted on four datasets demonstrate that BEACON significantly outperforms state-of-the-art baseline methods.

4.4. *Practical case studies*

In recent years, several studies have been conducted to investigate real-world applications of AI-based technology for predicting drug responses. For instance, Yi et al.[58] have developed a machine learning model to predict platinum resistance in the treatment of ovarian cancer. Their study retrospectively enrolled 102 OC patients from January 2006 to February 2018. In the validation cohort, the model achieved an AUC of 0.967, demonstrating its promising clinical application value. Buttarelli et al.[59] employed machine learning techniques to identify a novel gene signature that accurately predicts the response to first-line chemotherapy in patients with BRCA wild-type high-grade serous ovarian cancer, providing a useful tool for personalized treatment modalities. Dercle et al.[62] conducted a retrospective analysis using radiomics data based on machine learning methods to investigate the treatment sensitivity in 667 patients with metastatic colorectal cancer who were treated with irinotecan, 5-fluorouracil, and leucovorin alone or in combination with cetuximab. The experimental results revealed that radiomics data outperformed known biomarkers in providing early prediction of treatment sensitivity and could be used to guide decisions on the continuation or cessation of cetuximab treatment. These Practical Case Studies reveal that AI-based technology can accurately predict drug responses and identify patient populations that can benefit from treatment, which has important clinical significance for tailoring personalized treatment plans for patients.

5. Clinical application barriers and solutions

Despite the notable clinical application advantages demonstrated by AI-based prediction of cancer drug resistance, several obstacles persist in its practical clinical deployment, primarily manifesting in issues related to data, model performance, the inherent black-box nature of AI, as well as legal and ethical concerns.

Large-scale high-quality data forms the cornerstone for developing high-performance medical AI models. However, collecting sufficient high-quality data in the medical field poses significant challenges, primarily due to limitations in the discretized distribution of medical data, and inadequate data quality.

Specifically, even within the same hospital or medical center, medical data is dispersed across various information systems, such as the Hospital Information System (HIS), Laboratory Information System (LIS), and Picture Archiving and Communication System (PACS). The lack of unified standards and interfaces among these data systems makes it difficult to integrate and share data. Furthermore, aggregating data from different medical centers across countries or regions poses an even greater challenge. Similarly, the quality of data is also a crucial factor restricting the use of medical data. Some data may be subject to multiple noise interferences from equipment and the environment during the collection process, leading to issues such as data distortion. Moreover, some data suffer from severe modality deficiencies, challenging data integrity. To address these data issues, several corresponding solutions have been proposed successively. For instance, federated learning offers a solution to challenges such as discrete distribution. This method enables various medical centers to collaboratively train a shared machine learning model without duplicating or transferring their data. Data generation techniques, represented by generative adversarial networks (GANs), can tackle issues like missing data modalities and distorted data quality. These techniques generate entirely new data by learning from the distribution of existing data. Furthermore, international cooperation in establishing standards for data acquisition and quality control, as well as actively building multicenter medical big data repositories, are equally crucial solutions.

