Accepted Manuscript

Precision Oncology: Computational methods for multi-omics data integration to improve drug response prediction

Dr. Guna Gouru*

***Corresponding author:** Master of Science in Health Informatics, Department of Health Sciences, Northeastern University, Boston, MA, USA gouru.g@northeastern.edu

Abstract

Cancer heterogeneity presents a major obstacle to effective drug treatment, emphasizing the need for personalized approaches that can accurately predict drug responses. Advances in high-throughput technologies have driven precision medicine initiatives toward integrating multi-omics data, enabling more comprehensive understanding of tumor biology. However, integration of diverse omics layers poses challenges for computational modeling, as many traditional machine learning and statistical methods are not designed to capture complex, high-dimensional, and multi-modal data.

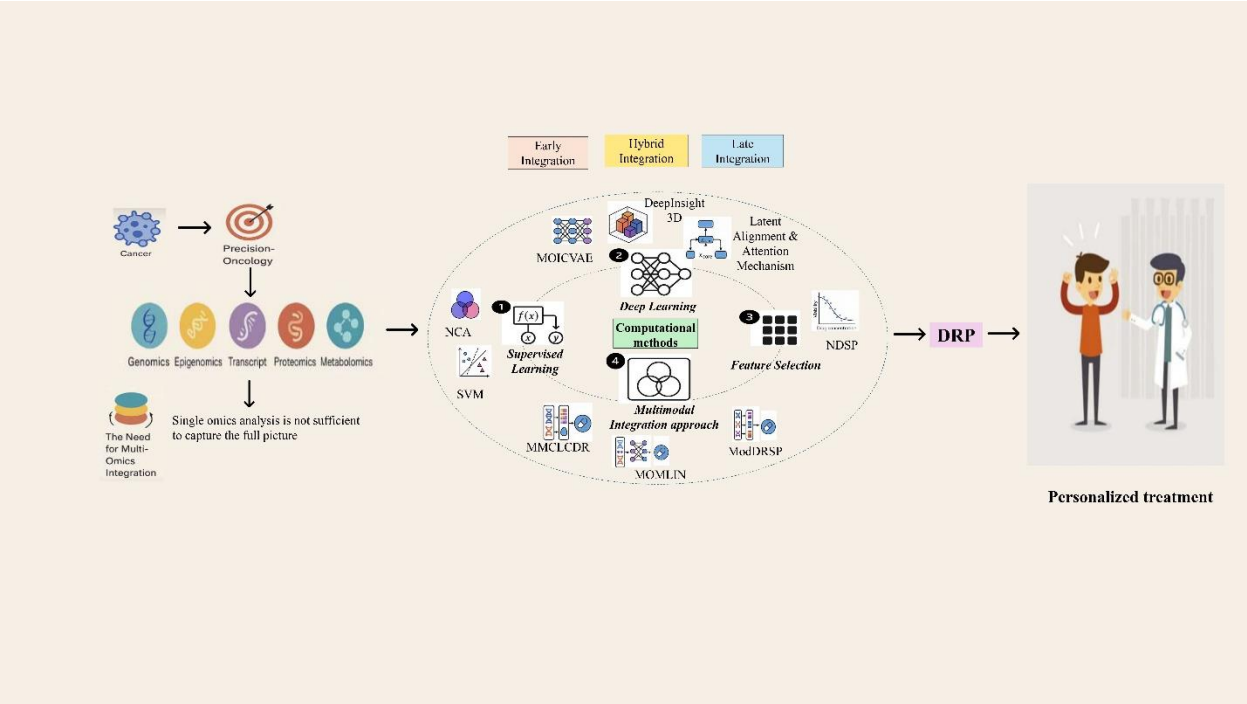
This review examines the studies that integrate multi-omics datasets, aiming to enhance Drug Response Prediction (DRP). Specifically, it outlines the most used omics types and computational approaches- classical machine learning model as well as advanced deep learning and multi-modal integration frameworks for improving DRP, detailing key methodologies and evaluation metrics such as AUC, F1 Score, and MSE, which assess model performance.

By summarizing the integrated omics data, computational methods, and challenges encountered, this review provides an in-depth overview of the existing landscape of precision medicine and future directions for advancing drug response prediction.

This peer-reviewed article has been accepted for publication but not yet copyedited or typeset, and so may be subject to change during the production process. The article is considered published and may be cited using its DOI.

10.1017/pcm.2025.10003

Graphical Abstract:



Keywords: Omics; Drug Response Prediction; Computational approaches; Precision medicine

Impact Statement

Drug sensitivity assessment is critical for optimizing personalized cancer treatment and advancing precision oncology. This review aims to provide a comprehensive and in-depth examination of computational methods used for multi-omics data integration to predict drug response in cancer patients, highlighting the challenges in integration of multi-omics. It further investigates how combining genomic, transcriptomic, epigenomic, proteomic, and metabolomic data can improve the understanding of tumor heterogeneity and enhance predictive accuracy. This integrated approach may significantly contribute to the development of more effective, tailored therapies in cancer treatment by advancing precision medicine.

1. Introduction and Background

Cancer is a highly complex and widespread disease, continuing to be a major global cause of mortality and a serious challenge to human health (Li et al. 2023). A key challenge in cancer treatment is cancer heterogeneity, the genetic and molecular variations among tumors, even within the same cancer type (Chen et al. 2024). This diversity complicates drug selection, as patients may respond differently to the same treatment (Chen et al. 2024).

Assessing drug sensitivity is essential for evaluating how well a drug works in individual patients, guiding treatment decisions. Therefore, developing predictive models for drug sensitivity plays a crucial role in advancing personalized medicine. Cancer is predominantly driven by genetic

https://doi.org/10.1017/pcm.2025.10003 Published online by Cambridge University Press

alterations, where variations in gene expression profiles and somatic mutations play a critical role in modulating therapeutic responses. Elucidating these molecular changes offers the potential to optimize treatment strategies and enhance efficacy at the individual patient level.(Wang et al. 2023).