Secondly, the instability of AI model performance is one of the key factors limiting its clinical application. The factors contributing to the differences in model performance are mainly manifested in three aspects: algorithmic differences, data-specific differences, and biological complexity. Specifically, different algorithms have distinct design principles, which determine their adaptability to different tasks. For example, the dual-layer graph convolution structure of DualGCN[24] enhances cross-modal information fusion, giving it an advantage in multi-omics data integration scenarios; PathDSP[25] focuses on the dynamic changes of signaling pathways through RNN and attention mechanisms, making it better at uncovering the temporal patterns of biological processes. Additionally, the hyperparameter settings and optimization strategies (such as learning rate and loss function selection) of algorithms can also lead to performance fluctuations. Furthermore, data collected from different medical centers or using different devices and methods exhibit certain heterogeneity, meaning there are differences in the distribution of feature data. This results in the excellent performance of the model being limited to specific datasets and lacking generalization across data from diverse populations in different regions. Moreover, current research often focuses solely on a single disease type, lacking diverse data encompassing different diseases, populations, and treatment stages, which leads to the lack of universal applicability of model performance. Furthermore, biological complexity is also a significant factor contributing to the differences in model performance. Cancer drug resistance involves changes at multiple levels, including genetic mutations, epigenetic modifications, alterations in protein expression, and metabolic reprogramming. Different algorithms have varying abilities to capture and analyze these complex factors, thereby affecting performance. On the other hand, tumor heterogeneity is widespread, with different parts and cellular subsets exhibiting distinct drug resistance mechanisms. Additionally, the interaction between the tumor microenvironment and tumor cells adds to the complexity of drug resistance, making it difficult for algorithms to accurately address, ultimately leading to differences in performance. The performance differences of the models in clinical settings can significantly impact doctors' ability to formulate precise treatment plans. When the prediction error of drug sensitivity exceeds 10%, it may lead to overtreatment (such as unnecessary chemotherapy) or undertreatment. For example, if an anticancer drug response prediction model reduces the predicted probability of an actually effective patient from 85% to 70%, it may cause doctors to mistakenly opt for conservative treatment. Currently, there are several solutions to improve the stability of model performance. For example, federated learning can be employed to leverage diverse data from multiple centers for distributed training, thereby reducing the impact of data heterogeneity. Another approach is to use ensemble learning, which combines the strengths of multiple algorithms to enhance the robustness of the model. Additionally, regularization techniques are used to constrain the complexity of the model and improve its stability against data variations and noise, further enhancing its robustness.

Then, the black-box nature of AI is also a significant factor hindering the development of medical AI. AI models, exemplified by deep learning, are trained in an end-to-end manner. This implies that data is input into the model to directly obtain corresponding results, but the reasons behind these results remain unknown. The lack of interpretability in the decision-making process of AI models often leads to distrust from doctors and patients, which severely restricts the deployment and application of AI models in clinical settings. In recent years, several methods have been developed to improve the interpretability of AI models. These include the interpretable heatmap method, which visualizes model parameters to reveal the importance of different features in the input data; the occlusion method, which assesses the importance of occluded parts to model decisions by observing changes in model output when input data is folded or masked; and the knowledge graph method, which represents data in a structured knowledge format for model training. Furthermore, some excellent interpretable deep learning frameworks have been proposed in genomic research. For example, Li et al.[101] constructed an interpretable deep learning framework based on the graph attention mechanism, which not only identifies key cancer genes but also explains how these genes interact and their roles in cancer development. Novakovsky et al.[102] proposed an interpretable and transparent neural network for genomics, which can predict transcription factor binding sites, chromatin accessibility, and de novo motifs. The predictions generated by it are transparent, providing both global (at the cell state level) and local (at the individual sequence level) biological insights into the data.

Finally, legal and ethical considerations are also pivotal factors limiting the development of medical AI models, particularly in the areas of data collection and model application. Data serves as the foundation for AI models, yet medical data often contains private information of patients, whose security is strictly regulated and protected by both ethics and law. Consequently, the process of collecting and using medical data must obtain approval from regulatory agencies, secure informed consent from patients, undergo rigorous data anonymization, and be stored in encrypted format. These steps lead to a prolonged and more challenging data collection process. Furthermore, the deployment and use of medical AI models in clinical settings similarly require approval from relevant regulatory agencies. For instance, in the United States, medical AI needs to obtain approval from the FDA. Additionally, there are still gaps in legal frameworks regarding the delineation of responsibility for entities involved in AI-related

medical accidents. To address ethical and legal obstacles, the primary step should be to improve laws and related systems in this field, clarifying the boundaries for the implementation of medical AI. At the same time, a comprehensive ethical review mechanism should be established to conduct rigorous ethical reviews of medical activities involving AI technology, ensuring that the use of technology conforms to ethical principles. Additionally, technical means such as data encryption and de-identification can be employed to protect patient-related information.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (No. 82404186), the China Postdoctoral Science Foundation (No. 2023M743947, China), the Postdoctoral Fellowship Program of CPSF (No. GZB20240869, China), the Hunan Provincial Natural Science Foundation of China (Nos.2024JJ6697 and 2025JJ40074) and the Changsha Natural Science Foundation (No. kq2403026, China).