Precision Oncology, a branch of Precision medicine, aims to deliver the most effective cancer treatment by tailoring the right therapy to the right patient, at the optimal dose and timing (Michele Araújo Pereira 2020; Schwartzberg et al. 2017). Various processes can disrupt genetic machinery at the DNA, RNA, or protein levels, resulting in changes to the expression of the protein encoded by the gene (Schwartzberg et al. 2017). To address this complexity, precision oncology uses high-throughput "-omics" technologies to capture cellular, molecular, and tissue level variations enabling more accurate predictions (Zhao et al. 2023). Achieving the objectives of precision oncology requires more comprehensive profiling of tumors at multiple biological layers. "Omics" encompasses a wide range of biological data types, including genomic information such as Copy number variations (CNV), Mutations and Single Nucleotide Polymorphisms (SNP's); epigenomic data such as DNA methylation; transcriptomic data such as mRNA sequencing; proteomic data such as proteins; lipid-related data (lipidomics) and metabolic compounds (metabolomics) (Llinas-Bertran et al. 2025). However, relying on a single layer of omics data is insufficient to establish precise connections between molecular changes and their phenotypic effects. Thus, the integration of multiple omics is essential to achieve a more complete understanding of cancer heterogeneity (Llinas-Bertran et al. 2025).

This literature review summarizes the type of omics used, and available data integration (DI) methods utilized for improving drug response prediction (DRP) aiming to address the following questions:

- What type of omics are used in selected studies for DRP?
- What methods and categories are utilized to improve DRP?
- What are some key challenges and future directions in improving DRP?

2. Methodology

This review aimed at a single objective: to evaluate computational methods used for multi-omics data integration in drug response prediction (DRP). The methodology was structured in two phases: (i) establishing a systematic search with a clearly defined initial query, and (ii) refining and expanding the search to capture additional relevant studies. The original PubMed query was: "multi-omics" OR "multiomics" OR "omics integration" AND "drug response" OR "drug sensitivity" OR "drug resistance" OR "drug efficacy" AND cancer OR neoplasm OR malignancy. This query was designed to identify studies published between 2021 and 2025 that combined at least two omics data types, or one omics type with other relevant modalities (e.g., clinical, imaging, chemical data), specifically in the context of predicting drug response in cancer.

Studies were excluded if they:

- Did not include at least two distinct omics data types, or one omics data type combined with other relevant data (e.g., clinical, imaging, or chemical data)
- Did not apply an explicit computational or statistical method for integrating omics or other relevant data types.

https://doi.org/10.1017/pcm.2025.10003 Published online by Cambridge University Press

The initial PubMed search returned 685 studies, of which 342 remained after applying the exclusion criteria. In the second phase, the focus was narrowed to studies describing or evaluating computational approaches for DRP. To ensure broader coverage of emerging methodologies, the search was refined using additional targeted queries:

1. "multi-omics" AND "data integration" AND "cancer" AND "drug response"
2. "multi-omics" AND "machine learning" AND "drug sensitivity"
3. "multi-omics" AND "deep learning" AND "drug response"

These supplementary searches identified 27 additional studies, of which 9 met the inclusion criteria by explicitly employing a computational method for DRP.

The final selected studies meeting the inclusion criteria were assessed based on:

- The study objective
- The omics data types that are integrated (Genomics, Epigenomics, Transcriptomics, Proteomics, Metabolomics, Lipidomics).

Figure 1 provides an overview of the methodological framework employed in this literature review. Additionally, several relevant articles ((Zitnik et al. 2019);(Chen and Zhang 2022); (Hasin et al. 2017); (Chakraborty et al. 2024)) were identified, though they did not align directly with the study's primary objective.

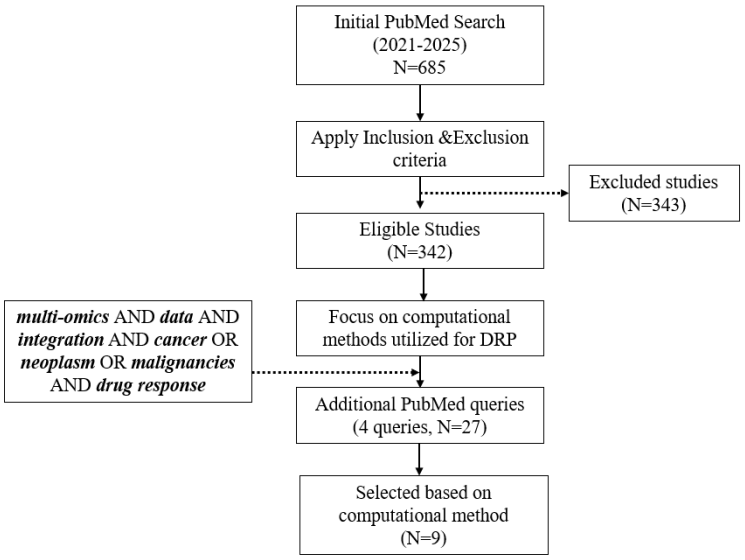


Fig 1 Framework illustrating the methodological approach of the review

3. Results

Drug response prediction (DRP) aims to predict how a drug affects the biological system, such as patient cells or model cell lines. Individual responses can differ significantly between individuals due to multiple factors, often due to genetic and molecular variations (Xiao et al. 2025). To gain more precise understanding of these effects, researchers analyze multi-omics datasets from both patient samples and cell lines. In this context, "drug response" refers to the range of biological or clinical outcomes observed following drug administration, which may include changes in cell viability, proliferation, or clinical efficacy, depending on the model system and study design. This

https://doi.org/10.1017/pcm.2025.10003 Published online by Cambridge University Press

definition highlights that drug response is a multifaceted, quantitative phenotype, shaped by numerous genetic, molecular, and environmental factors, and can differ substantially when measured in cell lines versus patient samples (Zhang and Nebert 2017). A key aspect of advancing personalized medicine is the ability to predict drug effectiveness for a group of patients who share similar molecular profiles (Hernandez-Lemus and Ochoa 2024). This approach aims to tailor treatments more accurately to individual patient's requirements, potentially enhancing outcomes and reducing adverse effects (Athieniti and Spyrou 2023).

3.1 Predominant omics used for DRP

This analysis was performed using R software, centers on determining the most utilized omic layers in selected studies. As shown in Figure 2a, transcriptomics emerged as the most frequently applied omic type, followed by genomics, epigenomics, and proteomics. Next, the bilateral pairings of the omic types found in the selected studies are examined. Figure 2b illustrates prominent combinations of multiple omics types utilized within individual studies.The network diagram illustrates the most frequently combined omic layers. Each node represents an individual omic type, with the size of the node corresponding to how often that omic layer was utilized across the selected studies. Line between nodes indicate that the two omic types were used together in at least one study, and the number on each line represents the frequency of that specific combination.