Author contributions

Conceptualization: Xinrui Long, Kai Huang, Kai Sun. Methodology: Xinrui Long, Kai Huang, Kai Sun. Visualization: Xinrui Long, Kai Huang, Sicen Lai. Supervision: Yuancheng Liu, Wangqing Chen, Juan Su, Ruhan Liu, Shuang Zhao. Writing − original draft: Xinrui Long, Kai Sun, Shuang Zhao, Kai Huang, Wangqing Chen. Writing − review & editing: Kai Huang, Kai Sun, Xiaoyu He.

Declaration of generative AI and AI-assisted technologies in the writing process

During the preparation of this work the author(s) used AI to polish the language. After using this tool, the author(s) reviewed and edited the content as needed and took full responsibility for the content of the publication.

Conflicts of interest

The authors declare no conflicts of interest.

References

1. Cao B, Bray F, Ilbawi A, Soerjomataram I. Effect on longevity of one-third reduction in premature mortality from non-communicable diseases by 2030: a global analysis of the Sustainable Development Goal health target. *Lancet Global Health* 2018;**6**:e1288−96.
2. Yen I, Shanahan F, Lee J, Hong YS, Shin SJ, Moore AR, et al. ARAF mutations confer resistance to the RAF inhibitor belvarafenib in melanoma. *Nature* 2021;**594**:418−23.
3. Hanahan D. Hallmarks of cancer: new dimensions. *Cancer Discov* 2022;**12**:31−46.
4. Dunn GP, Bruce AT, Ikeda H, Old LJ, Schreiber RD. Cancer immunoediting: from immunosurveillance to tumor escape. *Nat Immunol* 2002;**3**:991−8.
5. Vasan N, Baselga J, Hyman DM. A view on drug resistance in cancer. *Nature* 2019;**575**:299−309.
6. Chong CR, Jänne PA. The quest to overcome resistance to EGFR-targeted therapies in cancer. *Nat Med* 2013;**19**:1389−400.
7. Nussinov R, Tsai CJ, Jang H. Anticancer drug resistance: an update and perspective. *Drug Resist Updates* 2021;**59**:100796.
8. Bueschbell B, Caniceiro AB, Suzano PMS, Machuqueiro M, Rosário-Ferreira N, Moreira IS. Network biology and artificial intelligence drive the understanding of the multidrug resistance phenotype in cancer. *Drug Resist Updates* 2022;**60**:100811.
9. Mullard A. Stemming the tide of drug resistance in cancer. *Nat Rev Drug Discov* 2020;**19**:221−3.
10. Kriegeskorte N, Golan T. Neural network models and deep learning. *Curr Biol* 2019;**29**:R231−6.
11. Zheng H, Liu J, Cheng Q, Zhang Q, Zhang Y, Jiang L, et al. Targeted activation of ferroptosis in colorectal cancer *via* LGR4 targeting overcomes acquired drug resistance. *Nat Cancer* 2024;**5**:572−89.
12. Xu Z, Wang X, Zeng S, Ren X, Yan Y, Gong Z. Applying artificial intelligence for cancer immunotherapy. *Acta Pharm Sin B* 2021;**11**:3393−405.
13. Yang Y, Ye Z, Su Y, Zhao Q, Li X, Ouyang D. Deep learning for *in vitro* prediction of pharmaceutical formulations. *Acta Pharm Sin B* 2019;**9**:177−85.
14. Rajkomar A, Dean J, Kohane I. Machine learning in medicine. *N Engl J Med* 2019;**380**:1347−58.