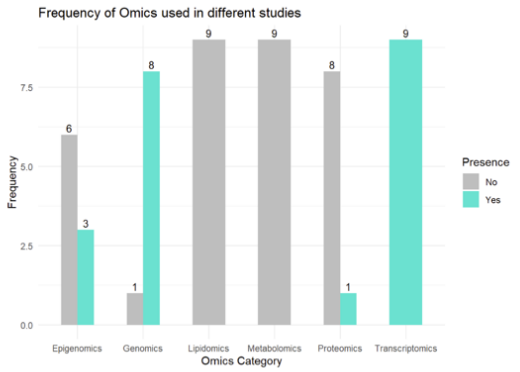


Fig 2a Representation of frequently used omics in selected studies

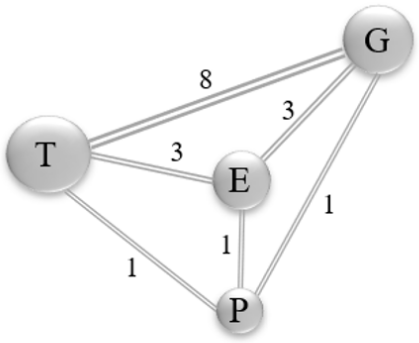


Fig 2b Co-occurrence of omic pairs in selected studies A*bbreviations.*
T: Transcriptomics **G:** Genomics
E: Epigenomics **P:** Proteomics

3.2 Computational methods used for DRP

Computational methods used for drug response prediction can be broadly categorized as under either classification or regression tasks, depending on the nature of the predicted outcome. Classification models assign samples to discrete categories such as 'sensitive' or 'resistant', based on thresholds applied to measured drug response values. Support Vector Machine (SVM) and Neighborhood Component Analysis (NCA) fall under this category. In contrast, regression models

https://doi.org/10.1017/pcm.2025.10003 Published online by Cambridge University Press

predict continuous quantitative measures such as IC50 (the concentration of drug needed to inhibit cell viability by 50%), AUC (area under the dose-response curve), or cell viability percentages. Multi-Omics Integrated Collective Variational Autoencoders (MOICVAE), Latent Alignment and Attention Mechanism, DeepInsight-3D, Novel Drug Sensitivity Prediction (NDSP), Multimodal Contrastive Learning for Cancer Drug Responses (MMCLCDR), Multi-modal and Omics Machine Learning Integration (MOMLIN), and Multi-stage Multi-modal Drug Representations (ModDRDSP) employ regression as their primary framework, directly modeling the continuous landscape of pharmacological response (Partin et al. 2023).

A key step in applying these computational methods is multi-omics data integration, which can be performed using early, late, or intermediate strategies. Early integration combines multiple omics datasets into a single table or graph format, which is then processed by a machine learning model. Late integration analyzes each omics layer independently and subsequently merges the individual predictions through an additional model. Intermediate integration allows the model to learn shared representations from multiple datasets simultaneously (Athieniti and Spyrou 2023).

Figure 3 displays a Sankey diagram illustrating the distribution of omics types across the reviewed studies, with the dots of each flow corresponding to the number of omics employed per study. Figure 4 illustrates the overall workflow for predicting drug response outcome. The effectiveness of these methods is evaluated based on their ability to accurately predict drug responses, with evaluation metrics for assessment (Jiang et al. 2025). Table 1 presents a summary of the selected studies, detailing the types of omics data integrated, the computational frameworks and algorithms employed, the datasets utilized, and the evaluation metrics applied. To further assist researchers in selecting appropriate datasets, a comparative summary of commonly used benchmark datasets is provided below Table 1a. A well-performing model should not only provide high prediction accuracy but also reveal biologically interpretable insights, such as the molecular mechanisms driving drug resistance or sensitivity.

3.2.1 Supervised Learning

SVM classify patients based on the response to the treatment. In these, patients were classified into different groups and a SVM model was trained using data from 153 patients for the feature extraction from gene expression and Immunohistochemistry data and classified new patients into Response or Resistance groups. A hyperplane was created to separate patients into two categories: Responders (those who benefited from the treatment) and non-responders (resistant) (Che et al. 2024).

NCA is a deep learning-based DRP model using multi-omics data to analyze breast cancer (BRCA) cell lines. After filtering poorly performing drugs, a dataset of 42 cell lines and 100 drug molecules was used. The model utilizes NCA for feature selection and a neural network regressor with Levenberg-Marquardt backpropagation for training, optimized with Bayesian optimization and 5-fold cross-validation. Additionally, K-means clustering was applied to categorize drugs in Olaparib and non-responders for the drugs such as Dabrafenib and Olaparib, and identifying outliers that negatively impacted model accuracy (Malik et al. 2021; Ruiz-Ramos et al. 2025).

https://doi.org/10.1017/pcm.2025.10003 Published online by Cambridge University Press

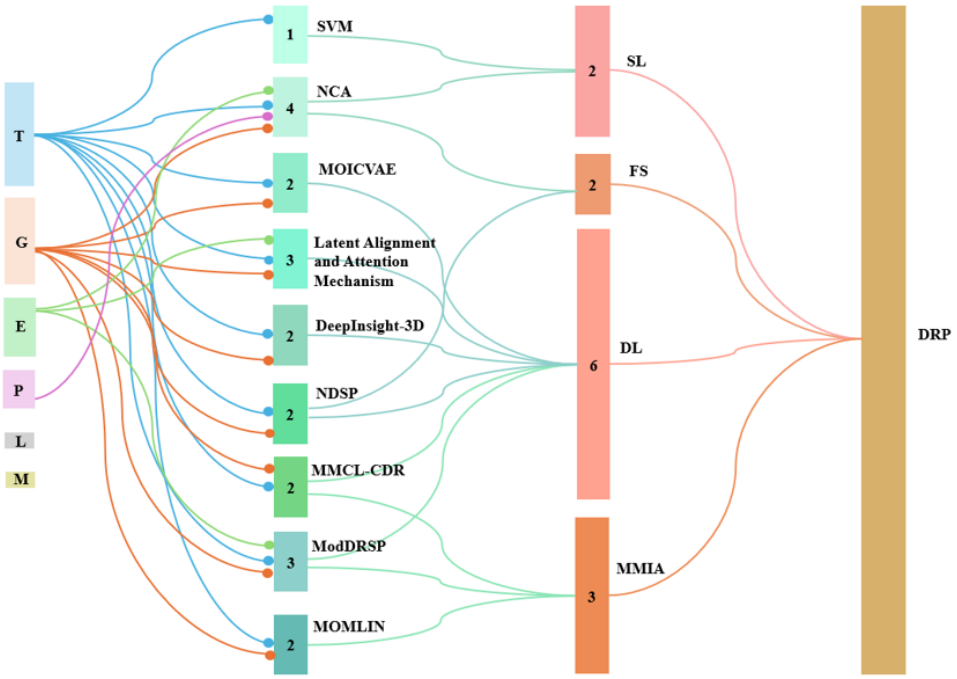


Fig 3 Sankey chart representing the omics and associated computational methods used in different studies. *Abbreviations.* **SL:** Supervised Learning **DL:** Deep Learning **FS:** Feature Selection **MMIA:** Multi modal Integration approach