15. Johannet P, Coudray N, Donnelly DM, Jour G, Illa-Bochaca I, Xia Y, et al. Using machine learning algorithms to predict immunotherapy response in patients with advanced melanoma. *Clin Cancer Res* 2021;**27**:131−40.
16. Zhu Y, Ouyang Z, Du H, Wang M, Wang J, Sun H, et al. New opportunities and challenges of natural products research: when target identification meets single-cell multiomics. *Acta Pharm Sin B* 2022;**12**:4011−39.
17. Gomes B, Ashley EA. Artificial intelligence in molecular medicine. *N Engl J Med* 2023;**388**:2456−65.
18. Lightbody G, Haberland V, Browne F, Taggart L, Zheng H, Parkes E, et al. Review of applications of high-throughput sequencing in personalized medicine: barriers and facilitators of future progress in research and clinical application. *Briefings Bioinf* 2019;**20**:1795−811.
19. Kim YR, Kim D, Kim SY. Prediction of acquired Taxane resistance using a personalized pathway-based machine learning method. *Cancer Res Treat* 2019;**51**:672−84.
20. Liu Q, Muglia LJ, Huang LF. Network as a biomarker: a novel network-based sparse Bayesian machine for pathway-driven drug response prediction. *Genes* 2019;**10**:602.
21. Parca L, Pepe G, Pietrosanto M, Galvan G, Galli L, Palmeri A, et al. Modeling cancer drug response through drug-specific informative genes. *Sci Rep* 2019;**9**:15222.
22. Mannheimer JD, Prasad A, Gustafson DL. Predicting chemosensitivity using drug perturbed gene dynamics. *BMC Bioinf* 2021;**22**:15.
23. Tan M, Özgül OF, Bardak B, Ekşioğlu I, Sabuncuoğlu S. Drug response prediction by ensemble learning and drug-induced gene expression signatures. *Genomics (San Diego, Calif)* 2019;**111**:1078−88.
24. Ma T, Liu Q, Li H, Zhou M, Jiang R, Zhang X. DualGCN: a dual graph convolutional network model to predict cancer drug response. *BMC Bioinf* 2022;**23**:129.
25. Tang YC, Gottlieb A. Explainable drug sensitivity prediction through cancer pathway enrichment. *Sci Rep* 2021;**11**:3128.
26. Lee SI, Celik S, Logsdon BA, Lundberg SM, Martins TJ, Oehler VG, et al. A machine learning approach to integrate big data for precision medicine in acute myeloid leukemia. *Nat Commun* 2018;**9**:42.
27. El-Deredy W, Ashmore SM, Branston NM, Darling JL, Williams SR, Thomas DG. Pretreatment prediction of the chemotherapeutic response of human glioma cell cultures using nuclear magnetic resonance spectroscopy and artificial neural networks. *Cancer Res* 1997;**57**:4196−9.
28. Liu R, Sun M, Zhang G, Lan Y, Yang Z. Towards early monitoring of chemotherapy-induced drug resistance based on single cell metabolomics: combining single-probe mass spectrometry with machine learning. *Anal Chim Acta* 2019;**1092**:42−8.
29. Liu R, Zhang G, Yang Z. Towards rapid prediction of drug-resistant cancer cell phenotypes: single cell mass spectrometry combined with machine learning. *Chem Commun* 2019;**55**:616−9.