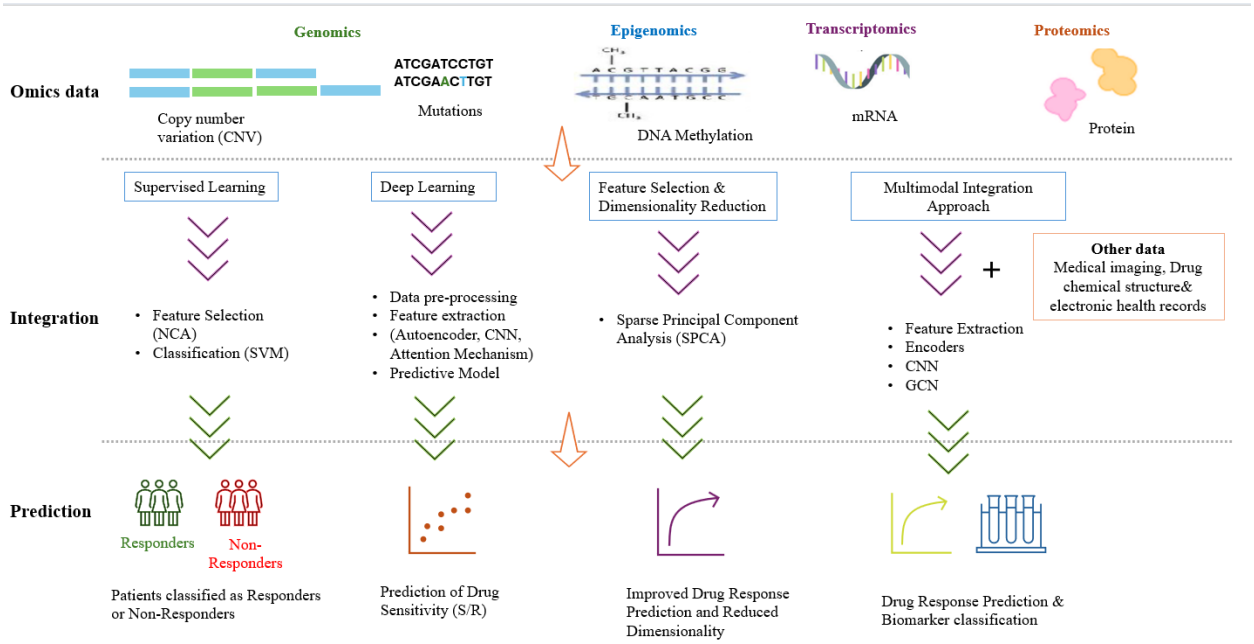


Fig 4 Multi-omics data integration Workflow for Drug Response Prediction

https://doi.org/10.1017/pcm.2025.10003 Published online by Cambridge University Press

Table 1 Computational methods used for multi-omics data integration, identified studies published in 2021-2025, including their method category, application objectives, datasets employed, evaluation metric and whether a specific cancer type was investigated. ***Abbreviations***. **GDSC:** Genomics of Drug Sensitivity in Cancer **DMSZ:** Leibniz Institute **CCLE:** Cancer Cell Line Encyclopedia **TCGA:** The Cancer Genome Atlas **GEO:** Gene Expression Omnibus **BC:** Breast Cancer **GC:** Gastric Cancer **AUC/AUROC:** Area Under Curve/ Area under Receiver Operating Characteristic Curve **AUPR:** Area Under Precision Recall Curve **MSE:** Mean Square Error **MAE:** Mean Absolute Error **RMSE:** Root Mean Square Error **CNV:** Copy number Variation **mRNA**: Ribo Nucleic acid **miRNA:** Micro Ribo Nucleic Acid **SNP:** Single Nucleotide Polymorphism **DNA:** Deoxyribose Nucleic Acid **CNA:** Copy Number Aberration **GCN:** Graph Convolutional Network **CNN:** Convolutional Neural Network **DMCN:** Deep message-crossing network **SMILES** Simplified molecular input line entry system **CVAE:** Conditional Variational Autoencoder **KNN:** K-Nearest Neighbor

Framework or Study name	Generic model(s) used	Modelling category	Predicted outcome	Data Modalities (omics + other)	Drug data & Integration strategy	Feature engineering	Evaluation metrics	Algorithm details	Primary database
Che et al. 2024	SVM	Supervised ML	Binary Classification (Sensitive vs Resistant)	**T:** mRNA sequencing + IHC	None (Early)	Feature extraction from gene expression vectors	AUC, F1 Score	Linear/Kernel SVM, patient level hyperplane separation, cross-validation	Direct Patient
Malik et al. 2021	NCA	Feature selection +Supervised ML	Classification (Responder vs non-responder)	**G:** CNV, Mutations, **T:** miRNA, **P:** Protein **E:** Methylation	None (Early)	NCA-based dimensionality reduction; Bayesian optimization	AUC	Neighborhood component analysis+ neural network regressor with k means clustering for drug classes	GDSC, TCGA (BRCA focus)
MOICVAE (Wang et al. 2023)	Autoencoder, CVAE, KNN	Deep learning	Binary classification (Sensitive/ Resistant) + survival regression	**G:** CNV, SNP, **T:** Gene expression + Clinical features(fro	None (Hybrid)	Multi-modal deep autoencoder+ CVAE latent fusion vector	AUC, Precision, Accuracy	Autoencoder fusion, CVAE classification, post-hoc KNN, Kaplan Meier	GDSC, CCLE, cBioPortal, METABRIC

https://doi.org/10.1017/pcm.2025.10003 Published online by Cambridge University Press