30. Ma Y, Ding Z, Qian Y, Shi X, Castranova V, Harner EJ, et al. Predicting cancer drug response by proteomic profiling. *Clin Cancer Res* 2006;**12**:4583−9.
31. Mian S, Ball G, Hornbuckle J, Holding F, Carmichael J, Ellis I, et al. A prototype methodology combining surface-enhanced laser desorption/ionization protein chip technology and artificial neural network algorithms to predict the chemoresponsiveness of breast cancer cell lines exposed to Paclitaxel and Doxorubicin under *in vitro* conditions. *Proteomics* 2003;**3**:1725−37.
32. Yanagisawa K, Toratani M, Asai A, Konno M, Niioka H, Mizushima T, et al. Convolutional neural network can recognize drug resistance of single cancer cells. *Int J Mol Sci* 2020;**21**:3166.
33. Steyaert S, Pizurica M, Nagaraj D, Khandelwal P, Hernandez-Boussard T, Gentles AJ, et al. Multimodal data fusion for cancer biomarker discovery with deep learning. *Nat Mach Intell* 2023;**5**: 351−62.
34. Majumdar A, Liu Y, Lu Y, Wu S, Cheng L. kESVR: an ensemble model for drug response prediction in precision medicine using cancer cell lines gene expression. *Genes* 2021;**12**:844.
35. Sharma A, Rani R. Ensembled machine learning framework for drug sensitivity prediction. *IET Syst Biol* 2020;**14**:39−46.
36. Naulaerts S, Menden MP, Ballester PJ. Concise polygenic models for cancer-specific identification of drug-sensitive tumors from their multi-omics profiles. *Biomolecules* 2020;**10**:963.
37. Choi J, Park S, Ahn J. RefDNN: a reference drug based neural network for more accurate prediction of anticancer drug resistance. *Sci Rep* 2020;**10**:1861.
38. Bazgir O, Zhang R, Dhruba SR, Rahman R, Ghosh S, Pal R. Representation of features as images with neighborhood dependencies for compatibility with convolutional neural networks. *Nat Commun* 2020;**11**:4391.
39. Sakellaropoulos T, Vougas K, Narang S, Koinis F, Kotsinas A, Polyzos A, et al. A deep learning framework for predicting response to therapy in cancer. *Cell Rep* 2019;**29**:3367−73.e4.
40. Chiu YC, Chen HH, Zhang T, Zhang S, Gorthi A, Wang LJ, et al. Predicting drug response of tumors from integrated genomic profiles by deep neural networks. *BMC Med Genom* 2019;**12**:18.
41. Carroll MJ, Parent CR, Page D, Kreeger PK. Tumor cell sensitivity to vemurafenib can be predicted from protein expression in a BRAF-V600E basket trial setting. *BMC Cancer* 2019;**19**:1025.
42. Wang Y, Wang Z, Xu J, Li J, Li S, Zhang M, et al. Systematic identification of non-coding pharmacogenomic landscape in cancer. *Nat Commun* 2018;**9**:3192.
43. Tan M. Prediction of anti-cancer drug response by kernelized multi-task learning. *Artif Intell Med* 2016;**73**:70−7.
44. Dorman SN, Baranova K, Knoll JH, Urquhart BL, Mariani G, Carcangiu ML, et al. Genomic signatures for paclitaxel and gemcitabine resistance in breast cancer derived by machine learning. *Mol Oncol* 2016;**10**:85−100.
45. Daemen A, Griffith OL, Heiser LM, Wang NJ, Enache OM, Sanborn Z, et al. Modeling precision treatment of breast cancer. *Genome Biol* 2013;**14**:R110.
46. Menden MP, Iorio F, Garnett M, McDermott U, Benes CH, Ballester PJ, et al. Machine learning prediction of cancer cell sensitivity to drugs based on genomic and chemical properties. *PLoS One* 2013;**8**:e61318.
47. Chen J, Wang X, Ma A, Wang QE, Liu B, Li L, et al. Deep transfer learning of cancer drug responses by integrating bulk and single-cell RNA-seq data. *Nat Commun* 2022;**13**:6494.
48. Lee K, Cho D, Jang J, Choi K, Jeong HO, Seo J, et al. RAMP: response-aware multi-task learning with contrastive regularization for cancer drug response prediction. *Briefings Bioinf* 2023;**24**:bbac504.
49. Sun R, Ge W, Zhu Y, Sayad A, Luna A, Lyu M, et al. Proteomic dynamics of breast cancer cell lines identifies potential therapeutic protein targets. *Mol Cell Proteomics* 2023;**22**:100602.
50. Jia K, Wang Y, Cao QI, Wang Y. Extensive prediction of drug response in mutation-subtype-specific LUAD with machine learning approach. *Oncol Res* 2023;**32**:409−19.
51. Wang Y, Yu X, Gu Y, Li W, Zhu K, Chen L, et al. XGraphCDS: an explainable deep learning model for predicting drug sensitivity from gene pathways and chemical structures. *Comput Biol Med* 2024;**168**: 107746.
52. Wiesweg M, Mairinger F, Reis H, Goetz M, Walter RFH, Hager T, et al. Machine learning-based predictors for immune checkpoint inhibitor therapy of non-small-cell lung cancer. *Ann Oncol* 2019;**30**: 655−7.
53. Nam Y, Koo H, Yang Y, Shin S, Zhu Z, Kim D, et al. Pharmacogenomic profiling reveals molecular features of chemotherapy resistance in IDH wild-type primary glioblastoma. *Genome Med* 2023;**15**:16.
54. Yu KH, Levine DA, Zhang H, Chan DW, Zhang Z, Snyder M. Predicting ovarian cancer patients' clinical response to platinum-based chemotherapy by their tumor proteomic signatures. *J Proteome Res* 2016;**15**:2455−65.
55. Qureshi R, Basit SA, Shamsi JA, Fan X, Nawaz M, Yan H, et al. Machine learning based personalized drug response prediction for lung cancer patients. *Sci Rep* 2022;**12**:18935.
56. Yang C, Jiang ZK, Liu LH, Zeng MS. Pre-treatment ADC image-based random forest classifier for identifying resistant rectal adenocarcinoma to neoadjuvant chemoradiotherapy. *Int J Colorectal Dis* 2020;**35**:101−7.
57. Xiong Q, Zhou X, Liu Z, Lei C, Yang C, Yang M, et al. Multiparametric MRI-based radiomics analysis for prediction of breast cancers insensitive to neoadjuvant chemotherapy. *Clin Transl Oncol* 2020;**22**:50−9.
58. Yi X, Liu Y, Zhou B, Xiang W, Deng A, Fu Y, et al. Incorporating SULF1 polymorphisms in a pretreatment CT-based radiomic model for predicting platinum resistance in ovarian cancer treatment. *Biomed Pharmacother* 2021;**133**:111013.
59. Buttarelli M, Ciucci A, Palluzzi F, Raspaglio G, Marchetti C, Perrone E, et al. Identification of a novel gene signature predicting response to first-line chemotherapy in BRCA wild-type high-grade serous ovarian cancer patients. *J Exp Clin Cancer Res* 2022;**41**:50.
60. Abraham JP, Magee D, Cremolini C, Antoniotti C, Halbert DD, Xiu J, et al. Clinical validation of a machine-learning-derived signature predictive of outcomes from first-line oxaliplatin-based chemotherapy in advanced colorectal cancer. *Clin Cancer Res* 2021; **27**:1174−83.
61. Lu W, Fu D, Kong X, Huang Z, Hwang M, Zhu Y, et al. FOLFOX treatment response prediction in metastatic or recurrent colorectal cancer patients via machine learning algorithms. *Cancer Med* 2020; **9**:1419−29.
62. Dercle L, Lu L, Schwartz LH, Qian M, Tejpar S, Eggleton P, et al. Radiomics response signature for identification of metastatic colorectal cancer sensitive to therapies targeting EGFR pathway. *J Natl Cancer Inst* 2020;**112**:902−12.
63. Ruffalo M, Thomas R, Chen J, Lee AV, Oesterreich S, Bar-Joseph Z. Network-guided prediction of aromatase inhibitor response in breast cancer. *PLoS Comput Biol* 2019;**15**:e1006730.
64. Herold T, Jurinovic V, Batcha AMN, Bamopoulos SA, Rothenberg-Thurley M, Ksienzyk B, et al. A 29-gene and cytogenetic score for the prediction of resistance to induction treatment in acute myeloid leukemia. *Haematologica (Roma)* 2018;**103**:456−65.
65. Gampenrieder SP, Rinnerthaler G, Hackl H, Pulverer W, Weinhaeusel A, Ilic S, et al. DNA methylation signatures predicting bevacizumab efficacy in metastatic breast cancer. *Theranostics* 2018; **8**:2278−88.
66. Guo L, Wang W, Xie X, Wang S, Zhang Y. Machine learning-based models for genomic predicting neoadjuvant chemotherapeutic sensitivity in cervical cancer. *Biomed Pharmacother* 2023;**159**:114256.
67. Nasimian A, Al Ashiri L, Ahmed M, Duan H, Zhang X, Rönnstrand L, et al. A receptor tyrosine kinase inhibitor sensitivity prediction model identifies AXL dependency in leukemia. *Int J Mol Sci* 2023;**24**:3830.
68. Kanda T, Wakiya T, Ishido K, Kimura N, Nagase H, Yoshida E, et al. Noninvasive computed tomography-based deep learning model