				m TCGA & BC data)				survival & immune marker analysis	
Latent Alignment and Attention Mechanism (Chen et al.2024)	DNN+ Attention	Deep learning	Regression (IC50/ Viability curves)	**G:** CNV, Mutations, **T:** Gene expression, **E:** DNA methylation	None (Hybrid)	Latent space alignment, imputation for missing features, attention module	MSE, F1 Score, AUROC	Multi modal alignment in shared latent space, feature attention weighing, deep neural predictor	DepMap
DeepInsight-3D (Sharma et al. 2023)	CNN	Deep learning	Classification + regression (drug response, IC50)	**G:** Somatic Mutations, **T:** Gene expression	None (Early)	Omics transformed into 3D image tensors, dimensionality via pixel mapping	AUC	CNN image-encoded omics, Class activation for gene/ pathway interpretability	GDSC, CCLE
NDSP (Liu & Mei 2023)	SPCA, Neural network	Feature selection + Deep learning	Regression (drug sensitivity)	**G:** CNV, Mutations, **T:** RNA sequencing	None (Hybrid)	SPCA, Similarity network fusion	Sensitivity, Specificity, Precision, Accuracy, F1 Score	Similarity network fusion, deep neural regressor	GDSC, EMBL-EBI, Cell Model Passports, GEO
MMCL CDR (Li et al. 2023)	Encoder, CNN, GCN	Multimodal integration +Deep learning	Regression (IC50/AUC prediction)	**G:** CNV, **T:** Gene expression + Morphological images(cell lines)	SMILES molecular graphs (GCN) (Hybrid)	Encoders for each modality, feature projection, contrasive learning	AUC, AUPR	CNN (image features), Encoder (omics), GCN (drug features), feature fusion via	GDSC, DMSZ, PubChem

https://doi.org/10.1017/pcm.2025.10003 Published online by Cambridge University Press

								contrastive learning	
MOMLIN (Rasid et al. 2024)	WMSCCA, Logistic regression	Multimodal integration	Regression + Classification (predict drug response, build biomarker networks)	**G:** Mutations, **T:** Gene expression + Clinical features, Pthway activity	None (Late)	Weighted Multi-Class Sparse Canonical Correlation (WMSCCA)	AUC	Multi-stage: WMSCCA-logistic regression classifier-biomarker network visualization	BC
ModDRDSP (Song et al. 2025)	Bi-GRU, DMCN, CellCNN	Multimodal integration + Deep learning + GNN	Regression (drug IC50, AUC prediction)	G: CNV, CNA , T: Gene expression, E: DNA methylation	SMILES (graph features), Drug Sensitivity assays (Hybrid)	KernelPCA, ConvMolFeaturizer, CellCNN	MSE, MAE, RMSE	Deep hierarchical bi-GRU (SMILES), DMCN (graph embedding), CellCNN (omics features); fused in ensemble	GDSC, CCLE, PubChem

Table 1a Overview of Benchmark Datasets for Drug Response Prediction

Dataset	Description	Strengths	Limitations
GDSC	Drug Sensitivity across cancer cell lines	Large-scale, including genomic profiles	Limited patient relevance
TCGA	Multi-omics data from patient tumors	Rich clinical context	No drug response data
CCLE	Genomic profiles of cell lines	Widely used, integrates with other datasets	Limited data coverage
DepMap	Functional genomics+ drug sensitivity	CRISPR/RNAi screens; target validation	Cell-line based

https://doi.org/10.1017/pcm.2025.10003 Published online by Cambridge University Press

3.2.2 Deep Learning

MOICVAE is designed to predict drug sensitivity in cancer patients by integrating multi-omics data. The datasets are processed into similar matrices to establish relationships between samples and omics features. The framework incorporates a Multi-Modal Deep Autoencoder (MDA) to fuse omics data, generating a fusion vector that encapsulates essential multi-omics patterns. This fusion vector is then passed into a Conditional Variational Autoencoder (CVAE), which encodes and reconstructs data representations to classify samples as Sensitive (S) or Resistant (R) to a given drug. The trained MOICVAE model is applied to TCGA (The Cancer Genome Atlas) samples, where it utilizes mRNA+ CNV data and combines with K-Nearest Neighbors (KNN) to predict drug response. Samples are categorized into S or R groups, indicating their likelihood of responding to treatment. Further differential analysis is conducted to assess the biological significance of these predictions. Kaplan-Meier survival plots compare overall survival between S and R groups, while tumor inflammatory scores highlight differences in tumor immune responses. Additionally, immune checkpoint markers such as HAVCR2, LAG3, TIGIT, and PDCD1 are analyzed to explore variations in the immune microenvironment, providing insights into the role of immune factors in drug resistance and sensitivity (Wang et al. 2023).

Latent Alignment and Attention Mechanism is another deep learning model which utilizes multiple biological data types to improve predictive accuracy. The framework begins with data preprocessing, where raw multi-omics data undergoes steps, such as overlapping, filtering, and imputation to manage missing values and ensure consistency across datasets. Following this, feature extraction is performed separately for each omics data type to capture key biological patterns effectively. The extracted features are then aligned in a latent space, preserving relationships between different data modalities while reducing dimensionality. To enhance predictive performance, an attention module is incorporated, allowing the model to focus on the most informative features that contribute to drug response. Finally, a predictive model, likely utilizing a deep neural network, processes the refined feature representations to estimate drug viability across different concentrations. The model outputs a dose-response curve, identifying whether a cancer sample is sensitive or resistant to a particular drug (Chen et al. 2024).

DeepInsight-3D is an advanced deep learning model designed to enhance the analysis of multi-omics data, particularly for predicting anticancer drug responses. It transforms multi-layered omics data into 3D image formats, enabling convolutional neural networks (CNNs) to effectively process and extract features. Unlike its predecessor, which handled single-layer data, DeepInsight-3D can accommodate multiple omics layers, improving the complexity and depth of models. The method includes two image construction strategies, either prioritizing the most informative layer or combining all layers equally. CNNs are then applied to identify patterns and classify the data, with an element decoder used to provide biologically relevant insights. This model excels in handling small sample sizes and offers interpretability through class-activation maps, which pinpoint crucial genes or pathways, making it a powerful tool for understanding drug responses in cancer treatment (Sharma et al. 2023).

3.2.3 Feature Selection

NDSP is one model that integrates multi-omics data to enhance DRP by utilizing deep learning and similarity network fusion approaches. The method begins by extracting drug targets using an improved sparse principal component analysis (SPCA) for different omics data types, such as RNA sequencing, CNA, and methylation. These extracted features are then used to construct sample

https://doi.org/10.1017/pcm.2025.10003 Published online by Cambridge University Press

similarity networks, which are subsequently merged to create a comprehensive representation of the data. The merged similarity networks are input into a deep neural network for training, reducing dimensionality and mitigating overfitting issues improving interpretability and accuracy in predicting drug sensitivity (Liu and Mei 2023).