predicts *in vitro* chemosensitivity assay results in pancreatic cancer. *Pancreas* 2024;**53**:e55−61.
69. Pomponio R, Tang Q, Mei A, Caron A, Coulibaly B, Theilhaber J, et al. An integrative approach of digital image analysis and transcriptome profiling to explore potential predictive biomarkers for TGFβ blockade therapy. *Acta Pharm Sin B* 2022;**12**:3594−601.
70. Mourragui SMC, Loog M, Vis DJ, Moore K, Manjon AG, van de Wiel MA, et al. Predicting patient response with models trained on cell lines and patient-derived xenografts by nonlinear transfer learning. *Proc Natl Acad Sci U S A* 2021;**118**:e2106682118.
71. Lamb J, Crawford ED, Peck D, Modell JW, Blat IC, Wrobel MJ, et al. The Connectivity Map: using gene-expression signatures to connect small molecules, genes, and disease. *Science* 2006;**313**: 1929−35.
72. Moor J. The dartmouth college artificial intelligence conference: the next fifty years. *AI Mag* 2006;**27**:87−91.
73. Samuel AL. Some studies in machine learning using the game of checkers. *IBM J Res Dev* 1959;**3**:210−29.
74. Jiang Y, Liu W, Li T, Hu Y, Chen S, Xi S, et al. Prognostic and predictive value of p21-activated kinase 6 associated support vector machine classifier in gastric cancer treated by 5-fluorouracil/oxaliplatin chemotherapy. *EBioMedicine* 2017;**22**:78−88.
75. Yang W, Soares J, Greninger P, Edelman EJ, Lightfoot H, Forbes S, et al. Genomics of Drug Sensitivity in Cancer (GDSC): a resource for therapeutic biomarker discovery in cancer cells. *Nucleic Acids Res* 2013;**41**:D955−61.
76. Barretina J, Caponigro G, Stransky N, Venkatesan K, Margolin AA, Kim S, et al. The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature* 2012;**483**:603−7.
77. Basu A, Bodycombe NE, Cheah JH, Price EV, Liu K, Schaefer GI, et al. An interactive resource to identify cancer genetic and lineage dependencies targeted by small molecules. *Cell* 2013;**154**:1151−61.
78. Shoemaker RH. The NCI60 human tumour cell line anticancer drug screen. *Nat Rev Cancer* 2006;**6**:813−23.
79. Subramanian A, Narayan R, Corsello SM, Peck DD, Natoli TE, Lu X, et al. A next generation connectivity map: L1000 platform and the first 1,000,000 profiles. *Cell* 2017;**171**:1437−52.e17.
80. Koleti A, Terryn R, Stathias V, Chung C, Cooper DJ, Turner JP, et al. Data Portal for the Library of Integrated Network-based Cellular Signatures (LINCS) program: integrated access to diverse large-scale cellular perturbation response data. *Nucleic Acids Res* 2018;**46**: D558−66.
81. Sinha S, Vegesna R, Mukherjee S, Kammula AV, Dhruba SR, Wu W, et al. PERCEPTION predicts patient response and resistance to treatment using single-cell transcriptomics of their tumors. *Nat Cancer* 2024;**5**:938−52.
82. Karampuri A, Jakkula BK, Perugu S. ResisenseNet hybrid neural network model for predicting drug sensitivity and repurposing in breast Cancer. *Sci Rep* 2024;**14**:23949.
83. Lo Vercio L, Amador K, Bannister JJ, Crites S, Gutierrez A, MacDonald ME, et al. Supervised machine learning tools: a tutorial for clinicians. *J Neural Eng* 2020;**17**:abbff2.
84. Jiang T, Gradus JL, Rosellini AJ. Supervised machine learning: a brief primer. *Behav Ther* 2020;**51**:675−87.
85. Azur MJ, Stuart EA, Frangakis C, Leaf PJ. Multiple imputation by chained equations: what is it and how does it work?. *Int J Methods Psychiatr Res* 2011;**20**:40−9.