3.2.4 Multi modal Integration approach

Multimodal data integration refers to the process of combining diverse data types from different sources to generate a comprehensive understanding of a subject (Hernandez-Lemus and Ochoa 2024). In this initially data is collected from different sources such as multi-omics data, medical imaging (histopathology, morphological, radiology), Drug chemical structure and electronic health records (EHRs). Then data preprocessing is done before the integration. Advanced computational models such as Machine Learning (ML), Artificial Intelligence (AI) integrate these diverse datasets. This integration allows for advanced data analysis to achieve key objectives ultimately aiding in personalized medicine and improved healthcare outcomes.(Llinas-Bertran et al. 2025)

MMCL-CDR is designed to enhance DRP by integrating multiple data modalities, including multi-omics, morphological images and chemical structure of the drug. In this model, for multi-omics, an Encoder (a neural network) that transforms high dimensional input into lower dimensional features is used. For morphological images of cells, a CNN is specifically used to extract features from morphology images. The extracted multi-omics and morphology images are passed through a projector for further transformation. An aggregation step combines both omics and image representations. For Molecular drug data, a Graph Convolutional Network (GCN) processes these molecular graphs to learn meaningful drug representations then Max Pooling is applied to summarize the extracted features. All the omics representation and Image representation are refined using Contrastive Learning, ensuring biologically relevant alignment. A fully connected neural network predicts whether a cancer cell line is resistant or sensitive to a particular drug based on the learned features (Li et al. 2023).

MOMLIN is a multi-modal framework integrating clinical features, multi-omics, and pathway activity to predict drug response and identify biomarker networks in cancer patients. It follows a three-stage process: First, Weighted Multi-Class Sparse Canonical Correlation Analysis (WMSCCA) selects sparse latent components that capture key predictive features. Next, a logistic regression model is trained using these components, by incorporating multi-omics data. Finally, biomarker networks are constructed through heatmap visualization and correlation analysis, revealing critical molecular signatures driving treatment response (Rashid and Selvarajoo 2024).

ModDRDSP is another tool that integrates multi-omics data, drug molecular structure to predict drug sensitivity response in cancer cell lines utilizing drug sensitivity, multi-omics and drug molecular structure data. Preprocessing involves KernelPCA for dimensionality reduction and ConvMolFeaturizer for molecular feature extraction. The framework employs deep learning models: deep hierarchical bi-directional GRU network (DSBiGRU) processes Simplified molecular input line entry system (SMILES) representations, a deep message-crossing network (DMCN) learns molecular graph embeddings, and CellCNN extracts multi-omics features (Han et al. 2023). These features are fused into multi-dimensional representation and analyzed using machine learning models to predict drug sensitivity (Song et al. 2025).

https://doi.org/10.1017/pcm.2025.10003 Published online by Cambridge University Press

4. Discussion

4.1 Omics utilization in DRP

This review revealed distinct trends in omics usage and computational methodologies for drug response prediction (DRP), with a predominant reliance on genomics, epigenomics, and transcriptomics, while proteomics remains underrepresented despite its potential for providing functional and dynamic insights into tumor biology. This imbalance likely reflects differences in data availability and technological maturity; however, the sparing use of proteomic and metabolomic data limits the ability to capture post-translational modifications and metabolic shifts essential for accurate drug response modeling. Addressing this gap represents a key opportunity for future research.

4.2 Computational methodologies

In parallel with data considerations, computational strategies for DRP continue to evolve. By categorizing existing approaches into supervised learning, deep learning, feature selection with dimensionality reduction, and multimodal integration, this review highlights both their complementary strengths and shared limitations.

Traditional supervised learning methods such as Support Vector Machines (SVM) and Neighborhood Component Analysis (NCA) have been widely applied. Yet, their effectiveness is constrained by the high dimensionality and small sample sizes typical of omics datasets, which predispose models to overfitting and poor generalizability. Moreover, their linear assumptions often fall short of capturing the complex nonlinear interactions that characterize cancer heterogeneity. (Liu and Mei 2023).

Deep learning frameworks like MOICVAE, latent alignment with attention mechanisms, and DeepInsight-3D address some of these challenges by enabling automatic feature extraction and nonlinear data fusion. These models have shown promise in capturing multi-modal biological interactions, but their success depends heavily on access to large, high-quality datasets. Furthermore, their limited interpretability—the "black-box" problem—remains a barrier to clinical adoption, as clinicians require mechanistic insights alongside predictive accuracy. (Liu and Mei 2023).

Feature selection and dimensionality reduction approaches such as sparse principal component analysis (SPCA) used in models like NDSP help mitigate issues of high dimensionality and noise. However, these approaches risk eliminating subtle but biologically meaningful signals, underscoring the ongoing challenge of balancing feature sparsity with biological relevance. Missing or incomplete data across omics layers further complicates integration, necessitating robust imputation and harmonization strategies to minimize bias and information loss.

Multimodal integration frameworks, such as MMCL-CDR, MOMLIN, and ModDRDSP, represent an important step toward holistic modeling by combining diverse data types, including clinical variables, imaging features, chemical drug structures, and multi-omics layers. While these approaches enhance the potential for comprehensive prediction, they face persistent hurdles in aligning heterogeneous data with different scales, distributions, and levels of reliability. Their complex architecture also demands extensive parameter tuning and large sample sizes to ensure stability. Moreover, model interpretability and prospective clinical validation remain underdeveloped, slowing their translational impact.

https://doi.org/10.1017/pcm.2025.10003 Published online by Cambridge University Press

4.3 Key Challenges to Clinical Translation

Taken together, the models, integration techniques, and methodological innovations reviewed here illustrate the transformative potential of informatics-driven approaches for optimizing therapeutic outcomes and advancing precision medicine (Shekhawat et al. 2025). At the same time, several fundamental challenges continue to limit clinical translation.

4.3.1 Data Heterogeneity

Despite considerable progress in multi-omics-based drug response prediction (DRP), several critical challenges hinder the full translation of computational models into clinical practice. Foremost among these is data heterogeneity, arising from the integration of diverse omics layers generated by different platforms, protocols, and laboratories. Variability in sample processing, measurement techniques, and batch effects introduce noise and bias that complicate model training and reduce reproducibility. The complexity increases when combining multi-omics data with auxiliary modalities such as clinical, imaging, or drug molecular information, necessitating rigorous normalization and harmonization strategies that remain an ongoing challenge (Jiang et al. 2025).