86. Wang XD, Reeves K, Luo FR, Xu LA, Lee F, Clark E, et al. Identification of candidate predictive and surrogate molecular markers for dasatinib in prostate cancer: rationale for patient selection and efficacy monitoring. *Genome Biol* 2007;**8**:R255.
87. Gottesman MM, Fojo T, Bates SE. Multidrug resistance in cancer: role of ATP-dependent transporters. *Nat Rev Cancer* 2002;**2**:48−58.
88. Zhong L, Ma CY, Zhang H, Yang LJ, Wan HL, Xie QQ, et al. A prediction model of substrates and non-substrates of breast cancer resistance protein (BCRP) developed by GA-CG-SVM method. *Comput Biol Med* 2011;**41**:1006−13.
89. Gantner ME, Di Ianni ME, Ruiz ME, Talevi A, Bruno-Blanch LE. Development of conformation independent computational models for the early recognition of breast cancer resistance protein substrates. *BioMed Res Int* 2013;**2013**:863592.
90. Estrada-Tejedor R, Ecker GF. Predicting drug resistance related to ABC transporters using unsupervised Consensus Self-Organizing Maps. *Sci Rep* 2018;**8**:6803.
91. Deng L, Fan Z, Xu H, Yu S. PDSM-LGCN: prediction of drug sensitivity associated microRNAs *via* light graph convolution neural network. *Methods* 2022;**205**:106−13.
92. Han M, Dai J, Zhang Y, Lin Q, Jiang M, Xu X, et al. Support vector machines coupled with proteomics approaches for detecting biomarkers predicting chemotherapy resistance in small cell lung cancer. *Oncol Rep* 2012;**28**:2233−8.
93. Shannon NB, Tan LLY, Tan QX, Tan JW, Hendrikson J, Ng WH, et al. A machine learning approach to identify predictive molecular markers for cisplatin chemosensitivity following surgical resection in ovarian cancer. *Sci Rep* 2021;**11**:16829.
94. Wu Y, Han W, Xu D, Wang X, Yang J, Lu Z, et al. Identification of subtype specific biomarkers of clear cell renal cell carcinoma using random forest and greedy algorithm. *Biosystems (Amst)* 2021;**204**: 104372.
95. Kim ST, Sa JK, Oh SY, Kim K, Hong JY, Kang WK, et al. Comprehensive molecular characterization of gastric cancer patients from phase II second-line ramucirumab plus paclitaxel therapy trial. *Genome Med* 2021;**13**:11.
96. Lin PC, Yeh YM, Chan RH, Lin BW, Chen PC, Pan CC, et al. Sequential and co-occurring DNA damage response genetic mutations impact survival in stage III colorectal cancer patients receiving adjuvant oxaliplatin-based chemotherapy. *BMC Cancer* 2021;**21**:217.
97. Abramson J, Adler J, Dunger J, Evans R, Green T, Pritzel A, et al. Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature* 2024;**630**:493−500.
98. Ni S, Kong X, Zhang Y, Chen Z, Wang Z, Fu Z, et al. Identifying compound−protein interactions with knowledge graph embedding of perturbation transcriptomics. *Cell Genom* 2024;**4**:100655.
99. Cao D, Chen M, Zhang R, Wang Z, Huang M, Yu J, et al. SurfDock is a surface-informed diffusion generative model for reliable and accurate protein−ligand complex prediction. *Nat Methods* 2025;**22**: 310−22.
100. Tao W, Lin X, Liu Y, Zeng L, Ma T, Cheng N, et al. Bridging chemical structure and conceptual knowledge enables accurate prediction of compound−protein interaction. *BMC Biol* 2024;**22**:248.
101. Li H, Han Z, Sun Y, Wang F, Hu P, Gao Y, et al. CGMega: explainable graph neural network framework with attention mechanisms for cancer gene module dissection. *Nat Commun* 2024;**15**: 5997.
102. Novakovsky G, Fornes O, Saraswat M, Mostafavi S, Wasserman WW. ExplaiNN: interpretable and transparent neural networks for genomics. *Genome Biol* 2023;**24**:154.