4.3.2 Model interpretability

Another prominent limitation is the lack of model interpretability, particularly for advanced deep learning and multimodal integration frameworks. Although powerful in capturing complex nonlinearities and interactions, they often fail to provide transparent biological explanations for their predictions. This opacity poses a barrier to clinical adoption, as clinicians require not only accurate predictions but also mechanistic insights to make informed therapeutic decisions and to trust model recommendations. Developing explainable AI approaches and embedding biological prior knowledge are therefore critical future directions to enhance model transparency and trustworthiness (Ennab and McHeick 2024).

4.3.3 Generalizability across cohorts

Many established models exhibit poor generalizability across independent cohorts and diverse patient populations. Overfitting to small, homogeneous datasets undermines their robustness in real-world applications, where genetic backgrounds and environmental exposures vary widely. This limitation is exacerbated by the lack of standardized response metrics in preclinical studies and heterogeneity in experimental conditions, such as assay types, drug concentrations, and cell line handling, which complicates comparisons across studies (Adam et al. 2020). Addressing this limitation requires larger, multi-institutional datasets, rigorous cross-cohort validation and privacy-preserving strategies such as federated learning.

5. Future directions

The integration of multi-omics data continues to offer immense potential, with deep learning methods expected to play a central role in uncovering complex nonlinear patterns (Shekhawat et al. 2025). A growing area of interest lies in single-cell multi-omics, which provides opportunities to dissect cellular heterogeneity and better understand molecular mechanisms in disease. Equally important are advances in visualization strategies, which can improve the accessibility and interpretation of results for a broader scientific audience. Establishing standardized repositories

https://doi.org/10.1017/pcm.2025.10003 Published online by Cambridge University Press

and collaborative platforms will be key to ensuring data availability, reproducibility, and more effective cross-study comparisons (Zhang 2024).

Another promising direction is the emphasis on model explainability. Tools such as SHAP values and attention mechanisms can help researchers interpret predictions, increasing trust and clinical relevance. Alongside this, harmonizing data across institutions and platforms remains a critical challenge for ensuring consistency and generalizability of findings. Emerging approaches like tabular-to-image conversion (e.g., DeepInsight) open new possibilities for applying CNN-based models in omics research, combining interpretability with predictive power (Sharma et al. 2019). Together, these advancements will help translate multi-omics research into more actionable insights for personalized medicine.

6. Conclusion

The integration of multi-omics data has significantly advanced drug response prediction, offering deeper insights into tumor heterogeneity and enabling more precise therapeutic strategies. Current approaches demonstrate the value of genomics, epigenomics, and transcriptomics in predictive modeling, but the limited use of proteomics and metabolomics restricts the ability to capture functional dynamics essential for understanding drug sensitivity and resistance. While computational methods from supervised learning to advanced deep learning and multimodal integration have shown considerable promise, they remain challenged by data heterogeneity, interpretability, and limited generalizability across cohorts.

Addressing these challenges will be critical for translating computational advances into clinical practice. Future research should prioritize the development of explainable and generalizable models, the harmonization of multi-institutional datasets, and the integration of underutilized omics layers. Combining methodological innovation with standardized pipelines, robust visualization tools, and collaborative data-sharing infrastructures will accelerate the clinical applicability of multi-omics-driven drug response prediction. These efforts will strengthen the foundations of precision oncology, enabling the delivery of more effective, personalized cancer treatments

Acknowledgements: This review was undertaken as part of the capstone project for the MSc in Health Informatics program at Northeastern University. I would like to express my sincere gratitude to Professor Jay Spitulnik, Program Director, for his valuable guidance and support throughout the development and writing of this review. I also extend my thanks to Lauri Fennell, Health Sciences Librarian, for her assistance in navigating reference management tools, and to Amy Lewontin, Collection Development Librarian, for her invaluable help in identifying Northeastern University's open access publishing agreements.

Author contribution: Guna Gouru: Conceptualization, Methodology, Formal Analysis, Writing-Original Draft Preparation, Review & Editing, Project Administration

Conflict of Interest Statement: The author declares no commercial or financial relationships that could be construed as a potential conflict of interest.

References

Adam G, Rampasek L, Safikhani Z, Smirnov P, Haibe-Kains B and Goldenberg A (2020) Machine learning approaches to drug response prediction: challenges and recent progress. *NPJ Precision Oncology* **4,** 19. https://doi.org/10.1038/s41698-020-0122-1.

Athieniti E and Spyrou GM (2023) A guide to multi-omics data collection and integration for translational medicine. *Computational and Structural Biotechnology Journal* **21,** 134-149. https://doi.org/10.1016/j.csbj.2022.11.050.

Chakraborty S, Sharma G, Karmakar S and Banerjee S (2024) Multi-OMICS approaches in cancer biology: New era in cancer therapy. *Biochimica et Biophysica Acta (BBA) Molecular Basis of Disease* **1870**(5)**,** 167120. https://doi.org/10.1016/j.bbadis.2024.167120.

Che G, Yin J, Wang W, Luo Y, Chen Y, Yu X, Wang H, Liu X, Chen Z, Wang X, Chen Y, Wang X, Tang K, Tang J, Shao W, Wu C, Sheng J, Li Q and Liu J (2024) Circumventing drug resistance in gastric cancer: A spatial multi-omics exploration of chemo and immuno-therapeutic response dynamics. *Drug Resistance Updates* **74,** 101080. https://doi.org/10.1016/j.drup.2024.101080.

Chen HO, Cui YC, Lin PC and Chiang JH (2024) An Innovative Multi-Omics Model Integrating Latent Alignment and Attention Mechanism for Drug Response Prediction. *Journal of Personalized Medicine* **14**(7)**,** 694. https://doi.org/10.3390/jpm14070694.

Chen Y and Zhang L (2022) How much can deep learning improve prediction of the responses to drugs in cancer cell lines? *Briefings in Bioinformatics* **23**(1)**,** 1-8. https://doi.org/10.1093/bib/bbab378.

Ennab M and McHeick H (2024) Enhancing interpretability and accuracy of AI models in healthcare: a comprehensive review on challenges and future directions. *Frontiers in Robotics and AI* **11,** 1444763. https://doi.org/10.3389/frobt.2024.1444763.

Han S, Fu H, Wu Y, Zhao G, Song Z, Huang F, Zhang Z, Liu S and Zhang W (2023) HimGNN: a novel hierarchical molecular graph representation learning framework for property prediction. *Briefings in Bioinformatics* **24**(5)**,** 1-11. https://doi.org/10.1093/bib/bbad305.

Hasin Y, Seldin M and Lusis A (2017) Multi-omics approaches to disease. *Genome Biology* **18,** 83. https://doi.org/10.1186/s13059-017-1215-1.

Hernandez-Lemus E and Ochoa S (2024) Methods for multi-omic data integration in cancer research. *Frontiers in Genetics* **15,** 1425456. https://doi.org/10.3389/fgene.2024.1425456.

Jiang W, Ye W, Tan X and Bao YJ (2025) Network-based multi-omics integrative analysis methods in drug discovery: a systematic review. *BioData Mining* **18,** 27. https://doi.org/10.1186/s13040-025-00442-z.

Li Y, Guo Z, Gao X and Wang G (2023) MMCL-CDR: enhancing cancer drug response prediction with multi-omics and morphology images contrastive representation learning. *Bioinformatics* **39**(12). https://doi.org/10.1093/bioinformatics/btad734.

Liu XY and Mei XY (2023) Prediction of drug sensitivity based on multi-omics data using deep learning and similarity network fusion approaches. *Frontiers in Bioengineering and Biotechnology* **11,** 1156372. https://doi.org/10.3389/fbioe.2023.1156372.

Llinas-Bertran A, Butjosa-Espin M, Barberi V and Seoane JA (2025) Multimodal data integration in early-stage breast cancer. *The Breast* **80,** 103892. https://doi.org/10.1016/j.breast.2025.103892.

https://doi.org/10.1017/pcm.2025.10003 Published online by Cambridge University Press

Malik V, Kalakoti Y and Sundar D (2021) Deep learning assisted multi-omics integration for survival and drug-response prediction in breast cancer. *BMC Genomics* **22**(1)**,** 214. https://doi.org/10.1186/s12864-021-07524-2.

Michele Araújo Pereira MKL, Patrícia Gonçalves Pereira Couto, Michele Groenner Penna, Luige Biciati Alvim, Thaís Furtado Nani, Maíra Cristina Menezes Freire, and Luiz Henrique Araúj (2020) Cancer Genomics in Precision Oncology: Applications, Challenges, and Prospects. In Masood N and Malik SS (eds), *Essentials of Cancer Genomic, Computational Approaches and Precision Medicine.* Singapore: Springer Nature Singapore Pte Ltd, 453-454.

Partin A, Brettin TS, Zhu Y, Narykov O, Clyde A, Overbeek J and Stevens RL (2023) Deep learning methods for drug response prediction in cancer: Predominant and emerging trends. *Frontiers in Medicine* **10,** 1086097. https://doi.org/10.3389/fmed.2023.1086097.

Rashid MM and Selvarajoo K (2024) Advancing drug-response prediction using multi-modal and -omics machine learning integration (MOMLIN): a case study on breast cancer clinical data. *Briefings in Bioinformatics* **25**(4). https://doi.org/10.1093/bib/bbae300.

Ruiz-Ramos J, Plaza-Diaz A, Puig-Campmany M, Sampol-Mayol C, Blazquez-Andion M, Serrano-Garcia-Calvo A, Sanz-Lopez N, Acebes-Roldan X and Juanes-Borrego A (2025) K-means clustering to identify high risk of early revisits in patients with drug-related problems attending the emergency department. *European Journal of Hospital Pharmacy* **0,** 1-6. https://doi.org/10.1136/ejhpharm-2024-004414.

Schwartzberg L, Kim ES, Liu, D. and Schrag D (2017) Precision Oncology: Who, How, What, When, and When Not? *American Society of Clinical Oncology educational book. American Society of Clinical Oncology. Annual Meeting* **37,** 160-179. https://doi.org/10.1200/EDBK_174176.

Sharma A, Lysenko A, Boroevich KA and Tsunoda T (2023) DeepInsight-3D architecture for anti-cancer drug response prediction with deep-learning on multi-omics. *Scientific Reports* **13**(1)**,** 2483. https://doi.org/10.1038/s41598-023-29644-3.

Sharma A, Vans E, Shigemizu D, Boroevich KA and Tsunoda T (2019) DeepInsight: A methodology to transform a non-image data to an image for convolution neural network architecture. *Scientific Reports* **9**(1)**,** 11399. https://doi.org/10.1038/s41598-019-47765-6.

Shekhawat RS, Kashyap P, Swain AK, Dhiman V, Sharma J, Kumar A and Yadav P (2025) Bioinformatics for Multi-Omics Data Integration. *Encyclopedia of Bioinformatics and Computational Biology (Second Edition)* **4,** 615-627. https://doi.org/10.1016/B978-0-323-95502-7.00159-7.

Song J, Wei M, Zhao S, Zhai H, Dai Q and Duan X (2025) Drug Sensitivity Prediction Based on Multi-stage Multi-modal Drug Representation Learning. *Interdisciplinary Sciences: Computational Life Sciences* **17,** 231-243. https://doi.org/10.1007/s12539-024-00668-1.

Wang C, Zhang M, Zhao J, Li B, Xiao X and Zhang Y (2023) The prediction of drug sensitivity by multi-omics fusion reveals the heterogeneity of drug response in pan-cancer. *Computers in Biology and Medicine* **163,** 107220. https://doi.org/10.1016/j.compbiomed.2023.107220.

Xiao M, Zheng Q, Popa P, Mi X, Hu J, Zou F and Zou B (2025) Drug molecular representations for drug response predictions: a comprehensive investigation via machine learning methods. *Scientific Reports* **15**(1)**,** 20. https://doi.org/10.1038/s41598-024-84711-7.

https://doi.org/10.1017/pcm.2025.10003 Published online by Cambridge University Press

Zhang G and Nebert DW (2017) Personalized medicine: Genetic risk prediction of drug response. *Pharmacology & Therapeutics* **175,** 75-90. https://doi.org/10.1016/j.pharmthera.2017.02.036.

Zhang J (2024) Emerging Trends in Multi-Omics Data Integration: Challenges and Future Directions. *Computational Molecular Biology* **14**(2)**,** 64-75. https://doi.org/10.5376/cmb.2024.14.0008.

Zhao Z, Jin T, Chen B, Dong Q, Liu M, Guo J, Song X, Li Y, Chen T, Han H, Liang H and Gu Y (2023) Multi-omics integration analysis unveils heterogeneity in breast cancer at the individual level. *Cell Cycle* **22**(20)**,** 2229-2244. https://doi.org/10.1080/15384101.2023.2281816.

Zitnik M, Nguyen F, Wang B, Leskovec J, Goldenberg A and Hoffman MM (2019) Machine Learning for Integrating Data in Biology and Medicine: Principles, Practice, and Opportunities. *Information Fusion* **50,** 71-91. https://doi.org/10.1016/j.inffus.2018.09.012.

https://doi.org/10.1017/pcm.2025.10003 Published online by Cambridge University Press